

Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets

David C. Wedge^{1,2,3,35*}, Gunes Gundem^{2,4,35}, Thomas Mitchell^{2,5,6,35}, Dan J. Woodcock¹, Inigo Martincorena², Mohammed Ghorii², Jorge Zamora², Adam Butler², Hayley Whitaker⁷, Zsofia Kote-Jarai⁸, Ludmil B. Alexandrov², Peter Van Loo^{2,9}, Charlie E. Massie^{6,10}, Stefan Drento^{1,2,9}, Anne Y. Warren¹¹, Clare Verrill^{3,12}, Dan M. Berney¹³, Nening Dennis¹⁴, Sue Merson⁸, Steve Hawkins⁶, William Howat¹¹, Yong-Jie Lu¹³, Adam Lambert¹², Jonathan Kay⁷, Barbara Kremeyer², Katalin Karaszi¹², Hayley Luxton⁷, Niedzica Camacho^{4,8}, Luke Marsden¹², Sandra Edwards⁸, Lucy Matthews¹², Valeria Bo¹⁵, Daniel Leongamornlert^{2,8}, Stuart McLaren², Anthony Ng¹⁶, Yongwei Yu¹⁷, Hongwei Zhang¹⁷, Tokhir Dadaev⁸, Sarah Thomas¹⁴, Douglas F. Easton¹⁸, Mahbubl Ahmed⁸, Elizabeth Bancroft^{8,14}, Cyril Fisher¹⁴, Naomi Livni¹⁴, David Nicol¹⁴, Simon Tavaré¹⁵, Pelvender Gill¹², Christopher Greenman¹⁹, Vincent Khoo¹⁴, Nicholas Van As¹⁴, Pardeep Kumar¹⁴, Christopher Ogden¹⁴, Declan Cahill¹⁴, Alan Thompson¹⁴, Erik Mayer¹⁴, Edward Rowe¹⁴, Tim Dudderidge¹⁴, Vincent Gnanapragasam^{5,20}, Nimish C. Shah⁵, Keiran Raine², David Jones², Andrew Menzies², Lucy Stebbings², Jon Teague², Steven Hazell¹⁴, Cathy Corbishley²¹, CAMCAP Study Group²², Johann de Bono⁸, Gerhardt Attard⁸, William Isaacs²³, Tapio Visakorpi²⁴, Michael Fraser²⁵, Paul C. Boutros^{26,27,28}, Robert G. Bristow^{25,27,29}, Paul Workman^{8,36}, Chris Sander³⁰, The TCGA Consortium²², Freddie C. Hamdy¹², Andrew Futreal², Ultan McDermott², Bissan Al-Lazikani^{8,36}, Andrew G. Lynch^{15,31,36}, G. Steven Bova^{23,24,36}, Christopher S. Foster^{32,33,36}, Daniel S. Brewer^{8,19,34,36}, David E. Neal^{6,20,36}, Colin S. Cooper^{8,19,36} and Rosalind A. Eeles^{8,14,36*}

Prostate cancer represents a substantial clinical challenge because it is difficult to predict outcome and advanced disease is often fatal. We sequenced the whole genomes of 112 primary and metastatic prostate cancer samples. From joint analysis of these cancers with those from previous studies (930 cancers in total), we found evidence for 22 previously unidentified putative driver genes harboring coding mutations, as well as evidence for *NEAT1* and *FOXA1* acting as drivers through noncoding mutations. Through the temporal dissection of aberrations, we identified driver mutations specifically associated with steps in the progression of prostate cancer, establishing, for example, loss of *CHD1* and *BRCA2* as early events in cancer development of *ETS* fusion-negative cancers. Computational chemogenomic (canSAR) analysis of prostate cancer mutations identified 11 targets of approved drugs, 7 targets of investigational drugs, and 62 targets of compounds that may be active and should be considered candidates for future clinical trials.

Prostate cancer is the most common solid cancer in men (diagnosed in 12%) and is often fatal (9% of male cancer deaths). It is difficult to manage clinically due to a poor current understanding of what dictates its highly variable natural history and of what underlies the development of castration-resistant disease¹. Extensive data on the structure of prostate cancer genomes have been published^{2–6}, including work from our own consortium^{7–10}. These studies have identified a number of genetically distinct subgroups, including cancers with *ERG*, *ETV1*, *ETV4*, *FLI1*, *SPOP*, *FOXA1* and *IDH1* alterations. Overlapping with these categories, cancers may have alterations in PI3K and DNA repair pathways, with the latter significantly

over-represented in advanced disease⁴. However, we have relatively limited understanding of the ordering of genetic events, with the exception that *ETS* gene alteration appears to represent an early event while mutations of *AR* are later, sometimes convergent events occurring in advanced and metastatic disease. Indeed, we have very little understanding of the evolution of mutational processes, the various genetic paths that cancers traverse on their way to progression, the levels of heterogeneity at different stages of development, or the effect of these factors on clinical outcome.

Gene status has been used in studies designed to improve the poor predictive value of conventional clinical markers (PSA,

A full list of affiliations appears at the end of the paper.

Gleason sum, stage) and to develop disease management strategies. For example, genetic alteration of *BRCA1* or *BRCA2*²¹, *PTEN* deletion¹², amplification of *AURKA* together with the *MYCN* gene¹³, and coordinated loss of *MAP3K7* and *CHD1*¹⁴ have been reported to have prognostic value. A number of commercial prognostic tests based on gene expression profiles are also available^{15–17}, and a classification framework has been proposed¹⁸. Improvements in the treatment of castration-resistant disease have been made through better targeting of AR regulation using abiraterone¹⁹ and enzalutamide²⁰, while PARP inhibitors are effective against cancers harboring *BRCA1* or *BRCA2* mutations and other defects in DNA repair pathways²¹. However, notable advances have been made recently through the retasking of approved drugs²². In this study, we use previously unpublished whole-genome DNA sequencing data in combination with published data to provide insights into the mechanism of progression of prostate cancer to lethal disease and to design new molecularly based strategies for drug targeting.

Results

We whole genome sequenced cancerous and matched normal samples from 87 primary prostate cancers from the UK and 5 from China, together with 10 hormone-naïve prostate metastases and 10 castration-resistant metastases from the USA. Our analysis (see Methods) yields insights into the nature and order of acquisition of driver alterations, genomic heterogeneity in primary and metastatic cancers, changes in mutational signatures during progression and potential drug targets. In addition, we identify coding and non-coding drivers by combining single nucleotide variants (SNVs) and small insertions/deletions (indels) in our dataset with those from

The Cancer Genome Atlas (TCGA)⁴ (425 samples), the COSMIC database²³ (243 samples) and Stand Up to Cancer²⁴ (SU2C-PCR, 150 samples) to give a combined dataset, hereafter referred to as the ‘joint dataset’, comprising 710 primary cancers and 220 metastases. Supplementary Table 1 summarizes the genes affected in both our study and the joint dataset.

For the 112 cancer–normal pairs in our cohort, we identified 392,753 SNVs, 54,952 indels and 10,921 chromosomal rearrangements (Fig. 1). The mean genome-wide substitution rate was 1.23/Mb, with a significant difference in mutational burden between the primary (0.99) and metastatic (2.30) samples ($P=4.4\times 10^{-15}$, Methods). Moreover, within the metastatic subset, mutation burden was higher in men treated with androgen deprivation therapy (ADT, also referred to as castration resistant prostate cancer, CRPC) than treatment-naïve men (2.98 vs. 1.61, $P=0.015$). There were also significantly more rearrangements in metastatic than in primary samples ($P=0.0059$), while the proportion of breakpoints attributed to a chromoplexy-like event²⁵ was indistinguishable between the two groups. Within the metastatic group, the ADT samples had more rearrangements than did the hormone-naïve ($P=0.027$), with no difference in the proportion of chromoplexy-like events (Fig. 1).

Genes of interest were determined through a comprehensive set of analyses to identify excess non-synonymous mutations in coding regions; excess missense mutations within a gene, indicative of an oncogenic driver; excess mutations in noncoding regions; regions with an excess of structural variants in either ETS⁺ or ETS⁻ cancers; and regions with recurrent copy number aberrations (CNAs) in either ETS⁺ or ETS⁻ cancers. Overall, we identified 73 genes with evidence for involvement in prostate cancer

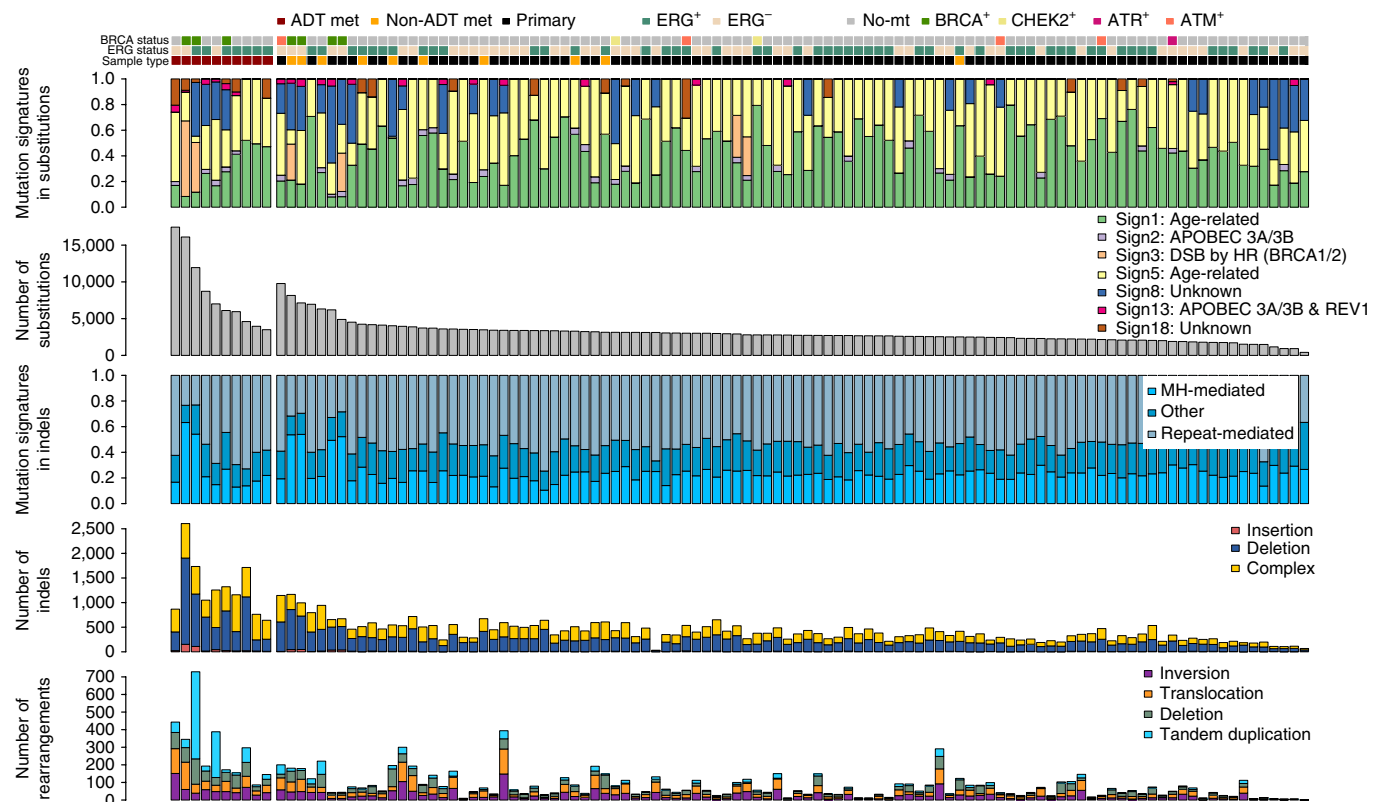


Fig. 1 | Mutational landscape of prostate cancers. From top to bottom: mutation status of DNA repair genes (No-mt indicates no mutation in these), ETS fusion status and sample type (met indicates metastasis); proportion of mutations assigned to each signature⁴⁸; number of SNVs identified in each sample; proportion of small indels associated with microhomology or repetitive regions; number of insertions, deletions and complex indels in each sample; and total number of structural variants in each sample, separated into inversions, translocations, deletions and tandem duplications. Sample ordering is reported in Supplementary Table 7.

