The long tail of oncogenic drivers in prostate cancer

Joshua Armenia1,2,19, Stephanie A. M. Wankowicz3,4,19, David Liu3,4,19, Jianjiong Gao1,2, Ritika Kundra1,2, Ed Reznik1,2, Walid K. Chatila1,2, Debyani Chakravarty1,2, G. Celine Han3,4, Ilسا Coleman5, Bruce Montgomery6, Colin Pritchard7, Colm Morrissey8, Christopher E. Barbieri9, Himisha Beltran10,11,12, Andrea Sboner9, Zafeiris Zafeiriou13, Susana Miranda10,13, Craig M. Bielski1,2, Alexander V. Penson1,2, Charlotte Tolonen4, Franklin W. Huang3,4, Dan Robinson14, Yi Mi Wu14, Robert Lonigro14, Levi A. Garraway3,4, Francesca Demichelis9, Philip W. Kantoff16, Mary-llen Taplin3, Wassim Abida16, Barry S. Taylor1,2,17, Howard I. Scher16, Peter S. Nelson5,6, Johann S. de Bono16, Mark A. Rubin9,11,12, Charles L. Sawyers1, Arul M. Chinnaiyan14, PCF/SU2C International Prostate Cancer Dream Team18, Nikolaus Schultz1,2,17,19* and Eliezer M. Van Allen3,4,19*

Comprehensive genomic characterization of prostate cancer has identified recurrent alterations in genes involved in androgen signaling, DNA repair, and PI3K signaling, among others. However, larger and uniform genomic analysis may identify additional recurrently mutated genes at lower frequencies. Here we aggregate and uniformly analyze exome sequencing data from 1,013 prostate cancers. We identify and validate a new class of E26 transformation-specific (ETS)-fusion-negative tumors defined by mutations in epigenetic regulators, as well as alterations in pathways not previously implicated in prostate cancer, such as the spliceosome pathway. We find that the incidence of significantly mutated genes (SMGs) follows a long-tail distribution, with many genes mutated in less than 3% of cases. We identify a total of 97 SMGs, including 70 not previously implicated in prostate cancer, such as the ubiquitin ligase CUL3 and the transcription factor SPEN. Finally, comparing primary and metastatic prostate cancer identifies a set of genomic markers that may inform risk stratification.

The genomic landscape of primary and metastatic prostate cancer has been robustly assessed through whole-exome sequencing (WES) of tumors and matched germline samples. These studies have identified multiple recurrently altered genes and pathways, including androgen signaling, DNA repair, and phosphoinositide 3-kinase (PI3K)–AKT signaling1,2. Additionally, they identified genomically distinct classes of prostate cancer, defined by ETS transcription family fusions3 or mutations in SPOP4, FOXA15, or IDH16. Nevertheless, prostate cancer harbors substantial interpatient genomic heterogeneity, and power analyses have suggested that larger WES studies may identify additional statistically significant mutated genes occurring at lower frequencies, indicating that the spectrum of genes involved in prostate cancer is incompletely defined1. As the aggregation and uniform meta-analysis of WES data has been transformative to research and clinical interpretation of germline genetics7, we hypothesized that mutual significance analysis using statistical and biological frameworks in a large and uniformly analyzed WES cohort might similarly identify new genes and pathways to refine the genomic landscape of prostate cancer.

