

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

David C. Wedge^{1,2,34*}, Gunes Gundem^{2,3,34}, Thomas Mitchell^{2,4,5,34}, Dan J. Woodcock¹, Inigo Martincorena², Mohammed Ghor², Jorge Zamora², Adam Butler², Hayley Whitaker⁶, Zsofia Kote-Jarai⁷, Ludmil B. Alexandrov², Peter Van Loo^{2,8}, Charlie E. Massie⁵, Stefan Dentro^{2,8}, Anne Y. Warren⁹, Clare Verrill¹⁰, Dan M. Berney¹¹, Nening Dennis¹², Sue Merson⁷, Steve Hawkins⁵, William Howat⁹, Yong-Jie Yu¹¹, Adam Lambert¹³, Jonathan Kay⁶, Barbara Kremeyer², Katalin Karaszi¹³, Hayley Luxton⁶, Niedzica Camacho^{7,3}, Luke Marsden¹³, Sandra Edwards⁷, Lucy Matthews¹³, Valeria Bo¹⁴, Daniel Leongamornlert⁷, Stuart McLaren², Anthony Ng¹⁵, Yongwei Yu¹⁶, Hongwei Zhang¹⁶, Tokhir Dadaev⁷, Sarah Thomas¹², Douglas F. Easton^{17,33}, Mahbubl Ahmed⁷, Elizabeth Bancroft^{7,12}, Cyril Fisher¹², Naomi Livni¹², David Nicol¹², Simon Tavaré¹⁴, Pelvender Gill¹³, Christopher Greenman¹⁸, Vincent Khoo¹², Nicholas Van As¹², Pardeep Kumar¹², Christopher Ogden¹², Declan Cahill¹², Alan Thompson¹², Erik Mayer¹², Edward Rowe¹², Tim Dudderidge¹², Vincent Gnanapragasam^{4,19}, Nimish C. Shah⁴, Keiran Raine², David Jones², Andrew Menzies², Lucy Stebbings², Jon Teague², Steven Hazell¹², CAMCAP study group, Johann de Bono⁷, Gerhardt Attard⁷, William Isaacs²⁰, Tapio Visakorpi²¹, Michael Fraser²², Paul C Boutros^{23,24,25}, Robert G Bristow^{22,24,26}, Paul Workman⁷, Chris Sander²⁷, The TCGA consortium²⁸, Freddie C. Hamdy¹³, Andrew Futreal², Ultan McDermott², Bissan Al-Lazikani^{7,35}, Andrew G. Lynch^{14,33,35}, G. Steven Bova^{21,20,33,35}, Christopher S. Foster^{29,30,33,35}, Daniel S. Brewer^{7,18,31,33,35}, David Neal^{5,19,33,35}, Colin S. Cooper^{7,18,33,35}, and Rosalind A. Eeles^{7,12,33,35*}

¹ Oxford Big Data Institute, University of Oxford, Old Road Campus, Oxford, OX3 7LF, UK

² Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK

³Memorial Sloan-Kettering Cancer Center, NY 10065, New York, USA

⁴Department of Urology, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK

⁵Uro-Oncology Research Group, Cancer Research UK, Cambridge Institute, Cambridge, CB2 0RE, UK

⁶Molecular Diagnostics and Therapeutics Group, University College London WC1E 6BT

⁷The Institute Of Cancer Research, London, SW7 3RP, UK

⁸Cancer Genomics, The Francis Crick Institute, London, NW1 1AT, UK

⁹Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK

¹⁰Oxford University Hospitals NHS Trust, John Radcliffe Hospital, Oxford, OX3 9DU, UK

¹¹Centre for Molecular Oncology, Barts Cancer Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AD, UK

¹²Royal Marsden NHS Foundation Trust, London and Sutton, SM2 5PT, UK

¹³The University of Oxford, Oxford, OX1 2JD, UK

¹⁴Statistics and Computational Biology Laboratory, Cancer Research UK Cambridge Institute, Cambridge, CB2 0RE, UK

¹⁵The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China

¹⁶Second Military Medical University, Shanghai, China 20043

¹⁷Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, CB1 8RN, UK

¹⁸Norwich Medical School, University of East Anglia, Norwich, NR4 7TJ, UK

¹⁹Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK

²⁰Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

²¹Institute of Biosciences and Medical Technology, BioMediTech, University of Tampere and Fimlab Laboratories, Tampere University Hospital, Tampere, FI-33520, Finland

²²Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada M5G 2M9

²³Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3

²⁴Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5G 1L7

²⁵Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada M5S 1A8

²⁶ Department of Radiation Oncology, University of Toronto, Toronto, Canada

²⁷cBio Center, Dana-Farber Cancer Institute & Harvard Medical School, Boston, MA 02215, USA

²⁸The TCGA consortium, National Cancer Institute @ NIH, Bethesda, MD20892, USA

²⁹University of Liverpool, Liverpool, UK

³⁰HCA Laboratories, London, WC1E 6JA, UK

³¹Earlham Institute, Norwich, NR4 7UH

³²School of Mathematics and Statistics/School of Medicine, University of St. Andrews,
KY16 9SS, UK

³³Joint PIs of CRUK-Prostate Cancer ICGC Project

³⁴Joint first authors made an equal contribution to this paper

³⁵Joint last authors made an equal contribution to this paper

*Correspondence should be addressed to: D.C.W (david.wedge@bdi.ox.ac.uk) & R.A.E
(Rosalind.eeles@icr.ac.uk)

Abstract

Prostate cancer (PCa) represents a significant 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 PCa samples. From joint analysis of these cancers with those from previous studies, 930 cancers in total, we identified evidence for 22 novel putative coding driver genes, as well as evidence for *NEAT1* and *FOXAI* acting as drivers through non-coding mutations. Through the temporal dissection of aberrations, we identified driver mutations specifically associated with steps in the progression of PCa, for example establishing loss of *CHD1* and *BRCA2* as early events in cancer development of *ETS* fusion negative cancers. Computational chemogenomic (CanSAR) analysis of PCa mutations identified eleven targets of current drugs, eight of investigational drugs and fifty four compounds that may be active and should be considered candidates for future clinical trials.

