nature communications

Article

https://doi.org/10.1038/s41467-026-72182-5

Recurrent mechanisms of biallelic epigenetic inactivation reveal new putative tumour suppressor genes in prostate cancer

Received: 19 February 2025

Accepted: 9 April 2026

Published online: 9 September 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

The inactivation of tumour suppressor genes is a key step in cancer development, and is usually achieved by homozygous loss. In prostate cancer, however, large genomic regions are often hemizygously lost, which complicates the identification of putative tumour suppressors in these regions. Here, we develop Epi2Hit, an integrative computational method that leverages whole genome sequencing, epigenomic profiling and gene expression to identify biallelic inactivation of tumour suppressor genes involving DNA methylation of promoter and enhancer regions of one allele and genomic loss of the other allele. We apply Epi2Hit to a cohort of 2,021 prostate cancers to discover tumour suppressor genes. In particular, we identify epigenetic biallelic inactivation of *ZFHX3* at a recurrence level similar to *TP53*. Biallelic inactivation of *ZFHX3*, a transcriptional repressor, leads to upregulation of oncogenes, including *MYC* and a shorter time to metastasis. Finally, we provide evidence that epigenetic silencing as 2nd hit is particularly enriched in regions with nearby essential genes, precluding homozygous loss.

Prostate cancer is the second most frequent type of cancer in men worldwide[1,2]. The disease is highly heterogeneous at the molecular level, with a paucity of somatic single nucleotide variant (SNVs) drivers but frequent structural variations (SVs), which commonly cause copy number losses[3]. Approximately 10% of the prostate cancer genome is recurrently lost in more than 20% of prostate cancer tumours, but only a subset of these harbour known prostate cancer driver genes[4–7].

Tumour suppressor genes (TSG) often undergo biallelic inactivation in various cancer types including prostate cancer[8–11]. The landmark Two-Hit hypothesis posited by Knudson forms the cornerstone of our understanding of TSGs. It suggests that the path towards malignancy is often paved by the sequential inactivation of both alleles of a gene[12]. This is true for prostate cancer, where TSGs such as *PTEN*, *TP53*, *CHD1* and *BRCA1/2* are frequently biallelically inactivated by a combination of copy number loss and SNVs[13–15]. Biallelic inactivation can, however, be highly complex, involving different inactivating germline and somatic events.

Epigenetic silencing by e.g. hypermethylation of promoter and enhancer regions of TSGs is an alternative mechanism of inactivation, often coinciding with genomic inactivation of the other allele. Biallelic inactivation involving copy number loss of one allele and epigenetic silencing of the other allele remains poorly explored except for a few prominent examples including *MLH1* (in colorectal cancer and endometrial cancer)[16,17], *BRCA1* (breast and ovarian cancer)[18–20], *MGMT* (glioma)[21,22] and *APC* (various cancers including prostate cancer)[23–26]. Several of these examples represent established prognostic or predictive biomarkers. Despite its recognised significance, a comprehensive methodology to elucidate the multifaceted nature of biallelic inactivation within this context has remained elusive, constraining our insights into disease progression and the implicated genes, as previous studies were limited to smaller gene panels, low-resolution methodologies, or lack of systematic integration of genomic alterations, DNA methylation, and expression[27,28].

To discover and explore the spectrum and magnitude of epigenetic biallelic inactivation in prostate cancer, we develop and apply a computational methodology, Epi2Hit, which integrates hemizygous genomic disruptions with hypermethylation at regulatory CpG sites to identify biallelic inactivation in TSGs. By leveraging information from somatic alterations with HM450 methylation array, ATAC-sequencing, Hi-C and RNA-sequencing, our approach provides complementary

e-mail: joachim.weischenfeldt@bric.ku.dk

support for the discovery of TSGs inactivated by genomic and epigenetic biallelic inactivation.

Results

Methodology to discover Epi2Hit in prostate cancer

Pan Prostate Cancer Group (PPCG) has assembled multiomics data from a clinically well-annotated and diverse collection of prostate cancers taken from 2021 patients. The sequence data has been analysed with standardised pipelines to identify whole-genome sequencing (WGS)-based SNVs, insertions and deletions (indels), copy number alterations (CNAs), SVs, 450 K array-based methylation patterns, and RNA-seq-based gene expression profiles[29]. Here we selected data from 991 primary prostate cancers, from which WGS data was available, to investigate genomic regions recurrently subjected to copy number loss (altered in at least 10% of the cohort). Consistently with previous findings, our analysis revealed several well-described genomic regions that exhibit high frequencies of homozygous loss, with particular enrichment on chromosomes 5, 10, 13 and 17[30,31] (representing 0.2% of the genome, dark blue, Fig. 1a). In contrast, we identified 8% of the genome being recurrent and almost exclusively hemizygously lost (light blue, Fig. 1a), in particular on chromosomes 2, 3, 5, 8 and 16[32–34] (Fig. 1a, b, Supplementary Table 1, Supplementary Data 1), suggesting that additional (epi) mutational processes could contribute to biallelic inactivation of putative TSG candidates in these otherwise hemizygously deleted regions.

Hypermethylation of regulatory CpG regions can cause transcriptional repression of TSGs including *MLH1*[35], *APC*[23] and *BRCA1*[20]. Motivated by these seminal findings, we asked whether any of the genes recurrently affected by hemizygous loss also exhibit an increase in promoter or enhancer methylation. To this end, we analysed methylation in regulatory regions associated with these genes using data obtained from the Illumina HumanMethylation450 BeadChip array for 1268 samples from the PPCG cohort, representing 318 donors with genomic, transcriptomic and methylation data (Supplementary Fig. 1a, methods). We assessed differential methylation by examining the distribution of β-values at each CpG site. Specifically, we considered a site to exhibit differential methylation if the distribution of β-values showed evidence of multimodality, characterised by two or more distinct peaks, using a KDE-optimisation approach (Supplementary Fig. 2a–d, methods). This approach allowed us to identify regions where methylation levels varied significantly across the sample set. We identified 22% ($n = 1187$) of the genes undergoing recurrent hemizygous loss ($n = 5492$) to also display differential methylation (Fig. 1c), representing candidate genes for biallelic inactivation through a combination of genetic and epigenetic mechanisms.

CpG regions that regulate transcription tend to be located in open chromatin, which we confirmed using available orthogonal ATAC-seq data from an independent cohort of primary prostate cancer samples (GSE188797)[36] (Fig. 1d), suggesting a more robust regulatory influence of CpGs in these regions. We confirmed a transcriptional repressive role of CpG hypermethylation using linear regression for key TSGs such as *APC*, *PTEN*, and *NOTCH1* (negative coefficients demonstrating anti-correlation between CpG methylation in open chromatin and transcription) (Fig. 1e).

To systematically discover and characterise known and poorly characterised TSGs undergoing biallelic inactivation involving hemizygous loss of one allele and CpG hypermethylation of the other allele in the PPCG cohort, we developed a computational integrative methodology, encompassing genomic, transcriptomic and epigenomic data with associated clinical annotations (from hereon termed "Epi2Hit", Fig. 2a, methods). We excluded CpG sites that correlated with stromal and immune cell content in the tumours, as well as those affected by global demethylation in partially methylated domains (PMDs)

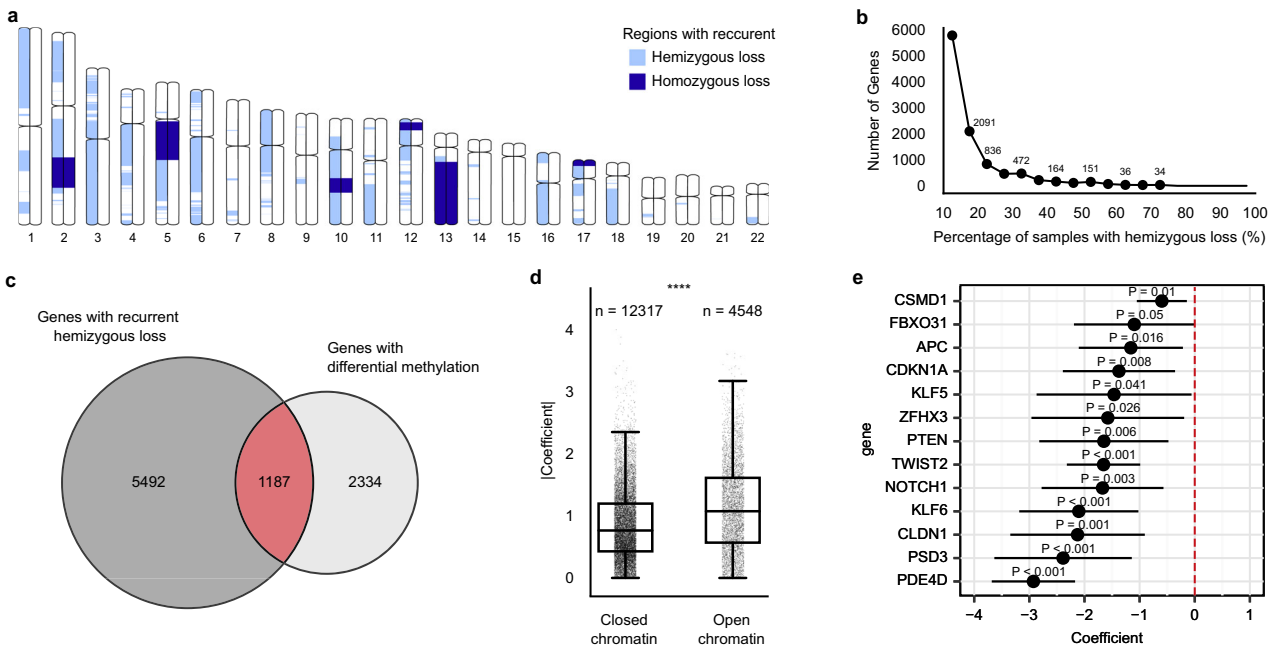

Fig. 1 | Prostate cancer is characterised by frequent hemizygous loss of putative TSGs. a Chromosome ideogram representing frequently (>10%) deleted regions (hemizygous loss - light blue, homozygous loss - dark blue) in the cohort. **b** The distribution of the number of genes with hemizygous loss across samples. The plot shows a decrease in the number of genes with hemizygous loss as the percentage of samples with affected genes increases, revealing the genes that are consistently affected across most of the samples. **c** Venn diagram showing the overlap between genes with recurrent hemizygous loss and differential methylation. The diagram highlights 1187 genes that exhibit both events. **d** Box plot of absolute coefficients of gene expression regressed on proximal CpG sites in closed ($n = 12,317$ CpGs) or open chromatin ($n = 4,548$ CpGs) P-value (****$P = 1.09 \times 10^{-115}$): two-sided Mann-Whitney U test). Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. **e** Forest plot represents the linear regression coefficients of top tumour suppressor genes ($n = 13$); dots indicate the estimated coefficient and horizontal lines show the 95% confidence interval. P values for each gene are shown in the plot. Source data are provided as a Source Data file.

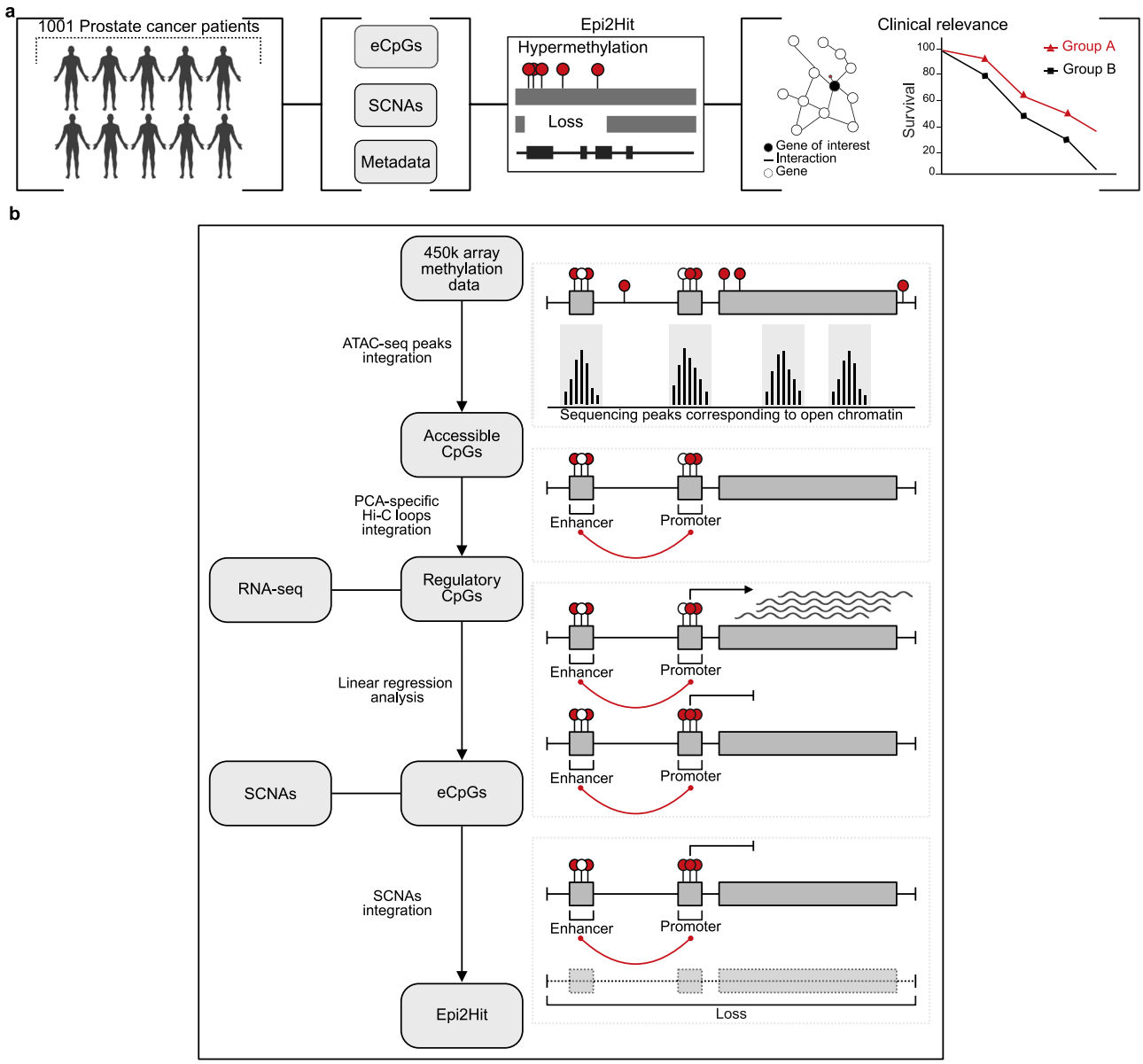

Fig. 2 | Epi2Hit pipeline. a Schematic overview of the Epi2Hit workflow. **b** Flowchart describes the steps of the pipeline for identification of eCpGs and Epi2Hit candidates. eCpGs are identified as CpG sites intersecting ATAC peaks (Accessible CpGs), annotated with the presence of promoter-enhancer loops (Regulatory CpGs). These CpGs are used for linear regression on gene expression to identify eCpGs. SCNA data is intersected with eCpGs to identify genes with biallelic inactivation involving epigenetic silencing. Created in BioRender. Weischenfeldt, J. (2026) https://BioRender.com/vfgm91z.

(Supplementary Fig. 1b). We further prioritised CpG sites located in open chromatin regions (using external ATAC-seq peak data). Regions were also annotated with external Hi-C based chromatin loops (GSE164347)[37], enhancer and promoter regions based on prostate-specific ChromHMM[38], to identify regulatory CpG, sites with a likely gene regulatory function (Fig. 2b, Supplementary Fig. 1c, d, Supplementary Fig. 2e). We regressed RNA-seq based gene expression on these regulatory CpGs to identify expression CpG sites whose methylation levels were significantly and negatively correlated with gene expression, and restricted our analysis to CpG sites displaying multimodal distribution, to select CpGs with differential methylation levels in prostate cancer epithelial cells across the cohort (eCpGs, $n = 300$, Supplementary Fig. 2c, d). These associations were stable when adjusting the model for tumour purity and fraction genome altered (FGA) (Supplementary Fig. 1e, f, Supplementary Table 2, methods). Finally, we integrated methylation levels of eCpGs with copy number losses from the same tumour, requiring the CpG site to overlap the copy number loss, to identify Epi2Hit candidate genes ($n = 155$, Fig. 2b, Supplementary Fig. 1c, d) and investigate their biological and clinical relevance in prostate cancer.

