nature

https://doi.org/10.1038/s41586-020-2135-x

Supplementary information

A genomic and epigenomic atlas of prostate cancer in Asian populations

In the format provided by the authors and unedited

Jing Li, Chuanliang Xu, Hyung Joo Lee, Shancheng Ren, Xiaoyuan Zi, Zhiming Zhang, Haifeng Wang, Yongwei Yu, Chenghua Yang, Xiaofeng Gao, Jianguo Hou, Linhui Wang, Bo Yang, Qing Yang, Huamao Ye, Tie Zhou, Xin Lu, Yan Wang, Min Qu, Qingsong Yang, Wenhui Zhang, Nakul M. Shah, Erica C. Pehrsson, Shuo Wang, Zengjun Wang, Jun Jiang, Yan Zhu, Rui Chen, Huan Chen, Feng Zhu, Bijun Lian, Xiaoyun Li, Yun Zhang, Chao Wang, Yue Wang, Guangan Xiao, Junfeng Jiang, Yue Yang, Chaozhao Liang, Jianquan Hou, Conghui Han, Ming Chen, Ning Jiang, Dahong Zhang, Song Wu, Jinjian Yang, Tao Wang, Yongliang Chen, Jiantong Cai, Wenzeng Yang, Jun Xu, Shaogang Wang, Xu Gao✉, Ting Wang✉ & Yinghao Sun✉

Supplementary Discussion

Clinical samples and histopathologic data

We analysed tumour and matched normal samples from 208 patients who underwent radical prostatectomy for clinically localized prostate cancer and provided informed consent. Each prostate sample was histologically evaluated by multiple experienced uropathologists to ensure that at least 90% of the sample was cancer and that the sample had a sufficiently large high-density tumour area for nucleic acid and protein extraction. Additional confirmation procedures are provided in the Methods.

Patient ages ranged from 50 to 88 years (median 69 years). Preoperative PSA levels ranged from 1.25 to 1,270.14 ng ml^{-1} (median 38.8 ng ml^{-1}). Gleason scores were distributed as follows: Gleason score 6: 17 patients (8%); 7: 96 (46%); 8–10: 94 (45%). Organ-confined carcinoma (pT2) was found in 105 patients; 93 showed extra-prostatic tumour extension (pT3); and 10 patients had advanced disease that had invaded bladder, rectum, or pelvic muscles (pT4). 15 patients (7%) had bone metastasis prior to surgery as determined by the referring urologist. 11% of patients who underwent lymph node dissection had positive lymph node invasion. The cohort also contained 18 patients who exhibited some level of metastasis at diagnosis, all of whom were hormone sensitive. 16 patients were treated with ADT before surgery, mostly to reduce the size of their primary tumour. The clinical and pathological characterization of the final cohort is summarized in Supplementary Data 2.

We applied four omics technologies to characterize isolated biomolecules from the Chinese prostate cancer cohort: whole genome sequencing (WGS, 112.03× on average per sample), whole genome bisulfite sequencing (WGBS, 29.5× average coverage per CpG per sample), RNA sequencing (138.51 M mapped paired-end reads on average) and miRNA sequencing (19.75 M mapped reads on average). The study design is diagrammed in Extended Data Fig. 1b.

Two cases exhibited extreme hypermutation phenotypes and were found to contain mutations in DNA repair genes. Patient 13, who also had a pre-surgery treatment history (Zoladex+bicalutamide), had a germline mutation in *MSH6* (3991C>T, Arg1331Ter). Patient 502, who also had bone metastasis, had a somatic indel in *MLH1* (Fig. 1). These outliers were excluded from the integrative analysis.

