nature medicine

RESOURCE

https://doi.org/10.1038/s41591-019-0579-z

Genome-wide germline correlates of the epigenetic landscape of prostate cancer

Kathleen E. Houlahan[1,2,3], Yu-Jia Shiah[1], Alexander Gusev[4,5], Jiapei Yuan[6], Musaddeque Ahmed[7], Anamay Shetty[4,8], Susmita G. Ramanand[6], Cindy Q. Yao[1], Connor Bell[9], Edward O'Connor[9], Vincent Huang[1], Michael Fraser[1], Lawrence E. Heisler[1], Julie Livingstone[1], Takafumi N. Yamaguchi[1], Alexandre Rouette[1], Adrien Foucal[1], Shadrielle Melijah G. Espiritu[1], Ankit Sinha[2,7], Michelle Sam[1], Lee Timms[1], Jeremy Johns[1], Ada Wong[1], Alex Murison[7], Michèle Orain[10], Valérie Picard[11], Hélène Hovington[11], Alain Bergeron[11], Louis Lacombe[11], Mathieu Lupien[1,2,7], Yves Fradet[11], Bernard Têtu[10], John D. McPherson[1], Bogdan Pasaniuc[12,13,14], Thomas Kislinger[2,7], Melvin L. K. Chua[7,15], Mark M. Pomerantz[9], Theodorus van der Kwast[7,16], Matthew L. Freedman[9,17,18], Ram S. Mani[6], Housheng H. He[2,7], Robert G. Bristow[2,7,19,20,21,22,23]* and Paul C. Boutros[1,2,3,14,24,25,26,27]*

Oncogenesis is driven by germline, environmental and stochastic factors. It is unknown how these interact to produce the molecular phenotypes of tumors. We therefore quantified the influence of germline polymorphisms on the somatic epigenome of 589 localized prostate tumors. Predisposition risk loci influence a tumor's epigenome, uncovering a mechanism for cancer susceptibility. We identified and validated 1,178 loci associated with altered methylation in tumoral but not nonmalignant tissue. These tumor methylation quantitative trait loci influence chromatin structure, as well as RNA and protein abundance. One prominent tumor methylation quantitative trait locus is associated with *AKT1* expression and is predictive of relapse after definitive local therapy in both discovery and validation cohorts. These data reveal intricate crosstalk between the germ line and the epigenome of primary tumors, which may help identify germline biomarkers of aggressive disease to aid patient triage and optimize the use of more invasive or expensive diagnostic assays.

Cancer is defined by a set of deregulated cellular processes, termed hallmarks[1], which ultimately arise from genomic and epigenomic aberrations[2,3]. There are three sources for these aberrations: environmental (for example, DNA-damaging xenobiotics), stochastic (for example, replication-associated mutations) and genetic (for example, inherited predisposition polymorphisms)[4]. Genome-wide association studies (GWASs) have revealed hundreds of germline variants associated with an elevated risk of cancer diagnosis[5–7]. Furthermore, some highly penetrant polymorphisms in tumor suppressor genes, including *RB1*, *APC*, *BRCA1* and *BRCA2*, induce unique mutational phenotypes, including epigenomic dysregulation[8–10].

Epigenetic aberrations associated with chemical modification of DNA provide additional modes of tumor-specific regulation[11]. Tumors can hijack epigenetic regulatory systems to silence tumor suppressors[12,13], and large-scale rewiring of DNA methylation is common in many cancer types[14]. Susceptibility loci are enriched at regulatory regions[15,16], and these loci can modulate the tumor epigenome[17]. For example, prostate tumors arising in men with deleterious germline *BRCA2* mutations show a genome-wide reduction

---

[1]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [3]Vector Institute, Toronto, Ontario, Canada. [4]Division of Population Sciences, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA. [5]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA. [7]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [8]University of Cambridge, Cambridge, UK. [9]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [10]Department of Pathology, Centre de recheche du CHU de Québec-Université Laval, Québec City, Québec, Canada. [11]Division of Urology, Centre de recheche du CHU de Québec-Université Laval, Québec City, Québec, Canada. [12]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [13]Department of Pathology and Laboratory Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [14]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [15]Division of Radiation Oncology, National Cancer Centre Singapore, Singapore, Singapore. [16]Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. [17]The Eli and Edythe L. Broad Institute, Cambridge, MA, USA. [18]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. [19]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [20]Division of Cancer Sciences, Faculty of Biology, Health and Medicine, University of Manchester, Manchester, UK. [21]The Christie NHS Foundation Trust, Manchester, UK. [22]Cancer Research UK Manchester Institute, Manchester, UK. [23]Manchester Cancer Research Centre, Manchester, UK. [24]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [25]Department of Urology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [26]Jonsson Comprehensive Cancer Center, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [27]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA, USA. *e-mail: robert.bristow@manchester.ac.uk; pboutros@mednet.ucla.edu

in methylation relative to sporadic tumors, which may account for their increased aggressivity[18].

These data suggest that common germline polymorphisms may influence the development of aggressive prostate tumors. GWASs have failed to identify loci robustly associated with prostate cancer survival[19], perhaps due to insufficiently large cohort sizes. Since single nucleotide polymorphisms (SNPs) can confer susceptibility by modulating DNA methylation[17], we reasoned that interrogating the more direct link between the germ line and methylation would yield associations with larger effect sizes than germline survival analyses. As has already been observed in neuroscience[20], germline-methylation analyses might overcome the limitations of small cohorts and identify loci otherwise not selected at genome-wide significance levels. Prognostic germline loci would be attractive minimally invasive biomarkers to aid the early clinical stratification of indolent versus aggressive disease, and provide prior probabilities to maximize the utility of more expensive fluid, tissue or radiologic assays.

We focus on prostate cancer—the second most common malignancy in men[21], with few known risk factors[22,23] and large molecular and clinical heterogeneity[24,25]. We compare germline whole genomes and tumor methylomes of 589 patients with localized prostate cancer ($n_{\text{discovery}} = 241$; $n_{\text{validation}} = 348$) and identify 7,590 validated *cis*-methylation quantitative trait loci (meQTLs) (that is, germline loci associated with altered methylation levels). Germline variants are not unique to the tumor; therefore, we introduce a novel class of functional variants: tumor meQTLs. These are loci associated with altered methylation in tumoral but not in nonmalignant tissue (that is, they have a larger effect in tumor versus reference tissues). We identify and validate 1,178 tumor meQTLs, and show that 17 of these show tumor-specific RNA or protein abundance changes (termed tumor meQTL-expression QTLs; eQTLs). Tumor meQTLs are enriched at tumor-specific regulatory regions in prostate cell lines and primary tumors, and preferentially target sites of chromatin looping. Some tumor meQTLs target known prognostic cancer driver genes, including *TCERG1L* and *AKT1*. Indeed, the tumor meQTL targeting *AKT1* was predictive of aggressive disease in both our discovery cohort (hazard ratio (HR) = 2.85; $P = 5.8 \times 10^{-3}$) and a validation cohort of 101 clinically matched samples (HR = 2.2; $P = 1.7 \times 10^{-2}$). Taken together, these data highlight how germline genotypes can modulate the tumor epigenome to contribute to the tumorigenesis of aggressive prostate cancers. This phenomenon may apply to other tumor types, providing a strategy to create robust, minimally invasive biomarkers for the early detection of aggressive disease.

Results

Prostate cancer susceptibility loci associated with tumor methylation dysregulation. We assembled 241 patients with treatment-naïve prostate cancer who had germline whole-genome sequencing (WGS) and methylation profiling by array, including 80 new genomes and 161 from the literature[25,26]. All patients had organ-confined clinically intermediate-risk disease and were treated by image-guided radiotherapy (IGRT) or surgery. The median clinical follow-up was 8.76 years. Identity-by-state clustering did not show population stratification (Extended Data Fig. 1a). Supplementary Table 1 summarizes this discovery cohort.

We sought to quantify polymorphisms that module specific epigenetic features of tumor evolution, termed tumor meQTLs, to distinguish them from meQTLs that exert effects in normal epithelial tissue. First, we validated previous work[17], showing evidence for the association of germline risk loci with tumor methylome alterations. We validated three of these five meQTLs ($P < 0.01$; see Methods; rs10934853:cg08044714, rs17021918:cg07677047 and rs339331:cg12892004; Extended Data Fig. 2a–e and Supplementary Table 2).

Next, we comprehensively analysed 160 validated germline susceptibility loci associated with prostate cancer incidence that account for 34.4% of familial risk[15,25,27–32] (analysis 1a; Fig. 1a and Supplementary Table 2). Each risk locus was tested for methylation associations methylome wide, identifying 79 meQTLs covering 30 loci and 77 probes ($P < 7 \times 10^{-10}$; Bonferroni-adjusted threshold; Spearman's correlation; Fig. 1b and Supplementary Table 2). Of these, 75/79 associations were in *cis*: the risk locus was located proximal to the methylated site (median distance: 11.5 ± 37.7 kilobase pairs (kbp); Extended Data Fig. 2f). There were four *trans* associations: rs2238776 (chromosome 22), associated with cg11491381 (*AVP*; chromosome 20); rs4976790 (chromosome 5), associated with cg05952543 and cg20792895 (*MKRN3*; chromosome 15); and rs7295014 (chromosome 12), associated with cg26860994 (*SND1*; chromosome 7). None of the risk variants within chromosome 8q24 (a well-characterized locus proven to be a major contributor to prostate cancer risk) were identified as meQTLs[33].

To validate these candidate meQTLs, we evaluated 348 cases from The Cancer Genome Atlas (TCGA) with tumor methylation data along with germline SNP array and whole-exome sequencing (WES) of blood samples (analysis 2; Fig. 1a)[34]. Following a stringent quality control and imputation process (Extended Data Fig. 1b), we estimated our SNP detection accuracy in this validation cohort to be 98.8% (Extended Data Fig. 1c–e). After imputation, 69/79 risk loci meQTLs were genotyped in the validation cohort, and 55/69 validated (23 loci and 55 methylation probes; false discovery rate (FDR) < 0.05; Spearman's correlation; Fig. 1c and Supplementary Table 2). Three *trans* associations—rs2238776-cg11491381, rs4976790-cg05952543 and rs4976790-cg20792895—replicated in this independent validation dataset ($\text{FDR} < 2.22 \times 10^{-2}$; Spearman's $|\rho| > 0.13$; Supplementary Table 2). Thus, 14% of known prostate cancer risk loci may influence risk by modulating tumor methylation.

We quantified the enrichment of validated meQTLs in transcription factor binding sites, or chromatin marks defined in four prostate cancer cell lines (LNCaP, PC3, 22Rv1 and VCaP) and one prostate epithelial cell line (RWPE-1) using chromatin immunoprecipitation sequencing (ChIP-Seq) data (see Methods and Supplementary Table 3). Risk meQTLs were enriched at active regulatory regions, including H3K27ac and H3K4me3 modifications indicative of active promoters, in all cell lines (FDR < 0.05; permutation test; $n = 10^5$; Extended Data Fig. 2g–k). We confirmed the enrichment of risk meQTLs at regulatory regions in 94 primary prostate samples[35]. Ten out of 23 loci overlapped at least one of the androgen receptor, H3K27ac, H3K4me3 or H3K27me3 sites in at least one patient ($P = 1.51 \times 10^{-3}$; permutation test; $n = 10^5$; Extended Data Fig. 2l). There was allele-specific H3K27ac modification at rs1983891 ($\beta = 0.72$; $P = 0.05$; logistic regression; Supplementary Table 4). Validated meQTL methylation targets were enriched in CpG islands on chromosomes 5 and 6 (odds ratio (OR) > 2.3; $\text{FDR} \leq 8.43 \times 10^{-3}$; Fisher's Exact Test; Supplementary Table 4). To distinguish tumor meQTLs from meQTLs (loci that affect methylation in prostate epithelial tissue), we considered 47 tumor-adjacent prostate samples from the TCGA (analysis 3; Fig. 1a). Each tumor-adjacent sample (henceforth referred to as reference) was confirmed to be morphologically normal by pathology review, and had no detectable prostate cancer mutations[34]. Of these 55 validated meQTLs, 52 were evaluated in reference tissue and 14 were tumor specific (defined as FDR > 0.05 in reference tissue and FDR < 0.05 in matched tumor tissue ($n = 47$); Spearman's correlation; Supplementary Table 2). Only 3/14 tumor meQTLs were proximal to a gene (within 1,500 base pairs (bp)), and none were significantly associated with messenger RNA (mRNA) changes (FDR < 0.05; Spearman's correlation; Supplementary Table 2). Finally, we identified meQTLs missed in the discovery cohort by conducting meQTL discovery in the TCGA cohort. We discovered 165 meQTLs (32 loci and 144 probes) in the TCGA cohort, of which 32 novel meQTLs were validated in

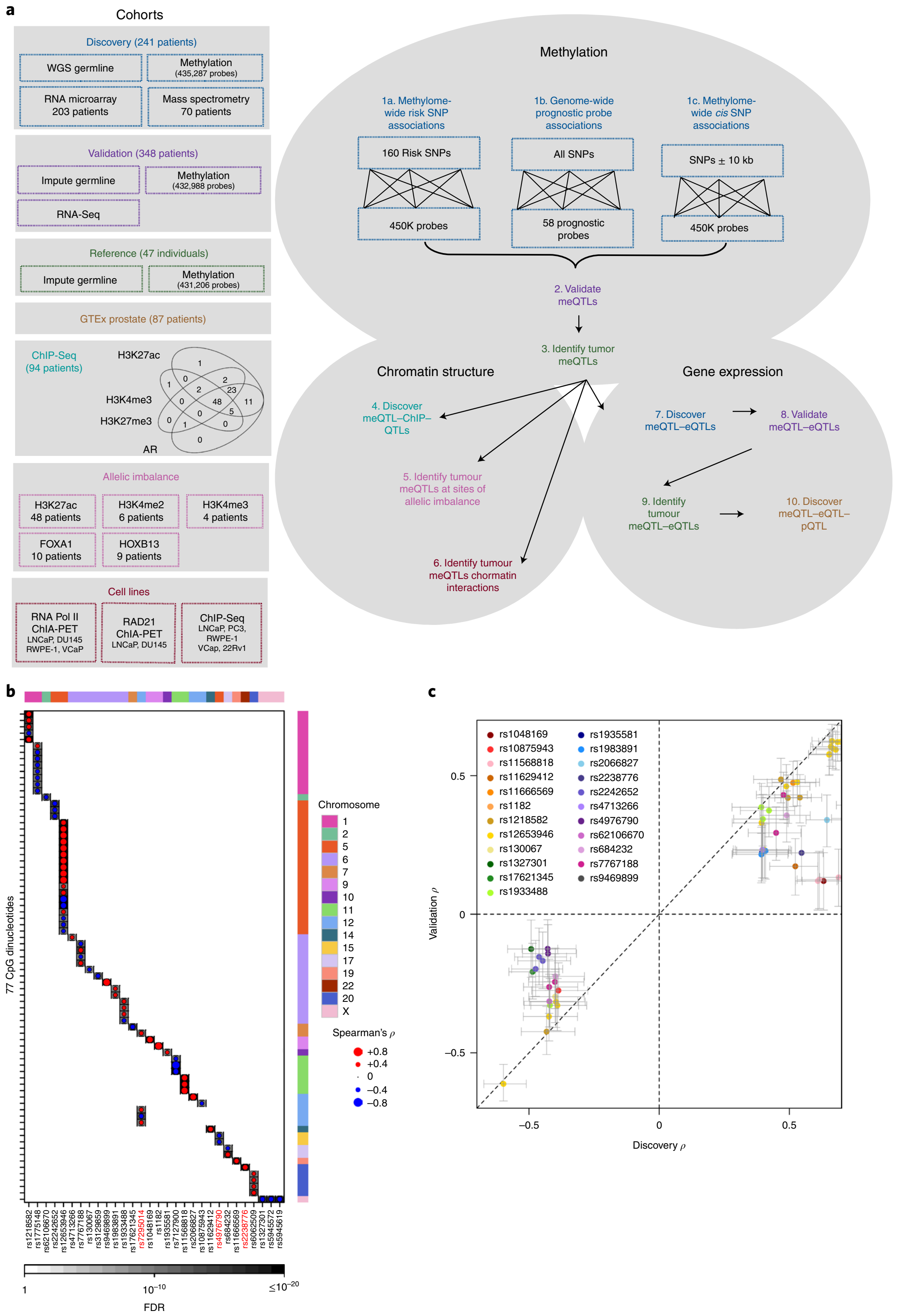


Fig. 1 | Prostate cancer susceptibility loci associated with tumor methylation dysregulation. a, Schematic of the datasets and workflow. The boxes along the left and their corresponding colors indicate the dataset used at each step. AR, androgen receptor. **b**, 30 risk loci (*x* axis) were associated with 77 methylation probes (*y* axis). Dot sizes represent Spearman's ρ magnitude, while colors indicate directionality. Background shading represents the FDR. The covariates along the top represent the chromosome of each SNP, while the covariates along the right indicate the chromosome of each methylation site. Red labels indicate SNPs involved in a *trans* meQTL. **c**, Comparison of Spearman's ρ in discovery and validation cohorts for 23 risk loci significantly associated (FDR < 0.05) with 55 methylation probes in the validation cohort. Error bars represent 95% confidence interval. The diagonal dotted line represents $y = x$.