We assembled and uniformly analyzed WES data from 1,013 tumors and matched germline prostate cancers (680 primary and 333 metastatic tumors) (Supplementary Table 1)1,2,4,7–9 that passed joint quality control parameters (Fig. 1a, Methods, Supplementary Fig. 1, Supplementary Table 2, and Supplementary Note). Patient characteristics, including age at diagnosis, Gleason score, and metastatic site, are shown in Table 1 and Supplementary Table 3. The mean nonsynonymous mutational load for primary and metastatic prostate cancers was 1.36 mutations/Mb and 2.93 mutations/Mb, respectively (Supplementary Fig. 2). As previously reported1,2,19, the mutational load was significantly higher in metastatic tumors ($P<0.001$, estimated 1.43 mutations/Mb higher mutational load adjusted for differences in tumor sequencing depth and tumor
Mutational significance analysis of point mutations and short indels using MutSig2CV\(^1\) and additional biological significance filters (Methods) identified 97 SMGs (Fig. 1a,b, Supplementary Figs. 3 and 4a–c, and Supplementary Table 4). As predicted by prior power analyses\(^2\), the majority of these new SMGs occurred in less than 5% of the overall cohort and could only be discovered in cohorts with over 900 samples (Supplementary Table 5). SMGs
Supplementary Table 6). New prostate cancer–specific SMGs in this category include well-known prostate cancer–associated genes, such as AR, SPOP, FOXA1, TP53, and PTEN (Fig. 1b and Supplementary Table 4). We identified 70 SMGs previously implicated in cancer but not previously reported as significantly altered in prostate cancer (1.3%). In this class of tumors, 5% had mutations in genes that encode SWI/SNF nucleosome-remodeling complex members (Fig. 1d), including ARID1A (1.6%), ARID4A (1%), ARID2 (1.3%), and SMARCA1 (1.1%), similar to observations made in other tumor types. In primary tumors, mutations in genes encoding epigenetic regulators and chromatin remodelers were significantly more common in tumors with-OUT previously known drivers (Supplementary Fig. 5), and an additional 9 SMGs not previously identified as recurrently altered in any cancer type (Fig. 1b and Supplementary Table 4).

We then integrated focal copy number events and available ETS fusion data to stratify these findings by pathway and function and developed a categorized set of SMGs in prostate cancer. Through this approach, we identified 20% of prostate cancer samples with mutations, frequently truncating, in epigenetic modifiers or chromatin-remodeling genes (Fig. 1c,d and Supplementary Table 6). Within this class of tumors, 5% had mutations in genes that encode SWI/SNF nucleosome-remodeling complex members (Fig. 1d), including ARID1A (1.6%), ARID4A (1%), ARID2 (1.3%), and SMARCA1 (1.1%), similar to observations made in other tumor types.

In primary tumors, mutations in genes encoding epigenetic regulators and chromatin modifiers were significantly associated with higher Gleason score (10% Gleason 3+4, 22% Gleason 8–10, P = 0.001, Fisher’s exact test). Furthermore, upon examination of the subset of our cohort for which ETS fusion status was available (n = 765), we found that alterations in epigenetic regulators and chromatin remodelers were significantly more common in tumors that lacked an ETS fusion (P = 1 x 10^{-4}, Fisher’s exact test) and in tumors without previously known drivers (ETS fusion, IDH1, SPOP, CUL3, or FOXA1 mutation) (P = 0.007, Fisher’s exact test) (Fig. 1c,d and Supplementary Table 6).

Our analysis also identified recurrently mutated genes in the ubiquitin–proteasome (USP) and ligase gene family, of which SPOP is a member, with mutations found in USP28 (1.4%), USP7 (1.2%), and CUL3 (1.3%) (Fig. 2a). CUL3 encodes part of a culin–RING-based (BTB–CUL3–RBX1) E3 ubiquitin ligase complex with SPOP, and mutations may affect degradation of prostate cancer tumorigenesis regulators, including AR, SRC-3, and TRIM24. CUL3 mutations were primarily in a hotspot (p.Met299Arg) and were mutually exclusive with SPOP mutations (Fig. 2a), although this cohort size was not sufficiently powered to establish statistical significance. CUL3-mutant tumors also exhibited copy number profiles similar to those of SPOP-mutant tumors, with losses at chromosomes 5q, 6q, and 13 (Fig. 2b and Supplementary Fig. 5). To confirm this finding in an orthogonal cohort, we identified nine additional somatic CUL3 mutations in an independent cohort of advanced prostate cancers (1.3% in the MSK-IMPACT data), including three p.Met299Arg alterations (Supplementary Fig. 6a).

In addition, the splicing pathway was altered in 4% of prostate tumors (Fig. 2c), most notably through hotspot mutations in SF3B1 (1.1%) and U2AF1 (0.5%). Mutations in SF3B1 mostly clustered around the highly conserved HEAT repeats in the C terminus (Fig. 2c), similar to what is observed in other cancer types. This alteration is thought to disrupt the recognition and binding of 3’ splice sites.