Epi2Hit-based discovery of tumour suppressors with epigenetic inactivation mechanisms

We systematically identified genomic-based biallelic inactivation events in our companion paper[39] including prostate cancer-specific TSGs from an external database[40] to discover known and poorly characterised TSGs also subjected to epigenetic-based inactivation. Among the TSGs that are inactivated in more than 2% of the PPCG cohort, we identified numerous TSGs with known and previously published tumour-suppressor function as Epi2Hit candidates (Fig. 3a som.loss/methyl), including *PSD3*[41] ($n = 48$, 5%), *ZFHX3*[42,43] ($n = 47$, 5%), *RBPMS*[44,45] ($n = 41$, 4%), *APC*[46,47] (n = 30, 3%), *PDE4D*[48–50] ($n = 20$, 2%) and *KLF5*[51,52] ($n = 17$, 2%). Consistent with a tumour-suppressive role, DepMap CRISPR dependency data (CERES scores) support that these

https://doi.org/10.1038/s41467-026-72182-5

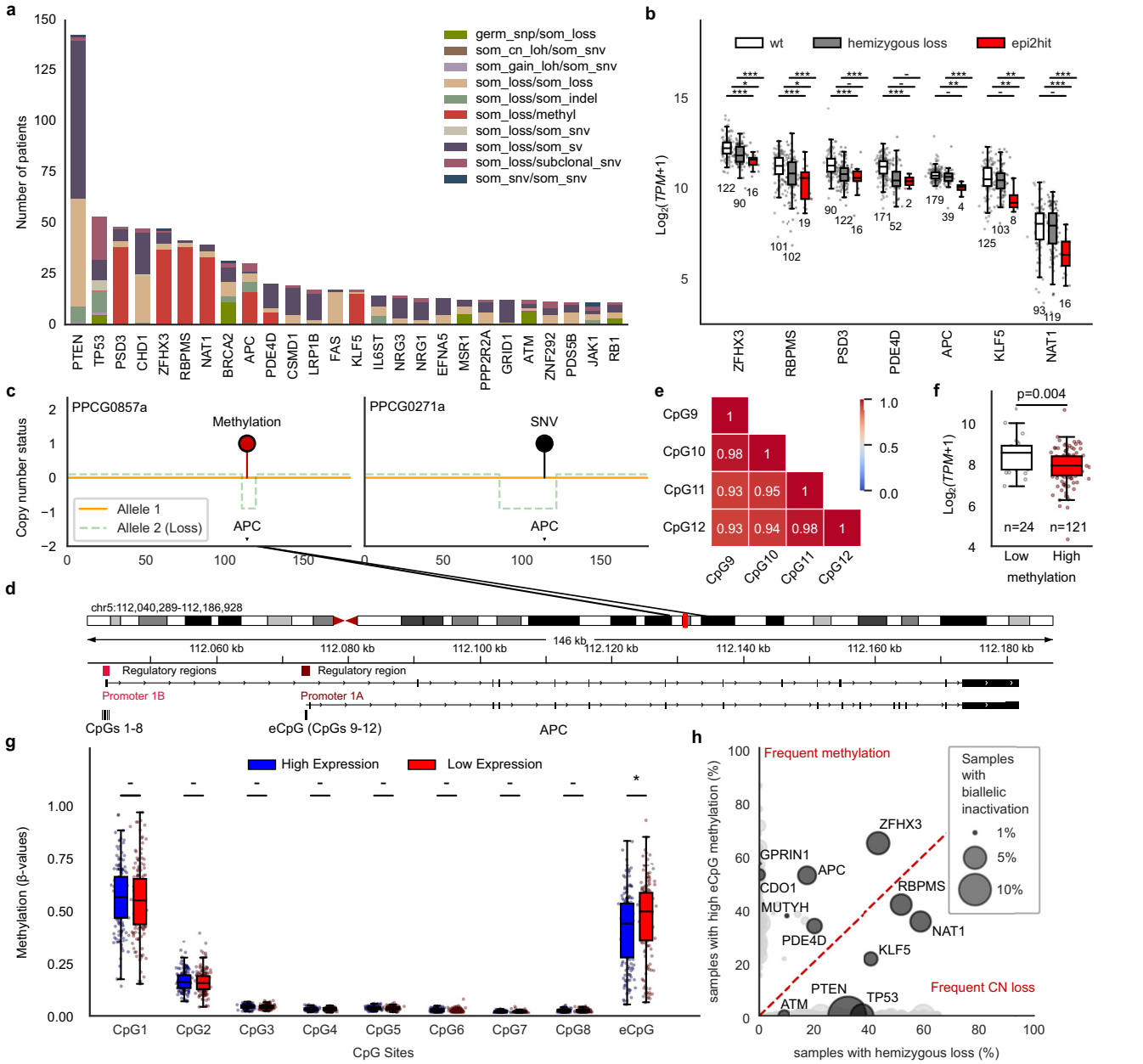

Fig. 3 | Epi2Hit workflow identifies TSGs susceptible to biallelic inactivation through the combination of hemizygous loss and hypermethylation. a Bar plot indicates the number of patients harbouring different types of biallelic inactivation for the 2% most frequent TSGs. **b** Box plots showing the expression of TSGs for wt (white), hemizygous loss (grey), Epi2Hit (red) groups of samples. Number of samples is indicated for each group. Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. P-value (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, - ns): two-sided Mann-Whitney U test for each comparing pair. **c** Somatic events leading to biallelic inactivation of the *APC* gene in two distinct cases. In case PPCG0857 *APC* is inactivated through a combined copy number loss and hypermethylation. In case PPCG0271 *APC* is inactivated by a copy number loss and an SNV. Y-axis displays the relative copy number across chromosome 5 and the *APC* locus. **d** IGV plot of the *APC* locus highlighting CpGs and eCpG sites (black). eCpG sites are located in a regulatory region. **e** Correlation matrix (Pearson) of four (cg23938220, cg02511809, cg24332422, cg15020645) *APC* CpG sites. Positive correlations are indicated in red. **f** Box plot shows *APC* expression levels between samples with *APC* High and *APC* Low eCpG methylation (two-sided Mann-Whitney U test, $P = 0.004$). Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. Number of samples is indicated for each group. **g** Box plots showing methylation (β-values) at different *APC* methylated CpGs (1-12) for high (above median, blue, $n = 168$) and low (below median, red, $n = 168$) expression groups. Each point represents one tumour sample. Centre line, median; box, 25–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR of the lower and upper quartiles; P-value (*$P < 0.05$ ($P = 0.01$), - ns): two-sided Mann-Whitney U test. **h** Scatter plot summarising sample-level frequencies of genetic and epigenetic inactivation. For each gene, the x-axis shows the percentage of tumours with hemizygous loss, and the y-axis shows the percentage of the same tumours with high eCpG methylation. Each point represents one gene, and bubble size reflects the percentage of tumours with biallelic hits (co-occurrence of hemizygous loss and hypermethylation in the same samples). Source data are provided as a Source Data file.

genes are generally non-essential across cancer cell lines, including prostate cancer lines (Score > (−1), Supplementary Table 3). To assess the transcriptional consequences of Epi2Hit, we compared gene expression levels between wild-type, hemizygous loss, and Epi2Hit samples (Fig. 3b), and identified a significant reduction in expression in Epi2Hit samples. We found a high frequency of hemizygous loss across multiple genes located on the chromosome 8p arm (Supplementary Fig. 3b). Chromosome 8p is deleted in more than 60% of prostate

cancers[39,53,54], but the actual TSG(s) have remained elusive due to the absence of recurrent homozygous losses. Although rare, we observed recurrent epigenetic inactivation of several genes including *NKX3-1* (1%), previously recognised as a candidate TSG in chr8p[55], suggesting that several TSGs in this region may contribute to tumour suppression through Epi2Hit.

APC is a known TSG, often subjected to promoter hypermethylation and gene expression downregulation[26]. The *APC* locus includes two promoters, promoter 1 A and promoter 1B, which regulate distinct transcripts and show different methylation patterns in normal and cancer tissue[23]. Promoter 1B generates three transcript isoforms that are predominant in normal colonic mucosa, whereas promoter 1 A, produces transcript 1 A and is more frequently methylated in cancer. In gastrointestinal malignancies, promoter 1 A is often hypermethylated while promoter 1B remains unmethylated, a pattern associated with reduced *APC* expression. The *APC* locus contains 12 regulatory CpGs (Fig. 3d), four of which we identified as eCpG (CpGs9-12) coinciding with promoter 1 A, which we found to display strong methylation correlation (Fig. 3c–e). Indeed, we found no correlation with gene expression when inspecting methylation levels of each of the first eight CpG sites associated with promoter 1B, whereas the methylation level of the eCpG identified by Epi2Hit pipeline (CpG9-12) displayed a strong anti-correlation with *APC* gene expression ($P = 0.01$) (Fig. 3f, g), confirming these as bona fide eCpGs and validating the general approach in our methodology. This finding also supports that *APC* is mediated predominantly through methylation of promoter 1 A in prostate cancer as in other malignancies, whereas promoter 1B remains unaffected.

We next tested the extent to which biallelic inactivated genes tended to be inactivated by hemizygous loss, methylation or both (Epi2Hit) at the sample level. For each gene, we plotted the percentage of samples with hemizygous loss and eCpG hypermethylation-based inactivation (Fig. 3h), with genes at the diagonal representing exclusively the Epi2Hit mechanism. While several TSGs displayed exclusive hemizygous loss (Fig. 3h) as well another genomic second hit (Fig. 3a), such as *PTEN* and *TP53*, we found several genes with frequent Epi2Hit including *PDE4D*, a gene with established suppressor function in prostate cancer[56,57], and *KLF5*, previously shown to be inactivated in prostate cancer by copy-number loss and epigenetic silencing[14](Supplementary Fig. 2f, Supplementary Fig. 3a).

To assess the reproducibility of our findings, we performed two levels of analysis. First, we observed robust concordance with MethylMix[58], a multi-state model that identifies differential methylation states predictive of gene expression. This analysis confirmed the presence of regulatory CpGs near *APC, KLF5* and *ZFHX3*. Next, we performed external validation by applying Epi2Hit to TCGA-PRAD[13] ($n = 493$ tumours) using matched methylation, RNA-seq, CNAs and SVs. This analysis confirmed our Epi2Hit discovery genes in the PPCG cohort, with 6 of the 7 top Epi2Hit TSGs among the top most frequently altered genes in the TCGA-PRAD cohort (Supplementary Fig. 3c, Supplementary Tables 4-5). Reassuringly, Epi2Hit confirmed the key candidates including *APC, KLF5* and *ZFHX3* (Supplementary Fig. 3d, methods).

Biallelic inactivation of *ZFHX3* drives *MYC* expression and is linked with more aggressive prostate cancer

One of the genes we identified as frequently inactivated as an Epi2Hit is *ZFHX3*. *ZFHX3* is a zinc finger homeobox transcriptional repressor, known to bind and repress the transcription of numerous genes[59–63]. It was previously reported to be a putative TSG in prostate cancers with downregulation of the gene associated with higher proliferation, cell migration and aggressiveness[42,64–66]. The biallelic inactivation encompassed a copy number loss of one allele and eCpG hypermethylation of the other allele, located in the gene's promoter region. Tumours with hemizygous loss showed higher eCpG methylation levels compared to tumours with copy-number neutral states (Supplementary Fig. 4a, b), confirming the observed co-occurrence of copy number loss and hypermethylation at the Epi2Hit genes.

To investigate the identified eCpGs at the *ZFHX3* locus and validate the biallelic inactivation, we took advantage of Oxford Nanopore-based long-read sequencing of six prostate cancer samples with and without biallelic inactivation of *ZFHX3*. Using phased allele-specific copy number and methylation-based analyses, we identified in Epi2Hit prostate cancer examples ($n = 3$, Supplementary Table 6) allele 1 to be lost (green dashed line, Fig. 4a top) and a concomitant strong allele-specific hypermethylation (5mC) at the eCpG (cg16563255) of allele 2 (yellow line, respectively, Fig. 4a top), with beta values ranging from 0.82-0.9. In contrast, we found no allele-specific differences at the copy number and methylation level of control prostate cancer cases ($n = 3$) without *ZFHX3* Epi2Hit (Fig. 4a bottom). We found the eCpG region of *ZFHX3* to form long-range loops to several other proximal regions including a hypomethylated promoter region of *ZFHX3*, suggesting that this eCpG could represent a critical enhancer element in regulating the transcription of *ZFHX3*. To further explore the epigenetic state of the *ZFHX3* locus, we extended the analysis to 5-hydroxymethylcytosine (5hmC), DNA modification associated with dynamic and transcriptionally permissive state[67,68]. In Epi2Hit tumours, the retained allele exhibited high 5mC and absence of 5hmC, consistent with stable transcriptional silencing. In contrast, Epi2Hit-WT tumours exhibited intermediate 5mC together with low 5hmC across both alleles, suggesting a more dynamic regulatory configuration (Supplementary Fig. 4c).

We further investigated whether inactivation of *ZFHX3* co-occurs with other frequent genetic alterations known to drive prostate cancer progression. We identified co-occurrence of *ZFHX3* inactivation with loss of *PTEN* ($q = 0.007$) or loss of *TP53* ($q < 0.0001$, Supplementary Table 7), in keeping with a cooperative role[69,70]. *MYC* gain exhibited a striking interaction with *ZFHX3* inactivation (odds ratio = 6, $q < 0.0001$, Supplementary Table 7), with significant upregulation of *MYC* in prostate cancer samples harbouring *ZFHX3* Epi2Hit ($p = 0.03$, Fig. 4b), suggesting that *ZFHX3* inactivation contributes to *MYC* overexpression. To further contextualise its evolutionary role, we used the mutational exclusivity metric (Gerhauser et al., 2018[14], Supplementary Fig. 4d, methods). This analysis showed that *ZFHX3* displays a higher exclusivity score, indicating that its inactivation tends to occur as an early event in prostate cancer. To explore the mechanistic basis of this regulation, we examined ZFHX3 binding within the *MYC* locus using orthogonal chromatin data from colon adenocarcinoma and neural stem cell[71,72], together with prostate adenocarcinoma cell[43]. *ZFHX3* binds to a specific site within a prostate-specific promoter region at the *MYC* locus (Fig. 4c, Supplementary Fig. 4e), marked by active chromatin (H3K4me3), in keeping with *ZFHX3* to act as transcriptional repressor of *MYC* expression by binding to regulatory elements of *MYC*[43].

We identified 11 additional prostate cancer-related genes with ZFHX3 binding sites near their promoter, and therefore likely to be direct targets of ZFHX3. We found genes associated with tumour-suppressive functions such as *CDH1* and *MUTYH*[73–79] to be downregulated in *ZFHX3* Epi2Hit tumours and genes that drive oncogenic processes (e.g., *SF3B1*, *PLEC*)[80–85] showed higher expression levels in *ZFHX3* Epi2Hit cancers (Fig. 4d, methods). These observations suggest a key transcriptional regulatory impact upon *ZFHX3* inactivation in driving prostate cancers.

