

# The molecular heterogeneity and clinical impact of genomic driver alterations in prostate cancer

---

Received: 20 October 2025

---

Accepted: 8 September 2026

---

Published online: 16 October 2026

---

 Check for updates

---

---

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

---

Prostate cancer is a common malignancy with a heterogeneous disease course. Prior studies have identified a paucity of point mutations in protein-coding genes and only a few clinically relevant driver alterations. Here, using the Pan Prostate Cancer Group whole genome sequencing dataset from 1001 men, we perform an integrative analysis of cancer driver genes by quantifying recurrent genomic alterations among small mutations, structural variants, and copy number alterations in 961 primary prostate tumour samples. We identify at least one recurrent somatic or germline alteration in 97% of tumours and identify putative driver genes and mechanisms of dysregulation including frequently altered regulatory elements. A prominent example is the activation of oncogenes by structural variation-mediated hijacking of onco-enhancers of the solute carrier gene family including *SLC30A4*. Other recurrent alterations including *FOXA1*, *MDM4*, *COL5A1*, *TP53*, and *BRCA1* stratify with metastatic disease, with additional drivers enriched in African-ancestry tumours. One quarter of tumours have acquired genome-informed therapeutic targets. Unsupervised analysis of alterations reveals four oncotypes with distinct genetic and clinical characteristics, including two clinically unfavourable oncotypes with shorter metastasis-free survival. Our study uncovers the full spectrum and recurrence patterns of genomic driver alterations in primary prostate cancers and delivers targets for biomarker development, mechanistic studies, and potential therapeutic interventions.

Prostate cancer is the most common malignancy in men that accounts for 10% of male cancer-related deaths and is expected to increase in incidence in the coming decades<sup>1</sup>. Its clinical trajectory is highly variable, with disease courses ranging from localised disease with long indolent phases that often require no therapy, to early-onset cancer that typically necessitates definitive treatment with curative intent<sup>2</sup> and life-threatening metastatic prostate cancer<sup>3</sup>. The management of newly diagnosed prostate cancer is primarily based on established clinical and pathological information along with small gene or mRNA panels. The most pressing unmet clinical needs include reliable prognostic approaches to minimize overtreatment and the identification of therapeutic targets<sup>4</sup>. This is especially challenging due to the

considerable molecular heterogeneity of the disease. Its genomic landscape is dominated by structural variation (SV) and copy-number alteration (CNA) drivers, while in contrast, drivers involving small mutations including single-nucleotide variants (SNVs) and insertions-deletions (indels) are relatively infrequent<sup>5,6</sup>. Hallmark driver alterations include *ETS* oncogene fusions in approximately 50% of prostate cancers, protein-coding mutations in *TP53*, *FOXA1*, and *SPOP*, genomic gains of the *MYC* oncogene and losses of tumour suppressor genes (TSGs) such as *PTEN* and *TP53*<sup>7–10</sup>. However, knowledge of drivers among non-coding mutations or SVs remains limited as most studies have used gene panels, exome-sequencing protocols or smaller WGS datasets that remain underpowered for genome-wide driver discovery.

---

✉ e-mail: [joachim.weischenfeldt@bric.ku.dk](mailto:joachim.weischenfeldt@bric.ku.dk); [juri.reimand@utoronto.ca](mailto:juri.reimand@utoronto.ca)

Recent WGS studies have observed 4–5 known drivers per genome across many cancer types<sup>11,12</sup>, reflecting a limited core set of hallmark pathways whose dysregulation is required for oncogenesis<sup>13</sup>. A substantial fraction of prostate cancers remain poorly explained, with no or only a single driver mutation detected<sup>5,7,10</sup>. Moreover, no study has systematically evaluated how these drivers functionally impact clinical outcomes, including metastasis. Thus there is an urgent need to deconvolve the molecular heterogeneity of prostate cancer and identify its different oncogenic driver routes. This will not only improve molecular understanding of the disease, but will also have direct implications for patient risk stratification. The premise of precision oncology is to accurately diagnose cancer subtypes and assign effective targeted therapies or surveillance protocols to each patient based on the genomic makeup of their tumour. Large WGS datasets coupled with detailed clinical outcome profiles are needed to advance this premise.

Here we identify and examine known and putative coding and non-coding cancer drivers in a large WGS dataset of 961 primary prostate tumour samples analysed with standardised pipelines in the Pan Prostate Cancer Group (PPCG) project<sup>14</sup>. We demonstrate the utility of integrating orthogonal layers of omics and clinical data to uncover driver genes and their involvement in integrated networks and oncogenic pathways, and ultimately, to identify molecular subtypes for risk-stratification of patient outcome.

## Results

### Recurrent genomic alterations in prostate cancer

We analysed 961 matched tumour-normal whole genomes from primary, localized, and predominantly untreated prostate cancers in the Pan Prostate Cancer Group (PPCG) cohort processed using standardized consensus pipelines<sup>14</sup>. The analysed subset comprised fresh-frozen samples from six contributing countries (Canada, Germany, UK, Australia, Denmark, France), with median tumour and normal sequencing depths of 69 and 41, respectively; matched RNA-seq data were available for 494 (51%) tumours, and one representative tumour sample per donor was selected (Supplementary Data 1). Key clinical and demographic features of the cohort, including donor age, PSA at tumour collection, Gleason grade group, pathological T-stage, metastasis during follow-up, and predicted ancestry, are summarised in Supplementary Fig. 1. Most of the cohort (709 or 74%) had the annotations required for metastasis-free survival analysis. The cohort was enriched for European ancestry (90%), with smaller groups for African (3.5%), East Asian (1.7%), Admixed American (0.6%), or unassigned (3.3%). Pathogenic germline risk variants in established prostate cancer predisposition genes were identified in 4.6% of donors, including germline variants in *BRCA1* or *BRCA2* in 2.2% of donors.

A cancer driver is a genomic alteration that provides a selective growth advantage to a cancer cell, and can be identified in large datasets of tumour samples as significantly enriched alterations compared to an expected background mutation rate<sup>15</sup>. To systematically discover such events, we identified statistically significant recurrent genomic alterations (RGAs) (FDR < 0.05) from six types of somatic alterations: CNAs including gains and losses, genes with non-synonymous small protein-coding mutations (SNVs, indels), regions with non-coding small mutations (SNVs or indels) or SV breakpoints, genes with biallelic inactivation (BI) caused by genomic alterations affecting both alleles, and genes with focal amplifications. RGAs were evaluated at the level of genes or prostate-relevant non-coding elements, depending on alteration class. Protein-coding SNVs and indels were assigned to coding exons, whereas non-coding SNVs and indels in regulatory elements were linked to nearby genes within  $\pm 3$  kb. SVs were evaluated analogously at the level of recurrent breakpoints. Recurrent CNAs were identified across broader genomic regions and assigned to a single index gene per recurrent region, based on statistical evidence and manual curation, to reduce redundancy from large

segments affecting multiple neighbouring genes. To facilitate high-confidence driver discovery, we only included SNV and indel calls identified by two or more PPCG variant calling pipelines<sup>14</sup>. SNV and indel, SV and CNA analyses were conducted using distinct robust algorithms to detect positive selection<sup>16–19</sup> (Methods). In addition to protein-coding genes, we studied 16,041 non-coding RNA genes, untranslated regions (UTRs), and 650,873 gene-regulatory and chromatin architectural elements derived from prostate tumour samples<sup>20–23</sup>, enabling systematic discovery of non-coding driver candidates. Altogether, the analysis included 2,372,041 SNVs or indels (median 2093 per tumour sample, including 17 protein-coding variants), 148,902 SVs (median 106) and 160,032 CNAs (median 114).

We first examined the major landmarks of the driver landscape in prostate cancer. In agreement with prior CNA studies<sup>5,7,10</sup>, PPCG genomes were dominated by 44 recurrent CNAs affecting established prostate cancer driver genes, such as 8q24 gains focused on *MYC* (16%), and losses of *NKX3-1* in 8p21 (58%), *PTEN* in 10q23 (35%), *TP53* in 17p13 (33%), and *RBI* in 13q14 (40%) (FDR < 0.05, one-sided binomial test, Fig. 1A, Supplementary Fig. 2A, Supplementary Data 2). Recurrent gains (3–7%) also affected genes known from other cancer types, such as *MDM4* in 1q32, a negative regulator of p53<sup>24</sup>, and *GATA2* in 3q21, a pioneer factor for androgen receptor (AR) binding that drives prostate cancer aggressiveness<sup>25,26</sup> as well as *TERT* in 5p15 and *CCND1* in 11q13 (Supplementary Fig. 2B). AR amplifications on Xq12 occurred in 1.1% of primary tumour samples (including one of 14 donors with prior anti-androgen therapy in PPCG), similar to an earlier study<sup>27</sup>. AR amplification following anti-androgen therapy is a hallmark of metastatic prostate cancer<sup>28,29</sup>, but is rare in primary untreated patients. SV analyses revealed 29 RGAs with recurrent SV breakpoints including rearrangements of ETS oncogenes in 51% of tumours and *PTEN*-disrupting SVs (16%). ETS alterations are found in 40–50% of prostate cancers<sup>10,30</sup>, with an elevated frequency in early-onset disease<sup>31</sup>. In agreement with the literature, we found only 20 genes with enriched protein-coding SNVs or indels (FDR < 0.05, dNdScv<sup>19</sup>), and 69 RGAs in genes or intergenic sites with non-coding SNVs and indels (FDR < 0.05, ActiveDriverWGS<sup>16</sup>). Also, 43 TSGs and 290 genes in total were altered by biallelic inactivation (BI) involving germline and somatic CNAs, SVs, SNVs or indels (permutation  $P$  < 0.05, Curveball<sup>18</sup>).

Besides CNA gains and losses, we characterised 53 RGAs that were identified in at least 1% of the cohort by gene- or locus-level integration of five alteration types (protein-coding SNVs or indels, SVs, amplifications, BI events, non-coding SNVs or indels) (Fig. 1B, Supplementary Data 3). Besides 57% (30/53) of the RGAs identified in prior prostate cancer genomics analyses<sup>7,8</sup> and 21% (11/53) known as drivers in other cancer types according to COSMIC Cancer Gene Census<sup>32</sup>, we identified 23% of genes or sites (12/53) as previously unreported driver candidates. 20 out of 53 RGAs were not ascertainable by standard panel-based methods or exome-sequencing and would therefore remain undetected, supporting the importance of WGS-based analyses. Several frequently altered genes (e.g., *PTEN*, *FOXA1*, *TP53*, *CHD1*) exhibited multiple alteration types, underscoring the value of our multi-layered approach in identifying key molecular drivers and advancing translational insights in prostate cancer.

### Diversity of genomic alterations in prostate cancer

Integrating different types of genomic alterations substantially increased our sensitivity to identify RGAs. Nearly all tumours (97%) had at least one RGA, with a mean of 8.4 RGAs per tumour (Fig. 2A, Supplementary Fig. 3, Supplementary Fig. 4). RGAs were dominated by CNAs (91%) and SVs (69%). In contrast, protein-coding driver mutations (SNVs, indels) occurred in only one third (36%) of the tumours. Biallelic TSG inactivations were found in 37% of tumours, oncogene amplifications in 4.8% of tumours, and 44% of tumours had at least one RGA involving non-coding SNVs and indels. When excluding CNAs, 88% of primary cancers in PPCG had at least one RGA with a mean of 2.7

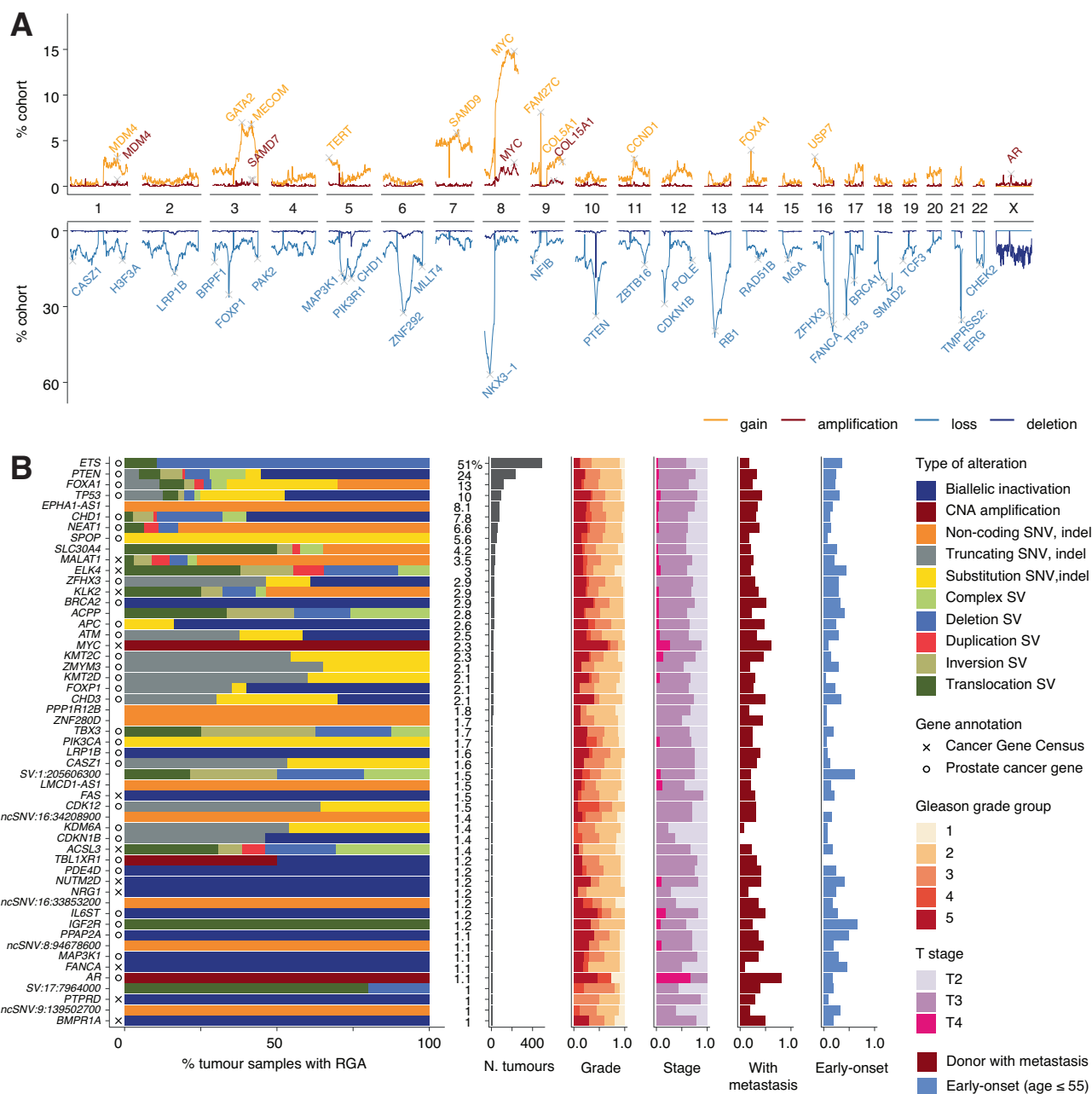


**Fig. 1 | Recurrent genomic alterations in prostate cancer. A** Frequencies of copy number alterations (CNAs) in the PPCG cohort of 961 primary tumour samples. Gains are shown above (3-4 copies in yellow, >4 copies in red; ploidy-adjusted) and losses are shown below (1 copy in pale blue, 0 copies in dark blue). Significantly recurrent alterations with most likely target driver genes are labelled (FDR < 0.05, one-sided binomial test). **B** Recurrent genomic alterations (RGAs) with at least 1% mutation frequency across 961 tumour samples. Colours indicate different types of

genomic alterations, as shown on the far right. To the right, the corresponding mutation frequency (y-axis) and numbers of tumour samples with RGAs (x-axis) followed by distributions of clinical annotations of RGAs including tumour grade (Gleason grade group), T-stage, donor with metastasis developed during follow-up, and donor age at diagnosis (i.e., early-onset prostate cancer). Source data for panels A-B are provided as Source Data files.

RGAs per tumour (median 2, 95% range 0–7 RGAs). Notably, no single variant type was able to identify all 494/961 (51.4%) tumours with fusions involving ETS transcription factors (*ERG* fusions in 47%, *ETV1* or *ETV4* in 2-3% tumours and single events in *FLI1* or *ETV5*) (Fig. 2B, Supplementary Data 4), underlining the value of multi-omics analyses.

We identified 20 genes enriched in protein-coding SNVs or indels using the dNdScv method<sup>19</sup> (FDR < 0.05), including well-defined prostate cancer genes (*TP53*, *SPOP*, *FOXA1*, *PTEN*) and less-characterised ones, such as *ZMYM3* coding for a scaffolding protein regulating BRCA1 recruitment to damaged chromatin<sup>33</sup>, or *CHD3*, a chromatin remodeling helicase (Fig. 2C, Supplementary Data 5, Supplementary Data 6).