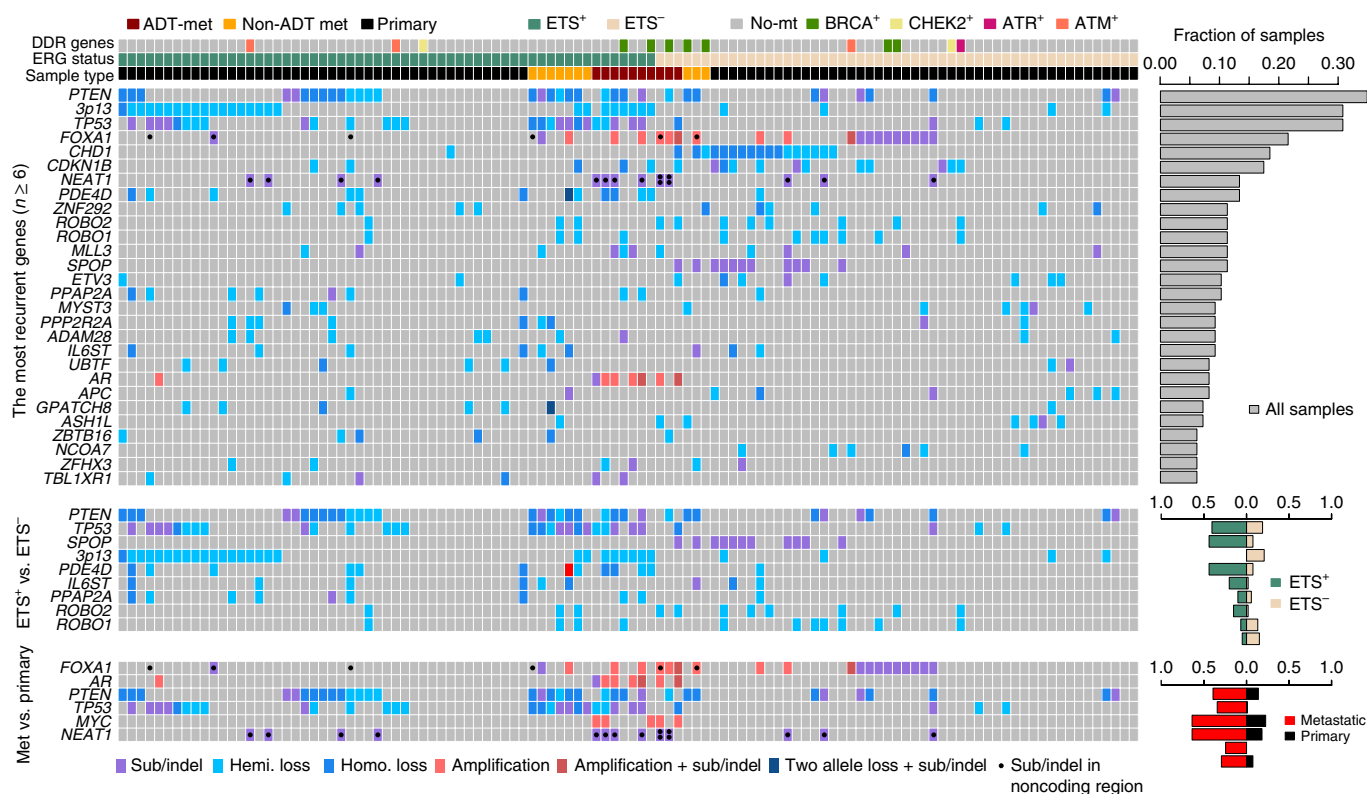


Fig. 2 | Landscape of driver genes in prostate cancer. Genes were identified using three different methods: top, genes that have undergone genetic aberration in at least 6 samples ($n=112$ biologically independent samples); middle, genes with aberrations enriched in either ETS⁺ or ETS⁻ cancers (Fisher's exact test for *PTEN*, *TP53*, *SPOP*, *3p13*, *PDE4D* and *PPAP2A*; *ROBO1* and *ROBO2* are in a region enriched for structural variants in ETS⁻ tumors; *IL6ST* is in a region enriched for structural variants in ETS⁺ tumors; $n=59$ ETS⁺, $n=53$ ETS⁻ biologically independent samples); bottom, genes enriched in metastatic (met) samples (Fisher's exact test, $n=20$ metastatic, $n=98$ primary biologically independent samples). Right, the fraction of samples bearing each type of aberration. DDR, DNA damage response; 'hemi. loss', loss of heterozygosity resulting from CNA; 'homo. loss', homozygous deletion resulting from CNA; 'two allele loss + sub/indel', genes in triploid regions bearing aberrations of all three gene copies. Sample ordering is reported in Supplementary Table 7.

development (Fig. 2, Table 1 and Supplementary Table 2). Based on a literature search, each gene was assigned a high, medium or low level of previous supporting evidence (Table 1 and Supplementary Table 2). In addition to 22 genes with little or no previous evidence of involvement in prostate cancer (Table 1, low previous evidence), we provide corroborating evidence for 8 more genes previously lacking strong evidence of driving prostate cancer (Table 1, medium previous evidence).

Coding driver mutations. We identified 28 genes with an excess of non-synonymous coding mutations, 5 of which are previously unknown drivers in prostate cancer (Supplementary Table 2). *TBL1XR1* was enriched in truncating SNVs and indels and is also located in a genomic region enriched for rearrangements in ETS⁺ cancers (chr3:172–179Mb) (Fig. 3). These rearrangements result in loss of heterozygosity (LOH) or, in one case, homozygous deletion, suggesting a cancer suppressor role for this gene. Another significantly mutated gene primarily affected by truncating mutations was *ZMYM3*, which encodes a component of CoREST, a transcriptional repressor complex including *REST* (RE-1 silencing transcription factor) and involved in suppression of neuronal differentiation-related genes in non-nervous tissues²⁶. In addition, two further CRPC samples from the SU2C-PCF study²⁴ had nonsense mutations and one sample in our study had a 70-kb exonic deletion in *REST*.

Two other genes with recurrent truncating mutations were *IL6ST* and *CASZ1* (Fig. 3). The latter is a putative cancer suppressor in neuroblastoma²⁷, while the former encodes glycoprotein 130, the signal-transducing subunit of the interleukin 6 (IL6) receptor.

The pattern of mutations we observe in the joint dataset for *IL6ST* is dominated by truncating events. Moreover, this gene is located in a genomic region recurrently rearranged in ETS⁺ cancers, resulting in either LOH or homozygous deletion (four cases of each), suggesting a cancer-suppressive role. *TBX3*, previously reported to harbor mutations in breast cancer²⁸, exhibited a mixed pattern of mutations with mostly missense mutations and two cancers harboring truncating events.

Analysis of missense mutations identified recurrent mutations in seven more genes, of which two are newly reported (Supplementary Table 2). *CNOT3* exhibited mutation hotspots altering two amino acid positions, Glu20Lys (4 of 932 samples) and Glu70Lys (5 of 932 samples), as well as a nonsense mutation in a single sample (Fig. 3). *CNOT3* has a known cancer-suppressive function in T-cell acute lymphoblastic leukemia²⁹. Enrichment for missense mutations was identified in *RPL11*, which encodes a ribosomal protein and putative cancer suppressor upstream of the MDM2–TP53 pathway³⁰. In contrast to findings of previous studies, the enrichment for missense mutations in both *CNOT3* and *RPL11* suggests oncogenic, rather than tumor suppressor, roles in prostate cancer. A comparison between coding mutations in metastatic and primary samples in the joint dataset identified enrichment in metastases for mutations in *TP53*, *AR*, *KMT2C*, *KMT2D*, *RB1*, *APC*, *BRCA2*, *CDK12*, *ZFHX3*, *CTNBN1* and *PIK3CB* (Supplementary Table 2), confirming previous studies^{3,24}.

Noncoding driver mutations. Analysis of noncoding portions of the genome identified two regions significantly enriched for

Table 1 | Putative driver genes

Gene	Mutation type(s)	Prior evidence	Prior evidence type	Evidence in our study
ADAM28	SV, CNA	Low	Biological evidence ⁵⁹	SVs and CNA in ETS ⁺
ANTXR2	SV, SNV/indel	Low	None	Clinical correlation
ASH1L	SV, SNV/indel	Low	²⁵	Truncating mutations, SVs in ETS ⁻
CDH12	SV	Low	None	Clinical correlation
FOXO1	CNA	Low	Biological evidence ⁶⁰	CNA in ETS ⁻
IL6ST	SV	Low	Biological evidence ⁶¹	dN/dS, SVs and CNA in ETS ⁺ , clinical correlation
LCE2B	SNV/indel	Low	None	dN/dS (missense)
MAP3K1	SV, CNA	Low	None	SVs, CNA in ETS ⁺
MYST3	SV	Low	²⁵	SVs in ETS ⁻ , RNA expression
NCOA7	SV	Low	None	SVs in ETS ⁻
NDST4	SNV/indel	Low	None	dN/dS (missense)
NEAT1	Noncoding	Low	Biological evidence ³¹	Noncoding
PDE4D	SV	Low	SNP data ⁶²	SVs and CNA in ETS ⁺
PPAP2A	SV	Low	SNP data ⁶²	SVs and CNA in ETS ⁺
PPP2R2A	SV	Low	Biological evidence ⁶³	SVs and CNAs in ETS ⁺
ROBO1	SV	Low	Biological evidence ⁶⁴	SVs in ETS ⁺
ROBO2	SV	Low	²⁵	SVs in ETS ⁺
RPL11	SNV/indel	Low	²⁵	dN/dS (missense)
SEN6	SV	Low	Biological evidence ⁴²	Enriched SVs, RNA expression
TBL1XR1	SNV/indel, SV	Low	Known AR co-regulator biological evidence ⁶⁵	dN/dS
USP28	SV, CNA, SNV/indel	Low	None	SVs, CNA, SNV/indel
ZNF292	SV, CNA SNV/indel,	Low	²⁵	Enriched SVs, homozygous deletions, truncating mutations
ARID1A	SNV/indel	Medium	⁶⁶	dN/dS
CASZ1	SNV/indel	Medium	COSMIC, TCGA and SU2C	dN/dS
CNOT3	SNV/indel	Medium	Mutated in leukemia ⁶⁷	dN/dS (missense)
LRP1B	SV, CNA	Medium	SNP data ⁶²	SVs and CNA in ETS ⁻
PIK3R1	SNV/indel	Medium	²⁴	dN/dS
RGMB	CNA	Medium	Deletions ³⁸	CNA in ETS ⁻
TBX3	SNV/indel	Medium	Known breast cancer gene	dN/dS
ZMYM3	SNV/indel	Medium	COSMIC SU2C	dN/dS

