Article

Integrated signatures define mutational processes in prostate cancer

https://doi.org/10.1038/s41586-026-10468-w

Received: 26 March 2025

Accepted: 26 March 2026

Published online: 09 September 2026

Open access

Check for updates

Andreas J. Gruber[1,74], André V. Olsen[2,3,74], Barbara Hernando[4,74], Kevin C. L. Cheng[5,6,74], Clarissa Gerhäuser[7], Marina Torres[4], Francesco Favero[2,3], Daria Kiriy[2,3], Ángel Fernández-Sanromán[4], Juan Maria Roldan-Romero[4], Lucy Barton[8], Diogo Pellegrina[5,6], G. Steven Bova[9], Daniel S. Brewer[10,11], Mark N. Brook[12], Benedikt Brors[13,14,15], Adam Butler[16], Géraldine Cancel-Tassin[17,18], Niall M. Corcoran[19,20,21,22,23,24], Olivier Cussenot[17], Abraham Gihawi[10,25], Etsehiwot G. Girma[2,3], Vincent J. Gnanapragasam[26,27,28,29], Anis A. Hamid[20,30], Vanessa M. Hayes[8,10,25,31], Housheng Hansen He[6,32], Christopher M. Hovens[19,20,22,23,33], Eddie L. Imada[34], G. Maria Jakobsdottir[35,36], Chol-Hee Jung[37], Francesca Khani[34,38], Zsofia Kote-Jarai[12], Philippe Lamy[39,40], Gregory Leeman[16], Massimo Loda[34,41,42], Pavlo Lutsik[7,43], Luigi Marchionni[34], Ramyar Molania[23,44], Anthony T. Papenfuss[23,44], Bernard Pope[20,37,45,46], Lucio R. Queiroz[34], Tobias Rausch[47], Brian Robinson[34,38], Atef Sahli[8], Karina D. Sørensen[39,40], Takafumi N. Yamaguchi[48], Sebastian Uhrig[13], Yaobo Xu[16], Claudio Zanettini[34], Pan Prostate Cancer Group (PPCG)*, Ronald Simon[49], Guido Sauter[49], Ros A. Eeles[12,50], Colin S. Cooper[10], Robert G. Bristow[8,35,51], David C. Wedge[8,36], Thorsten Schlomm[52], Geoff Macintyre[4,75] ✉, Jüri Reimand[5,6,53,75] ✉ & Joachim Weischenfeldt[2,3,52,75] ✉

Prostate cancer follows a long and heterogeneous disease course with incompletely understood aetiology[1]. Here we dissect the mutational processes shaping the genomes of 959 donors from the Pan Prostate Cancer Group and assess their clinical relevance. By integrating de novo extracted single-base substitution, insertion–deletion and copy-number signatures with six novel complex structural variant signatures, we identify eight integrated mutational footprints (IMFs) that collectively explain the mutational processes in 85% of primary prostate cancer genomes. IMFs were strongly influenced by regional biases in the genome, most prevalently androgen receptor-mediated mutagenesis and replication stress. Four IMFs, present in 37% of primary tumours, were significantly associated with shorter time to metastasis. These included reactive oxygen-species-driven mutagenesis and both canonical and non-canonical homologous recombination deficiency, the latter being enriched in patients of African ancestry. Extending to the metastatic setting, we found that IMFs predicted sensitivity to androgen receptor pathway inhibitors. Taken together, our study delineates the aetiologies and mutational processes that drive the genomic and clinical heterogeneity of prostate cancer, introduces IMFs as a unifying framework, and highlights their potential to improve both risk stratification and biomarker-guided treatment selection.

Prostate cancer is the second most common cancer globally[2], and the fifth leading cause of cancer death among men[3]. Despite its high prevalence and substantial burden on patients and society, the aetiology underlying the disease's clinical heterogeneity, from indolent to highly lethal, remains poorly understood. Prostate cancer is influenced by family history, inherited germline mutations and patient ancestry[4]. Mutations in DNA damage repair (DDR) genes such as *BRCA1* or *BRCA2*, associated with homologous recombination deficiency (HRD), are present in approximately 2% of localized disease and 15% of metastatic castration-resistant disease[5]. Other clinically relevant but rare DDR pathways (reviewed previously[6]) include mismatch repair deficiency (MMRd), identified in 2–3% of primary prostate cancer[7,8], and *CDK12* inactivation, associated with a high burden of somatic tandem duplications[9–11]. As a hormone-driven disease, androgen signalling is pivotal, mediated by activation of the steroid-binding androgen receptor (AR) transcription factor[1]. AR activation by testosterone and dihydrotestosterone causes widespread changes in transcriptional activation of AR-target genes, also termed the AR cistrome[12,13]. Moreover, both single-nucleotide variants (SNVs) and structural variants (SVs) accumulate at AR-binding sites (ARBSs)[14], with the latter associated with chromoplexy[15], creating chains of interchromosomal complex SVs (cSVs).

A list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: gmacintyre@cnio.es; juri.reimand@utoronto.ca; joachim.weischenfeldt@bric.ku.dk

Beyond these specific contributions, the broader landscape of oncogenic mutational processes and their underlying aetiologies remains largely undefined. To characterize mutational processes, mathematical methods are commonly used that identify mutational signatures, or footprints, that correspond to each specific mutational process. Methods have been developed to identify signatures based on SNVs and insertion–deletions (indels)[16] and, more recently, copy-number alterations (CNAs)[17] and cSVs[18]. Previous studies have mainly investigated these mutational signatures separately, identifying a few common mutational processes in prostate cancer including clock-like mutagenesis (SBS1 and SBS5), HRD (SBS3 and ID6), MMRd (SBS15, SBS21 and SBS44)[16,19–21] and complex genomic configurations, mainly chromoplexy (approximately 20%)[15,18] and chromothripsis (20–40%)[22,23]. Despite these important findings, a fundamental appreciation and understanding of the aetiologies and mutational processes driving prostate cancer remains unexplored.

The Pan Prostate Cancer Group (PPCG) has assembled a clinically well-annotated and diverse collection of whole-genome sequencing (WGS) data from fresh-frozen tumour specimens from 1,001 donors. The WGS data have been reanalysed with standardized pipelines annotating high-accuracy germline and somatic variant calls and driver mutations[24]. Here we present a comprehensive and integrative mutational signature analysis to describe the mutational processes that influence and shape the clinical outcome of primary prostate cancer.

Mutational signatures of prostate cancer

The PPCG cohort, comprising 1,209 prostate samples from 1,001 prostate cancer donors, provides the most comprehensive resource to date to investigate the footprints of somatic mutational processes in prostate cancer genomes (Fig. 1). Here we examined one representative primary tumour sample ($n = 959$) from each PPCG donor (Extended Data Fig. 1), with ages ranging from 32 to 88 years (mean, 60 years), and a median follow-up of 7 years. Of the 903 patients with follow-up information, 132 (14.6%) developed metastatic disease (median time to metastasis, 6 years; 24 patients had no available metastatic follow-up information)[24].

The genomic landscape comprised 4,678,132 somatic SNVs (range, 51–292,073), 291,531 indels (range, 1–62,203), 74,386 SVs (range, 0–876) and 55,279 CNAs (range, 0–1,023) (Fig. 1). We used a curated set of germline and somatic driver mutations to annotate patient tumour samples with mutations in known DDR genes, including those involved in homologous recombination (HR), base excision repair (BER), nucleotide excision repair (NER), non-homologous end-joining (NHEJ) and mismatch repair (MMR), tallying 21% of the cohort in total, in agreement with previous estimates[25,26].

To shed light on the mutational processes affecting prostate cancer genomes, we extracted four types of signatures de novo: single base substitution (SBS), indel (ID), cSV and copy-number (CX) signatures. In total, we extracted 22 SBS, 10 ID, 8 CX and 6 cSV signatures (Fig. 2a,b and Extended Data Figs. 2 and 3). Two of the extracted SBS signatures are considered artefactual according to the COSMIC signature annotations (SBS52 and SBS54) and were excluded from downstream analyses. Of the 899 samples with available copy-number profiles, we identified 321 (37.4%) as having CNA levels below the threshold required to identify a discernible pattern of chromosomal instability[17] (CIN) (<20 CNAs; Supplementary Methods), and were therefore classified as copy-number stable (hereafter, non-CIN).

Clock-like mutational processes, represented by SBS1, SBS5 and SBS40, were the most prevalent across the cohort as expected[16,27] (Extended Data Fig. 3c). Other established mutational processes detected included MMRd (2%), linked with the signatures SBS15, SBS21 and SBS44 (Fig. 2c) and *MSH2* mutations (Fig. 2d); reactive oxygen species (ROS) and BER-related mutagenesis (13.1%), represented by the SBS18 signature; and APOBEC mutagenesis (1.44%), caused by cytidine deaminase enzymes that are part of a viral defence mechanism, and associated with the SBS2 and SBS13 signatures[16] (Fig. 2c). While the prevalence of APOBEC mutagenesis is consistent with previous studies[16], we note that higher prevalences have been reported in other studies[28] (Supplementary Note 1).

We also found evidence for two additional SBS and ID signatures with limited previous evidence: ID83C, a signature with single-base deletions in homopolymers detected in 1.3% of samples; and SBS96D, which was detected in 3.9% of samples (Extended Data Fig. 3, further details are provided in Supplementary Note 2).

Motivated by the diverse SV patterns and associated genomic features found in prostate cancer (Supplementary Note 3 and Supplementary Figs. 1–5), we constructed prostate-specific cSV signatures de novo. In brief, we performed unsupervised clustering using multivariate analysis to identify cSV signatures based on a recent pan-cancer study approach[18]. In total, we identified six stable cSV signatures (Fig. 2a,b and Supplementary Table 1), which we tested for association with genomic features to assign putative aetiologies. This included somatic mutations in driver genes, deficiency in five DDR pathways, whole-chromosome CNAs, focal amplifications, WGD, tandem duplications, chromothripsis, chromoplexy and the unique patterns defining each individual signature. We also computed pairwise correlations between the signature types to determine evidence of convergence on an aetiology (Fig. 2c–e and Extended Data Fig. 4).

cSV signature 1 (cSV1) encoded a characteristic pattern of deletions in late-replicating regions, chromothripsis, inversions and local 2-jumps, suggesting replication stress as the putative aetiology (Fig. 2a). Moreover, cSV1 was associated with somatic alterations of *MYC* and *SPOP*, and with high activity levels of CX8 and CX11 (Fig. 2c–e), signatures linked to mid- and high-level amplifications through replication stress[17]. By contrast, cSV2 had characteristic patterns linked with SVs near ARBS and with chromoplexy formation[15,18], suggesting AR-driven transcriptional activity as the aetiology. cSV3 (unknown aetiology) was also associated with chromoplexy but characterized by deletions in early-replicating regions. cSV4 and cSV5 stratified patients by different sizes of tandem duplications and tandem duplicator phenotype (TDP)[29], and cSV4 was concurrently observed with biallelic inactivation of *CDK12* (Fig. 2d). Lastly, there was compelling evidence supporting HRD in tumours with cSV6, a complex signature characterized by a composite footprint of unbalanced translocations and foldback inversion. cSV6 was significantly correlated with the HRD-related mutational signatures SBS3, ID6 and CX3 (Fig. 2c), with deficiency in the HR pathway through biallelic inactivation of canonical genes (Fig. 2e), including *BRCA2* mutations (Fig. 2d). As expected, *BRCA2* mutations were also correlated with elevated activity of other well-established HRD signatures (SBS3, ID6, ID8 and CX3; Fig. 2d).

In summary, we identified the full spectrum of mutational signatures, including six novel cSV signatures, supporting the presence of several mutational processes in prostate cancer genomes.

Topographical biases of somatic mutations

Somatic alterations are not uniformly distributed across the genome but exhibit topographical biases—often in a tissue-specific manner[30,31]. To investigate the topographical biases of mutational processes in prostate cancer, we integrated our comprehensive set of signatures with a diverse range of common and prostate-specific genomic features.

We first tested the proximity of simple and complex SVs to SBS signatures, which showed a strong and distance-dependent co-occurrence of SVs with SNVs linked with APOBEC activity (false-discovery rate (FDR)-corrected $P = 1.9 \times 10^{-13}$, Kolmogorov–Smirnov test; Fig. 3a and Supplementary Fig. 6a) in the subset of patients with APOBEC activity (SBS2/SBS13, $n = 13$), as expected[19,32]. Across the different types of cSV, we identified an enrichment of chromothripsis breakpoints near APOBEC-associated SNVs (FDR-corrected $P = 3.03 \times 10^{-3}$,

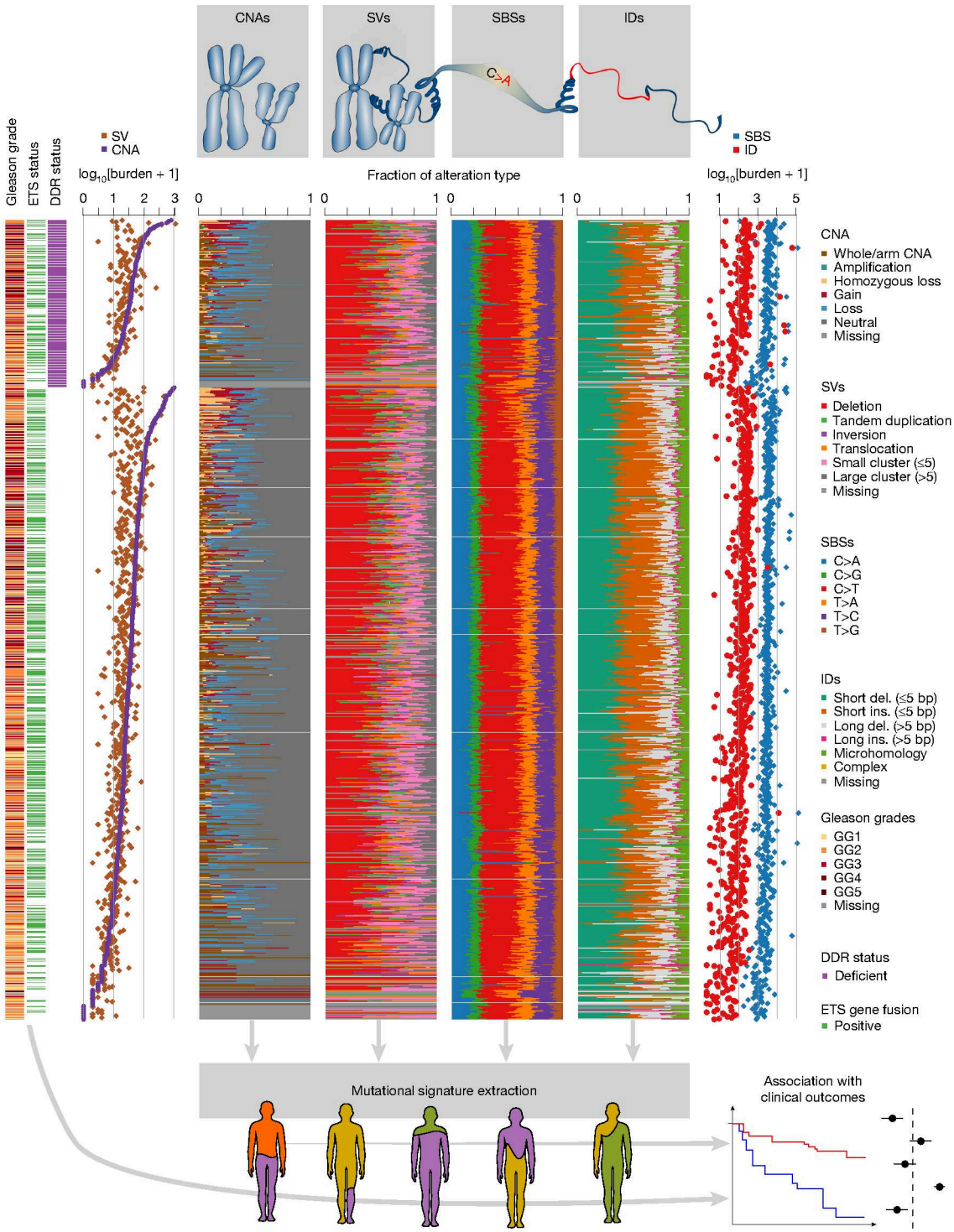


Fig. 1 | Overview of the WGS cohort. From left to right, the per-sample Gleason grade group, ETS status, deficiency in DDR mechanisms, CNA and SV burden, the percentages of CNAs, SVs, SBS and ID types, and the SBS and ID burden for 899 whole primary prostate cancer genomes (Extended Data Fig. 1). Bottom, outline of the approach to identify the major oncogenic mutational processes by integrating mutational signature analysis with driver gene status, mutational profiles and clinical information.

Kolmogorov–Smirnov test; Fig. 3b and Supplementary Fig. 6b,c), suggesting APOBEC-mediated editing at chromothriptic breakpoints. In support, a recent study used an *APOBEC3B*-knockout cell line to demonstrate that such APOBEC mutagenesis at chromothriptic breakpoints depends on the cytidine deaminase APOBEC3B[33]. Next, we examined the impact of replication and transcription. Highly transcribed genes (fourth quartile expression) and early-replicating regions were strongly enriched for both simple and complex SVs, including templated insertions, chromoplexy and tandem duplications (Fig. 3c), probably arising through fork-stalling and template switching or transcription–replication collision[18]. Finally, we examined AR binding to DNA, another well-known mutagenic process that can cause SVs in prostate cancers[14]. We found chromoplexy events to be specifically and uniquely enriched at tumour-specific ARBSs (TARBSs)[12] (Fig. 3c). By contrast, chromothripsis, was highly enriched at late-replicating regions, which are on average gene poor and transcriptionally inactive.

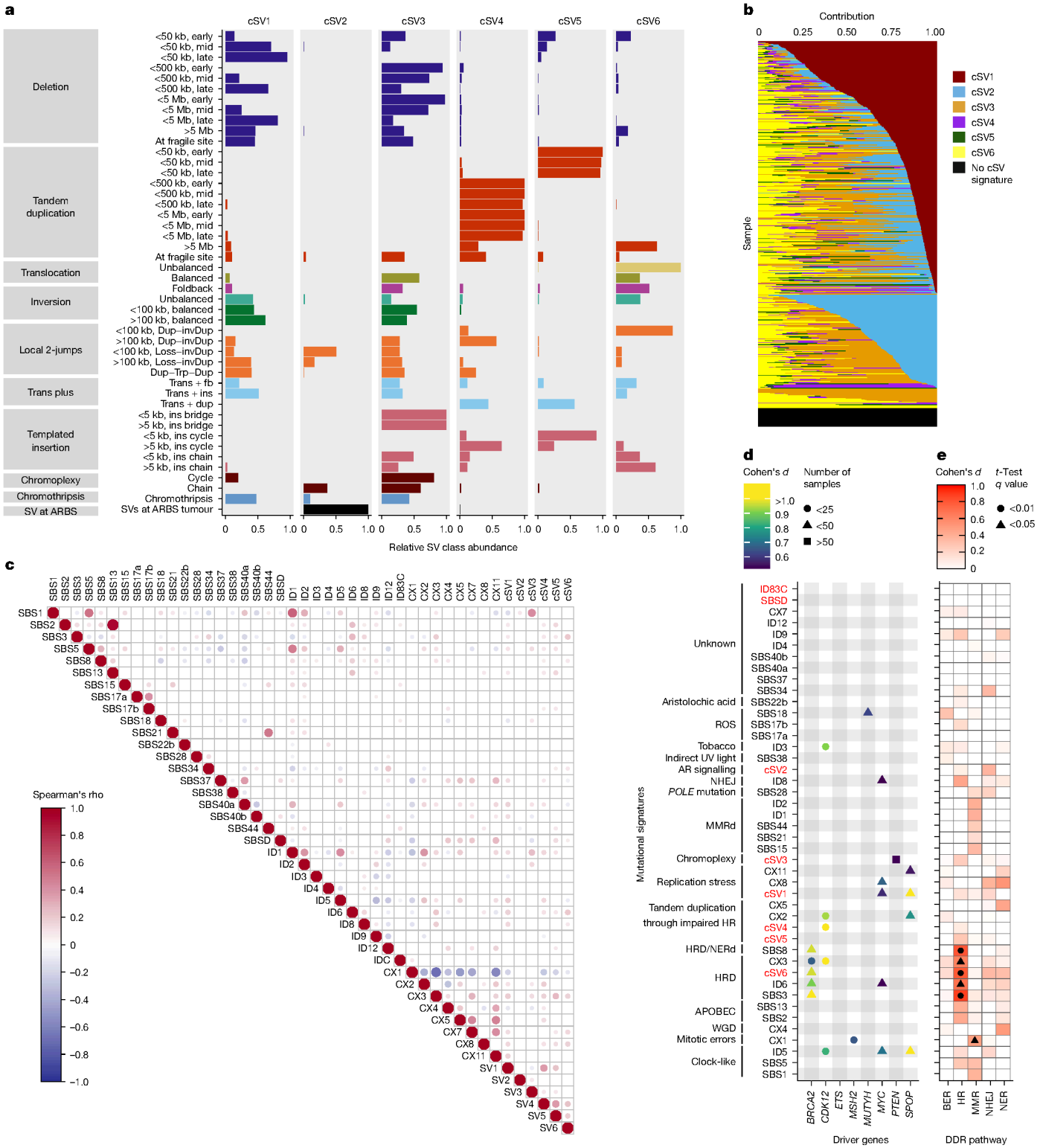


Fig. 2 | See next page for caption.