Recapitulation of the TCGA somatic mutation landscape with the CPGEA pipeline

The 13 Western prostate cancer cohorts used different technologies, platforms, and measurement depths, as well as bioinformatics tools (Supplementary Data 1). To ensure a meaningful comparison of the somatic mutation landscape between Chinese and Western prostate cancer cohorts, we applied the same bioinformatics pipeline to the CPGEA and TCGA cohorts. We decided to focus our efforts on obtaining and analysing TCGA data because it was the most representative study, and it was performed with the most comparable technology. We first ensured that our CPGEA pipeline applied to TCGA PRAD data recapitulated the reported TCGA results. TCGA raw data (phs000178.p6) for the 333 patients used in the original publication was downloaded from NCI's Genomic Data Commons (GDC) portal. Altogether, we obtained whole exome sequencing (WES) for 333 tumour-normal pairs (coverage ~100× using 2×76 bp reads) and RNA-seq for 332 tumour samples (2×48 bp reads). We also obtained whole genome sequencing (WGS) for 114 tumour-normal pairs. 19 pairs were "high pass" at ~45× coverage using 2×101 bp reads, and 100 pairs were "low pass" at ~7× coverage using 2×51 bp reads. We applied our bioinformatics pipelines to this TCGA data to estimate somatic mutations, gene fusions, and copy number abnormalities (CNAs), and compared them with results reported by TCGA.

For somatic mutation detection, we used 333 WES datasets aligned to hg38 by TCGA. The mutation burden per tumour was calculated based on the number of somatic mutations detected in exome sequencing space (32.7 Mb), and it was consistent with the TCGA report (median 0.70 per Mb from both pipelines, Extended Data Fig. 1c). The mutation frequencies of significantly mutated genes (SMGs) calculated by the CPGEA pipeline were strongly consistent with the frequencies TCGA reported (Extended Data Fig. 1f). For example, TCGA reported a *FOXA1* mutation rate of 4% (13 out of 333) in primary prostate cancers in their original publication. In the harmonized hg38 GDC portal, this number was updated to 6% (20 out of 333 tumours). Our pipeline estimated this rate to be 4% (13/333 tumors). Thus, our finding that *FOXA1* mutations are much more frequent in Chinese prostate cancer than in Western prostate cancer withstood reanalysis.

For somatic gene fusion detection, we used 332 RNA-seq datasets aligned to hg38 by TCGA. One patient (ID: TCGA-YL-A8SF) who had an *ERG* fusion in the original publication was missing from the GDC portal at the time of download. We applied the same pipeline based on

SOAPfuse software, but we did not apply panel of normal filtering because TCGA did not have RNA-seq data for matched normal samples and had only a limited number (43) of RNA-seq datasets from normal blood. We compared the frequencies of the 4 important fusion genes (*ERG*, *ETV1*, *ETV2*, and *FLI1*) reported in the original publication and found strong concordance between the two pipelines (Extended Data Fig. 1f). We estimated the total *ERG* fusion rate to be 42%, consistent with TCGA's report of 46%. We also note that we detected the *SCHLAP1-UBE2E3* fusion in the TCGA cohort (Fig. 2d), which was not previously reported in their original publication. We validated 35 out of 39 *SCHLAP1-UBE2E3* fusions in our CPGEA cohort (Extended Data Fig. 4b, Supplementary Data 6, website), suggesting that the *SCHLAP1-UBE2E3* fusion detected in the TCGA cohort could also be valid.

For somatic CNA detection, we used 114 WGS datasets downloaded from the legacy GDC portal. The original TCGA publication used Affymetrix SNP arrays for somatic CNA detection. To apply the CPGEA bioinformatics pipeline to TCGA samples, we used only the subset of samples with available WGS data (114 patients of 333) and compared the results to those generated using the Affymetrix array on the same patients. In addition, we identified recurrent CNA genomic lesions from the 114 array datasets using GISTIC2.0 with the parameters used in the original TCGA publication. We note that this approach does not completely remove many confounding factors due to differences in sequencing platform. Since the legacy WGS data was aligned to the hg19 genome assembly, and TCGA did not have WGS data aligned to hg38, we realigned all reads to the hg38 genome using our pipeline (see the supporting website). The CPGEA pipeline based on Control-FREEC and GISTIC2 was applied to the realigned WGS reads. The genome-wide view of the raw copy number alterations we detected was consistent with the TCGA report (Extended Data Fig. 1e). The recurrent genomic lesions with amplification or deletion reported by TCGA were recapitulated by our CPGEA pipeline (Extended Data Fig. 1d). The frequencies of gene-level amplification and deletion were also consistent between the two pipelines (Extended Data Fig. 1f).