Genes were identified in our study using several methods, detailed in the last column: dN/dS; enrichment for SVs or CNAs in ETS⁺ or ETS⁻ cancers; enrichment for truncating mutations or homozygous deletions; clinical correlation. From a PubMed literature search, prior evidence for each gene being a driver of prostate cancer was classified as 'low' if the gene has not been previously reported as playing a role in prostate cancer tumorigenesis or progression. Isolated alterations may have been observed or biological evidence for importance may have been presented as indicated in the prior evidence column. Prior evidence was classified as 'medium' for genes reported previously as playing a role in prostate carcinogenesis or progression but currently lacking statistical support based on genetic alterations. Evidence considered included presence of multiple genetic alterations, SNP associations, and known cancer genes in other tissues. The high confidence genes are those that are widely accepted to represent cancer genes and to be altered in prostate cancer, including genes such as *HRAS*, *SPOP* and *IDH1*. In each case, there are two or more of the following: statistical verification of higher incidence, biological experiments, clinical correlations, confirmation in multiple studies, recognition as cancer genes in other cancer types. dN/dS, non-synonymous:synonymous ratio calculated for all SNVs and indels; dN/dS (missense), non-synonymous:synonymous ratio calculated for missense SNVs only; SV, structural variant; CNA, copy number aberration; SNV, single nucleotide variant; indel, small insertion/deletion; ETS, E26 transformation-specific.

mutations. *NEAT1*, which produces a long noncoding RNA recently reported to be associated with prostate cancer progression³¹, was mutated in 13 of 112 ICGC samples, with significant over-representation in patients with metastatic disease (6 of 20 metastases vs. 7 of 91 primaries, Fisher's exact test, $P=0.012$; Fig. 3). Out of the metastatic tumors, *NEAT1* mutations were found only in patients who had undergone ADT, consistent with the link between high *NEAT1* expression and resistance to AR-targeting therapies³¹. Notably, two of these six samples had two separate *NEAT1* mutations. The *FOXO1* promoter also had significant evidence of selection. This gene modulates AR-regulated transcriptional signaling³² and has previously been found to harbor recurrent coding mutations⁵. In our series, we identified 14 samples with coding and 6 samples with

noncoding mutations, with two samples (PD14721a and PD12813a) bearing both a coding and a noncoding mutation. We also identified mutations in the promoter for *FOXO1*, a gene with known cancer-suppressive effect in prostate tumorigenesis³³, in three samples, but this was insufficient to reach statistical significance.

Structural variant enrichment in ETS⁺ and ETS⁻ cancers. The density of rearrangements varies across the genome as a result of various factors, including chromatin state, GC content, gene density, replication timing and repetitive sequence. To remove the effect of these factors, we segmented interbreakpoint distance across the genome separately in ETS⁺ and ETS⁻ cancers and identified regions with differential enrichment for rearrangements between the two

subtypes. The functional importance of many of these regions was supported by an excess of truncating mutations or CNAs.

In addition to regions previously identified as enriched for rearrangements in ETS⁺ cancers (*FOXP1*, *RYBP*, *SHQ1*, *PTEN* and *TP53*)^{34–37}, we identified two unreported regions. The region chr5:55–59Mb covers the genes *PPAP2A*, *PDE4D*, *MAP3K1* and *IL6ST* (Fig. 3). In *IL6ST*, we also detected significant enrichment for coding mutations, suggesting this is the main target of the aberrations. At chr3:171–178Mb, *TBL1XR1* was similarly enriched for both rearrangements and truncating mutations.

In ETS⁻ cancers, we confirmed a previously reported enrichment for rearrangements containing *CHD1*^{38,39}. A target of enriched rearrangements in the region chr1:149–158Mb is likely *ETV3*. In 5 of 9 cancers, *ETV3* was exclusively affected by these events (4 LOH by deletion and 1 by translocation). Additionally, 1 cancer had a truncating mutation (Arg413fs*3) and 2 had missense mutations (Ala73Val and Leu37Gln). In total, 12 patients had localized alteration, 10 of whom had ETS⁻ cancers. Moreover, in the joint dataset, there were four cancer samples with truncating mutations in this gene. In contrast to *ETV4*, the nature of variants in *ETV3* is indicative of a tumor-suppressive role in prostate cancer. Manual inspection of the recurrently rearranged region chr3:76–84Mb identified *ROBO1* and *ROBO2* as possible targets (Fig. 3). In total, 16 of 112 samples had an event affecting one or other of these genes, and in 4 samples both were affected. Previously implicated in pancreatic ductal adenocarcinoma⁴⁰, these two genes have not been previously reported in the context of prostate cancer.

Events enriched at chr6:80–114Mb indicate that *ZNF292* is a possible target. Eleven of 112 patients (5 ETS⁺ and 6 ETS⁻) had loss of at least one chromosome copy, and in 2 patients there was a homozygous loss specifically targeting *ZNF292*. Moreover, the joint dataset contained 5 of 932 samples with a truncating mutation, further suggesting a cancer-suppressive function for this gene in prostate cancer. Another gene affected by recurrent rearrangements on 6q was *SENP6*, which encodes a small ubiquitin-like modifier (SUMO)-specific protease that removes SUMO polypeptides from conjugated proteins⁴¹ and may influence AR function⁴². Of note, 4 of 5 rearrangements in this region affected *SENP6* only, leading to a significant reduction in expression; $P < 0.0001$ (Supplementary Fig. 1). Finally, located at chr6:126Mb, the nuclear receptor co-activator *NCOA7* was altered in 6 samples, with 1 sample having homozygous loss.

Further regions enriched for rearrangements in ETS⁻ cancers were chr2:133–144Mb (*LRP1B*), chr8:112–114Mb (*CSMD3*) and chr8:40–41Mb (*MYST3*). The first two genes are very large and fall within reported fragile sites⁴³. Nevertheless, preferential enrichment of breakpoints in ETS⁻ cancers may suggest either that underlying structure, such as AR binding sites or nucleosome structure, or that epistatic interactions between ETS fusion and other rearrangements affect the occurrence of rearrangements at these loci. Samples containing structural variants affecting *MYST3* were found to have significantly reduced RNA expression; $P = 0.025$ (Supplementary Fig. 1).

Timing of copy number aberrations. To identify routes to progression in prostate cancer, we developed an approach by which to order the occurrence of CNAs by combining information on the clonality of CNAs, timing relative to whole-genome duplication (WGD), and timing of homozygous deletions relative to neighboring hemizygous losses. Information from all tumors was combined using a Bradley–Terry model to give the most likely ordering of events. By applying a set of logical rules (see Methods), we deciphered the temporal ordering of subclonal CNAs within each cancer. In general, homozygous deletions appeared late in oncogenesis, corroborating previous findings that homozygous deletions are associated with advanced disease^{44–46}.

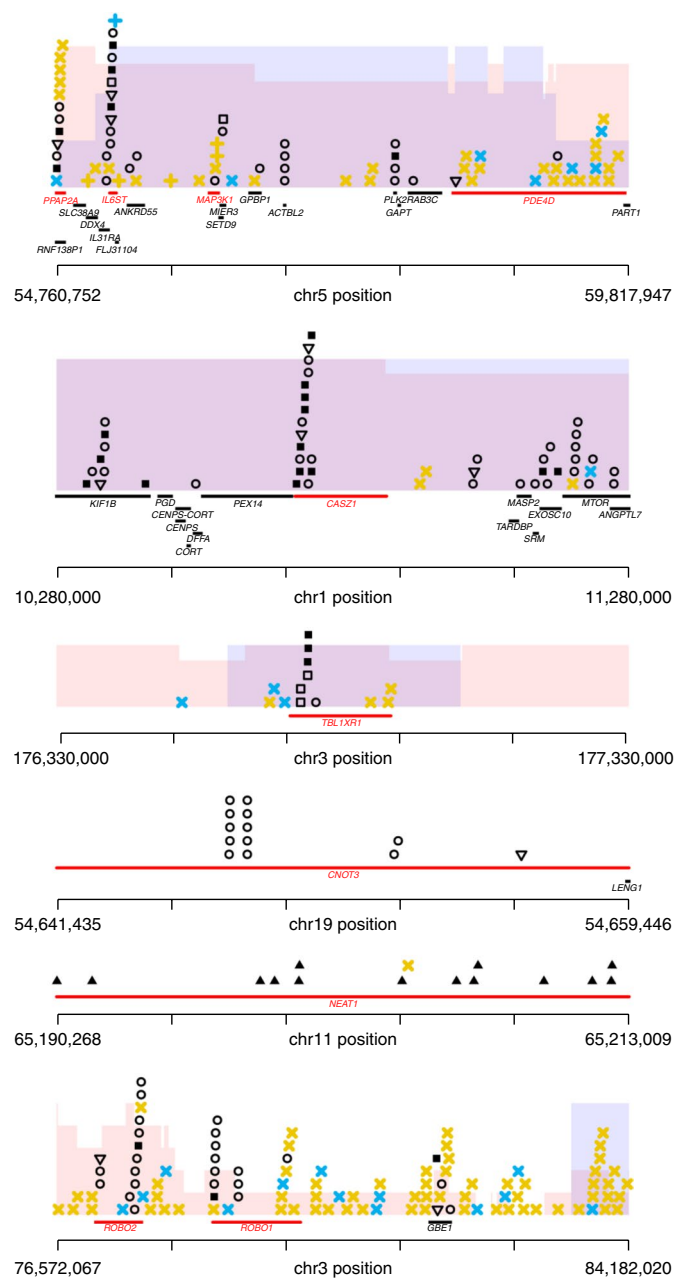


Fig. 3 | Putative new driver genes. Putative drivers are shown in red and genomic aberrations are displayed as follows: missense SNVs, circles; nonsense SNVs, open triangles; essential splice site mutations, open squares; indels, closed squares; noncoding mutations, closed triangles; simple structural variant, yellow cross; chromoplexy event, blue cross; region enriched for LOH with height proportional to the number samples containing LOH, pink shading; region enriched for homozygous deletions with height proportional to the number of samples containing homozygous deletion, blue shading.

Clear differences emerged in the evolution of ETS⁺ and ETS⁻ prostate cancers. Where present, the deletion between the *TMPRSS2* and *ERG* genes in ETS⁺ cancers was an early (generally clonal) event, as was gain of chr8q in the locus 112–137Mb (Fig. 4a). The earliest homozygous deletions in ETS⁺ cancers included chr5:55–59Mb, corroborating the rearrangements targeting *PPAP2A*, *PDE4D*, *MAP3K1* and *IL6ST*, and chr10:89–90Mb, which covers *PTEN* (Figs. 3 and 4a). In ETS⁻ cancers, losses at chr5:60–100Mb (*CHD1* and *RGMB*),