INTRODUCTION

Prostate cancer is the most common solid cancer in men (diagnosed in 12%) and 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 published²⁻⁶, including work from our own consortium⁷⁻¹⁰. 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, whilst 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, Gleason sum, stage) and to develop disease management strategies. For example, genetic alteration of *BRCA1/2*¹¹, *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 [ProlarisTM ¹⁵, DecipherTM ¹⁶ and OncotypeDxTM 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, whilst PARP inhibitors are effective against cancers harbouring *BRCA1/2* mutations and other defects in DNA repair pathways²¹. However, a significant recent improvement in the treatment of newly diagnosed advanced prostate cancer has been achieved through using docetaxel in combination with hormone therapy, a re-tasking of conventional therapy²².

In the present study, we use previously unpublished whole genome DNA sequencing data in combination with published data to provide new insights into the mechanism of progression of prostate cancer to lethal disease, and to design novel molecular-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 Shanghai, China together with 10 hormone-naïve prostate metastases and 10 castration-resistant metastases from the USA. Analysis (see Online Methods) reveals insights into the nature and order of acquisition of driver alterations, heterogeneity of genomic heterogeneity in primary and metastatic cancers, changes in mutational signatures during progression, and potential drug targets. In addition, we identified coding and non-coding drivers by combining single nucleotide variants (SNVs) and small insertions/deletions (indels) within our dataset with those from TCGA⁴ (425 samples), the COSMIC database²³ (243 samples) and Stand Up to Cancer²⁴ (SU2C-PCF, 150 samples) to give a combined dataset, hereafter referred to as the ‘joint dataset’, comprising 710 primary cancers and 220 metastases. Supplementary Table 1 summarises the genes affected in both our study and the joint dataset.

For the 112 cancer-normal pairs in our cohort, we identified 392,753 substitutions, 54,952 small insertion/deletions (indels) and 10,921 chromosomal rearrangements (Fig. 1). The mean genome-wide substitution rate across the whole dataset was 1.23/Mb, with a significant difference in mutational burden between the primary (0.99) and metastatic (2.30) samples ($P=4.4 \times 10^{-15}$, Online Methods). Moreover, within the metastatic subset, mutation burden was higher in men treated with androgen deprivation therapy (ADT or CRPC) with metastases than the treatment-naïve cases (2.98 vs 1.61, $P=0.015$). There were also significantly more rearrangements in metastatic than in primary samples ($P=0.0059$), whilst 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 identified through a comprehensive set of analyses to identify: excess of non-synonymous mutations in coding regions; excess missense mutations within a gene, indicative of an oncogenic driver; excess of mutations in non-coding regions; regions with an excess of structural variants in either ETS+ or ETS- cancers; and regions with recurrent copy number aberrations in either ETS+ or ETS- cancers. Overall, we identified 73 genes with evidence for involvement in prostate cancer development (Fig. 2, Table 1, Supplementary Table 2). Based on a literature search, each gene was assigned a high, medium or low level of previous supporting evidence (Table 1, 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 further genes previously lacking strong evidence of driving prostate cancer (Table 1, 'medium' previous evidence).

Coding drivers

We identified 28 genes with an excess of non-synonymous coding mutations, five of which are previously unknown drivers in prostate cancer (Supplementary Table 2). *TBLIXR1* 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 within our study had a 70kb 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*, which has previously been reported to harbor mutations in breast cancer²⁸, exhibited a mixed pattern of mutations with mostly missense mutations and two cancers harbouring truncating events.

Analysis of missense mutations identified mutations in seven further genes, of which two are newly reported (Supplementary Table 2). *CNOT3* exhibited mutation hotspots in two amino acid positions, p.E20K (4/932 samples) and p.E70K (5/932 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 leukaemia²⁹. Enrichment for missense mutations was identified in *RPL11* a ribosomal protein and a putative cancer suppressor upstream of the MDM2/TP53 pathway³⁰. In contrast to 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 the metastatic and primary samples within the joint dataset identified enrichment in metastases for mutations in *TP53*, *AR*, *KMT2C*, *KMT2D*, *RBI*, *APC*, *BRCA2*, *CDK12*, *ZFHX3*, *CTNNB1*, *PIK3CB* (Supplementary Table 2), confirming previous studies^{3,24}.

Non-coding drivers

Analysis of the non-coding components of the genome identified two regions with statistically significant enrichment for mutations. *NEAT1*, a lncRNA recently reported to be associated with PCa progression³¹, was mutated in 13/112 ICGC cases with a significant over-representation in patients with metastatic disease (6/20 metastases vs. 7/91 primaries, Fisher exact test, $P=0.012$, Fig. 3). Interestingly, out of the metastatic cases *NEAT1* mutations were found only in patients that had undergone ADT, consistent with the link between high *NEAT1* expression and resistance to AR-targeting therapies³¹. Notably, two of these six cases had two separate *NEAT1* mutations. The *FOXA1* promoter also had significant evidence of selection. This gene modulates AR-regulated transcriptional signalling³² and has been found to harbor recurrent coding mutations in previous studies⁵. In our series, we identified 14 samples with coding and 6 samples with non-coding mutations, with two samples (PD14721a and PD12813a) bearing both a coding and a non-coding mutation. Interestingly, we also identified mutations in the *FOXP1* promoter, 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 a variety of factors including chromatin state, GC content, gene density, replication timing and repetitive sequence. In order to remove the effect of these various factors, we segmented inter-breakpoint 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, two unreported regions were identified. 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. In chr3:171:178Mb, *TBLIXR1* is similarly enriched for both rearrangements and truncating mutations.

In ETS- cancers, we confirmed a previously reported enrichment for rearrangements containing *CHDI*^(38,39). A target of enriched rearrangements in the region chr1:149-158Mb is likely *ETV3*. In 5/9 cancers, *ETV3* was exclusively affected by these events (4 LOH by deletion and 1 by translocation). Additionally, one cancer had a truncating mutation (p.R413fs*3) and two had missense mutations (p.A73V and p.L37Q). In total, 12 patients had localised alteration, 10 of whom had ETS- cancers. In addition, eleven tumors, 8 of them ETS-, had LOH covering *ETV3*. Moreover, within the joint dataset, there are 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 PCa. Manual inspection of the recurrently rearranged region chr3:76-84Mb identified *ROBO1* and *ROBO2* as possible targets (Fig. 3). In total 16/112 samples had an event affecting one or other of these genes, and in four samples both were affected. Previously implicated in pancreatic ductal adenocarcinoma⁴⁰, these two genes have not been previously reported in the context of PCa.