The mutagenicity of AR binding was not restricted to SVs. In agreement with previous studies[14,34], we found enrichment of SNVs at ARBSs (Supplementary Note 4 and Supplementary Figs. 7–11). This occurred primarily in tumours exceeding a minimal AR activity threshold, suggesting additional factors contribute beyond this level. SNVs at ARBS may potentially disrupt AR binding and therefore contribute to cancer-promoting gene deregulation[14]. We next examined whether AR binding was linked with specific mutational signatures at ARBS, distinguishing between those enriched in tumours (TARBS) and those enriched in normal prostate (NARBS)[34]. We found several lines of evidence for AR binding associated with an increase in clock-like mutational signatures in tumours relative to their genome-wide activity. These included SBS5, SBS40a, ID5, SBS3 and SBS8 (fold change (FC) = 1.42, 1.47, 1.35, 1.40, 1.34; FDR-corrected $P = 4.5 \times 10^{-78}$, 2.8×10^{-69}, 9.8×10^{-5}, 3.8×10^{-22}, 1.4×10^{-19}, respectively, RM2 method[35]; Fig. 3d and Supplementary Note 4). Notably, while two of the clock-like mutational

Fig. 2 | cSV signatures in prostate cancer. a, Six cSV signatures. Each column (cSV1–6) represents extracted cSV signatures. The rows represent individual SV classes, with the major simple and cSVs (left-most column) and subtypes of SVs (second column), including duplications (dup), inversions (inv), triplications (trp), translocations (trans), insertions (ins) and fold-backs (fb), separated by size, occurrences at replication timing region (early, mid, late), fragile sites and TARBS. The horizontal coloured bars represent feature-normalized SV exposure for each cSV signature. **b**, cSV signature contribution for each sample in the cohort. The vertical and horizontal axes represent samples and relative contribution of signatures, respectively. Colours represent separate cSV signatures. **c**, Correlation analysis of scaled SBS, ID, CX and cSV signatures across samples (Spearman's rho). The dot sizes denote the correlation strength. Only significant ($P < 0.05$) correlations are indicated; non-significant results are blank. **d**,**e**, The heat map shows significant positive associations between signatures and genetic alterations in prostate cancer driver genes (**d**) and signatures and deficiency in DDR pathways (**e**). Mutational signatures are ordered by putative aetiology and novel signatures are highlighted in red. For **d**, a two-sided Welch's *t*-test was applied to compare means across groups. The dot colours denote the strength of differences (Cohen's *d*) in signature activities of tumour samples with and without the mutated gene. The dot shapes denote the number of samples with the mutated gene. For **e**, the samples were classified as deficient based on alterations in canonical genes in the pathway. A two-sided Welch's *t*-test was applied to compare means across groups. The dot colours denote the strength of differences (Cohen's *d*) in signature activities of deficient and proficient PPCG samples for a given DDR pathway. Dot shapes denote the level of significance. $n = 869$ (**a** and **b**), $n = 899$ (**c**) and $n = 959$ (**d** and **e**) tumour samples.

signatures were enriched at TARBS (SBS5 and SBS40a), we found a depletion of the clock-like signature SBS1 compared with both NARBS and genome background (Fig. 3e). SBS1 is associated with spontaneous deamination of 5-methylcytosine (5mC) at CpG dinucleotides leading to C to T transitions[36], and regulatory regions are enriched for CpG sites, where 5mC is associated with transcriptional repression[37]. We hypothesized that CpG methylation differences at ARBS could underpin the observed depletion of SBS1 signature. To this end, we performed Oxford Nanopore long-read sequencing of eight tumours in our cohort to jointly identify DNA nucleotide context and methylation changes, which showed an enrichment of unmethylated CpGs at TARBS compared to NARBS (average FC = 2.07, $P = 0.039$, Wilcoxon signed-rank test; Fig. 3f). This suggests that the AR mutagenicity at enhancers and promoters is conditioned on existing CpG methylation levels. In agreement, AR binding depends on the pioneering transcription factor FOXA1[38], which first binds to specific enhancers and promoters and recruits the demethylase TET1, therefore allowing AR to bind to these ectopic locations[39]. This provides a conceptual explanation for the observed depletion of SBS1 mutations at TARBS, probably due to open active chromatin and unmethylated CpG sites at ARBS in tumours (Fig. 3g).

We next investigated the influence of replication timing, derived from prostate cancer cell lines[40], on both mutational signatures and AR mutagenicity. We observed SNVs in late replicating regions to be 1.73-fold enriched ($P = 4.9 \times 10^{-323}$, Poisson test) compared with early-replicating regions, consistent with previous studies[41]. Most mutational signatures displayed similar enrichment in late replicating regions, with notable enrichment of the ROS-associated SBS18 in late replicating regions (median, 4.33-fold enrichment, $P = 4.9 \times 10^{-324}$, Poisson test; Extended Data Fig. 5a). By contrast, we found a relative enrichment of TARBS, and to a lesser extent NARBS, in early-replicating regions (Fig. 3h and Supplementary Fig. 10), compatible with the association with hypomethylated open active chromatin regions, generally found enriched in early-replicating regions[42]. We observed a near-uniform distribution of mutations across replication timing zones at TARBS with strong depletion of SBS1, which we propose is due to the epigenetic landscape of prostate cancer, with TARBS found preferentially at hypomethylated open active chromatin.

In summary, we found a broad impact of replication and epigenetic contexts in defining the topographical biases of both cSVs and SNVs, with AR-mediated somatic alterations strongly linked with early replication and active transcription and relative depletion of the ubiquitous SBS1 signature.

Diverse signatures converge on eight IMFs

Mutational processes can cause highly complex footprints in the genome, simultaneously involving signatures of SNVs, indels, SVs and CNAs. We therefore reasoned that combining all four levels of mutational signature (SBS, ID, CX and cSV) would better characterize the major mutational processes operating on prostate cancer genomes. For example, while processes such as chromothripsis and cytosine deamination can be potentially identified with a single signature type, robust identification of most mutational processes depends on multiple complementary mutational signatures, for example, HRD (SBS, ID, CX and cSV) and MMRd (multiple SBS and ID signatures).

To this end, we performed hierarchical clustering of the four levels of mutational signature from 521 samples of patients with CIN with activities quantified for all four mutational signature types (Extended Data Fig. 1), identifying eight clusters, which we term IMFs (Fig. 4a,b, Supplementary Figs. 12–13 and Extended Data Fig. 6). These IMFs represent the key mutational processes operating in prostate cancer. We quantified IMF activities within each tumour using linear combination decomposition. Using IMF activities, we pursued multiple association analyses to characterize the IMF aetiologies, including mutated genes (Fig. 4c), individual mutational signatures, specific chromosomal alterations, DDR pathway deficiencies, HRDetect[43] (HR deficiency score) and MSIsensor-pro[44] (MMR deficiency score), as well as tumour mutational burden (TMB), whole-genome duplication (WGD), cell cycle proliferation, AR pathway activity and hypoxia scores (Extended Data Fig. 7 and Supplementary Fig. 14).

IMF1 (MMRd) activity correlated with MMRd-related mutational signatures SBS15, SBS21 and ID2 (Spearman's rho = 0.14, 0.13 and 0.69, FDR-corrected $P = 0.006$, 0.01 and 1.45×10^{-72}, respectively; Extended Data Fig. 7a). IMF1 had higher activity in cases with microsatellite instability (Cohen's $d = 1.53$), mutations in *MSH2* (Cohen's $d = 1.32$) and deficiency in the MMR pathway (Cohen's $d = 0.65$; Extended Data Fig. 7d). Together, this supported an aetiology of MMRd for IMF1. Tumours with high IMF1 exhibited a near-uniform mutation rate across the genome (Extended Data Fig. 5b and Supplementary Fig. 8b).

IMF2 (ROS mutagenesis) was significantly associated with SBS18 (Spearman's rho = 0.32, FDR-corrected $P = 1.56 \times 10^{-12}$; Extended Data Fig. 7a), a signature linked with faulty BER and DNA damage induced by ROS[45]. In agreement, IMF2 was anti-correlated with hypoxia score (Spearman's rho = −0.13, FDR-corrected $P = 0.051$; Extended Data Fig. 7b). IMF2 was also associated with SBS8 (Spearman's rho = 0.32, FDR-corrected $P = 1.56 \times 10^{-12}$; Extended Data Fig. 7a), a signature with unknown aetiology but speculated to involve NER[46], a repair pathway that can be used in response to ROS-based damage[47] and late-replicating errors[48]. Both SBS8 and SBS18 were strongly enriched in late replicating regions (Extended Data Fig. 5a).

IMF3 (AR-mediated mutagenesis) displayed several marks of AR-mediated mutagenesis, including significant correlation with cSV2 and cSV3 (Spearman's rho = 0.34 and 0.48, FDR-corrected $P = 4.02 \times 10^{-15}$ and 1.75×10^{-30}; Extended Data Fig. 7a), and elevated mutation rate at TARBS (FDR-corrected $P = 1.61 \times 10^{-10}$, RM2 negative binomial test[35]; Supplementary Fig. 8b). IMF3 activities were higher in cases with *ETS* fusions (Cohen's $d = 0.75$, FDR-corrected $P = 1.65 \times 10^{-15}$, Welch's *t*-test; Fig. 4c), with *PTEN* mutations (Cohen's $d = 0.55$, FDR-corrected $P = 1.93 \times 10^{-5}$, Welch's *t*-test) and with mutations in one of two key AR regulators *TBL1XL1*[49] or *ZBTB16*[50] (Cohen's $d = 1.67$, 1.17; FDR-corrected $P = 0.02$,

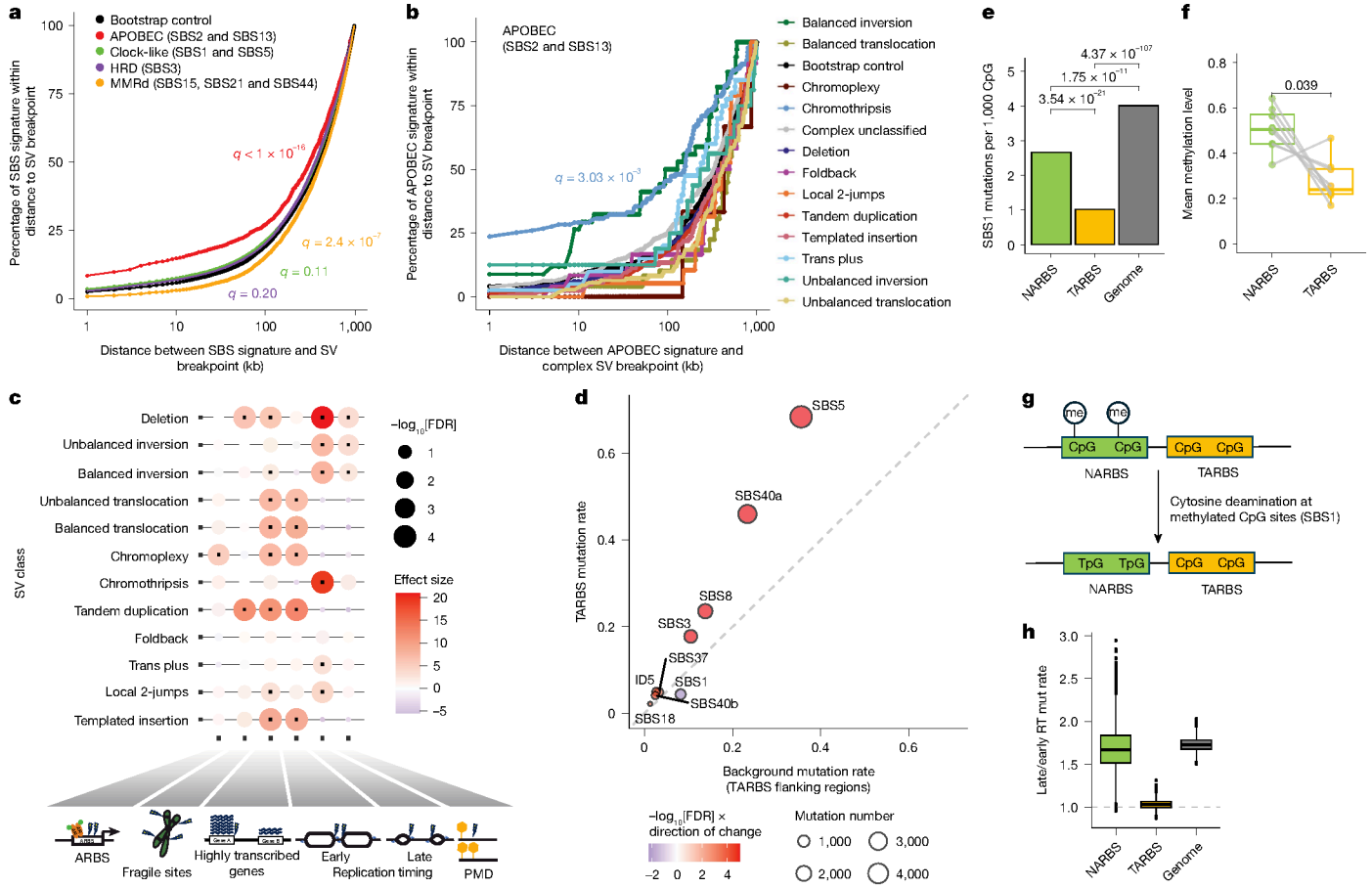


Fig. 3 | Regional biases of mutational processes acting in prostate cancers. **a**,**b**, SV breakpoint to SBS distance was measured as the percentage of SBS signatures (vertical axis) with an SV breakpoint within a certain genomic distance in kb (horizontal axis). Distance was compared to a bootstrap control (n = 1,000) using a two-sided Kolmogorov–Smirnov test; FDR-corrected P values (q) are shown. **a**, Distance analysis of SBS groups and the different SV classes. SBS signatures are grouped into known aetiologies. **b**, The distance analysis of APOBEC signatures SBS2 and SBS13 and the different SV classes (Supplementary Fig. 6). **c**, Enrichment of simple and complex SVs at six different genomic features (permutation test, n = 10,000). The vertical axis represents SV classes and the horizontal axis represents genomic features, including partially methylated domains (PMDs). The significance level (circle size), FDR-corrected significance (black dot) and effect size (circle colour) were assessed by permutation-based estimation using z scores. **d**, The mutation rate of SBS/ID signatures found to be significantly different between TARBS and flanks. The numbers represent signature-assigned TARBS and flank mutations (dot size). Statistical significance (colour) was calculated using a negative binomial test with RM2[35]; FDR-corrected P values are shown. **e**, SBS1 mutation rate in TARBS, NARBS and genome-wide, calculated as the number of SBS1 mutations per 1,000 CpG dinucleotides in the respective regions (two-sided Poisson tests, comparing the SBS1 frequency between regions). **f**, Methylation levels derived from Oxford Nanopore sequencing, showing the means of averaged ARBS methylation per case (dot size). The grey lines connect the same cases. Two-sided Wilcoxon signed-rank test. **g**, Schematic of the putative mechanism causing SBS1 depletion in TARBS compared with NARBS. **h**, The late to early replication timing (RT) rate ratio in genomic regions (bootstrapping n = 10,000). Mut, mutation. For **f** and **h**, the box plots indicate the first and third quartile (box limits) and the median (centre line), and the whiskers indicate the first or third quartile ±1.5× the interquartile range. n = 869 (**a**–**c**), n = 959 (**d**, **e** and **h**) and n = 8 (**f**) tumour samples.

0.002; Welch's t-test). We also found correlations with clock-like SBS1 and SBS40a signatures (Spearman's rho = 0.51 and 0.50, FDR-corrected $P = 3.21 \times 10^{-16}$, 1.92×10^{-9}; Extended Data Fig. 7a).

IMF4 (mitotic defects and replication stress) activity was significantly higher in cases with mutations in genes involved in the mitotic machinery, including *CDC16*[51] (homozygous loss) and *MZT2A/B*[52] and *TUBA3D/E*[53] (homozygous losses, located in close proximity) (FDR-corrected $P < 0.05$, Welch's t-test; Fig. 4c). Moreover, IMF4 displayed CNAs linked to replication stress with CX4 (associated with WGD), CX5 (impaired HR plus replication stress) and CX11 (mid-level amplifications through replication stress)[17], suggesting faulty replicated DNA after replication stress that persists into mitotic defects[54]. IMF4 also correlated with SBS37 (Spearman's rho = 0.35, FDR-corrected $P = 7.07 \times 10^{-16}$), a signature with unknown aetiology that accumulates in late-replicating regions[16] (Extended Data Fig. 7a), and showed the highest late-to-early mutation rate ratio among all mutational footprints (Extended Data Fig. 5b).

IMF5 (non-canonical impaired HR) had significantly higher activity in tumours with mutations in *CDK12* (Cohen's d = 1.87, P = 0.02, Welch's t-test; Fig. 4c), which has been shown to phenocopy impaired HR in the absence of mutations in *BRCA1/2*[55]. This suggests an aetiology of impaired HR, which we term non-canonical to differentiate from canonical impaired HR, which typically involves enrichment in *BRCA1/2*-mutant cases and a different pattern of DNA damage. In support, IMF5 activities were significantly correlated with cSV4, characterized by tandem duplications, and CX2, linked to impaired HR (Extended Data Fig. 7a). Notably, this mutational process was enriched in patients of African ancestry ($P = 5.36 \times 10^{-5}$, n = 32, Fisher's test; Supplementary Fig. 15), consistent with a previous study finding more tandem duplications in this group of patients[56]. We also found IMF5 to associate with increased hypoxia and TMB, and decreased AR pathway activity (Extended Data Fig. 7).

IMF6 (replication stress) showed significant correlation with cSV1 (Spearman's rho = 0.78, FDR-corrected $P = 2.39 \times 10^{-104}$; Extended Data

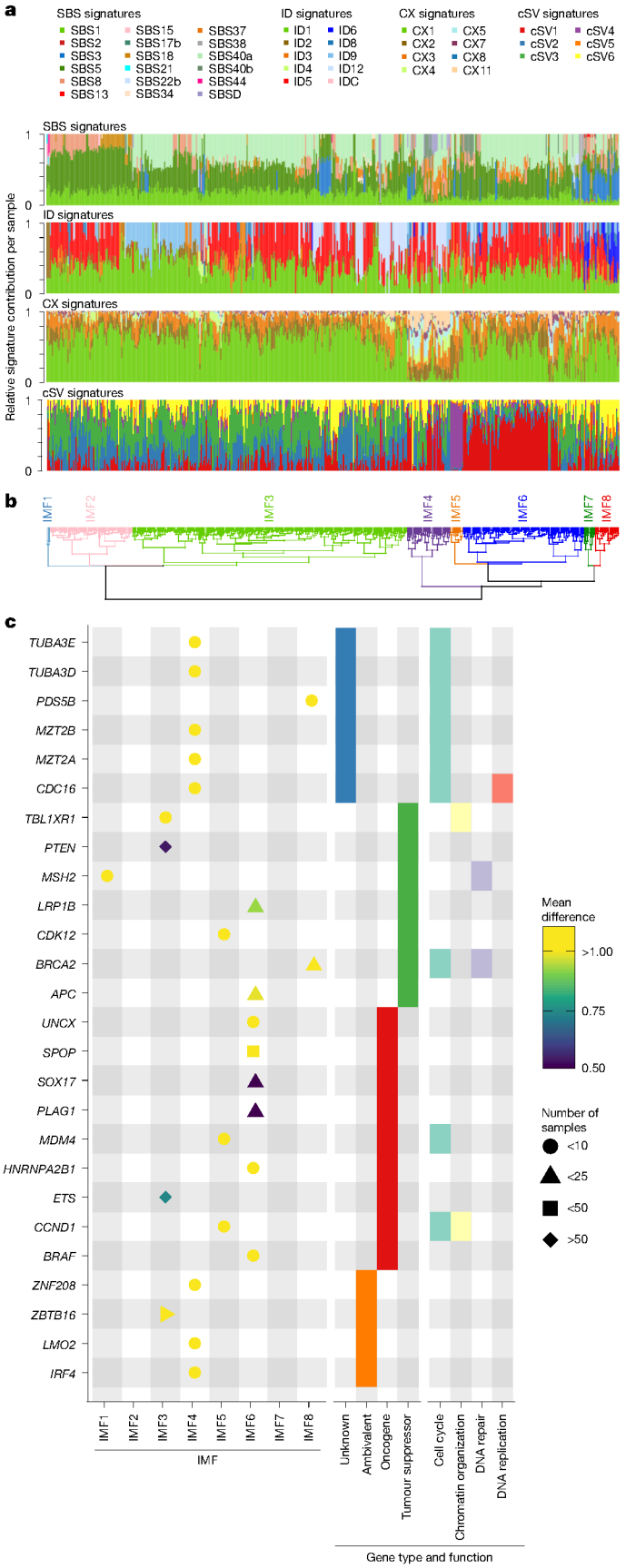


Fig. 4 | Mutational and chromosomal alteration signatures converge on eight integrated mutational footprints in prostate cancers. a, The activities of the four mutational signature types across 521 prostate cancer samples for SBS, ID, CX and cSV signatures. Only samples with more than 100 SNVs, 10 indels, 5 SVs and 20 CNAs were included (Extended Data Fig. 1). **b**, Hierarchical clustering of samples based on signature activities presented in **a**. **c**, Significant positive associations between IMF and cancer gene alterations. A two-sided Welch's *t*-test (FDR corrected) was applied to compare means across groups. The dot colours denote the strength of differences (Cohen's *d*) in IMF activities of tumour samples with and without alterations in the given gene. Dot shapes denote the number of mutated samples per gene. Gene annotations are summarized on the right. Associations with Cohen's $d < 0.5$ are not displayed.

Fig. 7a), a signature linked with chromothripsis and genomic instability in late-replicating regions, proliferation rate and with AR pathway activity and hypoxia (Spearman's correlation, FDR-corrected $P < 0.05$; Extended Data Fig. 7b). Moreover, IMF6 exhibited increased activity in cases with mutations in *SPOP* and WNT-pathway genes *APC* and *SOX17* (FDR-corrected $P < 0.05$, Welch's *t*-test; Fig. 4c). SPOP is a ubiquitin ligase, involved in preventing replication stress and genomic instability, and *SPOP*-mutant tumours are associated with dysregulation of PI3K, AR and WNT signalling[57].