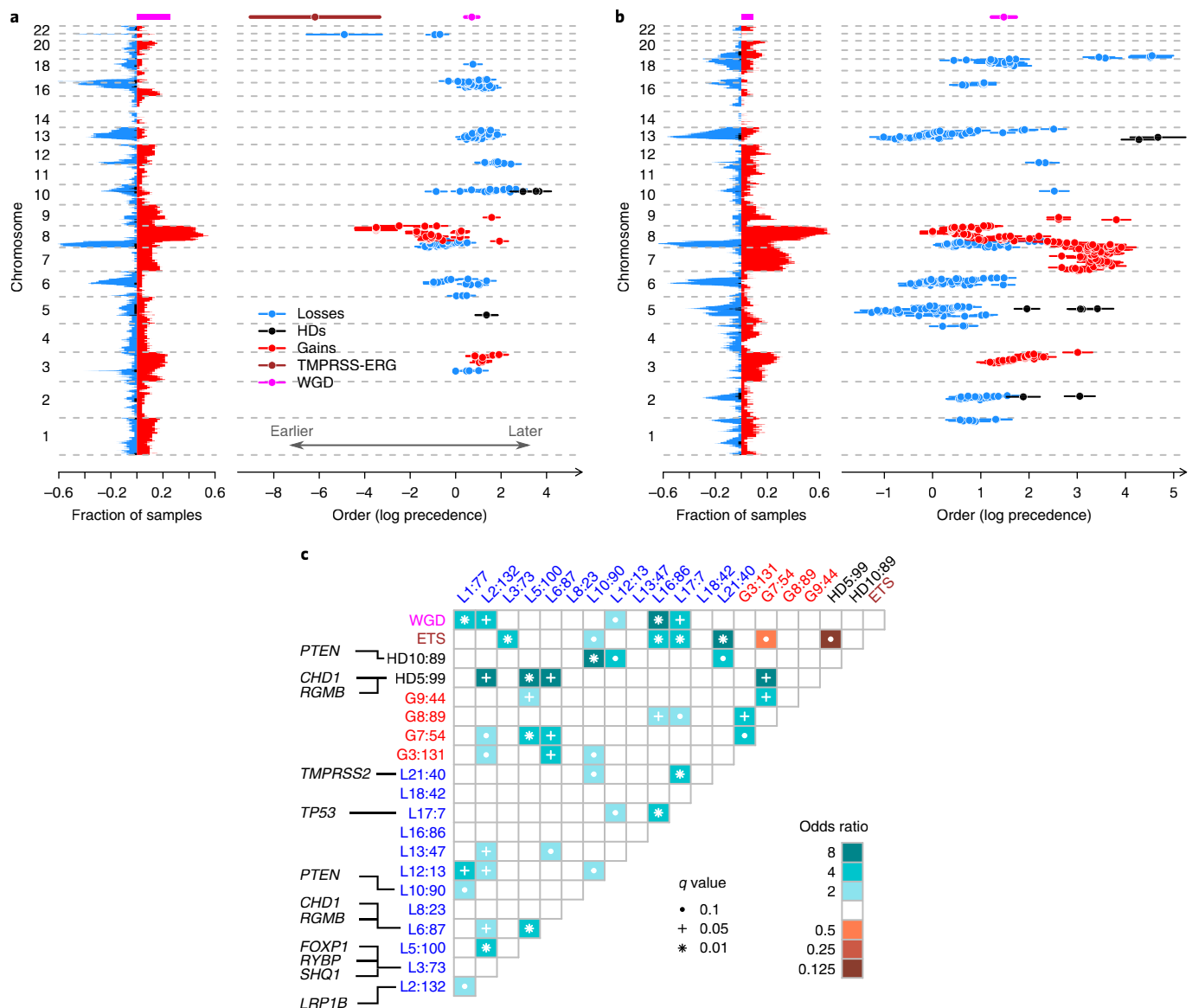


Fig. 4 | Temporal evolution of CNAs in ETS⁺ and ETS⁻ prostate cancer. (a,b) For ETS⁺ cancers ($n=45$ biologically independent primary cancer samples; **a**), and ETS⁻ cancers ($n=47$ biologically independent primary cancer samples; **b**), left plot shows the landscape of CNAs with genomic loci plotted against fraction of cancers. LOH is depicted in blue, homozygous deletions (HDs) in black, gains in red, *TMPRSS2-ERG* deletion in brown, and whole-genome duplication (WGD) in green. Right plot shows the temporal evolution of significantly recurrent ($P < 0.05$, permutation test with Benjamini-Hochberg procedure) CNAs by genomic locus over time (mean with 95% confidence intervals, log precedence relative to arbitrary reference). Lower values indicate earlier events. **(c)** Pairwise associations among CNAs. Recurrently aberrant regions with a false discovery rate < 0.1 are shown. Associations are indicated by odds ratio, with brown colors depicting mutually exclusive events and turquoise colors depicting correlated events. Genomic loci annotated by type of aberration (G, gain; L, loss; HD, homozygous deletion), chromosome, median position in Mb. For focal events, the putative target genes are annotated.

chr13:32–91Mb (which includes *BRCA2*, *RB1* and *FOXO1*) and chr6:73–120Mb are followed by losses at chr2:124–142Mb, then by gains at chr3:100–187Mb, and then whole chromosome gain of chr7 (Fig. 4b). Loss of *CHD1* has been previously implicated in the initiation of ETS⁻ prostate cancer, preventing *ERG* rearrangement in the prostate³⁸, and our data confirm the exclusivity between ETS positivity and homozygous loss of *CHD1* (Fig. 4c). In both ETS⁺ and ETS⁻ cancers, WGD correlated with loss of chromosomal segments at chr1:94Mb, chr2:140Mb, chr12:12Mb, chr16:85Mb and chr17:7Mb (Fig. 4c). From timing analysis, these losses appear to occur synchronously with WGD in most cases. Gains at chr8:101Mb occurred before WGD, gains at chr3:131Mb occurred synchronously, and gains at chr7:88Mb tended to follow WGD.

Timing of point mutations and indels. SNVs and indels were clustered according to their cancer cell fraction using a Bayesian Dirichlet process⁴⁷. The proportion of SNVs identified as subclonal showed considerable variation across cancers, but was significantly higher in primary than metastatic samples ($P=0.022$, Wilcoxon rank sum test, Fig. 5a), as was the proportion of subclonal indels ($P=0.00033$) and the fraction of the genome with subclonal CNAs ($P=0.0037$, Supplementary Fig. 2). This is apparent evidence for a bottleneck in acquiring metastatic potential rather than a response to treatment, since levels of heterogeneity in untreated metastases are no lower than in androgen-deprived metastases (Fig. 5a).

The levels of heterogeneity observed in SNVs and indels were correlated (Pearson $r=0.57$, $P=2.3 \times 10^{-9}$, Fig. 5a). Higher levels of

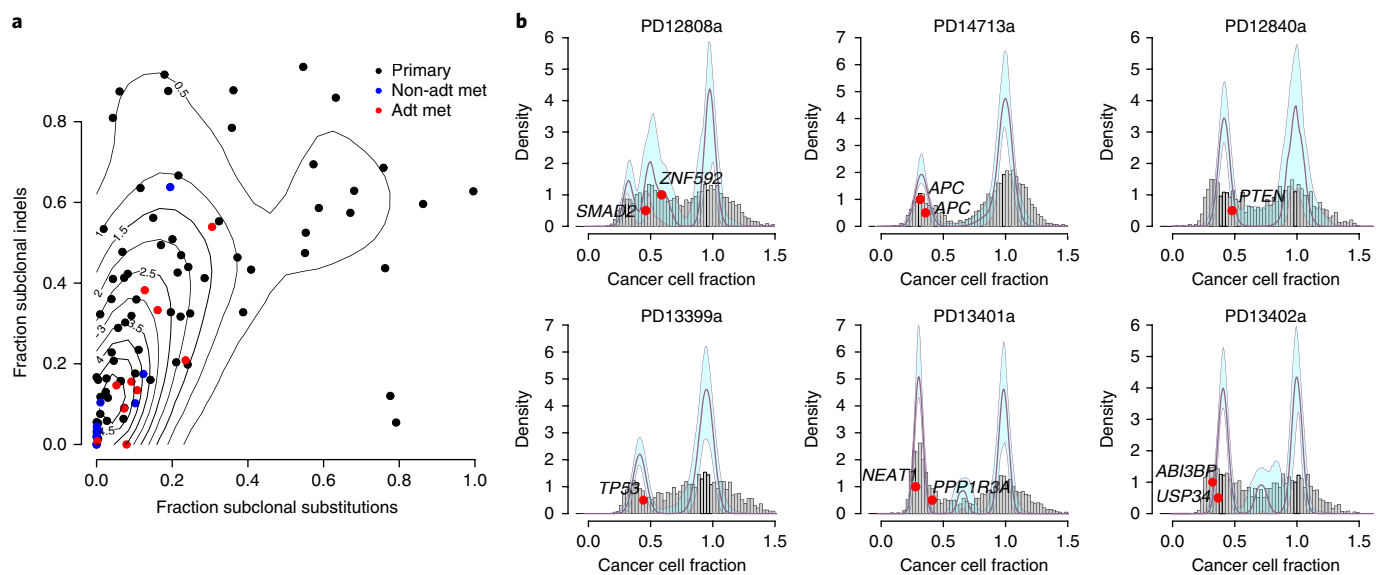


Fig. 5 | Heterogeneity and subclonal mutations. (a) Metastatic tumors have less heterogeneity than primary tumors, whether assessed from SNVs or indels. Each dot represents a different sample, colored by sample type. The x axis shows the fraction of SNVs that are subclonal and the y axis the fraction of indels that are subclonal. Contour lines calculated using R package *kde2d*; $n = 93$ biologically independent samples (10 ADT metastases, 9 hormone-naive metastases, 74 primary tumors). (b) Samples with multiple subclonal mutations in driver genes. Fraction of cancer cells carrying mutation is shown as gray histogram for all mutations and as red ovals for mutations in known driver genes. Mutations are clustered using a Dirichlet process as previously described⁴⁷, with thick plum-colored lines indicating fitted distribution and pale blue regions indicating 95% posterior confidence intervals. Peaks with a fraction close to 1 are clonal, whereas peaks at lower fractions indicate subclonal mutations.

heterogeneity were observed among indels than SNVs ($P = 2.4 \times 10^{-8}$). However, the possibility cannot be ruled out that variant calling of indels may have greater sensitivity for low-allele-frequency variants than calling of SNVs.

Driver SNVs were identified as clonal or subclonal in each sample according to the cluster to which they were assigned, with 84 classified as clonal and 22 (21%) as subclonal. Our power to detect subclonal mutations is limited by sequencing depth, and the real number of subclonal driver mutations is likely much higher. The driver mutations identified as subclonal included 2 mutations in *APC* in the same sample, PD14713a. Notably, this cancer has undergone clonal loss of one copy of chr5q, followed by mutations in *APC* in two different subclones (Fig. 5b and Supplementary Fig. 3), suggesting convergent evolution. Five other samples each had two subclonal drivers: PD12808a has a missense mutation in *ZNF292* and an essential splice site mutation in *SMAD2*; PD13401a has a nonsense mutation in *PPP1R3A* and a mutation in the promoter of *NEAT1*; PD13402a has a nonsense mutation in *USP34* and an essential splice site mutation in *ABI3BP* (Fig. 5b); PD12820a has a missense mutation in *USP48* and an essential splice site mutation in *ASXL2*; and PD13389a has a frameshift mutation in *PHF12* and an essential splice site mutation in *TBX3* (not shown). Subclonal mutations were also seen in several common drivers, including 1 in *TP53* (PD13339a) and 1 in *PTEN* (PD12840a). By contrast, *SPOP* was mutated in 10 samples, always clonally and always in ETS⁺ tumors (Fig. 2).

Mutational signatures. Analysis of the mutational signatures by non-negative matrix factorization (NMF) showed that, in addition to the ubiquitous ‘clock-like’ signatures 1 and 5, there was presence of the previously described⁴⁸ signatures 2, 3, 8, 13 and 18. Signature-3-positive samples were enriched for germline or somatic mutations in *BRCA1* and *BRCA2* genes (4 of 6 samples), as reported previously⁴⁸ (Fig. 1). However, the presence of high levels of microhomology-mediated deletions was even more strongly correlated with the presence of *BRCA* mutations (6 of 6 samples). Separating

the mutations into early clonal, late clonal and subclonal epochs (see Methods) showed that the proportion of signature 1 mutations decreases over time, suggesting an increase of cancer-associated mutagenic processes relative to innate processes ($P = 2.2 \times 10^{-16}$, test for trend in proportions).