Events enriched at chr6:80-114Mb indicate that *ZNF292* is a possible target. 11/112 patients (5 ETS+ and 6 ETS-) had loss of at least one chromosome copy and in two patients there was a homozygous loss specifically targeting *ZNF292*. Moreover, the joint dataset contained 5/932 samples with a truncating mutation, further suggesting a cancer suppressive function for this gene in PCa. Another gene affected by recurrent rearrangements on 6q was *SENP6*, a small ubiquitin-like modifier (SUMO)-specific protease that removes SUMO polypeptides from conjugated proteins⁴¹, and possibly plays a role in AR function⁴². Of note, 4/5 rearrangements in this region affected *SENP6* only, leading to a significant reduction in expression (Supplementary Fig. 1). Finally, located at chr6:126Mb, the nuclear receptor co-activator *NCOA7* was altered in six samples, with one sample having homozygous loss.

Further regions enriched in ETS- cancers were chr2:133-144Mb (*LRP1B*), chr8:112-114 (*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 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 (Supplementary Fig. 1).

Timing of copy number aberrations

In order to identify routes to progression in PCa, we developed a novel approach to order the occurrence of copy number aberrations by combining information on: the clonality of copy number aberrations; timing relative to whole genome duplication; 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 Online Methods), we deciphered the temporal ordering of the subclonal CNAs within each cancer. In general, homozygous deletions appear late in oncogenesis, corroborating previous findings that homozygous deletions are associated with more advanced disease⁴⁴⁻⁴⁶. Clear differences emerge in the evolution of PCa in the ETS+ and ETS- subsets. Where present, the deletion between the *TMPRSS2* and *ERG* genes in ETS+ cancers was an early (generally clonal) event, as was gain of chr8q within the locus 112 – 137Mb (Fig. 4a). The earliest homozygous deletions in ETS+ cancers include chr5:55Mb-59Mb, corroborating the rearrangements targeting *PPAP2A*, *PDE4D*, *MAP3K1* and *IL6ST*, and chr10:89Mb-90Mb, which covers *PTEN* (Figs. 3 and 4a).

In ETS- cancers losses at chr5:60–100Mb (*CHD1* and *RGMB*), chr13:32-91Mb (which includes *BRCA2*, *RBI* 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 cancers, preventing *ERG* re-arrangement in the prostate³⁸ and our data confirm the exclusivity between ETS positivity and homozygous loss of *CHD1* (Fig. 4c).

In both ETS-positive and ETS-negative cancers, whole genome duplication (WGD) was correlated with the 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 co-

synchronously with WGD in most cases. Gains at chr8:101Mb occurred prior to WGD, chr3:131Mb occurred synchronously, and gains at chr7:88Mb tended to follow WGD.

Timing point mutations and indels

Point substitutions and indels were clustered according to their cancer cell fraction (CCF) using a Bayesian Dirichlet process⁴⁷. The proportion of substitutions identified as subclonal showed considerable variation across cancers, but was significantly higher in primary than metastatic samples (Fig. 5a, $P=0.022$, Wilcoxon rank sum test), as was the proportion of subclonal indels ($P=0.00033$) and the fraction of the genome with subclonal copy number aberrations ($P=0.0037$, 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 substitutions and indels were correlated (Fig. 5a, Pearson $r = 0.57$, $P=2.3 \times 10^{-9}$). Higher levels of heterogeneity were observed amongst indels than substitutions ($P=2.4 \times 10^{-8}$). However, it cannot be ruled out that variant calling of indels may have greater sensitivity for low allele frequency variants than calling of point substitutions.

Driver mutations were identified as clonal or subclonal 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 to be much higher. The driver mutations identified as subclonal include two mutations in *APC* in the same sample, PD14713a. Interestingly, this cancer has undergone clonal loss of one copy of chr5q, followed by mutations in *APC* in 2 different subclones (Fig. 5b and Supplementary Fig. 3), suggesting convergent evolution. Five other

samples each have 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*; PD13389a has a frameshift mutation in *PHF12* and an essential splice site mutation in *TBX3* (not shown).

Subclonal mutations are also seen in several common drivers including one in *TP53* (PD13339a) and one in *PTEN* (PD12840a). On the other hand, *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 factorisation (NMF) revealed that, in addition to the ubiquitous ‘clock-like’ signatures 1 and 5, there was presence of the previously described signatures 2, 3, 8, 13 and 18⁴⁸. Signature-3-positive samples were enriched for germline/somatic mutations in *BRCA1/2* genes (4/6 samples) as reported previously⁴⁸ (Fig. 1). However, the presence of high levels of microhomology (MH)-mediated deletions was even more strongly correlated with the presence of *BRCA* mutations (6/6 samples). Separating the mutations into early clonal, late clonal and subclonal epochs, as described in Online Methods, revealed that the proportion of signature 1 mutations decreases over time, suggesting an increase of cancer-associated mutagenic processes relative to innate processes in normal cells ($P=2.2 \times 10^{-16}$, 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% (9/20) in metastases vs. 14% (14/92) in primaries, (Fisher exact test, $P=5.6 \times 10^{-3}$). Similarly, signature 18, which has been previously associated with failure of base excision repair and to the accumulation of

mutations from 8-Oxoguanine damage⁴⁹, was enriched in advanced disease, 40% (8/20) in metastases vs. 11% (10/92) in primaries (Fisher exact test, $P=3.8\times10^{-3}$). In a recent report of 560 breast cancer whole-genomes, signature 8 correlated with DNA damage repair deficiency⁵⁰. Androgen signalling 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, we have found that the proportion of mutations assigned to signature 8 is consistently higher amongst later appearing (subclonal) populations of cells ($55\% \pm 24\%$) than earlier (clonal) populations ($28\% \pm 12\%$) (t -test, $P=1.3\times10^{-4}$, 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 (8/20 metastases, 25/92 primaries, Fisher 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-naïve disease and CRPC in a larger dataset to explore the relevance of check-point inhibition as an alternative therapy for advanced prostate cancer.