We also identified SMGs in previously known prostate cancer–associated pathways, including AR, WNT/β-catenin, PI3K, and RAS–MAPK signaling. Within the AR/hormone signaling pathway, our analysis identified SPEN, which encodes a hormone-inducible transcription repressor, mutated in 2.4% of this cohort, mostly through truncating mutations (Fig. 3a,b). The SPEN protein is known to repress the estrogen receptor via NCOR2 by recruiting histone deacetylases and SRA, an RNA co-activator. SPEN expression is activated via estrogen and potentially other hormones, and its overexpression is associated with response to tamoxifen in breast cancer. SPEN mutations were significantly enriched in metastatic samples (q = 0.008, Fisher’s exact test) and clonal samples (Fig. 3a), suggestive of SPEN being a driver in advanced disease.

The PI3K pathway was altered in 25% of our samples, primarily as a result of homozygous loss and truncating mutations in PTEN (16%). Our analysis identified a new prostate cancer–associated gene in the PI3K pathway, PIK3R2 (1%), which, like PIK3R1, encodes a PI3K regulatory subunit. One of the PIK3R2 mutations (encoding p.Asp557Tyr) is paralogous to the known oncopgenic p.Asp560Tyr mutation in PIK3R1 (Supplementary Fig. 6b) and was also found in our validation cohort.

Genomic alterations in the WNT/β-catenin pathway were found in 10% of the cohort (Fig. 3c and Supplementary Table 6). For CTNNB1, while the majority of mutations clustered in the N-terminal domain (Fig. 3d), three residues, including a new p.Lys335His hotspot, clustered around the region of β-catenin that interacts with AXIN (Fig. 3e). The RAS–RAF–MAPK pathway was altered in 5% of samples (Supplementary Table 6), including SMGs in KRAS and BRAF, mostly due to established hotspot mutations not previously enriched for significance in prostate cancer.

As previously reported, we observed a significant number of inactivating alterations in DNA repair genes (16% of samples; Supplementary Table 6). New prostate cancer–specific SMGs in this pathway included MRE11A and PALB2. CDK12 was mutated primarily by truncating mutations (P < 0.001, binomial test), as previously observed in ovarian cancer. Of note, CDK12 missense variants significantly clustered in the kinase domain (P < 0.001, binomial test) (Supplementary Fig. 6c), suggesting putative functional relevance. Furthermore, 15 of 31 CDK12-mutant tumors (as well as 27 of 56 samples in the validation cohort) harbored two mutations in the gene, suggestive of frequent biallelic inactivation. Broadly, these results expand on SMGs in known cancer-associated pathways not previously implicated in prostate cancer and further delineate the genomic heterogeneity of mutations in the long tail of this disease.

Finally, we conducted a systematic comparison of primary and metastatic tumors to identify which events are associated with advanced disease (Fig. 4a and Methods). Genes with enrichment in metastatic samples included TP53, AR, PTEN, RB1, FOXA1, APC, and BRCA2 (Fig. 4a). Alterations in epigenetic regulators, including KMT2C and KMT2D, were also significantly enriched in metastatic tumors and in aggregate define a genomic signature of high-risk disease. Conversely, mutations in SPOP were significantly enriched in primary tumors (Fig. 4a). After correction for differences in mutational load, IDH1 and ZMYM3 mutations were also enriched in primary tumors (P = 0.01, mutation rate–adjusted permutation test). At the pathway level, PI3K, DNA repair, cell cycle, WNT/β-catenin,
and epigenetic regulators were significantly more frequently altered in metastatic than in primary tumors \( (P < 0.0001, \text{Fisher's exact test; Fig. 4b and Supplementary Table 7}) \).