IMF7 (APOBEC mutagenesis) was correlated with APOBEC-associated signatures SBS2 and SBS13, and with impaired HR-related mutational signatures SBS3 and ID6 (Spearman's correlation, FDR-corrected $P < 0.05$; Extended Data Fig. 7a), but showed no elevated activity in *BRCA2*-mutant cases. We examined whether these two distinct mutational processes tended to occur simultaneously or separately in the same tumour samples. To explore this, we evaluated the clonality of mutations assigned to SBS2/13 or SBS3/ID6 to test which signatures may have occurred preferentially early in tumour evolution (clonal) or later (subclonal). We found that APOBEC-related mutations often occurred clonally, while impaired HR mutations appeared subclonally (Supplementary Fig. 16), suggesting that APOBEC mutagenesis was the main aetiology and that SBS3/ID6 represented separate, subsequent processes.

IMF8 (impaired HR) was correlated with mutational signatures SBS3, ID6, cSV6 and CX3 related to impaired HR (Spearman's FDR-corrected $P < 0.05$; Extended Data Fig. 7a). IMF8 activity was also significantly higher in cases with *BRCA2* mutations, mutations in the HRD pathway and high HRdetect scores (FDR-corrected $P < 0.05$, Welch's *t*-test; Fig. 4c and Extended Data Fig. 7). Together, this evidence suggests an aetiology of canonical impaired HR.

To determine the dominant mutational process shaping each prostate cancer, we assigned samples to the IMF with the highest activity. Extended Data Table 1 summarizes the distribution of dominant mutational processes across PPCG tumours, highlighting AR, replication stress and ROS-mediated mutagenesis as the most recurrent mutational processes (39%, 28% and 21%, respectively; Extended Data Table 1).

Among the 321 non-CIN tumour samples (without CX but with available SBS and ID signatures), we reliably assigned a known mutational process to 244 of these non-CIN tumours, primarily clock-like mutational processes (Supplementary Note 5 and Supplementary Figs. 17–20). Of the remaining tumours without a detectable mutational process, 62 had low tumour purity (<20%) and low somatic variant burden. Overall, this results in 837 tumours (85%) with ascertainable signatures in our cohort (Supplementary Fig. 21).

In summary, by integrating four different types of mutational signature, we identified eight distinct mutational processes that, in combination with a subset found in non-CIN tumours, underpin 85% of primary prostate cancers in our dataset.

IMFs reflect clinical outcomes

We hypothesized that the distinct mutational processes captured by the IMFs may help explain differences in the age-at-diagnosis and clinical outcomes observed in patients with prostate cancer.

Previous studies have shown distinct mutational patterns in early-onset prostate cancer (55 years of age or younger at time of diagnosis)[19,58], but no specific mutational processes have been identified in patients with the late-onset disease. We found that IMF6 (replication stress) was predominant in late-onset tumours, whereas IMF8 (HRD) was enriched in early-onset cases (FDR-corrected $P = 3.58 \times 10^{-4}$ and $P = 0.0107$, respectively, Fisher's exact test; Extended Data Fig. 8a). Age was strongly correlated with cSV1, a key signature defining IMF6 (Extended Data Fig. 8b–d and Supplementary Figs. 22 and 23), and late-onset tumours frequently had alterations in replication-stress genes, including somatic variants in *SPOP* ($n = 39$) and *RIF1* ($n = 5$), germline variants in *RECQL4* ($n = 4$) and somatic *MYC* amplifications ($n = 40$; Extended Data Fig. 8e). These findings support the presence of a unique subset of patients with late-onset prostate cancer dominated by replication-stress processes.

To dissect the clinical heterogeneity of prostate cancer, we performed multivariate Cox proportional hazard analyses to compare metastasis-free survival (MFS) among patients according to their dominant IMF (Fig. 5a, Extended Data Table 1 and Supplementary Table 2), using patients with low chromosomal instability (non-CIN cases) as the reference group in the survival analyses given their reduced risk of disease progression compared with patients with genomically unstable tumours (Extended Data Fig. 9). Tumours primarily driven by either ROS-mediated mutagenesis (IMF2; hazard ratio = 5.26, $P = 1.4 \times 10^{-4}$, Wald test), AR-mediated mutagenesis (IMF3; hazard ratio = 4.05, $P = 0.0019$, Wald test), non-canonical impaired HR (IMF5; hazard ratio = 7.17, $P = 0.006$) or canonical impaired HR (IMF8; hazard ratio = 5.8, $P = 4.1 \times 10^{-4}$, Wald test) presented a high risk of metastasis, independent of age at diagnosis, Gleason grade group, replication timing ratio, tumour stage and TMB. MMRd (IMF1) and replication stress (IMF4 and IMF6) did not confer increased risk of metastasis compared with non-CIN cases. Notably, we found that IMFs were also able to risk-stratify patients within intermediate-risk Gleason grade groups GG2 or GG3, with GG2 tumours showing significantly increased risk with IMF8, and GG3 tumours with IMF2, IMF3 or IMF8 (Supplementary Fig. 24). Increased late-to-early replication timing ratio was associated with a decreased risk of metastasis (Fig. 5a), and with decreased biochemical recurrence risk in an external cohort (Supplementary Fig. 25). A similar Cox model analysis using individual mutational signatures did not reveal strong risk associations (Supplementary Figs. 26 and 27).

To determine whether intratumoural heterogeneity was also reflected in the dominant mutational processes, we analysed the mutational processes in an expanded series of tumour regions from four patients from PPCG with a high degree of intratumoural heterogeneity based on pathology reports (Fig. 5b and Supplementary Fig. 28). One patient (PPCG0851) had two genetically unrelated and spatially confined tumour foci with entirely distinct mutational processes. Two further patients (PPCG0056 and PPCG0440) exhibited both shared and regionally localized mutational processes, whereas all five analysed tumour samples in another patient (PPCG0058) were dominated by IMF3. Notably, we identified a tumour area (T1) defined by IMF2 in PPCG0440, a mutational process linked to ROS mutagenesis and altered oxygen metabolism.

We next sought to investigate whether the IMFs could also inform therapy selection in metastatic disease, in which the clinical need for robust predictive biomarkers is most urgent. To examine this, we used a WGS cohort of metastatic castration-resistant prostate cancers (mCRPC) from the Hartwig Medical Foundation study[59]. mCRPC is an advanced, lethal stage of prostate cancer where tumours progress despite androgen deprivation therapy (ADT). Current management relies on the sequential or combined use of AR pathway inhibitors (ARPIs) and taxane-based chemotherapies on a continuous ADT backbone. As the transition to second-line treatment often involves an empirical switch between these agents, we used this setting to evaluate the predictive power of IMFs. This context enabled us to control for first-line therapy and determine whether IMFs could independently predict response to the subsequent ARPI or taxane.

Overall, 240 patients from HMF with sufficient data were available for our analyses (Methods). Using these data, we conducted a target trial emulation design to mimic phase III randomized controlled biomarker trials[60], with patients stratified by median IMF activity. Patients were randomized in silico to an ARPI- or taxane-based regimen, with time to treatment failure (TTF) as the primary end point. Power analyses supported testing of IMF1, IMF4, IMF5 and IMF6 (Supplementary Tables 3 and 4), with IMF6 subsequently showing predictive capacity (Fig. 5c and Supplementary Table 5). In patients predicted to be sensitive (IMF6 > median), ARPI treatment was associated with a significantly reduced risk of treatment failure compared to taxane (hazard ratio = 0.21, 95% confidence interval (CI) = 0.05–0.79, $P = 0.021$, Wald test), whereas no significant difference was observed among patients predicted to be resistant.

In summary, IMFs not only stratify clinical outcomes but also show potential to predict therapeutic response in the metastatic setting, supporting their relevance for patient stratification and disease management. As our trial emulation analyses are retrospective and observational, with limited subgroup sizes, these findings should be viewed as hypothesis generating and warrant prospective biomarker-driven validation.

Discussion

The complete landscape of mutational processes in prostate cancer and how they contribute to clinical outcomes has remained unclear. Efforts have focused on characterizing and risk-stratifying patients with HRD owing to the increased risk and therapeutic potential. However, fewer than 3% of primary prostate cancers exhibit a deficiency in this repair mechanism[5]. Here we performed a large and comprehensive examination of the mutational processes shaping the genomes of 961 primary prostate cancers with diverse histopathological stages and clinical outcomes. We leveraged all genomic variants captured by WGS to dissect the genomic footprints caused by different mutational processes within the tumour. Through integrated analyses of multiple signature modalities including six novel cSV signatures, we identified eight mutational processes in prostate cancer that, together with clock-like mutational processes in non-CIN tumours, were able to explain the mutational processes underpinning genomic alterations in 85% of primary prostate cancers.

A key driver in prostate cancer is AR, the activity of which is mutagenic[15,58]. We found that mutations in more than a third of prostate cancer genomes with CIN were largely attributed to IMF3 (Extended Data Table 1), linked with AR-mediated mutagenesis. We observed substantial biases in the regional distribution of both SNVs and SVs, and we provide evidence that this is mainly driven by differences in replication timing, DNA methylation and AR binding, except in tumours with MMRd, which showed a nearly uniform distribution of SNVs, consistent with replication-coupled MMR with higher efficiency in early-replicating regions[61]. Signatures of MMRd, associated with replicative slippage and formation of microsatellite instability were detected with a prevalence similar to previous studies and with a dominance of mutations in *MSH2*[8]. Although IMF1 (MMRd) tumours exhibited CIN, we also found signatures of MMRd in chromosomally stable tumours.

Our study also identified a pervasive role for replication stress in prostate cancer, with dysfunctional DNA replication as the dominant mutational process in 28% of prostate cancers with CIN (IMF4 and IMF6). Although prostate cancer is generally considered to grow slowly, it may follow a pattern of punctuated equilibrium, where long periods of slow growth and relative genomic stability are interrupted by short episodes involving accelerated proliferation, genomic instability and replication stress[15,62–64]. Similar findings have been made in breast cancer, another hormone-sensitive cancer[65].

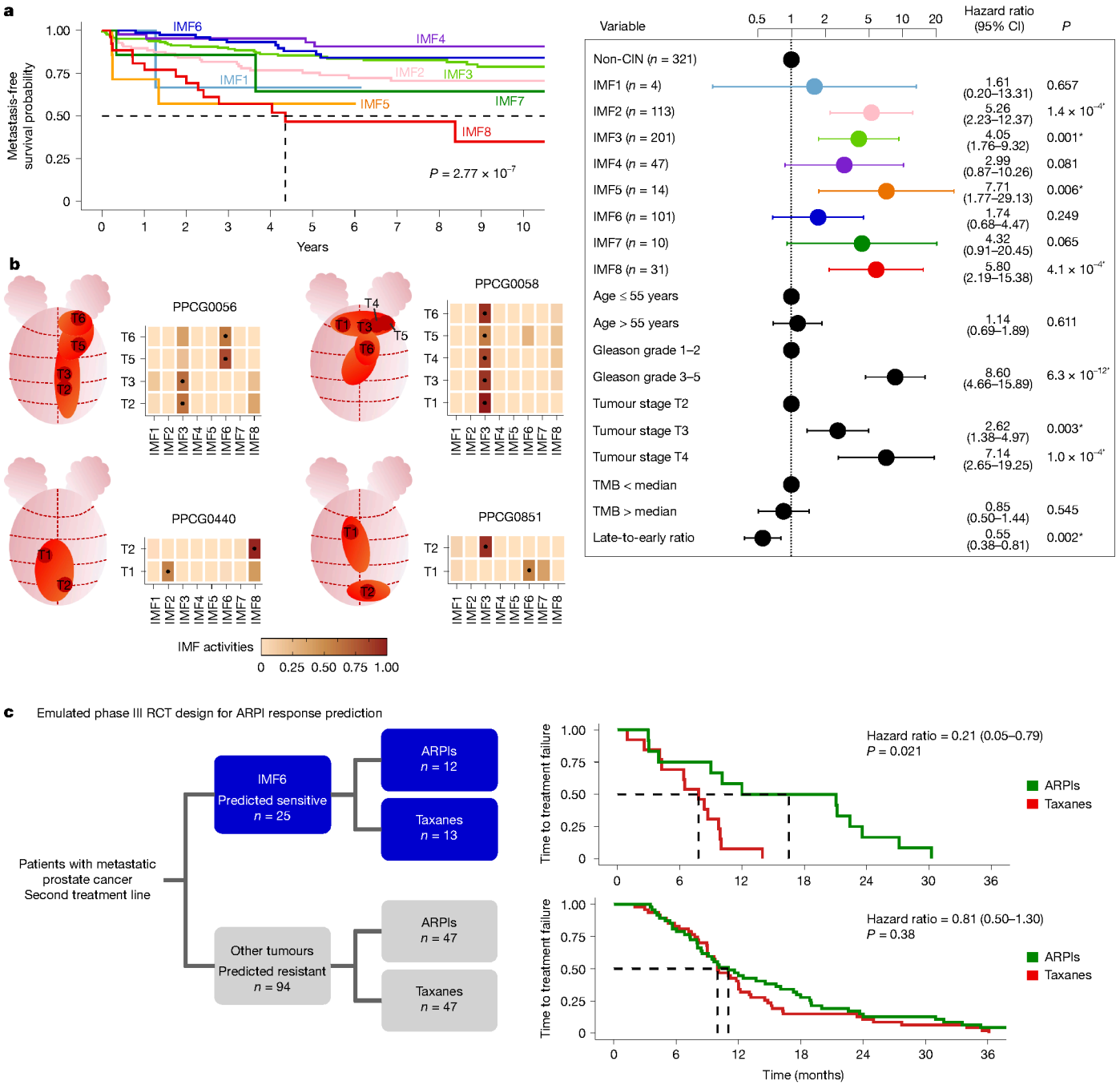


Fig. 5 | Eight dominant mutational processes explain the clinical heterogeneity in prostate cancer. a, MFS in PPCG samples with CIN stratified on the basis of their dominant mutational process. Kaplan–Meier curves (left) and a two-sided Cox proportional hazards model (right), showing MFS by dominant IMF. *P* values in the Kaplan–Meier curves were estimated using the log-rank test. The Cox proportional hazard model was corrected for the age at diagnosis, Gleason grade group, tumour stage, TMB and late-to-early mutation rate. Non-CIN samples ($n = 321$) were used as the reference group. The dots represent the mean hazard ratios and the bars represent the 95% CI. *$P < 0.05$. **b**, The activity levels of the eight IMFs in spatially confined tumour samples collected from the same patient. The black dots denote the dominant IMF per sample. Intratumour heterogeneity across patient samples was evaluated using Jaccard's distance (Supplementary Fig. 28). **c**, Assessment of IMF-based biomarkers for predicting therapy response in a real-world cohort of metastatic prostate cancer cases. Performance was assessed by emulating phase III randomized control biomarker trials. Kaplan–Meier curves show the TTF between patients with metastatic prostate cancer treated with second-line ARPIs or taxanes in biomarker-positive and biomarker-negative arms. *P* values, hazard ratios and 95% CIs were estimated using two-sided Cox proportional hazard models stratified by first-line therapy. No TTF differences were observed between patients treated with taxanes and ARPIs at the second treatment line (Extended Data Fig. 9). RCT, randomized controlled trial.

Although our study was not designed to develop a biomarker for risk stratification, we found that the IMFs were able to stratify primary prostate cancers that developed metastases. Using an approach to emulate phase III randomized control biomarker trials that we recently pioneered[60], we showed that IMF6 activity was predictive of reduced risk of treatment failure following ARPI. IMF6 is linked to *SPOP* mutations and replication stress, which corroborates recent findings of improved ARPI response in *SPOP*-mutated mCRPC[66] and may extend the scope of prediction beyond isolated *SPOP*-mutation-based approaches. Notably, only seven patients in the HMF cohort had *SPOP* alterations,

two of whom received these therapies at second line, further supporting that the predictive power of IMF6 is not limited to, but extends beyond, *SPOP*-mutated tumours. While these findings are encouraging towards addressing an urgent clinical unmet need, more extensive and well-powered prospective biomarker-driven studies are warranted.

Our study is not without limitations. For example, our approach is focused on identifying mutational processes causing alterations of the genome. Across the cohort, we identified 8.6% of primary prostate cancer samples without any detectable mutational processes, and found 22% to be non-CIN with primarily clock-like mutations. Several known oncogenic mutational processes involve dysregulation of epigenetic and/or transcriptional programs, and it is interesting to speculate that a subset of primary prostate cancers might be driven by such non-genetic processes. In support, a recent study found that epigenetic modifications, mediated by transcriptional silencing by Polycomb group proteins, could promote cancer initiation in an animal model system[67].

Together, our findings highlight that the complex nature of mutational processes, which often involve multiple types of genomic alterations (SNVs, indels, CNAs and SVs), requires a comprehensive integrative approach for accurate identification. We provide compelling evidence that such an approach is essential for deciphering the mutational processes shaping prostate cancers and for identifying patients developing more aggressive disease. Looking ahead, our study supports the implementation of WGS in clinical practice to improve risk stratification and potentially drive the development of predictive biomarkers for targeted therapy response.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10468-w.

1. Rebello, R. J. et al. Prostate cancer. *Nat. Rev. Dis. Primers* **7**, 9 (2021).
2. Bergengren, O. et al. 2022 update on prostate cancer epidemiology and risk factors—a systematic review. *Eur. Urol.* **84**, 191–206 (2023).
3. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
4. Wang, A. et al. Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants. *Nat. Genet.* **55**, 2065–2074 (2023).
5. Valsecchi, A. A. et al. Frequency of germline and somatic BRCA1 and BRCA2 mutations in prostate cancer: an updated systematic review and meta-analysis. *Cancers* **15**, 2435 (2023).
6. Lozano, R. et al. Genetic aberrations in DNA repair pathways: a cornerstone of precision oncology in prostate cancer. *Br. J. Cancer* **124**, 552–563 (2021).
7. Fang, B. et al. The somatic mutational landscape of mismatch repair deficient prostate cancer. *J. Clin. Med.* **12**, 623 (2023).
8. Abida, W. et al. Analysis of the prevalence of microsatellite instability in prostate cancer and response to immune checkpoint blockade. *JAMA Oncol.* **5**, 471–478 (2019).
9. Viswanathan, S. R. et al. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell* **174**, 433–447 (2018).
10. Antonarakis, E. S. et al. CDK12-altered prostate cancer: clinical features and therapeutic outcomes to standard systemic therapies, poly (ADP-Ribose) polymerase inhibitors, and PD-1 inhibitors. *JCO Precis. Oncol.* **4**, 370–381 (2020).
11. Wu, Y.-M. et al. Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell* **173**, 1770–1782 (2018).
12. Pomerantz, M. M. et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).
13. Wang, Q. et al. Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell* **138**, 245–256 (2009).
14. Morova, T. et al. Androgen receptor-binding sites are highly mutated in prostate cancer. *Nat. Commun.* **11**, 832 (2020).
15. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
16. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
17. Drews, R. M. et al. A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022).
18. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
19. Gerhauser, C. et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* **34**, 996–1011 (2018).
20. Yamaguchi, T. N. et al. The germline and somatic origins of prostate cancer heterogeneity. *Cancer Discov.* **15**, 988–1017 (2025).
21. 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).
22. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
23. Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364 (2017).
24. Jakobsdottir, G. M. et al. The Pan Prostate Cancer Group Dataset - Genomic, Transcriptomic and Methylation data for Clinical Discovery. *Sci. Data* https://doi.org/10.1038/s41597-026-08181-4 (in the press).
25. de Bono, J. S. et al. Central, prospective detection of homologous recombination repair gene mutations (HRRm) in tumour tissue from >4000 men with metastatic castration-resistant prostate cancer (mCRPC) screened for the PROfound study. *Ann. Oncol.* **30**, v328–v329 (2019).
26. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
27. Rodriguez-Fos, E. et al. Mutational topography reflects clinical neuroblastoma heterogeneity. *Cell Genom.* **3**, 100402 (2023).
28. Degasperi, A. et al. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
29. Menghi, F. et al. The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. *Cancer Cell* **34**, 197–210 (2018).
30. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
31. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
32. Sakofsky, C. J. et al. Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation. *PLoS Biol.* **17**, e3000464 (2019).
33. Maciejowski, J. et al. APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nat. Genet.* **52**, 884–890 (2020).
34. Pomerantz, M. M. et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. *Nat. Genet.* **52**, 790–799 (2020).
35. Lee, C. A., Abd-Rabbo, D. & Reimand, J. Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes. *Genome Biol.* **22**, 133 (2021).
36. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
37. Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J. & Jones, P. A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).
38. Parolia, A. et al. Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. *Nature* **571**, 413–418 (2019).
39. Lemma, R. B. et al. Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers. *Epigenet. Chromatin* **15**, 13 (2022).
40. Du, Q. et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun.* **10**, 416 (2019).
41. Stamatoyannopoulos, J. A. et al. Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
42. Keshet, I., Lieman-Hurwitz, J. & Cedar, H. DNA methylation affects the formation of active chromatin. *Cell* **44**, 535–543 (1986).
43. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
44. Jia, P. et al. MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability. *Genom. Proteom. Bioinform.* **18**, 65–71 (2020).
45. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836 (2019).
46. Jager, M. et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019).
47. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J.* **17**, 1195–1214 (2003).
48. Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun Biol.* **3**, 421 (2020).
49. Daniels, G. et al. Cytoplasmic, full length and novel cleaved variant, TBLR1 reduces apoptosis in prostate cancer under androgen deprivation. *Oncotarget* **7**, 39556–39571 (2016).
50. Hsieh, C.-L. et al. PLZF, a tumor suppressor genetically lost in metastatic castration-resistant prostate cancer, is a mediator of resistance to androgen deprivation therapy. *Cancer Res.* **75**, 1944–1948 (2015).
51. King, R. W. et al. A 20S complex containing CDC27 and CDC16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin B. *Cell* **81**, 279–288 (1995).
52. Hutchins, J. R. A. et al. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**, 593–599 (2010).
53. Parker, A. L., Kavallaris, M. & McCarroll, J. A. Microtubules and their role in cellular stress in cancer. *Front. Oncol.* **4**, 153 (2014).
54. Chan, K. L., Palmai-Pallag, T., Ying, S. & Hickson, I. D. Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nat. Cell Biol.* **11**, 753–760 (2009).
55. Riaz, N. et al. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun.* **8**, 857 (2017).
56. Gong, T. et al. Genome-wide interrogation of structural variation reveals novel African-specific prostate cancer oncogenic drivers. *Genome Med.* **14**, 100 (2022).
57. Blattner, M. et al. SPOP mutation drives prostate tumorigenesis in vivo through coordinate regulation of PI3K/mTOR and AR signaling. *Cancer Cell* **31**, 436–451 (2017).
58. 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).
59. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).