Indeed, *ZFHX3* inactivation was significantly enriched in patients with high Gleason grade group (GG3-5) compared to low Gleason grade group (GG1-2) (Chi-square test: $P = 0.0004$) (Fig. 4e). Moreover, patients with *ZFHX3* inactivation displayed shorter time to metastasis (log-rank test $P = 0.0007$, Fig. 4f), suggesting that *ZFHX3* inactivation is driving more aggressive disease, potentially through transcriptional upregulation of *MYC*.

https://doi.org/10.1038/s41467-026-72182-5

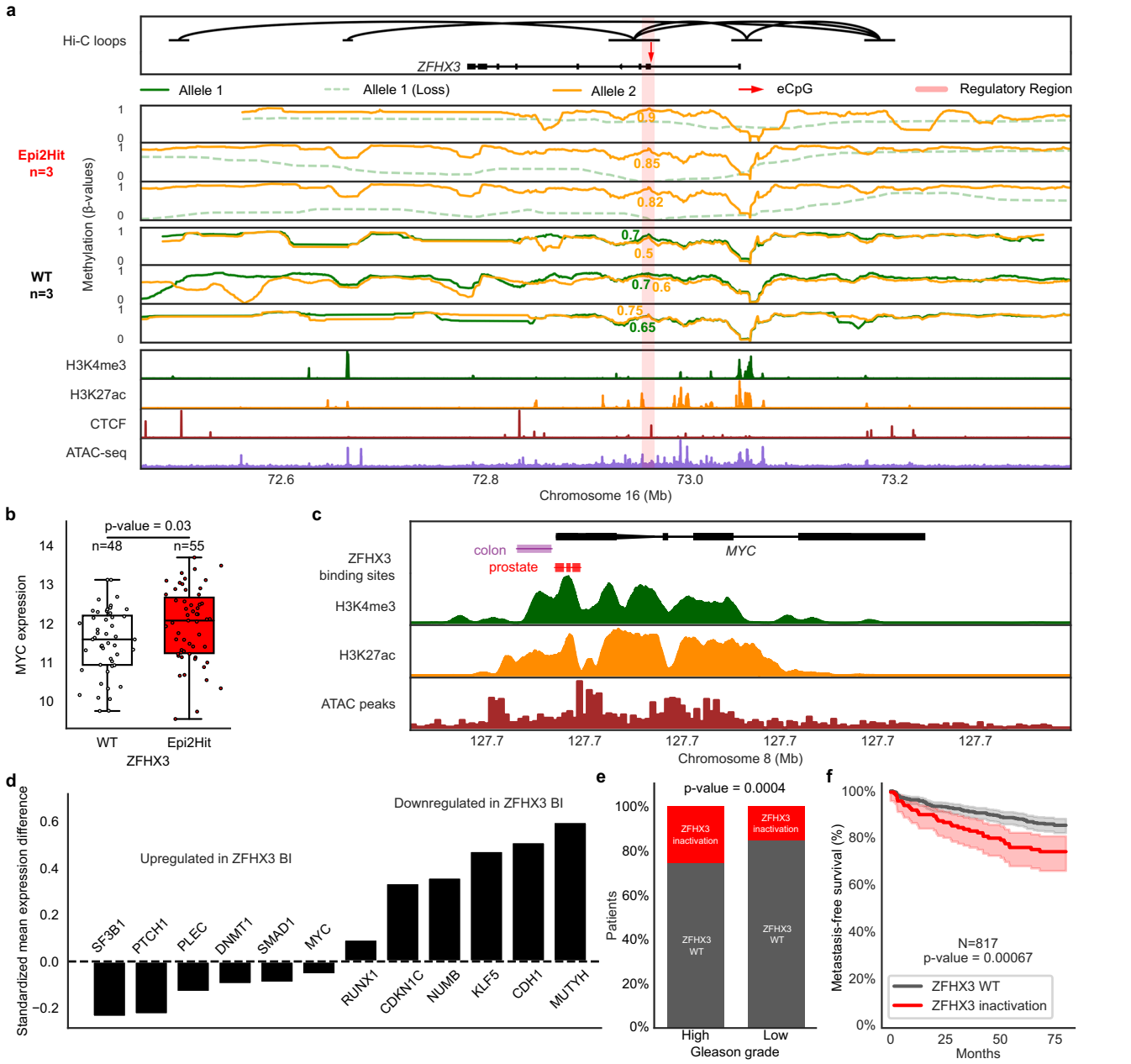

Fig. 4 | *ZFHX3* inactivation drives *MYC* upregulation. a Oxford Nanopore-based phased methylation and somatic copy number, and chromatin context at the ZFHX3 locus. Top: three Epi2Hit prostate cancers show loss of Allele 1 (green dashed) and allele-2 hypermethylation at the eCpG cg16563255 (red arrow; $\beta = 0.82–0.9$), with reduced ZFHX3 expression. Bottom: three Epi2Hit WT tumours show balanced allelic copy number and no allele-specific methylation. Hi-C loops illustrate chromatin interactions within the region. Prostate-specific chromatin marks - H3K4me3 and H3K27ac and CTCF, and prostate ATAC-seq peaks denote an active, accessible regulatory region overlapping the eCpG region. **b** Box plot shows *MYC* expression levels between *ZFHX3* WT and *ZFHX3* biallelic inactivation samples, illustrating a significant upregulation of MYC in *ZFHX3* biallelic inactivation cases (two-sided Mann-Whitney U test, $P = 0.03$). Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. Number of samples is indicated for each group. **c** ZFHX3 binding at the MYC promoter with conserved sites in colon and prostate. Chromatin marks (H3K4me3, H3K27ac) and ATAC-seq peaks confirm a prostate-specific active regulatory context. **d** Bar plot showing standardised mean expression differences (Cohen's *d*) for genes involved in prostate cancer progression, highlighting their differential expression in *ZFHX3* WT vs. biallelic inactivation cases. Positive values indicate higher expression in *ZFHX3* WT samples, whereas negative values reflect higher expression in *ZFHX3* biallelic inactivation cases. **e** Bar plot illustrating the distribution of *ZFHX3* wild-type (WT) and inactivation cases across high and low Gleason grade groups (High: $n = 324$ [WT = 241, inactivation = 83]; Low: $n = 493$ [WT = 417, inactivation = 76]; total $n = 817$). The plot shows a significant enrichment of *ZFHX3* inactivation in the high Gleason group compared to the low Gleason group (Chi-square test, $P = 0.0004$). **f** Kaplan–Meier survival curves illustrating metastasis-free survival (MFS) stratified by ZFHX3 status (WT, $n = 377$; inactivation, $n = 440$; among patients with available metastasis-free survival data, $n = 817$). Shaded bands indicate the confidence interval around the Kaplan-Meier estimate. Group differences were assessed using the log-rank test. Source data are provided as a Source Data file.

Interestingly, in a companion study, we identify *ZFHX3* inactivation in a subgroup of prostate cancer patients characterised by losses of several key TSGs including *TP53, PTEN* and *FANCA*, and by shorter metastasis-free survival[39], but it was not identified as an independent prognostic factor in multivariable Cox models (Supplementary Fig. 5f), suggesting that *ZFHX3* is correlated with other markers of aggressive disease.

https://doi.org/10.1038/s41467-026-72182-5

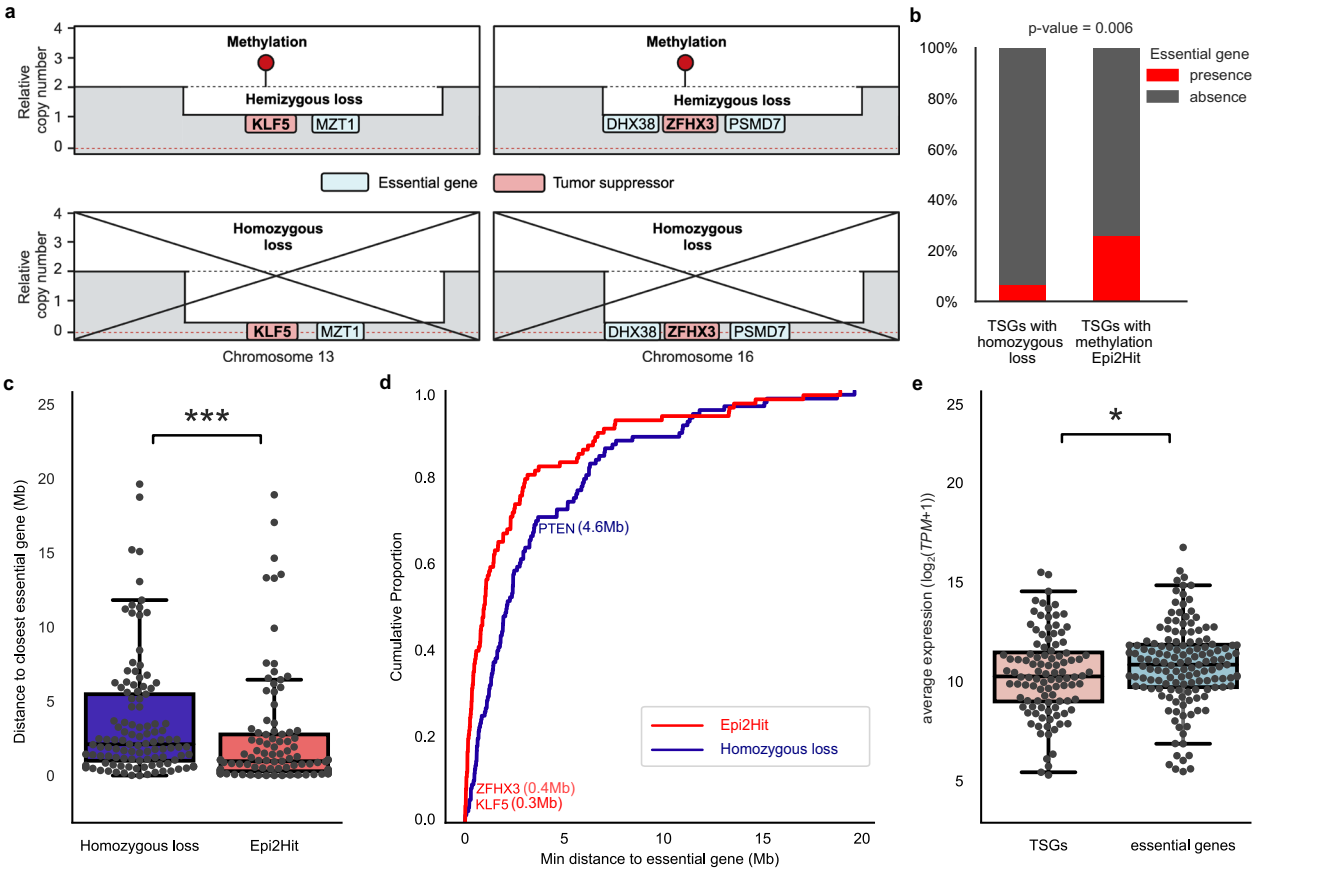

Fig. 5 | Nearby essential genes surrounding never-homozygous deleted tumour suppressor genes. a Schematic representation of the relationship between tumour suppressor genes (TSGs) and nearby essential genes on chromosomes 13 and 16. The presence of essential genes (blue) adjacent to TSGs (pink) limits the likelihood of homozygous deletions in these regions, leading to an increased probability of hemizygous loss and hypermethylation as alternative mechanisms of inactivation. Created in BioRender. Weischenfeldt, J. (2026) https://BioRender.com/hj1igem. **b** Stacked bar plot illustrating the distribution of TSGs with homozygous loss and TSGs with methylation (Epi2Hit) based on the presence (red) or absence (grey) of essential genes (homozygous loss: $n = 76$ [presence = 5, absence = 71]; methylation DoubleHit: $n = 46$ [presence = 12, absence = 34]; total $n = 122$) (from the DepMap database, with a score <−1.2, indicating high essentiality for cell survival). A Chi-square test of independence was used to assess the statistical significance of the observed differences, with a resulting $P = 0.006$. **c** Box plot comparing the distance to the closest essential gene for TSGs with homozygous loss (blue, $n = 113$) or Epi2Hit (red, $n = 103$). Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. ***$P < 0.001$ ($P = 0.00009$) (two-sided Mann–Whitney U test). **d** Empirical cumulative distribution function (ECDF) of the distance to the closest essential gene for TSGs affected in >2 patients by homozygous loss (blue) or Epi2Hit (red). *PTEN*, *KLF5* and *ZFHX3* are annotated. **e** Box plot comparing average gene expression ($\log_2$(TPM + 1)) of TSGs (pink, $n = 104$) and essential genes (blue, $n = 143$). Centre line, median; box, 25th–75th percentiles (IQR); whiskers, most extreme values within 1.5×IQR. *$P < 0.05$ ($P = 0.04$) (two-sided Mann–Whitney U test). Source data are provided as a Source Data file.

Proximity of essential genes influences tumour suppressor gene inactivation mechanisms

We sought to understand the difference in inactivating mechanisms observed in TSGs, such as *PTEN*, which frequently undergoes biallelic loss, compared to Epi2Hit genes such as *KLF5* and *ZFHX3* that exhibit loss of one allele and epigenetic silencing of the other allele. A previous study has suggested that TSGs subjected to biallelic loss are often located away from genes essential for cell survival[86]. We hypothesised that Epi2Hit genes would be less impacted by this constraint, since methylation of the wild-type allele would not affect the neighbouring genes compared to homozygous losses, which often encompass multiple neighbouring genes.

In contrast to homozygously deleted genes, we found Epi2Hit TSGs such as *KLF5* and *ZFHX3* indeed positioned in close proximity to common essential genes (Fig. 5a, methods). This proximity likely prevents their complete inactivation through double loss, favouring hemizygous deletion and hypermethylation as alternative inactivation mechanism. To further explore the enrichment of essential genes near TSGs undergoing Epi2Hit, we expanded our analysis to a broader list of genes identified by Epi2Hit. We observed a significant enrichment of essential genes near these TSGs (Chi-square test, $P = 0.006$, Fig. 5b), confirming that essential genes are more frequently located close to TSGs inactivated by Epi2Hits rather than by homozygous loss.

To corroborate this finding, we quantitatively assessed the genomic distance between each TSG and the closest essential gene. In support, Epi2Hit genes were significantly closer to essential genes than homozygous deleted TSGs (Fig. 5c,d). This association remained significant after adjusting for gene length and local gene density (Supplementary Fig. 5a, d, Supplementary Table 8-9, methods). Finally, to investigate whether expression differences might reflect selective constraints, we compared average expression levels between essential genes and TSGs. Essential genes showed significantly higher expression than TSGs (Fig. 5e), supporting the notion that preserving their expression may prevent adjacent homozygous loss. These findings underscores the importance of considering methylation as a second hit in the inactivation of TSGs, particularly those that are located in close proximity to essential genes. Moreover, it supports a model where proximity of essential genes constrain homozygous losses and favour alternative epigenetic mechanisms of inactivation, thereby avoiding the complete loss of function that could compromise tumour cell survival.

Discussion

Our study presents an integrative computational methodology to systematically identify biallelic inactivation of TSGs in cancer, integrating both genetic and epigenetic mechanisms - Epi2Hit. This workflow provides an orthogonal approach to deepen our understanding of prostate cancer initiation and progression by revealing how hemizygous deletions and hypermethylation of cis-regulatory regions cooperate to inactivate TSGs. Oncogenic mutational processes shaping prostate cancer genomes play a pivotal role[87]. In contrast, the impact of epigenetics is poorly understood. While previous studies have highlighted the role of hypermethylation and copy number loss in gene silencing, our approach identifies both previously known and understudied TSGs undergoing this complex inactivation process.

The strength of our methodology lies in its ability to identify putative Epi2Hit TSGs by leveraging CpG sites that are directly associated with gene expression changes, distinguishing them from non-functional CpGs. Our analysis of *APC* in prostate cancer further illustrates this principle: consistent with observations in other malignancies, epigenetic inactivation was driven predominantly by methylation of promoter 1A, while promoter 1B remained unaffected[23], underscoring the importance of promoter-specific regulation in tumour suppressor gene silencing. This precise identification of eCpGs, in conjunction with data from large cohorts, enables the detection of TSGs that are biallelically inactivated through hemizygous loss and hypermethylation, broadening our catalogue of potential cancer drivers. Notably, we identified known and understudied TSGs including *PSD3*, *PDE4D*, *RBPMS*, *APC* and *ZFHX3*, as undergoing Epi2Hit, emphasising the versatility of our approach in uncovering TSGs. *ZFHX3* Epi2Hit displayed a surprising recurrence, making it the fifth most frequent biallelic inactivated TSG in our analysis of primary prostate cancers across 1001 donors, superseding established TSGs such as *BRCA2* and *APC*. Our findings confirm and extend prior reports, showing a pervasive role in transcriptional regulation, with *ZFHX3* inactivation leading to relief of oncogene transcriptional repression such as *MYC*[43,88]. In support of an important and early driver role (methods, Supplementary Fig. 3d), we identified frequent early gains of *MYC* and loss of *ZFHX3* in an evolutionary Ordering-III, associated with a shorter time to metastasis in our companion paper[89].