Clinical correlates

CDH12 and *ANTXR2* alterations were significantly associated with time to biochemical recurrence (Benjamin-Hochberg adjusted $P = 0.0060$ (*CDH1*) & 0.012 (*ANTXR2*), HR = 9.3 & 7.7, Cox regression model, Fig. 6), and were significant predictors of biochemical recurrence independent of cofactors Gleason, PSA at prostatectomy, and pathological T-stage ($P = 0.00061$ (*CDH1*) & 0.0015 (*ANTXR2*), HR = 7.3 & 6.5, Cox regression model, Supplementary Table 4). A Cox regression model containing a combination of *CDH12*, *ANTXR2*, *SPOP*, *IL6ST*, *DLC1* & *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 is an independent predictor ($P = 0.0061$, HR = 3.6; Cox proportional hazards model). The number of substitutions detected was also an independent prognostic biomarker ($P=0.031$, HR=1.005; Cox proportional hazards model). The numbers of both samples and events within 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 objective identification of therapeutic intervention strategies. To this end, we applied established chemogenomic technologies using the canSAR knowledgebase⁵⁴ to map and pharmacologically annotate the cellular network of the prostate disease genes we 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 either directly interact with several of these proteins or directly affect their function (see Online Methods and Supplementary Fig. 5 for details). This resulted in a focussed prostate network of 163 proteins. We annotated each of these proteins with drug information and annotated them based on multiple assessments of ‘druggability’: the likelihood of the protein to be amenable to small molecule drug intervention. (Table 2 and Supplementary Table 5). We find that PCa driver genes are embedded in a highly druggable cellular network that contains eleven targets of approved therapies and eight targets of investigational drugs. As well as the Androgen Receptor (AR) and the Glucocorticoid Receptor (GR), the network contains targets of drugs

approved for other indications, several of which (e.g. BRAF, ESR1, RARA, RXRA, HDAC3) are under clinical investigation for PCa.

Eight proteins within the prostate network are targets of drugs currently 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 PCa due to the recently described role of DNA damage repair, particularly in advanced PCas^{21,55}. The network highlights targets of PI3 Kinase pathway inhibitors (PI3K, AKT1) that are undergoing clinical investigation in PCa, as well as IDH1 and MDM2 drug targets.

To give an indication of the potential of these drugs, we analysed the most recent systematic drug sensitivity data (GDSC-⁵⁶- <http://www.cancerrxgene.org/>). Eighteen drugs acting on our network were tested in GDSC on prostate cancer cell lines. Of these, 5 showed significant effect on growth inhibition in at least one cell line, and all 18 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 PCa therapy are also highlighted by 15 proteins that are under active chemical biology or drug discovery investigation (Table 2). These include the Inositol 1,4,5-trisphosphate receptor type 2 (IP3R2) and Menin (MEN1), a component of the MLL/SET1 histone methyltransferase complex. Mice with MEN1 mutations develop PCa⁵⁷ and recent data have shown that menin expression is involved in CRPC⁵⁸. A further 50 of the proteins are predicted to be druggable and therefore potentially amenable to drug discovery. These include the known PCa protein SPOP, the transcription activator BRG1 (SMARCA4); CDK12; and the CREB binding protein CREBBP.

In summary, we find that 74 of the 163 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

http://cansar.icr.ac.uk/cansar/publications/sequencing_prostate_cancers_identifies_new_cancer_genes_routes_progression_and_drug_targets/ (link via google chrome)

DISCUSSION

The analysis of whole genome sequence data from 112 prostate cancers has revealed many of the genetic factors underlying the processes of carcinogenesis, progression, metastasis and the acquisition of drug resistance. Supporting evidence has been provided for thirty candidate driver genes with limited or no previous support, including the non-coding drivers *NEAT1* and *FOXA1*.

Through the timing of genomic aberrations, we have a picture of the possible routes to progression in PCa. 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. Less heterogeneity of aberrations is observed in metastatic than primary cancers, likely resulting from a bottleneck in achieving metastatic potential.

We observe changes in the mutational processes operative upon cancers during progression. Signature 8 was operative to a greater extent at later stages, and was enriched in subclonal expansions, while signatures 13 and 18 were enriched in metastatic cancers. Cancers with

germline or somatic *BRCA1/BRCA2* mutations were enriched for signature 3, demonstrating the effect of double-strand repair defects throughout cancer evolution.

Survival analysis reveals that losses of *CDH12* and *ANTXR2* result in poorer recurrence-free survival. We identify 84 PCa associated proteins that are either targets for currently available drugs or new potential targets for therapeutic development.

Analysis of the whole-genome sequences of over a hundred prostate cancers has started to reveal the complex evolutionary pathways of these cancers. The early acquisition of individual 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 using the candidate driver genes here identified, in addition to previously known drivers, targets that can be exploited for immediate 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.

ONLINE METHODS

Patient Cohorts, Samples and Ethics

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 were collected as described previously^{59,60}. 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

<https://icgc.org/>. 20 men from PELICAN (Project to ELIminate lethal CANcer)⁶¹, an integrated clinical-molecular autopsy study of metastatic prostate cancer, were the subjects of the current study. Subjects consented to participate in the Johns Hopkins Medicine IRB-approved study between 1995 and 2005. (Supplementary Table 7). A17 had a germline *BRCA1* mutation, as previously reported⁶².

DNA preparation and DNA sequencing

DNA from whole blood samples and frozen tissue was extracted and quantified using a ds-DNA 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 50ng/μl in TE (10mM Tris/1mM EDTA), was between 1.8-2.0 with an OD 260/280. 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 50X for the cancer samples and 30X for matched controls⁵⁹.

The Burrows-Wheeler Aligner (BWA) was used to align the sequencing data to the GRCh37 reference human genome⁶³. Sequencing data have been deposited at the European Genome-phenome Archive (EGAS00001000262).