Signature 13, previously associated with the activity of the AID/APOBEC family of cytidine deaminases, was over-represented in advanced disease: 45% (9 of 20) in metastases vs. 14% (14 of 92) in primaries (Fisher’s exact test, $P = 5.6 \times 10^{-3}$). Similarly, signature 18, which has been previously associated with failure of base excision repair and the accumulation of mutations from 8-oxoguanine damage⁴⁹, was enriched in advanced disease: 40% (8 of 20) in metastases vs. 11% (10 of 92) in primaries (Fisher’s exact test, $P = 3.8 \times 10^{-3}$). In a recent report of 560 breast cancer whole genomes, signature 8 correlated with DNA damage repair deficiency⁵⁰. Androgen signaling is known to positively regulate multiple genes involved in DNA repair^{51, 52}, while androgen deprivation impairs DNA double-strand break repair⁵³. In support of these previous reports, the proportion of mutations assigned to signature 8 is consistently higher among later appearing (subclonal) populations of cells ($55\% \pm 24\%$) than earlier (clonal) populations ($28\% \pm 12\%$) (t -test, $P = 1.3 \times 10^{-4}$, Supplementary Table 3). The proportion of metastases with evidence for the action of signature 8 was higher than that for primary tumors, although not reaching statistical significance (8 of 20 metastases vs. 25 of 92 primaries, Fisher’s exact test $P = 0.28$). Increased prevalence of DNA-damage related genes in metastatic prostate cancer, as well as the observations made in this study, warrant an extensive study of mutational signatures in therapy-naive disease and CRPC in a larger dataset to explore the relevance of checkpoint inhibition as an alternative therapy for advanced prostate cancer.

Clinical correlates. *CDH12* and *ANTXR2* alterations were significantly associated with time to biochemical recurrence (Benjamini–Hochberg adjusted $P = 0.0060$ (*CDH12*) and 0.012 (*ANTXR2*), HR (hazard ratio) = 9.3 and 7.7, Cox regression model, Fig. 6) and were

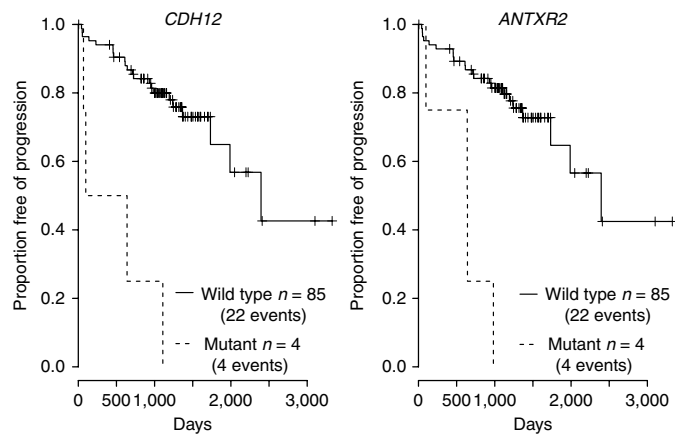


Fig. 6 | Clinical outcome. Kaplan-Meier plots of recurrently mutated genes for which there is a significant correlation with time to biochemical recurrence after prostatectomy: *CDH12* (left, $P=0.006$) and *ANTXR2* (right, $P=0.012$) (Cox regression model; Benjamini-Hochberg multiple testing correction). Clinical information was available for 89 prostatectomy samples with WGS data, with a median follow-up of 1,108 d in which there was biochemical recurrence in 26 patients. The mutations in both genes consisted of a frameshift deletion in one sample and structural variants in the remaining samples.

significant predictors of biochemical recurrence independent of cofactors Gleason, PSA at prostatectomy, and pathological T-stage ($P=0.00061$ (*CDH12*) and 0.0015 (*ANTXR2*), HR=7.3 and 6.5, respectively; Cox regression model, Supplementary Table 4). A Cox regression model containing a combination of *CDH12*, *ANTXR2*, *SPOP*, *IL6ST*, *DLC1* and *MTUS1* mutations was determined to be an optimal predictor of time to biochemical recurrence and was a significant improvement over a baseline model of Gleason, PSA at prostatectomy, and pathological T-stage (model χ^2 test, $P=0.00053$). The number of mutational signatures identified in a cancer was negatively correlated with time to biochemical recurrence in prostatectomy patients ($P=0.014$, HR=3.0; Cox proportional hazards model on number of processes greater than 3, Supplementary Fig. 4) and was an independent predictor ($P=0.0061$, HR=3.6; Cox proportional hazards model). The number of SNVs detected was also an independent prognostic biomarker ($P=0.031$, HR=1.005; Cox proportional hazards model). The numbers of both samples and events in this study are modest, and further analysis of larger cohorts is required to establish firmly these findings.

Druggable targets in the prostate cancer disease network. A key opportunity arising from systematic analyses of cancer genomics is the early identification of therapeutic intervention strategies. To this end, we applied established chemogenomic technologies using the canSAR knowledge base⁵⁴ to map and pharmacologically annotate the cellular network of the prostate disease genes identified in this study. We derived the network using curated protein-protein and transcriptional interaction data. We included the protein products of the genes identified in this study and other key proteins that directly interact with these proteins or affect their function (see Methods and Supplementary Fig. 5). This resulted in a focused prostate network of 156 proteins. Each protein was annotated according to multiple assessments of ‘druggability’; i.e., the likelihood of the protein being amenable to small molecule drug intervention (Table 2 and Supplementary Table 5). We found that prostate cancer driver genes were embedded in a highly druggable cellular network that contains 11 targets of approved therapies and 7 targets of investigational drugs. As well as the AR and the glucocorticoid receptor (GR), the network contains targets of drugs approved for other

Table 2 | Drug targets identified from CanSAR analysis

Target of approved drug

AR, BRAF, ESR1, HDAC3, **KCNH2**, MAP2K1, NR3C1, RARA, RARB, RARG, RXRA

Target of investigational drug

AKT1, ATM, MDM2, PDE4D, PIK3CA, PIK3CB, TP53

Target being investigated chemically

AHR, BRCA1, **CTNNB1, HRAS, IDH1, JUN, MAP3K1, MEN1, NCOR1, NCOR2**, NR4A1, **PIK3R1, PPP2R2A**

Predicted target by structure-based method

ANTXR2, APC, ARNT, ASH1L, BRCA2, CBFA2T2, CDH12, CDK12, CHD1, CREBBP, DLC1, DOCK10, ERG, ETV3, FOXA1, FOXG1, FOXO1, FOXO4, FOXPI, GATA1, GATA2, HDGF, HNF4A, IL6ST, KAT6A, KDM4A, KDM6A, KMT2C, KMT2D, NKX3-1, PIAS1, PIAS2, PTEN, RB1, RGM, RNF43, SKI, SMAD2, SMAD3, SMAD4, SMARCA4, SPDEF, SPOP, TBL1X, TBL1XR1, TBX3, TP73, ZBTB16, ZHX2

Proteins in bold typeface are derived from genes identified as prostate drivers in this study or proteins that have a significant known interaction with these proteins.

indications, several of which (for example, BRAF, ESR1, RARA, RXRA, HDAC3) are under clinical investigation for prostate cancer.

Seven proteins in the prostate network are targets of drugs now in clinical trials. In particular, the ataxia-telangiectasia mutated (ATM) inhibitor AZD-0156, currently in phase 1 for safety assessment, is a likely candidate for exploration in prostate cancer owing to the recently described role of DNA damage repair, particularly in advanced prostate cancers^{21, 55}. The network highlights targets of phosphatidylinositol-3-OH kinase pathway inhibitors (PI3K, AKT1) that are undergoing clinical investigation in prostate cancer, as well as IDH1 and MDM2 drug targets.

To give an indication of the potential of these drugs, we analyzed the most recent drug sensitivity data (GDSC⁵⁶, see URLs). Eighteen drugs acting on our network were tested in GDSC on prostate cancer cell lines. Of these, 5 showed significant effect on growth inhibition and the remaining 13 drugs showed weak activity in at least one cell line (Supplementary Table 6). However, to validate fully the potential of these drugs, extensive drug sensitivity testing needs to be performed in disease-relevant cancer models that correctly reflect the patient population.

Potential future opportunities for prostate cancer therapy are also highlighted by 13 proteins that are under active chemical biology or drug discovery investigation (Table 2). These include menin (MEN1), a component of the MLL–SET1 histone methyltransferase complex. Mice with *Men1* alterations develop prostate cancer⁵⁷, and recent data have shown that menin expression is involved in CRPC⁵⁸. A further 49 proteins are predicted to be druggable and therefore potentially amenable to drug discovery. These include the known prostate cancer protein SPOP, the transcription activator BRG1 (SMARCA4), CDK12 and the CREB binding protein CREBBP.

In summary, we find that 80 of the 156 proteins central to the prostate disease network are either targets of existing drugs or have the potential to be targeted in the future. To maintain an up-to-date view of this analysis, we provide a link to a live page in canSAR (see URLs).

Discussion

The analysis of whole-genome sequence data from 112 prostate cancers has highlighted many of the genetic factors underlying the processes of carcinogenesis, progression, metastasis and the acquisition of drug resistance. Supporting evidence has been provided for 30 candidate driver genes with limited or no previous support, including the noncoding drivers *NEAT1* and *FOXAI*.

Through the timing of genomic aberrations, we have a picture of the possible routes to progression in prostate cancer. Most driver mutations may occur either clonally or subclonally, but mutations in *SPOP* and *ETS* fusions occur early in cancer development and are exclusively clonal. Whereas the gain of 8q and *ETS* fusion appear to be sufficient to drive a dominant clonal expansion, *ETS*⁻ cancers typically need a combination of large-scale losses, acquired over an extended period of time. Known cancer drivers are frequently observed subclonally, and two competing drivers are seen in several cancers. Metastases have less genomic heterogeneity, likely resulting from a bottleneck in achieving metastatic potential.

We observe changes in the mutational processes operative upon cancers during progression. Signature 8 was enriched in subclonal expansions, and signatures 13 and 18 were enriched in metastatic cancers. Cancers with germline or somatic *BRCA1* or *BRCA2* mutations were enriched for signature 3, demonstrating the effect of double-strand repair defects throughout cancer evolution.

Losses of *CDH12* and *ANTXR2* result in poorer recurrence-free survival. We identify 69 prostate cancer-associated proteins that are either targets for existing drugs or new potential targets for therapeutic development.

Analysis of the whole-genome sequences of over 100 prostate cancers has started to reveal the complex evolutionary pathways of these cancers. The early acquisition of driver aberrations, including *ETS* fusions and whole genome duplications, strongly affects the acquisition of subsequent aberrations. Acquisition of individual mutations affects both the subsequent likelihood of metastasis and response to treatment. Network analyses identified, in addition to previously known drivers, targets that could be exploited for clinical investigation with existing drugs as well as targets for new drug discovery, giving potential for the results of genome analysis to be translated rapidly into therapeutic innovation and patient benefit.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at <https://doi.org/10.1038/s41588-018-0086-z>.