60. Thompson, J. S. et al. Predicting resistance to chemotherapy using chromosomal instability signatures. *Nat. Genet.* https://doi.org/10.1038/s41588-025-02233-y (2025).
61. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
62. Takahashi, N. et al. Replication stress defines distinct molecular subtypes across cancers. *Cancer Res. Commun.* **2**, 503–517 (2022).
63. Cuzick, J. et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *Lancet Oncol.* **12**, 245–255 (2011).
64. Maya-Mendoza, A. et al. High speed of fork progression induces DNA replication stress and genomic instability. *Nature* **559**, 279–284 (2018).
65. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
66. Swami, U. et al. SPOP mutations as a predictive biomarker for androgen receptor axis-targeted therapy in de novo metastatic castration-sensitive prostate cancer. *Clin. Cancer Res.* **28**, 4917–4925 (2022).
67. Parreno, V. et al. Transient loss of Polycomb components induces an epigenetic cancer fate. *Nature* **629**, 688–696 (2024).

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

[1]Department of Biology, University of Konstanz, Konstanz, Germany. [2]Biotech Research & Innovation Centre (BRIC) - University of Copenhagen, Copenhagen, Denmark. [3]Finsen Laboratory, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark. [4]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [5]Computational Biology Program, Ontario Institute for Cancer Research (OICR), Toronto, Ontario, Canada. [6]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [7]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [8]Manchester Cancer Research Centre and Cancer Research UK Manchester Institute, The University of Manchester, Manchester, UK. [9]Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [10]Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK. [11]The Earlham Institute, Norwich, UK. [12]The Institute of Cancer Research, London, UK. [13]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [14]German Cancer Consortium (DKTK), Core Center Heidelberg, and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [15]Medical Faculty Heidelberg and Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [16]Wellcome Sanger Institute, Hinxton, UK. [17]CeRePP (Centre de Recherche sur les Pathologies Prostatiques et Urologiques), Paris, France. [18]Sorbonne Université, GRC n°5 Predictive Onco-Urology, APHP, Tenon Hospital, Paris, France. [19]Collaborative Centre for Genomic Cancer Medicine, The Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia. [20]Department of Surgery, The University of Melbourne, Parkville, Victoria, Australia. [21]Department of Urology, Peninsula Health, Frankston, Victoria, Australia. [22]Division of Urology, Royal Melbourne Hospital, Parkville, Victoria, Australia. [23]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [24]Department of Urology, Western Health, Footscray, Victoria, Australia. [25]Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, New South Wales, Australia. [26]Department of Surgery, University of Cambridge, Cambridge, UK. [27]Department of Urology, Cambridge University Hospitals NHS Trust, Cambridge, UK. [28]University of Cambridge & Cambridge University Hospitals NHS Trust, Cambridge, UK. [29]Cambridge Urology Translational Research and Clinical Trials Office, Cambridge Biomedical Campus, Addenbrooke's Hospital, Cambridge, UK. [30]Genitourinary Oncology Service, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [31]School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa. [32]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [33]Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria, Australia. [34]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [35]Division of Cancer Sciences, The University of Manchester, Manchester, UK. [36]Christie Hospital, The Christie NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. [37]Melbourne Bioinformatics, The University of Melbourne, Parkville, Victoria, Australia. [38]Weill Cornell Medicine, New York, NY, USA. [39]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [40]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [41]The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [42]The New York Genome Center, New York, NY, USA. [43]Department of Oncology, KU Leuven, Leuven, Belgium. [44]Department of Medical Biology, The University of Melbourne, Melbourne, Victoria, Australia. [45]Australian BioCommons, The University of Melbourne, Parkville, Victoria, Australia. [46]Department of Medicine, Central Clinical School, Faculty of Medicine Nursing and Health Sciences, Monash University, Melbourne, Victoria, Australia. [47]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [48]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, USA. [49]Institute of Pathology, University Medical Centre Hamburg - Eppendorf (UKE), Hamburg, Germany. [50]Royal Marsden NHS Foundation Trust, London, UK. [51]Christie NHS Foundation Trust, Manchester, UK. [52]Department of Urology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany. [53]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [74]These authors contributed equally: Andreas J. Gruber, André V. Olsen, Barbara Hernando, Kevin C. L. Cheng. [75]These authors jointly supervised this work: Geoff Macintyre, Jüri Reimand & Joachim Weischenfeldt.

Pan Prostate Cancer Group (PPCG)

Andreas J. Gruber[1,74], André V. Olsen[2,3,74], Barbara Hernando[4,74], Kevin C. L. Cheng[5,6,74], Clarissa Gerhäuser[7], Lucy Barton[8], Diogo Pellegrina[5,6], G. Steven Bova[9], Daniel S. Brewer[10,11], Mark N. Brook[12], Benedikt Brors[13,14,15], Adam Butler[16], Géraldine Cancel-Tassin[17,18], Niall M. Corcoran[19,20,21,22,23,24], Olivier Cussenot[17], Abraham Gihawi[10,25], Etsehiwot G. Girma[2,3], Vincent J. Gnanapragasam[26,27,28,29], Anis A. Hamid[20,30], Vanessa M. Hayes[8,10,25,31], Housheng Hansen He[6,32], Christopher M. Hovens[19,20,22,23,33], Eddie L. Imada[34], G. Maria Jakobsdottir[35,36], Chol-Hee Jung[37], Francesca Khani[34,38], Zsofia Kote-Jarai[12], Philippe Lamy[39,40], Gregory Leeman[16], Massimo Loda[34,41,42], Pavlo Lutsik[7,43], Luigi Marchionni[34], Ramyar Molania[23,44], Anthony T. Papenfuss[23,44], Bernard Pope[20,37,45,46], Lucio R. Queiroz[34], Tobias Rausch[47], Brian Robinson[34,38], Atef Sahli[8], Karina D. Sørensen[39,40], Takafumi N. Yamaguchi[48], Sebastian Uhrig[13], Yaobo Xu[16], Claudio Zanettini[34], Alain Bergeron[54], Alan F. Rubin[23,44], Alastair D. Lamb[55], Alejandro Berlin[32], Aleksandra S. Ziuboniewicz[56], Andrew Erickson[56], Andrew Ryan[57], Andy G. Lynch[58,59], Angelo Corso Faini[34], Anne Nguyen[20], Anne Warren[60], Aravind Sankar[61], Arfa Mehmood[8,36], Balthasar C. Schlotmann[2,3], Beng Lim[20], Bethany K. Campbell[19,20,22,23], Breon Feran[23], Charlie E. Massie[62], Chen Hong[63], Chris Foster[12], Christoph Plass[7], Christopher Wirth[8], Colin Collins[64], Corinna Grima[20], Dan J. Woodcock[56], Daniel Barrowdale[61], Daniel J. Park[37], Emre Esentürk[56], Esther Baena[8], Fabian Falkenbach[65], Filippo Pederzoli[34], Giselle Kerry[61], Giuseppe N. Fanelli[34], Grace Hall[37], Harveer Dev[62], Holger Sültmann[14], Hubert Pakula[34], Ian Mills[56], Jan Korbel[47], Jennifer Ureta[23,44], Jessica Heilmann[7], Jingjing Zhang[10,66], Jocelyn Sietsma Penington[23], Jue Jiang[25], Justin Bedő[23,44], Justin S. Peters[20], Keiran Raine[16], Ken Chow[20], Kira Furlano[52], Lars Feuerbach[13,63], Lena Voith von Voithenberg[63], Marek Cmero[23,44], Markus Graefen[67], Matthew K. H. Hong[20], Matthew Wakefield[23,44], Matti Nykter[9], Melissa Cheung[26,27,28,29], Michael Borre[39,68], Michael J. Clarkson[20], Michael Kerger[20], Michael Scherer[7], Mohamed Omar[69], Nana Mensah[70], Natalie Kurganovs[20], Nening Dennis[12], Noora Al-Muftah[56], Parsa Pirhady[8], Patrick McCoy[20], Peter Georgeson[19,33], Peter Van Loo[71], Philip Dundee[20,22], Pier Vitale Nuzzo[34], Pierre Tennstedt[67], Pramit M. Phal[20], Radoslaw Lach[26,27,28,29,58], Rajbir N. Batra[7,62], Simon Pacey[8], Ronnie Rodrigues Pereira[8], Ruining Dong[23,44], Ryan Carelli[34], Ryan Stuchbery[20], Sam Norden[57], Samson Olofinsae[8], Silvia Rodrigues[34], Stefano Mangiola[23,44], Thomas Keane[61], Toby Milne-Clark[26,27,28,29,62], Tony Costello[20], Valeriia Haberland[10,72], Vivien Holmes[8], Weerachai Jaratlerdsiri[25,73], Wi Burns[12], Wikum Dinalankara[34], William J. Hutchison[23,44], Xiaotong He[8], Yidan Pan[71], Yong-Jie Lu[55], Yunfan Fu[23,44], Yves Fradet[54], Ronald Simon[49], Guido Sauter[49], Ros A. Eeles[12,50], Colin S. Cooper[10], Robert G. Bristow[8,35,51], David C. Wedge[8,36], Thorsten Schlomm[52], Geoff Macintyre[4,75], Jüri Reimand[5,6,53,75] & Joachim Weischenfeldt[2,3,52,75]

[54]CHU de Québec – Université Laval, Research Center, Oncology Axis, Québec, Quebec, Canada. [55]Barts Cancer Institute, Queen Mary University of London, London, UK. [56]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. [57]TissuPath, Mt Waverley, Victoria, Australia. [58]School of Mathematics and Statistics, University of St Andrews, St Andrews, UK. [59]School of Medicine, University of St Andrews, St Andrews, UK. [60]Department of Pathology, Cambridge University Hospitals, NHS Foundation Trust & University of Cambridge, Addenbrookes Hospital, Cambridge Biomedical Campus, Cambridge, UK. [61]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK. [62]Early Cancer Institute, Department of Oncology, University of Cambridge, Cambridge, UK. [63]German Cancer Research Center (DKFZ), Heidelberg, Germany. [64]Vancouver Prostate Centre, Vancouver, British Columbia, Canada. [65]Department of Urology, University Hospital Heidelberg, Heidelberg, Germany. [66]School of Engineering, Newcastle University, Newcastle upon Tyne, UK. [67]Martini-Klinik, University Medical Centre Hamburg - Eppendorf (UKE), Hamburg, Germany. [68]Department of Urology, Aarhus University Hospital, Aarhus, Denmark. [69]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [70]The Francis Crick Institute, London, UK. [71]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [72]Department of Computer Science, Brunel University of London, Uxbridge, UK. [73]Computational Genomics Group, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Camperdown, New South Wales, Australia.

Methods

PPCG cohort, WGS and variant calling

We used data from the PPCG consortium of primary prostate cancer samples from a total of 1,001 prostate cancer donors with 1,172 tumour samples. Informed ethical consent was obtained at clinical follow-up, and was consistent with local research ethics and International Cancer Genome Consortium (ICGC) guidelines (https://icgc.org/). Ethical approval was obtained from local research ethical committees (details are provided in the Supplementary Methods).

Analysis of SVs from genomic data

Simple and complex SV classification method. We conducted SV classification on the PPCG cohort using the cSVc tool[18], including also SVs near TARBS and chromothripsis, both abundant in prostate cancer. First, exact breakpoint estimation from our WGS short-read sequencing data was conducted by using split-read information. Specifically, for each sample, we used the median of soft-clipped reads from the corresponding tumour BAM file to obtain corrected breakpoint positions. Next, we used ClusterSV[18] to obtain clusters of SVs from the corrected SV data and we also used Battenberg[68] to obtain segmented CN files from tumour and normal read coverage files. For each sample, we generated CN files segmented by corrected SV breakpoints by using the corrected SV data, the CN segmentation file, CN coverage and the tumour BAM file. Finally, we conducted simple and complex SV classification (chromothripsis excluded) using the CN-SV segmentation file together with the corrected SV file, ClusterSV file and information about ploidy and purity.

To classify chromothripsis on the PPCG cohort, we used Shatterseek v.0.4, using SV and CN data. We used Shatterseek's recommended cut-off criteria to obtain high-confidence calls[69]. Multiple-testing correction was performed on the *P* values of three statistical tests indicative of chromothripsis (breakpoint enrichment test, exponential clustering test and the fragment joins test) and we used a *q*-value cut-off of 0.2 for the chromothripsis calls on each chromosome to confine the final high confidence call set.

To obtain the final SV dataset we merged the chromothripsis call set with the remaining simple and complex SV call set. Specifically, we merged calls only if a minimum of 90% of SV call sets spanning the chromothripsis cluster were marked as Complex Unclassified.

Tandem duplication detection. The presence of TDP samples was assessed by using three criteria described previously[70]. The criteria involve the proportion of tandem duplications, the total tandem duplication count and a TDP score[71]:

$$\text{TDP score} = -\frac{\sum_i |\text{Obs}_i - \text{Exp}_i|}{\text{TD}}$$

For a given sample, Obs_i denotes the observed tandem duplications for chromosome i, Exp_i denotes the expected tandem duplications for chromosome i and TD is the total number of tandem duplications. We calculated Exp_i by counting the tandem duplications per chromosome among all samples and dividing this by the sample size. Together, we classified TDP samples when the following criteria were met: tandem duplication proportion > 20%, total tandem duplication count > 50, and TDP score > −1.

Enrichment analysis of SV classes at specific genomic features were performed using a permutation testing framework as described in the Supplementary Methods.

De novo extraction and assignment of mutational signatures

SBS and ID signatures. For the de novo extraction of SBS and ID signatures we considered high-accuracy SNVs and stringently filtered (Tier 1) high-accuracy indels called and quality controlled as described previously[24]. SBS signatures were extracted de novo using SigProfilerExtractor (v.1.1.23)[72]. De novo extraction of SBS signatures (96 channels) and ID signatures (83 channels) was performed for 3 to 25 signatures for SBS and 1 to 25 signatures for ID using the following parameters: nmf_replicates=500;nmf_init = "random";min_nmf_iterations=10,000;max_nmf_iterations=1,000,000;export_probabilities=True;make_decomposition_plots=True;get_all_signature_matrices=True. We then decomposed every set of signatures (3 to 25 signatures for SBS and 1 to 25 signatures for ID) independently to known COSMIC signatures (v.3.4 for SBS and v.3.3 for ID) using the default parameters of the decompose_fit function of SigProfilerExtrator. To find the optimal number of SBS or ID signatures, respectively, we investigated each obtained set of signatures based on a number of features (Supplementary Methods).

Structural variant signatures (cSV). We extracted SV signatures by incorporating a non-negative matrix factorization (NMF) framework onto specific feature counts of the SV data. For deletions and tandem duplications, we grouped the SV classes into four different size groups: 0–50 kb, 50–500 kb, 500 kb to 5 Mb and >5 Mb. Each of these groups was further subdivided on the basis of breakpoint occurrence at distinct replication timing regions (early, mid and late) by applying bedtools pairToBed between SV data and replication timing data (see the 'Enrichment analysis of SV classes at specific genomic features' section of the Supplementary Information for details about assessment of replication timing status). In the same manner, we also incorporated a group that captured deletions and tandem duplications occurring at fragile sites (see the 'Enrichment analysis of SV classes at specific genomic features' section of the Supplementary Information for details about fragile site data). For balanced inversions, and the two local 2-jump classes Dup-invDup and Loss-invDup, we created two groups based on size with a 100 kb threshold. Templated insertion groups (chains, cycles and bridges) were divided by a 5 kb size threshold. Lastly, we included two novel features compared to a previous study[18], including chromothripsis and SV breakpoints in close proximity to TARBS, a unique feature of prostate cancer genomes. We counted SV breakpoints occurring at TARBS regardless of SV class (see the 'Enrichment analysis of SV classes at specific genomic features' section of the Supplementary Information for details about TARBS data). These groups were converted into a matrix that was used as input for the NMF (NMF package in R)[73]. For finding the optimal number of SV signatures, we applied a procedure used for finding the optimal number of rearrangement signatures in the Palimpsest R package[74]. In brief, based on 100 iterations, we determined the index for which the cophenetic values had the steepest decrease and used that index as the number of signatures. We normalized the SV signature values by dividing each value in a row with the row sum and subsequently visualized these values with ggplot2. We also extracted cSVs from WGS data of the HMF-Prostate cohort to quantify cSV signatures (see the 'Processing genomic data from the HMF-Prostate cohort' section).

CX signatures. CX signatures were identified and quantified from absolute copy-number profiles as previously described[17] (Supplementary Methods).

We computed the genome-wide distributions of the five fundamental copy-number features for each sample (the segment size, the difference in copy-number between adjacent segments, the lengths of oscillating copy-number segment chains, the breakpoint counts per 10 Mb, and the breakpoint counts per chromosome arm). Feature distributions observed in the cohort were then deconstructed into a total of eight signatures by applying a NMF. All of these signatures were already present in the pan-cancer compendium of CX signatures[17] (cosine similarity > 0.85). The final signature activities for all samples were computed using the linear combination decomposition function from YAPSA[75]

(R package, v.1.12.0) on the feature component-distributions given a predefined signature matrix. To ensure the robustness of the signature activities and to enable trust in small signature activities, we identified signature-specific thresholds by following the same procedure as in our previous work[17]. We next set activities to zero if they were below the signature-specific threshold. The sample-by-activity vector was finally normalized per sample. This approach was also applied to quantify CX signatures in the HMF-Prostate cancer cohort (see the 'Processing genomic data from the HMF-Prostate cohort' section).

Identifying integrated mutational footprints

We performed clustering of patient tumours defined by the complete spectrum of signatures to characterize the main mutational processes driving prostate cancer. For each sample, we combined the activities of each of the different types of mutational signatures into a single vector for clustering input, whereas only signatures being active in any of the clustered samples were considered. The data were scaled and centred before clustering, which was performed independently for samples with CIN and without CIN (non-CIN). Given that all tumours share a core set of clock-like mutational signatures[36] complemented with more distinct mutational processes during tumorigenesis, we opted to use hierarchical clustering given the natural hierarchical structure of the data. We tested a broad range of clustering parameters and empirically determined that the Ward-D2 algorithm in combination with Manhattan distance best identified known biological features of our cohort. This approach has also been chosen in other studies[76,77] to prioritize robust identification of known biology. In the case of clustering CIN samples, we excluded samples with fewer than 100 SNVs, fewer than 10 indels, or fewer than 5 SVs. We identified eight clusters of CIN samples (Fig. 4b) based on (1) the obtained minimum, median and maximum cluster stability metrics (Supplementary Fig. 12); (2) the silhouette scores in samples (Supplementary Fig. 13); and (3) biological expert knowledge. Each of these clusters characterized a distinct mutational process represented by the integration of mutational signatures. Thus, we referred to these processes as IMFs. In the case of non-CIN samples, we excluded samples with less than 100 SNVs, or less than 10 indels (244 cases; Supplementary Fig. 17). As performed for CIN samples, we explored cluster stability metrics to select 4 clusters for non-CIN samples (Supplementary Figs. 18 and 19).

To quantify the activity of the eight IMFs in a given tumour, we first generated an IMF definition matrix, where each IMF was defined across the 46 multi-signature space (22 SBS, 10 ID, 8 CX and 6 cSV signatures). This was done by averaging the multi-signature activities across all samples within each cluster (Extended Data Fig. 6). Then, given the IMF definitions, we applied the linear combination decomposition function from YAPSA[75] to the multi-signature vector for each tumour to compute the activity of each of the eight IMFs in the tumour sample. The resulting IMF activity vector is then normalized to sum to 1.

After quantifying the activities of the IMFs, each sample was classified according to its dominant mutational process, defined as the IMF with the highest activity. This stratification was applied for testing the potential of the IMF as biomarkers for predicting metastasis occurrence and treatment response (see the 'Predicting clinical outcomes in cases of primary prostate cancer' section).

Linking mutational processes with genomic features

Here we performed association analyses between activities of the integrated mutational footprints and different genomic features. We also correlated each individual mutational signature (SBS, ID, cSV and CX signatures) with different genomic features (detailed below). Only samples with available calls for a specific mutational signature and genomic feature were included in each analysis.

Spearman's correlation was applied to test the correlation of IMF activities with the 44 individual mutational signatures. All *P* values were corrected for multiple testing using the Benjamin–Hochberg method (Extended Data Fig. 7a).

Selection of 1,747 genes of interest. We compiled a list of 1,747 genes by integrating and curating the annotations provided in refs. 78–81. This set of genes included driver genes recurrently altered in prostate cancer[21,82], known pan-cancer driver genes, as well as genes reported to be involved in DNA damage response and repair (DDR) pathways, DNA replication, the cell cycle and chromatin organization. For all 1,747 genes, we identified their transcriptional start sites (TSS) on a GRCh37 reference genome using ENSEMBL. We selected the canonical TSS for each gene.

Mutation status of genes of interest. For each gene, we integrated germline and somatic alterations for classifying samples as mutant using the following criteria.

Oncogenes: samples with amplifications and somatic mutations that did not result in the truncation of the oncogene were classified as mutant. For oncogenes located in autosomal chromosomes (that is, *SPOP*, *IDH1* and *MYC*), amplifications were considered in cases in which the copy-number value at the transcriptional start site exceeded the sample ploidy by at least two copies. For oncogenes located in sex chromosomes (*AR*), the threshold was adjusted to require a gain of at least two copies more than the ploidy − 1. Given that the amplification of the enhancer region is frequently observed to be amplified after ADT treatment[83], we scanned a genomic region consisting of 0.5 Mb before the TSS and 1 Mb after the coding region of the *AR* oncogene.