Protein-coding driver genes were overall enriched in highly-disruptive alterations, including nonsense SNVs, frameshift indels, and essential splice site alterations ( $P = 1.0 \times 10^{-17}$ , odds ratio = 2.9, two-sided hypergeometric test), indicating that loss-of-function is a major consequence of small mutations in TSGs in prostate cancer. Mutational signature analysis suggested that indels were overrepresented among protein-coding driver mutations ( $P = 1.3 \times 10^{-46}$ , odds ratio = 5.5) and individual genes were also enriched, especially in the context of the indel signature ID5 of unknown etiology (Supplementary Fig. 5, Supplementary Data 7). All protein-coding driver genes found in this analysis have been identified in earlier surveys of prostate cancer

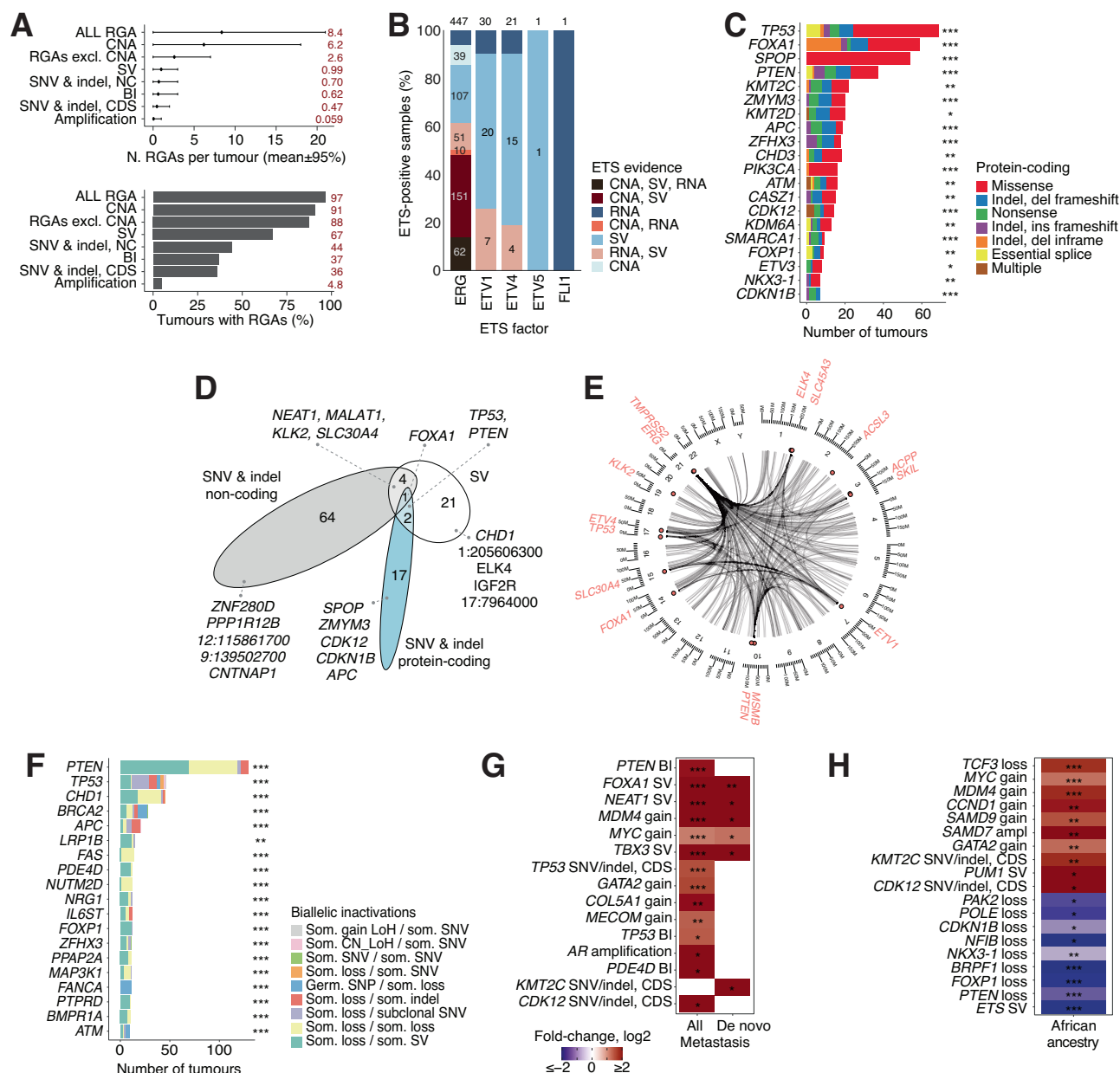


**Fig. 2 | Diversity of genomic alterations in prostate cancer.** **A** Overall frequency of RGAs in PPCG. Top: number of RGAs per prostate cancer genome. Points show mean values and error bars show 95% percentile ranges across 961 tumour samples. Bottom: Percentage of tumours with at least one of the RGAs presented in the top panel (CDS - protein-coding; NC - non-coding). **B** Mutation status of ETS family transcription factors in PPCG grouped by multi-omics evidence from CNAs, SVs and matching transcriptomics data. **C** Driver genes with protein-coding mutations across 961 tumour samples (P-values from the dNdScv method, adjusted for multiple testing, FDR < 0.05). **D** Euler diagram of 109 recurrent genomic alterations (RGAs) enriched in SVs, or coding or non-coding SNVs and indels across 961 tumour samples (P-values from the ActiveDriverWGS method, adjusted for multiple testing, FDR < 0.05). Top ten RGAs are listed for each subset. **E** Genes with recurrent translocations affecting at least 1% of tumour samples. The circo plot shows SV

hotspots (red circles) with adjacent driver genes labelled. **F** Established TSGs with significantly enriched BI events affecting at least 1% of tumour samples from integrative analysis of somatic (som) and germline (germ) alterations across 961 tumour samples (permutation P-values from the Curveball method; P < 0.05). LoH: loss of heterozygosity, SNP: single-nucleotide polymorphism. **G** Associations of RGAs in 103 metastatic prostate tumour samples (including 46 de novo metastatic tumours) compared with the 961 primary prostate tumours in PPCG (two-sided binomial tests, FDR < 0.05). **H** Associations of RGAs with 113 prostate tumour samples from patients of African ancestry, compared with 927 PPCG tumour samples from patients of non-African ancestry (two-sided binomial tests, FDR < 0.05). Asterisks indicate statistical significance (\* < 0.05, \*\* < 0.01, \*\*\* < 0.001). Source data for panels A–H are provided as Source Data files.

genomes<sup>7,8</sup>, suggesting a saturation of exome-based driver discovery due to the bias towards protein-coding regions.

In contrast, SV breakpoints and non-coding SNVs and indels are comparatively less well described in prostate cancer. Our driver analyses using ActiveDriverWGS<sup>6</sup> revealed 92 genes or other genomic elements enriched in non-coding mutations (SNVs, indels) or SVs (FDR < 0.05; Supplementary Data 8, Supplementary Data 9).

Examining genes with multiple types of alterations, we found five that were independently enriched in SVs and small mutations: three protein-coding genes *FOXA1*, *KLK2*, *SLC30A4* and two non-coding RNA genes *NEAT1* and *MALAT1* (Fig. 2D). The promoter of *KLK2*, an androgen-induced serine protease activating PSA to its enzymatically active form<sup>34</sup>, was significantly enriched in SNVs or indels in 15 tumours (FDR = 0.013), whereas its 3'UTR was enriched in SVs in 15 tumours

(FDR =  $7.4 \times 10^{-5}$ ) (Supplementary Fig. 6). Non-coding SVs at the *KLK2* locus showed a trend towards decreased gene expression ( $P = 0.033$ , FDR = 0.065), suggesting a possible regulatory effect. Reduced expression of *KLK2* has been associated with poor prognosis in ERG-negative prostate cancer<sup>35</sup>, however no robust associations were found in PPCG. Next, we examined mutational hotspots of SNVs and indels. The most significant SNV hotspot occurred in an intronic enhancer in a poorly studied zinc finger gene *ZNF280D* (16 tumours, FDR =  $2.8 \times 10^{-18}$ ; Supplementary Fig. 7A). The C>T SNV creates a de novo DNA binding site for the nuclear receptor transcription factors NR1I2 and NR1I3 (Supplementary Fig. 7B–D). NR1I2 is a regulator of drug metabolism associated with chemotherapy resistance in prostate cancer and is a putative target for androgen inhibition<sup>36,37</sup>. We found no evidence that the hotspot alters *ZNF280D* expression or correlates with NR1I2 or NR1I3 expression, suggesting that it may regulate a different target gene; however, the associated transcriptomic signature was enriched for energy metabolism pathways (Supplementary Fig. 7E). Additional mutational hotspots among RGAs occurred in protein-coding regions of *SPOP* and *FOXAI*, and a non-coding locus at chr9:139502700 (Supplementary Fig. 8). We also identified the long non-coding RNA (lncRNA) gene *EPHA1-AS1* with frequent non-coding SNVs and indels in 8.1% of tumours (FDR = 0.021) (Supplementary Fig. 9). Finally, RGAs in *FOXAI* were independently found in three driver analyses, including enrichment of protein-coding SNVs and indels, enrichment of non-coding SNVs and indels in 3'UTR, and via SVs enriched in a regulatory region downstream of *FOXAI* that harbor frequent chromatin long-range interactions (Supplementary Figs. 10A and 10B), reflecting three modes of *FOXAI* alteration in prostate cancer described previously<sup>38</sup>. Overall, RGAs with frequent SVs or non-coding SNVs and indels were enriched in prostate-specific enhancers or open-chromatin elements, suggesting gene regulatory roles (Supplementary Fig. 11). Among different types of SVs, translocations were enriched among RGAs compared to their genome-wide distribution ( $P = 1.8 \times 10^{-60}$ , OR = 2.4) whereas several RGAs such as *FOXAI*, *SLC30A4*, *PTEN*, *KLK2*, *TP53* and *IGF2R* were individually enriched in translocations (Fig. 2E, Supplementary Fig. 5). As further validation, we found 52 associations in which target genes of RGAs were differentially expressed in the presence of RGAs, based on matching transcriptomic data for PPCG tumours<sup>14</sup> (FDR < 0.05, Wilcoxon test) (Supplementary Fig. 12, Supplementary Data 10, Supplementary Data 11). As expected, genomic loss or biallelic inactivation was associated with decreased expression of TSGs including *PTEN*, *TP53* and *CHD1* while upregulation of ETS factors, including *ERG*, *ETV1* and *ETV4* was associated with SV-mediated fusions or CNAs.

Next, we examined biallelic inactivation of TSGs, motivated by frequent CNAs in prostate cancer that lead to chromosomal losses such as 8p, 5q, 10q and 16q. Genes with BI involved complete loss of function caused by deletion or disruption of both alleles. We integrated somatic SNVs and indels, SVs, CNAs and germline pathogenic variants across the genome using a statistical approach<sup>2,12,18</sup> ( $P < 0.05$ , Curveball permutation test). High-confidence BI genes included 19 known TSGs affected in at least 1% of the cohort while an extended list of 290 significant BI genes included 43 known cancer genes (Fig. 2F, Supplementary Fig. 13, Supplementary Data 12, Supplementary Data 13). Most arm-level recurrent CNA losses contained at least one TSG with significant BI (Fig. 1A). The most frequent BI events involved *PTEN* (13% of tumours) and *CHD1*, *TP53*, *BRCA2* and *APC* each found in 2–5% of tumours. Most BI events were highly gene-specific and involved CN loss of one allele while the other allele involved a germline or somatic SNV, indel or disruptive SV. In total, 10% of tumours with BI events of TSGs (29/295) had a pathogenic germline variant and a somatic second hit, including *BRCA2* in 1.1% of the cohort as well as *ATM* (0.5%), *TP53* (0.4%) and *RBI* (0.3%), elucidating the interplay of somatic alterations and genetic predisposition of prostate cancer.

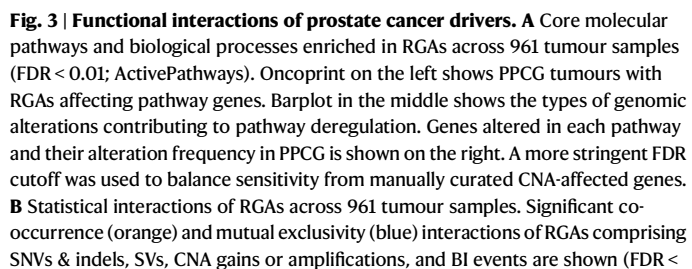
We compared the frequencies of RGAs between the primary cohort and metastatic tumours of PPCG, analysing 103 metastatic cases and 46 de novo metastatic cases. We found that metastatic tumours were enriched for several alterations associated with aggressive disease (FDR < 0.05, two-sided binomial test). Most notably, these included biallelic inactivations of *PTEN* seen in 50% of all metastases (FDR =  $3.5 \times 10^{-16}$ , 3.7-fold), *AR* amplifications seen in 5.8% of metastases (FDR = 0.016, 5.1-fold), as well as CNA gains of *MYC*, *MDM4*, and *COL5A1* (FDR = 0.0036, 11% of metastases), *TP53* alterations, and SVs in *FOXAI* and the lncRNA *NEAT1* (Fig. 2G, Supplementary Data 14). Several of these alterations were also enriched in de novo metastatic tumours, supporting a consistent association between these RGAs and metastatic disease.

To assess ancestry-linked differences in the RGA landscape, we compared 113 prostate cancers from donors of African ancestry<sup>39</sup> with 927 PPCG tumours of non-African ancestry, focusing on protein-coding SNVs and indels, SVs, and CNAs. We found that African-ancestry tumours were depleted for ETS alterations, with 13% frequency in samples of African ancestry and 56% in non-African ancestry (FDR =  $2.1 \times 10^{-16}$ ) and CNA losses including *PTEN* and *FOXPI*, while enriched for several alterations associated with more aggressive disease, including CNA gains of *MYC* seen in 40% of samples of African ancestry and 17% of non-African ancestry (FDR =  $6.7 \times 10^{-7}$ ), gains of *MDM4* and *GATA2*, as well as protein-coding mutations in *CDK12* and *KMT2C* (two-sided binomial test, FDR < 0.05; Fig. 2H, Supplementary Data 15). Overall, these results suggest that the ancestry-linked driver spectrum is shifted towards a more aggressive alteration profile in African-ancestry tumours.

### Functional interactions of prostate cancer drivers

Pathway and network analyses of cancer drivers complement gene-focused analyses by considering functionally related subsets of genes and non-coding elements<sup>40</sup>. To evaluate the functional importance of RGAs in prostate cancer, we integrated the main target genes of significant RGAs (protein-coding and non-coding SNVs and indels, SVs, CNAs, BI and amplification events) using pathway enrichment analysis. We found 268 significantly altered pathways involving 53 target genes of RGAs (FDR < 0.01, ActivePathways<sup>41</sup>; Fig. 3A, Supplementary Fig. 14, Supplementary Fig. 15, Supplementary Data 16). The pathways clustered into 12 major functional themes frequently affected by RGAs in PPCG that included hypoxia, prostate gland development, cell cycle, apoptosis, immune response, microRNA regulation, and homologous recombination repair. Of those, genes involved in cellular response to hypoxia showed genomic alterations in 76% of the PPCG cohort, in line with the prevalence and importance of hypoxia in prostate cancer characterised in our companion study<sup>42</sup>. Recurrent RGAs converged to multiple central pathways such as PI3K/AKT (60% of tumours), P53 (64%), AR (65%), and TGF-beta (33%). Frequent alterations in homologous recombination pathways (e.g. *BRCA1*, *BRCA2*, *ATM*, *POLE*, *RAD51B*) and immune response processes (e.g. *IL6ST*, *JAK1*) may define subsets of patients for further immuno-oncology investigation in the predominantly immune-cold and genomically stable landscape of prostate cancer.

We examined RGAs for patterns of co-occurrence or mutual exclusivity in the PPCG cohort and found 90 significant pairs of RGAs (two-sided hypergeometric test, FDR < 0.05) (Fig. 3B, Supplementary Data 17). *PTEN* and ETS factors were frequently co-altered (181 tumours vs 120 expected, FDR =  $4.6 \times 10^{-18}$ ) while both were mutually exclusive with oncogenic gains such as *MYC* and RGAs in *SPOP*, *CHD1*, and *FOXAI*, as expected<sup>15,43</sup>, collectively representing 80% of tumours in PPCG and supporting ETS+ and ETS- genomic subtypes of primary prostate cancer (Supplementary Fig. 16A). We examined the subset of 376 tumour samples lacking alterations in *SPOP* and ETS factors and found significant associations with 19 RGAs (FDR < 0.05), including enriched alterations in *MYC* and *FOXAI* (Supplementary Fig. 17, Supplementary



Non-coding SNVs and indels in gene-regulatory elements may contribute to gene regulation in cancer through long-range chromatin interactions<sup>16,21,44</sup>. To study this in detail, we integrated long-range chromatin interactions from prostate cancer<sup>21</sup> and recovered 41 distal interactions of RGAs and target genes (Fig. 3C, Supplementary Fig. 18, Supplementary Data 8). As the most frequent RGA, we identified an SV hotspot at a regulatory region downstream of *FOXAI* harboring long-range chromatin interactions to a distal gene,

*MIPOL1* (Supplementary Fig. 10B, C). Tumours with SVs showed increased expression of *MIPOL1*, suggesting that the SVs impact distal transcriptional regulation of *MIPOL1* (Supplementary Fig. 10D). Secondly, we found a frequently-altered subnetwork of long-range chromatin interactions involving the lncRNAs *NEAT1* and *MALAT1* that together have SNVs/indels or SVs in a large fraction (9.3%) of tumours in PPCG (Supplementary Fig. 19A-D). These two genes show strong expression correlation, suggesting a co-regulatory mechanism of transcriptional control (Supplementary Fig. 19F). Thus, putative driver mechanisms or localised mutational processes may be mediated by long-range chromatin interactions in prostate cancer.

To identify targetable RGAs in our cohort, we prioritised alterations with the highest clinical relevance from two established databases, OncoKB<sup>45</sup> and CiVic<sup>46</sup>. We focused on RGAs comprising BI of TSGs and curated driver variants of oncogenes and restricted our analysis to Level 1 evidence comprising biomarkers associated with response to FDA-approved therapies (Supplementary Data 19). We identified 221 tumour samples (23%) with a targetable alteration (Fig. 3D), including PI3K pathway targets *PTEN* and *PIK3CA* that are together highly mutated in PPCG representing 13% and 1.7% of primary tumours, respectively, but not yet approved for prostate cancer. Alterations in DNA repair pathways such as *BRCA2* (2.9% of PPCG primary tumours) are currently recommended for PARPi therapies to patients with metastatic castration-resistant prostate cancer and may also apply to patients with high-risk localised disease<sup>47</sup>. In addition, 12 tumours exhibited high tumour mutational burden (TMB,  $\geq 10$  non-synonymous SNVs per Mbp), a biomarker of immunotherapy response<sup>48</sup>. Of those, seven had BI in a known TSG, including four with *MSH2* BI (Supplementary Data 20). We also explored the full spectrum of potential drug targets across all evidence levels and cancer types, assigning most donors (69%) to a potential targeted therapy (Supplementary Fig. 20).