While several prior studies have explored biallelic inactivation through combined deletion and hypermethylation[27,28], they focused on small cohorts, targeted methylation panels, or preselected TSGs. In contrast, Epi2Hit performs genome-wide discovery, integrating gene expression and regulatory annotation (ATAC-sequencing, Hi-C loops) to systematically prioritise functionally relevant CpG sites. A methodological comparison (Supplementary Table 10) highlights how Epi2Hit advances previous efforts by providing a modular tool applicable across different cancer types, and by leveraging regulatory and transcriptional data to improve specificity and biological interpretability of epigenetic hits.

The applicability of the Epi2Hit approach is not limited to prostate cancer. The principles underlying our methodology can be extended to other tumour types, where similar interactions between genetic deletions and epigenetic silencing may play a pivotal role in cancer progression[27,90,91]. Epi2Hit is compatible with large-scale, multi-omic datasets such as those generated by The Cancer Genome Atlas (TCGA), and could be applied to these cohorts to further validate its utility across diverse malignancies. We anticipate that our methodology will also enable the discovery of TSGs in other cancer types, and develop a more comprehensive understanding of the cancer genome landscape across various malignancies.

The identification of epigenetic silenced genes also opens up avenues for targeted treatments. Specifically, the methylation component of biallelic inactivation presents an attractive target for epigenetic therapies. Preclinical studies in prostate cancer cells have shown that treatment with the DNA-demethylating agent decitabine, particularly in combination with S-adenosylmethionine (SAM), exerts synergistic anti-tumour effects by reducing proliferation and invasion while promoting reactivation of silenced tumour suppressor pathways[92]. In clinical settings, azacitidine has been associated with reduced PSA doubling time and decreased *LINE-1* methylation levels in men with chemo-naïve castration-resistant prostate cancer, suggesting systemic demethylation and a favourable safety profile[93]. A systematic review by Linnekamp et al. further highlights the therapeutic relevance of demethylating agents across a broad range of solid tumours, including prostate, colorectal, and lung cancers, with early-phase clinical trials showing clinical responses in subsets of patients[94]. Notably, in prostate cancer, PSA decline was associated with gene-specific demethylation such as reduced *GADD45A* methylation observed in responders[95]. Beyond systemic agents, a locus-specific demethylation strategy using a TALE–TET1 fusion protein was shown to reactivate a tumour suppressor gene and inhibit metastatic progression in prostate cancer models, underscoring the promise of targeted epigenetic editing approaches[96]. Together, these findings highlight the translational potential of reversing methylation-mediated silencing as a therapeutic avenue in prostate cancer and beyond.

Our methodology also paves the way for more sophisticated analyses using long-read sequencing technologies. While the 450k array provides valuable insights into methylation patterns, its probe design is biased toward CpG islands and promoter regions, potentially missing key regulatory elements such as enhancers, including super-enhancers[97], which are often located in open sea or intergenic regions[98] and tend to have a stronger regulatory effect on transcription than promoter elements alone[99]. As a result, some epigenetically silenced TSGs may go undetected in our analysis, representing a source for future studies.

In contrast, unbiased methods such as whole genome bisulphite or nanopore sequencing allow the investigation of all eCpGs outside the predefined regions. Moreover, since the array platform does not allow different haplotypes to be distinguished, we were not able to detect methylation in combination with small variants (SNVs and indels) or SV breakpoints. Long-read sequencing can capture these events more accurately, enabling the detection of intricate allele-specific epigenetic landscapes in conjunction with genomic alterations on the same DNA fragment that are missed by traditional short-read technologies. Future studies applying WGBS or long-read methylation profiling in large tumour cohorts will likely identify additional functionally silenced TSGs that remain undetected in 450K-based analyses.

Prostate cancer is characterised by frequent and recurrent copy number losses. We find that homozygous losses are depleted near essential genes compared to Epi2Hit events, suggesting that these losses are constrained by the proximity of essential genes, where Epi2Hit mechanisms tend to be favoured. This model further implies that additional TSGs inactivated by Epi2Hit may be located near essential genes.

In summary, the Epi2Hit workflow enabled us to discover TSG inactivation mechanisms including *ZFHX3*, a key regulator of *MYC*, and enables a more systematic investigation of epigenetics as a tumour-promoting mechanism in cancer. Its applications in other cancer types, potential therapeutic implications, and future integration with long-read sequencing underscore its value as a versatile and powerful approach in cancer genomics.

Methods

We used data from the PPCG consortium of primary prostate cancer samples from a total of 1001 prostate cancer donors. Informed written ethical consent was obtained at clinical follow-up, and was consistent with local research ethics and International Cancer Genome Consortium (ICGC) guidelines (https://icgc.org/). Ethical approval was obtained from local research ethical committees.

The full list of approving boards/committees:

Human Research Ethics Committee at Melbourne Health (HREC 2006.073; HREC 2011.009; HREC 2012.275; HREC 2012.220; HREC 2016.087) (Australia).

University Health Network Research Ethics Board and CHU de Québec–Université Laval Research Ethics Board (UHN 06-0822-CE; UHN 11-0024-CE; CHUQc-UL 2012-913:H12-03-192) (Canada).

UBC Clinical Research Ethics Board (UBC REB H21-03722) (Canada).

The National Committee on Health Research Ethics (reference no. 1302791) and notification to the Danish Data Agency (no. 1-16-02-330-13) (Denmark).

CPP Ile de France IV Institutional Review Board (IRB 00003835) (France).

Ethik-Kommission der Ärztekammer Hamburg (PV3552; PV4445; PV4679) (Germany).

NHS East of England–Cambridge Research Ethics Committee (REC 3/0180) (UK).

NHS East Midlands Research Ethics Committee (01/4/061) (UK).

NHS London Research Ethics Committee (CCR 2075) (UK).

University of Pretoria Human Research Ethics Committee (FWA00002567; IRB00002235; IORG0001762; approval #43/2010) (South Africa).

St Vincent's Hospital Human Research Ethics Committee (#SVH/12/231; #SVH/15/227) (Australia).

Human Research Protection Office, US Army Medical Research and Development Command (approvals E02371 and E03280 for genomic interrogation of SAPCS samples) (USA).

Dataset

The study leverages data from the Pan-Prostate Cancer Group (PPCG) cohort, a harmonised, clinically annotated dataset comprising 2,021 prostate cancer donors across seven countries. Multi-modal molecular profiling includes whole genome sequencing (WGS), 450 K DNA methylation arrays, RNA-seq, and where available, long-read Oxford Nanopore Technologies (ONT) sequencing. Detailed clinical annotations including Gleason Grade Group, PSA levels, treatment, and survival outcomes are described in Jakobsdottir et al. [29].

Of the 2021 donors, 1001 had WGS (1209 samples), 1490 had 450 K methylation arrays (1928 samples), and 1292 had RNA-seq (1594 samples) available after quality control.

The details of the PPCG cohort are described in the accompanying marker resource paper[29].

WGS data and variant calling

Whole-genome sequencing (WGS) data were generated across multiple international centres as part of the PPCG consortium and processed using standardised, containerised pipelines to ensure reproducibility and harmonisation across sites. Two centralised workflows for somatic and germline variant calling were applied uniformly, as described in Jakobsdottir et al. [29].

Allele-specific somatic copy number alterations (CNAs) were inferred using the Battenberg algorithm (v2.2.9, https://github.com/Wedge-lab/battenberg), which phases heterozygous SNPs using the 1000 Genomes Project reference panel. B-allele frequency (BAF) and logR ratios were then jointly modelled to identify clonal and subclonal CNAs. The method incorporates statistical testing to assess clonality and estimates cancer cell fraction (CCF), tumour purity, and ploidy for each sample.

CNA profiles were generated for all tumour–normal WGS pairs. CNA call sets were manually reviewed by two analysts and samples were classified as satisfactory ($n = 1128$), unsatisfactory ($n = 24$), over-fragmented ($n = 13$), or samples without detectable copy-number alterations ($n = 301$). Only high-confidence CNA profiles (satisfactory samples) were retained for downstream analyses in Epi2Hit.

Hemizygous loss on autosomes were defined as segments with total copy number = 1 and CCF ≥ 0.9, ensuring high-confidence clonal deletion events. These calls rely on Battenberg's allele-specific modelling and purity/ploidy-aware segmentation, providing robust and consistent annotation of deletion events for integration with methylation and gene expression data.

Further processing and harmonisation of the PPCG WGS data are described in detail in the accompanying marker resource paper[29]

DNA methylation

450 K array-based methylation data were processed using a standardised and publicly available RnBeads-based workflow[100,101] (https://github.com/panprostate/PPCG_DNA_methylation), which integrates preprocessing, normalisation, and QC steps designed to harmonise large multi-cohort data, as described in Jakobsdottir et al. [29]. Briefly, Raw IDAT files were imported and preprocessed using RnBeads v2.15.1 with background subtraction, dye-bias correction, and bead count-based filtering. Low-quality samples (median bead count <3 or high probe-wise detection p-values > 0.05) were flagged and excluded. Probes were filtered out if they overlapped known SNPs (MAF > 0.05), were known to be cross-reactive[102], or had detection p-values ≥ 0.05 in ≥1% of samples. Methylation data were normalised using a single-sample approach implemented via the SeSAMe method[103], re-implemented in the RnBeads framework. Background correction was performed using the "noobsb" method, which fits a sample-specific background distribution based on out-of-band signal intensities from Type I probes and regresses this from foreground intensities. Dye-bias correction was conducted using internal channel scaling. This normalisation strategy enabled robust cross-cohort comparison without requiring batch metadata. Technical consistency across countries and cohorts was evaluated using density plots of 65 SNP-tagging CpG probes. Additionally, beta-value distributions were visually inspected, and principal component analysis (PCA) of normalised values confirmed that primary axes of variation reflected biological rather than technical differences. To further mitigate population structure effects, genotype-derived SNP principal components were included in downstream analyses.

Reference methylation data for major immune and stromal cell types were compiled from public datasets and in-house FACS-sorted prostate epithelial populations. To distinguish among 11 major cell types, the 200 most differentially methylated CpG sites per cell type (100 hypermethylated and 100 hypomethylated), along with CpGs within 50 bp, were selected[104], resulting in a set of 2851 CpGs. Cell-type proportions were then estimated using the Houseman algorithm[105,106] as implemented in RnBeads, and normalised to sum to 100% per sample for downstream adjustment in methylation-expression analyses. Estimated cell-type proportions were used as covariates in downstream methylation-expression modelling within the Epi2Hit framework.

Further processing and harmonisation of the PPCG Methylation data are described in detail in the accompanying marker resource paper[29].

Principal component analysis is shown in Supplementary Fig. 1a

RNA sequencing data

RNA-seq data were uniformly processed across samples using a harmonised pipeline based on STAR for alignment to the GRCh37 reference genome, followed by transcript-level quantification with Salmon (v1.4.0) using GENCODE v19 annotations (https://github.com/panprostate/RNA).

RNA-seq data from the PPCG cohort, generated across several centres using diverse chemistries and platforms, underwent a multi-step normalisation strategy to correct for both known and unknown sources of technical variation. Lowly expressed genes were filtered, retaining 22760 transcripts for downstream analysis. Tumour purity was estimated using ESTIMATE[107,108] and singscore methods.

Normalisation was performed using the RUV-III method[109] with pseudo-replicates of pseudo-samples (PRPS), implemented in both supervised and unsupervised modes. In the supervised approach, known biological covariates (ERG status, tumour purity) guided the selection of negative control genes. In the unsupervised approach, mutual nearest neighbours (MNN) and KNN-based algorithms were used to generate pseudo-replicates and identify control genes based on low biological variability and high association with technical noise. RUV-III-PRPS outperformed other tested methods (CPM, TMM, upper-quartile, VST, ComBat) in preserving biological signals while removing technical artefacts. All normalisation code is publicly available at https://github.com/RMolania/RUVprps.

Further processing and harmonisation of the PPCG RNA data are described in detail in Jakobsdottir et al. [29].

Principal component analysis is shown in Supplementary Fig. 1a

Multi-omic data integration strategy

Due to the sparsity data representation, a stepwise integration strategy was employed to maximise the use of available data while minimising loss of statistical power. First, regulatory CpGs were identified using the full set of available DNA methylation profiles ($n = 1268$) together with ATAC-seq data from an independent cohort (GSE188797) and further annotated with external Hi-C-based chromatin loops (GSE164347). Next, expression-correlated CpGs (eCpGs) were determined within the subset of samples for which both methylation and RNA-seq data were available ($n = 1292$). Finally, candidate biallelic inactivation events were evaluated by integrating methylation data with WGS-derived copy number profiles ($n = 1001$). This modular approach enabled robust cross-modal analysis without restricting the study to only the samples profiled in all three data types.

Hi-C data

Hi-C data from prostate cancer were retrieved from the Gene Expression Omnibus (GEO) from data sets GSE164347[37].

For integration with Oxford Nanopore (ONT) long-read sequencing data, which was processed using the hg38 genome build, we performed liftover of the relevant Hi-C loop regions from hg19 to hg38 using pyliftover v0.4.1 to enable visualisation and plotting in the hg38 coordinate space.

ChIP-seq data

Processed ChIP-seq data for ZFHX3 was obtained from the Gene Expression Omnibus (GEO) (GSE49402, GSM1208743; Yan et al. 2013)[72] and from human neural stem cells (Pérez Baca et al., 2024)[71,72]. To provide prostate-specific evidence, we additionally used published ChIP-PCR data in C4-2B prostate cancer cells (Hu et al., 2019)[42,43], where primers targeting three regions of the MYC promoter (Regions A–C) were designed. These primer sequences were validated in silico using the UCSC isPcr tool against the hg38 assembly, which confirmed single, specific amplicons within the MYC promoter (chr8:127,735,377–127,735,930). The identified regions overlapped with ZFHX3 binding peaks detected in neural stem cells and colon cells, confirming promoter-proximal binding across distinct cellular contexts.

liftover from hg18 to hg19 and from hg19 to hg38 was performed using pyliftover 0.4.1.

Epigenetic marks (H3K4me3, H3K27ac) and CTCF binding sites at the loci were derived for prostate cancer cells (PC-3, C4-2B) from the ENCODE project[110]

ATAC sequencing data

To identify CpG sites located within open chromatin regions, processed ATAC-seq data generated from human tumour samples was retrieved from the Gene Expression Omnibus (GEO) from data set GSE188797[36]

Long-read sequencing

Tumour and matched normal genomic DNA from six prostate cancer donors were prepared for Oxford Nanopore long-read whole-genome sequencing using the Ligation Sequencing Kit V14 kit (SQK-LSK114). The libraries were sequenced on a PromethION P24 instrument with R10.4.1 flow cells (FLO-PRO114M). Average library sequencing depth was 30x and 18x, range 19.5x-48.6x and 6.0x-48.0x, for tumour and normals, respectively. Basecalling was performed with Dorado v1.0.0 with model hac@v5.2.0, while 5mC and 5hmC modifications were called concurrently with model sup@v5.2.0_5mC_5hmC@v1. The data was then aligned to hg38 and phased with the epi2me-labs wf-human-variation workflow v2.7.1, performing genome-wide allele specific CpG methylation calling (5mC and 5hmC) with modkit v0.3.3.

For *ZFHX3*, phased haplotypes allowed us to distinguish the deleted allele (allele 1, copy number = 0) from the retained allele (allele 2). The expected pattern for biallelic inactivation is (i) absence of methylation signal on the deleted allele and (ii) strong hypermethylation on the retained allele.

Identification of regulatory CpG sites (eCpGs)

DNA methylation data. Methylation data was obtained from the 450k array methylation dataset provided by the PPCG. The data included β-values corresponding to the methylation levels at various CpG sites across multiple samples. Samples that failed quality control were excluded from the analysis. The remaining 1268 primary tumour samples were used for the analysis.