In summary, we found that the results generated by the two sets of pipelines on TCGA data were strongly concordant, though not identical. Thus, all major conclusions made by TCGA hold when our pipeline is applied. This consistency is not surprising, as we used many of the same tools used by TCGA. However, this also reflects a significant effort in the field to develop more robust bioinformatics tools for reproducible results. Many of the other studies we included in this study represent state-of-the-art technology and analyses, and we are confident

in their major findings. However, because some used different technologies than were utilized for CPGEA (e.g., CNV arrays), we used high-level comparisons to compare CPGEA results to those cohorts. We firmly acknowledge the limitations of comparing across cohorts and that these are confounding factors in these comparisons.

Somatic mutation landscape

Across the CPGEA cohort, the mutation burden was 1.0 per Mb, corresponding to 36 non-synonymous mutations per cancer genome. This estimate is remarkably consistent with the mutation burden estimated from exome-seq in Western cohorts. Alexandrov signatures 2, 4, 6, 12 and 24 were each found in less than 5% of cases (Extended Data Fig. 1g, h).

Among the CNAs detected across the CPGEA cohort, the most frequent arm-level gains were 8q (20%), 3q (7%), 3p (5%), 7q (7%), 8p (5%), and 7p (6%). The most frequent losses were 8p (16%), 16q (7%) 13q (4%) and 18q (3%) (Fig. 2b, c, Extended Data Fig. 2a).

We detected 34,598 somatic structural variations (SVs), 29,454 of which were intra-chromosomal and 5,144 of which were inter-chromosomal (mean 166, 0 to 175 per genome). The intra-chromosomal changes included deletions (16,901), insertions (67), inversions (7,750), and tandem duplications (4,736) (Fig. 1, Extended Data Fig. 3a).

We categorized SVs into five tiers and identified candidate driver events, including events indicating enhancer hijacking (Extended Data Fig. 3d, Supplementary Data 5, Methods). For example, 5 tumours shared the same chromosome 15 inversion, which spanned a topologically associated domain (TAD) boundary (Extended Data Fig. 3c). Multiple genes located upstream of this inversion exhibited elevated expression in affected cancers (Extended Data Fig. 3c, Supplementary Data 5), supporting the hypothesis that a strong enhancer was repositioned by the inversion to upregulate genes in a neighbouring TAD.

Of the 88 fusion candidates selected for Sanger sequencing validation, 9 were inconclusive because of PCR reaction failure.

The striking difference in TMRPSS2-ERG fusions motivated us to compare the landscape of mutated genes between our Chinese cohort and those established in previous studies, which

mostly consisted of Caucasians. Altogether, we defined 11 core prostate cancer genes, 98 Western prostate cancer genes, and 115 Chinese prostate cancer genes, including 61 novel ones (Extended Data Fig. 4i). We also identified new prostate cancer driver genes, including *PTPRK, NFE2L3, USP48, MRRF, HEXDC* and *EBF1*, expanding those known from previous publications. Although *USP48* was not listed in the core 97-gene "long tail" of driver genes, it was reported with a significant q-value previously[1], underscoring the credibility of this gene being a true driver for prostate cancer.

We defined the spectrum of noncoding mutations by prioritizing the nearly 1 million somatic SNVs to obtain 41,109 potentially functional noncoding SNVs (Extended Data Fig. 5a). The recurrence of noncoding mutations did not differ much from that of coding mutations, suggesting an abundance of noncoding drivers. The 'long-tail' of recurrent noncoding mutations is consistent with the notion that they exert a relatively small effect size and have complex regulatory impact.