Variant Calling Pipeline

Substitutions, insertions and deletions were detected using the Cancer Genome Project Wellcome Trust Sanger Institute pipeline as described previously⁵⁹. In brief, substitutions were detected using CaVEMan with a cut-off 'somatic' probability of 95%. Post-processing filters were applied. Insertions and deletions were called using a modified version of Pindel⁶⁴. Variant allele frequencies of all indels were corrected by local realignment of unmapped

reads against the mutant sequence. Structural variants were detected using Brass⁵⁹. A positive ETS status was assigned if a breakpoint between *ERG*, *ETV1* or *ETV4* and previously reported partner DNA sequences was detected.

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 (<https://www.ebi.ac.uk/ega/studies/EGAS00001000262>). See Supplementary Table 7 for sample specific EGA accession codes.

Code availability

The CGP pipeline may be downloaded from <https://github.com/cancerit/dockstore-cgpgws>. The Battenberg pipeline may be downloaded from <https://github.com/Wedge-Oxford/battenberg>.

Mutation burdens

Mutation burdens were compared between primary and metastatic samples and between ADT and hormone-naïve samples using a negative binomial generalised 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 a novel approach to order the occurrence of copy number aberrations by combining three sources of information:

- Clonality of copy number aberrations
- Timing relative to whole genome duplication

- 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 PCa.

The Battenberg algorithm was used to detect clonal and subclonal somatic copy-number alterations (CNAs) and to estimate ploidy and cancer content from the next-generation sequencing data as previously described⁶⁵. Briefly, germline heterozygous SNPs were phased using Impute2, and a- and b- alleles were assigned. Data were segmented using piecewise constant fitting⁶⁶ 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 ASCAT⁶⁷.

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 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 within 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 sub-clonal cancer fractions, the second used implicit ordering of clonal HDs in relation to losses, and the third estimated the relative timing of whole genome duplication. The logical arguments used within each approach were considered in turn:

1. Battenberg algorithm-derived estimates for the cellular fraction and standard deviation 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 HDs 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.
2. Homozygous deletions have implicitly occurred after loss of heterozygosity at the same locus.
3. 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 prior to 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 prior to WGD. Otherwise, the gain occurred after WGD.

The above arguments allow us to gain insights into the order of copy number events within 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 copy number aberrations detected from each sample in random locations within the genome. The process was repeated for gains, LOH and HDs 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 from all the samples which 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 rationalised 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 TMRPSS2-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 R⁶⁸ 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 copy number aberrations 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 copy number aberrations. 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 CCFs and this sample was excluded from subclonality analysis. Point substitutions 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 within 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^{48,69}. 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, subs or indels) with a potential functional impact (missense, nonsense, start-lost, inframe, frameshift, or occurred in a non-coding transcript) or a CNA highlighted by the copy number aberration analysis that occurred in three or 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 discretised 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 Lasso⁷⁰, 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 novel oncogenes

The joint dataset was compiled from the aggregation of variants called within our samples with 3 other datasets, yielding a total of 930 samples, comprised of 710 primary and 220 metastatic samples:

- **TCGA**⁴, 425 primary cancer samples, whole exome sequencing with SureSelect Exome v3 baits on Illumina HiSeq 2000, average coverage ~100X
- **COSMIC database**²², 243 samples, curated set of mutations from several sources, <http://cancer.sanger.ac.uk/cosmic>
- **Stand Up to Cancer**²³ (SU2C-PCF), 150 metastatic castrate resistant samples, paired-end, whole exome sequencing with SureSelect Exome v4 baits on Illumina HiSeq2000, average coverage ~160X

To identify coding and non-coding drivers from substitutions 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. Non-coding 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 ($E_h = \sum_j r_j L_{j,h}$). 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_h as an offset. Both methods identify genes or non-coding regions with higher than expected mutation recurrence, correcting for gene length, sequence composition, mutation signatures acting across patients and for 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 PCa²⁵. To identify rearrangement drivers, we first used ChainFinder²⁵ to account for any bias towards regions with chromoplexy and identified ‘unique’ rearranged regions per sample taking the mid-point between all the breakpoints ChainFinder assigns to the same chromoplexy event. Next, separately aggregating the ICGC samples with and without ERG fusions, we calculated inter-breakpoint distance and performed piecewise constant fitting (PCF)⁶⁶ to identify genomic regions which were recurrently rearranged in multiple samples. Rearranged regions with potential functional impact were identified using two criteria: a minimum 3-fold difference in the number of SVs per MB of ERG+ and ERG- samples; region contains 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 82 protein products of the genes identified in this study (hereon referred to as Prostate Proteins) to seed a search for all possible interacting proteins in the canSAR interactome⁵⁴. This interactome contains merged and curated data from the IMeX consortium⁷¹, Phosphosite (<http://www.phosphosite.org/>), and other databases. It includes:

- 1) interactions where there were more than two publications reporting experiments demonstrating the binary interaction between the two proteins
- 2) interactions where there is 3D protein structural evidence of a direct complex
- 3) interactions where there are at least two publications reporting that one protein is a substrate of the other
- 4) 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 the following:

- A) interactions that were inferred from a large immunoprecipitation experiment without follow-up to demonstrate the specific binary interaction
- B) interactions inferred from text mining
- C) interactions inferred from co-occurrence in publications or from gene expression correlation.

The initial prostate cancer seeded network resulted in a large collection of 3366 proteins that have some experimental evidence of interacting with at least one Prostate Protein. When we added extra proteins into 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: Starting with the input (prostate protein) list, we obtained all possible first neighbours. We then computed, for each new protein, the proportion of its first neighbours 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 randomising 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 have corrected FDR p-values less than 0.05. (Supplementary Fig. 5). We performed network minimisation 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 163 proteins. Using canSAR's Cancer Protein Annotation Tool (CPAT)⁷², we annotated the 163 proteins with pharmacological and druggability data. We labelled proteins that are: 1) targets of approved drugs; 2) targets of drugs under clinical investigation, 3) targets of preclinical or discovery stage compounds that are active at concentrations equal to or less than 100 nM against the protein of interest 4) proteins that we predict to be druggable using our structural druggability prediction protocols⁷²⁻⁷⁵ but that have few or no published active inhibitors – these are potential targets for future drug discovery.