Annotation of CpG sites with gene information. Each site was annotated with the nearest gene(s) using the hg19 reference genome. This annotation was performed using the pybedtools package and the 'closest' function. The reference gene annotations were retrieved from the GENCODE v19 database. The CpG sites were matched to the closest genes, considering both strand orientation and distance. The output included all potential gene associations for each CpG site, facilitating further exploration of gene-specific methylation patterns.

Filtering based on methylation correlations. The dataset was subjected to a series of filtering steps to prioritise CpG sites with specific methylation patterns:

- Correlation with immune and stromal components: CpG sites were filtered based on their correlation with immune and stromal cells abundance. Pearson correlation coefficients were calculated for each site using methylation beta-values against a methylation-derived estimation of stromal and immune cell type proportions (described in detail in Jakobsdottir et al. [29]. Sites showing low correlation (correlation coefficient <0.4) with these components were retained, ensuring that the selected CpG sites primarily reflect methylation changes in tumour cells rather than in the surrounding stromal or immune cells.
- Correlation with global loss of methylation: Additionally, CpG sites were filtered based on their correlation with global loss of methylation estimated by the EpiCMIT-hypo score[111] (as described in Jakobsdottir et al.[29].), reflecting epigenetically-determined Cumulative MIToses which result in stochastic loss of methylation and the formation of partially methylated domains (PMDs).

ChromHMM annotation for regulatory elements. Filtered CpG sites were annotated with chromatin states using prostate-specific ChromHMM data from Pomerantz et al. [38]. ChromHMM annotations were intersected with CpG sites using pybedtools, and states were mapped to corresponding categories (e.g., Active Prostate Lineage-Specific Promoter). This step identified CpG sites within enhancer and promoter regions.

Merging CpG sites by methylation correlation. CpG sites annotated within the same regulatory region were merged if their methylation levels exhibited a Pearson correlation coefficient above 0.6. Merged sites were represented by the average methylation value and spanned the combined genomic coordinates of the individual CpG sites.

Hi-C data integration for chromatin loops. Hi-C data from prostate cancer (GEO; GSE164347) was used to annotate regulatory CpG sites with chromatin loop information. CpG sites overlapping loop anchor regions were identified, and corresponding loop IDs were assigned to these sites. Sites associated with loops were indicated by a 'Ls_' prefix in their identifiers.

Linear regression analysis to identify expression-associated CpG sites. We performed linear regression to identify CpG sites associated with gene expression (eCpGs) by regressing methylation levels against RNA-seq-based gene expression from the PPCG group.

- Regression modelling: ordinary Least Squares (OLS) regression was performed using the statsmodels package, treating methylation as the independent variable and gene expression as the dependent variable. Regression coefficients, p-values, confidence intervals, and R-squared values were extracted.
- Multiple testing correction: P-values were adjusted for multiple comparisons using the Benjamini-Hochberg False Discovery Rate (FDR) method from the statsmodels package (Ruiz-Arenas et al. [112]; Advani et al., *Nat Commun* 2024[113]; Ando et al., *Nat Commun* 2019[114]), Supplementary Data 2.
- eCpG identification: CpG sites with significant negative associations (adjusted p-value < 0.05) were classified as eCpGs. To ensure biological relevance, we further applied a regression coefficient threshold of < −0.4, which corresponds to an approximate 25% reduction in gene expression across the methylation range. Since gene expression was modelled as $\log_2(\mathrm{TPM}+1)$, a coefficient of −0.4 implies a moderate-to-strong transcriptional repression effect per unit increase in methylation. This level of effect size is consistent with enhancer and promoter methylation–expression associations previously reported in cancer cohorts[99] (Supplementary Data 2).

Sensitivity analyses for tumour purity and global genomic instability

Purity-adjusted methylation–expression regression. To assess the robustness of methylation-expression associations to tumour purity, we re-fit the regression models used to define eCpGs. In the baseline analysis, for each CpG–gene pair we modelled gene expression as a function of methylation using ordinary least squares (OLS):

$$expression = \alpha + \beta \cdot methylation + \varepsilon \quad (1)$$

$$expression = \alpha + \beta \cdot methylation + \gamma \cdot purity + \varepsilon \quad (2)$$

For all CpG-gene pairs considered in the main analysis we extracted the methylation coefficients (β) and R2 values from both models and compared their distributions. The direction of the methylation effect was preserved and the association remained significant after Benjamini–Hochberg false discovery rate (BH-FDR) correction ($q < 0.05$). Genome-wide concordance of methylation effect sizes between baseline and purity-adjusted models was summarised using Pearson and Spearman correlations (Pearson $r = 0.996$, Spearman $\rho = 0.996$) (Supplementary Fig. 1e).

To further account for global genomic instability, we additionally fit a two-covariate model including both tumour purity and fraction genome altered (FGA):

$$expression = \alpha + \beta \cdot methylation + \gamma \cdot purity + \delta \cdot FGA + \varepsilon \quad (3)$$

Methylation coefficients from this model were compared to the baseline model in the same way (Pearson r = 0.969, Spearman $\rho = 0.964$) (Supplementary Fig. 1f).

For loci highlighted in Fig. 3a (*ZFHX3, PDE4D, PSD3, KLF5, NAT1, APC* and *RBPMS*), we report locus-level regression results with and without purity adjustment, including methylation coefficients, p-values, FDR-adjusted q-values and R2 (Supplementary Table 2), to illustrate the stability of effect direction and magnitude at key Epi2Hit loci.

Categorisation of methylation Levels. Methylation levels for each feature were categorised into three groups: 'Not/Low_methyl', 'Medium_methyl', and 'Highly_methyl'. The classification was based on predefined thresholds (down_threshold=0.2 and up_threshold=0.7)[115–118] and the distribution of methylation values. Features with all values below the down_threshold were categorised as 'Not/Low_methyl', while those with values above the up_threshold were labelled 'Highly_methyl'. Features with methylation values spanning these thresholds were further analysed using kernel density estimation (KDE) and peak detection to identify potential multimodal distributions.

For features with intermediate methylation levels, a Gaussian kernel density estimation (KDE) was applied to estimate the probability density function (PDF) of the methylation values.

The KDE was performed using the scipy.stats.gaussian_kde function.

Peaks in the KDE were identified using the scipy.signal.find_peaks function. Depending on the number of detected peaks, the feature's methylation values were categorised accordingly:

To ensure reproducibility and robustness, we conducted a sensitivity analysis across three key KDE parameters: distance, height, and prominence, which define the minimum distance between peaks, the KDE height threshold, and the peak prominence threshold respectively.

A total of 305 CpG sites were selected from the prostate cancer cohort, each visually confirmed to exhibit bimodal or trimodal methylation distributions. We then systematically evaluated 125,000 parameter combinations generated from a three-dimensional grid (50 linearly spaced values from 0.01–0.5 for each parameter; see Supplementary Data 3).

Performance was measured by the number of CpGs for which the algorithm correctly detected 2–3 peaks. The optimal parameter range was defined as:

- distance_frac: 0.01–0.11
- height_quantile: 0.01–0.33
- prominence_frac: consistently optimal at 0.01

The midpoint values (distance_frac=0.06, height_quantile=0.17, prominence_frac=0.01) were selected as defaults for downstream analyses and recovered 94% (287/305) of the benchmark regions.

Pairwise KDE parameter performance is visualised in Supplementary Fig. 2d, where warmer (red) regions indicate higher peak detection accuracy across the parameter grid.

This parameter configuration was further validated on an independent luminal breast cancer cohort (TCGA-BRCA[119]), where an additional 450 manually selected CpG sites were analysed. Using the same settings, the KDE algorithm correctly identified 1, 2 or 3 peaks in 443/450 (98.4%) regions, confirming the sensitivity of the approach (Supplementary Data 4-5).

Depending on the number of detected peaks, the feature's methylation values were categorised accordingly:

- Single peak: a Gaussian Mixture Model (GMM) with three components was fitted to the methylation values using sklearn.mixture.GaussianMixture(https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html). The resulting

clusters were used to assign methylation categories based on the cluster means, with thresholds calculated from the median and minimum of these means.

We chose GMM for unimodal distributions based on its successful use in prior studies[116], where it stratified promoter methylation into binary states. In contrast, our method applies GMM at the CpG level and partitions the distribution into three states: 'Not/Low_methyl', 'Medium_methyl', and 'Highly_methyl', capturing partial methylation patterns frequently observed in cancer. GMM is only used when KDE fails to detect more than one peak, ensuring method selection is driven by the distributional shape.

A visual overview of the KDE vs. GMM logic, is shown in Supplementary Fig. 2a.

- Two peaks: features displaying a bimodal distribution were manually reviewed, with the lower and upper peak values serving as thresholds for categorising the methylation levels into *'Not/Low_methyl'*, *'Medium_methyl'*, and *'Highly_methyl'*.
- Three peaks: features with a trimodal distribution were also manually reviewed, using the median and minimum of the peak values as thresholds for classification.

If no peaks were detected or distributions were ambiguous, conservative fallback quantile thresholds (quant_low=0.3, quant_high=0.7) were used.

To identify genes that are biallelic inactivated we used our previously published methodology[14,39,120], expanding it to DNA methylation data.

Loss/Methylation: A heterozygous loss combined with high methylation (som_loss/methyl).

Validation of Epi2Hit in TCGA-PRAD

TCGA-PRAD data acquisition and preprocessing. For external validation, we applied Epi2Hit to an independent prostate cancer cohort from TCGA-PRAD. DNA methylation (450k array β-values), RNA-seq expression, and copy-number alteration (CNA) profiles for primary tumours (n = 493) were downloaded from cBioPortal(https://www.cbioportal.org/), and structural variant (SV) calls were obtained from TCGA-PRAD whole-genome sequencing samples in the NCI Genomic Data Commons (GDC(https://portal.gdc.cancer.gov/); accessed 22 July 2025). Only samples with matched methylation, expression and CNA data were retained. Gene expression values were transformed to $\log_2(\text{TPM} + 1)$.

Re-implementation of the Epi2Hit workflow in TCGA. We re-implemented the complete Epi2Hit pipeline on TCGA-PRAD using the same parameter settings as for PPCG. Briefly, CpG sites were annotated to genes, filtered (removal of CpGs correlated with immune/stromal estimates or global PMD-associated hypomethylation, restriction to prostate ChromHMM promoters/enhancers, intersection with prostate tumour ATAC-seq peaks, and prioritisation of CpGs at Hi-C loop anchors), and tested for association with gene expression to define eCpGs. CpG-level p-values were adjusted using the BH-FDR, and within-gene multiplicity was controlled using Simes-based aggregation[121] of eCpG p-values to obtain per-gene q-values. For each gene, methylation distributions at eCpGs were then modelled with KDE/GMM as described above to obtain discrete methylation states, which were integrated with CNAs and SVs to call Epi2Hit events (heterozygous loss plus hypermethylation).

Comparison of biallelic inactivation patterns across cohorts. For the most frequently disrupted tumour suppressor genes (TSGs), we quantified the number of samples in each cohort harbouring genomic biallelic hits versus Epi2Hit events These counts were summarised as bar plots for TCGA-PRAD and PPCG (Supplementary Fig. 3c, Supplementary Table 4-5).

MethylMix validation

Construction of gene-level methylation profiles. To compare Epi2Hit with a MethylMix framework for methylation-driven genes, we used the pre-eCpG set defined in the PPCG discovery cohort (prior to KDE/GMM-based peak detection).

Gene-level methylation and expression matrices were analysed with the MethylMix algorithm (R package MethylMix 2.0(https://bioconductor.org/packages/release/bioc/html/MethylMix.html)) using recommended settings.

Overlap with Epi2Hit and visualisation of methylation states. To quantify concordance, we intersected the list of methylation-driven genes with the set of genes harbouring Epi2Hit events in PPCG. Overall, 47% of methylation-driven genes were also identified by Epi2Hit, including key tumour suppressors such as *APC, KLF5, NAT1* and *ZFHX3*. For these four genes, we visualised the MethylMix beta mixture models as density curves overlaid on β-value histograms (Supplementary Fig. 3d), illustrating the presence of 2–3 distinct methylation states per gene.

Standardised mean difference analysis (Cohen's *d*)

We first used publicly available ChIP-seq data (GSE49402) to identify genes with ZFHX3 binding sites located within ±2 kb of their annotated transcription start sites, representing putative promoter-proximal targets of *ZFHX3*. From this list, we selected genes previously reported in the literature to have tumour-suppressive or oncogenic functions in prostate cancer. This final set of genes was used for expression analysis.

To quantify the differences in gene expression between *ZFHX3* wild-type (WT) and epigenetic biallelic inactivation (Epi2Hit) prostate cancer samples, we computed the standardised mean difference (Cohen's *d*) for the preselected genes. For each gene, we calculated the mean expression in the WT and Epi2Hit groups and divided their difference by the pooled standard deviation across the two groups. This approach allows comparison of expression differences across genes on a standardised scale, independent of absolute expression values.

The pooled standard deviation (s_p) was computed as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{4}$$

where n_1 and n_2 are the sample sizes of the WT and Epi2Hit groups, respectively, and s_1 and s_2 are their respective standard deviations.

Cohen's d is then:

$$d = \frac{x_1 - x_2}{s_p} \tag{5}$$

Positive values of d indicate higher expression in the WT group, and negative values indicate higher expression in the Epi2Hit group.

Essential genes analysis

To investigate the association between essential genes and tumour suppressor genes (TSGs) that are rarely or never inactivated by biallelic loss, we utilised data from the DepMap(https://depmap.org/portal/) database. Genes with scores below this threshold (<−1.2) were considered essential.

Calculation of proximity to essential genes

The distance between each TSG and the closest essential gene was calculated as the absolute genomic distance in megabases (Mb), considering both upstream (5′) and downstream (3′) directions. For each TSG, the minimum of the two distances (5′ or 3′) to the nearest essential gene was used. The presence of an essential gene near a TSG was annotated as True if the distance between the TSG and the nearest essential gene was less than the first quantile (Q1) of the mean distance (dTSG <Q1) across all TSGs. This criterion was used to assess the pattern of essential genes in relation to TSGs that are rarely or never inactivated by biallelic loss.

For comparative analyses, we included only TSGs affected by Epi2Hit or homozygous deletion in more than two tumour samples. Differences in the distance to essential genes between these groups were assessed using the two-sided Mann–Whitney U test.

To explore functional consequences, we also compared average gene expression ($\log_2(TPM+1)$) between essential genes and TSGs using RNA-seq data from the same cohort. Statistical significance was evaluated using the Mann–Whitney U test (Fig. 5e).

Structural covariates and multivariable models

To assess whether the proximity signal could be explained by structural genomic features and to address potential confounding, we assembled a unified gene-level dataset for all TSGs that included:

- Gene length (bp, converted to kb),
- Local gene density in a ± 250 kb window (gene_density_500kb),
- Distance to the centromere (dist_to_centromere, in kb),
- Proximity to essential genes at several DepMap CERES thresholds.

For each DepMap essentiality cutoff X∈{CERES < − 0.2, < − 0.5, < − 0.7, < − 1.0, < − 1.2, < − 1.5}, we computed:

- dist_CERES_X: distance (kb) to the nearest essential gene,
- count_CERES_X: number of essential genes within 500 kb,
- has_dist_CERES_X_within_500kb: indicator of ≥1 essential neighbour within 500 kb.

All distance-like variables were rescaled to kilobases for interpretability. These variables were used as predictors in multivariable logistic regression models with outcome Epi2Hit vs homozygous loss for each CERES threshold. Each model included gene length, local gene density, distance to centromere, number of nearby essential genes (count_CERES_X), distance to the nearest essential gene (dist_CERES_X) and the 500 kb proximity indicator. Odds ratios (ORs) and 95% confidence intervals are summarised in Supplementary Fig. 5a, b and Supplementary Table 8. At the main common-essential threshold (CERES <−1), both the number of essential neighbours and, to a lesser extent, their proximity remained significantly associated with Epi2Hit status after adjustment for gene length, gene density and chromosomal context.