### Structural variants hijack enhancers of solute carrier genes

Our RGA analysis identified a SV breakpoint hotspot at 15q21.1 in *SLC30A4* altered in 26 tumour samples in PPCG (FDR =  $1.9 \times 10^{-9}$ , ActiveDriverWGS). *SLC30A4* encodes solute carrier family 30 member 4, which belongs to the solute carrier (SLC) family of membrane transport proteins with critical roles in transporting nutrients, ions and drugs across cell membranes. *SLC30A4* is involved in zinc ion transport in prostate cells and has been implicated in prostate cancer<sup>49,50</sup>, however, it is not established as a cancer driver. Breakpoints in *SLC30A4* were associated with a high exclusivity score<sup>2</sup>, indicating these as likely early events in tumorigenesis (Supplementary Fig. 21). We found several lines of evidence suggesting that the main function of the SVs was not to alter *SLC30A4* per se. First, SV breakpoints affected both the gene body and the upstream enhancers<sup>23</sup>, marked by histone marks (H3K27ac, H3K4me3)<sup>51</sup> and binding of bromodomain and extra-terminal (BET) protein BRD4<sup>52</sup> (Fig. 4A). Second, most SVs were translocations between the *SLC30A4* locus and distant oncogenes including *ERG*, *ETV4*, *ETV1*, and *MYC* (Fig. 4A). SVs in *SLC30A4* were mutually exclusive with the consensus set of ETS alterations in PPCG (FDR = 0.027, hypergeometric test) as well as the most common fusion partners *TMPRSS2* and *ERG* (Fig. 3B, Supplementary Fig. 16C). We also noted a previous observation of a translocation juxtaposing *SLC30A4* to *ETV1* in metastatic prostate cancer<sup>53</sup>. This suggests a mechanism compatible with enhancer hijacking<sup>54,55</sup>. We found a similar pattern affecting *SLC45A3*, which is an established fusion partner of ETS oncogenes in prostate cancer<sup>56</sup>.

We pursued a genome-wide analysis to identify additional SVs between enhancers of SLC genes and distant oncogenes and found six additional SLC genes, including *SLC4A4* in three tumours and *SLC26A7*, *SLC12A1*, *SLC22A16*, *SLC2A12* and *SLC14A2* each in a single tumour, totalling 2.8% of prostate tumours (Fig. 4B–C; Supplementary Data 21). Hijacked enhancers are often highly active, resulting in high target gene expression<sup>11,54</sup>. We therefore hypothesised that SLC gene expression was associated with enhancer rearrangements. To this end, we leveraged transcriptomes from 68 normal tissues including prostate from the GTEx consortium<sup>57</sup>, which revealed a striking correspondence between the number of oncogenic SVs and the expression of respective SLC genes in normal prostate, with *SLC45A3* and *SLC30A4* among the most highly expressed and most frequently hijacked SLC genes (Fig. 4D). This pattern was also observed in prostate cancer transcriptomes from PPCG (Supplementary Fig. 21A). Further, 84% of SLC genes were upregulated in prostate tissues compared to non-prostate tissues (219 / 262

detectable genes,  $P = 2.04 \times 10^{-8}$ , binomial test). SLC genes were enriched for DNA-binding sites of AR, especially in recurrently rearranged SLC genes, suggesting AR-mediated regulation (Fig. 4B, Supplementary Fig. 21B, C, Supplementary Data 21). We also found SNVs near AR binding sites and SV breakpoints of *SLC30A4* (Fig. 1B), suggesting additional AR-mediated mutagenesis. Finally, we applied the CESAM algorithm<sup>11,55</sup> to ask if the juxtaposition of SLC enhancers was associated with oncogenic dysregulation. Indeed, SLC rearrangements were accompanied by significant upregulation of oncogenes *ERG*, *ETV1*, *ETV4*, *FLI1* and *SKIL* (Fig. 4E). We also identified *SLC30A4-MYC* hijacking, although the change in expression of *MYC* did not reach statistical significance. While *SLC45A3-ERG* SVs are known to result in fusion transcripts, most rearrangements in *SLC30A4* and other SLC genes resulted in enhancer-oncogene juxtaposition (Fig. 4F). Together, these findings suggest an AR-mediated mechanism by which SVs hijack tissue-specific onco-enhancers of SLC genes early in tumorigenesis to drive cancer gene upregulation in prostate cancer.

### Clinical and functional characteristics of RGA-based prostate cancer oncotypes

Finally, we investigated the associations between RGAs and clinical outcomes in prostate cancer. A multivariable Cox proportional-hazards (CoxPH) model predicting metastasis-free survival (MFS) using RGAs revealed six RGAs as prognostic factors (protein-coding mutations in *TP53* and *CHD3*, CNA gains in the oncogenes *COL5A1* and *MDM4*, *FOXAI* SVs, and *BRCA1* loss) when considering donor age at diagnosis, tumour T-stage and Gleason grade group as covariates (Wald test,  $P < 0.05$ ; Fig. 5A, Supplementary Data 22). When testing each RGA separately in clinically-adjusted CoxPH models with multiple-testing correction, we found seven associations with MFS (FDR < 0.05) and additional nominally significant RGAs (Supplementary Fig. 22, Supplementary Data 23). Significant findings included protein-coding mutations in *TP53* (FDR = 0.0027, HR = 3.3), CNA-driven gains of *MDM4* (FDR = 0.0041; HR = 3.8) and *COL5A1* (FDR = 0.0041; HR = 4.7), and genomic losses of TSGs *BRCA1*, *TP53* and *ZBTB16* (Supplementary Fig. 23). In *FOXAI*, the SV-enriched RGA at the chromatin interaction locus downstream of *FOXAI* was significantly associated with MFS (FDR = 0.0027, HR = 5.8) while coding and non-coding SNVs and indels in *FOXAI* were not associated (Supplementary Fig. 10E), suggesting that the clinical significance of different types of *FOXAI* alterations varies in prostate cancer. *COL5A1* gains were associated with differential expression of 112 genes in matching transcriptomes (FDR < 0.05) and enrichments of cell proliferation and DNA replication processes, suggesting a functional role of this RGA in aggressive prostate cancer (Supplementary Fig. 24). Next, we contextualised RGAs using donor age and identified SVs affecting *ETS*<sup>2,31</sup> and *IGF2R*, a gene involved in insulin signalling, to be enriched in early-onset prostate cancer (age  $\leq 55$  years) whereas several RGAs such as *SPOP* and *CHDI1*, both of which are mutually exclusive with ETS alterations, were enriched in late-onset cancer (Supplementary Fig. 25, Supplementary Data 24). We noticed that several RGAs associated with higher risk of metastasis were enriched in the subset of tumour samples with at least one missing clinical variable required for MFS analyses (252 or 26%). These included AR amplifications, gains of *MYC* and *MDM4*, and protein-coding mutations of *CDK12*, while ETS alterations were depleted (Supplementary Fig. 26), suggesting underestimates of our risk associations due to missing data. Finally, we found that higher overall RGA burden per tumour was also associated with shorter MFS when accounting for clinical covariates, with the strongest associations observed for CNA-based and SNV/indel-based RGAs (Supplementary Fig. 27), indicating the genomic complexity of high-risk tumours.

To ask whether RGA profiles distinguished tumour subsets with different metastatic risk, we performed hierarchical consensus

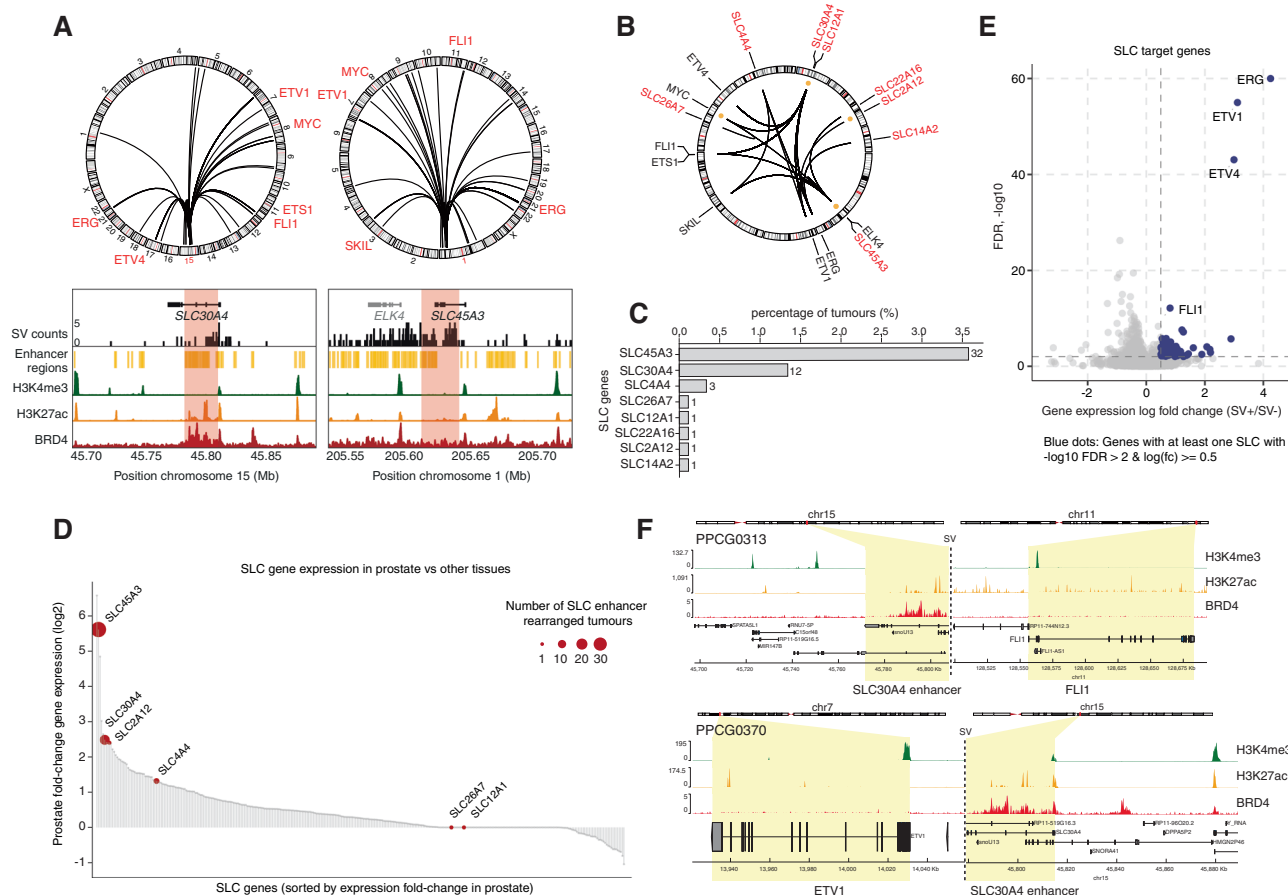


**Fig. 4 | Structural variants hijack enhancers of solute carrier genes.** **A** SVs in two SLC gene loci *SLC30A4* (left) and *SLC45A3* (right) across 961 tumour samples. From top to bottom: Circos plot of SVs juxtaposing SLC regulatory regions to distant genomic loci (oncogenes highlighted in red), gene annotation, binned breakpoint count (25 kbp bins), enhancer regions as defined previously in prostate cancer samples by chromHMM, and ChIP-seq peaks of H3K4me3, H3K27ac, and BRD4 in prostate cancer cells<sup>23,51,52</sup>. **B** Circos plot of SVs (arches) juxtaposing SLC regulatory regions (red) next to oncogenes (black). AR-regulated genes are highlighted with orange dots. **C** Percentage of PPCG tumours with SVs juxtaposing SLC regulatory regions next to an oncogene. **D** Gene expression level of SLC genes (x-axis) in normal prostate tissue from GTEx compared to other tissues. Log2 fold-change difference between median of 282 prostate tissue samples and median of 19,334 from 67 different non-prostate tissue samples from GTEx is shown. Size of red dots

corresponds to the number of PPCG tumours with SVs juxtaposing SLC regulatory regions next to an oncogene. **E** Analysis of SV-driven differential gene expression in enhancer hijacking events across 494 tumour samples with matching transcriptomic data ( $FDR < 0.05$ , CESAM). The volcano plot shows effect size on the x-axis ( $\log_2$  fold change comparing tumours with gene-proximal SVs versus control tumours without SVs) and significance of association between SVs and gene expression change on the y-axis ( $-\log_{10} FDR$ ). **F** Two examples of enhancer hijacking involving translocations between SLC enhancers and oncogenes: *SLC30A4* enhancer and *FLI1* (top) and *SLC30A4* enhancer and *ETV1* (bottom). Enhancer and oncogene regions marked in yellow. A vertical stippled line marks the SV breakpoint. Chromatin marks from top to bottom: H3K4me3, H3K27ac and BRD4 (see also A). Source data for (A–E) are provided as Source Data files.

clustering<sup>58</sup> of prostate tumour samples based on RGAs that we visualised as heatmaps and dendrograms (Fig. 5B, Supplementary Fig. 28, Supplementary Data 25). We identified four distinct subsets of tumours, which we refer to as oncotypes. Oncotypes showed a trending association with MFS when considering donor age at diagnosis, tumour T-stage and Gleason grade group as covariates ( $P = 0.063$ , CoxPH likelihood ratio test) and a significant separation based on MFS alone ( $P = 1.7 \times 10^{-14}$ , log-rank test). Similarly, the oncotypes showed significant differences in biochemical recurrence (BCR) (likelihood ratio  $P = 0.014$ ; log-rank  $P = 2.4 \times 10^{-17}$ ) (Fig. 5C–D). To validate the oncotypes, we first mapped CNA, SNV/indel, and SV-based RGAs to an independent cohort of prostate cancer samples from the Cancer Genome Atlas (TCGA-PRAD) cohort<sup>5</sup>, observing broadly similar alteration frequencies between the two cohorts (Supplementary Fig. 29). We then trained a random forest machine learning model for oncotype classification using RGAs in PPCG and classified prostate cancer samples from TCGA (Supplementary Fig. 30, Supplementary Data 26). As validation, the oncotype classification in TCGA was also associated with progression-free intervals<sup>59</sup> when adjusting for patient

age (likelihood ratio  $P = 0.0030$ ; log-rank  $P = 4.2 \times 10^{-4}$ ; Fig. 5E). As expected, oncotypes in PPCG were differentiated by clinical and genomic characteristics, explaining differences in risk (Fig. 5F–M).

Next, we examined the genomic and molecular characteristics of oncotypes in PPCG. The oncotypes appeared genomically distinct and were enriched with specific sets of RGAs, with 86 associations in total (two-sided hypergeometric test,  $FDR < 0.05$ ; Fig. 6A, Supplementary Data 27). We found two high-risk oncotypes, oncotype 4 and oncotype 2, with a 10-year metastasis incidence of 35% and 22%, respectively. Oncotype 4, termed “*TP53* / *LOH*” (93 samples), exhibited frequent CNA losses, including *PIK3R1* ( $FDR = 7.5 \times 10^{-52}$ , 90% tumours), *MAP3K1*, chr8p and *NKX3-1*, *TP53*, or *PTEN*, and significantly higher chromosomal instability ( $P = 2.5 \times 10^{-41}$ , one-sided Wilcoxon test; Fig. 5K). Oncotype 2, which we term “*MYC* / *BRCA2*” (198 samples), was enriched in gains of chr8q and *MYC* ( $FDR = 1.9 \times 10^{-18}$ , 40% tumours), *MDM4* and *GATA2*, biallelic inactivation of *CHD1* and *BRCA2*, and SNVs or indels in *CDK12*, *SPOP*, as well as *EPHA1-AS1*. In agreement with MFS associations, the two high-risk oncotypes were characterised by higher tumour grade, T-stage and prostate-specific antigen (PSA) levels