ACKNOWLEDGEMENTS

We acknowledge support from Cancer Research UK C5047/A22530, C309/A11566, C368/A6743, A368/A7990, C14303/A17197 and the Dallaglio Foundation, The NIHR support to The Biomedical Research Centre at The Institute of Cancer Research and The

Royal Marsden NHS Foundation Trust; 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); the Li Ka Shing foundation (DCW, DJW). The Academy of Finland and Cancer Society of Finland (GSB). We thank the National Institute for Health Research, Hutchison Whampoa Limited, University of Cambridge, 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, Dave Holland from the Infrastructure Management Team & Peter Clapham from the Informatics Systems Group at the Wellcome Trust Sanger Institute. DMB is supported by Orchid. We also acknowledge support from The Bob Champion Cancer Research Trust, The Masonic Charitable Foundation and The King Family. PW 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 and NIHR funding to the Biomedical Research Centre at Royal Marsden NHS Foundation Trust and Institute of Cancer Research. The authors would like to 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 would like to acknowledge support of the research staff in S4 who so carefully curated the samples and the follow-up data (Jo Burge, Marie Corcoran, Anne George, and Sara Stearn).

FIGURE LEGENDS

Figure 1. Mutational landscape of prostate cancers. From top-to-bottom: mutation status of DNA repair genes, ETS fusion status and sample type; proportion of mutations assigned to each signature⁴⁸; number of somatic point substitutions identified in each sample; proportion of small insertions/deletions associated with microhomology or repetitive regions; number of insertions, deletions and complex insertions/deletions in each sample; total number of structural variants in each sample, separated into inversions, translocations, deletions and tandem duplications. Sample ordering is reported in Supplementary Table 7.

Figure 2. Landscape of driver genes in prostate cancer. Genes were identified using three different methods: upper panel shows genes that have undergone genetic aberration in at least 6 samples; middle panel shows genes with aberrations enriched in either ERG+ or ERG- cancers (Fisher exact test for *PTEN*, *TP53*, *SPOP*, 3p13, *PDE4D*, *PPAP2A*; *ROBO1* and *ROBO2* are in a region enriched for SVs in ETS- tumors; *IL6ST* is in a region enriched for SVs in ETS+ tumors); lower panel shows genes enriched in metastatic samples (Fisher exact test). DDR = DNA damage response, ‘hemi.loss’ = loss of heterozygosity resulting from copy number change, ‘homo.loss’ = homozygous deletion resulting from copy number aberration, ‘two allele loss + sub/indel’ indicates genes in triploid regions bearing aberrations of all 3 gene copies. Sample ordering is reported in Supplementary Table 7.

Figure 3. Putative novel driver genes. Putative drivers are shown in red and genomic aberrations are displayed as: missense SNVs – circles; nonsense SNVs – open triangles; essential splice site mutations – open squares; indels – closed squares; non-coding mutations – closed triangles; simple SV - yellow cross; chromoplexy event – blue cross; region enriched for loss of heterozygosity, 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.

Figure 4. Temporal evolution of copy number aberrations in ETS+ and ETS- prostate cancer. For (a) ETS+ cancers, and (b) ETS- cancers: Left: The landscape of copy number aberrations with genomic loci plotted against fraction of cancers. Loss-of-heterozygosity is depicted in blue, homozygous deletions in black, gains in red, *TMPRSS2-ERG* deletion in brown and whole genome duplication in green. Right: The temporal evolution of significantly recurrent ($p < 0.05$) copy number aberrations by genomic loci over time (mean with 95% confidence intervals, log precedence relative to arbitrary reference). Lower values indicate earlier events (c) Pairwise associations among copy number aberrations. Recurrently aberrant regions with a false discovery rate < 0.1 are shown. Associations are indicated by odds ratio (OR) with brown colors depicting mutually exclusive events and blue-green 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.

Figure 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. x-axis = fraction of point substitutions that are subclonal, y-axis = fraction of indels that are subclonal, contour lines calculated using R package kde2d. (b) Samples with multiple subclonal mutations in driver genes. Fraction of cancer cells carrying mutation is shown as grey 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 subclonal fraction close to 1 are clonal, whereas peaks at lower subclonal fractions indicate subclonal mutations.

Figure 6. Clinical outcome. Kaplan-Meier plots for biochemical recurrence. Kaplan-Meier plots of recurrent mutated genes where there is a significant correlation with time to biochemical recurrence after prostatectomy ($P < 0.05$; Cox regression model; Benjamin-Hochberg multiple testing correction). Clinical information was available for 89 prostatectomy samples with WGS data, with a median follow up of 1108 days in which biochemical recurrence occurred in 26 patients. The mutations in both genes consisted of a frameshift deletion in one sample and structural variants in the remaining samples.

Table 1. Putative driver genes. 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: this would include genes such as *HRAS*, *SPOP*, *IDH1* etc. 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.

Table 2. Drug targets identified from CanSAR analysis. 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.

SUPPLEMENTARY FIGURE AND TABLE LEGENDS

Supplementary Figure 1. RNA expression of novel driver genes. Data are taken from the CamCaP study, which includes 40 samples from this study. All genes identified as novel within this study and bearing structural variants in at least 2 samples are shown. P-values are from Wilcoxon rank sum test.

Supplementary Figure 2. Heterogeneity of point substitutions and copy number aberrations. Each dot represents a different sample, colored by sample type. x-axis = amount of each genome that has subclonal CNA divided by the total amount of each genome that is copy-number aberrant, y-axis = fraction of point substitutions that are subclonal, contour lines calculated using R package kde2d.

Supplementary Figure 3. Multiple driver mutations in APC. (a) Mutations in *APC* gene in PD14713a are mutually exclusive, indicating that they have occurred in separate subclonal populations. Each blue or yellow string represents a forward or backward read, with somatic variants shown in red. Mutations in the left-hand and right-hand red boxed regions never occur on the same reads. (b) PD14713a is haploid on chromosome 5q, so the mutual exclusivity of APC mutations cannot be explained by occurrence in different chromosome

copies. Purple lines show total copy number, blue lines show minor allele specific copy number, called by the Battenberg algorithm.

Supplementary Figure 4. Prognostic biomarkers. (a) Kaplan-Meier plot of the number of mutational processes detected vs time to biochemical recurrence (b) Correlation between the number of processes detected and the number of substitutions detected. There was a significant positive correlation ($r = 0.31$, $P = 0.0027$; Spearman's correlation). 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 three. The number of substitutions detected was also an independent prognostic biomarker ($P = 0.031$, HR = 1.0005; Cox proportional hazards model).