Despite the drastic difference in mutational spectra, our study suggests that most genome-wide alteration signatures of prostate cancer are remarkably consistent between Chinese and Western prostate cancer. This disease has a relatively low mutation burden but is characterized by ample CNAs and chromosomal rearrangements (Figs. 1, 4). Although several new genes emerged as prostate cancer genes, we believe that the majority of genes and pathways implicated in prostate cancer have been discovered.

FOXA1 mutations in Chinese prostate cancer

Almost all *FOXA1* mutations in Chinese prostate cancer occurred within the hot spot. These mutations did not directly target the DNA binding domain; rather they influenced the loop domain, which may mediate AR interaction (Extended Data Fig. 6c). These results suggest that *FOXA1* somatic mutations in Chinese prostate cancer are likely oncogenic drivers that manifest their impact by modulating AR signalling. Consistent with this hypothesis, *FOXA1* expression was higher in tumours than in normal samples, and tumours with *FOXA1* mutations had higher expression than those with wild type *FOXA1* (Fig. 3b).

FOXA1 was mutually exclusive with ETS fusions. *SPOP* was also mutually exclusive with ETS fusions; however, the mutual exclusion test between *FOXA1* and *SPOP* did not pass the statistical significance cutoff (Extended Data Fig. 6f, h).

To dissect the downstream effects of *FOXA1* mutations, we examined differential gene activities based on *FOXA1* mutation patterns. The comparisons between the in-frame and frameshift groups and between the frameshift and missense groups revealed additional metabolic pathways.

Validation of *FOXA1* mutations

The discovery that *FOXA1* had the most recurrent mutations was further strengthened by the high rate of validation. Two independent methods were used to validate *FOXA1* mutations: Sanger sequencing and RNA-seq data. Putative somatic SNVs were validated by conventional Sanger-based sequencing analysis of PCR products obtained from the tumour WGA DNA. All Sanger validation (86 out of 90, 95.6%) details are listed in Extended Data Fig. 6a and the supporting website (http://www.cpgea.com). For RNA-seq validation, the missense mutations were validated by looking at RNA-seq BAM files in the IGV browser. For each *FOXA1* indel mutation, a new reference of 300 bp of flanking sequence was reconstructed around the insertion / deletion position. RNA-seq reads were realigned to these new references by Bowtie2 in each sample with perfect sequence matches on each end. Reads that perfectly matched and overlapped with the indel position were considered validation reads. The validation rate using RNA-seq data was 100% (63 out of 63).

DNA methylation abnormalities

At a global level, prostate cancer exhibited classic loss of DNA methylation: we found that the mean methylation levels of normal samples were 78.3% and of tumours were 68.8% ($P = 3.7\times10^{-32}$, two-sided Wilcoxon signed-rank test). Global CpG methylation in normal prostate tissue exhibited a classic bimodal distribution, whereas methylation in tumours was strongly shifted towards lower levels (Fig. 4a, Extended Data Fig. 7b).

Individual PMDs ranged from 100 kb to 7Mb (mean 489 kb). The degree of hypomethylation varied across tumours (Extended Data Fig. 7a).

At a focal level, we identified a total of 96,037 hypomethylated DMRs (hypoDMRs) after removing PMDs, covering 1,378,727 total CpGs. 80,793 hypomethylated DMRs (84%) were shared by ≤ 2 pairs, whereas 1,172 (1.2%) were shared by ≥10 tumours (Extended Data Fig. 8a). Similarly, we identified 17,131 hypermethylated DMRs (hyperDMRs) covering 739,986 total CpGs; 8,549 DMRs (50%) were shared by ≤ 2 pairs, and 4,214 (25%) were shared by ≥10 tumours (Extended Data Fig. 8b). We found that recurrent hyperDMRs overlapped CGIs significantly (Extended Data Fig. 8i).

289 CGIs displayed significant hypermethylation in tumours compared to normal controls, consistent with CIMP (Extended Data Fig. 8h, i). They were associated with 334 genes, including *GSTP1*, *RASSF1* (biomarkers used in ConfirmMDx[2-4]), and *CDKN2A*. Surprisingly, 33 CIMP$^+$ tumours were not associated with *IDH1*, *IDH2* or *TET1/2/3* mutations, but were associated with *SPOP* mutation[5] (Extended Data Fig. 8h, j).