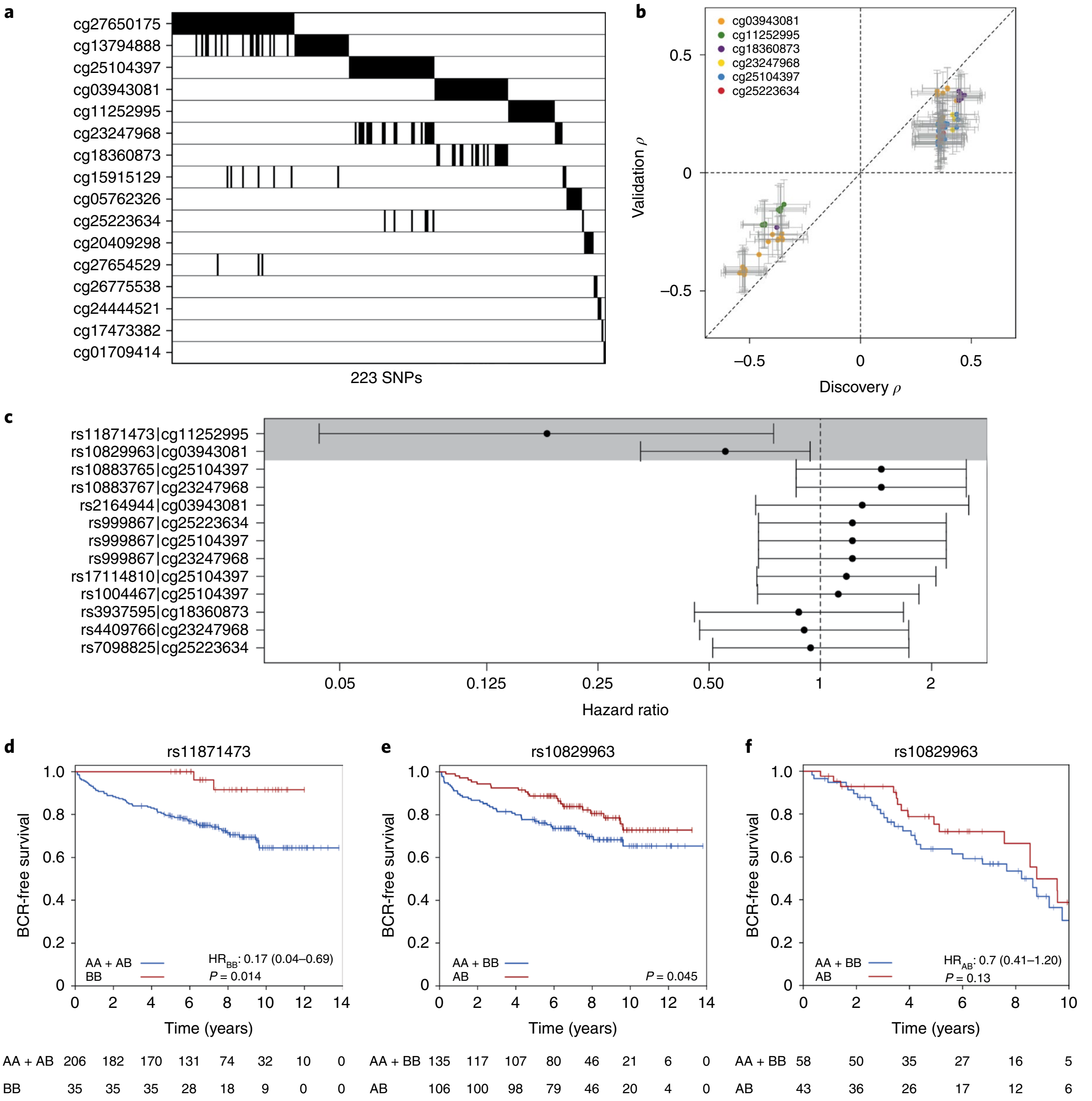
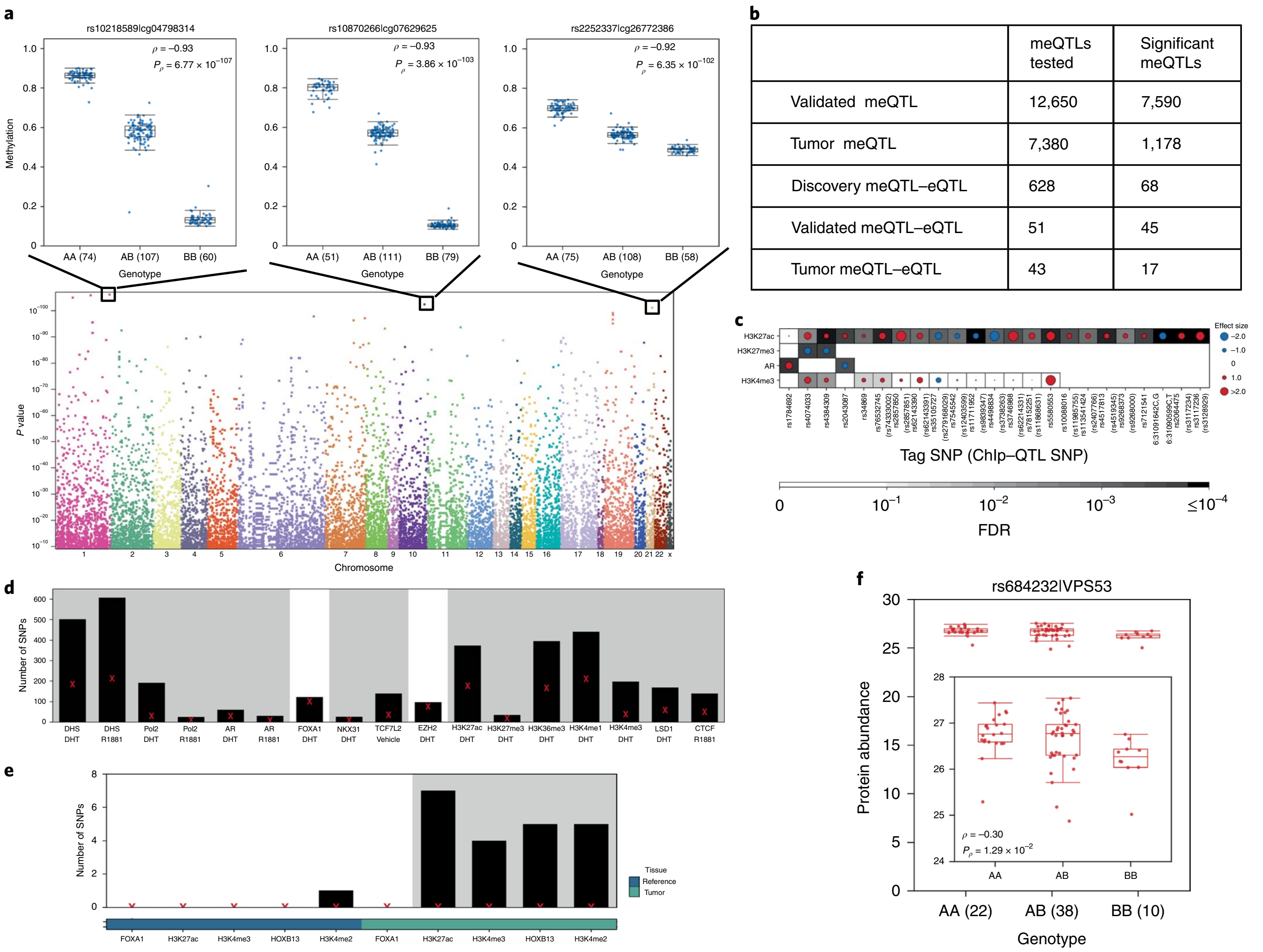


Fig. 2 | Germline variants associate with prognostic methylation levels. a, Summary of 223 significant SNPs (*x* axis) for each prognostic methylation probe (*y* axis). Black indicates that a SNP is significantly associated ($P<5\times10^{-8}$) with methylation levels at that probe. **b**, Comparison of Spearman's ρ in discovery and validation cohorts for 75 SNPs that were significantly associated (FDR < 0.05) with six methylation probes in the validation cohort. Error bars represent 95% confidence interval. The diagonal dotted line represents $y=x$. **c**, Two *cis*-meQTLs were prognostic. Dots and error bars represent HRs and 95% CIs, respectively, for the tag SNP from each of the seven haplotypes. The dotted line indicates HR = 1 and gray background shading indicates $P\leq0.05$. **d**, The homozygous alternative genotype of a haplotype on chromosome 17, associated with methylation of *ATP2A3*, gives a survival advantage. HR and *P* values are from the CoxPH model. **e**, A haplotype on chromosome 10, associated with methylation of *TCERG1L*, is co-dominantly associated with BCR. **f**, Co-dominant association with BCR replicated at rs10829963 in an independent cohort ($n=101$).

the discovery cohort (18 loci and 30 probes; FDR < 0.05; Spearman's correlation; Supplementary Table 2). These results expand our understanding of the role of risk loci in modulating tumor methylation, and suggest that we are probably underestimating the extent of this modulation.

Germline variants associate with prognostic methylation levels. On validating risk locus tumor meQTLs, we discovered novel locus candidates by identifying tumor meQTLs associated genome wide with tumor aggressivity. Germline loci that could delineate indolent from aggressive disease would provide a minimally invasive, early-detection biomarker filling an important clinical gap. We selected 58 methylation probes based on their association with biochemical recurrence (BCR), as defined by increasing prostate-specific antigen (PSA) levels following primary treatment—a trigger of salvage therapy and, when occurring within 18 months of primary treatment, a surrogate for prostate cancer-specific mortality[36] (Extended Data Figs. 1f and. 3a and Supplementary Table 5). We identified genome-wide candidate loci for each of the 58 prognostic methylation probes (analysis 1b; Fig. 1a) and discovered 292 meQTLs targeting 28% of these probes (16/58), covering 223 distinct loci ($P<5\times10^{-8}$; Spearman's correlation; Fig. 2a). For each of these loci, the presence of one or more alternative alleles was associated with significant changes in methylation.



b

	meQTLs tested	Significant meQTLs
Validated meQTL	12,650	7,590
Tumor meQTL	7,380	1,178
Discovery meQTL–eQTL	628	68
Validated meQTL–eQTL	51	45
Tumor meQTL–eQTL	43	17

Fig. 3 | Landscape of *cis*-tumor meQTLs. **a**, Bottom: identification of *cis*-tumor meQTLs methylome wide ($P < 3 \times 10^{-9}$; Bonferroni adjustment). Each point represents a SNP, ordered by chromosome along the *x* axis. The *y* axis gives *P* values from Spearman's correlation. Top: representative box plots, showing methylation (*y* axis) discretized by genotype (*x* axis), for three germline-methylation associations. Box plots represent median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the interquartile (IQR) range. Blue points represent methylation values. Numbers of samples with each genotype are given in parentheses on the *x* axes. **b**, Numbers of meQTLs tested and statistically significant (FDR < 0.05, based on Spearman's correlation) at each stage. **c**, Tumor meQTLs showed allele-specific androgen receptor binding or histone modification ($n = 30$; 23 unique SNPs). Circle size and color represent the magnitude and sign of the coefficient from the logistic regression model, respectively (that is, red indicates an alternate allele associated with increased binding, while blue indicates decreased binding). The background shading represents the FDR. The *x* axis labels show tag tumor meQTL-ChIP-QTL IDs, and SNP IDs in parentheses indicate the ChIP-QTL SNP in the case that the ChIP-QTL SNP is not the tag SNP. **d**, Tumor meQTLs were enriched at transcription factor binding sites and active regulatory elements in LNCaP cells. The *y* axis shows the number of tumor meQTLs that overlap each target/treatment pair. **e**, Tumor meQTLs were interrogated for overlap with sites of allelic imbalance at FOXA1, H3K27ac, H3K4me3, HOXB13 and H3K4me2 peaks in tumoral and reference tissue. The *y* axis shows the number of tumor meQTLs that overlap each target. The bottom covariate indicates sites of allelic imbalance in tumoral or reference tissue. In **d** and **e**, background shading indicates an FDR < 0.05 from permutation analysis ($n = 10^5$ permutations), and each red × represents the expected number of overlapping loci by chance. **f**, The tumor meQTL-eQTL-pQTL identified in this analysis is in linkage disequilibrium with the risk locus, rs684232, which is also a pQTL for VPS53. Box plots represent median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR range. Red points represent protein abundance values. Numbers of samples with each genotype are given in parentheses on the *x* axis.

The TCGA dataset[34] was used to validate these prognostic meQTLs, providing genotype information for 151/292 loci (analysis 2; Fig. 1a). Of these, 113/151 meQTLs were validated, representing seven *cis*-haplotypes covering six methylation probes (FDR < 0.05; Fig. 2b and Supplementary Table 6). These included 35 loci associated with methylation of cg25104397 ($\text{FDR} \leq 4.92 \times 10^{-2}$; Spearman's $|\rho| = 0.11–0.25$), 17 associated with cg23247968 ($\text{FDR} \leq 1.24 \times 10^{-3}$; $|\rho| = 0.18–0.25$) and six associated with cg25223634 ($\text{FDR} \leq 2.02 \times 10^{-2}$; $|\rho| = 0.13–0.20$; Supplementary Table 6). These three probes are located within 41 bp on chromosome 10, within an open sea region of *C10orf26*. Their methylation was highly correlated, and six loci were associated with all three sites (Extended Data Fig. 3b,c). We used paired tumor/reference samples to determine whether these meQTLs were tumor specific (that is, $\text{FDR}_{\text{tumor}} < 0.05$ and $\text{FDR}_{\text{reference}} > 0.05$; Spearman's correlation; analysis 3; Fig. 1a) and identified 38 tumor meQTLs ($\text{FDR}_{\text{reference}} > 0.25$), all of which target methylation at two sites: cg18360873 and cg03943081 (Supplementary Table 6). These methylation probes are located 5′ and 3′, respectively, of *TCERG1L*—an epigenetic driver event in prostate cancer[26].

Given the prognostic value of the methylation sites targeted by these meQTLs, we evaluated the prognostic value of the loci themselves. As seen earlier, not all risk meQTLs were tumor specific, suggesting that meQTLs that have a role in reference tissue are also biologically important. Therefore, we considered tag SNPs for all seven haplotypes involved in validated meQTLs. Two *cis*-meQTLs were predictive of BCR (HR = 0.554 and 0.180; $P = 2.92 \times 10^{-2}$ and 1.73×10^{-2}; Cox proportional hazards (CoxPH) model; Fig. 2c–e), and one (rs10829963) showed the same survival trend in an independent cohort[26,27,37,38] of 101 clinically matched patients (HR = 0.70; $P = 0.13$; Fig. 2f). The validation cohort was insufficiently powered to test rs11871473 ($n_{BB} = 11$; $1 - \beta = 0.44$). Taken together, these results suggest that the germ line may shape tumor aggressivity via tumor methylation dysregulation.

Landscape of *cis*-tumor meQTLs. All validated tumor meQTLs were in *cis* associations (that is, the loci were within 59,151 bp of the methylation site). To quantify *cis*-tumor meQTL frequency in prostate cancer, we systematically evaluated loci within a 10-kbp window around each of the 434,504 methylation probes (analysis 1c; Fig. 1a). We identified 169,562 loci associated with the methylation status of 3.3% of all CpGs quantified (14,287 distinct probes; $P < 3 \times 10^{-9}$; representing 1.5×10^{7} independent tests; Spearman's correlation; Fig. 3a). These associations are not driven by variants affecting the hybridization of probes on the methylation array (Extended Data Fig. 1g,h).