Supplementary Figure 5. Processes used in the canSAR analysis.

Supplementary Figure 6. dN/dS analysis. QQ plot of the P -values derived from dN/dS analysis show no sign of inflation. Four genes (of 20,184) had p -values = 0 and are not shown on this plot.

Supplementary Table 1. Summary of sample specific genetic aberrations. Worksheet 1 reports the number of genetic aberrations identified in each gene this study, separated by type: essential splice site, frameshift, inframe, missense, nonsense, silent, stop-lost, homozygous deletion, rearrangement. We also report the number of samples with multiple hits resulting in homozygous loss. Worksheet 2 reports the number of genetic aberrations identified in each gene across the joint dataset. Worksheet 3 reports the samples within our dataset that bear homozygous losses in each gene. Worksheet 4 contains results of dNdScv analysis to identify coding drivers. Novel drivers are shown with red text. Significant q -values ($FDR < 0.1$) from analysing either missense SNVs or all SNVs and indels are shown

with green shading. Effect sizes (dN/dS values) are reported separately for missense, nonsense, essential splice site and indel variants. Worksheet 5 contains results of NBR analysis to identify non-coding drivers. Significant regions ($FDR < 0.1$) are shown with blue / purple shading.

Supplementary Table 2. Classification of Driver Genes. See Table 1.

Supplementary Table 3. Mutational signature analysis. For each sample, the number of somatic mutations attributed to each signature is shown in worksheet 1. Worksheet 2 reports the number of mutations assigned to each signature, broken down into temporal categories (early, late, clonal, subclonal) as described in Online Methods.

Supplementary Table 4. Survival analyses. Cox regression model results for the presence of recurrent in genes with time to biochemical recurrence after prostatectomy as endpoint. Multivariate analysis was performed taking into account cofactors Gleason (6-9), PSA at prostatectomy, and pathological T-stage (T2, T3). Clinical information was available for 89 prostatectomy samples with WGS data, with a median follow up of 1108 days in which biochemical recurrence occurred in 26 patients. Red background shading indicates features that have a significant association with outcome in both univariate analyses after multiple testing correction and multivariate analyses. Dark grey shading indicates features that are only significant without multiple testing correction. Light grey shading indicates features that are significant in univariate but not in multivariate analyses.

Supplementary Table 5. Druggability analysis. Genes identified using the CanSAR software. Genes are color-coded: bright green = target of an approved drug; dark green = target of an investigational drug; yellow = target that is being investigated chemically; red = no chemical information in public databases, but predicted to be druggable using our structure-based method.

Supplementary Table 6. Drug sensitivity data. Of the drugs identified through CanSAR analysis, 18 are reported in the Genomics of Drug Sensitivity in Cancer database. Of these, 5 showed significant effect on growth inhibition in at least one cell line, and all 18 showed weak activity in at least one cell line.

Supplementary Table 7. Clinical and Molecular Details of Cancers Subject to DNA Sequencing. The order of samples in Fig. 1 and Fig. 2 are in the columns *fig1_order* and *fig2_order*.

Supplementary Table 8. Details of samples included and excluded from the subclonal analysis.

REFERENCES

1. Attard, G. *et al.* Prostate cancer. *Lancet* **387**, 70-82 (2016).
2. Weischenfeldt, J. *et al.* Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159-70 (2013).
3. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239-43 (2012).
4. Cancer Genome Atlas Research, The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011-25 (2015).
5. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet* **44**, 685-9 (2012).
6. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-20 (2011).
7. 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**, 1521-32 (2014).
8. Cooper, C.S., Eeles, R., Wedge, D.C. & Van Loo, P. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. **47**, 367-72 (2015).
9. Boutros, P.C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* **47**, 736-45 (2015).
10. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-7 (2015).
11. 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**, 186-93 (2015).
12. Kluth, M. *et al.* Concurrent deletion of 16q23 and PTEN is an independent prognostic feature in prostate cancer. *Int J Cancer* **137**, 2354-63 (2015).

13. 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).
14. Rodrigues, L.U. *et al.* Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. *Cancer Res* **75**, 1021-34 (2015).
15. 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**, 245-55 (2011).
16. Klein, E.A. *et al.* Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. *Urology* **90**, 148-52 (2016).
17. Bostrom, P.J. *et al.* Genomic Predictors of Outcome in Prostate Cancer. *Eur Urol* **68**, 1033-44 (2015).
18. Luca, B.-A. *et al.* DESNT: A Poor Prognosis Category of Human Prostate Cancer. *European Urology Focus*.
19. 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**, 152-60 (2015).
20. 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**, 509-21 (2015).
21. Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med* **373**, 1697-708 (2015).
22. 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**, 1163-77 (2016).
23. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-11 (2015).
24. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-28 (2015).
25. Baca, S.C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-77 (2013).
26. 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).
27. Liu, Z. *et al.* CASZ1, a candidate tumor-suppressor gene, suppresses neuroblastoma tumor growth through reprogramming gene expression. *Cell Death Differ* **18**, 1174-83 (2011).
28. Fischer, K. & Pflugfelder, G.O. Putative Breast Cancer Driver Mutations in TBX3 Cause Impaired Transcriptional Repression. *Front Oncol* **5**, 244 (2015).
29. 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**, 186-90 (2013).
30. Sasaki, M. *et al.* Regulation of the MDM2-P53 pathway and tumor growth by PICT1 via nucleolar RPL11. *Nat Med* **17**, 944-51 (2011).
31. Chakravarty, D. *et al.* The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat Commun* **5**, 5383 (2014).
32. Yang, Y.A. & Yu, J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes Dis* **2**, 144-151 (2015).
33. Takayama, K. *et al.* Integrative analysis of FOXP1 function reveals a tumor-suppressive effect in prostate cancer. *Mol Endocrinol* **28**, 2012-24 (2014).
34. 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**, 130-41 (2013).