Clustering tumours by miRNA expression

We identified 4 clusters of tumours based on miRNA expression levels (Extended Data Fig. 9h), which showed strong correlation with iCluster results (P = 0.0003, Fisher's exact test). miRNA clusters 2 and 3 had higher CNA and epimutation burdens than miRNA clusters 1 and 4, while there were no significant differences in mutation burden across clusters (Extended Data Fig. 9i). miRNA cluster 2 was also enriched for CIMP+ tumours (P = 0.0002, Fisher's exact test), suggesting that miRNA expression in these tumours might be affected by DNA methylation alterations. Indeed, we found that tumour suppressor miRNAs, such as *miRNA-200b*, were epigenetically repressed in those tumours (Extended Data Figs. 8g, 9j). Tumours with miRNA cluster 2 had shorter BCR than other tumours (Extended Data Fig. 9k), similar to CIMP+ tumours.

Molecular subtypes of Chinese prostate cancer

Subtype B displayed the lowest CNA, mutation, and epimutation rates, although it was enriched for *ZFHX3* mutations (7%) and exhibited an expression pattern reminiscent of stem cell pathways. Subtype C was characterized by 8q amplification and mutations in *MYCBP2* (17%) and exhibited upregulated expression of *ERG* fusion target genes, although *ERG* fusions

were only found in 16.7% patients in this subtype. Subtype D combined the CNA profiles of subtypes A and C and thus had the highest CNA burden as well as the highest epimutation rate, although the mutation rate was similar to that of subtypes A and C. Subtype D exhibited higher expression of genes in the MMR pathway and HR activation, consistent with the high CNA load. Patients in subtype D also had the shortest BCR (Extended Data Fig. 9e, $P = 0.04$, log-rank test), underscoring the potential prognostic value of molecular taxonomy. Subtype D did not feature any characteristic mutations. Key features of each subtype were highlighted in Extended Data Fig. 9b–d.

Oncogenic pathways

To define the mutational landscape in a pathway centric manner, we compiled 12 pathways important in prostate cancer and compared the genetic lesions in these pathways between Chinese and Western prostate cancer. 10 pathways were directly chosen from Sanchez-Vega et al[6], which assigned genes to pathways based on a combined revision of pathway analyses from previous TCGA marker papers published between 2008 and 2017, a review of the scientific literature[6] (including but not limited to the references in Sanchez-Vega et al.'s Table S2), and expert curation. We then added the AR pathway, HR pathway, and DNA repair pathway based on our own curation and additional literature. Specifically, AR pathway information was taken from https://cbio.mskcc.org/cancergenomics/prostate/pathways/. The HR pathway was combined into the DNA repair pathway to remove redundancy after careful review of the literature[7].

We observed a similar pattern in the Cell Cycle pathway (24% versus 27%) as in the AR pathway, underscoring the importance of these oncogenic pathways in prostate cancer.

Important druggable genes including *EGFR*, *BRAF*, *KRAS* and *TP53* did not seem to play dominant roles in prostate cancer.

Supplementary Information References

1 Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

2 Van Neste, L. et al. Risk score predicts high-grade prostate cancer in DNA-methylation positive, histopathologically negative biopsies. *Prostate* **76**, 1078–1087 (2016).

3 Stewart, G. D. et al. Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: results of the MATLOC study. *J. Urol.* **189**, 1110–1116 (2013).

4 Partin, A. W. et al. Clinical validation of an epigenetic assay to predict negative histopathological results in repeat prostate biopsies. *J. Urol.* **192**, 1081–1087 (2014).

5 Maruyama, R. et al. Aberrant promoter methylation profile of prostate cancers and its relationship to clinicopathological features. *Clin. Cancer Res.* **8**, 514–519 (2002).

6 Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).

7 Dietlein, F., Thelen, L. & Reinhardt, H. C. Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. *Trends Genet.* **30**, 326–339 (2014).