Within a given cancer type, the ability to redefine mutational significance with rapidly expanding sample sizes may identify new biologically and clinically relevant genes and pathways not previously appreciated. This study has leveraged this strategy to identify new driver genes and pathways potentially implicated in the pathogenesis of prostate cancer. While many of the significantly altered genes and pathways are mutated at low frequencies, given the incidence of prostate cancer, these alterations still impact large patient populations. In addition, whereas expanded analysis of primary indolent prostate cancer suggests near saturation for gene discovery\(^\text{28}\), this analysis, which includes more advanced cases, has identified new biologically and clinically relevant events and creates an opportunity to prospectively assess a metastasis-associated genomic marker for clinical stratification in localized prostate cancer.

Combined statistical and biological significance analysis enabled a focused assessment of the SMGs identified herein, and efforts to functionally characterize this long tail of SMGs in prostate cancer may inform their relative phenotypic effects on oncogenicity, metastatic potential, and response characteristics to known or emerging prostate cancer therapeutics. Indeed, many of the genes identified through statistical analysis alone are of unknown function and
suggest that even larger sample sizes paired with functional analysis will be necessary to discriminate which are relevant to prostate cancer oncogenesis. Subsequent studies that harmonize even larger prostate cancer molecular cohorts through uniform genomic analysis may also orthogonally validate these findings and further mitigate technical differences, such as the stochastic effects of sequencing on variant detection, when analyzed in aggregate. Overall, our analysis demonstrates the utility of uniform genomic analysis in a single cancer type at a larger scale than previously reported, thereby redefining the molecular landscape of prostate cancer and providing a rationale to revisit mutational significance in other cancer types as data generation scales by orders of magnitude.
Fig. 4 | Enrichment of genomic alterations in metastatic tumors. a, Most genomic alterations are enriched in metastatic disease. Alteration percentages in metastatic samples (n = 333) are shown on the x axis, and those in primary samples (n = 680) are shown on the y axis. The significance of enrichment (two-sided Fisher’s test q value or weighted permutation test) is shown by the size of the dots. Genes in bold have a significant enrichment of mutations using Fisher’s test and weighted permutation test correcting for mutation burden. b, Pathway alteration frequencies in metastatic disease as compared to primary disease. A sample was considered altered in a given pathway if at least a single gene in the pathway had a genomic alteration. P values indicate the level of significance (two-sided Fisher’s exact test).

Methods
Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0083-2.

Received: 19 January 2017; Accepted: 26 January 2018; Published online: 2 April 2018

References
1. Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. Cell 161, 1215–1228 (2015).
2. Cancer Genome Atlas Research Network. Molecular taxonomy of primary prostate cancer. Cell 163, 1011–1025 (2015).
3. Tomlins, S. A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310, 644–648 (2005).
4. Barbier, C. E. et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. 44, 685–689 (2012).
5. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505, 495–501 (2014).
6. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
7. Beltran, H. et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. Nat. Med. 22, 298–305 (2016).
8. Kumar, A. et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nat. Med. 22, 369–378 (2016).
9. Raca, S. C. et al. Punctuated evolution of prostate cancer genomes. Cell 153, 666–677 (2013).
10. Hieronymus, H. et al. Copy number alteration burden predicts prostate cancer relapse. Proc. Natl. Acad. Sci. USA 111, 11139–11144 (2014).
11. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).
12. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499, 43–49 (2013).
13. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507, 315–322 (2014).
14. Theurillat, J.-P. P. et al. Ubiquitome analysis identifies dysregulation of effector substrates in SPOP-mutant prostate cancer. Science 346, 85–89 (2014).
15. Geng, C. et al. Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover. Proc. Natl. Acad. Sci. USA 110, 6997–7002 (2013).
16. Yuan, W.-C. et al. A Cullin3–KLHL20 ubiquitin ligase–dependent pathway targets PML to potentiate HIF-1 signaling and prostate cancer progression. Cancer Cell 20, 214–228 (2011).
17. Groner, A. C. et al. TRIM24 is an oncogenic transcriptional activator in prostate cancer. Cancer Cell 29, 846–858 (2016).
18. Boysen, G. et al. SPOP mutation leads to genomic instability in prostate cancer. eLife 4, e09207 (2015).
19. Abida, W. et al. Prospective genomic profiling of prostate cancer across disease states reveals germline and somatic alterations that may affect clinical decision making. JCO Precis. Oncol. https://doi.org/10.1200/PO.17.00029 (2017).
20. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. 23, 703–713 (2017).
21. Dolatshad, H. et al. Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia 29, 1092–1103 (2015).
22. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163, 506–519 (2015).
23. Papaemmanuil, E. et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N. Engl. J. Med. 365, 1384–1395 (2011).
24. McHugh, C. A. et al. The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature 521, 232–236 (2015).
25. Shi, Y. et al. Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. Genes Dev. 15, 1140–1151 (2001).
26. Légaré, S. et al. The estrogen receptor cofactor SPEN functions as a tumor suppressor and candidate biomarker of drug responsiveness in hormone-dependent breast cancers. Cancer Res. 75, 4351–4363 (2015).
27. Kuchay, S. et al. FBXL2- and PTPL1-mediated degradation of p110-free p85β regulatory subunit controls the PI(3)K signalling cascade. Nat. Cell. Biol. 15, 472–480 (2013).
28. Fraser, M. et al. Genomic hallmarks of localized, non-indolent prostate cancer. Nature 541, 359–364 (2017).