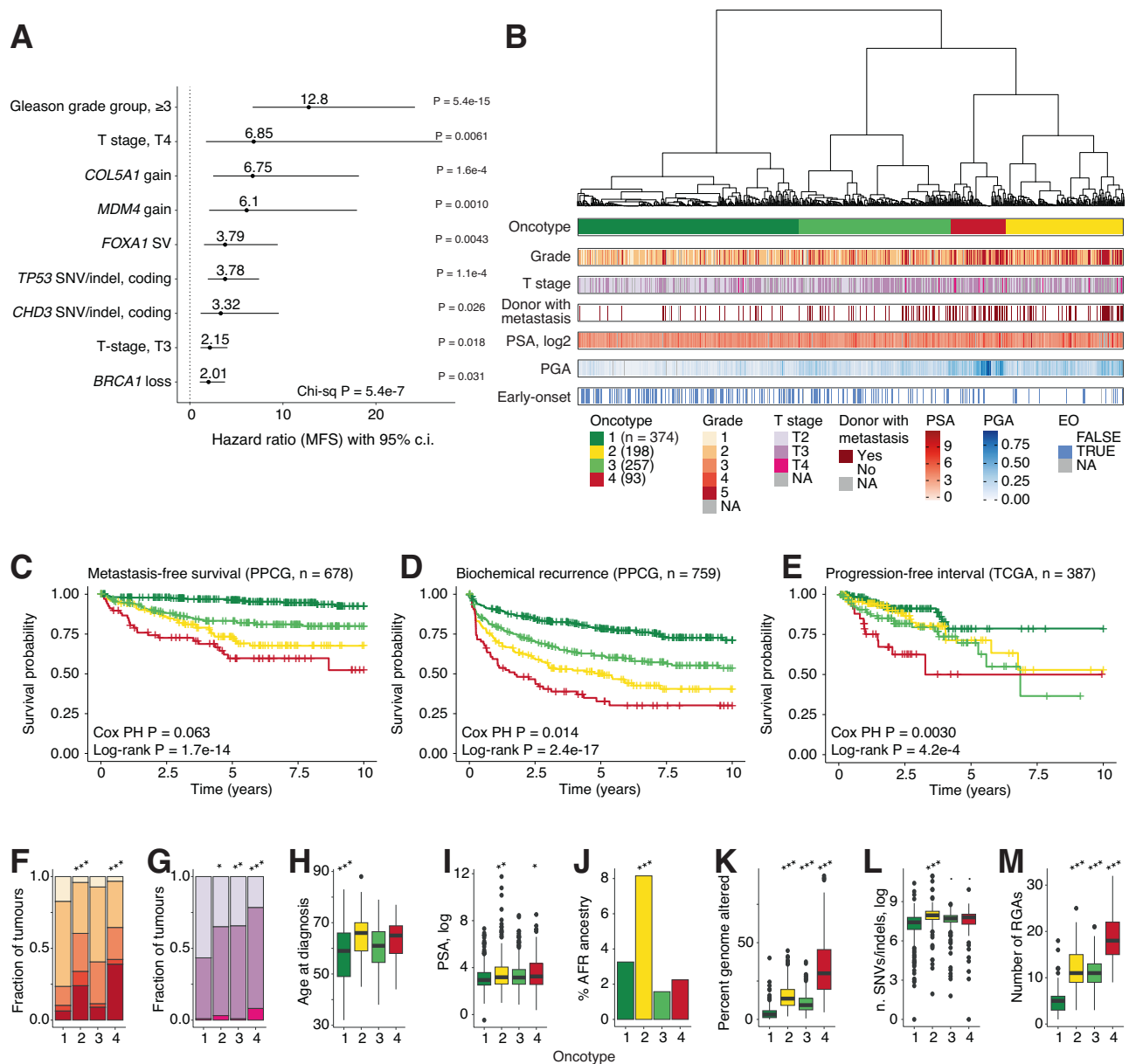


**Fig. 5 | Clinical characteristics of prostate cancer drivers.** **A** RGAs associated with metastasis-free survival (MFS) from multivariable Cox proportional-hazards (CoxPH) regression models using 709 tumours with complete clinical annotations, accounting for clinical covariates including tumour grade (Gleason grade group), T-stage, and donor age at diagnosis. RGAs identified by univariable screening (chi-square test, unadjusted  $P < 0.1$ ) were preselected for inclusion in the multivariable CoxPH model. Significant RGAs and clinical covariates from the multivariable model are reported (two-sided Wald tests, unadjusted  $P < 0.05$ ). Points show hazard ratios (HRs), and error bars indicate 95% confidence intervals. The chi-square  $P$ -value reflects a non-directional likelihood-ratio test comparing the full model with a reference model including only clinical variables. **B** Four oncotypes of prostate cancer identified by consensus hierarchical clustering of RGA profiles across 922 tumour samples with at least one RGA, visualised as a dendrogram. Colour strips below the dendrogram show clinical and genomic annotations (top to bottom): grade (Gleason grade group), T-stage, donor with metastasis developed during follow-up, prostate specific antigen (PSA), genomic instability (percent genome altered, PGA), and early-onset (EO) disease defined as  $\leq 55$  years at age of diagnosis). **C–E** Kaplan-Meier plots of the four oncotypes in PPCG in the context of metastasis-free survival (**C**; 678 tumour samples) and biochemical recurrence (BCR) (**D**; 759 tumour samples). Survival analysis using CoxPH regression was adjusted for Gleason grade group (1–2 versus 3–5), T-stage, and donor age at diagnosis ( $\leq 55$  years versus  $> 55$  years, defining early- and late-onset disease,

respectively). The  $P$ -value from the likelihood ratio test on the CoxPH model adjusted for clinical covariates and the log-rank  $P$ -value with no clinical adjustment are shown. **E** Kaplan-Meier plot showing progression-free intervals of the four oncotypes in the validation cohort from TCGA (387 tumours). The  $P$ -value from the likelihood ratio test of TCGA samples accounts for early-onset status.

**F–M** Characterisation of oncotypes in the context of clinical and genomic variables. To evaluate significance of each oncotype, tumours within the oncotype were compared to all other tumours combined (one-sided Wilcoxon rank-sum tests; lower tail for age and higher tail for other variables). African ancestry was evaluated with a one-sided hypergeometric test. Asterisks indicate statistical significance (\* $< 0.05$ , \*\* $< 0.01$ , \*\*\* $< 0.001$ ). Boxplots show the median (centre line), interquartile range (IQR; box, 25th–75th percentiles), whiskers extend to the most extreme values within 1.5-fold IQR, and points indicate outliers. **F** Grade for 892 tumour samples ( $P_2 = 2.4 \times 10^{-12}$ ,  $P_4 = 1.2 \times 10^{-9}$ ). **G** T-stage for 790 tumour samples ( $P_2 = 0.011$ ,  $P_3 = 0.0042$ ,  $P_4 = 7 \times 10^{-6}$ ). **H** Donor age at diagnosis for 898 tumour samples ( $P_1 = 1.9 \times 10^{-11}$ ). **I** Prostate specific antigen (PSA) at tumour collection for 907 samples ( $P_2 = 0.0084$ ,  $P_4 = 0.034$ ). **J** African vs non-African ancestry for 895 tumour samples ( $P_2 = 9 \times 10^{-4}$ ). **K** Percent genome altered for 916 tumour samples ( $P_2 = 1.4 \times 10^{-26}$ ,  $P_3 = 1.7 \times 10^{-5}$ ,  $P_4 = 2.6 \times 10^{-41}$ ). **L** Number of SNVs/indels for 922 tumour samples ( $P_2 = 1.5 \times 10^{-18}$ ,  $P_3 = 0.054$ ,  $P_4 = 0.077$ ). **M** Number of RGAs for 922 tumour samples ( $P_2 = 5.2 \times 10^{-17}$ ,  $P_3 = 2.1 \times 10^{-14}$ ,  $P_4 = 1.1 \times 10^{-44}$ ). Source data for (**A**, **E**, **K–M**) are provided as Source Data files.

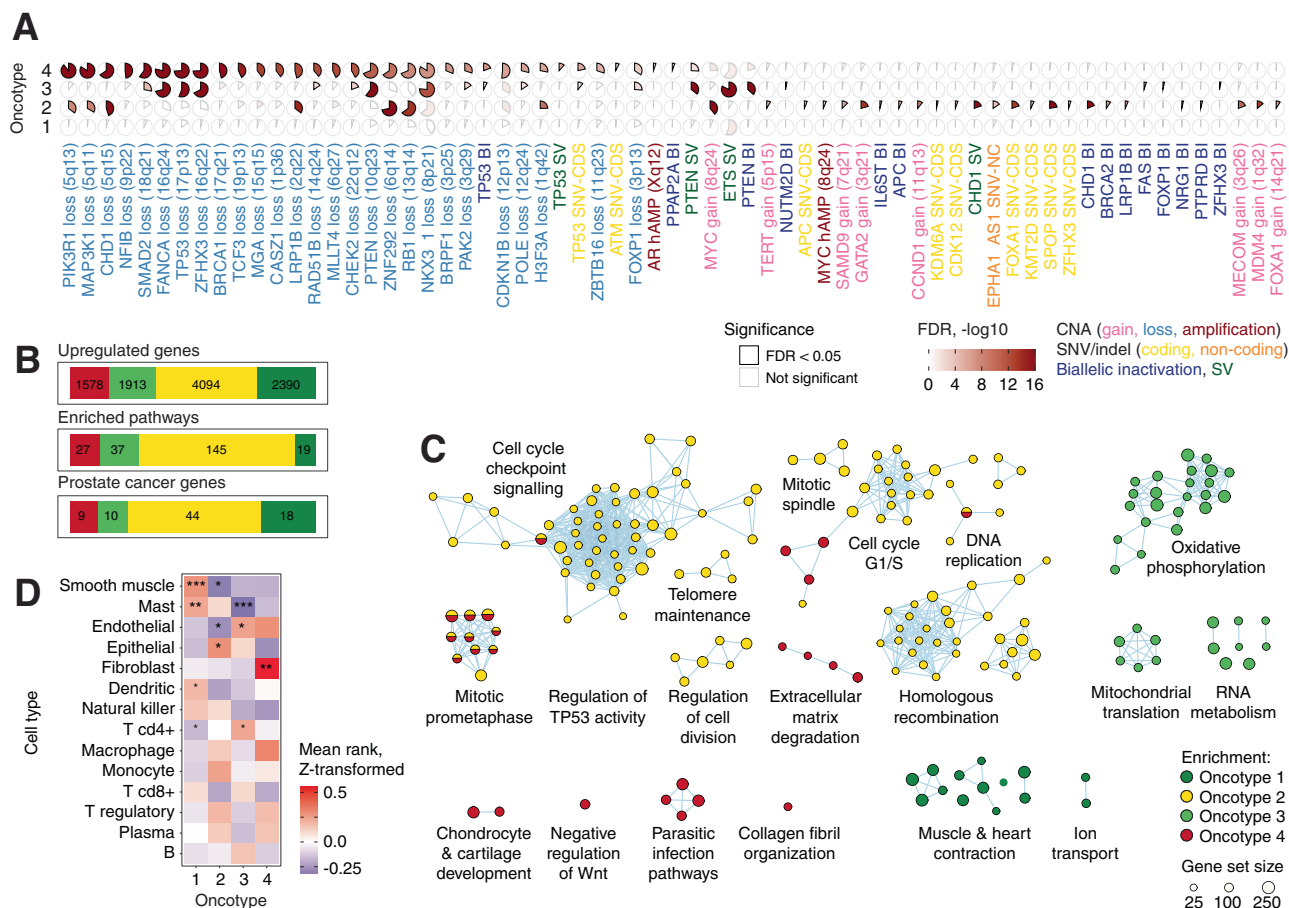


**Fig. 6 | Functional characteristics of RGA-based oncotypes.** **A** RGAs enriched in each oncotype relative to all other oncotypes combined across 922 tumour samples (one-sided hypergeometric tests,  $FDR < 0.05$ ). Pie charts show the fraction of tumour samples within each oncotype harbouring the respective RGA. Solid-coloured pie charts indicate significant RGA enrichments, and greyed-out pie charts indicate non-significant associations (CDS.mut: protein-coding SNVs, indels; NC.mut: non-coding SNVs, indels; amp: amplification). **B** Transcriptomics analyses of each oncotype relative to other oncotypes combined, comparing up-regulated genes ( $FDR < 0.05$ ) and enriched pathways across 473 tumour samples with oncotype assignments and matching transcriptomics data. Colours indicate oncotypes. **C** Pathway enrichment analysis of oncotypes across 473 tumour samples with oncotype assignments and matching transcriptomics data. Significant pathways and processes visualised as an enrichment map ( $FWER < 0.05$ , ActivePathways). The network represents enriched

processes and pathways as nodes coloured by the relevant oncotypes, while edges between nodes indicate similar pathways or processes sharing many genes, organised as functional themes with characteristic processes annotated. Multiple-testing correction was performed using Holm family-wise error rate (FWER) in ActivePathways. **D** TME characteristics of oncotypes across 473 tumour samples with oncotype assignments and matching transcriptomics data, based on cell type deconvolution. Cell type signatures in each oncotype were compared with those in all other oncotypes combined and adjusted for multiple testing ( $FDR < 0.05$ , two-sided Wilcoxon tests). The colour scale shows the mean rank-normalised z-score of each TME signature within each oncotype, with red indicating higher and blue indicating lower relative abundance compared with the cohort-wide distribution. Asterisks indicate statistical significance (\*  $< 0.05$ , \*\*  $< 0.01$ , \*\*\*  $< 0.001$ ). Source data for (A–D) are provided as Source Data files.

(Fig. 5F–I). Oncotype 2 was enriched in PPCG donors of African ancestry (Fig. 5J), consistent with the disproportionately higher burden of aggressive prostate cancer reported in men of African ancestry<sup>60</sup>. In contrast, the largest and lower-risk oncotype 1, termed “RGA-desert” (374 samples), had only a 5% incidence of metastasis within ten years and was characterised by fewer genomic alterations and RGAs and no oncotype-specific RGAs were detected (Fig. 5M), however it was characterised by earlier age of diagnosis and lower grade and stage (Fig. 5F, G, H, M). Lastly, the intermediate-risk oncotype 3 with 257 samples was enriched in *ETS* alterations ( $FDR = 1.9 \times 10^{-31}$ , 82% tumours) and SVs, losses and BI events of *PTEN*.

Finally, we characterised the oncotypes at a molecular level using transcriptomics and tumour microenvironment (TME) signatures of corresponding tumour samples, comparing each oncotype with other oncotypes combined (Fig. 6B–D, Supplementary Data 28). The higher-risk oncotype 2 was the most transcriptomically distinct and involved four thousand upregulated genes and 145 enriched pathways (adjusted  $P$ -value  $< 0.05$ , Supplementary Data 29), including DNA replication,

cell proliferation, and homologous recombination processes. The second higher-risk oncotype 4 was enriched in processes related to extracellular matrix degradation and proliferation. We used cell type deconvolution to assess TME characteristics, which showed enrichment of fibroblasts in oncotype 4, consistent with cancer-associated fibroblasts that have been linked to poor prognosis and tumour progression in solid cancers<sup>61</sup>. Oncotype 2 was enriched in epithelial signatures that likely reflect malignant cells, suggesting that these are mainly TME-depleted tumours<sup>62</sup>. In contrast, the lower-risk oncotype 1 was enriched for muscle regulation pathways and characterised by TME signatures of smooth muscle and mast cells, potentially reflecting a low tumour purity and relatively higher contribution of normal cells to these indolent tumours. This suggests that its low metastatic potential reflects the relative absence of high-risk genomic, transcriptomic and microenvironmental features rather than a single dominant indolent program. Lastly, the intermediate-risk oncotype 3 was enriched in energy metabolism and RNA processing pathways. Comparison of the four oncotypes with integrated mutational

footprints (IMFs)<sup>63</sup> showed that oncotype 2 was enriched for IMF6, characterised by replication stress, as well as IMF5, characterised by non-canonical impaired homologous recombination (Supplementary Fig. 31) in line with the transcriptomic and pathway-level signatures of this oncotype (Fig. 6B). In contrast, the oncotype 1 was enriched in the non-CIN subset of genomically stable tumours. Collectively, these data suggest complex interactions between driver alteration patterns and clinical characteristics in prostate cancer.

## Discussion

As a genetic disease, prostate cancer is initiated and propagated by genomic alterations, in particular SVs and CNAs. Here, we investigated a large WGS dataset of prostate cancers to identify recurrent genomic alterations as cancer drivers. Besides large CNAs, most prostate cancers have only a handful of driver alterations and a sizable fraction have none at all, and few established prognostic signals beyond overall genomic instability are known. Therefore, analysing whole cancer genomes with well-annotated clinical parameters is critical for clinical translation, therapy development, and mechanistic understanding. Here we extend the landscape of prostate cancer drivers by identifying recurrent complex SVs, non-coding SNVs and indels, oncogenic duplications, and biallelic inactivations of TSGs, many of which are supported by complementary clinical, transcriptomic, or pathway-level evidence. Unsupervised clustering of prostate cancers based on drivers reveals clinically relevant subtypes, including aggressive tumours defined by either recurrent oncogene alterations or loss of heterozygosity of TSGs. Compared to variants in protein-coding regions, the genomic territory outside of protein-coding regions (~98%) controlling transcription, replication, and genome maintenance remains poorly explored. This is particularly important in prostate cancer, which has an abundance of non-coding SVs and few protein-coding SNVs and indels. We found several recurrent SV hotspots at the solute carrier (SLC) genes, including *SLC30A4*. SLC family proteins transport organic molecules in and out of the cell<sup>64</sup> and play important roles in exocrine glandular organs such as the prostate<sup>65</sup>. Here, we provide evidence that the primary driver mechanism of the recurrent SVs at SLCs is enhancer hijacking to upregulate distal oncogenes. We show that the frequency of SLC enhancer hijacking is proportional to SLC gene expression in prostate tissue. We also find the most frequently hijacked SLC enhancers to be AR-regulated, which is known to be mutagenic, suggesting that AR binding contributes to the SV-mediated enhancer hijacking. It is plausible that AR binding at SLC enhancers facilitates both elevated transcriptional activity and SV formation. Our findings also support the utility of WGS in identifying the non-coding regulatory driver alterations in cancer.

As part of PPCG, our study has benefited from a large and uniformly processed dataset with extensive clinical data, enabling us to answer key questions concerning the clinical relevance of cancer drivers in prostate cancer, especially in relation to tumour metastasis. We note that sequencing depth and intratumoral heterogeneity are likely to lead to an underestimation of this driver landscape in prostate cancer. In companion studies, we explore clonal and subclonal mutations through trajectories of tumour evolution and their causal roles in evolutionary divergence in prostate cancer<sup>66,67</sup>. Significant recurrence of genomic alterations is a reliable indicator of positive selection, however sequencing artefacts or unknown locus-specific mutational processes may contribute false positives to driver analyses. For increased stringency, we focused on high-confidence variant calls supported by two workflows, thereby removing lower-confidence or sub-clonal variants. Epigenetics can also drive cancer gene dysregulation including suppression of TSGs, which we explore in our companion study<sup>68</sup>. Patient ancestry is an important determinant of the driver alteration types, for example entailing much lower frequencies of ETS fusions and higher rates of *FOXAI* in men of black African ancestry<sup>39</sup>. Our

PPCG cohort is enriched in European ancestry and future studies in diverse ancestries<sup>39,69</sup> will enable a more complete picture of genomic alterations driving prostate cancer.