Received: 6 March 2017; Accepted: 22 February 2018;

Published online: 16 April 2018

References

- Attard, G. et al. Prostate cancer. *Lancet* **387**, 70–82 (2016).
- 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).
- Grasso, C. S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
- Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
- Barbieri, C. E. et al. Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
- Berger, M. F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
- Lalonde, E. et al. Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. *Lancet Oncol.* **15**, 1521–1532 (2014).
- Cooper, C. S. et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).
- Boutros, P. C. et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
- Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
- Castro, E. et al. Effect of BRCA mutations on metastatic relapse and cause-specific survival after radical treatment for localised prostate cancer. *Eur. Urol.* **68**, 186–193 (2015).
- Kluth, M. et al. Concurrent deletion of 16q23 and *PTEN* is an independent prognostic feature in prostate cancer. *Int. J. Cancer* **137**, 2354–2363 (2015).
- Mosquera, J. M. et al. Concurrent *AURKA* and *MYCN* gene amplifications are harbingers of lethal treatment-related neuroendocrine prostate cancer. *Neoplasia* **15**, 1–10 (2013).
- Rodrigues, L. U. et al. Coordinate loss of *MAP3K7* and *CHD1* promotes aggressive prostate cancer. *Cancer Res.* **75**, 1021–1034 (2015).
- 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).
- Klein, E. A. et al. Decipher genomic classifier measured on prostate biopsy predicts metastasis risk. *Urology* **90**, 148–152 (2016).
- Boström, P. J. et al. Genomic predictors of outcome in prostate cancer. *Eur. Urol.* **68**, 1033–1044 (2015).
- Luca, B.-A. et al. DESNT: a poor prognosis category of human prostate cancer. *Eur. Urol. Focus.* <https://doi.org/10.1016/j.euf.2017.01.016> (2017).
- Ryan, C. J. et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naïve men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. *Lancet Oncol.* **16**, 152–160 (2015).
- Loriot, Y. et al. Effect of enzalutamide on health-related quality of life, pain, and skeletal-related events in asymptomatic and minimally symptomatic, chemotherapy-naïve patients with metastatic castration-resistant prostate cancer (PREVAIL): results from a randomised, phase 3 trial. *Lancet Oncol.* **16**, 509–521 (2015).
- Mateo, J. et al. DNA-repair defects and olaparib in metastatic prostate cancer. *N. Engl. J. Med.* **373**, 1697–1708 (2015).
- James, N. D. et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *Lancet* **387**, 1163–1177 (2016).
- Forbes, S. A. et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
- Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
- Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
- Svensson, C. et al. REST mediates androgen receptor actions on gene repression and predicts early recurrence of prostate cancer. *Nucleic Acids Res.* **42**, 999–1015 (2014).
- Liu, Z. et al. *CASZ1*, a candidate tumor-suppressor gene, suppresses neuroblastoma tumor growth through reprogramming gene expression. *Cell Death Differ.* **18**, 1174–1183 (2011).
- Fischer, K. & Pflugfelder, G. O. Putative breast cancer driver mutations in *TBX3* cause impaired transcriptional repression. *Front. Oncol.* **5**, 244 (2015).
- De Keersmaecker, K. et al. Exome sequencing identifies mutation in *CNOT3* and ribosomal genes *RPL5* and *RPL10* in T-cell acute lymphoblastic leukemia. *Nat. Genet.* **45**, 186–190 (2013).
- Sasaki, M. et al. Regulation of the MDM2-P53 pathway and tumor growth by *PICT1* via nucleolar *RPL11*. *Nat. Med.* **17**, 944–951 (2011).
- Chakravarty, D. et al. The oestrogen receptor alpha-regulated lncRNA *NEAT1* is a critical modulator of prostate cancer. *Nat. Commun.* **5**, 5383 (2014).
- Yang, Y. A. & Yu, J. Current perspectives on *FOXA1* regulation of androgen receptor signaling and prostate cancer. *Genes Dis.* **2**, 144–151 (2015).
- Takayama, K. et al. Integrative analysis of *FOXP1* function reveals a tumor-suppressive effect in prostate cancer. *Mol. Endocrinol.* **28**, 2012–2024 (2014).
- Krohn, A. et al. Recurrent deletion of 3p13 targets multiple tumour suppressor genes and defines a distinct subgroup of aggressive ERG fusion-positive prostate cancers. *J. Pathol.* **231**, 130–141 (2013).
- Carver, B. S. et al. Aberrant ERG expression cooperates with loss of *PTEN* to promote cancer progression in the prostate. *Nat. Genet.* **41**, 619–624 (2009).
- King, J. C. et al. Cooperativity of *TMPRSS2-ERG* with *PI3*-kinase pathway activation in prostate oncogenesis. *Nat. Genet.* **41**, 524–526 (2009).
- Kluth, M. et al. Clinical significance of different types of p53 gene alteration in surgically treated prostate cancer. *Int. J. Cancer* **135**, 1369–1380 (2014).
- Burkhardt, L. et al. *CHD1* is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. *Cancer Res.* **73**, 2795–2805 (2013).
- Liu, W. et al. Identification of novel *CHD1*-associated collaborative alterations of genomic structure and functional assessment of *CHD1* in prostate cancer. *Oncogene* **31**, 3939–3948 (2012).
- Biankin, A. V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
- Heun, P. SUMO organization of the nucleus. *Curr. Opin. Cell Biol.* **19**, 350–355 (2007).
- Kaikkonen, S. et al. SUMO-specific protease 1 (*SEN1*) reverses the hormone-augmented SUMOylation of androgen receptor and modulates gene responses in prostate cancer cells. *Mol. Endocrinol.* **23**, 292–307 (2009).
- Smith, D. I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett.* **232**, 48–57 (2006).

44. Taylor, B. S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
45. Williams, J. L., Greer, P. A. & Squire, J. A. Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. *Cancer Genet.* **207**, 474–488 (2014).
46. Chen, Z. et al. Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. *Nature* **436**, 725–730 (2005).
47. Bolli, N. et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
48. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
49. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
50. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
51. Polkinghorn, W. R. et al. Androgen receptor signaling regulates DNA repair in prostate cancers. *Cancer Discov.* **3**, 1245–1253 (2013).
52. Goodwin, J. F. et al. DNA-PKcs-mediated transcriptional regulation drives prostate cancer progression and metastasis. *Cancer Cell* **28**, 97–113 (2015).
53. Tarish, F. L. et al. Castration radiosensitizes prostate cancer tissue by impairing DNA double-strand break repair. *Sci. Transl. Med.* **7**, 312re11 (2015).
54. Tym, J. E. et al. canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.* **44**, D938–D943 (2016). D1.
55. Leongamornlert, D. et al. Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. *Br. J. Cancer* **110**, 1663–1672 (2014).
56. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).
57. Seigne, C. et al. Characterisation of prostate cancer lesions in heterozygous Men1 mutant mice. *BMC Cancer* **10**, 395 (2010).
58. Malik, R. et al. Targeting the MLL complex in castration-resistant prostate cancer. *Nat. Med.* **21**, 344–352 (2015).
59. Rudnicka, C. et al. Overexpression and knock-down studies highlight that a disintegrin and metalloproteinase 28 controls proliferation and migration in human prostate cancer. *Medicine (Baltimore)* **95**, e5085 (2016).
60. Zhang, H. et al. FOXO1 inhibits Runx2 transcriptional activity and prostate cancer cell migration and invasion. *Cancer Res.* **71**, 3257–3267 (2011).
61. Malinowska, K. et al. Interleukin-6 stimulation of growth of prostate cancer in vitro and in vivo through activation of the androgen receptor. *Endocr. Relat. Cancer* **16**, 155–169 (2009).
62. FitzGerald, L. M. et al. Identification of a prostate cancer susceptibility gene on chromosome 5p13q12 associated with risk of both familial and sporadic disease. *Eur. J. Hum. Genet.* **17**, 368–377 (2009).
63. Zhao, W., Cao, L., Zeng, S., Qin, H. & Men, T. Upregulation of miR-556-5p promoted prostate cancer cell proliferation by suppressing PPP2R2A expression. *Biomed. Pharmacother.* **75**, 142–147 (2015).
64. Parray, A. et al. ROBO1, a tumor suppressor and critical molecular barrier for localized tumor cells to acquire invasive phenotype: study in African-American and Caucasian prostate cancer models. *Int. J. Cancer* **135**, 2493–2506 (2014).
65. Daniels, G. et al. TBLR1 as an androgen receptor (AR) coactivator selectively activates AR target genes to inhibit prostate cancer growth. *Endocr. Relat. Cancer* **21**, 127–142 (2014).
66. Jones, S. et al. Somatic mutations in the chromatin remodeling gene *ARID1A* occur in several tumor types. *Hum. Mutat.* **33**, 100–103 (2012).
67. Collart, M. A., Kassem, S. & Villanyi, Z. Mutations in the *NOT* genes or in the translation machinery similarly display increased resistance to histidine starvation. *Front. Genet.* **8**, 61 (2017).

Acknowledgements

The authors thank those men with prostate cancer and the subjects who have donated their time and their samples to the Cambridge, Oxford, The Institute of Cancer Research, John Hopkins and University of Tampere BioMediTech Biorepositories for this study. We also acknowledge support of the research staff in S4 who so carefully curated the samples and the follow-up data (J. Burge, M. Corcoran, A. George and S. Stearn). We thank M. Stratton for discussions when setting up the CR-UK Prostate Cancer ICGC Project. We acknowledge support from Cancer Research UK C5047/A14835/A22530/A17528, C309/A11566, C368/A6743, A368/A7990, C14303/A17197 (Z.K.-J., S. Merson, N.C., S.E., D.L., T. Dadaev, M.A., E.B., J.B., G.A., P.W., B.A.-L., D.S.B., C.S.C., R.A.E.), the Dallaglio Foundation (CR-UK Prostate Cancer ICGC Project and Pan Prostate Cancer Group), PC-UK/Movember (Z.K.-J.), the NIHR support to The Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust (Z.K.-J., N.D., S. Merson, N.C., S.E., D.L., T. Dadaev, S. Thomas, M.A., E.B., C.F., N.L., D.N., V.K., N.A., P.K., C.O., D.C., A.T., E.M., E.R., T. Dudderidge, S. Hazell, J.B., G.A., P.W., B.A.-L., D.S.B., C.S.C., R.A.E.), Cancer Research UK funding to The Institute of Cancer Research and the Royal Marsden NHS Foundation Trust CRUK Centre, the National Cancer Research Institute (National Institute of Health Research (NIHR) Collaborative Study: “Prostate Cancer: Mechanisms of Progression and Treatment (PROMPT)” (grant G0500966/75466) (D.E.N., V.G.), the Li Ka Shing Foundation (D.C.W., D.J.W.) and the Academy of Finland and Cancer Society of Finland (G.S.B.). We thank the National Institute for Health Research, Hutchison Whampoa Limited, University of Cambridge and the Human Research Tissue Bank (Addenbrooke’s Hospital), which is supported by the NIHR Cambridge Biomedical Research Centre; The Core Facilities at the Cancer Research UK Cambridge Institute, Orchid and Cancer Research UK, D. Holland from the Infrastructure Management Team, and P. Clapham from the Informatics Systems Group at the Wellcome Trust Sanger Institute. D.M.B. is supported by Orchid. C.V.’s academic time was supported by the NIHR Oxford Biomedical Research Centre (Molecular Diagnostics Theme/Multimodal Pathology sub-theme). We also acknowledge support from the Bob Champion Cancer Trust, The Masonic Charitable Foundation successor to The Grand Charity, The King Family and the Stephen Hargrave Trust (C.S.C., D.S.B.). P.W. is a Cancer Research Life Fellow. We acknowledge core facilities provided by CRUK funding to the CRUK ICR Centre, the CRUK Cancer Therapeutics Unit and support for canSAR C35696/A23187 (P.W., G.A.).

Author contributions

R.A.E., C.S.C., D.E.N., D.S.B., C.S.F., G.S.B., A.G.L., P.W., B.A.-L., D.C.W., F.C.H. and D.F.E. designed the study. R.A.E., C.S.C., D.C.W. and D.S.B. wrote the paper. M.F., P.C.B., R.G.B. and all other authors contributed to revisions. Z.K.-J., H.W., C.E.M., D.E.N., V.G., A.G.L., R.A.E., F.C.H., G.S.B., A.Y.W., C.S.F., C.V., D.M.B., N.D., S. Merson, S. Hawkins, W.H., Y.-J.L., A.L., J.K., K.K., H.L., L. Marsden, S.E., L. Matthews, A.N., Y.Y., H.Z., S.T., E.B., C.F., N.L., S. Hazell, D.N., P.G., V.K., N.V.A., P.K., C.O., D.C., A.T., E.M., E.R., T. Dudderidge, C.C., W.I., T.V. and N.C.S. coordinated sample collection, pathology review and processing. D.C.W., G.G., T.M., I.M., D.J.W., D.S.B., M.G., J.Z., A.B., L.B.A., S.D., B.K., N.C., V.B., D.L., S. McLaren, T. Dadaev, M.A., S.T., C.G., K.R., D.J., A.M., L.S., J.T., A.F., P.V.L. and U.M. supported, directed and performed the analyses. C.S. and the TCGA, J.d.B. and G.A. provided data for the meta-analysis. D.F.E., A.G.L., G.S.B., C.S.F., D.S.B., D.E.N., C.S.C. and R.A.E. are joint principal investigators for the CR-UK Prostate Cancer ICGC Project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-018-0086-z>.