Tumour suppressor genes: only samples with biallelic inactivation or homozygous deletion (copy-number value < 0.5) of the gene were considered as mutant, with the following exceptions: (1) for *TP53*, samples with germline variants or single hotspot driver mutations; (2) for *CDK12*, samples with single hotspot driver mutations; and (3) for DDR-related genes, samples with germline variants due to its impact on cancer risk. For *BRCA1* mutants, we also evaluated the presence of reversion mutations rescuing the HRD phenotype. Rescue was defined by a mutation in either *TB53BP1*, *RIF1* or *MAD2L2*[84,85]. None of the *BRCA1* mutant samples presented reversion mutations.

To test for differences in signature activities between samples with and without mutations in any of the 1,747 key genes, we divided the samples into two groups: one group with a mutation and one without. We then used Welch's *t*-test to perform a test of equality of means between the two groups. Only positive associations were shown to mitigate compositional data effects. To ensure high-confidence associations, we discarded gene associations lacking a moderate/large effect size (Cohen's $d < 0.5$). Finally, we further filtered out genes with less than three samples mutated for IMF associations (Fig. 4c).

In the case of individual mutational signatures, we limited the association analysis to well-known prostate cancer genes[21,82], filtered out genes with less than ten samples mutated, and the analysis was performed for the whole cohort and also after stratifying samples on the basis of age at diagnosis (Fig. 2d and Supplementary Fig. 29).

Deficiency in DDR mechanisms. Here we evaluated differences in signature activity between samples deficient and proficient for five different DDR mechanisms: HR, MMR, NHEJ, NER and BER. To do so, we first evaluated the mutation status of genes involved in each DDR pathway. Samples with inactivation of at least one of the DDR-related genes were classified as deficient; otherwise, as proficient. The genes evaluated for each repair pathway were as follows: HR: *BRCA1, BRCA2, PALB2*; MMR: *MLH1, MSH2, MSH6, PMS2*; BER: *MUTYH, NTHL1, POLD1, POLE, WRN*; NHEJ: *NBN, RAD50, ATM*; and NER: *DDB2, ERCC2, ERCC3, ERCC4, ERCC5, POLD1, POLE, XPC, XPA, CUL3*.

For each repair mechanism, Welsh's *t*-tests were performed to compare signature activities between deficient and proficient samples.

P values were corrected for multiple testing using the Benjamini–Hochberg method. Cohen's *d* was computed to assess the size of the difference between groups (Fig. 2e, Supplementary Fig. 22 and Extended Data Fig. 7).

Assigning mutational signatures to mutations. For each mutation type in a sample, SigProfilerExtractor[72] outputs the posterior probability of a mutation being generated from a signature given the mutation type. The probability is computed by SigProfilerExtractor from the estimated exposure of each signature in the sample, trinucleotide profilers of the de novo extracted signatures, and the abundance of each mutation type in the sample. The SigProfilerMatrixGenerator package was used for classifying SNVs into 96 categories and for classifying indels into 83 categories based on the sequence context[86]. We then assigned mutations to the signature with the highest probability given their mutation type, as previously done in other studies[87–89]. We note that flat signatures, such as SBS5 and SBS40a, are intrinsically difficult to distinguish because their trinucleotide profiles have low information content and high similarity to one another[16,90].

WGD. To test whether the WGD status associated with the activity of a signature, we compared activities between samples with and without WGD by applying a Welch's *t*-test. Only samples with activities higher than zero were included in the analysis. *P* values were corrected for multiple testing using the Benjamini–Hochberg method. Cohen's *d* was computed to assess the size of the difference between groups (Extended Data Fig. 4b).

Whole-chromosome CNAs. Copy-number segments were classified as a whole-chromosome alteration if it was larger than 95% of the whole chromosome length covered and had a copy-number value greater or smaller than 2. We then compared signature activities between samples with and without whole-chromosome CNAs using a Welch's *t*-test. *P* values were corrected for multiple testing using the Benjamini–Hochberg method. Cohen's *d* was computed to assess the size of the difference between groups.

Focal amplifications. Focal amplifications were defined as segments with a copy-number value equal or higher than 8. We then compared the activities of signatures between samples with and without focal amplifications using a Welch's *t*-test. *P* values were corrected for multiple testing using the Benjamini–Hochberg method. Cohen's *d* was computed to assess the size of the difference between groups.

Microsatellite instability. Microsatellite instability (MSI) was evaluated using MSIsensor-pro v.1.3.0 with the default parameters[44]. In brief, MSIsensor-pro scan was first executed on the reference genome to obtain homopolymers and microsatellites. Next, MSIsensor-pro was applied on tumour and normal BAM files with the homopolymer and microsatellite information file, and with coverage threshold set to 15 and coverage normalization set to 1. Samples with a score higher than 10 were considered as MSI-Hi.

HRDetect. We used HRDetect[43] to detect the presence of samples with HRD. In brief, we combined all four mutation data type (SV, CNA, SNV and indels) as input for HRDetect to infer six HRD-associated features (microhomology deletions, SBS3 and SBS8 signature, SV3 and SV5 signature). These features were then used to compute an HRD probability score using the default parameters. We considered all samples with a probability score > 0.7 to have HRD.

Kataegis. The detection of kataegis was performed as described previously[24]. We counted the number of kataegis events per sample and tested the Spearman correlation to the signature activities. Multiple-testing correction was performed according to the Benjamini–Hochberg method.

Cell cycle progression scores. RNA-sequencing (RNA-seq) data were processed and harmonized before downstream analysis as described previously[24]. For this analysis, we used the high-confidence set of gene expression data (tier 1) RUV-III PRPS-normalized data (v.1.4/K10). Cell cycle progression (CCP) scores, also known as Prolaris scores, were computed for each sample. For each of the 31 genes in the CCP signature[63], $\log_2$-normalized expression values were centred and scaled relative to the median expression of the specific gene across all samples. The resulting values were then summed to produce a single CCP score per sample. All 31 signature genes were available in the tier 1 data release used for this analysis.

AR pathway activities. The pathway scores were computed from AR and prostate cancer-related gene sets using the singscore package (https://doi.org/10.18129/B9.bioc.singscore) v.3.21 as described in our companion paper[24] using the singscore HALLMARK_ANDROGEN_RECEPTOR as a representation of AR activity.

Hypoxia scores. A hypoxia score was computed as described in our companion paper[24] using the 'Buffa' hypoxia signature[91]. In brief, the hypoxia score was computed for each RNA-seq sample by median-centring gene expression across the cohort, scaling to unit variance and summing across signature genes to yield a single hypoxia score.

ARBS mutation rate enrichment. Chromatin immunoprecipitation–sequencing (ChIP–seq) data of prostate tumour-specific AR binding (TARBS, 9,181 sites), normal prostate-specific AR binding (NARBS, 2,690 sites), *FOXA1* (19,735 sites) and HOXB13 (66,104 sites) were obtained from ref. 12. TARBS were overlapped with FOXA1 ChIP–seq peaks and HOXB13 ChIP–seq peaks to obtain co-binding sites (Supplementary Fig. 11a). Assay for transposase-accessible chromatin using sequencing (ATAC–seq; 111,884 sites) data were obtained from ref. 92. ATAC–seq peaks were overlapped with TARBS to obtain $TARBS^+$ ATAC–seq sites. ATA–seq peaks without TARBS overlap were sampled without replacement by selecting sites that match the ATAC–seq signal of each $TARBS^+$ ATAC–seq site—this forms a set of $TARBS^-$ ATAC–seq sites with matching site number and ATAC–seq signal distribution for comparison to $TARBS^+$ ATAC–seq sites (Supplementary Fig. 7c). Similarly, TARBS were downsampled to match the number and ATAC–seq signal of NARBS for an additional validation (Supplementary Fig. 7b). Chromatin state data were obtained from ref. 34. Enhancer regions were overlapped with TARBS to obtain a set of $TARBS^+$ enhancers, and enhancers without TARBS overlapped were sampled to the same number as $TARBS^-$ enhancers for comparison (Supplementary Fig. 7d).

To analyse mutation rate uniformly, we defined ARBS to be 400 bp regions using the midpoint of each ChIP–seq peak as the midpoint of each site. 400 bp is the median of ChIP–seq peak length. We also observed that the mutation rate starts to increase around the 400 bp boundary. 400 bp flanks on both sides of the defined ARBS were used as control. A negative binomial regression model, RM2 (ref. 35), was used to compare ARBS mutation rate to flanking regions. Sequence trinucleotide ratio and megabase-scale background mutation rate were accounted for as covariates in the model. Mutation rate of SBS and ID signatures was calculated by considering subsets of mutations assigned to individual signatures. ARBS mutation rate analysis was performed on samples with at least 100 SNVs, excluding samples with all SNVs attributed to artifact signatures ($n = 959$).

ARBS methylation level. Eight samples were sequenced on the Oxford Nanopore Promethion N24 machine. Methylation levels were called using the modkit package (https://github.com/nanoporetech/modkit)

and methrix package (https://github.com/CompEpigen/methrix). Uncovered sites and sites overlapping SNPs were removed, as well as sites with a coverage below 20× in 6 out of the 8 samples. In each sample, the methylation level of each CpG site was calculated as the number of methylated reads divided by the total reads. The get_region_summary function in methrix was then used to generate average methylation levels of CpG sites located in TARBS and NARBS in each sample. Two-sided Wilcoxon rank-sum tests were used to compare TARBS and NARBS methylation levels.

Replication timing–mutation rate association. Early, mid and late replication timing regions were defined as described in the 'Enrichment analysis of SV classes' section of the Supplementary Information. For each sample, the mutation rates of each replication timing were calculated as the number of SNVs and indels per Mb. replication timing associations were calculated as the ratio of late to early mutation rates.

To account for sequence-context biases that may differ across replication timing regions, we generated an expected background using SigProfilerSimulator. Within each sample, SNVs and indels were randomly assigned to a new genomic position that preserved their original context (trinucleotide for SNVs and indel-context class for indels). The expected mutation rates for replication timing regions were then computed from the simulated mutations (SNVs + indels per Mb), and a simulated late-to-early replication timing ratio was derived for each sample. Corrected replication timing ratios were obtained by dividing the observed late-to-early ratio by the corresponding simulated ratio. When estimating replication timing association of individual samples, two screening steps were implemented to avoid unreliable estimation of replication timing association due to low mutation number. Firstly, samples with less than 100 mutations were excluded, secondly, samples with less than 10 mutations in either early or late replication timing regions were excluded ($n = 959$).

Replication timing association of SBS and ID signatures of individual samples was estimated by considering subsets of mutations assigned to each signature as described in the signature assignment section. The same mutation number screening procedure was applied. For background correction in signature analyses, Sigprofiler simulated mutations were assigned to their most probable mutational signature using the same procedure applied to observed mutations and signature-specific corrected replication timing ratios were computed analogously.

Owing to low mutation counts in individual samples when restricting regions to TARBS or NARBS, regional mutation rates were estimated by summing mutations from all samples. A genome mutation rate for all samples was also calculated with the same method applied to compare with TARBS and NARBS. TARBS/NARBS replication timing mutation rates for each IMF were calculated by summing mutations from all samples belonging to each IMF.

Signature timing. To investigate the temporal dynamics of APOBEC- and HRD-associated mutagenesis in prostate tumours with evidence of IMF7 activity, we requantified the contributions of SBS2, SBS13 (APOBEC), SBS3 and ID6 (HRD) in clonal and subclonal epochs. For each donor, clonal and subclonal mutations were analysed separately and signature activities (A_{clonal} and $A_{subclonal}$, respectively) were estimated using non-negative least squares (NNLS; R nnls package v.1.5), following the framework applied in PCAWG[93]. To compare activity between epochs, we calculated a fold change (FC) statistic for each signature, defined previously[93]:

$$FC = \frac{(A_{subclonal}/(1 - A_{subclonal}))}{(A_{clonal}/(1 - A_{clonal}))}$$

where A_{clonal} and $A_{subclonal}$ represent the re-estimated activities in the clonal and subclonal epochs, respectively. This statistic reflects dynamic changes in signature activity, with FC > 1 indicating higher activity in the subclonal period.

Predicting clinical outcomes in cases of primary prostate cancer

Preparation of clinical data. The collection and harmonization of clinical data from the PPCG cohort is described previously[24] (Supplementary Table 5). Histopathological data included tumour stage and Gleason grade group. Tumour stages were simplified by omitting additional stage identifier after the main identifier (that is, stage T3b was simplified to stage T3). For survival analyses, tumour stage was then categorized into three groups (T2, T3 and T4); Gleason grade group (GG) 1 to 5 was divided into two groups (≤2 and >2); and age at diagnosis was split into two categories (≤55 and >55 years). MFS was used as the primary end point for testing prognostic utility of mutational scenarios. MFS was defined as the number of days from the date of sample collection and to the date of first metastasis.

Survival analyses. Kaplan–Meier estimates (function survfit from the survival[94] R package) and two-sided Cox proportional hazard models (function coxph from the survminer[95] R package) were used for survival analysis across patients based on their dominant integrated mutational footprint (IMF; see the 'Identifying integrated mutational footprints' section). Cox proportional hazard models were corrected by age at diagnosis, tumour stage, TMB, late-to-early mutation rate and Gleason grade. Wald test was used to evaluate statistical significance of regression coefficients obtained from Cox proportional hazard models.

Given the well-known clinical impact of CIN[96], we first evaluated differences in MFS between tumours with (≥20 CNAs) and without (<20 CNAs) detectable CIN (Extended Data Fig. 1). As expected, tumours without CIN showed longer MFS times compared to tumours with CIN. Therefore, tumours without CIN were used as reference in the survival analysis comparing MFS across dominant IMFs (Extended Data Table 1).

Furthermore, survival analyses were also performed based on the activities of individual signatures within each tumour. Two-sided Cox proportional hazard models were performed for the whole cohort, and after stratifying samples by the presence or absence of CIN. Cox proportional hazard models were thus used to evaluate the prediction capacity of each mutational signature for MFS (Supplementary Fig. 26). In this case, the study site (country) of origin was also included as covariate to avoid putative batch effects. To ensure robust associations, we filtered out signatures active in less than five samples and only included samples with a minimum number of somatic alterations per type (≥100 SNVs for SBS signature association analyses, ≥10 indels for ID signatures, ≥5 SVs for cSV signatures and ≥20 CNAs for CX signatures).

Associations with clinical variables. To test for associations between signatures and clinical variables, we applied a Welch's *t*-test to compare signature activities between samples that did and did not acquire metastasis, as well as between samples with low (≤2) and high (>2) Gleason grade groups (Supplementary Fig. 27) and study site of tumour collection (Supplementary Fig. 30). *P* values were corrected for multiple testing using the Benjamini–Hochberg method, while Cohen's *d* was computed to assess the size of the difference between groups. Comparison analyses were performed for the whole cohort and after stratifying samples based on the presence of CIN. To ensure robust associations, we filtered out signatures active in less than 5 samples and only included samples with a minimum number of somatic alterations per type (≥100 SNVs for SBS signature association analyses, ≥5 indels for ID signatures, ≥5 SVs for cSV signatures and ≥20 CNAs for CX signatures).

Replication of findings in the TCGA-PRAD cohort. We downloaded both raw and processed WGS data, as well as clinical data, from the TCGA-PRAD cohort (May 2025, https://portal.gdc.cancer.gov/). After excluding samples with fewer than 100 SNVs, fewer than 10 indels and

fewer than 5 SVs, the final dataset comprised 290 TCGA-PRAD samples, of which 232 (80%) exhibited high levels of CIN (≥20 CNAs).

We quantified activities of the new SBS96D and ID83 signatures extracted in the PPCG cohort in the TCGA-PRAD dataset. Consensus calls for SNVs and indels were used to quantify activities of SBS and ID signatures (see the 'De novo extraction and assignment of mutational signatures' section). Replication timing ratio was calculated in the cohort as described above.

The clinical history of each patient was downloaded from the GDC data portal in the form of an XML file, from which we collected overall survival, date of last follow-up, date of biochemical recurrence, tumour stage, Gleason grade, age at diagnosis and PSA. The time to biochemical recurrence was calculated as the time from diagnosis to the biochemical recurrence or death. The time to biochemical recurrence was censored if they were calculated from the last follow-up or treatment end date.

Kaplan–Meier estimates and two-sided Cox proportional hazard models were used for survival analysis across subgroups. Cox proportional hazard models were corrected by Gleason grade. Age at diagnosis was not included as covariate as most TCGA-PRAD samples were from individuals aged over 55 years, and we therefore violated the proportional hazards assumption.

Identifying and benchmarking with simple rearrangement signatures. Simple SV-size-based rearrangement signatures were identified using Palimpsest v.2.0.0 with the default parameters, using PPCG SV calls from the PPCG cohort. Ten simple rearrangement signatures were identified. The six cSV signatures were compared pairwise to the ten simple rearrangement signatures using Spearman correlation.

Demonstrating clinical utility in the HMF-Prostate cohort. We investigated the potential of our integrated mutational footprints as biomarkers for predicting resistance and/or sensitivity to different common therapies for prostate cancer. To achieve this, we used the Hartwig Medical Foundation (HMF) cohort—a real-world retrospective cohort of treated patients with metastatic prostate cancer[59].

Processing genomic data from the HMF-Prostate cohort

Before testing the clinical performance of IMF-based biomarkers, we downloaded and processed genomic data from this cohort to quantify mutational signatures. IMF activities were next quantified to assign the dominant mutational process, as previously described. Supplementary Table 3 shows the number and frequency of samples assigned to each dominant IMF.

SBS and ID signatures. We downloaded SNVs and indels from HMF, which were called using an in-house pipeline previously described[97]. Each sample VCF file was input to the SigProfilerAssignment tool, which generated the sample-by-mutation type matrix to quantify SBS and ID signatures. We limited the signature set to the SBS and ID signatures extracted in the PPCG cohort.

cSV signatures. We implemented a workflow to systematically analyse cSV signatures in prostate cancer samples, using variant calls from Purple and GRIDSS provided by the HMF pipeline v.5. Each sample aliquot was processed independently; for each aliquot, we extracted purity and ploidy values from the Purple estimates ([sample_id].purple.purity.tsv).

We retained only high-confidence GRIDSS variants (PASS), excluding variants overlapping centromeres or telomeres. SCNA and SV calls were used to extract cSV signatures.

CX signatures. We downloaded PURPLE-derived copy-number profiles from HMF, and processed them as previously described[60]. We then computed the genome-wide distributions of five fundamental copy-number features for each sample, quantified activities of CX signatures extracted in the PPCG cohort using the linear combination decomposition function from YAPSA[75], and applying the signature-specific threshold for shrinking to zero low-level activities (see the 'CX signatures' section for further details).

Emulating phase III randomized control trials for therapy response prediction

We aimed to emulate phase III randomized controlled biomarker trials, classifying patients as biomarker-positive or biomarker-negative and assigning them to either the therapy of interest (experimental arm) or an alternative standard-of-care (SoC; control arm). For each IMF, patients were classified as biomarker positive if their IMF activity exceeded the median among cases with non-zero activity; otherwise, they were classified as biomarker negative. Only patients with clinical response data (enabling calculation of TTF) and with sufficiently high-quality genomic data to compute all four mutational signature types were included ($n = 240$). TTF was computed as previously described[60].

Clinical response data covered the full treatment history, although records before enrolment were acknowledged to be less accurate and comprehensive than post-enrolment information. We evaluated the predictive power of the IMFs in the second-line mCRPC setting, in which patients commonly transition empirically between ARPIs and taxanes following first-line therapy. Our analysis focused on these two common therapy types, aiming to move beyond the conventional sequential switching paradigm toward a precision, biomarker-guided approach to therapy selection at second line. No significant differences in TTF were observed between these two therapies (Extended Data Fig. 9), supporting the use of biomarker-positive versus biomarker-negative comparisons to evaluate predictive capacity in this setting.

To decide which therapy–biomarker combinations to analyse, we performed power analyses in line with the Consolidated Standards of Reporting Trials (CONSORT) statement to ensure sufficient statistical power for clinical assessment. To identify the required cohort size to have sufficient power, we performed one-tailed power calculations ($\beta = 0.8$, $\alpha = 0.05$) using censoring and prediction ratio data from the cohort and fixing the hazard ratio at 5 for testing resistance and 1/5 for testing sensitivity (based on previous knowledge[60]). Supplementary Table 4 summarizes the results of the power analyses.

We then compared ARPI-treated patients and taxane-treated in the second-line setting using two-sided Cox proportional hazards models for both biomarker-positive and biomarker-negative cases, with TTF as the primary end point (function coxph from the survival package in R). Cox proportional hazards models were stratified by treatment therapy at first line to control for potential confounding effects of treatment sequencing. This ensures that TTF differences are not attributable to switches of the therapy type at the second line. No information of Gleason grade and tumour stage was available in this cohort. Kaplan–Meier survival curves (survfit function from the survival package in R) were generated to represent differences in treatment effectiveness across treatment arms in a univariate mode. Supplementary Table 5 shows the results of the survival analyses.

Reporting summary

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

Data availability

The PPCG dataset comprises WGS, RNA-seq, DNA methylation, clinical and histopathological data from 2,021 prostate cancer donors. Owing to the sensitive nature of human genomic and 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) under the study ID

EGAS00001002876 and are described in the Reporting Summary and our companion paper[24]. Access to raw, processed and clinical data from the Hartwig Medical Foundation can be requested at https://www.hartwigmedicalfoundation.nl/en/data/data-access-request/. XML files used to construct the TCGA clinical data can be accessed through the Genomic Data Commons portal, and the procedure for requesting access to controlled genomic data is outlined online (https://gdc.cancer.gov/access-data/obtaining-access-controlled-data). 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/mutational-signatures). A static code repository is available at Zenodo[98] (https://doi.org/10.5281/ZENODO.18754193).

68. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
69. Isidro, C.-C., Xi, R. & Park, P. J. ShatterSeek: an R package for the detection of chromothripsis from next-generation sequencing (NGS) data (2022).
70. Xing, R. et al. Whole-genome sequencing reveals novel tandem-duplication hotspots and a prognostic mutational signature in gastric cancer. *Nat. Commun.* **10**, 2037 (2019).
71. Menghi, F. et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl Acad. Sci. USA* **113**, E2373–E2382 (2016).
72. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom.* **2**, 100179 (2022).
73. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinform.* **11**, 367 (2010).
74. Shinde, J. et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics* **34**, 3380–3381 (2018).
75. Hübschmann, D. et al. Analysis of mutational signatures with yet another package for signature analysis. *Genes Chromosomes Cancer* **60**, 314–331 (2021).
76. Akamine, Y. et al. Application of hierarchical clustering to multi-parametric MR in prostate: differentiation of tumor and normal tissue with high accuracy. *Magn. Reson. Imaging* **74**, 90–95 (2020).
77. Strauss, T. & von Maltitz, M. J. Generalising Ward's method for use with Manhattan distances. *PLoS ONE* **12**, e0168288 (2017).
78. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385 (2018).
79. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
80. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
81. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
82. Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
83. Takeda, D. Y. et al. A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell* **174**, 422–432 (2018).
84. Boersma, V. et al. MAD2L2 controls DNA repair at telomeres and DNA breaks by inhibiting 5′ end resection. *Nature* **521**, 537–540 (2015).
85. Isono, M. et al. *BRCA1* directs the repair pathway to homologous recombination by promoting 53BP1 dephosphorylation. *Cell Rep.* **18**, 520–532 (2017).
86. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).
87. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
88. Poulsgaard, G. A., Sørensen, S. G., Juul, R. I., Nielsen, M. M. & Pedersen, J. S. Sequence dependencies and mutation rates of localized mutational processes in cancer. *Genome Med.* **15**, 63 (2023).
89. Adler, N. et al. Mutational processes of tobacco smoking and APOBEC activity generate protein-truncating mutations in cancer genomes. *Sci. Adv.* **9**, eadh3083 (2023).
90. Medo, M., Ng, C. K. Y. & Medová, M. A comprehensive comparison of tools for fitting mutational signatures. *Nat. Commun.* **15**, 9467 (2024).
91. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
92. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, 6413 (2018).
93. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
94. Therneau, T. M. Survival: survival analysis https://doi.org/10.32614/CRAN.package.survival (2023).
95. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S. Drawing survival curves using 'ggplot2'. (2021).
96. McGranahan, N., Burrell, R. A., Endesfelder, D., Novelli, M. R. & Swanton, C. Cancer chromosomal instability: therapeutic and diagnostic challenges. *EMBO Rep.* **13**, 528–538 (2012).
97. Martínez-Jiménez, F. et al. Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
98. Gruber, A., Vidas Olsen, A., Hernando, B. & Cheng, K. Integrated signatures define mutational processes in prostate cancer. *Zenodo* https://doi.org/10.5281/ZENODO.18754193 (2026).

Acknowledgements This work was performed as part of the Pan-Prostate Cancer Group (PPCG). We thank the patients, staff and scientists who contributed to this study; Y. Li for providing feedback and discussions on cSVs; R. Houlston and R. Culliford for supporting us with information about an external prostate cancer cohort; and those men with prostate cancer and the individuals who have donated their time and samples to the datasets used in this study. J.W. was supported by the Danish Cancer Society R147-A9843, Danish Cancer Society R374-A22518, Danish Council for Independent Research 8020-00282, Danish Council for Independent Research 3101-00177A, Novo Nordisk Foundation NNF20OC0060141 and Sygeforsikringen Danmark 2022-0198; J.R. was supported by the Canadian Institutes of Health Research (CIHR) Project Grant (PJT-162410) and the Investigator Award from the Ontario Institute for Cancer Research (OICR); funding to OICR is provided by the Government of Ontario. A.J.G. was supported by the University of Konstanz; G.M. and B.H. are 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 (CEX2024-001442-S) by the Spanish Ministry of Science and Innovation (MCIN/AEI/ 10.13039/501100011033). G.M. and B.H. were 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 (100010434; LCF/BQ/PR23/11980033). This work is also supported by a Cancer Research UK RadNet Manchester award (C1994/A28701), a CRUK Manchester Institute Core Senior Investigator award (C5759/A27412) and a Prostate Cancer UK FASTMAN Centre of Excellence award (MA-COE128-002). Part of the research presented in this Article 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; Prostate Cancer UK (MA-TIA23-002, MA-ETNA19-003, TLD-CAF22-011); the Dallaglio Foundation; Prostate Cancer Research; the Bob Champion Cancer Research Trust; Movember; Big C Cancer Charity; the Masonic Charitable Foundation; the King Family; the Stephen Hargrave Foundation; the Alan Boswell Group; The Bob Willis Fund; and the Bedford Memorial Trust. This work was also supported in part by grants awarded to K.D.S. from The Novo Nordisk Foundation (NNF20OC0059410, NNF21OC0071712), The Danish Cancer Society (R352-A20573) and Independent Research Fund Denmark (9039-00084B). B.P. was supported by a Victorian Health and Medical Research Fellowship from the Victoria State Government, Australia. This research was also supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. 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). This work was also supported by NHMRC project grants 1104010 (C.M.H., T.C., N.M.C.) and 1047581 (C.M.H., T.C., N.M.C.), as well as a federal grant from the Australian Department of Health and Ageing to the Epworth Cancer Centre, Epworth Hospital (T.C., N.M.C., C.M.H.). In carrying out this research, we received funding and support from Australian Prostate Cancer Research and the University of Melbourne, Australia. Genomic sequencing and interrogation of Southern African Prostate Cancer Study (SAPCS) data were 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 USA 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 US National Institute of Health (NIH) National Cancer Institute (NCI) Award (1R01CA285772-01) and a US Prostate Cancer Foundation (PCF) Challenge Award (2023CHAL4150). V.M.H. was further supported by the Petre Foundation through the University of Sydney Foundation (Australia). V.J.G. acknowledges infrastructure support from the NIHR Cambridge Biomedical Research Centre (BRC-1215–20014). The French Prostate ICGC project was funded by Institut National de la Santé et de la Recherche Médicale (INSERM) and Institut National du Cancer (INCa); grant INSERM CV_2011/023 (C18), with additional support from LYric (grant INCa-4662). Analysis of the Australian samples was funded through a PRECEPT programme grant, co-funded by Movember and the Australian Federal Government. We also acknowledge support from Prostate Cancer Research, Prostate Cancer UK, the Bob Champion Cancer Research Trust, Movember, Big C Cancer Charity, the Masonic Charitable Foundation, the Prostate Cancer Research Centre, the King Family, The Andy Ripley Memorial Fund and the Stephen Hargrave Foundation, the NIHR support to The Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. D.K. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement number 101034290 (EMERALD International PhD Programme for Medical Doctors). The study was funded by the ICGC EOPC (01KU1001 A) and ICGC-Data Mining (01KU1505A) projects on early-onset prostate cancer by the German Federal Ministry of Education and Research (BMBF). We acknowledge the DKFZ, MPI and EMBL Genomics Core Facilities.

Author contributions A.J.G., A.V.O., B.H., K.C.L.C., C. Gerhäuser, M.T., F. Favero, D.K., A.F.-S., J.M.R.-R. and L.B. contributed to the formal analysis presented in this study. J.W., A.J.G., A.V.O., B.H., K.C.L.C., T.S., G.M. and J.R. wrote the manuscript. A.J.G., A.V.O., B.H. and K.C.L.C. produced and contributed to the visualizations of the study. J.W. conceived and designed the study. J.W., J.R. and G.M. 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 organization and conduct of the study and critiqued the output for important intellectual content. R. Simon and G.S. contributed samples and pathology workup. A.J.G., A.V.O., B.H., K.C.L.C., C. Gerhäuser, M.T., F. Favero, D.K., A.F.-S., J.M.R-R., L.B.,

D.P., G.S.B., D.S.B., M.N.B, B.B., A. Butler, G.C.-T., N.M.C., O.C., A.G., E.G.G., V.J.G., A.A.H., V.M.H., H.H.H., C.M.H., E.L.I., G.M.M.J., C.-h.J., F.K., Z.K.-J., P. Lamy, G.L., M.L., P. Lutsik, L.M., R.M., A.T.P., B.P., L.R.Q., T.R., J.K., B.R., A. Sahli, K.D.S., T.N.Y., S.U., Y.X., C.Z., N.A.-M., E.B., D.B., J.B., A. Bergeron, A. Berlin, M.B., W.B., B.K.C., R.C., M. Cheung, K.C., M.J.C., M. Cmero, C.C., A.C.F., T.C., N.D., H.D., W.D., R.D., P.D., C.F., A.E., E.E., F. Falkenbach, G.N.F., B.F., L.F., Y. Fradet, Y. Fu, K.F., P.G., M.G., C. Grima, V. Haberland, G.A.H., X.H., J.H., V. Holmes, C.H., M.K.H.H., W.J.H., W.J., J.J., T.K., M.K., G.K., N.K., R.L., A.D.L., B.L., Y.-J.L., A.G.L., S.M., C.E.M., P.M., A.M., N.M., I.M., T.M.-C., A.N., S.N., P.V.N., M.N., S.O., M.O., H.P., Y.P., D.J.P., F.P., J.P., J.S.P., P.M.P., P.P., C.P., K.R., S.R., R.R.P., A.F.R., A.R., A. Sankar, B.C.S., R. Stuchbery, H.S., P.T., J.U., R.R.N.B., M.S., S.P., P.V.L., L.V.v.V., M.W., A.W., C.W., D.J.W., J.Z., A.S.Z., R. Simon, G.S., R.A.E., C.S.C., R.G.B., D.C.W., T.S., G.M., J.R. and J.W. provided access to data and/or contributed to gathering, processing and curating data. All of the authors had access to all data in the study. All of the authors contributed to the review and the editing of the manuscript. All of the authors approved the manuscript before submission.

Competing interests V.J.G. has received honoraria from Janssen and JEB as a speaker; and is a member of an external expert committee to AstraZeneca, UK. R.A.E. has received honoraria from GU-ASCO, Janssen, University of Chicago and Dana Farber-Cancer Institute as a speaker; has received educational honorarium from Bayer and Ipsen; is a member of an external expert committee to AstraZeneca, UK; is a member of active surveillance for the Movember Committee; is a member of the scientific advisory board of Our Future Health; and undertakes private practice as a sole trader. G.M. is co-founder, director and shareholder of Tailor Bio. The other authors declare no competing interests.

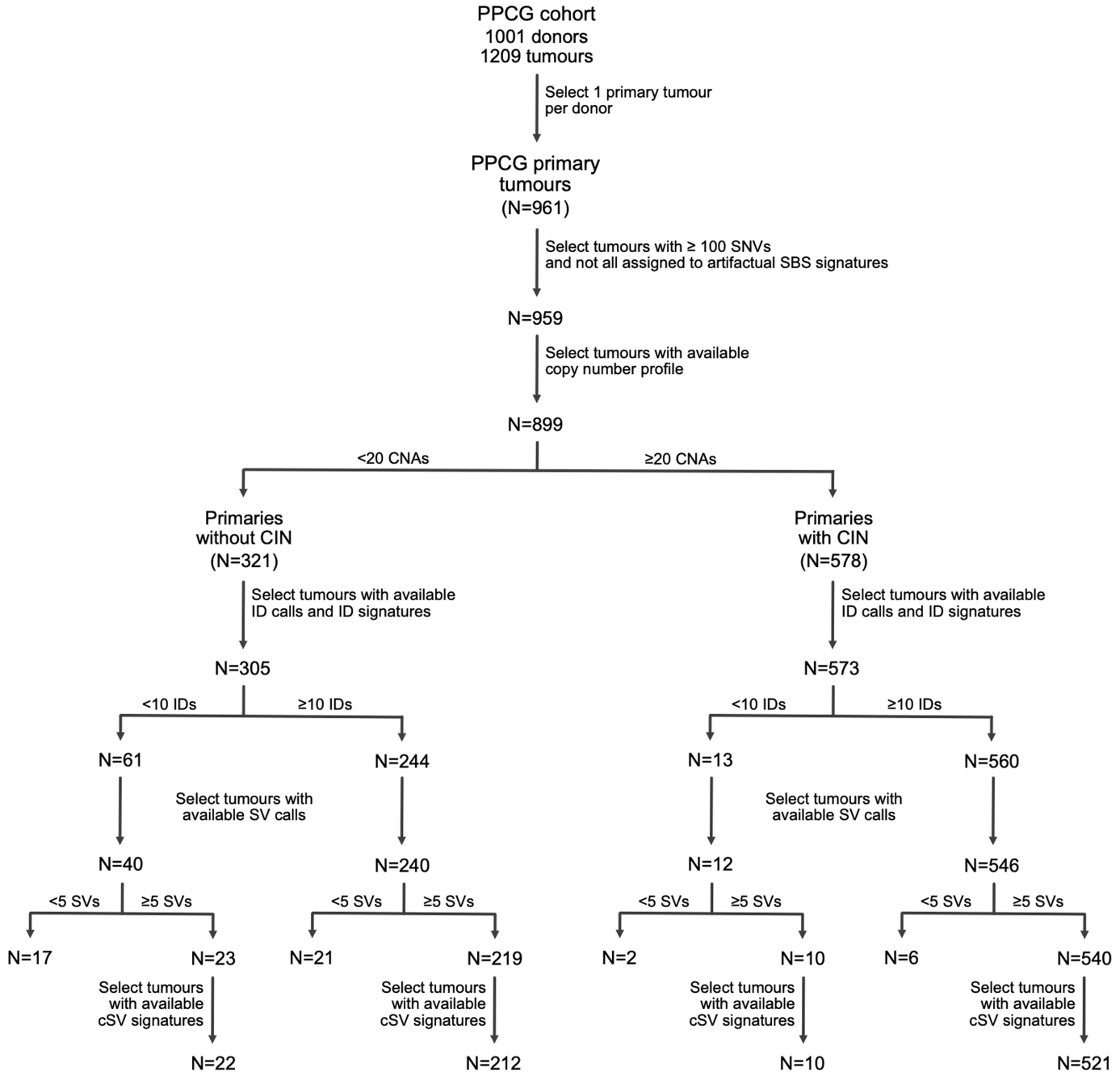
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10468-w.

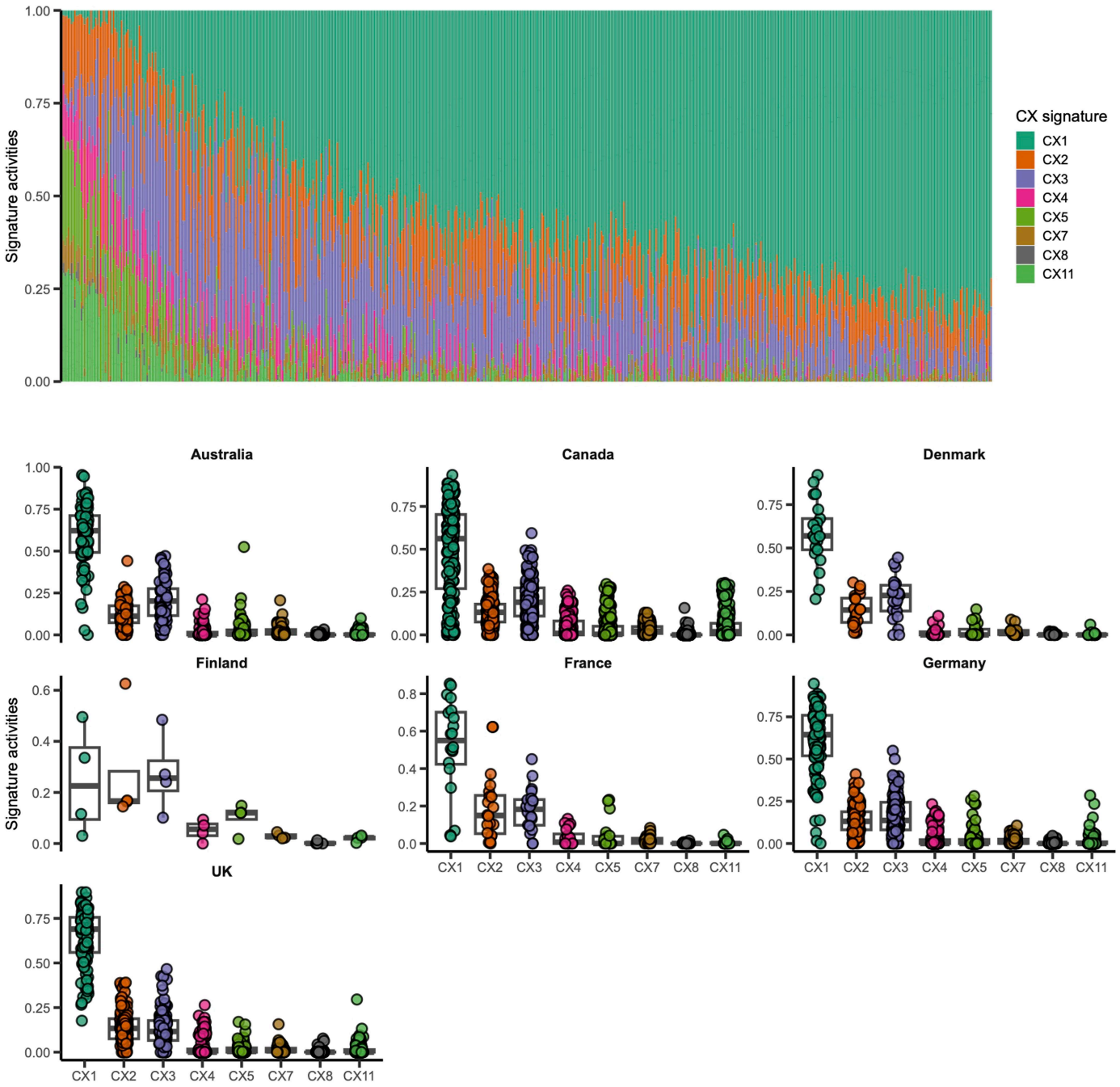
Correspondence and requests for materials should be addressed to Geoff Macintyre, Jüri Reimand or Joachim Weischenfeldt.

Peer review information *Nature* thanks Christopher Barbieri, who co-reviewed with Andrea Sboner, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

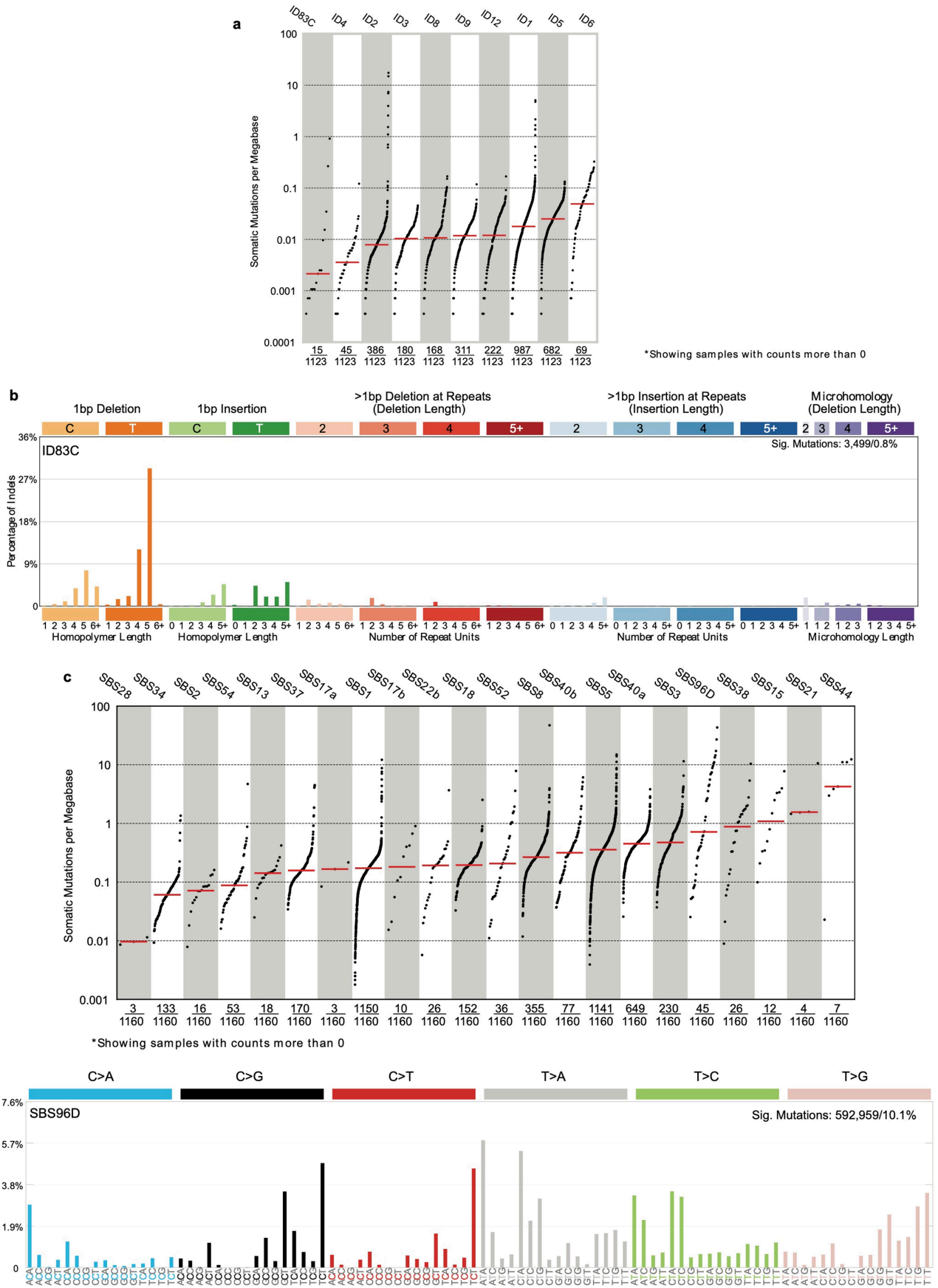
Reprints and permissions information is available at http://www.nature.com/reprints.