Precision oncology aims to tailor treatment to the distinct genomic features of a patient's cancer, enhancing therapeutic efficacy while minimizing toxicity. WGS emerges as the most comprehensive platform for profiling prostate cancer in clinical settings. The benefits include a higher genomic resolution at reduced processing costs and the identification of the next frontier of candidate cancer drivers among infrequent and non-coding alterations. These insights are valuable for identifying potential therapeutic targets or biomarkers in smaller, molecularly defined subsets of cancer patients. Detailed inventories of genomic events responsible for tumour initiation, progression, therapy response, and metastasis enable more accurate patient profiling, aiding in the selection of clinical trials or active surveillance strategies. Simultaneously, systematic genome editing of RGAs combined with phenotypic analysis will deepen our understanding of the genetic and molecular pathways driving aggressive prostate cancer, ultimately paving the way for innovative targeted therapies. Our study of coding and non-coding recurrent genomic alterations in primary prostate cancer and their clinical associations supports further basic and translational research.

## Methods

### Ethics

We used data from the PPCG consortium of primary prostate cancer samples from a total of 1,001 prostate cancer donors. Written informed consent was obtained at clinical follow-up, and all procedures were conducted in accordance with local regulations and International Cancer Genome Consortium (ICGC) ethical guidelines (<https://icgc.org/>). Ethical approval was obtained from local research ethical committees. The full list of approving boards/committees includes: 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).

### Somatic SNVs and indels and structural variants (SVs)

Somatic SNVs and indels from index samples of primary prostate cancers from 961 donors were retrieved from PPCG consensus variant call sets of the human genome hg19<sup>14</sup>. We selected high-confidence variants detected by two indel or SNV calling pipelines in PPCG and excluded two hypermutated samples with >30 SNVs or indels per megabase. Functional effects of SNVs and indels on protein-coding genes were mapped using the ANNOVAR software (ver. 2019-10-24)<sup>70</sup> using RefSeq hg19 gene definitions. Mutational signatures of SNVs and indels were retrieved from the PPCG consensus dataset<sup>63</sup> and each SNV and indel was assigned the signature with the highest probability. Somatic SVs were also retrieved from the PPCG consensus dataset for

961 primary prostate cancers<sup>14</sup>. SVs in ETS family genes were considered jointly as a separate composite driver event as described below. For driver analyses, SVs were defined as two breakpoints (SVBPs) comprising the start and end coordinates of the SV.

### ETS driver status

ETS alteration status was annotated separately from other RGA analyses based on PPCG ETS consensus calls that were integrated across SV, CNA, and transcriptomics datasets that comprised ETS genes and translocation partners (*ERG*, *TMPRSS2*, *SLC45A3*, *ETV4*, *ETV1*). Alterations of multiple types were merged and CNA gains at the *TMPRSS2/ERG* were not included. We compiled a consensus mutation status of ETS family of transcription factors (*ERG*, *ETV1*, *ETV4*, *ETV5* and *FLI1*) using fusion predictions from our SV and CNA calls<sup>14</sup> and from PPCG RNA-seq data<sup>14</sup>. Tumour samples were labelled as ETS positive if at least one of the ETS family genes was predicted to have a fusion.

### Genes and genomic regions

Genomic regions for driver discovery included protein-coding exons, untranslated regions (5'UTR, 3'UTR) and non-coding RNA genes from Gencode v19, and regulatory and chromatin architectural genomic elements from prostate cancer epigenomics studies, with 725,000 genomic elements in total. Gene coordinates were derived from Gencode v19 and genes with duplicated symbols were removed. Protein-coding exons (CDS) and UTRs were merged into unique consecutive regions per gene. We included ~14,000 non-coding RNA genes from the lincRNA, antisense, miRNA, and sense-intronic classes and removed very long ncRNAs (>50 kbp). We also derived full gene coordinates to annotate non-coding elements and detect SV-based drivers. To define non-coding genomic elements, we subtracted CDS regions from non-coding regions and removed small remaining fragments (<50 bp). Several classes of non-coding regulatory elements were used. Open-chromatin elements of prostate cancer samples in the TCGA prostate cancer cohort were retrieved from a previous ATAC-seq study<sup>22</sup> and mapped from hg38 to hg19 using the liftOver software<sup>71</sup>. Androgen receptor (AR) binding sites from normal prostates (NARBS) and prostate tumours (TARBS) were retrieved from an earlier study (GSE70079)<sup>20</sup>. Chromatin states of prostate cancers were retrieved from the ChromHMM profiles of an earlier study (GSE130408)<sup>23</sup> and filtered to include active chromatin elements (ActEnh\_nonProst, ActProm\_nonProst, ActEnh\_Prost, ActProm\_Prost, BivEnh, BivProm, PrimEnh\_nonProst, PrimEnh\_Prost) while repressive and heterochromatin states (HetChrom, RepChrom) were excluded. Chromatin loop anchors identified from prostate cancer cell lines were retrieved from an earlier study (GSE164347)<sup>21</sup>. Non-coding genomic regions were annotated to adjacent genes within ±3 kbp and to distal sites using long-range chromatin interactions, permitting multiple genes per region. Distal annotations were based on long-range chromatin interactions<sup>21</sup> such that one anchor of the interaction overlapped the non-coding region and the other anchor overlapped a gene transcription start site (TSS) of the target gene within ±1 kbp. Chromosomal interactions with adjacent genes were excluded. To interpret drivers, we collected prostate cancer driver genes from two studies<sup>7,8</sup>, and for driver genes of other cancer types, we used the Cancer Gene Census (CGC) database<sup>32</sup> (downloaded 2024-09-18). For pathway analyses, gene sets corresponding to Gene Ontology biological processes and Reactome molecular pathways were retrieved from the g:Profiler software<sup>72</sup> (downloaded 2024-09-17).

### Copy-number alterations (CNAs)

Somatic CNAs for chromosomes 1-22 and X were retrieved from the consensus dataset for 961 primary cancers from PPCG<sup>14</sup>. Tumour samples were classified based on ploidy estimates as diploid (ploidy <2.5) or whole-genome duplicated (WGD). Short segments were filtered (<1 kbp). Segments were annotated based on total copy number

(minor + major allele CN) as balanced, gained (min 3 copies for diploid, 6 for WGD), lost (1 for all), with homozygous deletion (0 for all), or with high-number amplification (min 5 for diploid, 8 for WGD), or high-number amplification in chrX (3 for diploid, 4 for WGD). To study chromosomal instability, percent genome altered (PGA) scores were derived as fractions of autosomal genomes in any non-balanced states. We annotated the CN status of each gene in each tumour sample, requiring at least 50% of gene sequence in a non-balanced CN state, excluding ChrX genes with homozygous loss events.

### Somatic and germline biallelic inactivation (BI)

To systematically identify genes with BI events from the PPCG WGS cohort<sup>14</sup>, we used our previously published methodology<sup>2,12</sup> expanding its previous definitions and generalized the method in the python module PyBiallelic. Analyses were performed on 961 primary cancers in PPCG. Briefly, to identify tumour suppressor deficiencies, we defined BI as a gene locus GA/B, where alleles A and B are genetically altered, leading to a genetic Gmut/mut state. The BI assessment includes three genetic inactivation event types consisting of somatic deletions ("Loss"), somatic SVs ("Break") and somatic or germline SNVs ("Mutation"). Given a heterozygous GA/B locus, we required a loss of the A allele of the gene, leading to a hemizygous G/B state, and genetic inactivation of the remaining B allele, specifically requiring the second event to overlap the loss on the A allele, leading to BI. We considered four classes of BI: Loss/Mutation: A heterozygous loss combined with a damaging germline SNP (germ\_snp/som\_loss) or nonsynonymous driver mutations of the B allele (som\_loss/som\_snp), including subclonal SNVs (som\_loss/subclonal\_snp). This category also comprises loss of heterozygosity (LOH), where the A allele carrying a driver somatic SNV is amplified, resulting in a copy-neutral LOH (som\_cn\_loh/som\_snp) or an amplification with LOH (som\_gain\_loh/som\_snp). Loss/Loss: Two deletion events that overlap an exon, resulting in a copy-number derived allele count of 0 for both the A and B alleles (som\_loss/som\_loss). Loss/Break: A heterozygous loss combined with SVs where one or both breakpoints are situated in an exon of the B allele (som\_loss/som\_sv). Mutation/Mutation: Two nonsynonymous driver somatic SNVs of the same gene, requiring both to be either stopgain or out-of-frame SNVs. Next, we inferred the germline mutation to occur on the A allele and the somatic mutation on the B allele, with the assumption that two independent driver mutation events are highly unlikely to occur on the same allele. To test the significance of a BI gene, given its mutational background in the cohort, we constructed a binary model using the curveball method (version 0.0.1)<sup>18</sup>. In this model, a randomized background is defined by permutations of the observed data, maintaining the mutational load for each tumour sample and the number of donors with single or BI of a specific gene constant. The observed binary matrix was created with columns for allele A and allele B for each gene. A single hit results in 1/0 for the A/B columns, respectively, while a BI results in 1/1. We applied this randomization process 10,000 times and computed the p-value by evaluating the number of times the background showed a higher occurrence of BI than the observed data, divided by the number of randomizations. We selected known cancer driver genes in prostate cancer<sup>7,8</sup> or other cancer types<sup>32</sup> having significant BI (curveball  $P < 0.05$ , permutation test) in at least five tumour samples for further analyses. PyBiallelic with a suite of tools for handling and parsing multiple types of alterations is available below: (<https://bitbucket.org/weischenfeldt/biallelic.py>; <https://biallelic-py.readthedocs.io/en/latest/>).

### Driver analysis of protein-coding mutations

Small mutations in protein-coding genes (SNVs/indels) were studied for positive selection and driver gene predictions using the dNdScv method (version 0.0.1.0)<sup>19</sup>. Results for the global\_cv metric were adjusted for multiple testing using FDR across tested genes and filtered for significance (FDR < 0.05). Protein-coding impact of SNVs/

indels was derived from dNdScv and further annotated as substitutions (missense, deletion inframe) or truncations (nonsense, stop loss, deletion frameshift, essential splice, insertion frameshift), while excluding silent mutations.

**Driver analysis of regions with non-coding SNVs/indels and SVs**  
Driver analysis of non-coding small mutations (SNVs and indels) and SV breakpoints was performed with the ActiveDriverWGS method (version 1.2.1)<sup>16</sup> that employs a Poisson regression model to identify genomic elements with enriched somatic SNVs and indels relative to a flanking background sequence and trinucleotide-based mutation frequencies. Driver analysis was conducted separately for small mutations (SNVs, indels) and SV breakpoints (encoded as 1-bp indels in the model) across ncRNA genes and non-coding regulatory elements. For SVs in protein-coding genes, we analysed entire gene sequences rather than protein-coding exons. The background sequence parameter in ActiveDriverWGS was defined for the analyses of different alteration classes. For SNVs and indels, the default value ( $\pm 50$  kbp) was used. For SVs, a wider background window of  $\pm 500$  kbp around genomic elements was used due to the sparser genome-wide distribution of SVs. P-values were adjusted for multiple testing using the Benjamini-Hochberg FDR across all tested elements and SNV/indel and SV analyses. Significant elements were selected (FDR < 0.05).

### Identification of recurrent SNV/indel hotspots

Recurrent SNV/indel hotspots were defined from single-base and indel events centred on the mutated genomic position. We used a 3-bp hotspot window comprising the focal base and one flanking base on either side, expanding the hotspot boundaries accordingly when overlapping indels extended beyond this interval. We retained hotspots present in at least 1% of the cohort. Each hotspot was then annotated as overlapping a predefined protein-coding RGA, a non-coding RGA, tested but not significant (based on our ActiveDriverWGS analyses), or not tested because it did not overlap any predefined genomic element. Hotspots were annotated based on the RGA in which they were located, or, where no RGA annotation was available, using adjacent-gene annotations from the consensus PPCG variant annotation datasets.

### Merging and annotating RGAs

Following driver analysis, significant non-coding elements were merged based on overlapping coordinates to reduce redundancy from input sets of genomic elements. Merging was done separately for drivers from small mutations (SNVs, indels), and drivers from SVs, and the merged drivers were assigned the superset of associated tumour samples and the minimal FDR value from the respective merged elements. Target genes of non-coding elements were annotated manually, giving priority to known driver genes in prostate cancer, driver genes involved in other cancer types, or other protein-coding genes. Top RGAs with min 1% frequency in PPCG were selected by integrating multiple alteration types (coding SNVs and indels, non-coding SNVs and indels, SVs, BIs, amplifications) and annotated using prostate cancer genomics studies<sup>7,8</sup> and the Cancer Gene Census database for known cancer genes. We also compared top RGAs with genes targeted by the Illumina TruSight Oncology 500 (TSO500) sequencing panel ([https://www.illumina.com/content/dam/illumina-marketing/documents/products/gene\\_lists/gene\\_list\\_trusight\\_oncology\\_500%20ctDNA%20with%20fusions.xlsx](https://www.illumina.com/content/dam/illumina-marketing/documents/products/gene_lists/gene_list_trusight_oncology_500%20ctDNA%20with%20fusions.xlsx)), to evaluate if the RGAs would be detectable using standard panel-based approaches. Similarly, we evaluated RGAs with protein-coding SNVs or indels in the top list to assess applicability of whole-exome sequencing for RGA discovery. We also asked which types of non-coding genomic elements were over- or underrepresented among the elements found as RGAs, combining regions enriched in SNVs/indels and SVs (ActiveDriverWGS, FDR < 0.05) and excluding protein-coding exons for SNVs/indels and gene

bodies for SVs. We used a permutation test by sampling sets of genomic elements randomly from the total set of testable genomic elements to derive a background distribution and evaluated the proportions of genomic element types in the true set of results relative to the background distribution. Empirical P-values were adjusted for multiple testing using FDR and filtered for significance (FDR < 0.05). Median expected values with 95% confidence interval from the background distribution were reported.

### Enrichment analysis of mutational signatures and SV classes in drivers

Statistical interactions of SNVs/indels in RGAs and mutational SBS and ID signatures were analysed with hypergeometric tests that compared the genome-wide distribution of a given signature and its presence in a specific gene or non-coding RGA. Similarly, SV-based RGAs were analysed for types of SV events (translocation, duplication, deletion, inversion). The analysis was limited to SBS or ID signatures with at least 1000 SNVs or indels, and RGAs with at least 10 altered tumour samples. The analysis excluded known artefactual mutational signatures according to the COSMIC database and the previously unreported SBS96D signature identified in prostate cancer<sup>63</sup>. Additional sets of SNVs/indels in driver genes were also analysed for SBS signatures: truncation-inducing mutations (nonsense, frameshift); substitution-inducing mutations (missense, non-frameshift), and non-coding mutations. Resulting P-values were adjusted for multiple testing and filtered for significance (FDR < 0.05).

### Analysis of recurrent copy-number alterations (CNAs)

CNA calls from the consensus PPCG dataset were used, with small CNA segments removed (<1 kbp). Two strategies were used to find recurrent CNAs. First, we used the GISTIC2 software<sup>17</sup> (version 2.0.24) on copy-number segments for 961 primary tumour samples. The following parameters were used for GISTIC: “-smallmem 0 -broad 1 -brlen 0.5 -conf 0.90 -armpeel 1 -gcm extreme”. Significant CNAs from GISTIC2 were selected (residual FDR < 0.05). Secondly, in order to find single genes or focal regions with enriched CNAs, we used a custom algorithm that first identified individual genes with recurrent CNAs and secondly grouped these into adjacent gene sets that were frequently co-altered by the same CNA segments. The algorithm involved three steps. First, we computed the enrichment of CNA events per gene using a binomial statistic parameterised by the total number of CNA events in all protein-coding genes and all tumour samples in PPCG, selecting recurrently altered genes (FDR < 0.05). Second, we tested pairs of significantly altered genes per chromosome based on the copy-number segments altering the genes using hypergeometric tests and selected gene pairs that were often altered by the same CNAs (FDR < 0.05). Third, pairs of significantly co-altered genes were structured as a non-directional interaction network (i.e. graph) and connected components were identified using a recursive approach to derive genomic regions with frequent CNAs as defined by gene sets. The analysis was conducted separately for three types of CNA events (gains, losses, high-number amplifications) while homozygous losses were analysed separately in the context of biallelic inactivation. Finally, results from GISTIC2 and co-CNA analyses were reviewed manually. Index genes were then selected for most recurrently altered genomic regions based on manual curation by prioritising prostate cancer genes<sup>7,8</sup> or genes from other cancer types<sup>32</sup> with the highest alteration rates in the regions, considering CNAs of matching types (gains or losses) and filtering a few adjacent redundant regions as well as *TMPPRS/ERG* deletions that were curated separately.

### Comparison of RGAs in primary and metastatic tumours

We compared the prevalence of RGAs between the 961 primary tumours and metastatic tumours, including all 103 metastatic cases and the subset of 46 de novo metastatic cases of the PPCG cohort. For

a subset of cases with multiple samples per donor, the sample with the highest purity was selected. For each RGA observed in at least 1% of the primary cohort, we defined tumour-level alteration status using the same driver definitions as in the primary analyses, including protein-coding SNVs/indels, non-coding SNVs/indels, SV breakpoints, CNA gains, losses and high-level amplifications, and BI events. For each driver, metastatic enrichment was tested by binomial test using the alteration frequency in the primary cohort as the background expectation. Separate analyses were performed for all metastatic tumours and for de novo metastatic tumours. P-values were adjusted across all tested RGAs across the two analyses using the Benjamini-Hochberg FDR procedure.