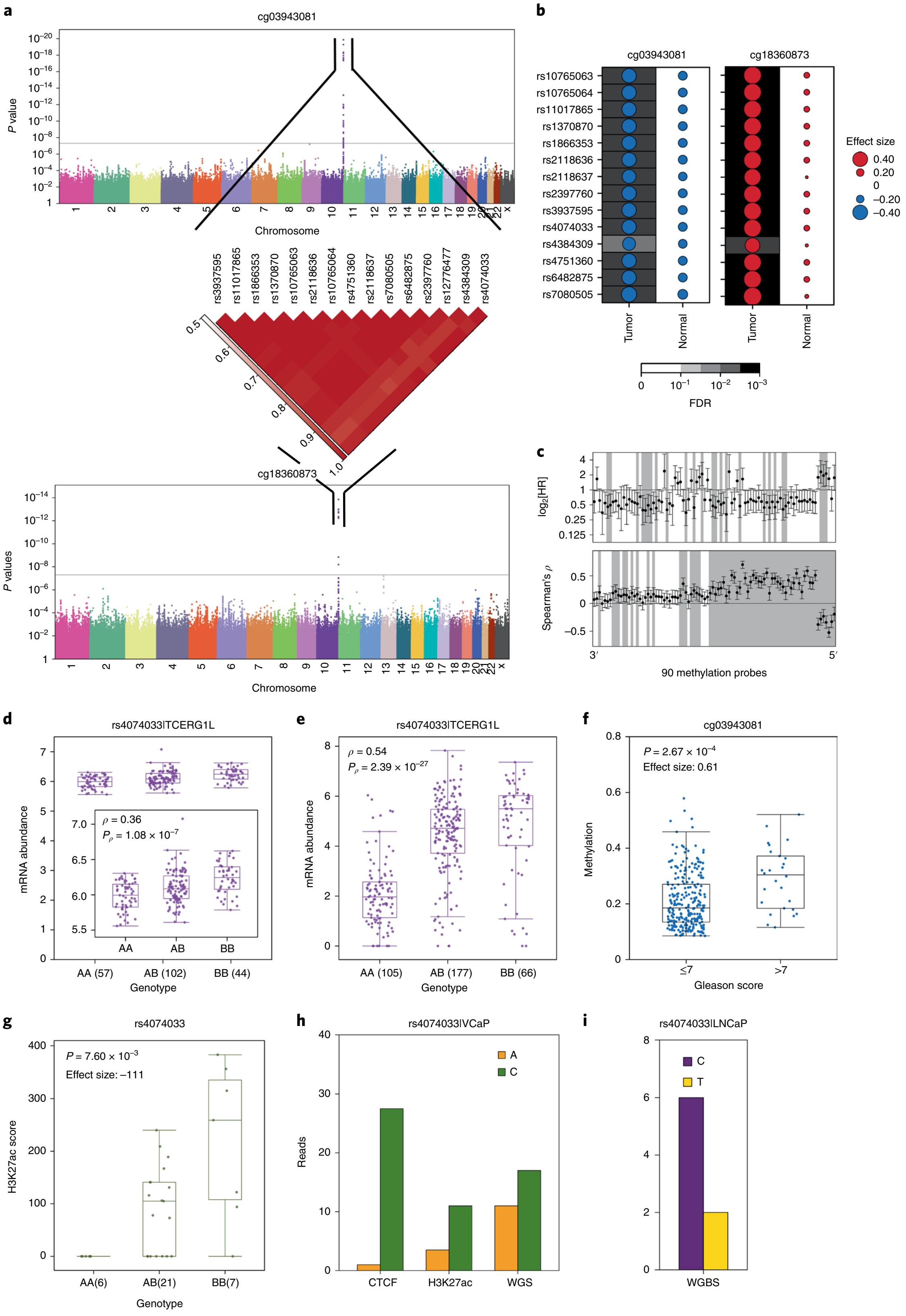
We validated the locus with the lowest *P* value for each probe, provided it was genotyped on the TCGA platform (12,650 loci; analysis 2; Fig. 1a and Supplementary Table 7), and 7,590/12,650 (60%) *cis*-meQTLs were validated in this independent cohort (FDR < 0.05; Spearman's correlation; Fig. 3b). Of the 7,590 validated meQTLs, 7,380 had genotype and methylation data for the 47 reference prostate samples[34] (analysis 3; Fig. 1a). One-third (1,178/7,380) were consistent with tumor meQTLs, meaning that they had associations in tumor tissue ($\mathrm{FDR}_{\mathrm{tumor}} < 0.05$ in matched tumor samples) but not reference prostate epithelial tissue ($\mathrm{FDR}_{\mathrm{reference}} \geq 0.05$), or that they had opposite effects in tumor and reference tissue (234 meQTLs; $\mathrm{sign}(\rho_{\mathrm{tumor}}) \neq \mathrm{sign}(\rho_{\mathrm{reference}})$; Fig. 3b and Extended Data Fig. 4a). Almost half (546/1,178) of these tumor meQTLs were differentially methylated between tumoral and reference tissue, suggesting that at least a subset of tumor meQTLs target dysregulated methylation sites (Extended Data Fig. 4b). These probes were enriched for open seas in intergenic regions on chromosome 6 (OR > 1.07; $\mathrm{FDR} \leq 3.10 \times 10^{-2}$; Fisher's exact test; Supplementary Table 8). In contrast, CpG islands in promoter regions on chromosome X were significantly depleted of validated associations (OR < 0.86; $\mathrm{FDR} \leq 2.19 \times 10^{-3}$; Fisher's exact test; Supplementary Table 8). The depletion on chromosome X may result from less accurate imputation on this chromosome (Extended Data Fig. 1e).

Tumor meQTLs target active regulatory regions and sites of chromatin looping. To determine whether tumor meQTLs influence transcription regulation, we quantified the enrichment of tumor meQTLs at regulatory regions in four prostate cancer cell lines (LNCaP, PC3, VCaP and 22Rv1) and one prostate epithelial cell line (RWPE-1) (Supplementary Table 3). Tumor meQTLs were enriched at androgen receptor and CTCF bindings sites in LNCaP, as well as active enhancers and promoters, as seen by enrichment at H3K27ac, H3K4me1 and H3K4me3 marks ($\mathrm{FDR} < 1 \times 10^{-26}$; permutation test; $n = 10^{5}$; Fig. 3d). Enrichment at regulatory regions was replicated in PC3, VCaP, 22Rv1 and RWPE-1 cell lines (Extended Data Fig. 4c–f). Tumor meQTLs were more strongly enriched at the repressive chromatin mark H3K27me3 in the epithelial cell line RWPE-1 ($\mathrm{FDR} < 1 \times 10^{-26}$; Extended Data Fig. 4c) than the cancer cell lines PC3 (FDR = 0.36; Extended Data Fig. 4d) or LNCaP (FDR = 0.006; Fig. 4d), suggesting that a subset of these sites may activate during tumorigenesis. To confirm this enrichment at active regulatory regions, we considered H3K27ac, H3K27me3, H3K4me3 and androgen receptor ChIP-Seq data from 94 primary prostate cancer samples[35] (analysis 4; Fig. 1a). Tumor meQTLs were significantly enriched at H3K27ac and H3K4me3 sites in all primary samples and at androgen receptor binding sites in 84% of samples (FDR < 0.05; permutation test; $n = 10^{5}$; Extended Data Fig. 4g). To identify specific tumor meQTLs modulating chromatin structure, we tested for allele-specific androgen receptor binding and H3K27ac, H3K27me3 and H3K4me3 histone modifications (that is, ChIP–QTLs) in primary prostate cancer samples[35] (analysis 4; Fig. 1a). We discovered 30 tumor meQTL–ChIP–QTLs (at 23 unique loci) targeting one of the four marks (FDR < 0.05; H3K27ac = 22; H3K27me3 = 2; androgen receptor = 2; and H3K4me3 = 4; logistic regression; Fig. 3c). The variant rs2043087 is located within *ALDH1A2* (a prostate cancer tumor suppressor[39]) and is associated with increased H3K27ac ($\beta = 1.27$; $\mathrm{FDR} = 6.04 \times 10^{-3}$) but decreased androgen receptor binding ($\beta = -1.49$; $\mathrm{FDR} = 9.00 \times 10^{-3}$).

We further characterized a high-confidence subset of 59 tumor meQTLs associated with BCR ($P < 0.05$; CoxPH model; Supplementary Table 7). To support the tumor-specific role of tumor meQTLs in modulating protein–DNA interactions, we identified sites of allelic imbalance in transcription factor binding and histone modification genome wide in paired tumor and reference

Fig. 4 | Tumor meQTL associated with *TCERG1L* regulation. a, A haplotype on chromosome 10 was strongly associated with methylation probes at both the 5′ and 3′ ends of *TCERG1L*. The Manhattan plots present *P* values (*y* axes; Spearman's correlation) for the association of each SNP. The *x* axes are ordered by chromosome, with methylation at both the 5′ (cg03943081) and 3′ (cg18360873) ends of *TCERG1L*. The gray lines represent the Bonferroni adjustment. All associated SNPs were in strong linkage disequilibrium. A linkage disequilibrium plot shows pairwise D′ values between all associated SNPs. **b**, *TCERG1L* meQTLs were tumor specific. Dot size and color reflect the magnitude and directionality, respectively of Spearman's ρ between genotype and methylation at probes 5′ and 3′ of *TCERG1L*. The background shading indicates the FDR. **c**, The genotype at rs4074033 was associated with methylation levels of 64/90 probes spanning *TCERG1L*. Bottom: forest plot showing the Spearman's correlations and 95% CIs for association of each methylation probe (*x* axis) and genotype at rs4074033. The horizontal line represents $\rho = 0$. Top: forest plot showing the $\log_2$[HR] values and 95% CIs for association of methylation probes with BCR using a CoxPH model. The horizontal line represents HR = 1. Gray shading in both plots indicates significant association (FDR < 0.05). **d,e**, The alternative (B) allele exhibits a dominant effect resulting in increased mRNA abundance of *TCERG1L* in both the discovery (**d**) and validation cohort (**e**). Purple points represent mRNA abundance values. Numbers of samples with each genotype are given in parentheses. Difference in abundance levels was quantified using Spearman's correlation. **f**, Methylation 5′ of *TCERG1L* (cg03943081) was significantly associated with Gleason score in the discovery cohort. The effect was quantified by Mann–Whitney *U*-test, and the effect size represents the fold change. Blue points represent methylation levels. **g**, The alternative allele showed increased H3K27ac modification in the discovery cohort. The effect was quantified by Mann–Whitney *U*-test (AA versus AB + BB), and the effect size represents differences in medians. Green points represent H3K27ac peak signals. **h**, The alternative allele at rs4074033 preferentially shows H3K27ac modification and is preferentially bound by CTCF. The VCaP cell line is heterozygous at rs4074033 (that is, genotype AC). The *y* axis shows the number of reads with each allele at rs4074033 from CTCF and H3K27c ChIP-Seq and WGS data. **i**, The alternative allele at rs4074033 is methylated in LNCaP cell lines (genotype: CC). The *y* axis shows the number of reads with the methylated C allele versus the unmethylated T allele (from whole-genome bisulfite sequencing). Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5× IQR.

samples (analysis 5; Fig. 1a). Sites of allelic imbalance reflect loci with high regulatory potential. Specifically, we discovered sites of allelic imbalance in tumor and reference samples for FOXA1, HOXB13, H3K27ac, H3K4me3 and H3K4me2. We observed a strong enrichment of tumor meQTLs at H3K27ac, H3K4me3, HOXB13 and H3K4me2 sites in tumor ($FDR < 0.01$; permutation test; $n = 10^5$) but not reference samples ($FDR > 0.19$), supporting their tumor-specific role (Fig. 3e, Extended Data Fig. 4h and Supplementary Table 7). Next, we explored the impact of tumor meQTLs on chromatin structure, specifically RAD21 and RNA



polymerase II (RNA Pol-II) chromatin loops in LNCaP, DU145, VCaP and RWEP-1 cell lines (analysis 6; Fig. 3a). Fourteen tumor meQTLs overlapped with RNA Pol-II peaks in at least one cell line, and most (12/14) were involved in chromatin looping (Extended Data Fig. 4i and Supplementary Table 7). Eleven overlapped with RAD21 binding sites and 9/11 were involved in chromatin loops. Seven overlapped with RNA Pol-II and RAD21 sites, suggesting that these tumor meQTLs are targeting active enhancer–promoter interactions. These results show that tumor meQTLs preferentially target *cis*-regulatory elements in a tumor-specific manner. Tumor meQTL mechanisms are probably myriad, including disrupting androgen receptor binding (rs1784692 or rs2043087), deregulating RNA Pol-II looping (rs3747623 or rs1867529) and others not recognized in this first study.

Tumor meQTLs drive an aggressive gene expression program. DNA methylation can directly dysregulate transcription, thus we quantified tumor meQTLs' modulation of the transcriptome (microarray profiling) of 203 patients in the discovery cohort. We focused on validated tumor meQTLs with methylation sites proximal to a gene (within 1,500 bp; 628 associations; analysis 7; Fig. 1a). We identified 68 tumor meQTLs associated with mRNA abundance in the discovery cohort (termed tumor meQTL–eQTLs; FDR < 0.05; Spearman's $|\rho| = 0.20–0.55$), of which 45 also associated with mRNA abundances in the TCGA validation cohort (analysis 8; FDR < 0.05; $|\rho| = 0.11–0.75$; Fig. 3b and Supplementary Table 7). Utilizing RAD21 and RNA Pol-II chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) profiling of prostate cancer cell lines, we identified additional targets for 17 tumor meQTLs (distance between locus and target: 0–148.5 megabase pairs; median = 13.9 megabase pairs), four of which were significantly associated with mRNA abundance of five transcripts (Extended Data Fig. 4j). We discovered a significant tumor meQTL–eQTL association targeting *MINCR*—a MYC-induced long non-coding RNA that has been implicated in Burkitt lymphoma and gallbladder cancer[40,41]. Only three of these eQTLs could be tested in TCGA; two out of three validated (rs2456274:FAM57A and rs1225741:ELOVL2; FDR < 0.05; Extended Data Fig. 4j), and one was previously reported[15]. We confirmed that 17/43 tumor meQTL–eQTLs were tumor specific at the RNA level using prostate epithelial eQTL statistics from the Genotype-Tissue Expression (GTEx) project[42] (analysis 9; FDR > 0.05; Fig. 3b, Table 1 and Supplementary Table 7). These 17 were not enriched in any specific pathway; however, 6/10 genes involved in these tumor meQTL–eQTLs were differentially abundant in tumor versus reference tissue (FDR < 0.05; Extended Data Fig. 4k).

As an exploratory analysis, we tested whether the ten genes in these 17 tumor meQTL–eQTLs were dysregulated at the protein level (analysis 10; Fig. 1a). We exploited a dataset of 70 tumors with mass spectrometric quantitation of protein abundances[38]. Only 3/10 transcripts had their protein abundances quantified, and the small sample size led to very low statistical power ($1-\beta < 0.39$). Nevertheless, vacuolar protein sorting-associated protein 53 homolog (VPS53) was a strong tumor meQTL ($P = 6.95 \times 10^{-12}$; $\rho = -0.42$) that associated with both RNA (FDR = 8.22×10^{-3}; $\rho = -0.25$) and, more modestly, protein abundances ($P = 4.27 \times 10^{-2}$; $\rho = -0.24$; Extended Data Fig. 4l,m). This tumor meQTL–eQTL–protein QTL (pQTL) is of particular interest because rs2456274 is in linkage disequilibrium with the risk locus rs684232 ($D' = 1$; $P = 1.29 \times 10^{-2}$; $\rho = -0.30$; Fig. 3f), which has been reported as an eQTL for *VPS53* (ref. [15]). Thus, the discovery of tumor meQTLs has recapitulated a known risk locus, confirming the value of this approach in identifying novel susceptibility loci.