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

Correspondence and requests for materials should be addressed to D.C.W. or R.A.E.

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

¹Oxford Big Data Institute, University of Oxford, Oxford, UK. ²Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. ³Oxford NIHR Biomedical Research Centre, Oxford, UK. ⁴Memorial Sloan-Kettering Cancer Center, New York, NY, USA. ⁵Department of Urology, Addenbrooke’s Hospital, Cambridge, UK. ⁶Uro-Oncology Research Group, Cancer Research UK, Cambridge Institute, Cambridge, UK. ⁷Molecular Diagnostics and Therapeutics Group, University College London, London, UK. ⁸The Institute of Cancer Research, London, UK. ⁹Cancer Genomics, The Francis Crick Institute, London, UK. ¹⁰Early Detection Programme, Cancer Research UK Cambridge Centre, Department of Oncology, University of Cambridge, Cambridge, UK. ¹¹Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. ¹²Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. ¹³Centre for Molecular Oncology, Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK. ¹⁴Royal Marsden NHS Foundation Trust, London and Sutton, UK. ¹⁵Statistics and Computational Biology Laboratory, Cancer Research UK Cambridge Institute, Cambridge, UK. ¹⁶The Chinese University of Hong Kong, Shatin, Hong Kong, China. ¹⁷Second Military Medical University, Shanghai, China. ¹⁸Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. ¹⁹Norwich Medical School, University

of East Anglia, Norwich, UK. ²⁰Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. ²¹St George's Healthcare NHS Trust, London, UK. ²²Full lists of members and affiliations appear in the Supplementary Note. ²³Johns Hopkins School of Medicine, Baltimore, MD, USA. ²⁴Institute of Biosciences and Medical Technology, BioMediTech, University of Tampere and Fimlab Laboratories, Tampere University Hospital, Tampere, Finland. ²⁵Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada. ²⁶Ontario Institute for Cancer Research, Toronto, ON, Canada. ²⁷Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. ²⁸Department of Pharmacology and Toxicology, University of Toronto, Toronto, ON, Canada. ²⁹Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada. ³⁰cBio Center, Dana-Farber Cancer Institute & Harvard Medical School, Boston, MA, USA. ³¹School of Mathematics and Statistics/School of Medicine, University of St. Andrews, Fife, UK. ³²University of Liverpool, Liverpool, UK. ³³HCA Laboratories, London, UK. ³⁴Earlham Institute, Norwich, UK. ³⁵These authors contributed equally: David C. Wedge, Gunes Gundem and Thomas Mitchell. ³⁶These authors jointly supervised this work: Bissan Al-Lazikani, Paul Workman, Andrew G. Lynch, G. Steven Bova, Christopher S. Foster, Daniel S. Brewer, David E. Neal, Colin S. Cooper and Rosalind A. Eeles. *e-mail: david.wedge@bdi.ox.ac.uk; rosalind.eeles@icr.ac.uk

Methods

Patient cohorts, samples and ethics. We have complied with all relevant ethical regulations. We collected 92 cancer samples from prostatectomy patients treated at The Royal Marsden NHS Foundation Trust, London, at the Addenbrooke's Hospital, Cambridge, at Oxford University Hospitals NHS Trust and at Changhai Hospital, Shanghai, China as described previously⁶⁸. Clinical details for the patients are shown in Supplementary Table 7. Ethical approval was obtained from the respective local ethics committees and from The Trent Multicentre Research Ethics Committee. All patients were consented to ICGC standards (see URLs). Twenty men from PELICAN (Project to ELIminate lethal CANcer)⁶⁹, an integrated clinical-molecular autopsy study of metastatic prostate cancer, were subjects in the current study. Subjects consented to participate in the Johns Hopkins Medicine IRB-approved study between 1995 and 2005 (Supplementary Table 7). Subject A17 had a germline *BRCA1* mutation, as previously reported⁷⁰.

DNA preparation and DNA sequencing. DNA from whole blood samples and frozen tissue was extracted and quantified using a dsDNA assay (UK-Quant-iT PicoGreen dsDNA Assay Kit for DNA) following the manufacturer's instructions with a fluorescence microplate reader (Biotek SynergyHT, Biotek). Acceptable DNA had a concentration of at least 50 ng/μl in TE (10 mM Tris, 1 mM EDTA) and an OD_{260/280} between 1.8 and 2.0. WGS was performed at Illumina, Inc. (Illumina Sequencing Facility, San Diego, CA, USA) or the BGI (Beijing Genome Institute, Hong Kong), as described previously, to a target depth of 50× for the cancer samples and 30× for matched controls⁸. The Burrows-Wheeler Aligner was used to align the sequencing data to the GRCh37 reference human genome⁷¹.

Variant calling pipeline. SNVs, insertions and deletions were detected using the Cancer Genome Project Wellcome Trust Sanger Institute pipeline as described previously⁸. In brief, SNVs were detected using CaVEMan with a cut-off 'somatic' probability of 95%. Postprocessing filters were applied. Insertions and deletions were called using a modified version of Pindel⁷². Variant allele frequencies of all indels were corrected by local realignment of unmapped reads against the mutant sequence. Structural variants were detected using Brass⁸. A positive ETS status was assigned if a breakpoint between *ERG*, *ETV1* or *ETV4* and previously reported partner DNA sequences was detected.

Mutation burdens. Mutation burdens were compared between primary and metastatic samples and between ADT and hormone-naive samples using a negative binomial generalized linear model (GLM), implemented with the R package MASS. Sample type was found to be an independent predictor of number of SNVs, as was age at time of sampling.

Timing of copy number events. We developed an approach to order the occurrence of CNAs by combining three sources of information: (i) clonality of CNAs; (ii) timing relative to whole-genome duplication, and; (iii) timing of homozygous deletions relative to neighboring hemizygous losses. Information from all tumors was combined using a Bradley-Terry model to give the most likely ordering of events during progression of prostate cancer.

The Battenberg algorithm was used to detect clonal and subclonal somatic CNAs and to estimate ploidy and cancer content from the next-generation sequencing data as previously described⁷³. Briefly, germline heterozygous SNPs were phased using IMPUTE2 and a- and b-alleles were assigned. Data were segmented using piecewise constant fitting⁷⁴ and subclonal copy-number segments were identified via a *t*-test as those with b-allele frequencies that differed significantly from the values expected of a clonal copy number state. Ploidy and cancer purity were estimated with the same method used by ASCAT⁷⁵.

In this cohort, we defined WGD samples as those that had an average ploidy greater than 3. For tumors that had not undergone WGD, gains were defined as those regions that had at least one allele with copy number greater than 1, while losses were defined as those segments that had undergone LOH. For tumors that had undergone WGD, losses were called in those segments with at least one allele with copy number of less than 2, whereas gains were called for those with an allelic copy number greater than 2. An extension of this logic was used for subclonal copy number segments; the evolving cellular fraction was always defined as that which deviated away from overall ploidy (defined as 2 for non-WGD samples and 4 for WGD samples). For example, if 75% of cells in a non-WGD tumor have a copy number of 3 + 1 at a given genomic loci, with the remaining 25% of cells having a copy number of 2 + 1, then we assume there has been clonal gain to 2 + 1 and then a subclone containing 75% of cells has undergone a further gain.

Three independent approaches were used to extract evolutionary data from each cancer sample. The first involved ordering clustered subclonal cancer fractions, the second used implicit ordering of clonal homozygous deletions in relation to losses, and the third estimated the relative timing of whole genome duplication. The logical arguments used in each approach are considered in turn. First, Battenberg algorithm-derived estimates for the cellular fraction and s.d. of each subclonal aberration were input to a Markov chain Monte Carlo hierarchical Bayesian Dirichlet process to group linked events together in an unsupervised manner. This defined clusters of different cell populations, each present at a calculated cancer cell fraction. The pigeonhole principle was then

used to determine the hierarchical relationship between these clusters. Using this process, gains, losses and homozygous deletions were ordered with the following caveat to ensure that only independent events are ordered: if there was a clonal and subclonal gain (or loss) at the same locus, then only the clonal or initial gain (or loss) was ordered. Second, homozygous deletions have implicitly occurred after loss of heterozygosity at the same locus. Finally, the parsimony principle was used to define the allele counts that correspond to early and late changes in relation to WGD. For losses, if the minor allele copy number equals 0, then the loss occurred before WGD. Otherwise the loss occurred after WGD. Regarding gains, if the major allele copy number is twice or greater than ploidy, then the gain occurred before WGD. Otherwise, the gain occurred after WGD.

The above arguments allow us to gain insights into the order of copy number events in each individual tumor sample. To establish a consensus order across a cohort of tumor samples requires the ordering data to be integrated across all samples. As specific copy number events (location of breakpoints and the individual copy number states) tend to be unique to individual samples, we defined reference copy number segments that occurred recurrently. These were then used to build an overall contingency table.

The reference genomic segments were defined as regions that were recurrently aberrant. Regions of significant recurrence (false detection rate (FDR), $P < 0.05$) were determined by performing 100,000 simulations, placing the CNAs detected from each sample in random locations within the genome. The process was repeated for gains, LOH and homozygous deletions and the randomly generated copy number landscape compared to that arising from this cohort provided significance levels. Each significantly aberrant region was initially segmented using all breakpoints from all the events that contributed to that region. For instance, the significantly enriched region for LOH:chr8:0–44Mb contains over 300 breakpoints drawn from all the samples that contain LOH at chromosome 8p. We computed significantly recurrent regions and reference segments for both ETS⁺ and ETS⁻ sample subgroups.

Performing pair-wise comparisons between all segmented results using the Bradley-Terry method described below proved computationally expensive, and therefore the total number of segments used in the pairwise comparison was rationalized by grouping reference segments to make combined segments of minimum length 1 Mb.

We then considered each tumor sample in turn. If any copy number event overlapped the reference genomic segments and was ordered in relation to any other event (that also overlapped regions of significance), those overlapped reference segments were ordered in comparison to other overlapped reference segments. In addition to these reference segments, the *TMPRSS2-ERG* deletion was ordered more stringently by considering only those segments that could result in the gene fusion, and not merely overlap the locus. In this manner, a contingency table of contests was constructed using reference genomic segments as the variables. We built contingency tables for both ETS⁺ and ETS⁻ tumor samples to determine whether their evolutionary trajectory differed significantly.

An implementation of the Bradley-Terry model of pairwise comparison in R^{26} with bias reduced maximum likelihood estimated the ability or overall order of each individual reference segment.

Subclonal analysis. The fraction of each cancer genome with subclonal CNAs was calculated as the total amount of the genome with subclonal CNA, as identified by the Battenberg algorithm, divided by the total amount of the genome that had CNAs. One sample (PD13397a, Supplementary Table 8) was identified as having very low cellularity, as it had a completely flat copy number profile and only 411 identified SNVs. Since CNAs could not be called in this sample, it was not possible to adjust allele frequencies to cancer cell fractions (CCFs), and this sample was excluded from subclonality analysis. SNVs and indels were separately clustered using a Bayesian Dirichlet process, as previously described⁴⁷. Clonal variants are expected to cluster at a CCF close to 1.0. However, in 18 tumors (Supplementary Table 8), there was no cluster in the range 0.95–1.05. The likely cause of a shift in CCF is inaccuracy in copy number calling, and these samples therefore failed quality control and were excluded from subclonality analysis. From Markov chain Monte Carlo sampling carried out in the Dirichlet process model, the posterior probability of each variant having a CCF below 0.95 was estimated. Variants with a probability above 80% were designated as 'subclonal', those with probability below 20% were designated 'clonal', and those with intermediate probabilities were designated as 'uncertain'. The fraction of subclonal variants used in Fig. 5 and Supplementary Fig. 2 was then calculated after excluding uncertain variants.