### African ancestry analysis of RGAs

To compare RGAs across ancestry groups, we analysed a published independent cohort of 113 treatment-naïve prostate cancers from men of African ancestry<sup>39</sup>, comparing these with the PPCG cohort of 927 non-African tumours, after excluding samples with predicted African ancestry and those of uncertain ancestry from the PPCG cohort. For each RGA, we aggregated tumour counts by alteration class from the earlier study, including protein-coding SNV/indel drivers, non-coding SNV/indel drivers, SVs, gains, losses and high-level amplifications. CNAs were annotated based on GISTIC2 calls (gain=1; amplification=2; loss=1). ETS alterations were handled as a composite event across canonical fusion partners. We then tested whether each RGA was enriched in the African ancestry cohort relative to the non-African PPCG cohort using a binomial test, treating the non-African frequency from PPCG as the background expectation. RGAs observed at <1% frequency in both cohorts were excluded, and P-values were adjusted across all tested RGAs using the Benjamini-Hochberg FDR procedure (FDR < 0.05).

### Transcriptomics analyses

We used the PPCG consensus transcriptomics dataset (version 1.2) that was processed using a harmonised pipeline (<https://github.com/panprostate/>) and the RUV-III method<sup>73</sup> as described in the PPCG data resource publication<sup>14</sup>. We selected the subset of 494 transcriptomes that were derived from a tumour sample with a matching WGS sample ID, in order to reduce the effect of intratumoral heterogeneity on integrating genomic and transcriptomic information. Genes with duplicated or missing symbols were removed. For each driver alteration, we compared gene expression levels of target genes of tumours with and without alterations, limiting the analysis to cases with at least three altered tumours having transcriptomics and WGS data. We separately tested drivers based on non-coding SNVs or indels, SV breakpoints, BI events, CNA gains, losses, and amplifications. Gene expression levels were compared using non-parametric Wilcoxon rank-sum tests, adjusted for multiple testing across the set of genes tested for expression differences, and significant results were reported (FDR < 0.05). To determine genes from Cancer Gene Census<sup>32</sup> and prostate cancer genes<sup>7,8</sup> with significantly correlated expression with *EPHA1-AS1*, we first derived a background distribution of Spearman correlation coefficients from one million randomly sampled gene pairs using the PPCG transcriptomics dataset that included lncRNA expression levels. 2.5% and 97.5% quantiles from the background distribution of correlation coefficients were used to determine the cancer genes showing significant negative or positive correlation with *EPHA1-AS1*, respectively (empirical  $P < 0.05$ ).

### Pathway enrichment analysis of RGAs

We performed an integrative analysis of molecular pathways and biological processes affected by RGAs using the ActivePathways method (version 2.0.4)<sup>41</sup> that used as input a matrix of FDR values with RGAs as rows and the types of alterations as columns. ActivePathways assigned a higher priority to genes affected by multiple alteration

types and allowed us to evaluate the types of genomic alterations significantly contributing to each altered pathway. This analysis included CNAs, coding SNVs/indels, non-coding SNVs/indels, BI events, and amplifications. ActivePathways was run with parameters defining gene set sizes (min 25, max 250), Benjamini-Hochberg FDR as the multiple-testing correction method, cutoff (value 0.5) to account for FDR-adjusted input P-values. Biological processes from Gene Ontology and molecular pathways from Reactome were used for the analysis. Significantly enriched pathways were selected using a more stringent cutoff to balance sensitivity from analysing manually-curated genes from large and frequent CNAs (FDR < 0.01). The results were visualised as an enrichment map<sup>74</sup> in Cytoscape software. Subnetworks corresponding to sets of related pathways were annotated manually to identify major functional themes<sup>75</sup>. Representative pathways from major themes were selected and visualised as oncoprint plots displaying the genes and alteration types in each pathway.

### Co-occurrence and mutual exclusivity of RGAs

We tested pairs of driver alterations for statistical interactions (i.e., co-occurrence or mutual exclusivity) using hypergeometric tests over the sets of tumour samples affected by specific alterations. This analysis focused on drivers from SNV/indel, SVBP, BI, and CNA gain and amplification analyses. We excluded CNA losses due their high frequency and co-occurrence patterns in a subset of tumours. Alterations with 1% frequency in the cohort were selected and multiple types of alterations per gene or locus were collapsed. All pairs of alterations were tested and significant interactions were selected after multiple-testing corrections across all RGA-RGA pairs (FDR < 0.05). To highlight the long tail of recurrent events in ETS-/SPOP-negative tumours, we asked which RGAs were enriched or depleted in the corresponding subset of 376 tumours. RGAs were collapsed across alteration types, RGAs with at least 1% frequency in this subset were included, and the RGAs were visualised as an oncoprint plot. Associations with ETS-/SPOP- status were detected using two-sided hypergeometric tests, P-values were adjusted for multiple testing and significant findings were reported (FDR < 0.05).

### Distal interactions of RGAs

We evaluated the potential distal target genes of RGAs detected from analyses of non-coding SNVs and indels or SV breakpoints. Distal target genes were detected through long-range chromatin interactions from prostate cancer tissues<sup>21</sup> as described above. 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<sup>31</sup>, and BRD4 binding sites were derived from prostate cancer cells (LNCaP) from a previous study<sup>52</sup>. The interactions were visualised as an interaction network in Cytoscape and the most frequently altered RGAs were further investigated. *MIPOL1* was identified as the distal target of the SV hotspot at the *FOXAI* locus, and its expression was compared between SV-positive and SV-negative tumour samples using RUV-III normalised PPCG transcriptomics data in matching samples and two-sided Wilcoxon tests were used to evaluate significance. As the other example, the lncRNAs with RGAs *NEAT1* and *MALAT1* were associated through long-range chromatin interactions and were further studied using co-expression analyses in PPCG tumour samples using Spearman correlation tests.

### Drug target analyses of RGAs

Annotated clinical evidence for drug-variant associations was obtained from OncoKB<sup>45</sup> (<https://www.oncokb.org/>, “Actionable Genes”, downloaded Dec. 14th 2024) and CiViC<sup>46</sup> (<https://civcdb.org/>, “Evidence”, “nightly build”, downloaded Dec. 14th 2024). We considered only Evidence Levels 1-4 for OncoKB drug-variant associations and assigned drugs with evidence for “Prostate” or “All Solid Tumors” for prostate cancer relevant drugs. We considered CiViC drug-variants evidence

levels A-E and that fulfilled the following criteria: `evidence_type == 'Predictive'` & `evidence_direction == 'Supports'` & `significance == 'Sensitivity/Response'` & `variant != 'Wildtype'`. CIViC drug-variants with “disease” evidence for “Prostate” or “Solid Tumor” were considered prostate cancer relevant drugs.

### Transcriptomic and pathway-level analyses of specific RGAs

To characterise the transcriptional and functional consequences of specific RGAs in *ZNF280D* and *COL5A1*, we compared matched transcriptomic profiles of tumours carrying each RGA against tumours lacking that RGA using matching transcriptomics data from PPCG. Transcriptome-wide differential expression signatures were assessed using two-sided Wilcoxon rank-sum tests, *P*-values were adjusted across all tested genes, and significant genes were reported (FDR < 0.05). Separately, all gene-level *P* values were used for pathway enrichment analysis with the ActivePathways method<sup>41</sup>, using GO Biological Process and Reactome gene sets and restricting analyses to gene sets containing 25–250 genes. The background gene universe for ActivePathways analysis was defined by the set of genes represented in the PPCG transcriptomics data having any pathway annotations. Significant pathways were selected following the default multiple testing correction method in ActivePathways (family-wise error rate (FWER) < 0.05). For *COL5A1* analyses, we excluded genes co-gained with *COL5A1* from transcriptomic and pathway-level analyses to avoid confounding effects.

### Clinical analyses of RGAs

Clinical information of the 961 primary tumour samples was derived from the PPCG consensus dataset<sup>14</sup>. Metastasis-free survival (MFS) intervals were right-censored at the 10 year time point. Donors with missing values or zero values in time intervals were excluded from MFS analyses. We determined the associations of RGAs with MFS given clinical covariates as categorical variables: Gleason grade group (<3 vs ≥3), T-stage (T2, T3, T4), and age of diagnosis (≤55 years as early-onset vs >55 years as late-onset disease). Donors lacking any clinical covariate values, time intervals or event status values, or tumours with T1 stage were excluded from these analyses. After filtering missing values, RGAs with at least 10 tumour samples altered in the PPCG cohort were considered for MFS analyses. Different alteration types were considered separately. Four major types of analyses were conducted. First, we asked if each of the RGAs was individually associated with MFS when adjusted for clinical covariates. For each RGA, Cox proportional hazards (CoxPH) regression models were used for MFS analyses using the presence of the RGA and clinical covariates (T-stage, Gleason grade group, early-onset status; as described above). The significance of each RGA was determined using ANOVA and chi-square tests that compared RGA-based, clinically-adjusted MFS models to control MFS models comprising only clinical covariates. Clinically-adjusted hazard ratio (HR) values and the 95% confidence intervals (i.e., 1.96 standard errors) of the RGAs were derived from model coefficients. *P*-values were adjusted for multiple testing across the subset of RGAs for which the tests were conducted, and significant associations were highlighted (FDR < 0.05). Second, we conducted a multivariable analysis of MFS. We first determined RGAs that were individually associated with MFS in RGA-informed univariate CoxPH models without adjusting for clinical covariates (chi-square tests, *P* < 0.1). This subset of significant RGAs was then used together with clinical covariates (T-stage, Gleason grade group, early-onset status; as described above) in a single multivariable CoxPH model of MFS. From this CoxPH model we selected significant RGAs and clinical covariates using an unadjusted *P*-value threshold (Wald test, *P* < 0.05), with no multiple-testing applied due to the single-model nature of this test. The results were visualised as a forest plot with HR values and 95% confidence intervals (i.e., 1.96 standard errors). To associate RGAs with donor age at diagnosis, we used univariate logistic regression models with binary age values as

model response (≤55 years as early-onset (EO); >55 years as late-onset (LO)). We tested each RGA for which at least 10 tumour samples were available with age information, conducted multiple testing corrections and reported significant results after multiple-testing correction across all selected RGAs (FDR < 0.05), and grouped the results as EO-enriched or LO-enriched based on model coefficients. Third, we split the cohort based on median number of RGAs per tumour sample (i.e., median-dichotomised RGA burden) and asked if this RGA burden was associated with MFS when adjusting for clinical covariates (grade, stage, age; as above) using CoxPH models. Different types of alterations as well as all alterations combined were tested and *P*-values from log-likelihood tests and HR values with 95% confidence intervals were reported. Fourth, we asked if tumours missing any of the clinical variables required for MFS analysis (metastasis status, metastasis time interval, Gleason grade group, T-stage, age) were associated with certain RGAs. Enrichments or depletions of RGAs in the subset of tumours with missing clinical data were tested using two-sided hypergeometric tests and adjusted for multiple testing across the RGAs, and significant results were reported (FDR < 0.05). We separately examined RGAs at the *KLK2* locus in the context of MFS and separately analysed SVs and non-coding SNVs/indels as well as the two alteration types combined. No significant associations with MFS were found cohort-wide while too few clinically-annotated ETS-negative samples were available for analysis. Lastly, clinical and genomic annotations of tumour samples and donors were visualised as histograms along with genomic information such as sequencing depth, number of small variants, and CNA prevalence as percent genome altered, with counts and median values shown for categorical and numerical data, respectively.

### Discovery of oncotypes using RGA-based hierarchical clustering

To discover groups of tumour samples with similar profiles of RGAs, we performed consensus hierarchical clustering of the tumour samples. To define RGA-based oncotypes, we first constructed a binary tumour-by-RGA matrix using RGAs present in at least 1% of tumours, excluding a small subset of tumours where no RGAs were found. When multiple alteration types mapped to the same gene in a given sample, they were collapsed to a single binary event. We then performed consensus hierarchical clustering using the ConsensusClusterPlus method (version 1.62.0)<sup>58</sup>. Consensus clustering was applied using Euclidean distance and Ward.D2 linkage with 10,000 bootstrap resampling iterations, subsampling 80% of tumours and 80% of RGAs per iteration. The number of clusters was evaluated across multiple numbers (*k* = 2–10) using consensus metrics and silhouette scores, and the final solution was selected at four clusters, based on stable silhouettes and a significant drop of per-tumour consensus scores in the next potential solution. Patient-level cluster membership and cluster certainty were then extracted from the consensus clustering output. The oncotypes and RGA profiles were visualised as a heatmap, a dendrogram, silhouette and consensus scores, as well as clinical and genomic annotations including Gleason grade group, T-stage, whether the donor developed a metastasis during the follow-up period, PSA levels at tumour collection, genomic instability score (i.e., percent genome altered), and age-based grouping (early vs late-onset disease). The association of oncotypes with MFS and biochemical recurrence (BCR) in PPCG was tested using CoxPH models adjusted by clinical covariates (Gleason grade group, T-stage, donor age) as described above. Log-likelihood tests as well as log-rank tests were used to evaluate the survival differences of oncotypes. Each oncotype was also compared with all other oncotypes combined for several genomic and clinical annotations. Continuous variables (Gleason grade group, T-stage, age at diagnosis, prostate-specific antigen, percent genome altered, number of SNVs/indels, and number of RGAs) were tested using one-sided non-parametric Wilcoxon rank-sum tests, with the lower tail used for age and the upper tail used for the remaining

variables. Categorical annotations, such as African ancestry, were tested using one-sided hypergeometric tests. To determine the RGAs enriched in each oncotype, we compared the RGA frequency in a given oncotype to all other oncotypes combined using one-sided hypergeometric tests for higher frequency, adjusted the results for multiple-testing across the RGAs tested, and reported significant results for RGAs (FDR < 0.05). Using a similar strategy, we evaluated the overlap of our RGA-based oncotypes and integrated mutational footprints based on mutational processes in prostate cancer<sup>63</sup>. Two-sided hypergeometric tests were used to determine over- or under-representation of the two sets of annotations, adjusted for multiple testing across all tested annotation pairs, and significant results were reported (FDR < 0.05).

### Transcriptomic and pathway-level characterization of oncotypes

Transcriptomic differences between RGA-based oncotypes were evaluated using matching RNA-seq data from the PPCG cohort. Analyses were restricted to 473 tumour samples with both RNA sequencing data and assigned driver-based oncotypes from WGS data. For each oncotype, gene expression levels were compared between tumours assigned to that oncotype and all remaining tumours using two-sided Wilcoxon rank-sum tests. Differential expression analyses were performed independently for each oncotype across all expressed genes. Effect sizes were reported as log<sub>2</sub> fold changes computed from the difference in mean RUV-III-normalized expression between oncotype-positive and remaining tumours. *P*-values were adjusted for multiple testing across all oncotypes and genes, and genes with FDR < 0.05 were considered significantly differentially expressed. Biological pathways associated with each driver-based oncotype were identified using the ActivePathways method<sup>41</sup>. For each oncotype, gene-level differential expression *P*-values were used as input for pathway enrichment analyses of Gene Ontology Biological Process and Reactome gene sets. Analyses were performed separately for each oncotype, considering gene sets containing 25–250 genes. Resulting pathway-level *P*-values across all oncotypes were adjusted using the default Holm family-wise error rate (FWER) procedure and filtered for significance (FWER < 0.05). Significant pathways were visualised as enrichment maps in Cytoscape, with pathway nodes annotated according to the oncotypes contributing to each enrichment and curated manually into functional themes.

### Tumour microenvironment (TME) analysis of oncotypes

TME composition was assessed in a subset of PPCG tumours for which matching transcriptomics data was available. Cell-type deconvolution profiles of prostate tumour samples (version 4) were retrieved from the TME working group of PPCG<sup>14</sup>. Briefly, deconvolution of PPCG transcriptomics data was performed using the BayesPrism method (version 2.2.2)<sup>76</sup> and a collection of signatures derived from single-cell transcriptomics datasets of prostate samples<sup>77–82</sup>. Each cell type and cell state was represented by at least 200 cells. Single-cell data was filtered for protein-coding genes excluding genes in X and Y chromosomes and ribosomal and mitochondrial genes. Default settings for Gibbs sampling and optimization were used, together with thresholds of 0.1 for the outlier fraction and 0.01 for the outlier cut. To analyse TME features of the four oncotypes, we restricted the analysis to the subset of 473 tumours with both oncotype assignment and matched transcriptomic data. To interpret oncotypes in the context of TME for each cell-type signature, we compared cell type scores across oncotypes using two-sided Wilcoxon rank-sum tests, contrasting each oncotype against all other oncotypes combined. Multiple-testing correction across all cell-type and oncotype combinations was performed and significant results are highlighted (FDR < 0.05). For visualisation, TME scores were rank-normalised across samples and z-transformed

from the rank-normalised values, and the heatmap shows the mean rank-normalised z-score for each oncotype.

### Cis expression analysis to identify enhancer hijacking at SLC genes

We used the structural alteration mapping (CESAM) approach to identify SLC genes involving enhancer hijacking events. CESAM was applied as previously described<sup>55</sup> with the following modifications. We obtained contact domains and loop regions from prostate tumour tissue<sup>21</sup>. The genome was binned using contact domains (TADs) in which the presence or absence of breakpoints was annotated. In addition, the presence or absence of loops was also considered for each bin. We then genotyped each contact domain bin for breakpoints (rows) in each tumour sample (column). We applied a linear regression model in which the normalised gene expression was regressed on the genotype matrix of breakpoints, contact domains and loops, using the specificities of the sample countries and the first three principal components of the gene expression matrix as covariates. For transcriptomics studies of SLC genes, we retrieved the GTEx v10 dataset representing gene-level TPM values by tissue (GTEx\_Analysis\_v10\_RNASeQCv2.4.2\_gene\_tpm.gct.gz) from the GTEx portal (release version 10). Gene expression differences in prostate tissues were calculated as log<sub>2</sub> fold-change difference between median of 282 prostate tissue samples and median of 19,334 samples representing 67 different non-prostate tissue samples.