Table 1 | Significant tumor meQTL-eQTLs

SNP	Methylation probe	Gene
rs1225741	cg13351621	*SYCP2L*
rs16934152	cg13558087	*POLR1E*
rs2456274	cg08881796	*VPS53*
rs2570972	cg08367326	*AMIGO1*
rs3761188	cg09328228	*PABPC1L*
rs3761188	cg15588266	*PABPC1L*
rs3764509	cg14963724	*CNDP2*
rs3807032	cg24330456	*RNF39*
rs3807033	cg05563515	*RNF39*
rs3807033	cg17322683	*RNF39*
rs3807033	cg23793213	*RNF39*
rs3849767	cg18264728	*DAB2*
rs4147470	cg03997398	*ABLIM3*
rs4147470	cg04669407	*ABLIM3*
rs9261309	cg13918754	*RNF39*
rs9261309	cg20249327	*RNF39*
rs9295763	cg20249327	*ELOVL2*

Tumor meQTL associated with *TCERG1L* regulation. To further characterize novel loci of interest, we focused on tumor meQTLs targeting prognostic methylation sites within and 5′ to *TCERG1L* (that is, identified in analysis 1b; Fig. 1a). *TCERG1L* was previously identified as a strong epigenetic driver of aggressive prostate cancer (HR = 2.90; 95% confidence interval (CI): 1.30–6.30; $P = 0.007$; $n = 130$)[26] and its paralog, *TCERG1*, is recurrently mutated in prostate cancer[43]. Furthermore, *TCERG1L* promoter hyper-methylation has been reported in colorectal cancer[44,45]. In the discovery cohort, methylation of *TCERG1L* was strongly associated with a 15-locus region on chromosome 10q26.3 adjacent to and inside of its gene body ($P < 4.35 \times 10^{-9}$; Spearman's $|\rho| = 0.42–0.58$; Fig. 4a). These loci were in strong linkage disequilibrium and were associated with both the 5′ and 3′ probes even when correcting for tumor cellularity (Extended Data Fig. 5a,b). The haplotype had opposite effects on the 5′ and 3′ probe (that is, the alternative allele was associated with decreased methylation at the 5′ probe but increased methylation at the 3′ probe) (Extended Data Fig. 5a). Concordantly, methylation at these two probes was anti-correlated and had opposing effects on patient outcome (Extended Data Fig. 5c–e). The *TCERG1L* meQTL was confirmed to be tumor specific at the 3′ and 5′ probes ($\mathrm{FDR}_{\mathrm{reference}} > 0.14$; Spearman's $|\rho| = 0.08–0.27$; permutation $P = 0.11$; see Methods, Fig. 4b and Extended Data Fig. 5f).

To further interrogate the *TCERG1L* tumor meQTL, we assessed the methylation profile of 90 probes spanning *TCERG1L*. Methylation of 64/90 probes was significantly associated with the tag SNP rs4074033 (Fig. 4c and Supplementary Table 9), and 25/90 were associated with BCR (FDR < 0.05; CoxPH model), expanding *TCERG1L* methylation from an epigenetic driver[26] to a tumor meQTL driver.

Additionally, tumor meQTLs in *TCERG1L* correlated with mRNA abundance: the non-reference allele was dominantly associated with increased *TCERG1L* mRNA in our discovery cohort and the TCGA validation cohort ($P = 2.67 \times 10^{-8}$ and 4.53×10^{-26}, respectively; Mann–Whitney *U*-test; effect size = −0.38 and −2.87, respectively; Fig. 4d,e). While rs4074033 was identified as a tumor-specific meQTL, it was significantly associated with *TCERG1L* mRNA abundance in reference tissue—an association also observed in GTEx[42] (Extended Data Fig. 5g). Out of genotype, tumor methylation and tumor mRNA abundance, tumor methylation was the strongest prognostic measure (HR = 1.68; 95% CI = 1.01–2.78;

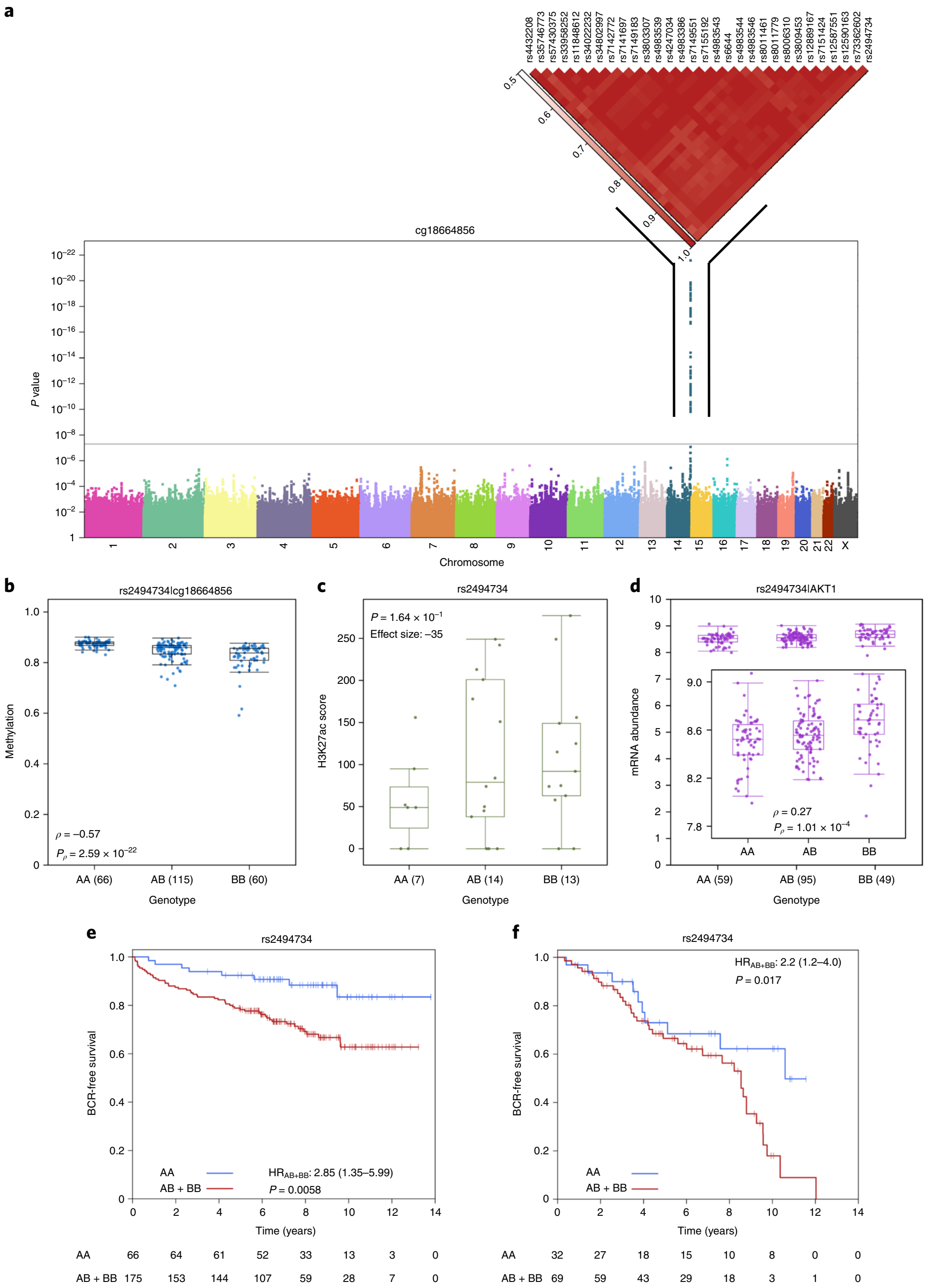


Fig. 5 | Tumor meQTL associated with *AKT1* regulation. a, A haplotype on chromosome 14 was strongly associated with methylation of a probe within the gene body of *AKT1*. The Manhattan plot represents *P* values from Spearman's correlation, as outlined in Fig. 4a. **b**, The alternative allele was associated with decreased methylation of cg18774856. This was effect quantified by Spearman's correlation. Box plots represent median values and 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR range. Blue points represent methylation values. Numbers of samples with each genotype are given in parentheses. **c**, The alternative allele showed increased H3K27ac modification in this region. This was effect quantified by Mann–Whitney *U*-test (AA versus AB + BB). The effect size gives the difference in medians. Green points represent H3K27ac ChIP-Seq signals. **d**, The alternative allele was associated with increased mRNA abundance of *AKT1*. The effect was quantified by Spearman's correlation. Purple points represent mRNA abundance. **e**, The presence of the alternative allele confers a survival disadvantage, as presented in a Kaplan–Meier plot with time along the *x* axis and the estimated proportion of individuals without a BCR event on *y* axis. The HR from a CoxPH model is also presented, along with the number of individuals without an event in each group at each time point (bottom). **f**, The alternative allele at rs2456274 was dominantly associated with rapid BCR in an independent cohort ($n = 101$).

$P = 0.05$; Extended Data Fig. 5h), concordant with the literature[26], suggesting that tumorigenic dysregulation is targeted at methylation. Methylation is also significantly associated with Gleason score in the discovery and validation cohorts (fold $\text{change}_{\text{discovery}} = 0.61$; $P_{\text{discovery}} = 2.67 \times 10^{-4}$; fold $\text{change}_{\text{validation}} = 0.87$; $P_{\text{validation}} = 1.14 \times 10^{-2}$; Mann–Whitney *U*-test; Fig. 4f and Extended Data Fig. 5i).

Next, we evaluated the effect of *TCERG1L* germline-dependent tumor methylation on chromatin organization—specifically, H3K27ac modifications[46]. In agreement with the mRNA abundance data, three SNPs within the haplotype (rs12776477, rs4384309 and rs4074033) were located within 100 bp of a H3K27ac peak, and the alternative allele dominantly increased the peak score ($\text{median}_{AA} - \text{median}_{AB+BB} = -111$; Mann–Whitney *U*-test; $P = 7.60 \times 10^{-3}$; Fig. 4g and Extended Data Fig. 5j). As further confirmation, H3K27ac modification was negatively correlated with 5′ methylation of the gene (Spearman's $\rho = -0.60$; $P = 2.65 \times 10^{-4}$; Extended Data Fig. 5k), and was replicated in an independent cohort[35] ($\beta = 1.82$; $P = 7.65 \times 10^{-3}$; logistic regression). The alternative allele was significantly associated with decreased H3K27me3 modifications ($\beta = -1.72$; $P = 3.40 \times 10^{-4}$) and increased H3K4me3 modifications ($\beta = 1.66$; $P = 7.51 \times 10^{-4}$) (Fig. 3c). Finally, across eight cell lines, only cell lines with at least one alternative allele showed CTCF binding (Extended Data Fig. 5l). In VCaP prostate cancer cells, which are heterozygous at rs4074033, the alternative allele was preferentially bound by CTCF and preferentially subject to H3K27ac modification (Fig. 4h). The two alleles of rs4074033 differ by an A–C transversion, with the C allele harboring a CpG not present in the A allele. This CpG is methylated in LNCaP cell lines (Fig. 4i). This methylation is consistent with differential CTCF binding, which is associated with altered poly(ADP-ribose) polymerase 1 activity and subsequently DNA (cystosine-5)-methyltransferase 1 activity[47]. Taken together, these data show that germline loci in *TCERG1L* may influence the methylation and chromatin organization of the gene via CTCF binding in a tumor-specific manner, supporting reports of *TCERG1L* as an epigenetic driver of aggressive prostate cancer[26].

Tumor meQTL associated with *AKT1* regulation. Next, we screened other driver genes that account for prostate cancer aggression and observed a similar link between germline loci, tumor methylation and histone organization for *AKT1*, which with *MYCN* is sufficient to transform prostate epithelial cells into adenocarcinomas[48], and is associated with an elevated risk of prostate cancer incidence[49–51]. We discovered an association between a 30-locus haplotype both 5′ and spanning into the oncogene *AKT1* on chromosome 14 and a methylation probe within a CpG island in the gene body (cg18664856; Fig. 5a). The alternative allele additively decreased the methylation of this probe, which was quantified using the tag SNP rs2494734 (Spearman's $\rho = -0.57$; $P = 2.59 \times 10^{-22}$; Fig. 5b). The meQTL was robust to correction for tumor cellularity, was validated in the TCGA cohort, and was tumor specific (Spearman's $\rho_{\text{tumor}} = -0.39$; $\text{FDR}_{\text{tumor}} = 0.015$; $\rho_{\text{reference}} = -0.31$; $\text{FDR}_{\text{reference}} = 0.054$; permutation $P = 0.06$; see Methods, Extended Data Fig. 6a–c and Supplementary Table 7). Furthermore, the alternative allele dominantly associated with increased H3K27ac modification[46] (effect size = −35; $P = 0.164$; Mann–Whitney *U*-test; Fig. 5c) and H3K27ac modification was negatively correlated with cg18664856 methylation (Spearman's $\rho = -0.39$; $P = 2.76 \times 10^{-2}$; Extended Data Fig. 6d,e). Because methylation of cg18664856 was also negatively correlated with *AKT1* mRNA abundance (Spearman's $\rho = -0.38$; $P = 1.54 \times 10^{-5}$; Extended Data Fig. 6f), we checked the effect of the rs2494734 genotype on *ATK1* mRNA abundance. Consistently, the alternative allele was additively associated with increased *AKT1* mRNA abundance (Spearman's $\rho = 0.27$; $P = 1.01 \times 10^{-4}$; Mann–Whitney *U*-test; Fig. 5d). This effect was validated in the TCGA cohort (Spearman's $\rho_{\text{tumor}} = 0.17$; $P_{\text{tumor}} = 1.74 \times 10^{-3}$; Extended Data Fig. 6g). While no association was seen in the TCGA reference tissue ($\rho_{\text{reference}} = -0.06$; $P_{\text{reference}} = 0.68$; Extended Data Fig. 6h), this eQTL was reported in GTEx[42] with a *P* value above the genome-wide significance level ($P > 1.62 \times 10^{-10}$; Supplementary Table 7).

Finally, given the robust literature on *AKT1*'s oncogenic functions and therapeutic value[52], we tested the effect of the rs2494734 genotype on survival. The alternative allele was dominantly associated with an increased risk of relapse (HR = 2.85; 95% CI: 1.35–5.99; $P = 5.80 \times 10^{-3}$; CoxPH model; Fig. 5e), and was validated in an independent cohort of 101 patients[26,27,37,38] (HR = 2.2; 95% CI: 1.2–4.0; $P = 0.017$; Fig. 5f). These findings highlight another example of the interplay between germline haplotypes and tumor methylation, regulating downstream gene expression and impacting the clinical behavior of prostate cancer.

Discussion

Tumor meQTLs occur when a germline locus influences the epigenetic profile of a tumor but not its predecessor nonmalignant cells. The resulting regulatory effects can ripple through the central dogma, facilitating interactions between the germ line and the somatic tissue, leading to tumorigenesis decades after birth. While specific driver mutations can be driven by environmental or replicative factors, they arise in the context of the germline genome that biases towards or against them[4]. Understanding this interaction can help identify determinants of disease susceptibility and aggressivity. This is particularly important in prostate cancer where current clinical factors do not fully predict the interpatient heterogeneity in tumor behavior and treatment response. As was first presented by Heyn et al.[17], and is confirmed here, some GWAS loci modulate risk via dysregulation of DNA methylation. Measuring this direct effect of the germ line on methylation generates large effect sizes, overcoming the power limitations of small cohorts. We validate this aspect of germline modulators of tumor methylation by re-identifying the rs684232 haplotype—a previously reported risk locus[17]. Furthermore, we identify novel loci predictive of aggressive disease, including loci targeting prostate cancer driver events, such as *TCERG1L*[26] and *AKT1* (refs. [48–51]). Interestingly, not all risk meQTLs were tumor specific. MeQTLs detected in reference tissue may facilitate tumor initiation (that is, modulating pre-neoplastic methylation), while tumor meQTLs may facilitate tumor progression (that is, modulating oncogenic methylation).

The mechanisms by which germline loci affect tumor methylation are largely unknown, and there are probably many. First, the most direct would be a SNP breaking a methylated CpG dinucleotide; in our data, this accounted for only 0.1% of tumor meQTLs. Second, SNPs can influence CTCF binding, supported by their enrichment and allele-specific binding at these sites (Figs. 3d and 4h). Changes in CTCF binding can impact local methylation by modulating poly(ADP-ribose) polymerase 1 activity, and subsequently DNA (cystosine-5)-methyltransferase 1 activity[47]. Third, SNPs can create or destroy DNA motifs that alter protein binding affinities, thereby promoting or antagonizing methylation[53]. Finally, tumor meQTLs may represent a secondary effect of the germline-modulating processes that co-occur with methylation changes (for example, chromatin modifications).

The cohort analyzed here was modest in size relative to contemporary GWASs ($n = 589$ patients), yet we identified and validated 7,590 meQTLs and 1,178 tumor-specific meQTLs, suggesting that they are very widespread in prostate cancer. The tumor meQTLs reported represent tag loci and require fine mapping to determine the casual loci. Additionally, cell type composition can play a role in meQTL identification, as different cell types can have different methylation profiles[54]. For example, loci modulating the tumor microenvironment might alter measured methylation unrelated to methylation in cancer cells. Our approach focused on *cis* associations (loci proximal to the methylation site) due to their strong signal. We also detected *trans* tumor meQTLs, despite being underpowered to explore these. Larger cohorts are needed to quantify the *trans* influences of the germ line on the tumor epigenome, suggesting we are currently underestimating the influence of germline aberrations on the tumor epigenome and gene expression.