Acknowledgements
We thank the patients for participating in this study. We also thank the Broad Cancer Genome Analysis and Data Sciences groups for analysis methodology and computational support. This work was supported by the SU2C-PCF Prostate Cancer International Dream Team, Prostate Cancer Foundation Young Investigator Awards (B.S.T., C.P., N.S., and E.M.V.A.), Prostate Cancer Foundation-V Foundation Challenge Award (E.M.V.A., P.S.N., and J.S.d.B.), NIH K08CA188615 (E.M.V.A.), NCI P50-CA097186 and NCI P50-CA92629 SPOREs in Prostate Cancer, the Marie-Josée and Henry R. Kravis Center for Molecular Oncology, a National Cancer Institute Cancer Center Core Grant (P30-CA008748), and the Robertson Foundation (B.S.T. and N.S.).

Author contributions
J.A., S.A.M.W., N.S., E.M.V.A., D.L., J.G., R.K., E.R., W.K.C., D.C., G.C.H., C.E.B., A.S., C.M.B., A.V.P., C.T., F.D., M.A.R., and B.S.T. contributed with algorithm development and analysis of genomic data. R.L., L.A.G., I.C., B.M., C.P., C.M., H.B., Z.Z., S.M., F.W.H., D.R., Y.M.W., P.W.K., M.-E.T., W.A., H.I.S., P.S.N., J.S.d.B., M.A.R., C.L.S., and A.M.C. developed the patient cohort, obtained tumor biopsies, performed molecular testing for metastatic cases, and carried out data interpretation of the overall cohort. J.A., S.A.M.W., D.L., N.S., and E.M.V.A. performed final aggregate cohort assembly, mutation review, interpretation, and manuscript preparation.

Competing interests
E.M.V.A. is a consultant for Tango Therapeutics and Genome Medical.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0078-z.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to N.S. or E.M.V.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Cohort collection and quality control. Samples were included in this study if tumor and matched germline whole-exome sequencing raw data (BAM or Fastq files) were accessible and met downstream quality-control characteristics (see "Quality control"). These cohorts were identified through review of the literature and expert review (Supplementary Table 1). All cohorts had institutional review board approval for access from the original studies, listed in the citation. We obtained the whole-exome sequencing BAM files from all samples. All samples underwent uniform alignment through the same version of the Picard pipeline. Details of versioning parameters for all alignment with the Picard task were provided in the Supplementary Note. All tumor samples were required to have at least 50× mean target coverage, and all paired normal samples were required to have at least 30× mean target coverage. Mean target coverage across the cohorts for tumors was 104.7× and for normal samples was 103.8×. ComEst was used to estimate the level of contamination with foreign DNA55. All samples had ComEst scores lower than 5%, and the mean ComEst value was 0.6%.

Clinical data. All clinicopathological annotations were obtained from the original papers24,25,26. All primary tumors were treatment naive; all metastatic tumors were castration resistant.