Prostate-specific essentiality

To evaluate whether these patterns hold in a prostate-specific context, we repeated the analysis using only prostate cancer cell lines (22RV1, LNCaP, DU145) from DepMap. Essentiality scores were recalculated across the same CERES thresholds, and prostate-specific proximity measures (count_CERES_X, dist_CERES_X, has_dist_CERES_X_within_500kb) were recomputed for each TSG. Multivariable logistic models with the same covariates (gene length, gene density, distance to centromere and essential-gene proximity) were then refitted. The number and proximity of prostate-essential neighbours remained strong predictors of Epi2Hit versus homozygous loss across thresholds (Supplementary Fig. 5c, d, Supplementary Table 9), supporting a genuine constraint by essential neighbours' effect rather than structural artefacts or dependence on a particular essentiality definition.

Copy number loss frequency (ideogram-level)

For each ideogram interval (chromosome, start, end; hg19 autosomes), we calculated the proportion of tumour samples (n = 991) with overlapping homozygous loss (norm_total = −2) or hemizygous loss (norm_total = −1) in the segmentation data. Overlap was defined as segment_start <region_end and segment_end > region_start. Homozygous loss (hom %) and hemizygous-only loss (het %; excluding samples that also carried a homozygous loss in that region) were computed separately. Their sum (loss %) represented the total fraction of samples with any loss in the interval, without double-counting. Only intervals with loss % > 10% of samples were retained for ideogram plotting.

The chromosome ideogram was generated using the Python package *quyuan* (v1.1.3; https://github.com/tcztzy/quyuan).

Copy number loss frequency (chromosome-level)

For each autosomal chromosome, we determined the proportion of tumour samples (n = 991) with at least one segment meeting the loss criterion anywhere on that chromosome. Homozygous loss (hom %, norm_total = −2) and hemizygous loss (het %, norm_total = −1) were calculated separately, and the union of these sample sets (loss %) represented the total percentage of tumours with any loss on that chromosome, without double-counting those carrying both types of loss.

Exclusivity score of copy number alterations

To quantify the tendency of driver genes to occur independently rather than co-occur with other driver genes across cancer samples, we applied the exclusivity score metric, as described by Gerhauser et al. (Cancer Cell, 2018)[14]. This score estimates the degree to which a given gene alteration is exclusive or co-occurs with other gene alterations.

For each driver gene, the exclusivity score is calculated as the mean inverse of the number of co-altered driver genes in samples where the gene is present. Specifically, for each sample with an alteration of the target gene, the reciprocal sum of driver genes with the alteration is computed and the exclusivity score is calculated as the arithmetic mean of the reciprocal values across all relevant samples.

Formally, the exclusivity score is defined as:

$$Exclusivity\ Score = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{c_i} \tag{6}$$

Where:

n is the number of samples containing the target gene,

c_i is the number of driver genes altered in a sample

The exclusivity score ranges from 0 to 1. Higher values indicate greater exclusivity (occurring as the only driver alteration), while lower values suggest frequent co-occurrence with other driver genes.

Survival analysis

Lifelines package (lifelines.KaplanMeierFitter and lifelines.CoxPHFitter) (https://github.com/lifelines/lifelines) was used for survival analysis across subgroups.

Statistics & reproducibility

All statistical tests were two-tailed, unless otherwise specified. For parametric data, we used tests assuming normal distribution. For non-parametric data, we used distribution-free tests. Significance thresholds and specific tests are detailed in the figure legends and Methods.

The Epi2Hit pipeline was implemented in Python and comprises modular steps for methylation processing, expression-methylation correlation, probe filtering and merging, gene annotation, chromatin state mapping, and integration with Hi-C, ATAC-seq, and long-read

sequencing data. Default parameters and thresholds (CNV, KDE, regression covariates) are detailed in Supplementary Table 11.

The pipeline supports parallel processing through the joblib library (e.g., n_jobs=16) to accelerate steps such as correlation-based probe merging. All analyses were run on a shared Linux server with 64 CPU cores and 512 GB of RAM. Jobs were typically submitted with a memory request of up to 40 GB RAM. Runtime for the full pipeline applied to the PPCG cohort (~1000 samples and ~450,000 probes) was approximately 3-4 h, depending on data modalities and filtering settings. Due to its modular and parallelizable structure, the pipeline scales efficiently with increasing cohort size and is readily applicable to other tumour types and multi-omic datasets.

Epi2Hit is distributed as a Docker image available at https://hub.docker.com/r/al0ka/epi2hit. The container includes all necessary dependencies and project scripts and is available for both x86_64 (linux/amd64) and ARM64 (linux/arm64) architectures. Users can obtain the image by executing "docker pull al0ka/epi2hit:latest" and then run the workflow by following the included documentation. The image was tested using the included example data on an Ubuntu 24.04.3 LTS machine with 4GB of RAM and 4 CPU cores. The analysis consisted of three steps: the initial step (eCpG identification - run-epi2hit_py) took approximately 7 min, the second step (generating metadata.yaml) was nearly instantaneous, completing in about 1 s, and the final step (running biallelic_py) took approximately 1 min.

All data analyses were performed using Python 3.8 (pandas v1.1.5, numpy v1.19.5, lifelines 0.27.7, scipy v1.5.4, and scikit-learn v0.23.2, statsmodels v0.13.2, joblib v1.2.0, tqdm v4.64.1, plotnine v0.10.1, pybedtools v0.10.0, pyBigWig v0.3.22, statannot v0.2.3) and and R 4.2.2 (MethylMix v2.41.0).

The plots and visualisations were generated using the seaborn v0.11.0, matplotlib v3.6.3, and additional figures were created using BioRender (BioRender.com) and quyuan (v1.1.3).

Reporting summary

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

Data availability

PPCG dataset comprises whole-genome sequencing (WGS), RNA-seq, DNA methylation, clinical, and histopathological data from 2,021 prostate cancer donors. Due to the sensitive nature of human genomic/clinical data and participant-consent restrictions, PPCG data are provided under controlled access and are available only for approved research purposes. Raw genome sequencing data have been deposited in the European Genome-Phenome 1360 Archive (EGA) under the study ID EGAS00001002876 and is described in the Reporting Summary and our companion manuscript[29]. In addition, PPCG working project data (e.g., variant calls, pipeline outputs and intermediate analyses) and harmonised RNA-seq/methylation data are made available to approved projects through secure transfer (Globus), and clinical/histopathological data are hosted in a GDPR-compliant secure environment (REDCap); access is limited to approved PPCG researchers. Access to PPCG controlled data requires submission of a Project Concept Form, available at https://panprostate.org, and approval by the PPCG Steering Committee. Requests are submitted by emailing the completed form to PPCG@icr.ac.uk. We will respond to the data access request within 14 days. We will provide guidance on the application process; access, once granted, is provided for the duration of the approved project under the applicable data access agreement/terms. To support reproducibility without redistribution of controlled PPCG data, we provide a compact, self-contained example dataset in the project repository that is sufficient to run the Epi2Hit workflow end-to-end and reproduce representative outputs. Specifically, the repository includes a TCGA-PRAD chromosome 8 example dataset (450 K methylation β-values for chr8, a matched gene expression matrix, prostate ATAC-seq peaks, ChromHMM segmentation, and Hi-C loops), together with example mutation calls and copy-number segment inputs used for downstream biallelic analyses. These files are provided under the repository's data/ directory and are used by analysis/TCGA_run.ipynb (and the containerised demo) to reproduce the main workflow outputs on the example dataset. The publicly available data used in this study are available in the NCBI Gene Expression Omnibus (GEO) database under accession code GSE164347 (prostate Hi-C loops): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164347[37]. The publicly available data used in this study are available in the GEO database under accession code GSE188797 (tumour ATAC-seq peaks): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE188797[36]. The publicly available data used in this study are available in the GEO database under accession code GSE49402 (GSM1208743, ZFHX3 ChIP-seq): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49402[72]. The publicly available data used in this study are available from the ENCODE Project database (epigenetic marks (H3K4me3, H3K27ac) and CTCF tracks derived for prostate cancer cells (PC-3, C4-2B)): https://www.encodeproject.org/[110]. The remaining data are available within the Article, Supplementary Information or Source Data file. Source data are provided with this paper.

Code availability

Code used to generate results and plots are available at https://github.com/panprostate/Epi2Hit. A static version of the code is available at https://doi.org/10.5281/ZENODO.18788787[122].

References

1. Cancer today. https://gco.iarc.who.int/today/.
2. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clinicians* **71**, 209–249 (2021).
3. Quigley, D. A. et al. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **175**, 889 (2018).
4. Sun, J. et al. DNA copy number alterations in prostate cancers: a combined analysis of published CGH studies. *Prostate* **67**, 692–700 (2007).
5. Taylor, B.S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
6. Liu, W. et al. Deletion of a small consensus region at 6q15, including the MAP3K7 gene, is significantly associated with high-grade prostate cancers. *Clin. Cancer Res.* **13**, 5028–5033 (2007).
7. Lapointe, J. et al. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res.* **67**, 8504–8510 (2007).
8. Gläsker, S. et al. Reconsideration of biallelic inactivation of the VHL tumour suppressor gene in hemangioblastomas of the central nervous system. *J. Neurol. Neurosurg. Psychiatry* **70**, 644–648 (2001).
9. Kwabi-Addo, B. et al. Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. *Proc. Natl. Acad. Sci. USA* **98**, 11563–11568 (2001).
10. Alers, J. C. et al. Identification of genetic markers for prostatic cancer progression. *Lab. Invest.* **80**, 931–942 (2000).
11. Camacho, N. et al. Appraising the relevance of DNA copy number loss and gain in prostate cancer using whole genome DNA sequence data. *PLoS Genet.* **13**, e1007001 (2017).
12. Knudson, A. G. Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. USA.* **68**, 820–823 (1971).
13. Cancer Genome Atlas Research Network The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
14. Gerhauser, C. et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996–1011 (2018).

15. Wedge, D. C. et al. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. *Nat. Genet.* **50**, 682–692 (2018).
16. Carnevali, I. W. et al. MLH1 promoter methylation could be the second hit in lynch syndrome carcinogenesis. *Genes* **14**, 2060 (2023).
17. Plotkin, A., Olkhov-Mitsel, E. & Nofech-Mozes, S. MLH1 methylation testing as an integral component of universal endometrial cancer screening—a critical appraisal. *Cancers* **15**, 5188 (2023).
18. Stefansson, O. A. et al. BRCA1 promoter methylation status in 1031 primary breast cancers predicts favorable outcomes following chemotherapy. *JNCI Cancer Spectrum* **4**, pkz100 (2020).
19. Zhang, L. & Long, X. Association of BRCA1 promoter methylation with sporadic breast cancers: evidence from 40 studies. *Sci. Rep.* **5**, 1–12 (2015).
20. Kondrashova, O. et al. Methylation of all BRCA1 copies predicts response to the PARP inhibitor rucaparib in ovarian carcinoma. *Nat. Commun.* **9**, 1–16 (2018).
21. Rivera, A. L. et al. MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro. Oncol.* **12**, 116 (2010).
22. Leske, H. et al. MGMT methylation pattern of long-term and short-term survivors of glioblastoma reveals CpGs of the enhancer region to be of high prognostic value. *Acta Neuropathologica Commun.* **11**, 1–12 (2023).
23. Zhu, L., Li, X., Yuan, Y., Dong, C. & Yang, M. APC Promoter methylation in gastrointestinal cancer. *Front. Oncol.* **11**, 653222 (2021).
24. Van der Auwera, I. et al. Aberrant methylation of the Adenomatous Polyposis Coli (APC) gene promoter is associated with the inflammatory breast cancer phenotype. *Br. J. Cancer* **99**, 1735–1742 (2008).
25. Wang, Y., Fan, C., Yu, J. & Wang, X. APC methylation predicts biochemical recurrence of patients with prostate cancer: a meta-analysis. *Int. J. Clin. Exp. Med.* **8**, 15575 (2015).
26. Chen, Y. et al. APC gene hypermethylation and prostate cancer: a systematic review and meta-analysis. *Eur. J. Hum. Genet.* **21**, 929–935 (2013).
27. Zardo, G. et al. Integrated genomic and epigenomic analyses pinpoint biallelic gene inactivation in tumors. *Nat. Genet.* **32**, 453–458 (2002).
28. Nishiki, H. et al. Integrated analysis of somatic DNA variants and DNA methylation of tumor suppressor genes in colorectal cancer. *Int. J. Mol. Sci.* **26**, 1642 2025).
29. Maria Jakobsdottir, G. The pan prostate cancer group: an international consortium analysing multiomic data for clinical discovery. *Nature Scientific Data* (The University of Melbourne, 2026).
30. Velez, M. G. et al. Differential impact of tumor suppressor gene (TP53, PTEN, RB1) alterations and treatment outcomes in metastatic, hormone-sensitive prostate cancer. *Prostate Cancer Prostatic Dis.* **25**, 479–483 (2021).
31. Ku, S.-Y., Gleave, M. E. & Beltran, H. Towards precision oncology in advanced prostate cancer. *Nat. Rev. Urol.* **16**, 645–654 (2019).
32. Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. *Cancer Genet.* **207**, 474–488 (2014).
33. Möller, K. et al. Chromosome 5 harbors two independent deletion hotspots at 5q13 and 5q21 that characterize biologically different subsets of aggressive prostate cancer. *Int. J. Cancer* **148**, 748–758 (2021).
34. van Dekken, H. et al. Genetic evaluation of localized prostate cancer in a cohort of forty patients: gain of distal 8q discriminates between progressors and nonprogressors. *Lab. Invest.* **83**, 789–796 (2003).
35. Gausachs, M. et al. MLH1 promoter hypermethylation in the analytical algorithm of Lynch syndrome: a cost-effectiveness study. *Eur. J. Hum. Genet.* **20**, 762–768 (2012).
36. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, 1898 (2018).
37. Hawley, J. R. et al. Reorganization of the 3D genome pinpoints noncoding drivers of primary prostate tumors. *Cancer Res.* **81**, 5833–5848 (2021).
38. Pomerantz, M. M. et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat. Genet.* **52**, 790 (2020).
39. Pellegrina, D. et al. The molecular heterogeneity and clinical impact of genomic driver alterations in prostate cancer. *Nat Commun.* (in press).
40. Tumor suppressor gene database (TSGene) Home. https://bioinfo.uth.edu/TSGene/.
41. Zhu, X., Liu, J., Xu, X., Zhang, C. & Dai, D. The Pleckstrin and Sec7 domain-containing gene as a novel epigenetic modification marker in human gastric cancer and its clinical significance. *Int. J. Oncol.* **46**, 195–204 (2015).
42. Hu, Q. et al. ZFHX3 acts as a tumor suppressor in prostate cancer by targeting FTO-mediated m6A demethylation. *Cell Death Discov.* **10**, 284 (2024).
43. Hu, Q. et al. ZFHX3 is indispensable for ERβ to inhibit cell proliferation via MYC downregulation in prostate cancer cells. *Oncogenesis* **8**, 28 (2019).
44. Rabelo-Fernández, R. J. et al. Reduced RBPMS levels promote cell proliferation and decrease cisplatin sensitivity in ovarian cancer cells. *Int. J. Mol. Sci.* **23**, 535 (2022).
45. Yu, J. et al. RBPMS inhibits bladder cancer metastasis by down-regulating MYC pathway through alternative splicing of ANKRD10. *Commun. Biol.* **8**, 367 (2025).
46. Bruxvoort, K. J. et al. Inactivation of Apc in the mouse prostate causes prostate carcinoma. *Cancer Res* **67**, 2490–2496 (2007).
47. Powell, S. M. et al. APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
48. Rahrmann, E. P. et al. Identification of PDE4D as a proliferation promoting factor in prostate cancer using a Sleeping Beauty transposon-based somatic mutagenesis screen. *Cancer Res.* **69**, 4388–4397 (2009).
49. Powers, G. L. et al. Phosphodiesterase 4D inhibitors limit prostate cancer growth potential. *Mol. Cancer Res.* **13**, 149–160 (2015).
50. Geybels, M. S. et al. Epigenomic profiling of prostate cancer identifies differentially methylated genes in TMPRSS2:ERG fusion-positive versus fusion-negative tumors. *Clin. Epigenet.* **7**, 128 (2015).
51. Xing, C. et al. Klf5 deletion promotes Pten deletion-initiated luminal-type mouse prostate tumors through multiple oncogenic signaling pathways. *Neoplasia* **16**, 883–899 (2014).
52. Li, J. et al. KLF5 is crucial for androgen-AR signaling to transactivate genes and promote cell proliferation in prostate cancer cells. *Cancers (Basel)* **12**, 748 (2020).
53. Bergerheim, U. S. R., Kunimi, K., Peter Collins, V. & Ekman, P. Deletion mapping of chromosomes 8, 10, and 16 in human prostatic carcinoma. *Genes Chromosomes Cancer* **3**, 215–220 (1991).
54. Macoska, J. A. et al. Evidence for three tumor suppressor gene loci on chromosome 8p in human prostate cancer. *Cancer Res.* **55**, 5390–5395 (1995).
55. He, W. W. et al. A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* **43**, 69–77 (1997).
56. Gulliver, C., Huss, S., Semjonow, A., Baillie, G. S. & Hoffmann, R. Loss of PDE4D7 expression promotes androgen independence, neuroendocrine differentiation and alterations in DNA repair:

implications for therapeutic strategies. *Br. J. Cancer* **129**, 1462–1476 (2023).
57. Böttcher, R. et al. Human phosphodiesterase 4D7 (PDE4D7) expression is increased in TMPRSS2-ERG-positive primary prostate cancer and independently adds to a reduced risk of post-surgical disease progression. *Br. J. Cancer* **113**, 1502–1511 (2015).
58. Gevaert, O. *MethylMix*. (Bioconductor, 2017). https://doi.org/10.18129/B9.BIOC.METHYLMIX.
59. Mao, L., Huang, W., Zou, P., Dang, X. & Zeng, X. The unrecognized role of tumor suppressor genes in atrial fibrillation. *Gene* **642**, 26–31 (2018).
60. Miura, Y. et al. Cloning and characterization of an ATBF1 isoform that expresses in a neuronal differentiation-dependent manner (*). *J. Biol. Chem.* **270**, 26840–26848 (1995).
61. Zhao, D. et al. Transcription factor ZFHX3 regulates calcium influx in mammary epithelial cells in part via the TRPV6 calcium channel. *Biochem. Biophys. Res. Commun.* **519**, 366–371 (2019).
62. Parsons, M. J. et al. The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven axis. *Cell* **162**, 607–621 (2015).
63. Ninomiya, T. et al. Regulation of the alpha-fetoprotein gene by the isoforms of ATBF1 transcription factor in human hepatoma. *Hepatology* **35**, 82–87 (2002).
64. Sun, X. et al. Deletion of atbf1/zfhx3 in mouse prostate causes neoplastic lesions, likely by attenuation of membrane and secretory proteins and multiple signaling pathways. *Neoplasia* **16**, 377–389 (2014).
65. Dayoub, A. et al. Inactivation of PTEN and ZFHX3 in mammary epithelial cells alters patterns of collective cell migration. *Int. J. Mol. Sci.* **24**, 313 (2022).
66. Grasso, C. S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
67. Pastor, W. A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14**, 341–356 (2013).
68. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**, 45–68 (2014).
69. Sun, X. et al. Additive effect of Zfhx3/Atbf1 and Pten deletion on mouse prostatic tumorigenesis. *J. Genet. Genomics* **42**, 373 (2015).
70. Zhang, L. et al. Prognostic effect of coexisting TP53 and ZFHX3 mutations in non-small cell lung cancer patients treated with immune checkpoint inhibitors. *Scand. J. Immunol.* **94**, e13087 (2021).
71. Pérez Baca, M. D. R. et al. Haploinsufficiency of ZFHX3, encoding a key player in neuronal development, causes syndromic intellectual disability. *Am. J. Hum. Genet.* **111**, 509–528 (2024).
72. Yan, J. et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801–813 (2013).
73. Takayama, K.-I. et al. RUNX1, an androgen- and EZH2-regulated gene, has differential roles in AR-dependent and -independent prostate cancer. *Oncotarget* **6**, 2263 (2015).
74. The hallmarks of CDKN1C (p57, KIP2) in cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1816**, 50–56 (2011).
75. Belle, V. A., McDermott, N., Meunier, A. & Marignol, L. NUMB inhibition of NOTCH signalling as a therapeutic target in prostate cancer. *Nat. Rev. Urol.* **11**, 499 (2014).
76. Che, M. et al. Opposing transcriptional programs of KLF5 and AR emerge during therapy for advanced prostate cancer. *Nat. Commun.* **12**, 1–15 (2021).
77. Imtiaz, H. et al. Genetic polymorphisms in CDH1 and Exo1 genes elevate the prostate cancer risk in Bangladeshi population. *Tumor Biol.* https://doi.org/10.1177/1010428319830837 (2019).
78. Stone, L. Fatal interaction: a new target identified. *Nat. Rev. Urol.* **14**, 259–259 (2017).
79. Paller, C. J. et al. Pan-cancer interrogation of MUTYH variants reveals biallelic inactivation and defective base excision repair across a spectrum of solid tumors. *JCO Precis. Oncol.* (2024) https://doi.org/10.1200/PO.23.00251.
80. Buckup, M. et al. Plectin is a regulator of prostate cancer growth and metastasis. *Oncogene* **40**, 663 (2021).
81. Sheng, T. et al. Activation of the hedgehog pathway in advanced prostate cancer. *Mol. Cancer* **3**, 29 (2004).
82. Spliceosome component SF3B1 as novel prognostic biomarker and therapeutic target for prostate cancer. *Transl. Res.* **212**, 89–103 (2019).
83. Gravina, G. L. et al. Increased levels of DNA methyltransferases are associated with the tumorigenic capacity of prostate cancer cells. *Oncol. Rep.* **29**, 1189–1195 (2013).
84. Vo, B. T., Cody, B., Cao, Y. & Khan, S. A. Differential role of Sloan–Kettering Institute (Ski) protein in Nodal and transforming growth factor-beta (TGF-β)-induced SMAD signaling in prostate cancer cells. *Carcinogenesis* **33**, 2054 (2012).
85. Qiu, X. et al. MYC drives aggressive prostate cancer by disrupting transcriptional pause release at androgen receptor targets. *Nat. Commun.* **13**, 1–17 (2022).
86. Pertesi, M. et al. Essential genes shape cancer genomes through linear limitation of homozygous deletions. *Commun. Biol.* **2**, 1–11 (2019).
87. Andreas, J. G. et al. Integrated signatures reveal dominant mutational processes in prostate cancer. *Nature* https://doi.org/10.1038/s41586-026-10468-w (2026).
88. Sun, X. et al. Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nat. Genet.* **37**, 407–412 (2005).
89. Fernández-Sanromán, A. et al. Evolutionary trajectory and momentum shape prostate cancer outcomes. *Nat Commun.* (in press).
90. Tokumaru, Y. et al. Biallelic inactivation of the RIZ1 gene in human gastric cancer. *Oncogene* **22**, 6954–6958 (2003).
91. Guerrero-Preston, R. et al. Key tumor suppressor genes inactivated by 'greater promoter' methylation and somatic mutations in head and neck cancer. *Epigenetics* **9**, 1031–1046 (2014).
92. Schmidt, T. & Sticht, C. The simultaneous treatment of PC-3 cells with the DNA-demethylating agent decitabine and S-adenosylmethionine leads to synergistic anticancer effects. *Genes* **15**, 15121634 (2024).
93. Sonpavde, G. et al. Azacitidine favorably modulates PSA kinetics correlating with plasma DNA LINE-1 hypomethylation in men with chemonaïve castration-resistant prostate cancer. *Urol. Oncol.* **29**, 682–689 (2011).
94. Linnekamp, J. F., Butter, R., Spijker, R., Medema, J. P. & van Laarhoven, H. W. M. Clinical and biological effects of demethylating agents on solid tumours - A systematic review. *Cancer Treat. Rev.* **54**, 10–23 (2017).
95. Singal, R. et al. Phase I/II study of azacitidine, docetaxel, and prednisone in patients with metastatic castration-resistant prostate cancer previously treated with docetaxel-based therapy. *Clin. Genitourin. Cancer* **13**, 22–31 (2015).
96. Li, K. et al. Manipulation of prostate cancer metastasis by locus-specific modification of the CRMP4 promoter region using chimeric TALE DNA methyltransferase and demethylase. *Oncotarget* **6**, 10030–10044 (2015).
97. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).
98. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

99. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
100. Assenov, Y. et al. Comprehensive analysis of dna methylation data with rnbeads. *Nat. Methods* **11**, 1138–1140 (2014).
101. Muller, F. et al. Rnbeads 2.0: comprehensive analysis of dna methy1467 lation data. *Genome Biol.* **20**, 1–12 (1466).
102. Chen, L. et al. Genetic and epige1470 netic regulation of the organic cation transporter 3, slc22a3. *Pharmacogenomics* **13**, 110–120 (2013).
103. Zhou, W., Triche, T. J., Peter, W. & Laird, H. Sesame: reducing 1473 artifactual detection of dna methylation by infinium beadchips in genomic deletions. *Nucleic Acids Res.* **46**, e123–e123 (1474).
104. Moss, J. et al. Com1504 prehensive human cell-type methylation atlas reveals origins of circulating cell-free 1505 dna in health and disease. *Nat. Commun.* **9**, 5068 (1503).
105. Chen, L. et al. Genetic and epigenetic regulation of the organic cation transporter 3, SLC22A3. *Pharmacogenomics J.* **13**, 110–120 (2013).
106. Andres Houseman, E. et al. 1508 Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC 1509 Bioinforma.* **13**, 1–16 (2012).
107. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
108. Foroutan, M. et al. Single sample scoring of molecular phenotypes. *BMC Bioinforma.* **19**, 404 (2018).
109. Molania, R. et al. Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat. Biotechnol.* **41**, 82–95 (2023).
110. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
111. Duran-Ferrer, M. et al. The proliferative history shapes the DNA methylome of B-cell tumors and predicts clinical outcome. *Nat. Cancer* **1**, 1066–1081 (2020).
112. Ruiz-Arenas, C. et al. Identification of autosomal cis expression quantitative trait methylation (cis eQTMs) in children's blood. *eLife* **11**, 65310 (2022).
113. Advani, J. et al. QTL mapping of human retina DNA methylation identifies 87 gene-epigenome interactions in age-related macular degeneration. *Nat. Commun.* **15**, 1972 (2024).
114. Ando, M. et al. Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. *Nat. Commun.* **10**, 2188 (2019).
115. Ma, X., Wang, Y.-W., Zhang, M. Q. & Gazdar, A. F. DNA methylation data analysis and its application to cancer research. *Epigenomics* **5**, 301–316 (2013).
116. Zhang, W., Flemington, E. K., Deng, H.-W. & Zhang, K. Epigenetically silenced candidate tumor suppressor genes in prostate cancer: Identified by modeling methylation stratification and applied to progression prediction. *Cancer Epidemiol. Biomark. Prev.* **28**, 198–207 (2019).
117. Colaneri, A. et al. Expanded methyl-sensitive cut counting reveals hypomethylation as an epigenetic state that highlights functional sequences of the genome. *Proc. Natl. Acad. Sci. USA* **108**, 9715–9720 (2011).
118. Liu, H. et al. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res.* **44**, 75–94 (2016).
119. Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
120. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium Author correction: pan-cancer analysis of whole genomes. *Nature* **614**, E39 (2023).
121. Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986).
122. Kiriy, D. *Recurrent Mechanisms of Biallelic Epigenetic Inactivation Reveal New Putative Tumour Suppressor Genes in Prostate Cancer*. (Zenodo, 2026). https://doi.org/10.5281/ZENODO.18788787.

Acknowledgements

This work was performed as part of the Pan-Prostate Cancer Group (PPCG). We thank the patients, staff and scientists who contributed to this study. The following grants supported this work: J.W. was supported by the Danish Cancer Society #R147-A9843, Danish Cancer Society #R374-A22518, Danish Council for Independent Research #8020-00282, Danish Council for Independent Research #3101-00177 A, Novo Nordisk Foundation #NNF200C0060141, Sygeforsikringen Danmark # 2022-0198. D.K. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement number 101034290 (EMERALD International PhD Programme for Medical Doctors). This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant to J.R. (grant no. PJT-162410) and the Investigator Award to J.R. from the Ontario Institute for Cancer Research (OICR). Funding to OICR is provided by the Government of Ontario. This work was supported by the University of Konstanz and an Exploration grant of the Boehringer Ingelheim Foundation to A.J.G. G.M and B.H. is hosted by the Centro Nacional de Investigaciones Oncológicas (CNIO), which is supported by the Instituto de Salud Carlos III and recognised as a 'Severo Ochoa' Centre of Excellence (ref. CEX2019-000891-S) by the Spanish Ministry of Science and Innovation (MCIN/AEI/ 10.13039/501100011033). B.H. was also supported by a Spanish Ministry of Science and Innovation grant PID2019-111356RA-I00 and PID2023-151298OB-I00 (MCIN/AEI/ 10.13039/501100011033). B.H's postdoctoral research contract was supported by philanthropists via the 'Amigos/as del CNIO' Programme. B.H. was also supported by 'La Caixa' Foundation (ID 100010434; LCF/BQ/PR23/11980033). G.M. was also supported by a Spanish Ministry of Science and Innovation grants PID2019-111356RA-I00 and PID2023-151298OB-I00 (MCIN/AEI/ 10.13039/501100011033). This work is supported by a Cancer Research UK RadNet Manchester award [C1994/A28701], a CRUK Manchester Institute Core Senior Investigator award [C5759/A27412] and a Prostate Cancer UK FASTMAN Centre of Excellence award (MA-COE128-002). Part of the research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia. We acknowledge support from Cancer Research UK C5047/A29626, C5047/A22530, C309/A11566, C368/A6743, A368/A7990, and C14303/A17197; We acknowledge support from Prostate Cancer UK (MA-TIA23-002, MA-ETNA19-003, TLD-CAF22-011), the Dallaglio Foundation; and Prostate Cancer Research. We also acknowledge support from the Bob Champion Cancer Research Trust, Movember, Big C Cancer Charity, the Masonic Charitable Foundation, the King Family, the Stephen Hargrave Foundation, Alan Boswell Group, The Bob Willis Fund and the Bedford Memorial Trust. This work was supported in part by grants awarded to K.D.S from The Novo Nordisk Foundation (grant no. NNF20OC0059410, NNF21OC0071712), The Danish Cancer Society (grant no. R352-A20573), and Independent Research Fund Denmark (grant no. 9039-00084B). Bernard Pope was supported by a Victorian Health and Medical Research Fellowship from the Victoria State Government, Australia. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. ML's work is supported by the National Cancer Institute grants P50CA211024, P01CA265768, R01 CA259200, the USA Department of Defence (DoD) grants DoD PC160357, PC200390, as well as the Prostate Cancer Foundation (22CHAL05). This work was supported by NHMRC project grants 1104010 (CMH, AJC, NMC) and 1047581 (CMH, AJC, NMC), as well as a federal grant from the Australian Department of Health and Ageing to the Epworth Cancer Centre, Epworth Hospital (AJC, NMC,

CMH). In carrying out this research, we received funding and support from Australian Prostate Cancer Research and the University of Melbourne, Australia. Genomic sequencing and interrogation of Southern African Prostate Cancer Study (SAPCS) data was supported by the Ancestry and Health Genomics Laboratory at the University of Sydney through National Health and Medical Research Council (NHMRC) of Australia funding, including a Project Grant (APP1165762) and Ideas Grants (APP2001098, APP2010551), and U.S.A. Congressionally Directed Medical Research Programmes (CDMRP) Prostate Cancer Research Programme (PCRP) funding, including an Idea Development Award (PC200390, TARGET Africa) and HEROIC Consortium Award (PC210168 and PC230673, HEROIC PCaPH Africa1K). Further support for SAPCS analytical costs was provided by the U.S.A. National Institute of Health (NIH) National Cancer Institute (NCI) Award (1R01CA285772-01) and a U.S.A. Prostate Cancer Foundation (PCF) Challenge Award (2023CHAL4150). Hayes was further supported by the Petre Foundation via the University of Sydney Foundation (Australia). V.J.G acknowledges infrastructure support from the NIHR Cambridge Biomedical Research Centre (BRC-1215–20014). The French Prostate ICGC project was funded by Institut National de la Santé et de la Recherche Médicale (INSERM) and Institut National du Cancer (INCa); grant INSERM CV_2011/023 (C18), with additional support from LYric (grant INCa-4662). Analysis of the Australian samples was funded via a PRECEPT programme grant, co-funded by Movember and the Australian Federal Government. N.M.C also acknowledge support from Prostate Cancer Research, Prostate Cancer UK, the Bob Champion Cancer Research Trust, Movember, Big C Cancer Charity, the Masonic Charitable Foundation, the Prostate Cancer Research Centre, the King Family, The Andy Ripley Memorial Fund, and the Stephen Hargrave Foundation. Acknowledgement: The authors would like to thank those men with prostate cancer and the subjects who have donated their time and samples to the data sets used in this study. The study was funded by the ICGC EOPC (01KU1001 A) and ICGC-Data Mining (01KU1505A) projects on early-onset prostate cancer by the German Federal Ministry of Education and Research (BMBF). The authors acknowledge the DKFZ, MPI and EMBL Genomics Core Facilities.