These data reveal a novel mechanism through which the germline genome influences the somatic landscape of a tumor. These germline–somatic interactions can be exploited to identify prognostic germline loci that might be minimally invasive biomarkers, to aid the triage of patients to more expensive tissue- or radiology-based assays. These data support further exhaustive study of germline–somatic interactions in prostate and other tumor types.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0579-z.

Received: 17 September 2018; Accepted: 13 August 2019;
Published online: 7 October 2019

References

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
4. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
5. Tomlinson, I. P. et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008).
6. Peterson, G. M. et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* **42**, 224–228 (2010).
7. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
8. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* **1**, 157–162 (2001).
9. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumourigenesis. *Cell* **61**, 759–767 (1990).
10. Nik-Zainal, S. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
11. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
12. Reynolds, P. A. et al. Tumour suppressor p16^{INK4A} regulates polycomb-mediated DNA hypermethylation in human mammary epithelial cells. *J. Biol. Chem.* **281**, 24790–24802 (2006).
13. Suzuki, H. et al. Epigenetic inaction of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nat. Genet.* **36**, 417–422 (2004).
14. Saghafinia, S. et al. Pan-cancer landscape of aberrant DNA methylation across human tumors. *Cell Rep.* **25**, 1066–1080 (2018).
15. Whitington, T. et al. Gene regulatory mechanisms underpinning prostate cancer susceptibility. *Nat. Genet.* **48**, 387–397 (2016).
16. Cowper-Sal-lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
17. Heyn, H. et al. Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep.* **24**, 331–338 (2014).
18. Taylor, R. A. et al. Germline *BRCA2* mutations drive prostate cancers with distinct evolutionary trajectories. *Nat. Commun.* **8**, 13671 (2017).
19. Szulkin, R. et al. Genome-wide association study of prostate cancer-specific survival. *Cancer Epidemiol. Biomarkers Prev.* **24**, 1796–1800 (2015).
20. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
21. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
22. Klotz, L. et al. Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. *J. Clin. Oncol.* **33**, 272–277 (2015).
23. D'Amico, A. V. et al. Cancer-specific mortality after surgery or radiation for patients with clinically localized prostate cancer managed during the prostate-specific antigen era. *J. Clin. Oncol.* **21**, 2163–2172 (2003).
24. Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
25. Cooper, C. S. et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).
26. Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
27. Espiritu, S. G. et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell* **173**, 1003–1013 (2018).
28. Lin, D. W. et al. Genetic variants in the *LEPR*, *CRY1*, *RNASEL*, *IL4*, and *ARVCF* genes are prognostic markers of prostate cancer-specific mortality. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1928–1936 (2011).
29. Eeles, R. A. et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41**, 1116–1121 (2009).
30. Eeles, R. A. et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45**, 385–391 (2013).
31. Lévesque, E. et al. Steroidogenic germline polymorphism predictors of prostate cancer progression in the estradiol pathway. *Clin. Cancer Res.* **20**, 2971–2983 (2014).
32. 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).
33. Matejcic, M. et al. Germline variation at 8q24 and prostate cancer risk in men of European ancestry. *Nat. Commun.* **9**, 4616 (2018).
34. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
35. Stelloo, S. et al. Integrative epigenetic taxonomy of primary prostate cancer. *Nat. Commun.* **9**, 4900 (2018).
36. Jackson, W. C. et al. Intermediate endpoints after postprostatectomy radiotherapy: 5-year distant metastasis to predict overall survival. *Eur. Urol.* **74**, 413–419 (2018).
37. Bhandari, V. et al. Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* **51**, 308–318 (2019).
38. Sinha, A. et al. The proteogenomic landscape of curable prostate cancer. *Cancer Cell* **35**, 414–427 (2019).
39. Kim, H. The retinoic acid synthesis gene *ALDH1a2* is a candidate tumor supporessor in prostate cancer. *Cancer Res.* **65**, 8118–8124 (2005).
40. Doose, G. et al. MINCR is a MYC-induced lncRNA able to modulate MYC's transcriptional network in Burkitt lymphoma cells. *Proc. Natl Acad. Sci. USA* **112**, E5261–E5280 (2015).
41. Wang, S. et al. Long non-coding RNA MINCR promotes gallbladder cancer progression through stimulating EZH2 expression. *Cancer Lett.* **380**, 122–133 (2016).
42. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
43. Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
44. Yi, J. M. et al. Genomic and epigenomic integration identifies a prognostic signature in colon cancer. *Clin. Cancer Res.* **17**, 1535–1545 (2011).
45. Yi, J. M. et al. DNA methylation biomarker candidates for early detection of colon cancer. *Tumour Biol.* **33**, 363–372 (2012).
46. Kron, K. J. et al. TMPRSS2–ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat. Genet.* **49**, 1336–1345 (2017).
47. Zampieri, M. et al. ADP-ribose polymers localized on Ctcf–Parp1–Dnmt1 complex prevent methylation of Ctcf target sites. *Biochem. J.* **441**, 645–652 (2012).
48. Lee, J. K. et al. N-Myc drives neuroendocrine prostate cancer initiated from human prostate epithelial cells. *Cancer Cell* **29**, 536–547 (2016).
49. Kwon, E. M. et al. Genetic polymorphisms in inflammation pathway genes and prostate cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **20**, 923–933 (2011).
50. Karyadi, D. M. et al. Confirmation of genetic variants associated with lethal prostate cancer in a cohort of men from hereditary prostate cancer families. *Int. J. Cancer* **136**, 2166–2171 (2015).
51. Liu, J. M. et al. Association between single nucleotide polymorphisms in *AKT1* and the risk of prostate cancer in the Chinese Han population. *Genet. Mol. Res.* **16**, gmr16019469 (2017).
52. Song, M. et al. AKT as a therapeutic target for cancer. *Cancer Res.* **79**, 1019–1031 (2019).
53. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
54. Bernstein, B. E. et al. The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).

Acknowledgements

The authors thank all members of the Boutros laboratory, as well as K. Kron and A. Meng, for helpful suggestions and support. The results described here are based in part on data generated by the TCGA Research Network (http://cancergenome.nih.gov/). This study was conducted with the support of Movember through Prostate Cancer Canada and with the additional support of the Ontario Institute for Cancer Research, funded by the Government of Ontario. We thank the Princess Margaret Cancer Centre Foundation and Radiation Medicine Program Academic Enrichment Fund for support (to R.G.B.). R.G.B. is the recipient of a Canadian Cancer Society Research Scientist Award. This work was supported by Prostate Cancer Canada and is proudly funded by the Movember Foundation (grant RS2014-01 to P.C.B.; grant RS2014-02 to M.L.; and grant RS-2016-01 to H.H.H.). P.C.B. was supported by a Terry Fox Research Institute New Investigator

Award and a CIHR New Investigator Award. H.H.H. was supported by CIHR operating grant 142246 and CCSRI grant 703800. This project was supported by Genome Canada through a Large-Scale Applied Project contract to P.C.B., R. Morin and S. P. Shah. K.E.H. was supported by a CIHR Vanier Fellowship. R.S.M. acknowledges funding from the Prostate Cancer Research Program Impact Award from the US Department of Defense (W81XWH-17-1-0675), as well as the Individual Investigator Research Award from CPRIT (RP190454). M.L.F. acknowledges funding from NIH (5R01CA193910), the Challenge Award from the Prostate Cancer Foundation, and the H.L. Snyder Medical Foundation. B.P. acknowledges funding from the National Human Genome Research Institute (R01HG009120). This work was supported by the NIH/NCI under award number P30CA016042, and by an operating grant from the National Cancer Institute Early Detection Research Network (1U01CA214194-01) to P.C.B. and T.K.

Author contributions

A.Shetty, M.F., M.S., L.T., J.J., A.W., M.O., V.P., H.H. and A.Sinha prepared the samples. B.T. and T.v.d.K. performed the pathology analyses. K.E.H., Y.-J.S. and M.A. performed the statistical and bioinformatics analyses. A.G., J.Y., S.G.R., C.Q.Y., V.H., L.E.H., Y.-J.S., J.L., T.N.Y., S.M.G.E., A.R., A.F., A.M., C.B. and E.O'C. processed the data. K.E.H. wrote the first draft of the manuscript. K.E.H., R.G.B. and P.C.B. initiated the project. M.L.K.C., M.M.P., J.D.M., M.L., T.K., B.P., M.L.F., R.S.M., H.H.H., R.G.B. and P.C.B. supervised the research. Y.F., B.T., A.B. and L.L. generated tools and reagents. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0579-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0579-z.

Correspondence and requests for materials should be addressed to R.G.B. or P.C.B.

Peer review information Kate Gao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Discovery patient cohort. All patients had pathologically confirmed prostate cancer and were hormone naïve at the time of therapy. All patients were treated with either IGRT or radical prostatectomy (surgery). Single ultrasound-guided needle biopsies were obtained for the IGRT cohort before the start of therapy, as previously described[26]. Fresh-frozen radical prostatectomy specimens were obtained from the University Health Network (UHN) Pathology BioBank or from the Genito-Urinary BioBank of the Centre Hospitalier Universitaire de Québec—Université Laval (CHUQ). In accordance with the local Research Ethics Board and International Cancer Genome Consortium guidelines, whole blood was collected (and informed consent obtained) at the time of clinical follow-up. Previously collected tumor tissue was utilized based on UHN Research Ethics Board-approved study protocols (UHN 06–0822-CE, UHN 11–0024-CE and CHUQ 2012–913:H12–03–192). Two genitourinary pathologists (T.v.d.K. and B.T.) independently evaluated scanned hematoxylin and eosin-stained slides to confirm Gleason scores and tumor cellularity for all tumor specimens. The clinical T category was reported using standard National Comprehensive Cancer Network criteria (https://www2.tri-kobe.org/nccn/guideline/urological/english/prostate.pdf). Serum PSA was reported based on the reading at the time of diagnosis (measured in ng ml^{-1}). The discovery cohort consisted of samples from 161 cases previously characterized[26,27], along with 80 new cases collected and processed in the same manner. These additional 80 cases were chosen to match the clinical features of the original 130 (that is, similar age, Gleason score, tumor stage, proportion of BCRs, and time to BCR). For patients treated with IGRT, BCR was defined as a rise in the PSA concentration of at least 2.0 ng ml^{-1} above the nadir. The nadir refers to the stable PSA level that follows a slight rise directly after radiotherapy. For patients treated with surgery, BCR was defined as two consecutive postsurgery PSA measurements over 0.2 ng ml^{-1}, or triggered salvage therapy.

Sample processing. At UHN, selected prostate samples were cut into 60 µm × 10 µm sections, with a hematoxylin and eosin-stained 4-µm section every ten cuts. Hematoxylin and eosin-stained sections were marked by a genitourinary pathologist (T.v.d.K. or B.T.) to indicate areas suitable for macro-dissection (that is, >70% tumor cellularity). Manual macro-dissection was performed using sterile scalpel blades, and DNA was obtained by phenol:chloroform extraction, as previously reported[26]. DNA was extracted from whole blood using an ArchivePure DNA Blood Kit (5 PRIME) at the Applied Molecular Profiling Laboratory at the Princess Margaret Cancer Centre. At CHUQ, histology quality control was performed the same way as described above, and when the tumor area was considered large enough, two cores with 1 mm diameter were taken from the tumoral zone using a sterile biopsy punch (Miltex). Tissues were immediately disrupted in ATL buffer using a Minilys homogeneizer (Bertin Technologies). DNA was finally extracted from the lysate using a QIAmp DNA Mini Kit (Qiagen). The same kit was used to generate DNA extractions on blood samples from this site. All DNA samples were quantified using a Qubit 2.0 Fluorometer (Life Technologies) and assessed for purity using a NanoDrop ND-1000 spectrophotometer.

Methylation array data generation. Methylation microarray data generation was carried out as previously described[26]. Briefly, Illumina Infinium HumanMethylation450 BeadChip kits were used to assess global methylation, using 500 ng of input genomic DNA, at the McGill University and Génome Québec Innovation Centre. All samples were processed from fresh-frozen prostate cancer tissue.

Methylation array data analysis. Methylation microarray data were processed using the R statistical environment (version 3.2.3), as outlined elsewhere[55]. Briefly, raw methylation intensity levels were preprocessed using Dasen[56] and filtered according to detectability above background noise, non-CpG methylation and cross-hybridization using the DMRcate package (version 1.6.53). Chromosome location, probe position and gene symbol were annotated using the IlluminaHumanMethylation450kanno.ilmn12.hg19 package (version 0.6.0).

WGS. WGS was conducted as previously reported[26]. Briefly, sequencing libraries were prepared using 50 ng genomic DNA and enzymatic reagents from KAPA Library Prep Kits (KAPA Biosystems; catalog number: KK8201) according to the protocols described for end repair, A-tailing and adapter ligation[57]. Sequencing was carried out using a HiSeq 2000 platform (Illumina), and samples were sequenced to a minimum coverage depth of 30× and a median coverage of 44.2 ± 4.7× (s.d.).

mRNA microarray generation. Total RNA was extracted from alternating adjacent sections, using the mirVana miRNA Isolation Kit (Life Technologies), according to the manufacturer's instructions, as described previously[26]. In total, three batches were profiled at two locations. For batch 1 samples, 150 ng total RNA was assayed on the Affymetrix Human Gene 2.0 ST array (HuGene 2.0 ST) at the Centre for Applied Genomics (Hospital for Sick Children, Toronto, Ontario, Canada). For samples in batches 2 and 3, 100 ng total RNA was assayed on the Affymetrix Human Transcriptome Array 2.0 (HTA 2.0) and HuGene 2.0 ST, respectively, at the London Regional Genomics Centre (Robarts Research Institute, London, Ontario, Canada).

WGS data analysis. Raw sequencing reads were aligned to the human reference genome GRCh37 using BWA-mem (version 0.7.12+)[58] at the lane level (Supplementary Table 1). Picard (version 1.92) merged these lane-level BAMs from the same library and marked duplicates. Picard was also used to merge library-level BAMs from the same sample without marking duplicates. Local realignment and base quality recalibration were completed on tumor/normal pairs together with the Genome Analysis Toolkit (GATK version 3.4.0+) (Supplementary Table 1)[59]. Normal samples were extracted, headers corrected (SAMtools version 0.1.9)[60] and files indexed (Picard version 1.92) into individual sample-level BAMs.

mRNA abundance analysis. Raw mRNA data were downloaded from GSE107299 and preprocessed under R (version 3.2.5). Background correction, normalization algorithms and annotation were implemented in the oligo (version 1.34.2) package from the BioConductor (version 3.2) open-source project. The Robust multichip average algorithm was applied to the raw intensity data[61]. Probes were mapped to Entrez gene ID using custom CDF files (version 20) for the HTA 2.0 and HuGene 2.0 ST arrays from http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp. The sva package (version 3.18.0) was used to correct for batch effects between different arrays. mRNA abundance levels from HuGene 2.0 ST and HTA 2.0 were combined into one dataset based on Entrez gene IDs. The mRNA abundance levels were averaged among duplicated Entrez gene IDs. Entrez gene IDs were then converted into gene symbols and chromosome locations, based on the human reference genome GRCh37 from the UCSC table browser (download date: 2 August 2016).

Identification of germline SNPs. GATK (version 3.4.0+) (Supplementary Table 1) was used to call germline SNPs by first running HaplotypeCaller on the realigned and recalibrated tumor/normal pair (Supplementary Table 1). Next, VariantRecalibrator and ApplyRecalibration were applied to ensure high-quality calls. GATK Best Practices filters were applied to the resulting Variant Call Format files. We only considered biallelic SNPs in this analysis, and 98.54% of autosomal SNPs (4,894,225/4,966,931) had all three genotypes.

Candidate risk meQTL replication. We conducted a candidate meQTL analysis to replicate the eight prostate meQTLs reported by Hyen et al.[17]. Associations were tested using Spearman's correlation. Spearman's correlation tested the additivity of the alternative allele (that is, the correlation between methylation and the genotype coded 0 (homozygous reference), 2 (heterozygous) and 3 (homozygous alternative)). We considered a significant threshold of $P<0.01$ (Bonferroni adjustment). Three of the eight meQTLs could not be tested in this cohort as the probes were filtered out during methylation processing (cg20129853, cg13762704 and cg02340056; Supplementary Table 2).