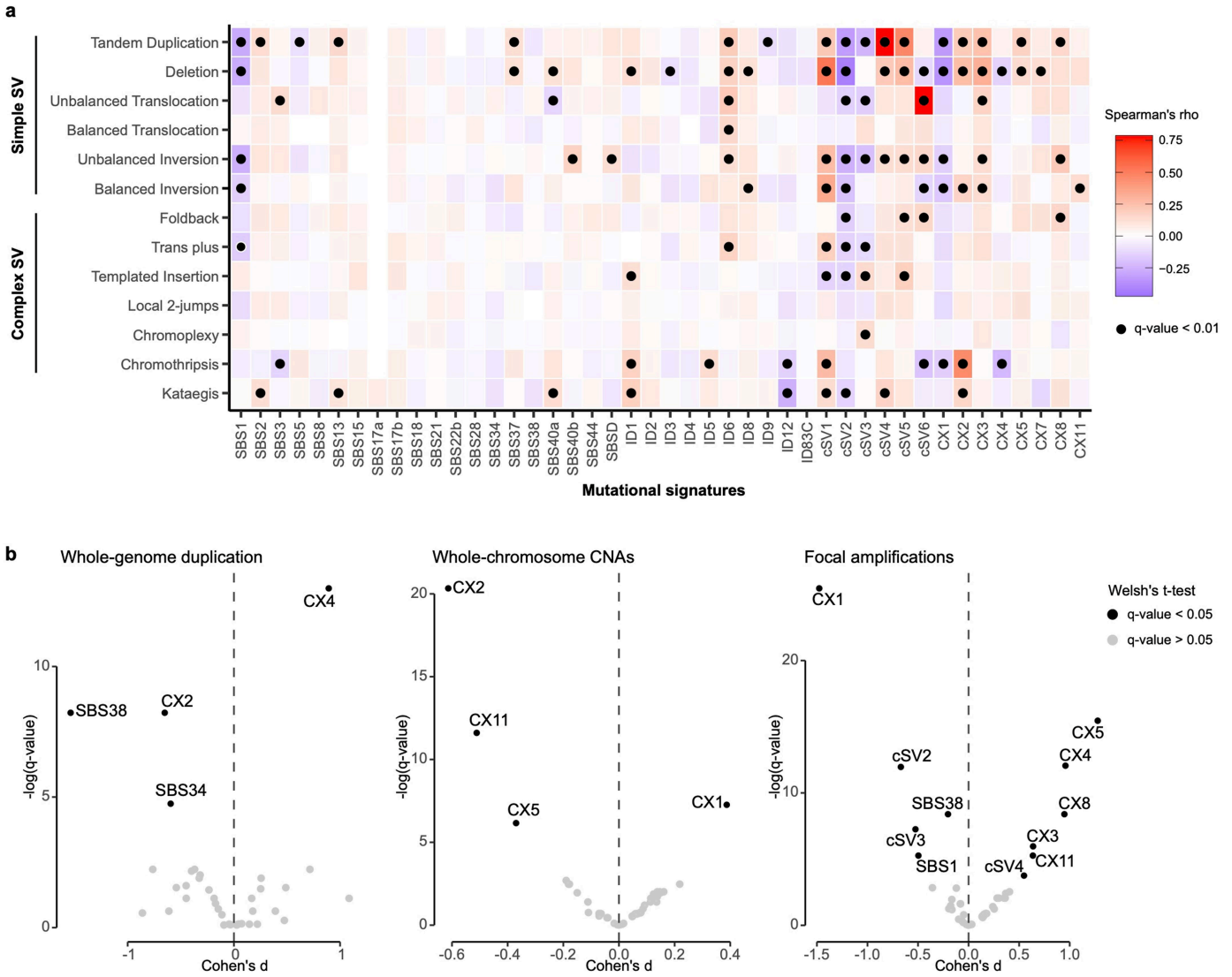
Extended Data Fig. 1 | Remark diagram.



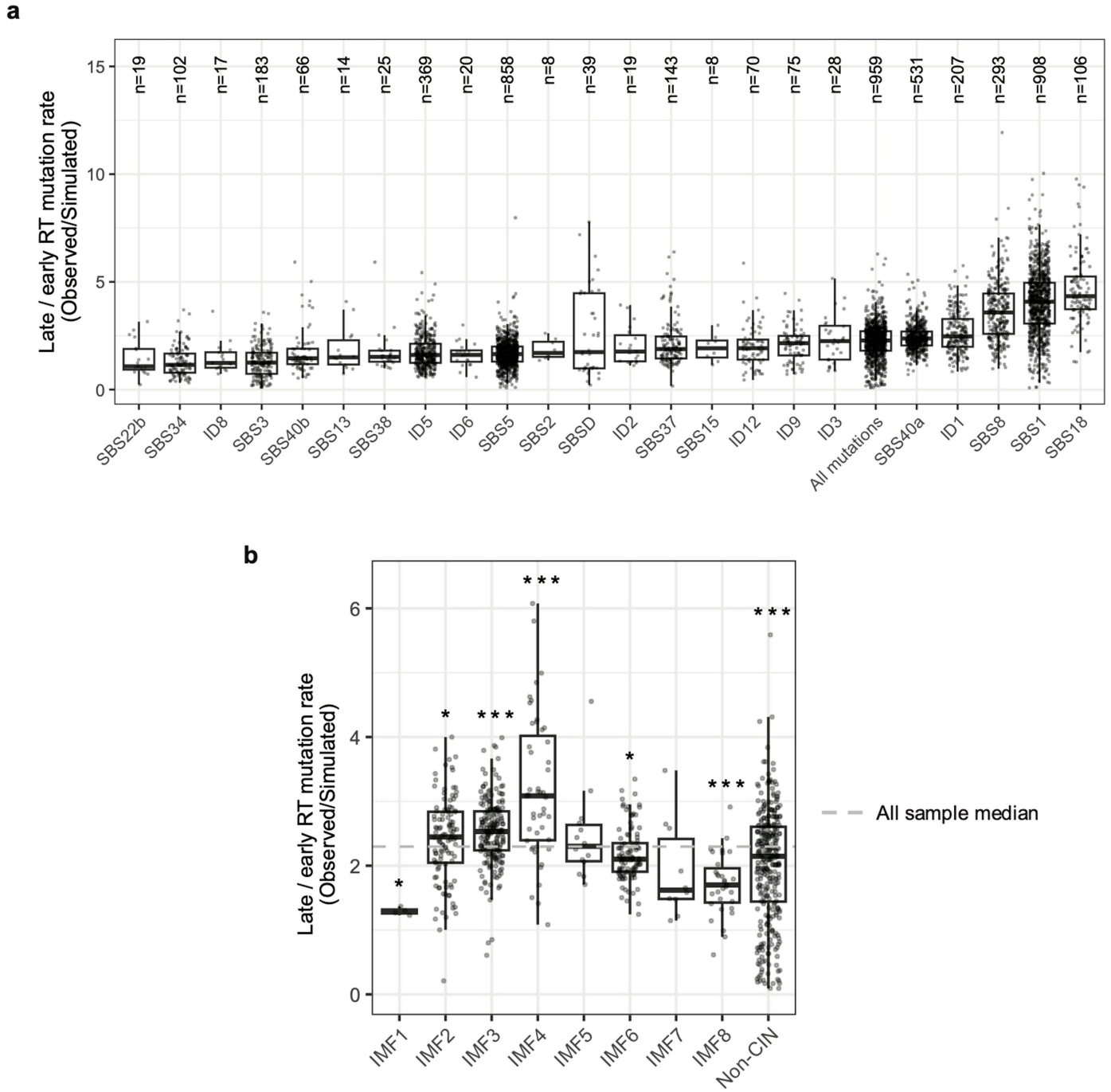
Extended Data Fig. 2 | Overview of extracted copy number signatures. **Top:** barplot indicates distribution of CX signature activities across the whole cohort (n = 578). Samples are ordered by decreasing the number of CNAs from left to right. **Bottom:** boxplots denote the distribution of signature activities across samples collected from a specific country (Australia: n = 75; Canada: n = 222; Denmark: n = 21; Finland: n = 4; France: n = 20; Germany: n = 112; UK: n = 124). Boxplot boxes indicate first quartile, median and third quartile. Whiskers indicate the first or third quartile ± 1.5 times the interquartile range.



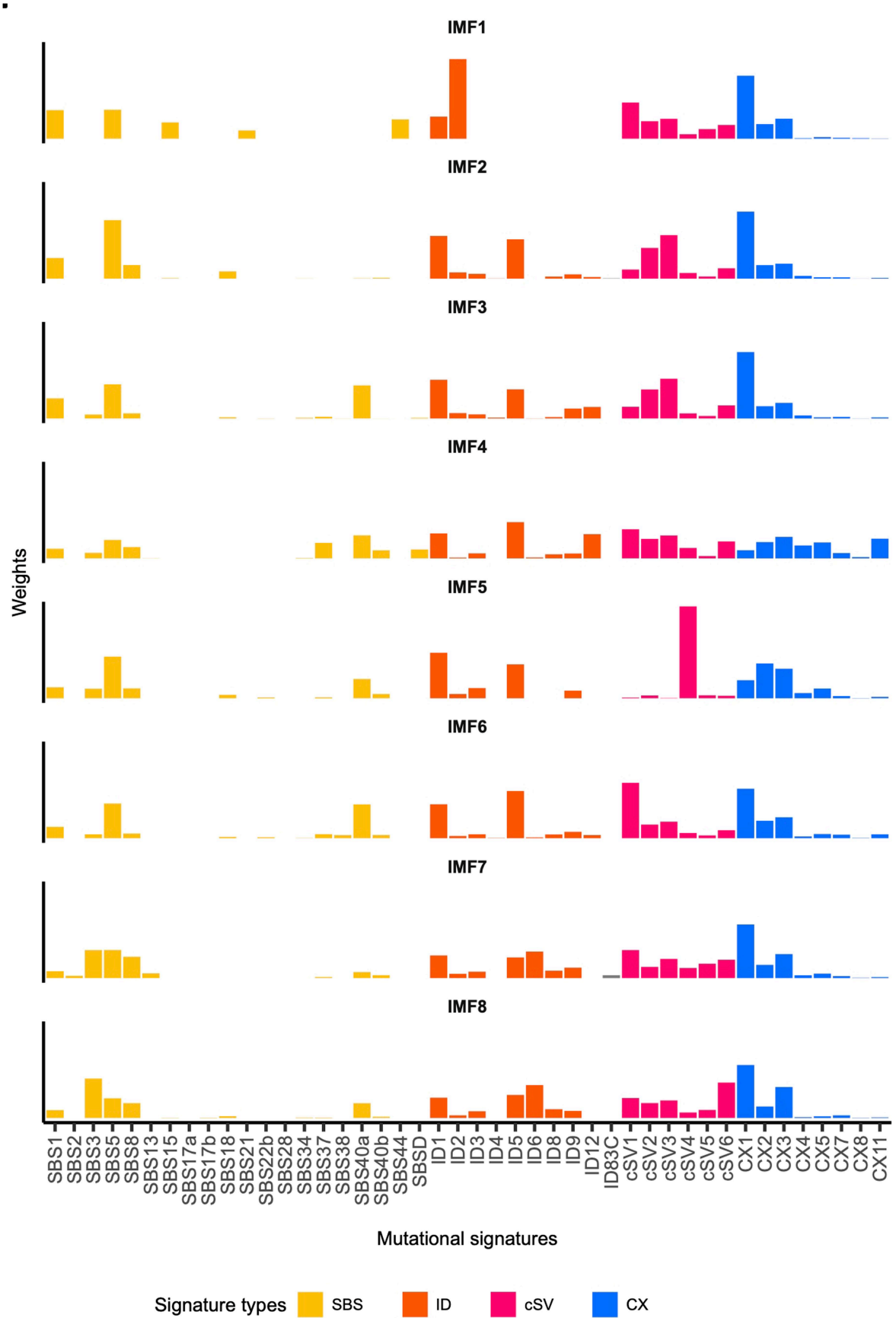
Extended Data Fig. 3 | Overview of extracted indel (ID) and single base substitution (SBS) signatures. a) Distribution of the number of somatic insertions and deletions per megabase assigned to each ID signature per sample (n = 1,123). **b**) Barplot showing the novel signature ID83C (also referred to as IDC). **c**) Distribution of the number of somatic single base substitutions per megabase assigned to each SBS signature per sample (n = 1,160). **d**) Barplot showing the novel signature SBS96D (also referred to as SBSD).



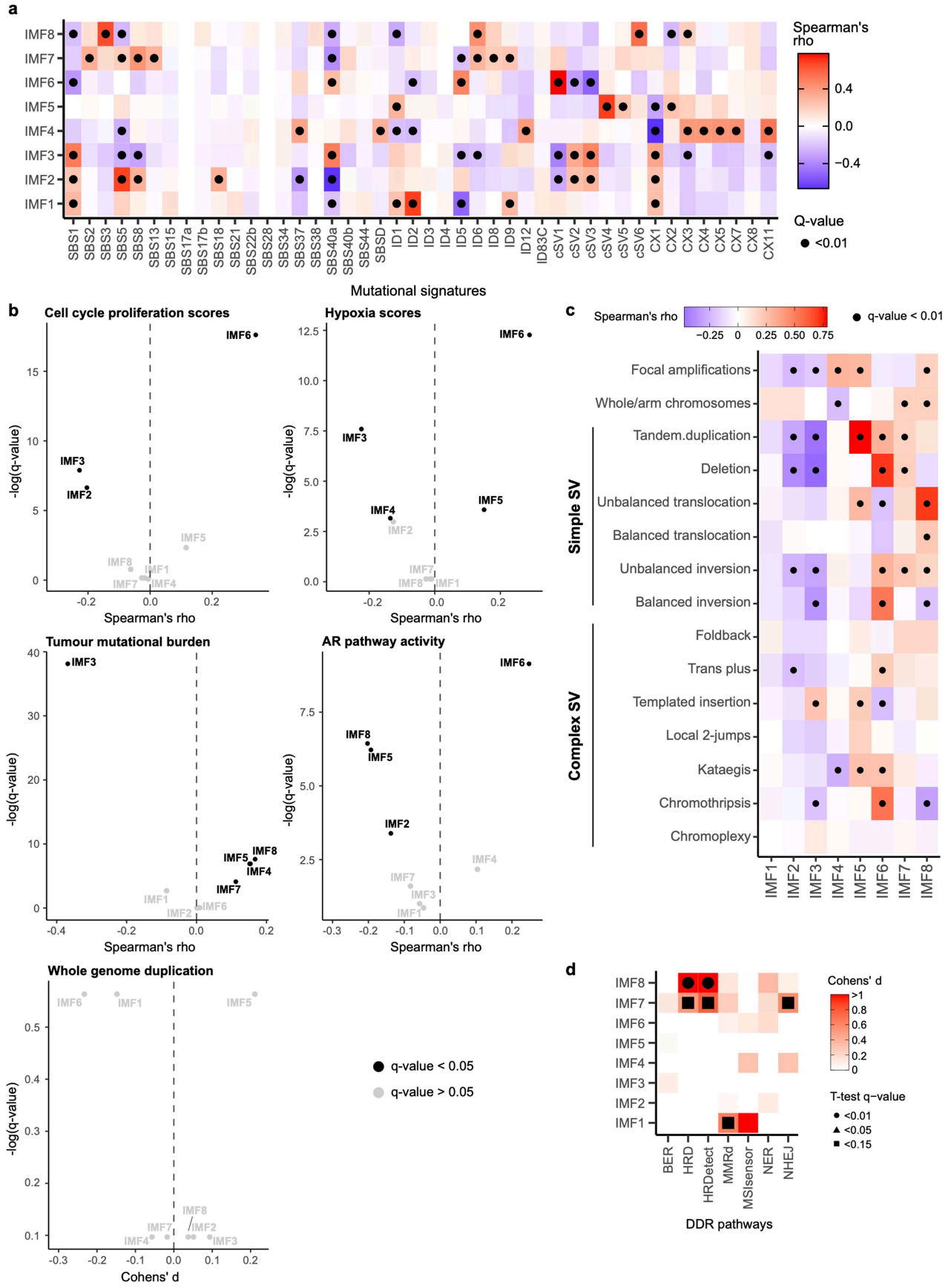
Extended Data Fig. 4 | Associations of mutational signatures with genomic features. a) Heatmap showing significant associations between signature activities and the presence of simple and complex SVs, and kataegis (n = 959). **b**) Scatter plots showing strength of differences (Cohen's d) in signature activities of tumour samples with and without whole genome duplication (right), whole chromosome CNAs (middle), and focal amplifications (left) (n = 959). P-values were corrected for multiple testing using the Benjamini-Hochberg method. Black dots represent significant correlations.



Extended Data Fig. 5 | Association of mutational signatures and replication timing. a) Late-to-early mutation rate ratios are calculated for each signature in each sample. To prevent unreliable estimation of replication timing (RT) mutation rate, samples are required to have at least 10 mutations in both early and late regions of the particular signature to be included. Only signatures with at least 5 qualified samples are shown. **b**) Comparison of mutation rate RT association in IMF. Box plots show late to early mutation rate ratio of each sample. Dotted line denotes the median rate ratio of all samples. Asterisks indicate significant differences between an IMF and other groups by two-sided Wilcoxon tests. ***FDR < 0.001, **FDR < 0.01, *FDR < 0.05. Sample sizes for panel b: IMF1 = 4, IMF2 = 113, IMF3 = 201 250, IMF4 = 47, IMF5 = 14, IMF6 = 101 110, IMF7 = 10, IMF8 = 31, non-CIN = 323. Boxplots boxes indicate first quartile, median and third quartile. Whiskers indicate the first or third quartile ±1.5 times the interquartile range.

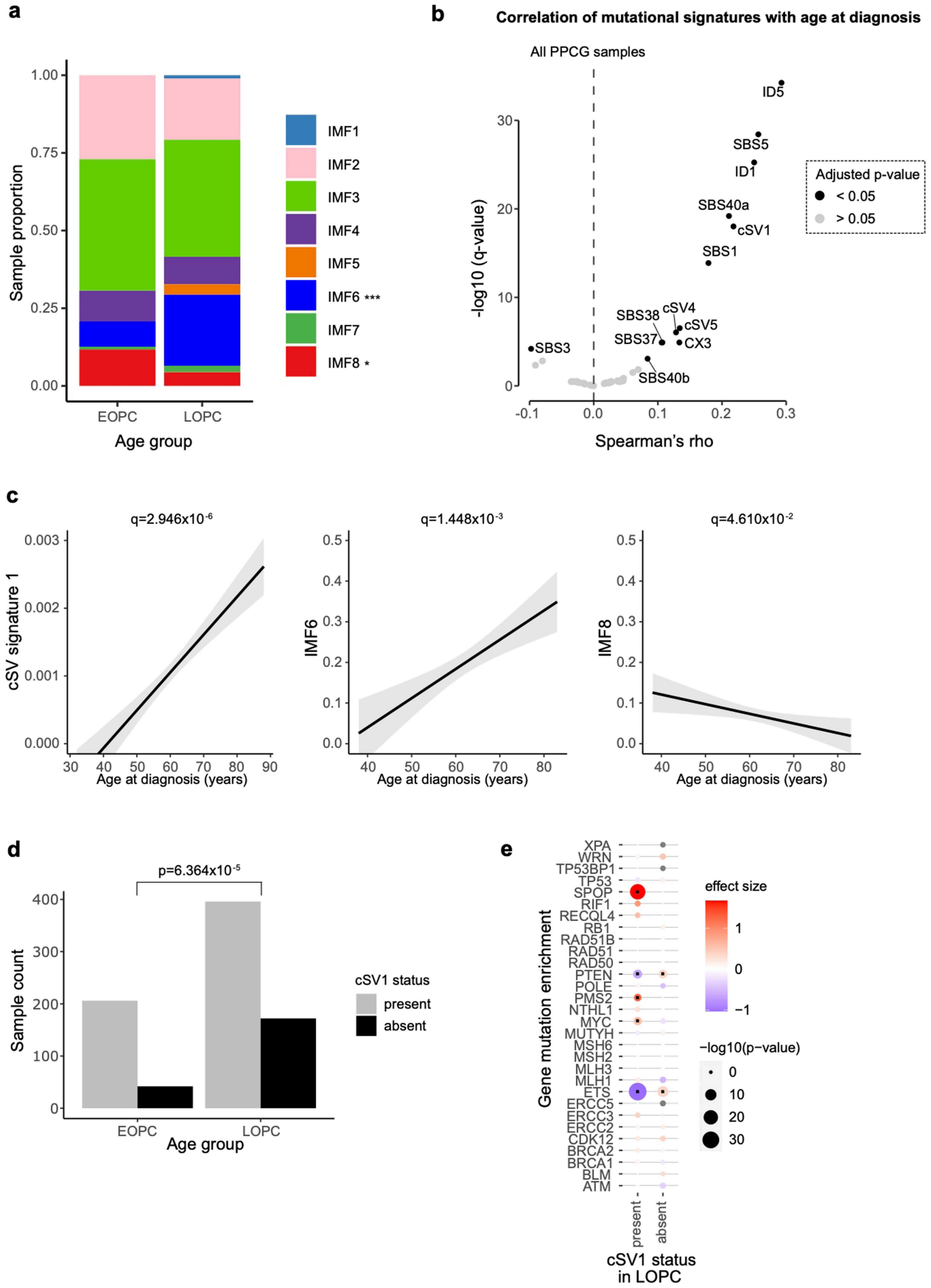


Extended Data Fig. 6 | Definitions of integrated mutational footprints. Barplot displaying the weights of each of the 44 mutational signatures per integrated mutational footprint (IMF) extracted in the PPCG prostate cancer cohort. Weights are normalized per signature type to enable comparison across the contribution of signatures per signature type.