### Recurrent SVs, epigenetic marks, and expression of SLC genes

We intersected recurrent SV hotspots at SLC genes with epigenetic marks (H3K4me3, H3K27ac) and CTCF binding sites (derived from prostate cancer cells PC-3, C4-2B) from the ENCODE project<sup>51</sup>, and BRD4 binding sites derived from prostate cancer cells (LNCaP)<sup>52</sup>. We also examined enhancer regions identified in prostate cancer samples<sup>23</sup>. We identified SLC-oncogene SVs as one breakpoint within 20 kbp of SLC enhancers and the other breakpoint within 20 kbp of known prostate cancer oncogenes (*ERG*, *ETV4*, *ETV1*, *FLII*, *SKIL*, *MYC*, *ETV5*, *AR*, *ETS1*, *ELK4*). DNA-binding sites of AR were obtained from an earlier study<sup>20</sup>. Binding sites in normal prostate tissue and prostate tumours were both considered and analysed jointly. For each SLC gene, we determined AR binding sites in transcription start sites (TSS) as defined through BestRefSeq transcript in GRCh37<sup>83</sup>. To find SLC genes potentially regulated by AR, SLC genes were grouped by the presence or absence of AR binding sites (ARBS) located within ±1 kbp of their TSSs. Mean expression of each SLC gene in prostate tumour samples in PPCG was calculated based on RUV-III-normalised values across PPCG transcriptomics data as described above. SLC genes lacking RNA-seq data were excluded from the analysis. A reference distribution of RNA expression of SLC genes lacking ARBS was constructed using the *ecdf* function in the R stats package. Mean expression values of SLC genes having ARBS in their TSSs were compared to the defined reference empirical distribution. The subset of ARBS-present SLC genes with significantly higher expression relative to the reference distribution (*P* < 0.05) were predicted to be regulated by AR activity. Prostate SLC gene expression fold-change from GTEx data was performed as described in the transcriptomics analysis section above.

### Analysis of recurrent non-coding SNVs and indels in transcription factor (TF) binding motifs

To explore the potential functional effect of non-coding SNVs and indels in the *ZNF280D* RGA in the context of TF binding motifs, we detected the presence of TF binding motifs using the FIMO software (version 5.4.1)<sup>84</sup> within a sequence window of ±25 bp around each SNV or indel of interest. FIMO analysis was performed on both the reference sequence (i.e. without SNV/indel) and the altered sequence (i.e. with SNV/indel) using default parameters of FIMO and the core set of TF binding motifs from the HOCOMOCO database (v11)<sup>85</sup>. A motif was

considered to be potentially lost due to the SNV/indel if it was only detected in the reference sequence but not in the altered sequence, based on default significance cutoffs of FIMO. Vice versa, a motif was considered to have a potential de novo gain through the SNV or indel if the motif was only detected significantly in the altered sequence but not in the reference sequence.

### Processing of TCGA genomic alterations for external validation

To validate oncotypes in an independent cohort, we analysed WGS and clinical data from The Cancer Genome Atlas prostate adenocarcinoma (TCGA-PRAD) cohort<sup>5</sup>. Somatic SNVs and indels were obtained from VarScan2 somatic variant calls, SVs from Manta calls, and CNAs from GATK4 copy-number segment data retrieved through the Genomic Data Commons portal. Only variants passing the respective caller quality filters (FILTER = PASS) were retained for analysis. SVs involving non-standard chromosomes were excluded. Hypermutated tumours (>30 SNVs/indels per megabase) were excluded. After filtering, 387 TCGA tumours were retained for analysis. For patients represented by multiple tumour aliquots, a single representative sample was selected by choosing the alphabetically first tumour sample identifier. To facilitate direct comparison with the PPCG cohort, analyses of TCGA genomic alterations were restricted to RGAs identified in the PPCG discovery analysis, excluding BI events that were unavailable for TCGA. Protein-coding SNVs/indels in the respective RGAs were annotated using dNdScv<sup>19</sup>. For RGAs enriched in non-coding SNV/indels or SV breakpoints, genomic coordinates of PPCG driver elements were converted from hg19 to hg38 using liftOver and overlapping TCGA variants were identified using genomic interval overlap analyses. ERG fusion status was obtained from TCGA annotations available through cBioPortal<sup>5,86</sup> and represented as a composite ETS driver event for RGA analyses. CNA segments supported by more than five probes were annotated as losses, gains, or high-level gains using segment mean thresholds of -0.2, 0.4, and 0.8, respectively. CNA index genes corresponding to PPCG RGAs were then annotated using the same strategy as in PPCG, requiring at least 50% of the gene sequence to overlap a segment of the corresponding CNA state.

### Validation of driver-based prognostic oncotypes in TCGA

To classify tumours in the external TCGA validation cohort according to the four driver-based oncotypes identified in PPCG, we trained supervised machine-learning models using the PPCG cohort as a reference set. Binary driver matrices were constructed for PPCG and TCGA cohorts using the harmonized set of RGAs. To reduce sparsity and improve model robustness, only RGAs altered in at least 1% of PPCG tumours (excluding BI events) were retained, resulting in a set of 83 recurrent RGAs for classification. Tumours lacking previously assigned oncotypes in PPCG were excluded from model training. Random forest classifiers were trained using RGA status as predictor variables and PPCG oncotypes as outcome classes. Model hyperparameters were optimized using five-fold cross-validation with class balancing by down-sampling and Cohen's kappa statistic as the performance metric. The final tuned random forest model was then applied to assign TCGA tumours to the previously defined PPCG driver-based oncotypes. The prognostic relevance of the four externally assigned driver-based oncotypes was evaluated using progression-free interval (PFI) data from the TCGA-PRAD cohort, where 75% of PFI events represented biochemical recurrence. Clinical outcome data and donor age were obtained from the integrated TCGA pan-cancer clinical data resource, and follow-up times were right-censored at 10 years. CoxPH regression models were fitted with oncotypes as a categorical variable and oncotypes 1 as the reference group, corresponding to the lowest-risk group in PPCG. Because detailed clinicopathological variables used in the PPCG survival analyses were unavailable for the TCGA cohort, models were adjusted only for early-onset disease status (age at diagnosis ≤55 years). To assess the

prognostic contribution of oncotypes assignments, models including early-onset status and oncotypes were compared with models containing only early-onset status using likelihood ratio tests from the CoxPH model. Hazard ratios and 95% confidence intervals were derived from model coefficients, and statistical significance was determined using chi-square tests of the nested models.

### Statistics and reproducibility

No statistical method was used to predetermine sample size. The analysed cohort comprised all available primary prostate cancer samples with WGS data passing PPCG consensus quality control. Sex was not considered in the analyses because prostate cancer is a male-specific disease. Individual analyses were restricted to samples with the required genomic, transcriptomic, or clinical data. Samples and variants were excluded only according to predefined quality-control criteria or data availability requirements. WGS samples failing PPCG consensus quality control were excluded before this study, as described in the PPCG data resource publication<sup>14</sup>. High-confidence SNV and indel analyses were limited to variants detected by at least two independent PPCG variant-calling pipelines. No samples or variants were excluded based on study outcomes. The study was not randomized and investigators were not blinded because this was a retrospective computational analysis of existing genomic and clinical datasets. Statistical tests, sample sizes, and significance measures are reported in the corresponding Methods sections and figure legends. P-values were adjusted for multiple testing where appropriate. Reproducibility was assessed through independent validation and orthogonal analyses. Findings from WGS data were further evaluated using transcriptomic, pathway, tumour microenvironment, and clinical analyses. Key findings were also examined in independent datasets, including TCGA primary prostate cancers, metastatic PPCG tumours, and an external cohort of prostate cancers from men of African ancestry. The RGA-defined oncotypes were independently validated in the TCGA cohort using a random forest classifier trained on PPCG samples, reproducing significant associations with progression-free intervals.

### 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 2021 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 Archive (EGA) (<https://ega-archive.org/studies/EGAS00001002876>) under the study ID EGAS00001002876 as described in the PPCG data resource publication<sup>14</sup>. 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](mailto: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. The TCGA-PRAD whole-genome sequencing variant calls were derived from controlled-access TCGA data. Access can be obtained through the National

Cancer Institute Genomic Data Commons (GDC) following approval by the relevant dbGaP Data Access Committee. The publicly available datasets on prostate cancer epigenomics used in this study are available from the NCBI Gene Expression Omnibus (GEO) database under accession codes GSE164347 [ref.<sup>21</sup>] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164347>), GSE70079 [ref. 20] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70079>), GSE209889 [ref. 52] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE209889>) and GSE130408 [ref. 23] (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130408>). Single-cell transcriptomics datasets used for TME analyses are available from GEO or the European Genome-phenome Archive (EGA) under accession codes GSE176031 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176031>) [ref. 77], GSE137829 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137829>) [ref. 78], EGAS00001005115 (<https://www.ega-archive.org/studies/EGAS00001005115>) [ref.<sup>79</sup>], EGAS00001005787 (<https://www.ega-archive.org/studies/EGAS00001005787>) [ref. 80], GSE141445 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE141445>) [ref. 81], and GSE150692 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150692>) [ref. 82]. ATAC-seq peak calls for TCGA-PRAD [ref.<sup>22</sup>] are available from GDC (<https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>). GTEx transcriptomics data (release 10) are available from the GTEx portal (<https://www.gtexportal.org/home/downloads/adult-gtex%23qtl>). Gencode gene annotations (version 19) are available from [https://www.encodegenes.org/human/release\\_19.html](https://www.encodegenes.org/human/release_19.html). Epigenetic marks profiled in prostate cancer cells are available from the ENCODE data portal (<https://www.encodeproject.org/>). Annotated clinical evidence for drug-variant associations is available from the databases OncoKB<sup>45</sup> (<https://www.oncokb.org/>) and CiViC<sup>46</sup> (<https://civcdb.org>). The remaining data are available within the Article or Supplementary Information. Source data are provided with this paper.

## Code availability

The source code for reproducing analyses and figures is available at Github (<https://github.com/panprostate/ppcg-drivers>). A static code repository is available at Zenodo (<https://doi.org/10.5281/zenodo.22262026>)<sup>87</sup>.

## References

- James, N. D. et al. The Lancet Commission on prostate cancer: planning for the surge in cases. *Lancet* **403**, 1683–1722 (2024).
- Gerhauser, C. et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996–1011.e8 (2018).
- van Dessel, L. F. et al. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nat. Commun.* **10**, 5251 (2019).
- Rebello, R. J. et al. Prostate cancer. *Nat. Rev. Dis. Prim.* **7**, 9 (2021).
- Cancer Genome Atlas Research Network The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
- Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
- 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).
- Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
- Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
- Fraser, M. et al. Somatic driver mutation prevalence in 1844 prostate cancers identifies ZNRF3 loss as a predictor of metastatic relapse. *Nat. Commun.* **12**, 6248 (2021).
- Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
- Jakobsdottir, G. M. et al. The Pan Prostate Cancer Group Dataset - genomic, transcriptomic and methylation data for clinical discovery. *Sci. Data* (2026).
- Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
- Zhu, H. et al. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. *Mol. Cell* **77**, 1307–1321 (2020).
- Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
- Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).
- Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- Pomerantz, M. M. et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).
- Hawley, J. R. et al. Reorganization of the 3D genome pinpoints noncoding drivers of primary prostate tumors. *Cancer Res.* **81**, 5833–5848 (2021).
- Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
- Pomerantz, M. M. et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat. Genet.* **52**, 790–799 (2020).
- Finch, R. A. et al. Mdmx is a negative regulator of p53 activity in vivo. *Cancer Res.* **62**, 3221–3225 (2002).
- He, B. et al. GATA2 facilitates steroid receptor coactivator recruitment to the androgen receptor complex. *Proc. Natl. Acad. Sci. USA* **111**, 18261–18266 (2014).
- Rodriguez-Bravo, V. et al. The role of GATA2 in lethal prostate cancer aggressiveness. *Nat. Rev. Urol.* **14**, 38–48 (2017).
- Merson, S. et al. Focal amplification of the androgen receptor gene in hormone-naïve human prostate cancer. *Br. J. Cancer* **110**, 1655–1662 (2014).
- Edwards, J., Krishna, N. S., Grigor, K. M. & Bartlett, J. M. S. Androgen receptor gene amplification and protein expression in hormone-refractory prostate cancer. *Br. J. Cancer* **89**, 552–556 (2003).
- Linja, M. J. et al. Amplification and overexpression of androgen receptor gene in hormone-refractory prostate cancer. *Cancer Res.* **61**, 3550–3555 (2001).
- Tomlins, S. A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
- Weischenfeldt, J. et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
- Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
- Leung, J. W. C. et al. ZMYM3 regulates BRCA1 localization at damaged chromatin to promote DNA repair. *Genes Dev.* **31**, 260–274 (2017).
- Williams, S. A., Xu, Y., De Marzo, A. M., Isaacs, J. T. & Denmeade, S. R. Prostate-specific antigen (PSA) is activated by KLK2 in prostate

- cancer ex vivo models and in prostate-targeted PSA/CLK2 double transgenic mice. *Prostate* **70**, 788–796 (2010).
35. Bonk, S. et al. Reduced CLK2 expression is a strong and independent predictor of poor prognosis in ERG-negative prostate cancer. *Prostate* **80**, 1097–1107 (2020).
  36. Chen, Y., Tang, Y., Wang, M.-T., Zeng, S. & Nie, D. Human pregnane X receptor and resistance to chemotherapy in prostate cancer. *Cancer Res.* **67**, 10361–10367 (2007).
  37. Zhang, B. et al. Pregnane X receptor as a therapeutic target to inhibit androgen activity. *Endocrinology* **151**, 5721–5729 (2010).
  38. Parolia, A. et al. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* **571**, 413–418 (2019).
  39. Jaratlerdsiri, W. et al. African-specific molecular taxonomy of prostate cancer. *Nature* **609**, 552–559 (2022).
  40. Reyna, M. A. et al. Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **11**, 729 (2020).
  41. Paczkowska, M. et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* **11**, 735 (2020).
  42. Jakobsdottir, G. M. et al. Hypoxia spurs evolution of aggressive prostate cancer. *Nat. Commun.* (2026). [Placeholder reference].
  43. Gan, W. et al. SPOP promotes ubiquitination and degradation of the ERG oncoprotein to suppress prostate cancer progression. *Mol. Cell* **59**, 917–930 (2015).
  44. He, B. et al. Diverse noncoding mutations contribute to deregulation of cis-regulatory landscape in pediatric cancers. *Sci. Adv.* **6**, eaba3064 (2020).
  45. Suehnholz, S. P. et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. *Cancer Discov.* **14**, 49–65 (2024).
  46. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
  47. Lotan, T. L. et al. Homologous recombination deficiency (HRD) score in germline BRCA2- versus ATM-altered prostate cancer. *Mod. Pathol.* **34**, 1185–1193 (2021).
  48. Samstein, R. M. et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
  49. Singh, C. K. et al. Analysis of zinc-exporters expression in prostate cancer. *Sci. Rep.* **6**, 36772 (2016).
  50. Henshall, S. M. et al. Expression of the zinc transporter ZnT4 is decreased in the progression from early prostate disease to invasive prostate cancer. *Oncogene* **22**, 6005–6012 (2003).
  51. Encode Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
  52. Li, M. et al. LSD1 inhibition disrupts super-enhancer-driven oncogenic transcriptional programs in castration-resistant prostate cancer. *Cancer Res.* **83**, 1684–1698 (2023).
  53. Quigley, D. A. et al. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **174**, 758–769.e9 (2018).
  54. Northcott, P. A. et al. Enhancer hijacking activates GF11 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
  55. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
  56. Tomlins, S. A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595–599 (2007).
  57. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
  58. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
  59. Liu, J. et al. An integrated TCGA Pan-Cancer Clinical Data Resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018).
  60. Gong, J. et al. Genetic and biological drivers of prostate cancer disparities in Black men. *Nat. Rev. Urol.* **21**, 274–289 (2024).
  61. Chen, Y., McAndrews, K. M. & Kalluri, R. Clinical and therapeutic relevance of cancer-associated fibroblasts. *Nat. Rev. Clin. Oncol.* **18**, 792–804 (2021).
  62. Bagaev, A. et al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell* **39**, 845–865.e7 (2021).
  63. Gruber, A. J. et al. Integrated signatures define mutational processes in prostate cancer. *Nature* <https://doi.org/10.1038/s41586-026-10468-w> (2026).
  64. Khan, A., Liu, Y., Gad, M., Kenny, T. C. & Birsoy, K. Solute carriers: the gatekeepers of metabolism. *Cell* **188**, 869–884 (2025).
  65. Bai, X., Moraes, T. F. & Reithmeier, R. A. F. Structural biology of solute carrier (SLC) membrane transport proteins. *Mol. Membr. Biol.* **34**, 1–32 (2018).
  66. Fernández-Sanromán, Á. et al. Evolutionary trajectory and momentum shape prostate cancer outcomes. *Nat. Commun.* (2026). [Placeholder reference]
  67. Esentürk, E. et al. Causes of evolutionary divergence in prostate cancer. *Nat. Commun.* (2026). [Placeholder reference]
  68. Kiri, D. et al. Recurrent mechanisms of biallelic epigenetic inactivation reveal new putative tumour suppressor genes in prostate cancer. *Nat. Commun.* **17**, 9646 (2026).
  69. Li, J. et al. A genomic and epigenomic atlas of prostate cancer in Asian populations. *Nature* **580**, 93–99 (2020).
  70. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
  71. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
  72. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).
  73. Molania, R. et al. Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat. Biotechnol.* **41**, 82–95 (2023).
  74. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
  75. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).
  76. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517 (2022).
  77. Song, H. et al. Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nat. Commun.* **13**, 141 (2022).
  78. Dong, B. et al. Single-cell analysis supports a luminal-neuroendocrine transdifferentiation in human prostate cancer. *Commun. Biol.* **3**, 778 (2020).
  79. Wu, S. Z. et al. Cryopreservation of human cancers conserves tumour heterogeneity for single-cell multi-omics analysis. *Genome Med.* **13**, 81 (2021).
  80. Tuong, Z. K. et al. Resolving the immune landscape of human prostate at a single-cell level in health and cancer. *Cell Rep.* **37**, 110132 (2021).
  81. Chen, S. et al. Single-cell analysis reveals transcriptomic remodellings in distinct cell types that contribute to human prostate cancer progression. *Nat. Cell Biol.* **23**, 87–98 (2021).