Risk locus associations. A list of 160 germline polymorphisms associated with prostate cancer risk was cultivated from the literature[15,25,27–32]. SNPs from these studies were chosen if they were associated with the risk of prostate cancer or prognosis of patients with prostate cancer. Associations were tested using Spearman's correlation. Spearman's correlation tested the additivity of the alternative allele (that is, the correlation between the event and the genotype coded 0 (homozygous reference), 2 (heterozygous) and 3 (homozygous alternative)). Significant associations were defined as FDRs < 0.05. We chose Spearman's correlation to avoid violating distributional assumptions made in linear models, given that methylation data do not follow traditional distributions. Additionally, we selected Spearman's correlation over the previously reported multivariate random forest selection frequency method[17], given the subsequent genome- and methylome-wide approaches in this work (see the 'Discovery prognostic genome wide association studies' and 'Discovery *cis* germline–methylation associations' sections in the Methods). This approach was too computationally intensive to apply to the 1×10^7 independent tests conducted in the following sections, so for consistency we applied Spearman's correlation for all associations. However, we implemented the multivariate random forest selection frequency method[17] to confirm a subset of our high-confidence hits (see 'Validation of germline–methylation associations').

Survival analysis. Survival analysis was conducted with the R statistical environment (version 3.3.1). Where the assumption of proportional hazards applied, a CoxPH model was implemented to test the association between methylation (that is, the median dichotomized *m* value) and BCR, as defined previously[26]. Probes with a *P* value of $<1\times10^{-4}$ were carried forward to the analysis. For survival analysis of *TCERG1L* methylation levels, cutp from the survMisc (version 0.4.5) package was used to determine a dichotomization threshold, to replicate thresholds used in previous work[28]. Survival associations were validated in an independent cohort of 101 clinically matched primary samples[26,27,37,38].

Discovery prognostic germline–methylation associations. Genome-wide associations were tested for all 58 prognostic methylation probes. Germline SNPs were filtered based on a minimum allele frequency of >0.1 and Hardy–Weinberg equilibrium violation ($P>1\times10^{-8}$). Associations between the remaining SNPs and the 58 prognostic methylation probes were evaluated using the R-plugin feature of the plink software (version 1.07), to implement a Spearman's correlation test[62]. Spearman's correlation tested the additivity of the alternative allele (that is, the correlation between the event and the genotype coded 0 (homozygous reference), 2 (heterozygous) and 3 (homozygous alternative)). Manhattan plots were generated to visualize the results for each SNP. Quantile–quantile plots were generated to assess bias in the model fit. A stringent Bonferroni adjustment was applied to correct for multiple hypothesis testing; therefore, SNPs with $P < 5 \times 10^{-8}$ were considered significantly associated. Linkage disequilibrium was calculated and visualized using Haploview (version 4.2)[63]. Pairwise linkage disequilibrium was quantified using D′, and haplotypes were defined according to Gabriel et al.[64].

Discovery *cis* germline–methylation associations. All methylation probes were tested for *cis* germline–methylation associations by looking at SNPs that were in a ±10 kb window around the probe. Associations were tested using Spearman's correlation, as outlined above (see 'Discovery prognostic germline-methylation associations' in Methods). Associations were deemed significant for P values $< 3 \times 10^{-9}$, as this represented the Bonferroni threshold (1.5×10^{7} independent tests).

TCGA validation cohort. The TCGA prostate adenocarcinoma data were used as a validation cohort[34]. Concordance between SNP6 microarray (SNP6) genotypes and WES of blood sample calls was evaluated, and only samples with >80% concordance were retained (348 samples; three samples were excluded for concordance < 80%: TCGA-HC-7738, TCGA-EJ-7312 and TCGA-EJ-5505). Genotypes were imputed using the Sanger Imputation Service (pre-phasing with Shapeit2)[65], PBWT[66] and the Haplotype Reference Consortium (release 1.1) panel[67]. The accuracy of the imputed genotypes was evaluated against WES blood sample calls. A median accuracy of 0.988 was estimated. Genotypes were imputed a second time using combined SNP6 and WES calls, with the same imputation pipeline as described above. In the event that SNP6 and WES disagreed on the genotype at a particular position, the WES call was used. A final list of 40,405,505 SNPs were then available for validation studies.

Validation of germline–methylation associations. Associated SNPs from the discovery cohort ($P < 5 \times 10^{-8}$ from genome-wide analysis; P value $< 4 \times 10^{-9}$ from the *cis* germline–methylation analysis) were tested in the imputed TCGA cohort using the same Spearman's correlation method outlined above. False discovery adjustment was applied to the remaining SNPs, and associations were considered to validate if the FDR was <0.05 and the directionality of the Spearman's ρ was consistent in the discovery and validation cohorts. We implemented the multivariate random forest selection frequency method from Hyen et al.[17] for 59 high-confidence tumor meQTLs, and found that all 59 had a q value of 0 (calculated as the proportion of null models with a random forest selection frequency greater than that of the fit model, as described in Hyen et al.[17]), supporting the validity of our approach (Supplementary Table 7).

Tumor-specific germline–methylation associations. Tumor-specific germline methylation associations were determined using the TCGA tumor and reference methylomes[34]. Similar to the discovery phase, associations were tested using a Spearman's correlation test. Associations were considered tumor specific if the FDR was <0.05 in the tumor and ≥0.05 in the reference in a subset of samples with both tumor and reference methylation profiling ($n = 47$). Tumor specificity was further confirmed for the two stated examples (*TCERG1L* and *AKT1*) via a permutation test. The normal Spearman's ρ was compared with a distribution of tumor Spearman's ρ based on 1,000,000 random subsets of 47 tumor samples. P values were calculated based on the number of iterations where the normal $|\rho|$ was larger than the tumor $|\rho|$. To identify differentially methylated regions (DMRs) between tumoral and normal tissue, raw intensity values were re-normalized together using Dasen[56], and DMRs were identified using the R package DMRcate (version 1.12.1) with default parameters.

ChIP-Seq data analysis. A subset of 34 samples in the discovery cohort were profiled by H3K27ac ChIP-Seq, as previously described[46]. Peak bed files and raw FASTQs for H3K27ac ($n = 92$), H3K27me3 ($n = 76$), androgen receptor ($n = 88$) and H3K4me3 ($n = 56$) were downloaded for an independent cohort from the Gene Expression Omnibus (GSE120738)[35]. Tumor meQTLs overlapping each target were identified using the downloaded bed files. Here, we considered all SNPs within the same haplotype as the tag tumor meQTL. The raw FASTQ files were aligned using bwa (version 0.7.15), and the aligned BAM files from each target were merged for each patient (that is, H3K27ac, H3K27me3, H3K4me3 and androgen receptor BAMs from the same patient were merged). Using the merged BAM files, patients were genotyped at overlapping sites of interest using GATK (version 3.4.0+) HaplotypeCaller. Differential binding analysis was conducted using logistic regression to quantify the contribution of genotype on binding variation. We considered the loci significant if the FDR was <0.05. For each tumor meQTL, we tested all SNPs within the tumor meQTL haplotype, reporting the SNP with the minimum P value. ChIP-Seq data for LNCaP, PC3, 22Rv1, VCaP and RWPE-1 cell lines were downloaded from the sources outlined in Supplementary Table 3 (refs. [46,68–80]).

Regulatory region enrichment analysis. To detect whether these probes are enriched in certain chromosomes, genomic locations and CpG classes, we applied Fisher's exact test followed by multiple test correction (FDR). The methylation promoter region (transcription start site 200, transcription start site 1,500 and the 5′ untranslated region), gene body (first exon and gene body and the 3′ untranslated region) and intergenic region were defined as previously described[81]. Enrichment at transcription factor binding sites and regulatory elements was conducted with previously published ChIP-Seq data from primary tumors[35] and LNCaP, PC3, 22Rv1, VCaP and RWPE-1 cell lines[46,68–80] (Supplementary Table 3). If multiple target/treatment pairs existed, the median number of overlapping SNPs was used. Enrichment was quantified using a permutation test that randomly sampled 23 SNPs when interrogating risk loci meQTLs and 1,031 SNPs when interrogating *cis*-tumor meQTLs genome wide from a list of observed SNPs in our cohort. P values were calculated as the number of null iterations with at least the same number of SNPs overlapping ChIP-Seq peaks as tumor meQTLs, divided by the total number of iterations (10^5). P values were FDR adjusted to account for multiple hypothesis testing. For novel *cis*-tumor meQTLs, we considered the full haplotypes of the tag SNPs (that is, the tumor meQTLs or randomly sampled SNPs), and considered the haplotype overlapping if at least one SNP within the haplotype overlapped with the ChIP-Seq peaks.

Allele imbalance ChIP-Seq analysis. Prostate tissue was collected from 48 patients with localized primary prostate adenocarcinoma. Each patient yielded a sample of the adenocarcinoma and a sample from the surrounding nonmalignant prostate tissue. We performed ChIP-Seq for H3k27ac ($n = 48$), H3k4me2 ($n = 6$), H3k4me3 ($n = 4$), FOXA1 ($n = 10$) and HOXB13 ($n = 9$) on these samples, as well as germline SNP genotyping from blood. Germline variants were phased and imputed to the Haplotype Reference Consortium panel[67]. Mapping and aligning was performed using bwa[58]; allele-specific reads were processed according to the WASP pipeline[82] to remove mapping bias; peaks were identified using the MACS2 software[83]. Allele-specific read counts were generated by the GATK ASEReadCounter[59]. We tested for allele-specific signal using a haplotype beta-binomial test that accounts for read overdispersion. Beta-binomial overdispersion parameters were estimated for each individual/experiment from the aligned allele-specific counts, and were found to be consistently low (<0.01). For each peak and individual, haplotype-specific read counts were merged across all heterozygous read-carrying sites in the peak for a single measure of allele specificity. Every SNP within 100 kb of the peak center and containing at least one heterozygous individual was then tested for allelic imbalance. All heterozygous individuals were tested together under the expectation of a consistent allele-specific effect. Each test was performed once for samples from normal tissue, tumoral tissue, or both, as well as a test for differences in imbalance between tumoral and normal tissues. Finally, peaks were considered 'imbalanced' in each of these four test categories if any of the variants tested for that peak exhibited an allele-specific signal at a 5% FDR.

Overlap between SNP and peak anchor regions. The ENCODE accession numbers for RAD21 ChIA-PET data from LNCaP and DU145 cells are ENCLB189DLP and ENCLB678KEV, respectively. ChIA-PET2 was utilized to process the raw data and obtain the intrachromosomal interactions[84]. Peaks with interactions represent a subgroup of the total peaks identified from the ChIA-PET data. We employed intersectBed (bedtools) to overlap the coordinates of SNP sites and peak regions. Overlap analyses of SNPs with total peaks or interaction peaks are summarized in Supplementary Table 7.

Prediction of the potential targets of risk loci. Peak anchors that overlapped with loci regions were acquired. Genes located in the paired peak anchors were predicted as potential targets of these risk loci.

Pathway enrichment. Genes harboring tumor meQTL–eQTLs were processed using g:Profiler[85] (version r1741_e90_eg37; significance set at the FDR; output set to the generic enrichment map; Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and Reactome databases; background set to all annotated genes; minimum number of genes per pathway: 2).

Germline–RNA (eQTL) and germline–protein (pQTL) associations. Germline–RNA associations were tested for tumor meQTLs. These associations were first interrogated in the discovery cohort using a Spearman's correlation test ($n = 203$), and then validated in the TCGA PRAD RNA-Seq cohort[34]. For stringency, only tumor-specific eQTLs that were not observed as GTEx prostate epithelial eQTLs[42] were retained. Tumor-specific associations were defined as FDRs > 0.05 from published GTEx results, where the FDR was applied over the candidate list of eQTLs ($n = 87$). Germline–protein associations were identified in an exploratory analysis of 70 primary prostate cancers[38] using Spearman's correlation test.

CTCF mechanism. VCaP CTCF and H3K27ac ChIP-Seq and WGS BAM files were downloaded from ENCODE. Whole-genome bisulfite sequencing FASTQ files were downloaded from the Gene Expression Omnibus (accession GSE86832) for three replicates. FASTQ files were aligned using Bismark[86] (version 0.15.0), with one mismatch allowed in a seed alignment.

Data visualization. Visualizations were generated in the R statistical environment (version 3.3.1), with the lattice (version 0.24-30), latticeExtra (version 0.6-28) and BPG (version 5.6.23) packages[87]. Haplotypes were visualized using Haploview (version 4.2)[63].

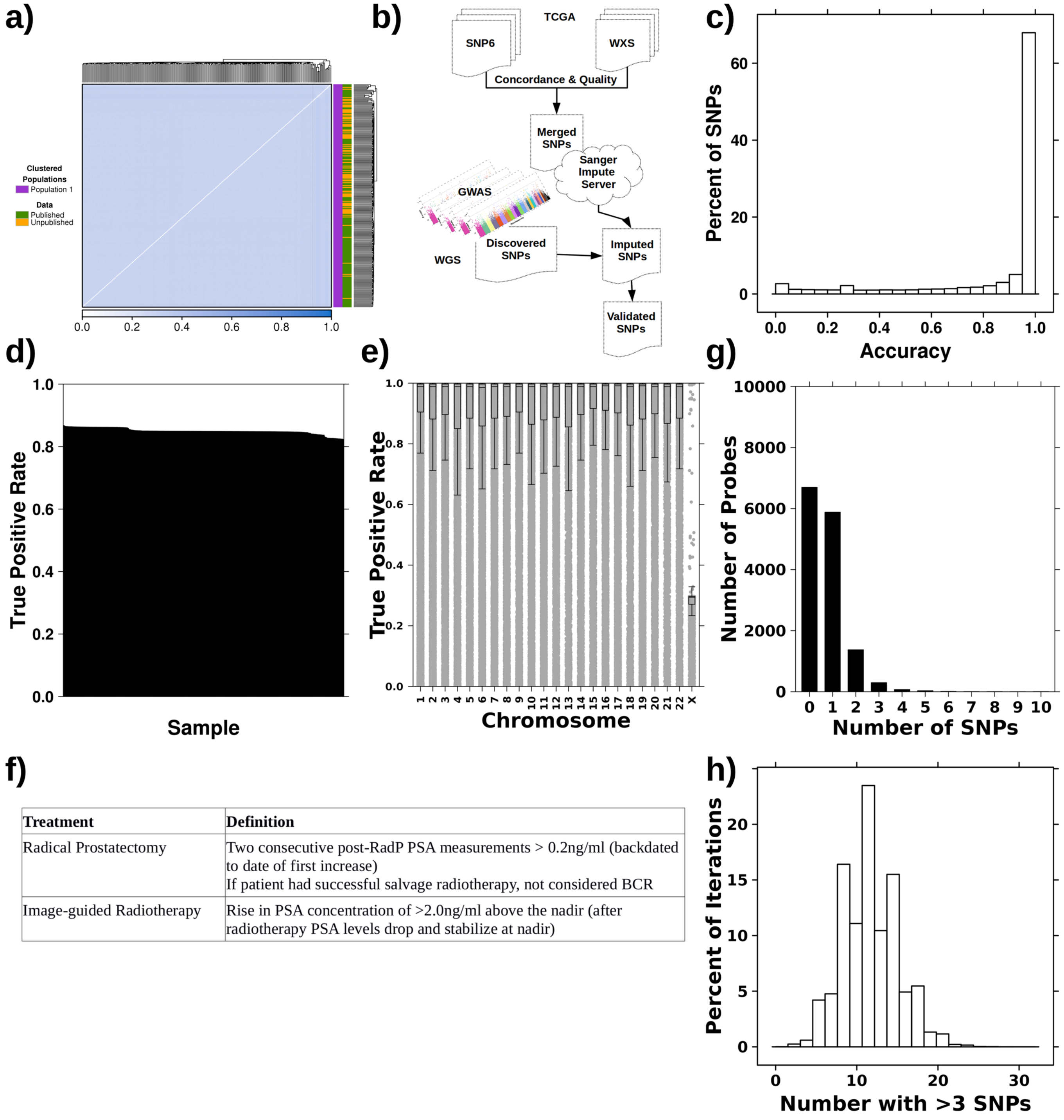
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Methylation data are available in the Gene Expression Omnibus under accession GSE84043. Raw sequencing data are available in the European Genome-phenome Archive under accession EGAS00001000900. Processed variant calls are available through the ICGC Data Portal under the project PRAD-CA (https://dcc.icgc.org/projects/PRAD-CA). TCGA WGS/WES data are available from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/projects/TCGA-PRAD). Primary sample ChIP-Seq data were retrieved from the Gene Expression Omnibus under accession GSE120738. Cell line data sources are outlined in Supplementary Table 3. Detailed information on experimental design can be found in the Life Sciences Reporting Summary.