Author contributions

D.K., F.F., C.G., and P.L. contributed to the formal analysis and visualisation presented in this study. F.G.R.G., A.L., J.S.J., and J.H. performed nanopore sequencing. D.K. and J.W. wrote the manuscript. J.W. conceived and designed the study. C.S.C., D.C.W., R.A.E., D.B., and Z.K.-J. made a substantial contribution to the organisation and conduct of the study and critiqued the output for important intellectual content. A.J.G., A.V.O., B.H., K.C.C., D.P., G.M., S.B., D.S.B., R.G.B., M.B., B.B., A.B., G.C., N.M.C., O.C., A.G., E.G.G., V.J.G., A.A.H., V.M.H., H.H., C.M.H., E.L.L.I., M.J., C.H.J., F.K., Z.K-J., P.L., G.G.L., M.L., L.M., R.M., A.P., B.P., L.R.Q., T.R., B.R., A.S., K.D.S., S.S.U., D.C.W., Y.X., T.Y., C.Z., T.S., J.R., and J.W. and all other members of the The Pan Prostate Cancer Group provided access to data and/or contributed to gathering, processing and curating data. All authors had access to all data in the study. All authors contributed to the review and the editing of the manuscript. All authors approved the manuscript before submission.

Competing interests

V.J.G. is honoraria from Janssen and JEB as a speaker. Member of external expert committee to Astra Zeneca UK. R.A.E. is Honoraria from GU-ASCO, Janssen, University of Chicago, Dana Farber Cancer Institute USA as a speaker. Educational honorarium from Bayer and Ipsen, member of external expert committee to Astra Zeneca UK and Member of Active Surveillance Movember Committee. Member of the SAB of Our Future Health. Undertakes private practice as a sole trader at The Royal Marsden NHS Foundation Trust and 90 Sloane Street SW1X 9PQ and 280 Kings Road SW3 4NX, London, UK. G.M. is co-founder, director and shareholder of Tailor Bio Ltd. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-72182-5.

Correspondence and requests for materials should be addressed to Joachim Weischenfeldt.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Daria Kiriy[1,2], Francesco Favero[1,2], Clarissa Gerhäuser[3], Jessica Heilmann[3], Pavlo Lutsik[3,4], Francisco German Rodriguez Gonzalez[1,2], Alessio Locallo[1,2], Jakob Schmidt Jespersen[1,2], Andreas J. Gruber[5], Andre V. Olsen[1,2], Barbara Hernando[6], Kevin CL Cheng[7,8], Diogo Pellegrina[7], Geoff Macintyre[6], G. Steven Bova[9], Daniel S. Brewer[10,69], Robert G. Bristow[11,12,13], Mark N. Brook[14], Benedikt Brors[15,16], Adam Butler[17], Géraldine Cancel-Tassin[18,70], Niall M. Corcoran[19,20,21,22,23,24,25], Olivier Cussenot[18], Ros A. Eeles[14,26], Abraham Gihawi[10,27], Etsehiwot G. Girma[1,2], Vincent J. Gnanapragasam[28,29], Anis Hamid[19,30], Vanessa M. Hayes[10,27,31,32], Housheng Hansen He[7,33], Christopher M. Hovens[19,20,22,23,25,34], Eddie L. Imada[35], G. Maria Jakobsdottir[12,13], Chol-Hee Jung[36], Francesca Khani[35,37], Zsofia Kote-Jarai[14,26], Philippe P. Lamy[38,39],

Gregory Leeman[17], Massimo Loda[35,40,41], Luigi Marchionni[35], Ramyar Molania[23], Anthony T. Papenfuss[23], Bernard Pope[19,25,34,36,42,43], Lucio R. Queiroz[35], Tobias Rausch[44], Brian Robinson[35,37], Atef Sahli[31], Karina D. Sørensen[38,39], Sebastian Uhrig[15], David C. Wedge[12,31], Yaobo Xu[17], Takafumi N. Yamaguchi[45], Claudio Zanettini[35], Pan Prostate Cancer Group (PPCG)*, Colin S. Cooper[10], Thorsten Schlomm[46], Jüri Reimand[7,8,47] & Joachim Weischenfeldt[1,2,46] ✉

[1]The Finsen Laboratory, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark. [2]Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark. [3]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [4]Department of Oncology, KU Leuven, Leuven, Belgium. [5]University of Konstanz, Konstanz, Germany. [6]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [7]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. [8]Computational Biology Program, Ontario Institute for Cancer Research (OICR), Toronto, ON, Canada. [9]Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland; Tays Cancer Center, Tampere University Hospital, Tampere, Finland. [10]Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK. [11]Manchester Cancer Research Centre and Cancer Research UK Manchester Institute, The University of Manchester, 555 Wilmslow Road, Manchester, UK. [12]Christie Hospital, The Christie NHS Foundation Trust, Manchester, UK. [13]Division of Cancer Sciences, The University of Manchester, The University of Manchester, 555 Wilmslow Road, Manchester, UK. [14]The Institute of Cancer Research, 123 Old Brompton Road, London, UK. [15]Division Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [16]Medical Faculty Heidelberg and Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [17]Wellcome Sanger Institute, Hinxton, Cambridge, UK. [18]CeRePP (Centre de Recherche sur les Pathologies Prostatiques et Urologiques), 4 rue de la Chine, Paris, France. [19]Department of Surgery, The University of Melbourne, Parkville, VIC, Australia. [20]Division of Urology, Royal Melbourne Hospital, Parkville, VIC, Australia. [21]Department of Urology, Peninsula Health, Frankston, VIC, Australia. [22]Collaborative Centre for Genomic Cancer Medicine, The Victorian Comprehensive Cancer Centre, Parkville, VIC, Australia. [23]Walter and Eliza Hall Institute of Medical Research, 1G Royal Pde, Parkville, VIC, Australia. [24]Department of Urology, Western Health, Gordon Street, Parkville, VIC, Australia. [25]Department of Surgery, The Collaborative Centre for Genomic Cancer Medicine, The University of Melbourne, Grattan Street, Parkville, VIC, Australia. [26]Royal Marsden NHS Foundation Trust, London, UK. [27]Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia. [28]Division of Urology, Department of Surgery, University of Cambridge, Cambridge, UK. [29]Cambridge Urology Translational Research and Clinical Trials Office, Addenbrooke's Hospital, Cambridge Biomedical Campus, Cambridge, UK. [30]Genitourinary Oncology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [31]Manchester Cancer Research Centre, The University of Manchester, 555 Wilmslow Road, Manchester, UK. [32]School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa. [33]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada. [34]Department of Clinical Pathology, The University of Melbourne, Parkville, VIC, Australia. [35]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York-Presbyterian Hospital, New York, NY, USA. [36]Melbourne Bioinformatics, The University of Melbourne, Parkville, VIC, Australia. [37]Weill Cornell Medicine, 1300 York Avenue, New York, NY, USA. [38]Department of Molecular Medicine, Aarhus University Hospital, Palle Juul-Jensens Boulevard 99, Aarhus, DK, Denmark. [39]Department of Clinical Medicine, Aarhus University, Palle-Juul Jensens Boulevard 99, Aarhus, DK, Denmark. [40]The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [41]The New York Genome Center, New York, New York, NY, USA. [42]Australian BioCommons, The University of Melbourne, North Melbourne, VIC, Australia. [43]Department of Medicine, Central Clinical School, Faculty of Medicine Nursing and Health Sciences, Monash University, Melbourne, VIC, Australia. [44]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, Heidelberg, Germany. [45]Jonsson Comprehensive Cancer Center, University of California - Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA, USA. [46]Department of Urology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, Berlin, Germany. [47]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [69]Present address: The Earlham Institute, Norwich, UK. [70]Present address: Sorbonne Université, GRC n°5 Predictive Onco-Urology, APHP, Tenon Hospital, 4 rue de la Chine, Paris, France. ✉e-mail: joachim.weischenfeldt@bric.ku.dk

Pan Prostate Cancer Group (PPCG)

Daria Kiriy[1,2], Francesco Favero[1,2], Clarissa Gerhäuser[3], Jessica Heilmann[3], Pavlo Lutsik[3,4], Andreas J. Gruber[5], Andre V. Olsen[1,2], Barbara Hernando[6], Kevin CL Cheng[7,8], Diogo Pellegrina[7], Geoff Macintyre[6], Marina Torres[6], Angel Fernández-Sanromán[6], Juan Maria Roldan-Romero[6], Lucy Barton[31], G. Steven Bova[9], Daniel S. Brewer[10,69], Mark N. Brook[14], Benedikt Brors[15,16], Adam Butler[17], Géraldine Cancel-Tassin[18,70], Niall M. Corcoran[19,20,21,23,24,25], Olivier Cussenot[18], Abraham Gihawi[10,27], Etsehiwot G. Girma[1,2], Vincent J. Gnanapragasam[28,29], Anis Hamid[19,25,30], Vanessa M. Hayes[10,27,31,32], Housheng Hansen He[7,33], Christopher M. Hovens[19,20,22,23,25,34], G. Maria Jakobsdottir[12,13], Chol-Hee Jung[36], Francesca Khani[35,37], Zsofia Kote-Jarai[14,26], Philippe P. Lamy[38,39], Gregory Leeman[17], Massimo Loda[35,40,41], Luigi Marchionni[35], Ramyar Molania[23], Anthony T. Papenfuss[23], Bernard Pope[19,25,34,36,42,43], Lucio R. Queiroz[35], Tobias Rausch[44], Brian Robinson[35,37], Atef Sahli[31], Karina D. Sørensen[38,39], Takafumi N. Yamaguchi[45], Sebastian Uhrig[15], Yaobo Xu[17], Claudio Zanettini[35], Noora Al-Muftah[48], Esther Baena[31], Daniel Barrowdale[49], Justin Bedő[23], Alain Bergeron[50], Alejandro Berlin[51], Michael Borre[39], Wi Burns[14], Bethany K. Campbell[22,23,25], Ryan Carelli[35], Melissa Cheung[52], Ken Chow[19,25], Michael J. Clarkson[19], Marek Cmero[23,53], Colin Collins[54], Angelo Corso Faini[35], Tony Costello[19,25], Nening Dennis[14], Harveer Dev[55], Wikum Dinalankara[35], Ruining Dong[23,53], Philip Dundee[19,20,25], Chris Foster[14], Andrew Erickson[48], Emre Esentürk[48], Fabian Falkenbach[56], Giuseppe N. Fanelli[35], Breon Feran[23], Lars Feuerbach[15,57], Yves Fradet[50], Yunfan Fu[23,53], Kira Furlano[46], Peter Georgeson[22,34], Markus Graefen[58], Corinna Grima[19], Valleriia Haberland[10,59], Grace A. Hall[36], Xiaotong He[31], Vivien Holmes[31], Chen Hong[57], Matthew KH Hong[19,25], William J. Hutchison[53], Eddie L. Imada[35], Weerachai Jaratlerdsiri[27,60], Jue Jiang[27], Thomas Keane[49], Michael Kerger[19,53], Giselle Kerry[49], Natalie Kurganovs[19,25],

Radoslaw Lach[52,61], Alastair D. Lamb[62], Beng Lim[19], Yong-Jie Lu[62], Andy G. Lynch[61], Stefano Mangiola[23,53], Charlie E. Massie[55], Patrick McCoy[19,25], Arfa Mehmood[12,31], Nana Mensah[63], Ian Mills[48], Toby Milne-Clark[52,55], Anne Nguyen[19,25], Sam Norden[64], Pier Vitale Nuzzo[35], Matti Nykter[9], Samson Olofinsae[31], Mohamed Omar[35], Hubert Pakula[35], Yidan Pan[65], Daniel J. Park[36], Filippo Pederzoli[35], Jocelyn Penington[23], Justin S. Peters[19,25], Pramit M. Phal[19,25], Parsa Pirhady[31], Christoph Plass[3], Keiran Raine[17], Silvia Rodrigues[35], Ronnie Rodrigues Pereira[31], Alan F. Rubin[23,53], Andrew Ryan[64], Aravind Sankar[49], Balthasar C. Schlotmann[1,2], Ryan Stuchbery[19,25], Holger Sültmann[66], Avigail Taylor[48], Pierre Tennstedt[58], Jennifer Ureta[23,53], Peter Van Loo[65], Lena Voith von Voithenberg[57], Matthew Wakefield[23,53], Anne Warren[52], Christopher Wirth[31], Dan J. Woodcock[48], Jingjing Zhang[10,67], Aleksandra S. Ziuboniewicz[48], Ronald Simon[68], Guido Sauter[68], Ros A. Eeles[14,26], Colin S. Cooper[10], Robert G. Bristow[11,12,13], David C. Wedge[12,31], Thorsten Schlomm[46], Jüri Reimand[7,8,47], Joachim Weischenfeldt[1,2,46] ✉ & Rajbir R. N. Batra[3,55]

[48]University of Oxford, Nuffield Department of Surgical Sciences, Headington, Oxford, UK. [49]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK. [50]CHU de Québec – Université Laval, Research Center, Oncology Axis, 2260 boulevard Henri Bourassa, Montreal, QC, Canada. [51]Princess Margaret Cancer Centre, University Health Network, University of Toronto, Toronto, ON, Canada. [52]Department of Surgery, Urology, University of Cambridge & Cambridge University Hospitals NHS Trust, Cambridge Urology Translational Research and Clinical Trials (Office), Cambridge Biomedical Campus, Addenbrooke's Hospital Site, S Wards Building, Keith Day Road, Cambridge, UK. [53]Department of Medical Biology, The University of Melbourne, Melbourne, VIC, Australia. [54]Vancouver Prostate Centre, 2660 Oak Street, Vancouver, BC, Canada. [55]Early Cancer Institute, Department of Oncology, University of Cambridge, Hills Road, Cambridge, UK. [56]Department of Urology, University Hospital Heidelberg, Im Neuenheimer Feld 672, Heidelberg, Germany. [57]German Cancer Research Center (DKFZ), Heidelberg, Germany. [58]Martini-Klinik, University Medical Centre Hamburg - Eppendorf (UKE), Martinistraße 52, Hamburg, Germany. [59]Department of Computer Science, Brunel University of London, Kingston Lane, Uxbridge UB8 3PH, UK. [60]Computational Genomics Group, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW 2050, Australia. [61]School of Medicine and School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. [62]Barts Cancer Institute, Queen Mary University of London, Chartherhouse Square, London, UK. [63]The Francis Crick Institute, 1 Midland Road, London, UK. [64]TissuPath, 32 Ricketts Road, Mt Waverley, VIC, Australia. [65]The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX, USA. [66]German Cancer Consortium (DKTK), Core Center Heidelberg and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [67]School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [68]Institute of Pathology, University Medical Centre Hamburg - Eppendorf (UKE), Hamburg, Germany. [69]Present address: The Earlham Institute, Norwich, UK. [70]Present address: Sorbonne Université, GRC n°5 Predictive Onco-Urology, APHP, Tenon Hospital, 4 rue de la Chine, Paris, France. joachim.weischenfeldt@bric.ku.dk