Variant calling. We restricted the analysis to considering sites in the common pool of bases that covered the bait sets used in the respective source projects, creating an intersected BED file using the bedtools intersect tool (Supplementary Fig. 7 and Supplementary Table 8) (http://bedtools.readthedocs.io/en/latest/content/tools/intersect.html). Single-nucleotide variants (SNVs) were called with MuTect (version 1.1.6)56, using the intersected BED file. Unfiltered MuTect mutation calls are located in Supplementary Table 9.

Artifacts introduced by DNA oxidation during sequencing or the formalin fixation process were removed when appropriate31. Specifically, regarding artifacts from formalin fixation, formalin fixation introduces multiple types of DNA damage including deamination, which converts cytosine to uracil and leads to downstream mispairing in PCR. C>T/G>A. Because the deamination occurs before ligation of palindromic Illumina adapters, likely deamination artifacts will have a read orientation bias. We used this read orientation to identify artifacts and calculate a Phred-scaled Q score for FPPE artifacts42.

To further reduce the number of low-confidence mutations with potential strand bias, we performed a Fisher’s exact test on each called mutation site in aggregate to identify variants occurring significantly more frequently in one read direction than in the other. A false discovery rate threshold, measured by Benjamini–Hochberg, of <0.0001 was used. In addition, all SNVs were required to have an allelic fraction of ≥0.01 to be called.

Indels were called with Strelka (version 1.0.11)28. SNVs and indels were also filtered through a large panel of normal samples to extract additional poor calls. Any mutations in hotspot genes, defined by cancerhotspots.org34, initially called by MuTect but subsequently filtered out, were rescued for the final variant list. When possible, we used ERG fusion calls defined according to the original source data2. For the 126 additional TCGA samples that were not part of the TCGA manuscript, we derived ERG fusion2. For 305 prostate cancer samples that were analyzed by TCGA, we compared the segments with any known copy number profiles generated by RecapSeg to those from TCGA data (Supplementary Fig. 8). We generated a scatterplot to compare the segment means of matched segments >200 kb from the SNP6 and RecapSeg data, resulting in a Pearson correlation of 0.92. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0. In addition, we called the allelic copy number of well-known prostate cancer genes, accounting for purity and ploidy, obtained from FACETS (version 0.5.10)60 (genes examined: TP53, APC, PTEN, RB1, BRCA2, CDKN1B, FANCA, ATM, AR). We performed manual review of copy number calls for selected oncogenes and tumor suppressors. All data are available for visualization and analysis in the CBioPortal for Cancer Genomics at http://www.cbioportal.org/61.

Mutation and copy number burden. Mutational burden was calculated as the number of mutations over the number of bases covered per sample and is reported as mutations per megabase. Copy number burden was calculated as the fraction of the genome altered using copy number segments with ≥0.2 [as previously defined]. A multivariate linear regression adjusting for purity and coverage was used to evaluate the difference in mutational and copy number burden in metastatic and primary tumors. Additional information is provided in the Supplementary Note.

Mutational significance analysis. All mutations that passed quality control were analyzed using MutsigCV to identify significantly mutated genes (SMGs). MutsigCV integrates three separate significance assessment methods: MutsigCV; MutsigFN, which looks at the functionality of a mutation in a gene; and MutsigCL, which looks at the clustering of mutations within the gene, specifically looking for hotspot mutations. Both MutsigFN and MutsigCL measure significance based on permutations. SMGs fell within two different categories: (i) q values less than 0.1 and altered in at least 10 samples and (ii) q values between 0.1 and 0.25, altered in at least 10 samples, and in known cancer-related genes. Additionaly, genes with low median allelic fraction (<0.1) were removed from the SMG list. Genes encoding proteins >1,500 amino acids in length (except for cancer-related genes) and those with a fraction of truncating variants larger than 50% of the total mutations, indicating a putative tumor suppressor were also removed from the SMG list. Genes with low prevalence in prostate cancer (median expression below the bottom tertile in TCGA RNA-seq) were also removed from the SMG list. Finally, genes with at least five oncogenic variants (according to OncoKB; http://oncokb.org/) but that were not previously included in the SMG list were added to the SMG list.