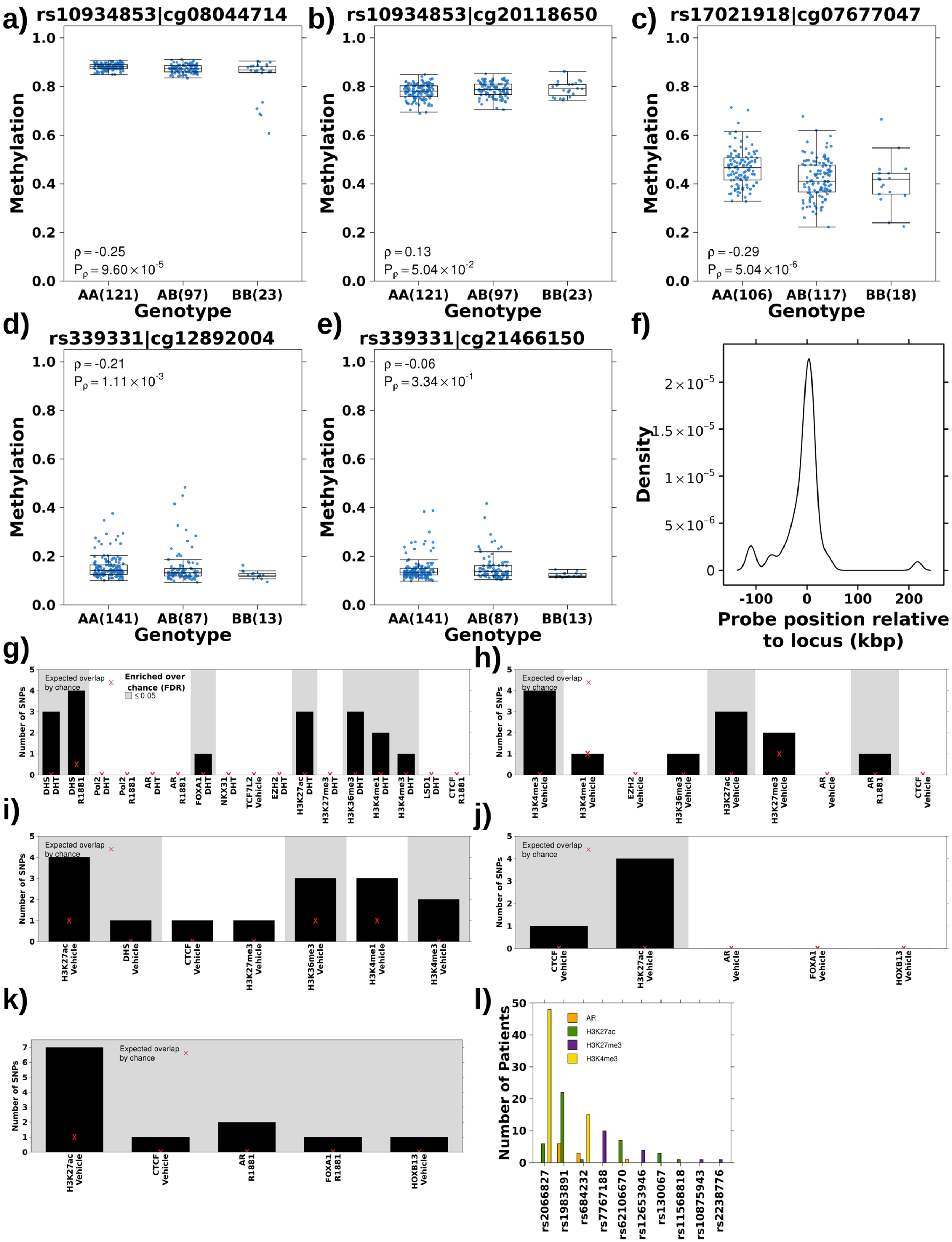
References

55. Shiah, Y.-J., Fraser, M., Bristow, R. G. & Boutros, P. C. Comparison of pre-processing methods for infinium HumanMethylation450 BeadChip array. *Bioinformatics* **33**, 3151–3157 (2017).
56. Pidsley, R. et al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
57. Fisher, S. et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
60. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Irizarry, R. A. et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
62. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
63. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
64. Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
65. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
66. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
67. The Haplotype Reference Consortium A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
68. Yu, J. et al. An integrated network of androgen receptor, polycomb, and *TMPRSS2–ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
69. Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390–394 (2011).
70. Tan, P. Y. et al. Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol. Cell. Biol.* **32**, 399–414 (2012).
71. Hazelett, D. J. et al. Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* **10**, e1004102 (2014).
72. Jin, H. J. et al. Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program. *Nat. Commun.* **5**, 3972 (2014).
73. Xu, K. et al. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338**, 1465–1469 (2012).
74. Zhang, X. et al. Integrative functional genomics identifies an enhancer looping to the *SOX9* gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res.* **22**, 1437–1446 (2012).
75. Chen, Y. et al. ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nat. Med.* **19**, 1023–1029 (2013).
76. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
77. Liang, Y. et al. LSDI-mediated epigenetic reprogramming drives CENPE expression and prostate cancer progression. *Cancer Res.* **77**, 5479–5490 (2017).
78. Sutinen, P. et al. SUMOylation modulates the transcriptional activity of androgen receptor in a target gene and pathway selective manner. *Nucleic Acids Res.* **42**, 8310–8319 (2014).
79. Taberlay, P. C. et al. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).
80. Rickman, D. S. et al. Oncogene-mediated alterations in chromatin conformation. *Proc. Natl Acad. Sci. USA* **109**, 9083–9088 (2012).
81. Mehrmohamadi, M. et al. Integrative modelling of tumour DNA methylation quantifies the contribution of metabolism. *Nat. Commun.* **7**, 13666 (2016).
82. Van de Geijn, B. et al. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
83. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
84. Li, G. et al. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.* **45**, e4 (2016).
85. Reimand, J. et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
86. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
87. P'ng, C. et al. BPG: seamless, automated and interactive visualization of scientific data. *BMC Bioinformatics* **20**, 42 (2019).



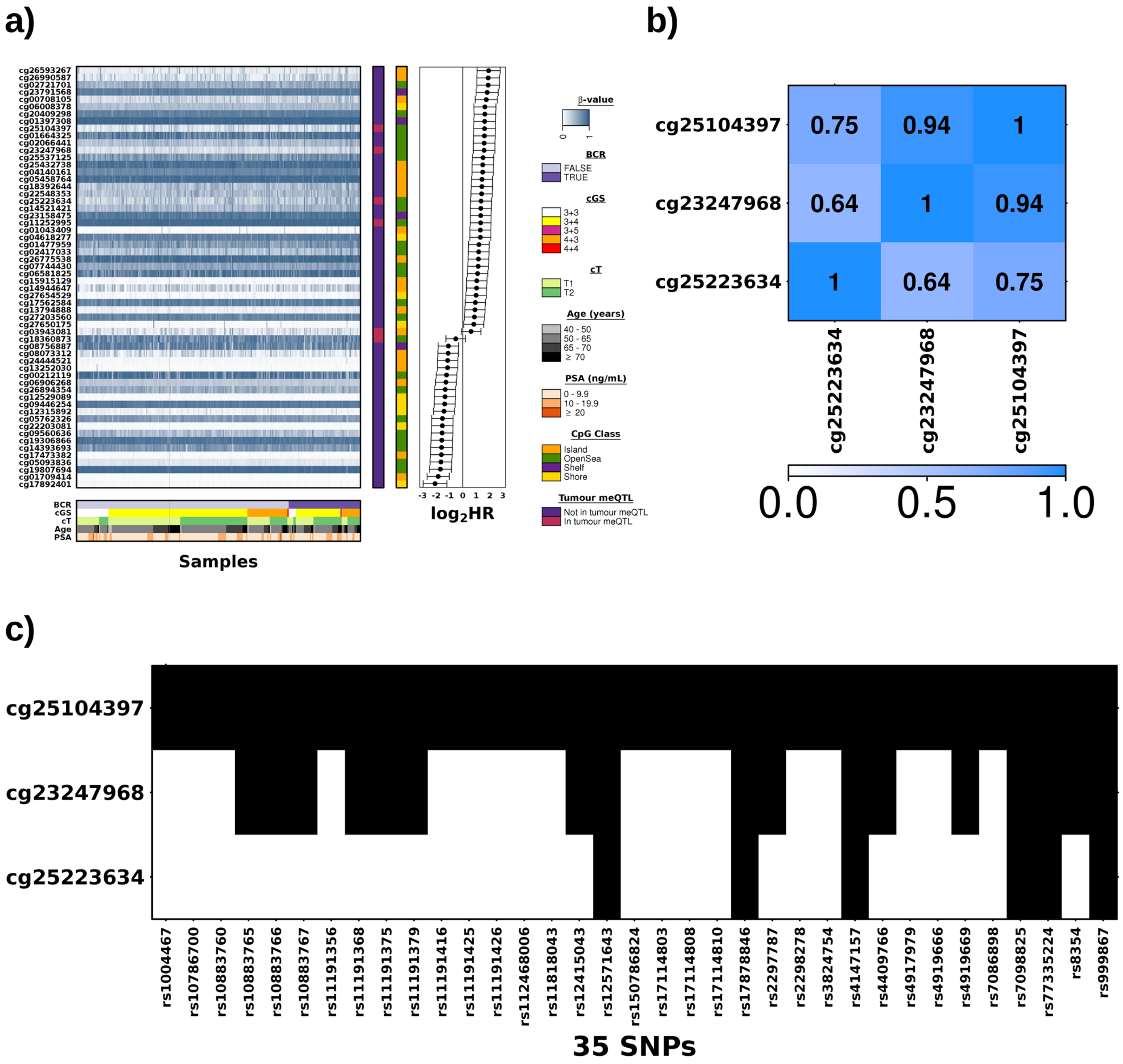
Treatment	Definition
Radical Prostatectomy	Two consecutive post-RadP PSA measurements > 0.2ng/ml (backdated to date of first increase) If patient had successful salvage radiotherapy, not considered BCR
Image-guided Radiotherapy	Rise in PSA concentration of >2.0ng/ml above the nadir (after radiotherapy PSA levels drop and stabilize at nadir)

Extended Data Fig. 1 | Data analysis and quality controls. **a**, Identity-by-state clustering showed no evidence of population stratification. The heat map shows the identity-by-state values for all pairwise comparisons. The first covariate along the right shows the cluster provided by plink (version 1.9). The second covariate indicates whether the sample was previously published or novel to the present study. **b**, Validation dataset workflow. SNP6 array and WES calls were tested for >80% concordance, and merged and additional genotypes were imputed using the Sanger Impute Server. meQTLs were validated in the imputed cohort using the same Spearman's correlation test. **c**, Estimates of per-SNP imputation accuracy by comparison of imputation calls from SNP6 alone with WES genotypes. **d**, The accuracy per sample was consistently above 0.8, with a median per-sample accuracy of 0.849. **e**, The accuracy per chromosome was consistent with no chromosomal bias, with the exception of chromosome X. **f**, Definition of BCR following radical prostatectomy and image-guided radiotherapy (IGRT). **g**, Number of probes with 0–10 SNPs within 50 bp of each methylation probe. Multiple SNPs within this region could effect hybridization of the probe. **h**, Null distribution of probes with more than three SNPs within 50 bp of the probe. The distribution was generated by randomly sampling 12,650 probes 10^6 times.

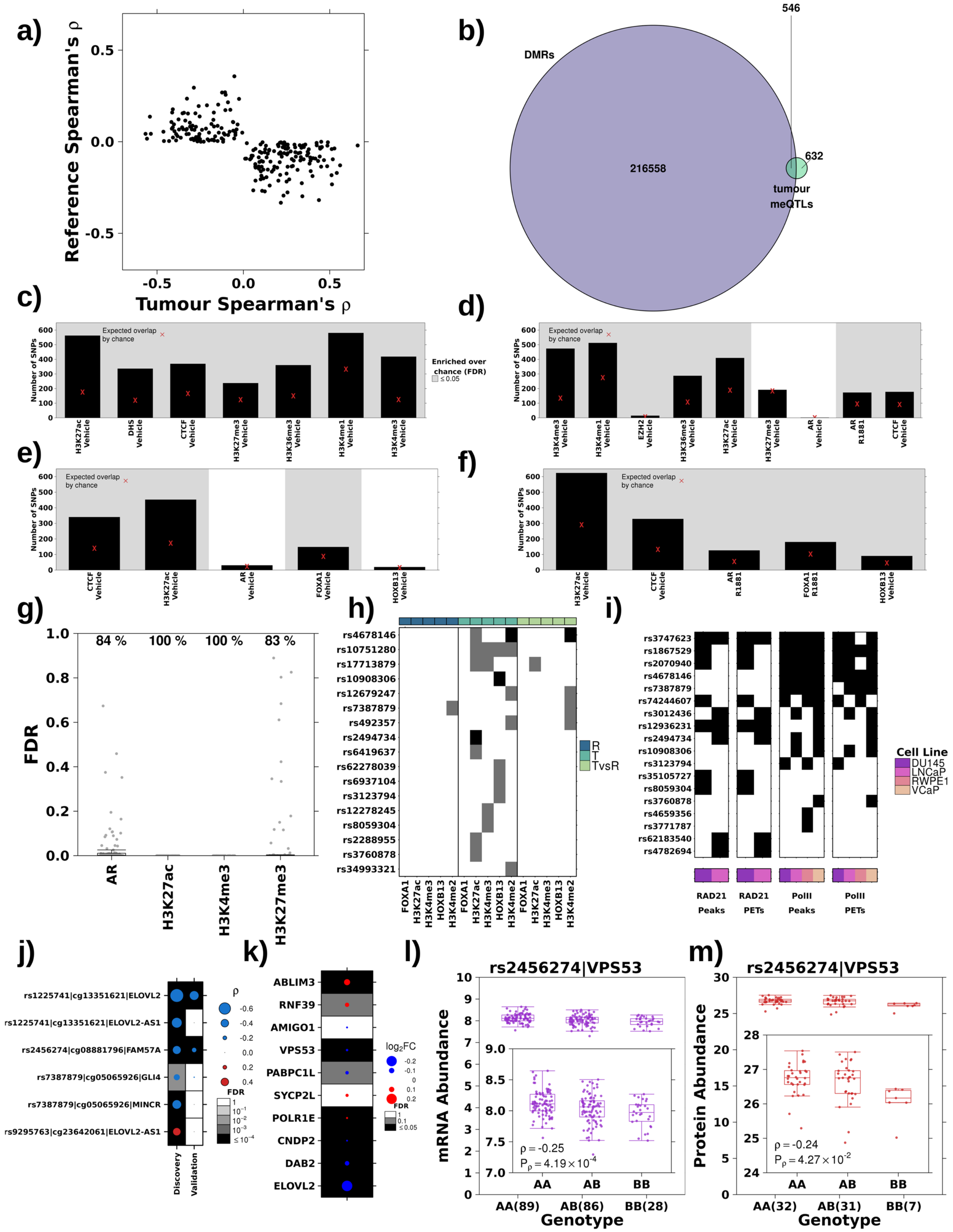


Extended Data Fig. 2 | see figure caption on next page.

Extended Data Fig. 2 | Characterizing risk meQTLs. a–**e**, Three out of five of the meQTLs reported by Heyn et al.[17] validated in this cohort ($P < 0.01$; Spearman's correlation). Box plots represents median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR range. Blue points represent DNA methylation. Numbers of samples with each genotype are given in parentheses. **f**, Distribution of distances between loci and probes, with respect to each locus. **g**–**l**, Overlap of risk SNP meQTLs and regulatory regions in LNCaP (**g**), PC3 (**h**), RWPE-1 (**i**), 22Rv1 (**j**) and VCaP cell lines (**k**) and primary samples (**l**). In **g**–**k**, bar plots show the numbers of tumor meQTLs that overlap each target/treatment pair. Background shading indicates FDRs < 0.05 based on permutation analysis ($n = 10^5$ permutations). Each red × reflects the number of overlapping SNPs expected by chance alone. In **l**, the bar plot shows the number of patients with either androgen receptor ($n = 88$ patients), H3K27ac ($n = 92$), H3K27me3 ($n = 56$) or H3K4me3 ($n = 76$) peaks overlapping each risk meQTL.