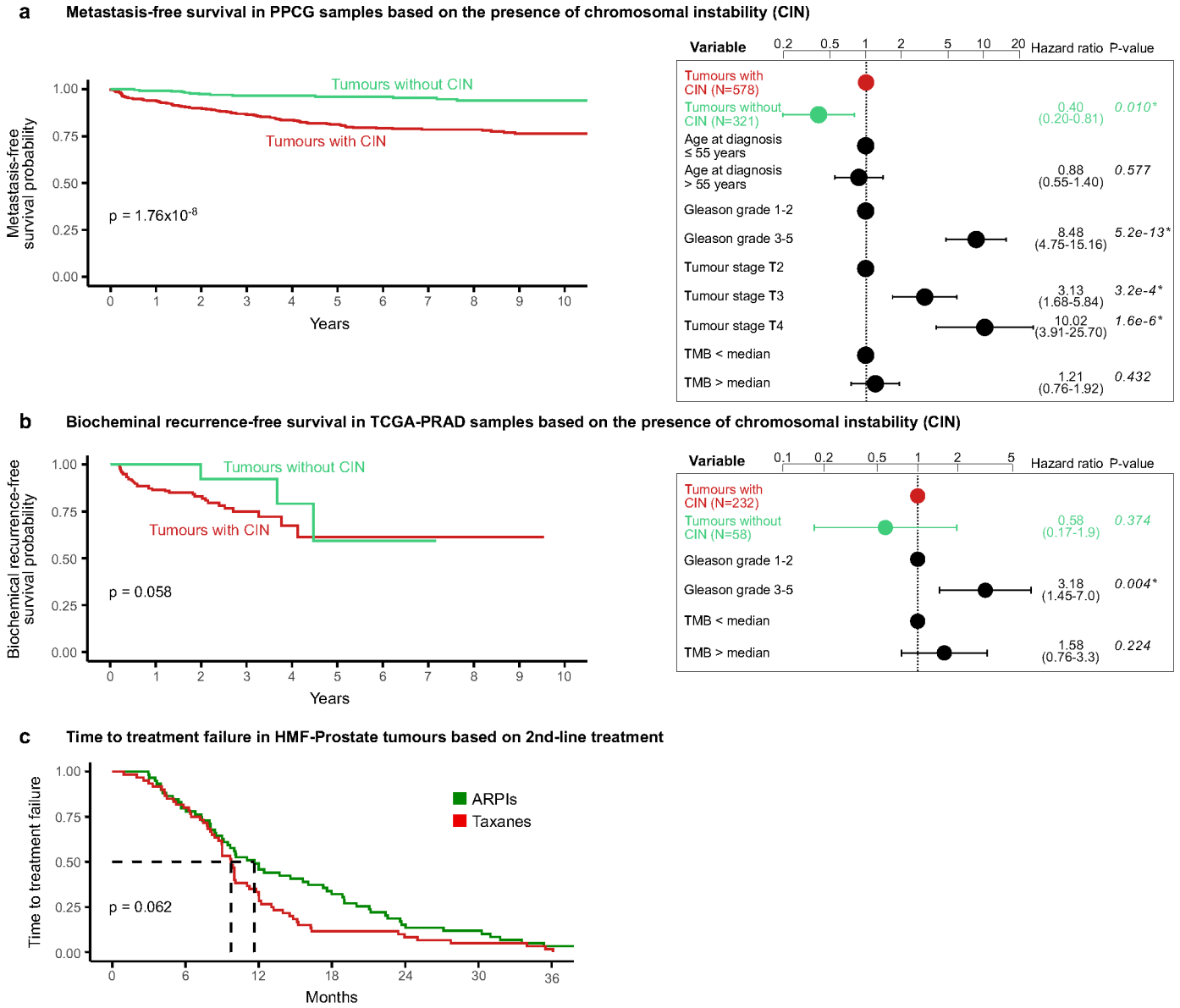
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Association of integrated mutational footprints with genomic features. a) Heatmap showing the strength of correlation (Spearman's rho) between activities of integrated mutational footprints (IMF) and individual mutational signatures. P-values were corrected for multiple testing using the Benjamini-Hochberg method. Black dots represent significant correlations. **b**) Scatter plots showing strength of correlation of IMF activities with cell cycle proliferation scores, hypoxia scores, tumour mutational burden, AR pathway activity, and the presence of whole genome duplication. Associations were evaluated using Spearman's correlations, except for WGD status, where a two-sided Welch's t-test was used to compare group means. P-values were corrected for multiple testing using the Benjamini-Hochberg method. Black dots represent significant correlations. **c**) Heatmap showing significant correlations between IMF activities and the number of simple and complex SVs, kataegis, whole/arm-level chromosome alterations, and focal amplifications. **d**) Heatmap showing the strength of correlation (cohens' d) between IMF activities and the presence of deficiency in DNA damage repair (DDR) pathways. A two-sided Welch's t-test was applied to compare means across groups. P-values were corrected for multiple testing using the Benjamini-Hochberg method. Non-significant results are represented as white fields. Panels a-d: n = 521 tumour samples.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Age-related impact of mutational processes on prostate cancer. a) Barplot showing proportions of samples included in the eight IMFs according to age at diagnosis. Fisher's exact test was applied to compare the distribution between early-onset (EOPC) and late-onset prostate cancers (LOPC). FDR-corrected p-values: IMF1 = 0.710, IMF2 = 0.229, IMF3 = 0.605, IMF4 = 0.710, IMF5 = 0.221, IMF6 = 2.86×10^{-3}, IMF7 = 0.710, IMF8 = 0.042. Asterisk indicates significant differences. **b**) Scatterplots of the Spearman correlation between signature activities and age at diagnosis for all samples with quantified signature activities. P-values were corrected for multiple testing using the Benjamini-Hochberg method. Black dots represent significant correlations. **c**) Multiple linear regression plot showing the relationship between age at diagnosis and cSV1 signature exposure (left), and the relationship between age at diagnosis and IMF6/IMF8 (centre/right). Black line: linear regression line, grey area: 95% confidence band. FDR-corrected p-value (q) obtained from multiple linear regression of cSV signatures against age/IMF. **d**) Barplot showing the number of early-onset (EOPC) and late-onset (LOPC) prostate cancer samples with high cSV1 (upper quartile range) or low cSV1 signature exposure. P-value obtained from Fisher's exact test. **e**) Enrichment analysis of driver mutations in LOPC with the presence or absence of cSV1 signature. P-value obtained from Fisher's exact test. Circle size denotes -log10(p-value), colours indicate effect size (log odds ratio), and black dots denote FDR significance. Gene list contains 45 well-known genes in prostate cancer and only genes that are found in at least one sample in the subsets are depicted. Panels a,c: n = 521 tumour samples (EOPC: n = 248; and LOPC: n = 568). Panel b: n = 959 tumour samples. Panel d-e: n = 248 EOPC and n = 568 LOPC. Fisher's exact tests were two-sided.



Extended Data Fig. 9 | Survival time comparisons in primary and metastatic prostate cancer samples. a) Kaplan-Meier curves (left) and two-sided Cox proportional hazards model (right) showing metastasis-free survival for prostate tumours from PPCG with and without chromosomal instability (CIN). The p-value in Kaplan-Meier curves was estimated by the log-rank test. The Cox proportional hazards model (right) was corrected for the age at diagnosis, Gleason grade group, tumour stage and tumour mutational burden (TBM). Dots represent the mean hazard ratios with bars representing the 95% CI. **b**) Kaplan-Meier curves (left) and two-sided Cox proportional hazards model (right) showing biochemical recurrence-free survival for prostate tumours from TCGA with and without chromosomal instability (CIN). The p-value in Kaplan-Meier curves was estimated by the log-rank test. The Cox proportional hazards model (right) was corrected for the Gleason grade group, and tumour mutational burden (TBM). Tumour stage was not available for this cohort, and most patients were older than 55 years old at diagnosis. Dots represent the mean hazard ratios with bars representing the 95% CI. **c**) Kaplan-Meier curve showing time to treatment failure for HMF metastatic prostate tumours treated with ARPIs (N = 59) and taxanes (N = 60) at second line. The p-value in Kaplan-Meier curves was estimated by the log-rank test.

Extended Data Table 1 | Description of the eight major mutational processes in prostate cancer

IMF	Cases with IMF as dominant	Proposed aetiology	Enriched in cancer genes	Risk of metastasis	Therapy management utility
IMF1	4 (0.1%)	Mismatch repair deficiency	MSH2	Low	
IMF2	113 (21%)	Reactive oxygen species		High	
IMF3	201 (39%)	AR-mediated mutagenesis	ETS, PTEN, TBL1XR1, ZBTB16	Intermediate	
IMF4	47 (9%)	Mitotic defects & replication stress	IRF4, LMO2, CDC16, MZT2A/B, TUBA3D/E, ZNF208	Low	
IMF5	14 (3%)	Non-canonical impaired HR	CDK12, CCND1, MDM4	High	
IMF6	101 (19%)	Replication stress	SPOP, ARC, SOX17, BRAF, HNRNPA2B1, LRP1B, UNCX	Low	Predicts response to ARPIs
IMF7	10 (2%)	APOBEC mutagenesis		Intermediate	
IMF8	31 (6%)	Canonical impaired HR	BRCA2	High	
NoCIN				Low	

Table of the eight major mutational processes in prostate cancer (IMF1-8), including the number of patients with each IMF as dominant, proposed aetiology, enriched cancer genes, metastasis risk, and proposed therapy management in the metastatic setting based on our emulated phase 3 clinical trial in metastatic patients. IMFs are derived from samples with all four mutational signature types quantified (n=521, Supplementary Note 5). IMF, integrated mutational footprints.

nature portfolio

Corresponding author(s): Joachim Weischenfeldt, Jüri Reimand, Geoff Macintyre

Last updated by author(s): Mar 23, 2026

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection: No dedicated software was used for data collection

Data analysis	Data analysis was performed using standard and peer-reviewed computational tools in Python and R, together with specialised bioinformatic software for mutational signature extraction, single nucleotide variant, insertion-deletion, structural variant analysis, and copy-number quantification. Germline and somatic variant calling is described in our companion paper Jakobsdottir et al. All software used in the study is listed below. Programming languages include Python 3.10 and R 4.3.1. Computational tools, algorithms, and packages: SigProfilerExtractor v1.1.23 (SBS and ID de novo mutational signature extraction) SigProfilerMatrixGenerator v1.3.14 SigProfilerAssignment v1.1.0 YAPSA v1.12.0 NMF R package v0.23 (SV signature extraction framework) Palimpsest v2.0.0 (simple SV signatures) Shatterseek v0.4 (chromothripsis detection) ClusterSV (as implemented in Li et al. 2020 pipeline) Battenberg copy-number segmentation v2.2.9 Purple /GRIDSS (HMF pipeline v5) for external metastatic cohort data processing SigProfilerSimulator for replication-timing background corrections singscore v3.21

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

PPCG dataset includes WGS, RNA sequencing, DNA methylation, clinical, and histopathological data from 2,021 prostate cancer donors. Due to the sensitive nature of these human genomic and clinical data and the participant-consent restrictions governing their use, all PPCG data are controlled-access and can be shared only for approved research purposes.
Access to PPCG data requires submission of a Project Concept Form, available at https://panprostate.org, and approval by the PPCG Steering Committee. Completed forms should be emailed to PPCG@icr.ac.uk.
Publicly accessible repositories and accession codes
Aligned whole-genome sequencing BAM files for contributing cohorts are available from the European Genome–Phenome Archive (EGA) under the study ID EGAS00001002876 (https://ega-archive.org/studies/EGAS00001002876)
Controlled-access components:
WGS working project data (variant calls, pipeline outputs, intermediate analyses) are hosted at the Institute of Cancer Research and made available through the Globus data transfer service to approved PPCG members; all files are encrypted prior to upload.
Harmonised RNA-seq and methylation data are similarly available through Globus for approved projects.
Clinical and histopathological data are hosted in REDCap, a GDPR-compliant secure environment at the Institute of Cancer Research; access is granted only to approved PPCG researchers.
Additional external datasets and databases used
The study additionally utilised publicly available external genomic and epigenomic resources, including:
- HMF clinical and raw data: Provided by HMF as part of data access request DR-215
- TCGA raw data: Provided as part of data access project ID 40834
- TCGA somatic variant calls acquired from Genomics Data Commons, https://portal.gdc.cancer.gov/, May 2025
- 1000 Genomes Project haplotype reference panel (for phasing)
- ExAC, gnomAD, dbSNP (for germline and indel filtering)
- COSMIC coding sequence dataset (for BBDuK filtering)
- Gene Expression Omnibus (GEO) accession GSE70079 (Pomerantz et al. 2015)
- ATACseq data (Corces et al. 2018, PMID: 30361341).
- GEO reference methylation datasets: GSE35069, GSE71837, GSE49667, GSE74877, GSE86258, GSE87797 (cell-type references)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex is not relevant for prostate cancer.
Reporting on race, ethnicity, or other socially relevant groupings	The study includes genomic, clinical, and demographic data from 2,021 prostate cancer donors recruited across multiple international cohorts (UK, Germany, Denmark, France, Australia, Canada, USA, and Southern Africa). Donor genetic ancestry was characterised using germline SNP-based ancestry inference across the five 1000 Genomes super-population groups (EUR, AFR, EAS, SAS, AMR) to support analyses of genomic diversity and disease heterogeneity. Ancestry was therefore defined using genetic markers, not self-reported race or ethnicity. This approach avoids conflating social constructs with biological attributes and limits potential misclassification. Ancestry information was used exclusively to examine mutational, genomic, and molecular variation across populations, with the goal of improving biological understanding of prostate cancer in diverse groups. Analyses were performed at the

population level and did not attribute biological differences to socially constructed racial categories.
Overall, socially relevant groupings were included only when scientifically justified, defined using rigorous genomic methods, and reported in a manner that avoids stigmatization or over-interpretation.

Population characteristics

The full PPCG study cohort consists of 2,021 prostate cancer donors enrolled across international PPCG partner institutions. The age at tumour collection ranged from 32 to 90 years, with a median of 62 years. Clinical covariates collected for each donor included pretreatment PSA levels (median 7.97 ng/mL; range 0.40–3,459 ng/mL), clinical and pathological T category, Gleason Grade Group, outcome measures including biochemical relapse, distant metastasis, overall survival, and cancer-specific survival, where available.
Disease severity and pathological features spanned the full clinical spectrum:
- Gleason Grade Group distribution: GG1 (14%), GG2 (37%), GG3 (20%), GG4 (9.2%), GG5 (17%).
- Pathological T categories: pT2 (44%), pT3 (48%), pT4 (1.6%).
- Metastatic status: 9.5% of donors exhibited metastatic disease; 4.2% presented with de novo metastases.
Treatment-related covariates included radical prostatectomy (n = 1,847), radiation therapy (n = 113), and TURP procedures (n = 22).
Longitudinal clinical follow-up, including relapse status, was available for most donors (1,279 donors (68%) did not relapse; 602 (32%) experienced relapse).
Genotypic and multi-omic covariates were also part of the dataset: all donors underwent whole-genome sequencing, RNA sequencing, and/or DNA methylation profiling, with germline and somatic variant calls, copy-number alterations, structural variants, and genetic ancestry estimates derived from 4,235 ancestry-informative markers
The cohort includes donors from seven countries and reflects a diversity of ancestries (European, East Asian, South Asian, Admixed American, African).

Recruitment

Recruitment was performed independently by each contributing clinical center according to local ethical approvals, institutional review board (IRB) regulations, and International Cancer Genome Consortium (ICGC) guidelines. Donors were enrolled at the time of prostate cancer diagnosis or treatment, primarily through radical prostatectomy, biopsy, or autopsy-based rapid tissue donation programs, depending on the cohort

Ethics oversight

Ethical approval for all contributing cohorts within the PPCG was obtained from the relevant institutional research ethics committees. The full list of approving boards/committees:
Human Research Ethics Committee at Melbourne Health (HREC 2006.073; HREC 2011.009; HREC 2012.275; HREC 2012.220; HREC 2016.087) (Australia)
University Health Network Research Ethics Board and CHU de Québec–Université Laval Research Ethics Board (UHN 06-0822-CE; UHN 11-0024-CE; CHUQc-UL 2012-913:H12-03-192) (Canada)
UBC Clinical Research Ethics Board (UBC REB H21-03722) (Canada)
The National Committee on Health Research Ethics (reference no. 1302791) and notification to the Danish Data Agency (no. 1-16-02-330-13) (Denmark)
CPP Ile de France IV Institutional Review Board (IRB 00003835) (France)
Ethik-Kommission der Ärztekammer Hamburg (PV3552; PV4445; PV4679) (Germany)
NHS East of England–Cambridge Research Ethics Committee (REC 3/0180) (UK)
NHS East Midlands Research Ethics Committee (01/4/061) (UK)
NHS London Research Ethics Committee (CCR 2075) (UK)
University of Pretoria Human Research Ethics Committee (FWA00002567; IRB00002235; IORG0001762; approval #43/2010) (South Africa)
St Vincent's Hospital Human Research Ethics Committee (#SVH/12/231; #SVH/15/227) (Australia)
Human Research Protection Office, US Army Medical Research and Development Command (approvals E02371 and E03280 for genomic interrogation of SAPCS samples) (USA)

As indicated in the manuscript, informed ethical 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.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size

The final harmonized PPCG cohort consisted of 2,021 donors, including 1,001 donors with WGS (1,172 samples), 1,292 donors with RNA-seq (1,594 samples), and 1,490 donors with DNA methylation data (1,928 samples), spanning multiple disease stages, ancestries, and international contributing cohorts. Sample size was not predetermined by statistical power calculations, because the PPCG represents a consortium-level aggregation of all existing high-quality prostate cancer genomic datasets available from contributing groups.
Details about the samples analysed are presented in Supplementary Figure 1.
External dataset include
HMF-Prostate Phase III Randomised Control Study: 240 patients

HMF-Prostate somatic variant analysis: 467 patients
TCGA-PRAD somatic variant analysis: 387 patients

Data exclusions

A small number of samples were excluded based on predefined, objective quality-control criteria established by the PPCG consortium before analysis. Whole-genome sequencing samples were removed if they exhibited DNA contamination greater than 3%, mismatched tumour–normal pairing, duplicate assays, or insufficient sequencing coverage (tumour <30× or normal <20×). In total, 24 WGS samples were excluded for contamination, mis-matching, or duplication. For variant calling, additional exclusions were applied at the variant level according to reproducible filtering procedures: SNVs located in ENCODE blacklisted regions, recurrent artefacts (Sanger panel of normals), or previously reported artefact-prone genes were removed. Indels were tiered using gnomAD allele frequencies, and only variants passing caller-specific filters were retained. Structural variants arising from technical artefacts (transposable element insertions, pseudogene mappings, fold-back artefacts) were filtered using PCAWG-derived exclusion lists.
For copy-number analyses, samples were excluded if automated QC and expert review deemed the CNA profiles unsatisfactory (n = 24). Low-information samples (e.g., true quiet genomes or insufficient reads per chromosome copy) were flagged for limited interpretability but not removed unless specifically required by an analysis.
All exclusion criteria were defined prior to analysis, applied uniformly across cohorts, and unrelated to the biological hypotheses of the study. Further details are provided in our companion manuscript Jakobsdottir et al., 2025a and presented in Supplementary Figure 1.

Replication

All analyses in this study were designed to ensure reproducibility across multiple independent levels of genomic data and validation procedures. The mutational signatures were extracted using established, peer-reviewed computational approaches with 500–1,000+ NMF replicates, stability metrics, silhouette analyses, and cross-cohort verification to confirm robust and reproducible signature solutions.
To verify that integrated mutational footprints (IMFs) were reproducible, we performed:
Independent clustering stability assessments (minimum, median, and maximum stability metrics), Silhouette score evaluation,
Cross-validation using biological plausibility (DDR deficiency, AR activity, replication stress, chromothripsis, and known prostate cancer drivers), and re-quantification of IMF activities in an external metastatic cohort (Hartwig Medical Foundation) using independent variant calling pipelines and confirming expected associations, including predictive capacity of IMF6 for therapy response. Where multiple tumour regions were available for a subset of patients, replication of mutational processes across spatially distinct tumour regions was directly assessed. Three of four patients showed reproducible IMF patterns across clonally related tumour foci. Finally, the major mutational processes were validated in the external TCGA-PRAD whole-genome dataset, where replication timing biases, HRD-associated signatures, and CIN-related effects reproduced the patterns observed in the PPCG cohort. All attempts at replication were successful. No findings in the manuscript failed replication or were deemed non-reproducible.

Randomization

Randomization was not applicable because this study did not involve experimental interventions, treatment allocation, or prospective assignment of samples to experimental groups. Instead, the work is a retrospective, observational analysis of whole-genome sequencing and clinical data from primary prostate cancer donors. All analyses were performed on pre-existing tumour samples collected through the PPCG consortium and external cohorts, with no experimental manipulation, controlled interventions, or conditions requiring randomized assignment. The study design therefore does not include any procedures where randomization would be relevant or meaningful.

Blinding

Blinding was not applicable because this study did not involve experimental interventions, treatment assignment, or subjective outcome assessment. The work is a retrospective, computational analysis of whole-genome sequencing data and associated clinical annotations from existing prostate cancer cohorts. All analyses were performed on pre-existing molecular and clinical datasets using automated bioinformatic pipelines (e.g., mutational signature extraction, clustering, and genomic feature association analyses). These procedures do not involve investigator-dependent scoring or assessments that could be influenced by knowledge of sample identity. Therefore, no elements of the study required blinding, and its absence does not affect the validity or reproducibility of the results.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration: n/a

Study protocol	n/a
Data collection	Clinical data were collected retrospectively from participating PPCG cohorts through site-specific clinical systems and harmonized in a central, secure REDCap instance hosted at the Institute of Cancer Research. Collected variables included age at diagnosis or tumour collection, pretreatment PSA, AJCC clinical and pathological T category, Gleason Grade Group, treatment pathway (e.g., radical prostatectomy, radiotherapy, TURP), and longitudinal outcomes. Harmonization procedures and cohort-level metadata are described in the PPCG data descriptor; all data access is controlled per PPCG governance (Project Concept Form approval; DAC processes for controlled repositories)
Outcomes	Prespecified clinical outcomes included: Biochemical relapse after curative-intent primary therapy (defined according to standard criteria: AUA for post-prostatectomy and "Phoenix" definition for primary radiotherapy), Distant metastasis (at diagnosis or during follow-up), Overall survival and cancer-specific survival (where available). Time origin for survival analyses was date of sample collection. Outcome definitions and cohort-level follow-up summaries are provided in the PPCG data descriptor and companion analyses.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*