Comparison of genomic alterations between primary and metastatic tumors. Enrichment analysis of mutations and copy number alterations observed in metastatic tumors as compared to primary tumors was performed by tabulating the frequency of mutations or copy number events observed in either metastatic or primary prostate cancer and performing a two-sided Fisher’s exact test on a set of biologically relevant cancer-related genes (n = 650 genes)2. Multiple-hypothesis testing correction was performed using the Benjamini–Hochberg method. To adjust for differences due to increased mutation load in metastatic tumors, we also performed a modified Fisher’s exact test, a permutation test where the probability of mutation in each sample is weighted by the mutation rate in that sample, and a simulation of 10,000 permutations performed with a two-sided P value calculated as the proportion of those permutations with the observed or more extreme outcome. This directly corrects for differential observed mutation rates between primary and metastatic tumors and represents the null hypothesis that mutations are equally likely to be found in primary versus metastatic tumors, adjusting for differences in mutation rate. We were able to perform this mutational-rate-based adjustment in genes where the only events were mutations. In cases where functional events included both mutations and copy number changes (for example, PTEN), we performed only a Fisher’s exact test.

Clonality analysis. Clonality of mutations was estimated as cancer cell fraction (CCF)27,28 and implemented in the FACETS algorithm27. Additional information is provided in the Supplemental Note.

Statistical analysis. Two-tailed Fisher’s exact tests were used to assess enrichment of alterations in epigenetic regulators and chromatin remodelers in ET5-fusion-negative tumors. Association of mutation burden and fraction of the genome altered with metastasis status, age at diagnosis, and Gleason score was evaluated using Mann–Whitney Wilcoxon tests and permutation tests. All statistical analyses were performed using R version 3.3.1 (https://www.r-project.org/).

Validation datasets. To validate mutations detected in this study cohort, we queried cancer panel data from two sources: (i) Foundation Medicine, 204 patients with prostate cancer29 (mutation calling for this cohort was obtained as previously described and data are available in phs001179) and (ii) clinical sequencing data from 706 samples from Memorial Sloan Kettering patients (MSK-IMPACT)23 (mutation calling for this cohort was obtained as previously described and data are available from the paper at cbioPortal).

Code availability. Bioinformatics tools used in the analysis of this dataset are publicly available. Any that are not are available upon request.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. BAM files are accessible as described for the original cohorts (Supplementary Table 1). In addition, all mutation calls and clinical annotation were deposited into cbioPortal for analysis and visualization: http://www.cbioportal.org/study?id=prad_p1000.

References
28. Cibulskis, K. et al. ComEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27, 2601–2602 (2011).
29. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).
30. Costello, M. et al. Discovery and characterization of artificial mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage. Nat. Genet. 45, 177–183 (2013).
31. Van Allen, E. M. et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat. Med. 20, 682–688 (2014).
32. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumors–normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
33. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. 34, 155–163 (2016).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.
35. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
36. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
37. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
38. Cheng, D. T. et al. Memorial Sloan Kettering—Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
39. Chakravarty, D. OncoKB: a precision oncology knowledge base. *Precis. Oncol.* https://doi.org/10.1200/PO.17.0011 (2017).
40. McGranahan, N. et al. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54 (2015).
41. Hartmaier, R. J. et al. High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis. *Cancer Res.* **77**, 2464–2475 (2017).
## Experimental design

1. **Sample size**
   - Describe how sample size was determined.
   - Methods (page 12, paragraph 1)

2. **Data exclusions**
   - Describe any data exclusions.
   - Methods (page 12, paragraph 1-3)

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   - N/A

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   - N/A

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - N/A

   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

7. **Software**
   - Describe the software used to analyze the data in this study.
   - Supplementary Information file and Methods (pages 12-15)
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.

| N/A |

9. Antibodies

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

| N/A |

10. Eukaryotic cell lines

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

| N/A |

b. Describe the method of cell line authentication used.

| N/A |

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

| N/A |

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.

| N/A |

### 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.

| N/A |

**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.

| Supplementary table 3. |