Mutational spectra. The mutational spectra, defined by the triplets of nucleotides around each mutation of each sample, were deconvoluted into mutational processes as previously described^{68,77}. Clonal and subclonal variants were separated as defined above. Further separation of clonal mutations was performed for mutations in genomic regions that had undergone copy number gains. These mutations were classified as 'early' or 'late' depending whether their observed allele frequencies were more likely to indicate their presence on 2 or 1 chromosome copies, respectively, as assessed by binomial probability. Assignment of mutations to mutational signatures was carried out on each subset of mutations (early, late, clonal, subclonal), as well as on all mutations from each sample (Supplementary Table 3).

Clinical survival analyses. A Cox regression model was fitted to 71 features: every gene with mutations (breakpoints, substitutions or indels) with a potential functional impact (missense, nonsense, start-lost, in-frame, frameshift, or occurred in a noncoding transcript) or a CNA highlighted by the CNA analysis that occurred in three of more prostatectomy patients. The endpoint was biochemical recurrence. *P*-values were adjusted for multiple testing using the Benjamini–Hochberg method. Multivariate analyses were performed on all genes found to be significant using discretized Gleason (6, 7, 8 or 9), pathological T-stage (T2, T3) and PSA at prostatectomy as cofactors. Gene selection for the optimal predictor of time to biochemical recurrence was determined using Lasso⁷⁸, a shrinkage and selection method for linear regression, starting with all genes that had a significant association with time to biochemical recurrence. Standard algorithms were used for survival analyses and statistical associations.

Identifying new oncogenes. The joint dataset was compiled from the aggregation of variants called in our samples with three other datasets, yielding a total of 930 samples, comprising 710 primary and 220 metastatic samples: (i) TCGA⁴, comprising 425 primary cancer samples, whole-exome sequencing with SureSelect Exome v3 baits on Illumina HiSeq 2000, average coverage ~100×; (ii) COSMIC database²², comprising 243 samples, curated set of mutations from several sources (see URLs); (iii) Stand Up to Cancer²³ (SU2C-PCF), comprising 150 metastatic castration-resistant samples, paired-end, whole-exome sequencing with SureSelect Exome v4 baits on Illumina HiSeq2000, average coverage ~160×.

To identify coding and noncoding drivers from SNVs and indels, we used two previously described methods⁵⁰. Coding drivers on the joint dataset (930 cancers) were identified using dNdScv, a dN/dS method designed to quantify positive selection in cancer genomes. dNdScv models somatic mutations in a given gene as a Poisson process. Inferences on selection are carried out separately for missense substitutions, truncating substitutions (nonsense and essential splice site mutations) and indels and then combined into a global *P*-value per gene. Noncoding recurrence was studied using NBR. Both dNdScv and NBR model the variation of the mutation rate across the genome using a negative binomial regression with covariates. First, Poisson regression is used to obtain maximum-likelihood estimates for the 192 rate parameters (r_j) describing each of the possible trinucleotide substitutions in a strand-specific manner. $r_j = n_j/L_j$, where n_j is the total number of mutations observed across samples of a given trinucleotide class (j) and L_j is the number of available sites for each trinucleotide. These rates are used to estimate the total number of mutations across samples expected under neutrality in each element considering the mutational signatures active in the cohort and the sequence of the elements ($E_n = \sum_j r_j L_{j,n}$). This estimate assumes no variation of the mutation rate across elements in the genome. Second, a negative binomial regression is used to refine this estimate of the background mutation rate of an element, using covariates and E_n as an offset. Both methods identify genes or noncoding regions with higher than expected mutation recurrence, correcting for gene length, sequence composition, mutation signatures acting across patients and the variation of the mutation rate along the genome. A QQ plot confirmed that *P*-values obtained from this method in this cohort were not subject to inflation and consequent over-calling of driver genes (Supplementary Fig. 6).

Chromoplexy, characterized by highly clustered genomic breakpoints that occur in chains and are sometimes joined by deletion bridges, has been shown to be prevalent in prostate cancer²⁵. To identify rearrangement drivers, we first used ChainFinder²⁵ to account for any bias toward regions with chromoplexy and identified ‘unique’ rearranged regions per sample taking the midpoint between all the breakpoints ChainFinder assigns to the same chromoplexy event. Next, separately aggregating the ICGC samples with and without ERG fusions, we calculated interbreakpoint distance and performed piecewise constant fitting (PCF)⁷⁴ to identify genomic regions that were recurrently rearranged in multiple samples. Rearranged regions with potential functional impact were identified using two criteria: a minimum threefold difference in the number of SVs per megabase of ERG⁺ and ERG⁻ samples, and presence of at least one gene with multiple samples with truncating events—i.e., homozygous deletion, stop codon, frameshift indel or essential splice site mutation. In addition, several identified regions were significantly enriched for LOH in either ETS⁺ or ETS⁻ samples, from copy number analysis (see above). The variants identified in key regions are depicted in Fig. 3.

Chemogenomics annotation of the prostate cancer network. To construct the network, we used the 71 protein products of the 73 genes identified in this study (henceforth referred to as prostate proteins) to seed a search for all possible interacting proteins in the canSAR interactome³⁴. This interactome contains merged and curated data from the IMEx Consortium⁷⁹, Phosphosite (see URLs) and other databases. It includes (i) interactions where there were more than two publications reporting experiments demonstrating the binary interaction between the two proteins, (ii) interactions where there is 3D protein structural evidence of a direct complex, (iii) interactions where there are at least two publications reporting that one protein is a substrate of the other, and (iv) interactions where there are at least two papers reporting that one protein is the product of a gene under the direct regulatory control of the other. It excludes (i) interactions that were inferred from a large immunoprecipitation experiment without follow-up to demonstrate the specific binary interaction, (ii) interactions inferred from text mining, and

(iii) interactions inferred from co-occurrence in publications or from gene expression correlation.

The initial prostate cancer seeded network resulted in a large collection of 3,290 proteins that have some experimental evidence of interacting with at least one prostate protein. When we add extra proteins to the network, we wanted to ensure that we only add proteins that are more likely to function primarily through interaction with the proteins in the network rather than just be generic major hubs. To this end, we carried out the following steps. First, starting with the input (prostate protein) list, we obtained all possible first neighbors. We then computed, for each new protein, the proportion of its first neighbors that are in the original input list. To define the proteins that are most likely to function through our network, we calculated the chances of these proportions occurring in a random network. We did this by randomizing our interactome 10,000 times and computing how often the observed proportions can be achieved by chance (empirical *P*-value). We corrected the *P*-values for multiple testing and retained only proteins that had corrected FDR $P < 0.05$ (Supplementary Fig. 5). We performed network minimization to maintain only proteins that are strongly connected to more than one prostate protein or whose only connection is to one of the prostate proteins. We identified a prostate cancer network of 156 proteins. Using canSAR's Cancer Protein Annotation Tool (CPAT)⁸⁰, we annotated the 156 proteins with pharmacological and druggability data. We labeled proteins that are (i) targets of approved drugs, (ii) targets of drugs under clinical investigation, (iii) targets of preclinical or discovery stage compounds that are active at concentrations equal to or less than 100 nM against the protein of interest, or (iv) proteins that we predict to be druggable using our structural druggability prediction protocols^{80–83} but that have few or no published active inhibitors—these are potential targets for future drug discovery.

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

Code availability. Alignment and variant calling was performed using analysis pipelines in the Cancer Genome Project (CGP) at the Wellcome Trust Sanger Institute. Software versions applied to each sample are listed in Supplementary Table 9. The CGP pipelines can be downloaded (see URLs). Chromoplexy was called using ChainFinder version 1.0.1. ChainFinder can be downloaded (see URLs). The Battenberg algorithm was used to call clonal and subclonal CNAs in all samples. The Battenberg pipeline can be downloaded (see URLs). Putative drug targets were identified using CanSAR version 3.0. Data analysis was carried out using R, version 3.0.0.

Data availability. Sequencing data that support the findings of this study have been deposited in the European Genome-phenome Archive with the accession code EGAS00001000262. Sample-specific accession codes are listed in Supplementary Table 7.

URLs. GDSC, <http://www.cancerrxgene.org>; canSAR, http://cansar.icr.ac.uk/cansar/publications/sequencing_prostate_cancers_identifies_new_cancer_genes_routes_progression_and_drug_targets/; ICGC, <https://icgc.org/dockstore-cgpwgs>; <https://github.com/cancerit/dockstore-cgpwgs>; Battenberg pipeline, <https://github.com/Wedge-Oxford/battenberg>; Phosphosite, <http://www.phosphosite.org/>; ChainFinder, <http://archive.broadinstitute.org/cancer/cga/chainfinder/>; COSMIC database, <http://cancer.sanger.ac.uk/cosmic>.

References

- Mao, X. et al. Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis. *Cancer Res.* **70**, 5207–5212 (2010).
- Liu, W. et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat. Med.* **15**, 559–565 (2009).
- Nickerson, M. L. et al. Somatic alterations contributing to metastasis of a castration-resistant prostate cancer. *Hum. Mutat.* **34**, 1231–1241 (2013).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
- Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
- Nilsen, G. et al. Copynumber: efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
- Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107**, 16910–16915 (2010).
- Firth, D. & Turner, H. L. Bradley–Terry models in R: the BradleyTerry2 package. *J. Stat. Softw.* **48**, 1–21 (2012).
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).

78. Shin, S., Fine, J. & Liu, Y. Adaptive estimation with partially overlapping models. *Stat. Sin* **26**, 235–253 (2016).
79. Orchard, S. et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).
80. Patel, M. N., Halling-Brown, M. D., Tym, J. E., Workman, P. & Al-Lazikani, B. Objective assessment of cancer genes for drug discovery. *Nat. Rev. Drug Discov.* **12**, 35–50 (2013).
81. Bulusu, K. C., Tym, J. E., Coker, E. A., Schierz, A. C. & Al-Lazikani, B. canSAR: updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.* **42**, D1040–D1047 (2014).
82. Mitsopoulos, C., Schierz, A. C., Workman, P. & Al-Lazikani, B. Distinctive behaviors of druggable proteins in cellular networks. *PLoS Comput. Biol.* **11**, e1004597 (2015).
83. Workman, P. & Al-Lazikani, B. Drugging cancer genomes. *Nat. Rev. Drug Discov.* **12**, 889–890 (2013).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

We used Lawrence et al (PMID 24390350) as a guide to appropriate sample size. This paper reports that out merged dataset (930 samples) should be sufficient to detect 90% of driver genes with prevalence 2% with 90% power, given the average mutation rate within our samples.

2. Data exclusions

Describe any data exclusions.

Samples were excluded from the subclonal analysis if they had a poor copy number profile.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No experiments were used within this study

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not applicable

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not applicable. All samples were run through common analysis pipelines.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted
- A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

The CGP pipeline was used to call SNVs, SVs and indels, and is available from <https://github.com/cancerit/dockstore-cgpngs>. The Battenberg algorithm was used to call copy number aberrations and is available from <https://github.com/Wedge-Oxford/battenberg>. We have included a section on 'Code availability' on page 22.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* [guidance for providing algorithms and software for publication](#) may be useful for any submission.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Not applicable.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in this study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used within this study

b. Describe the method of cell line authentication used.

NA

c. Report whether the cell lines were tested for mycoplasma contamination.

NA

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

NA

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animal studies were used

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Relevant information is provided in Supplementary Table 7, i.e. age, Gleason score, T stage, PSA, primary / metastasis