Patterns of Structural Variation Define Prostate Cancer Across Disease States

Meng Zhou[1,2]*, Minjeong Ko[4]*, Anna C. Hoge[4], Kelsey Luu[4], Yuzhen Liu[4], Magdalena L. Russell[4], William W. Hannon[4], Zhenwei Zhang[1,5], Jian Carrot-Zhang[1,2,3], Rameen Beroukhim[1,2], Eliezer M. Van Allen[1,2,6], Atish D. Choudhury[1,3], Peter S. Nelson[4,7], Matthew L. Freedman[1,3,8], Mary-Ellen Taplin[1,3]†, Matthew Meyerson[1,2,3]†, Srinivas R. Viswanathan[1,2,3]†, Gavin Ha[4,7]†

[1] Department of Medical Oncology, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215

[2] Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142

[3] Harvard Medical School, 25 Shattuck St, Boston, MA 02115

[4] Public Health Sciences and Human Biology Divisions, Fred Hutchinson Cancer Center, 1100 Fairview Ave. N, Seattle, WA 98109

[5] Department of Pathology, UMass Memorial Medical Center, 1 Innovation Dr. #2, Worcester, MA 01605

[6] Center for Cancer Genomics, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215

[7] Department of Genome Sciences, University of Washington, 1959 Pacific St, Seattle, WA, 98195

[8] Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215

* These authors contributed equally to this work.

† These authors jointly supervised this work.

Correspondence:

Gavin Ha
1100 Fairview Ave N
Seattle, WA 98019
1-206-667-2802
gha@fredhutch.org

Srinivas Viswanathan
450 Brookline Ave
1-857-215-0573
Boston MA 02215
srinivas.viswanathan@dfci.harvard.edu

Matthew Meyerson
450 Brookline Ave
Boston MA 02215
Matthew_meyerson@dfci.harvard.edu

Mary-Ellen Taplin
450 Brookline Ave
Boston MA 02215
Mary_Taplin@dfci.harvard.edu

ABSTRACT

The complex genomic landscape of prostate cancer evolves across disease states under therapeutic pressure directed toward inhibiting androgen receptor (*AR*) signaling. While significantly altered genes in prostate cancer have been extensively defined, there have been fewer systematic analyses of how structural variation shapes the genomic landscape of this disease across disease states. We uniformly characterized structural alterations across 278 localized and 143 metastatic prostate cancers profiled by whole genome and transcriptome sequencing. We observed distinct significantly recurrent breakpoints in localized and metastatic castration-resistant prostate cancers (mCRPC), with pervasive alterations in noncoding regions flanking the *AR, MYC*, *FOXA1*, and *LSAMP* genes enriched in mCRPC and *TMPRSS2-ERG* rearrangements enriched in localized prostate cancer. We defined nine subclasses of mCRPC based on signatures of structural variation, each associated with distinct genetic features and clinical outcomes. Our results comprehensively define patterns of structural variation in prostate cancer and identify clinically actionable subgroups based on whole genome profiling.

INTRODUCTION

Over the past decade, genomic sequencing studies have progressively sharpened our view of the genetic landscape of prostate cancer (1). Such studies have defined key driver genes in prostate cancer and have enabled the deployment of therapeutic agents in molecularly-defined disease subsets, including potent androgen receptor (*AR*)-targeted therapies (2, 3), poly (ADP-ribose) polymerase (PARP) inhibitors in *BRCA1/2*-altered prostate cancers, and immune checkpoint inhibitors in cancers with microsatellite instability (4–7).

To date, most cancer genomic studies have employed whole exome sequencing (WES) and have thus been focused on mutations or copy number alterations that occur within the protein-coding regions of genes, which represent only 1-2% of the genome. More recent studies applying whole

genome sequencing (WGS) to prostate and other cancers have identified previously underappreciated recurrent alterations in regulatory (non-coding) regions of the genome and have illuminated complex mechanisms of genomic alterations – driven by structural variants (SVs) – that are difficult to discern by WES; in the case of prostate cancer, most of these studies have focused on localized disease, the disease state in which tissue is most readily accessible for profiling (8–22). There remains a need for continued high-resolution genomic discovery efforts in prostate cancer.

In addition to efforts characterizing entire cancer genomes, recent studies have illustrated the importance of molecularly profiling prostate cancer across disease states. While many localized prostate cancers can be cured with surgery or radiotherapy, a substantial portion of higher-risk cancers recur and progress to metastatic disease, which is incurable. Recurrent prostate cancer may have a long natural history, during which time a patient may receive several lines of therapy – with androgen deprivation therapy (ADT) as a backbone – that may shape the cancer's genomic landscape (23).

Indeed, while hormone-refractory castration-resistant prostate cancer (CRPC) has been less extensively profiled than localized prostate cancer, several studies have indicated that CRPCs display genomic landscapes distinct from treatment-naïve disease (24, 25). A cardinal hallmark of CRPC is the reactivation of *AR* signaling in the face of maximal ADT (22, 26–28). This may occur via diverse mechanisms, including the production of constitutively active *AR* splice variants (*AR-V*s) and activating mutations or copy number amplifications of the *AR* gene (29–31) or of regulatory elements distal to the gene body (13, 15, 32). To date, the relative contribution of each of these mechanisms in driving *AR* reactivation in CRPC has not been systematically explored. Also needed is a more global map of significant hotspots of structural variation in prostate cancer genomes, drawn within a rigorous statistical framework.