35. Carver, B.S. *et al.* Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat Genet* **41**, 619-24 (2009).
36. King, J.C. *et al.* Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation in prostate oncogenesis. *Nat Genet* **41**, 524-6 (2009).
37. Kluth, M. *et al.* Clinical significance of different types of p53 gene alteration in surgically treated prostate cancer. *Int J Cancer* **135**, 1369-80 (2014).
38. Burkhardt, L. *et al.* CHD1 is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. *Cancer Res* **73**, 2795-805 (2013).
39. Liu, W. *et al.* Identification of novel CHD1-associated collaborative alterations of genomic structure and functional assessment of CHD1 in prostate cancer. *Oncogene* **31**, 3939-48 (2012).
40. Biankin, A.V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399-405 (2012).
41. Heun, P. SUMO organization of the nucleus. *Curr Opin Cell Biol* **19**, 350-5 (2007).
42. 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).
43. 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**, 474-88 (2014).
46. Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. *Nature* **436**, 725-30 (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**, 415-21 (2013).
49. Pilati, C. *et al.* Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* (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**, 1245-53 (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**, D938-43 (2016).
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**, 1663-72 (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**, D955-61 (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**, 344-52 (2015).

59. 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* (2015).
60. Mao, X. *et al.* Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis. *Cancer Res* **70**, 5207-12 (2010).
61. Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**, 559-65 (2009).
62. Nickerson, M.L. *et al.* Somatic alterations contributing to metastasis of a castration-resistant prostate cancer. *Hum Mutat* **34**, 1231-41 (2013).
63. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
64. 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**, 2865-71 (2009).
65. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
66. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
67. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-5 (2010).
68. Firth, D. & Turner, H.L. Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software* **48**(2012).
69. 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**, 246-59 (2013).
70. Shin, S., Fine, J. & Liu, Y. Adaptive Estimation with Partially Overlapping Models. *Stat Sin* **26**, 235-253 (2016).
71. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* **9**, 345-50 (2012).
72. 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).
73. 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**, D1040-7 (2014).
74. Mitsopoulos, C., Schierz, A.C., Workman, P. & Al-Lazikani, B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLoS Comput Biol* **11**, e1004597 (2015).
75. Workman, P. & Al-Lazikani, B. Drugging cancer genomes. *Nat Rev Drug Discov* **12**, 889-90 (2013).
76. 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).
77. Zhang, H. *et al.* FOXO1 inhibits Runx2 transcriptional activity and prostate cancer cell migration and invasion. *Cancer Res* **71**, 3257-67 (2011).
78. 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**, 155-69 (2009).
79. 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**, 368-77 (2009).
80. 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**, 142-7 (2015).

81. 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**, 2493-506 (2014).
82. 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**, 127-42 (2014).
83. Jones, S. *et al.* Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. *Hum Mutat* **33**, 100-3 (2012).
84. 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).

Author contribution

RAE, CSC, DN, DSB, CSF, SB, AGL, PW, BA, DCW, FCH and DFE designed the study.

RAE, CSC, DCW and DB wrote the paper, and all other authors contributed to revisions.

ZKJ, HW, CEM, DN, VG, AGL, RAE, FCH, SB, AYW, CSF, CV, DMB, ND, SM, SH, WH, Y-JL, AL, JK, KK, HL, LM, SE, LMatthews, AN, YY, HZ, ST, EB, CF, NL, SH, DNicol, PG, VK, NVA, PK, CO, DC, AT, EM, ER, TD, NCS, coordinated sample collection, pathology review and processing.

DCW, GG, TM, IM, DJW, DSB, MG, JZ, AB, LGB, SD, BK, NC, VB, DL, SM, TD, MA, STavare, CG, KR, DG, AM, LS, JT, AF, UM, supported, directed and performed the analyses.

CS and the TCGA, JdeB and GA provided data for the meta-analysis.

TABLE 1

gene	Mutation type(s)	Previous evidence	Prior evidence	Evidence in our study
ADAM28	SV, CNA	low	⁷⁶ biological evidence	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	⁷⁷ biological evidence	CNA in ETS-
IL6ST	SV	low	⁷⁸ biological evidence	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	non-coding	low	³¹ biological evidence	non-coding
PDE4D	SV	low	⁷⁹ SNP data	SVs and CNA in ETS+
PPAP2A	SV	low	⁷⁹ SNP data	SVs and CNA in ETS+
PPP2R2A	SV	low	⁸⁰ biological evidence	SVs and CNAs in ETS+
ROBO1	SV	low	⁸¹ biological evidence	SVs in ETS+
ROBO2	SV	low	²⁵	SVs in ETS+
RPL11	SNV/indel	low	²⁵	dN/dS (missense)
SENP6	SV	low	⁴² biological evidence	enriched SVs, RNA expression
TBL1XR1	SNV/indel,SV	low	⁸² known AR co-regulator biological evidence	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	⁸³	dN/dS
CASZ1	SNV/indel	medium	COSMIC, TCGA and SU2C	dN/dS
CNOT3	SNV/indel	medium	⁸⁴ Mut. in leukemia	dN/dS (missense)
LRP1B	SV, CNA	medium	⁷⁹ SNP data	SVs and CNA in ETS-
PIK3R1	SNV/indel	medium	²⁴	dN/dS
RGMB	CNA	medium	³⁸ deletions	CNA in ETS-
TBX3	SNV/indel	medium	known breast cancer gene	dN/dS
ZMYM3	SNV/indel	medium	COSMIC SU2C	dN/dS

Table 2

Target of approved drug
AR, BRAF, ESR1, HDAC3, KCNH2, MAP2K1, NR3C1, RARA, RARB, RARG, RXRA
Target of investigational drug
AKT1, ATM, LRRK2, MDM2, PDE4D, PIK3CA, PIK3CB, TP53
Target being investigated chemically
AHR, BRCA1, CTNNB1, HRAS, IDH1, ITPR1, ITPR2, 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, ESCO1, ETV3, FOXA1, FOXG1, FOXO1, FOXO4, FOXP1, GATA1, GATA2, HDGF, HNF4A, IL6ST, KAT6A, KDM4A, KDM6A, KMT2C, KMT2D, NEDD4L, NKX3-1, PIAS1, PIAS2, PTEN, RB1, RNF43, SKI, SMAD2, SMAD3, SMAD4, SMARCA4, SPDEF, SPOP, TBL1X, TBL1XR1, TBX3, TP73, ZBTB16, ZHX2