82. Crowley, L. et al. A single-cell atlas of the mouse and human prostate reveals heterogeneity and conservation of epithelial progenitors. *Elife* **9**, e59465 (2020).
83. Thibaud-Nissen, F. et al. P8008 the NCBI Eukaryotic Genome Annotation Pipeline. *J. Anim. Sci.* **94**, 184–184 (2016).
84. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
85. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).
86. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
87. Reimand, J. Panprostate/ppcg-drivers: Pellegrina, Kiri, Cheng et al. *Nat. Commun.* <https://doi.org/10.5281/ZENODO.22262026> (2026).

## Acknowledgements

The authors would like to thank those men with prostate cancer and the subjects who have donated their time and tumour samples to the data sets used in this study. We also acknowledge support of the research staff in the groups who so carefully curated the tumour samples and the follow-up data.

## Author contributions

J.R., D.P., D.K., K.C.L.C., F.F., E.M., Z.P.K., B.H., A.V.O., Y.F., W.J., W.J.H., P.Lu., C.G., and J.W. contributed to the formal analysis presented in this study. J.R. and J.W. wrote the manuscript with contributions from D.P., D.K., and K.C.L.C. J.R., D.P., D.K., K.C.L.C., E.M., F.F., and J.W. produced and contributed to the visualisations of the study. J.R. conceived and designed the study. J.R. and J.W. supervised the work. C.S.C., D.C.W., R.A.E., D.S.B., Z.K.-J., and R.G.B. made a substantial contribution to the organisation and conduct of the study and critiqued the output for important intellectual content. D.P., D.K., K.C.L.C., F.F., E.M., Z.P.K., B.H., A.V.O., A.S., Y.F., P.Lu., W.J., W.J.H., G.S.B., D.S.B., M.B., B.B., A.B., G.C.-T., N.M.C., O.C., A.G., E.G.G., V.J.G., A.H., H.H., E.L.I., G.M.J., C.H.J., F.K., Z.K.-J., P.La., G.L., M.L., R.M., L.R.Q., T.R., B.R., K.D.S., S.U., Y.X., T.N.Y., C.Z., S.M., L.V.v.V., L.M., A.T.P., B.P., V.M.H., T.S., R.G.B., C.M.H., R.A.E., C.S.C., C.G., G.M., A.J.G., D.C.W., J.W., J.R., and all other members of the Pan Prostate Cancer Group provided access to data and/or contributed to gathering, processing and curating data. A full list of Pan Prostate Cancer Group members can be found in the Supplementary Information. All authors had access to all data in the study. All authors contributed to the review and editing of the manuscript and approved the manuscript before submission.

## Funding

(D.K.): 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). (B.H.): hosted by the Centro Nacional de Investigaciones Oncológicas (CNIO), which is supported by the Instituto de Salud Carlos III and recognized 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). (Y.F.): The Lorenzo and Pamela Galli Medical Research Trust. (D.S.B.): We 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. (G.C.): 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). (N.M.C.): Analysis of the Australian samples was funded via a PRECEPT programme grant, co-funded by Movember and the Australian Federal Government (Principal Investigator Niall M. Corcoran). (O.C.): 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). (V.J.G.): infrastructure support from the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). (V.M.H.): 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 Programs (CDMRP) Prostate Cancer Research Program (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 USA. Prostate Cancer Foundation (PCF) Challenge Award (2023CHAL4150). Hayes was further supported by the Petre Foundation via the University of Sydney Foundation (Australia). (H.H.): H.H. is supported by CIHR project grants (142246, 152863, 152864, 159567 and 438793), Terry Fox New Frontiers Program Project Grants (1090 and 1124), and Tier 1 Canadian Research Chair in RNA Medicine. (G.J.): This research was co-funded by the NIHR Manchester Biomedical Research Centre (NIHR203308). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. (M.L.): M.L.'s work is supported by the National Cancer Institute grants P50CA211024, P01CA265768, R01 CA259200, the USA Department of Defense (DoD) grants DoD PC160357, PC200390, as well as the Prostate Cancer Foundation (22CHAL05). (R.M.): The Lorenzo and Pamela Galli Medical Research Trust. (K.D.S.): 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). (L.M.): AB's work was conducted with the support of Movember through Prostate Cancer Canada. (A.T.P.): a National Health and Medical Research Council (NHMRC) Investigator Grant (2026643) and funding from the Lorenzo and Pamela Galli Medical Research Trust. The research also benefited from support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support. (B.P.): 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. (T.S.): The German ICGC program was funded by the German Federal Ministry of Education and Research (BMBF). (R.G.B.): 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). (C.M.H.): 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. (R.A.E.): We acknowledge support from Cancer Research UK C5047/

A14835/A22530/A17528, C309/A11566, C368/A6743, A368/A7990, C14303/A17197, Prostate Cancer UK (MA-TIA23-002, the Dallaglio Foundation (CR-UK Prostate Cancer ICGC Project and Pan Prostate Cancer Group), PC-UK/Movember, the NIHR support to The Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. (C.S.C.): 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. (D.C.W): This research was co-funded by the NIHR Manchester Biomedical Research Centre (NIHR203308). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work is supported by a Cancer Research UK RadNet Manchester award (C1994/A28701). (C.G.): The study was funded by the ICGC EOPC (O1KU1001 A) and ICGC-Data Mining (O1KU1505A) 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. (G.M.): G.M. is hosted by the Centro Nacional de Investigaciones Oncológicas (CNIO), which is supported by the Instituto de Salud Carlos III and recognized as a ‘Severo Ochoa’ Centre of Excellence (ref. CEX2019-000891-S) by the Spanish Ministry of Science and Innovation (MCIN/AEI/ 10.13039/501100011033). 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). (A.J.G.): This work was supported by the University of Konstanz and an Exploration grant of the Boehringer Ingelheim Foundation to A.J.G. (J.W.): JW 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 #3101-00177A, Novo Nordisk Foundation #NNF200C0060141, Sygeforsikringen Danmark # 2022-0198. (J.R.): This work was supported by the Canadian Institutes of Health Research (CIHR) Project Grant to Jüri Reimand (grants no. PJT-162410, PJT-197925) and the Investigator Award to Jüri Reimand from the Ontario Institute for Cancer Research (OICR). Funding to OICR is provided by the Government of Ontario.

## Competing interests

(V.J.G.): Honoraria from Janssen and JEB as a speaker. Member of external expert committee to AstraZeneca UK. (R.A.E.): 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 AstraZeneca 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.): 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/s41467-026-77900-7>.

**Correspondence** and requests for materials should be addressed to Joachim Weischenfeldt or Jüri Reimand.

**Peer review information** *Nature Communications* thanks Arul Chinnaiyan, Francisco Martinez-Jimenez, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Diogo Pellegrina<sup>1,50</sup>, Daria Kiri<sup>2,3,50</sup>, Kevin C. L. Cheng<sup>1,4,50</sup>, Francesco Favero<sup>2,3</sup>, Enrichetta Miletto<sup>2,3</sup>, Zoe P. Klein<sup>1,5</sup>, Barbara Hernando<sup>6</sup>, Andre V. Olsen<sup>2,3</sup>, Atef Sahli<sup>7</sup>, Yunfan Fu<sup>8,9</sup>, Pavlo Lutsik<sup>10,11</sup>, Weerachai Jaratlerdsiri<sup>12</sup>, William J. Hutchison<sup>8,9</sup>, G. Steven Bova<sup>13</sup>, Daniel S. Brewer<sup>14,15</sup>, Mark Brook<sup>16</sup>, Benedikt Brors<sup>17,18,19</sup>, Adam Butler<sup>20</sup>, Géraldine Cancel-Tassin<sup>21,22</sup>, Niall M. Corcoran<sup>23,24,25</sup>, Olivier Cussenot<sup>21</sup>, Abraham Gihawi<sup>12,26</sup>, Etsehiwot G. Girma<sup>2,3</sup>, Vincent J. Gnanapragasam<sup>27</sup>, Anis A. Hamid<sup>24</sup>, Housheng Hansen He<sup>28</sup>, Eddie Luidy L. Imada<sup>29</sup>, G. Maria Jakobsdottir<sup>30,31</sup>, Chol-hee Jung<sup>32</sup>, Francesca Khani<sup>29</sup>, Zsafia Kote-Jarai<sup>16,33</sup>, Philippe Lamy<sup>34,35</sup>, Gregory Leeman<sup>20</sup>, Massimo Loda<sup>29,36</sup>, Ramyar Molania<sup>8,9</sup>, Lucio R. Queiroz<sup>29</sup>, Tobias Rausch<sup>37</sup>, Brian Robinson<sup>29</sup>, Karina D. Sørensen<sup>34,35</sup>, Sebastian Uhrig<sup>17</sup>, Yaobo Xu<sup>20</sup>, Takafumi N. Yamaguchi<sup>38</sup>, Claudio Zanettini<sup>29</sup>, Pan Prostate Cancer Group\*, Stefano Mangiola<sup>39</sup>, Lena Voith von Voithenberg<sup>10</sup>, Luigi Marchionni<sup>29</sup>, Anthony T. Papenfuss<sup>8,9</sup>, Bernard Pope<sup>24,40,41</sup>, Vanessa M. Hayes<sup>12,26,42,43</sup>, Thorsten Schlomm<sup>44</sup>, Robert G. Bristow<sup>45,46,47</sup>, Christopher M. Hovens<sup>48</sup>,

**Ros A. Eeles** <sup>16,33</sup>, **Colin S. Cooper**<sup>26</sup>, **Clarissa Gerhauser** <sup>10</sup>, **Geoff Macintyre** <sup>6</sup>, **Andreas J. Gruber** <sup>49</sup>,  
**David C. Wedge** <sup>7</sup>, **Joachim Weischenfeldt** <sup>2,3,44</sup>  & **Jüri Reimand** <sup>1,4,5</sup> 

<sup>1</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada. <sup>2</sup>Biotech Research & Innovation Centre (BRIC)—University of Copenhagen, Copenhagen, Denmark. <sup>3</sup>Finsen Laboratory, Copenhagen University Hospital—Rigshospitalet, Copenhagen, Denmark. <sup>4</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. <sup>5</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. <sup>6</sup>Spanish National Cancer Research Centre (CNIO), Madrid, Spain. <sup>7</sup>Manchester Cancer Research Centre, The University of Manchester, Manchester, UK. <sup>8</sup>Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia. <sup>9</sup>Department of Medical Biology, The University of Melbourne, Melbourne, VIC, Australia. <sup>10</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>11</sup>Department of Oncology, KU Leuven, Leuven, Belgium. <sup>12</sup>Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, NSW, Australia. <sup>13</sup>Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. <sup>14</sup>Department of Metabolic Health, Norwich Medical School, University of East Anglia, Norwich, UK. <sup>15</sup>The Earlham Institute, Norwich Research Park, Norwich, UK. <sup>16</sup>The Institute of Cancer Research, London, UK. <sup>17</sup>Division Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>18</sup>German Cancer Consortium (DKTK), Core Center Heidelberg, and National Center for Tumor Diseases (NCT), Heidelberg, Germany. <sup>19</sup>Medical Faculty Heidelberg and Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. <sup>20</sup>Wellcome Sanger Institute, Cambridge, UK. <sup>21</sup>CeRePP (Centre de Recherche sur les Pathologies Prostatiques et Urologiques), Paris, France. <sup>22</sup>GRC n°5 Predictive Onco-Urology, APHP, Tenon Hospital, Sorbonne Université, Paris, France. <sup>23</sup>Department of Urology, Royal Melbourne Hospital, Parkville, VIC, Australia. <sup>24</sup>Department of Surgery, The University of Melbourne, Parkville, VIC, Australia. <sup>25</sup>Department of Urology, Western Health, Victoria, Australia. <sup>26</sup>Norwich Medical School, University of East Anglia, Norwich, UK. <sup>27</sup>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, Cambridge, UK. <sup>28</sup>Department of Medical Biophysics, Princess Margaret Cancer Centre, University Health Network; University of Toronto, Toronto, ON, Canada. <sup>29</sup>Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA. <sup>30</sup>Division of Cancer Sciences, The University of Manchester, Manchester, UK. <sup>31</sup>Christie Hospital, The Christie NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. <sup>32</sup>Melbourne Bioinformatics, The University of Melbourne, Parkville, VIC, Australia. <sup>33</sup>The Royal Marsden NHS Foundation Trust, London, UK. <sup>34</sup>Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. <sup>35</sup>Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. <sup>36</sup>Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. <sup>37</sup>Genome Biology, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>38</sup>Jonsson Comprehensive Cancer Center, University of California—Los Angeles, Los Angeles, CA, USA. <sup>39</sup>Adelaide University, Adelaide, SA, Australia. <sup>40</sup>Melbourne Bioinformatics, The University of Melbourne, Bedford Street, North Melbourne, Victoria, Australia. <sup>41</sup>Australian BioCommons, The University of Melbourne, North Melbourne, VIC, Australia. <sup>42</sup>Manchester Cancer Research Centre, University of Manchester, Manchester, UK. <sup>43</sup>School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa. <sup>44</sup>Department of Urology, Charité—Universitätsmedizin Berlin, Berlin, Germany. <sup>45</sup>Manchester Cancer Research Centre and Cancer Research UK Manchester Institute, The University of Manchester, Manchester, UK. <sup>46</sup>Christie NHS Foundation Trust, Manchester, UK. <sup>47</sup>Div Cancer Sciences, Faculty of Biology, Medicine & Health, University of Manchester, Manchester, UK. <sup>48</sup>Department of Surgery, The Collaborative Centre for Genomic Cancer Medicine, The University of Melbourne, Parkville, VIC, Australia. <sup>49</sup>University of Konstanz, Konstanz, Germany. <sup>50</sup>These authors contributed equally: Diogo Pellegrina, Daria Kiri, Kevin C. L. Cheng. \*A list of authors and their affiliations appears at the end of the paper.  e-mail: [joachim.weischenfeldt@bric.ku.dk](mailto:joachim.weischenfeldt@bric.ku.dk); [juri.reimand@utoronto.ca](mailto:juri.reimand@utoronto.ca)

## Pan Prostate Cancer Group

**Diogo Pellegrina** <sup>1,50</sup>, **Daria Kiri**<sup>2,3,50</sup>, **Kevin C. L. Cheng**<sup>1,4,50</sup>, **Francesco Favero** <sup>2,3</sup>, **Enrichetta Miletto**<sup>2,3</sup>,  
**Zoe P. Klein** <sup>1,5</sup>, **Barbara Hernando** <sup>6</sup>, **Andre V. Olsen**<sup>2,3</sup>, **Atef Sahli**<sup>7</sup>, **Yunfan Fu**<sup>8,9</sup>, **Pavlo Lutsik**<sup>10,11</sup>,  
**Weerachai Jaratlerdsiri** <sup>12</sup>, **William J. Hutchison** <sup>8,9</sup>, **G. Steven Bova**<sup>13</sup>, **Daniel S. Brewer** <sup>14,15</sup>, **Mark Brook** <sup>16</sup>,  
**Benedikt Brors** <sup>17,18,19</sup>, **Adam Butler**<sup>20</sup>, **Géraldine Cancel-Tassin** <sup>21,22</sup>, **Niall M. Corcoran**<sup>23,24,25</sup>, **Olivier Cussenot**<sup>21</sup>,  
**Abraham Gihawi** <sup>12,26</sup>, **Etsehiwot G. Girma**<sup>2,3</sup>, **Vincent J. Gnanapragasam** <sup>27</sup>, **Anis A. Hamid** <sup>24</sup>,  
**Housheng Hansen He** <sup>28</sup>, **Eddie Luidy L. Imada**<sup>29</sup>, **G. Maria Jakobsdottir** <sup>30,31</sup>, **Chol-hee Jung** <sup>32</sup>, **Francesca Khani** <sup>29</sup>,  
**Zsafia Kote-Jarai** <sup>16,33</sup>, **Philippe Lamy** <sup>34,35</sup>, **Gregory Leeman** <sup>20</sup>, **Massimo Loda** <sup>29,36</sup>, **Ramyar Molania**<sup>8,9</sup>,  
**Lucio R. Queiroz** <sup>29</sup>, **Tobias Rausch** <sup>37</sup>, **Brian Robinson** <sup>29</sup>, **Karina D. Sørensen** <sup>34,35</sup>, **Sebastian Uhrig** <sup>17</sup>,  
**Yaobo Xu** <sup>20</sup>, **Takafumi N. Yamaguchi** <sup>38</sup>, **Claudio Zanettini** <sup>29</sup>, **Stefano Mangiola**<sup>39</sup>, **Lena Voith von Voithenberg** <sup>10</sup>,  
**Luigi Marchionni** <sup>29</sup>, **Anthony T. Papenfuss** <sup>8,9</sup>, **Bernard Pope** <sup>24,40,41</sup>, **Vanessa M. Hayes** <sup>12,26,42,43</sup>,  
**Thorsten Schlomm** <sup>44</sup>, **Robert G. Bristow** <sup>45,46,47</sup>, **Christopher M. Hovens** <sup>48</sup>, **Ros A. Eeles** <sup>16,33</sup>, **Colin S. Cooper**<sup>26</sup>,  
**Clarissa Gerhauser** <sup>10</sup>, **Geoff Macintyre** <sup>6</sup>, **Andreas J. Gruber** <sup>49</sup>, **David C. Wedge** <sup>7</sup>, **Joachim Weischenfeldt** <sup>2,3,44</sup>   
& **Jüri Reimand** <sup>1,4,5</sup> 

A list of members and their affiliations appears in the Supplementary Information.