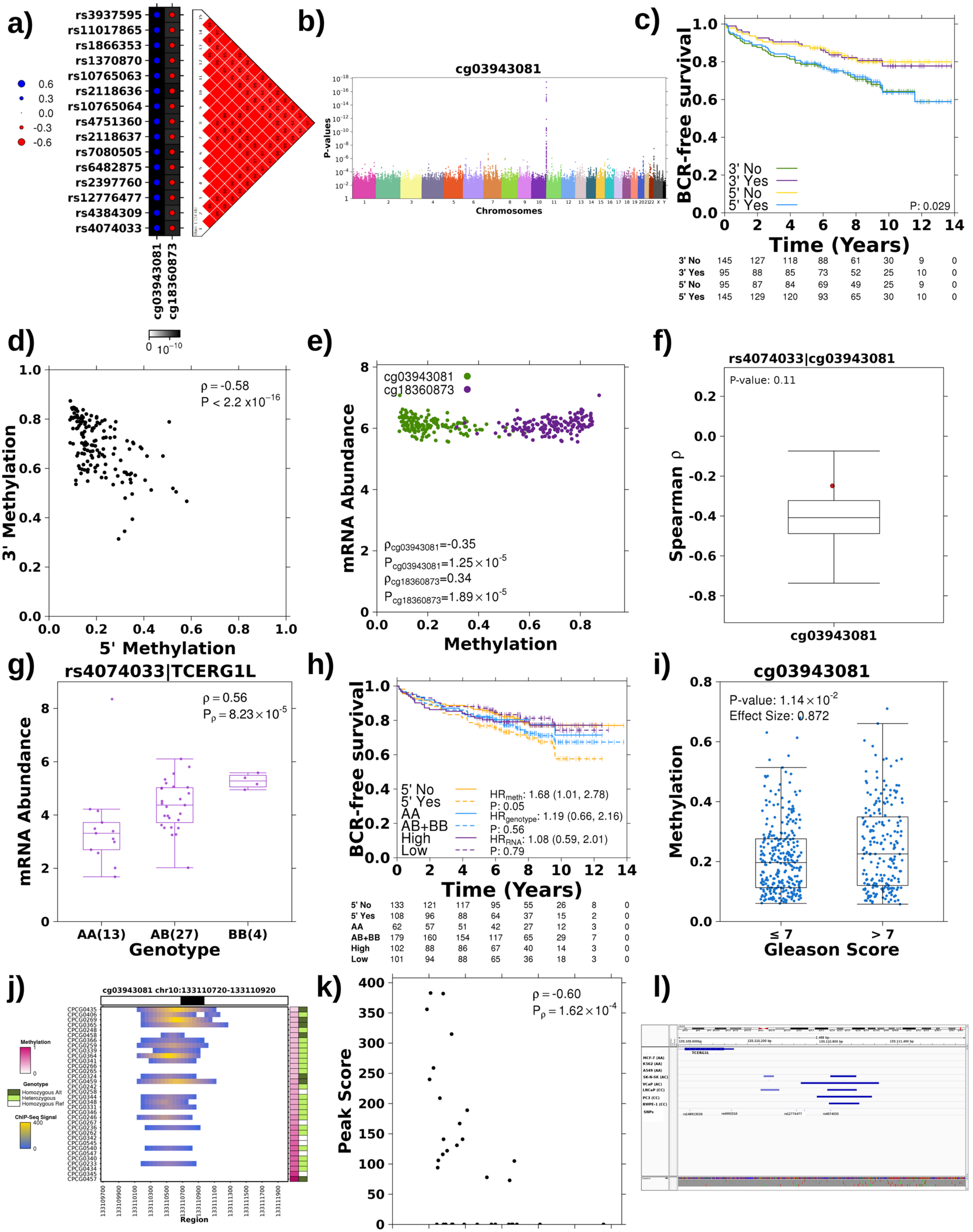


Extended Data Fig. 3 | Characterizing meQTLs targeting prognostic methylation sites. **a**, Methylation *β* values for the 58 selected prognostic methylation probes (rows). Each column represents an individual, and clinical cohorts of the individuals are presented along the bottom. The covariates on the right indicate whether or not the probe was identified as a tumor meQTL, and the CpG class of the probe. The forest plot on the far right depicts HR and 95% CI values considering BCR as endpoint, as determined by the CoxPH model. cGS, clinical Gleason score; cT, clinical T category. **b**, Three probes, located within an open sea region on chromosome 10 within c10orf26, were highly correlated (Spearman's correlation). **c**, Methylation of all of these probes (*y* axis) was associated with the same six SNPs (*x* axis). Black indicates that the SNP was significantly associated with methylation of the probe ($P < 5 \times 10^{-8}$).



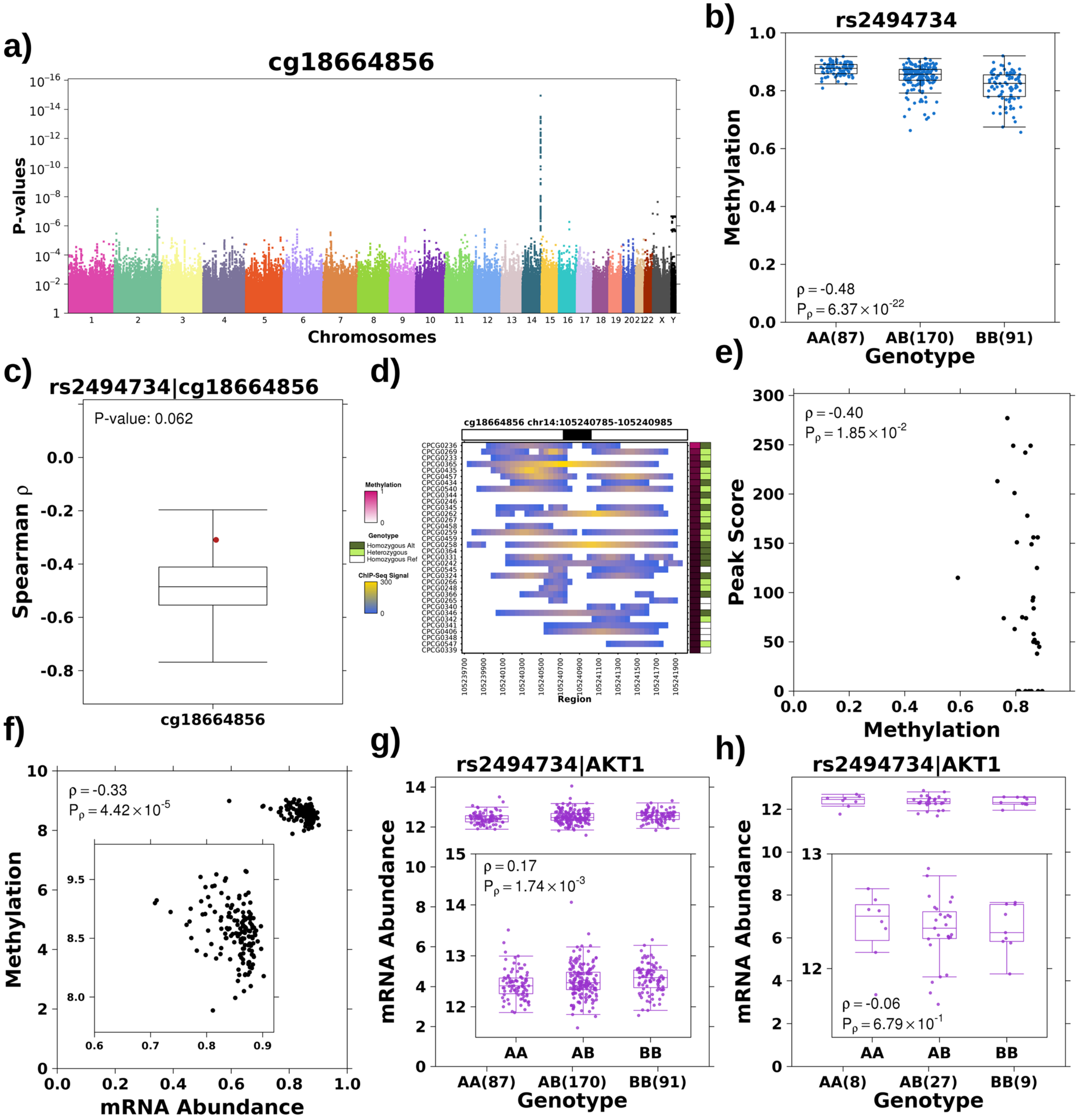
Extended Data Fig. 4 | see figure caption on next page.

Extended Data Fig. 4 | Characterizing tumor meQTLs. a, A subset of tumor meQTLs had opposite effects in tumoral and reference tissue ($n = 234$). **b**, A subset of tumor meQTLs had overlapped DMRs between tumoral and reference tissue. **c**–**g**, Tumor meQTLs were enriched at active regulatory elements in RWPE-1 (**c**), PC3 (**d**), 22Rv1 (**e**) and VCaP cell lines (**f**) and primary samples (**g**). **c**–**f** show bar plots of the numbers of tumor meQTLs that overlap each target. Gray shading indicates significant enrichment ($FDR < 0.05$; $n = 10^5$ permutations). Each red × represents the expected number of overlapping SNPs by chance. In **g**, the box plot shows per-sample FDRs quantifying the enrichment of tumor meQTLs overlapping each target (10^5 permutations). The numbers above indicate the percentages of samples with significant enrichment ($FDR < 0.05$). **h**, Tumor meQTLs overlap allelic imbalance loci in FOXA1, H3K27ac, H3K4me3, HOXB13 and H3K4me2 ChIP-Seq. Black indicates that the tag SNP overlaps the target, while gray indicates that the SNP in strong linkage disequilibrium with the tag SNP overlaps the target. The covariate indicates whether allelic imbalance was identified in tumor, reference, or tumor versus reference analysis. **i**, Tumor meQTLs overlap with RAD21 and RNA Pol-II chromatin loops. Black indicates that the SNP overlaps with the RAD21 or RNA Pol-II ChIA-PET peak or intrachromosomal loops from paired-end tags (PETs). The covariate indicates the cell line. **j**, Long-range gene targets were identified for 17 tumor meQTLs. Dot sizes and colors show the magnitude and direction of Spearman's ρ, respectively. Background shading indicates the FDR. **k**, Six of the ten genes in tumor meQTL–eQTLs were differentially abundant in tumor versus reference tissue ($FDR < 0.05$). Dot sizes and colors indicate the $\log_2$[fold change] magnitude and sign, respectively. Background shading indicates the FDR. **l**,**m**, rs2456274 was associated with mRNA (**l**) and protein (**m**) abundance of *VPS53* (Spearman's correlation). Box plots represent median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR range. Purple and red points represent mRNA and protein abundance, respectively. Numbers of samples with each genotype are givecoule effectn in parentheses.



Extended Data Fig. 5 | see figure caption on next page.

Extended Data Fig. 5 | Characterizing *TCERG1L* tumor meQTLs. a, Haplotype strongly associated with 5′ and 3′ methylation of *TCERG1L*. Dot sizes and colors represent the magnitude and directionality of Spearman's ρ, respectively. Background shading reflects the P value. Pairwise D′ values are shown to the right (solid red means that D′ = 1). **b**, Manhattan plot showing P values after adjusting for tumor cellularity. **c**, Methylation of 5′ and 3′ probes of *TCERG1L* showed opposite effects on BCR. P values were determined by log-rank test. **d**, Methylation of 5′ and 3′ probes was negatively correlated (Spearman's correlation). **e**, Methylation of the 5′ probe is negatively correlated with mRNA abundance of *TCERG1L* (Spearman's correlation), while methylation of the 3′ probe is positively correlated. **f**, meQTLs are stronger in tumoral tissue than reference tissue. The box plot represents the bootstrapped distribution of Spearman's ρ in tumoral tissue ($n_{\text{sampled}} = 47$; $n_{\text{permutations}} = 10^6$), and shows median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR. The red dot represents Spearman's ρ in the reference tissue (the P value is the proportion of iterations where $|\rho_{\text{tumor}}| < |\rho_{\text{reference}}|$). **g**, The *TCERG1L* eQTL is seen in reference tissue (Spearman's correlation). Purple points represent mRNA abundance. Numbers of samples with each genotype are given in parentheses. **h**, *TCERG1L* methylation (cg03943081) is the strongest prognostic measure compared with genotype and mRNA abundance of *TCERG1L* (CoxPH model). **i**, Methylation is significantly associated with Gleason score in the validation cohort (Mann–Whitney U-test; effect size = fold change). **j**, The tag SNP (rs4074033) overlaps H3K27ac histone modifications. The heat map shows the ChIP-Seq peak signal intensity for each patient (y axis) against the spanning region of the SNP (x axis). The covariate along the top indicates ±100 bp around the SNP (black). The covariates on the right indicate methylation β values and genotypes of the patients. **k**, The H3K27ac ChIP-Seq peak signal is negatively correlated with methylation (cg03943081; Spearman's correlation). **l**, Integrative Genomics Viewer screen shot of CTCF ChIP-Seq peaks across eight cell lines, showing only the cell lines that are heterozygous or homozygous for the alternative allele show CTCF binding.



Extended Data Fig. 6 | Characterizing AKT1 tumor meQTLs. **a**, Manhattan plot showing *P* values after adjusting for tumor cellularity. **b**, meQTL validates in the TCGA cohort (Spearman's correlation). Box plots represent median values with 0.25 and 0.75 quantiles. Whiskers represent 1.5× the IQR range. Blue points represent methylation values. **c**, The meQTL is stronger in tumoral than reference tissue. The box plot represents the bootstrapped distribution of Spearman's ρ in tumoral tissue ($n_{\text{sampled}} = 47$; $n_{\text{permutations}} = 10^6$). The red dot represents Spearman's ρ in reference tissue (*P* values represent the proportion of iterations where $|\rho_{\text{tumor}}| < |\rho_{\text{reference}}|$). **d**, The meQTL overlaps H3K27ac modification. The heat map shows the H3K27ac ChIP-Seq signal as previously outlined in Extended Data Fig. 5. **e**, Scatterplot showing negative correlation between the H3K27ac peak score and methylation of cg18664856 (Spearman's correlation). **f**, Methylation at cg18664856 is negatively correlated with mRNA abundance of *AKT1*. **g**, Association between the rs2494734 genotype and mRNA abundance of *AKT1* replicated in TCGA (Spearman's correlation). Purple points represent mRNA abundance. **h**, The association is weaker in reference tissue, as quantified using Spearman's correlation.

nature research

Corresponding author(s): Dr. Paul Boutros

Last updated by author(s): Jul 28, 2019

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 Authors & Referees 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	No software used.
Data analysis	R (v3.2.1), GATK (v3.4.0+), Picard (v1.92), BWA-mem (v0.7.12+), samtools (0.1.9), DMRcate (v1.6.53), IlluminaHumanMethylation450kanno.ilmn12.hg19 (v0.6.0), oligo (v1.32.0), sva (v3.14.0), hugene20sttranscriptcluster.db (v2.13.0), hta20sttranscriptcluster.db (v8.3.1), plink (v1.07), Haploview (v4.2), BPG (v5.6.23), Sanger Imputation Service, latticeExtra (v0.6-28), lattice (v0.24-30), Bismark (v0.15.0), bedtools (v2.27.1)

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/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

Methylation data are available in the Gene Expression Omnibus under accession GSE84043. Raw sequencing data are available in the European Genome-phenome Archive under accession EGAS00001000900 (https://www.ebi.ac.uk/ega/studies/EGAS00001000900). Processed variant calls are available through the ICGC Data Portal under the project PRAD-CA (https://dcc.icgc.org/projects/PRAD-CA). TCGA WGS/WXS data are available at Genomic Data Commons Data Portal (https://gdc-portal.nci.nih.gov/projects/TCGA-PRAD). Primary samples ChIP-Seq data was retrieved from Gene Expression Omnibus under accession GSE120738. Cell line data sources are outlined in Supplementary Table 4.

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	A discovery cohort of 241 samples was used in this analysis. TThis sample size provided a median power, as determined from an one-way ANOVA using Cohen's medium effect size, greater than 80%.
Data exclusions	No data was excluded for the discovery analysis. Samples with concordance <80% between SNP6 array and WXS calls were excluded in the validation cohort.
Replication	An independent cohort of 348 samples, TCGA PRAD, was used to verify the reproducibility of the discovered associations.
Randomization	Samples were allocated to groups based on genotype. Genotypes were not known prior to inclusion of individuals into study.
Blinding	All participants were recruited into the study based only on clinical features. Therefore, neither genotype nor methylation profile was known upon inclusion.

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
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	All relevant population characteristics have been provided in supplementary table 1
Recruitment	New cases included in this publication were chosen to match clinical characteristics, i.e. similar age, Gleason score, tumour stage, proportion of biochemical recurrences (BCR) and time to BCR, of previously published cases (Fraser et al, Nature 2018; Espiritu et al, Cell 2018; Bhandari et al, Nature Genetics 2019; Sinha et al, Cancer Cell 2019).
Ethics oversight	In accordance with local Research Ethics Board (REB) and International Cancer Genome Consortium (ICGC) guidelines, whole blood and informed consent was collected at the time of clinical follow-up. Previously collected tumour tissue was utilized based on UHN REB approved study protocols (UHN 06-0822-CE, UHN 11-0024-CE, CHUQ 2012-913:H12-03-192).

Note that full information on the approval of the study protocol must also be provided in the manuscript.