49 In this study, we performed linked-read WGS on 36 mCRPC tumor-normal pairs. We combined
50 these data with WGS and whole transcriptome sequencing (RNA-Seq) data from previously
51 described localized and metastatic CRPC cohorts (9, 13, 15, 33). We then established a
52 harmonized workflow for the integrative genomic analysis of 278 localized and 143 metastatic
53 CRPC samples, interrogated both hotspots and genome-wide patterns of structural variation, and
54 evaluated their consequences.

55 ## RESULTS

56 ### WGS analysis of localized and metastatic prostate cancer cohorts

57 We performed linked-read whole genome sequencing on 36 biopsy specimens from 33 mCRPC
58 patients and matched blood normal controls. After quality control, 17 tumor samples were
59 excluded based on insufficient tumor purity and/or contamination, reflecting the challenge of
60 obtaining high-purity metastatic biopsies, particularly from bone lesions (34). We included only
61 samples with tumor purity > 15% in downstream analyses so as to increase confidence in
62 structural variant calls (**Methods, Figure 1A, Table S1**). We re-analyzed a linked-read WGS
63 dataset of 23 samples published previously (15), resulting in a total of 42 linked-read WGS
64 samples from 38 patients with mean coverage of 34X (range 21X - 54X) and 33X (range 25X -
65 45X) for tumor and normal samples, respectively (**Table S1A**). The mean molecule length was
66 29 kB and 34 kB in tumor and normal samples, respectively (**Table S1A**).

67 We further combined these data with 101 mCRPC samples sequenced with standard short-read
68 sequencing, published previously (13). This resulted in the generation of a final combined cohort
69 of 143 tumor-normal pairs, which were uniformly analyzed for copy number and structural
70 alterations via a harmonized pipeline (**Figure 1A**). Fifty-four samples (37.8% of 143 samples)
71 were collected at castration resistance, prior to receiving treatment of second-generation
72 androgen receptor signaling inhibitor (ARSi) such as abiraterone and/or enzalutamide ("pre-

73 treatment"), while the remaining 89 samples (62.2% of 143 samples) were collected at
74 progression ("post-treatment", **Figure 1B**, **Table S1B**). We analyzed the somatic single nucleotide
75 variant (SNVs), insertion-deletions (indels), copy number alterations (CNAs), and SVs in the
76 combined cohort and identified recurrent somatic alterations in each of these classes (**Figure 1A,**
77 **Methods**).

78 A total of 2,315,452 SNVs and indels were called, with a mean tumor mutation burden (TMB) of
79 2.82 mutations per million bases (Mb). We confirmed that known driver genes of prostate cancer
80 were enriched for non-synonymous mutations, including *TP53*, *RB1*, *PTEN*, *FOXA1*, *CDK12*, *AR*
81 and *SPOP* among known COSMIC Cancer Gene Census genes (dndscv, $q \leq 0.1$, **Table S1C and**
82 **S1D, Methods**). We detected an average of 272 (range 96-833) SV events per sample. Based
83 on breakpoint orientations, SV events were classified into deletions, inversions, tandem
84 duplications, inter-chromosomal translocations, and intra-chromosomal translocations, while
85 intra-chromosomal translocations were further divided into balanced and unbalanced events
86 based on copy number information (**Methods**). Chromoplexy was detected in 53 samples (37.1%
87 of 143 samples) while chromothripsis was detected in 37 samples (25.9%); these events were
88 not mutually exclusive (Fisher's exact test, log-odds=1.417, p-value=0.612). Ten cases (7.0%)
89 harbored a genome-wide tandem duplicator phenotype (TDP), all of which had *CDK12*
90 inactivating alterations, as recently reported (15, 35). We found that TDP was mutually exclusive
91 with ETS rearrangements (Fisher's exact test, log-odds ratio=0.133, p=0.043) and chromothripsis
92 (log-odds ratio=0.301, p-value=0.007), as previously reported (10, 13, 15, 35).

93 Analysis of CNA events across the genome revealed amplification and deletion peaks in the
94 regions of known prostate cancer genes (10, 13, 15, 24). Many oncogenic drivers of mCRPC,
95 such as *AR* and *MYC*, are within peaks of amplification across the cohort, while tumor
96 suppressors such as *PTEN, TP53,* and *KMT2C* were found within deletion peaks (**Figure S1C**,
97 **Table S1E and S1F**), consistent with prior reports (10, 13, 15, 36).

98 ## Recurrent somatic structural variants in prostate cancer-associated genes

99 Structural variants may either activate or inactivate gene function, depending on the location of
100 the breakpoints and the specific class of SV. We analyzed the potential impact of SVs called
101 across our combined cohort, distinguishing between those with predicted inactivating ("gene
102 transecting events") and activating ("gene flanking events") effects (**Figure 1C, Figure S1C,**
103 **Table S1G and S1H**). Frequent gene transecting alterations were observed at the *TTC28* (37.1%
104 of 143 samples), *LSAMP* (31.5%), and *PTPRD* (23.8%) loci, which have not been extensively
105 studied in prostate cancer, though they have been reported in callsets for certain cohorts (28).
106 Rearrangements involving *TTC28* were predominantly inter-chromosomal translocations between
107 the gene body and various non-recurrent partner loci (**Figure S2E**). This likely represents
108 retrotransposon activity, given that the *TTC28* locus harbors an active L1 retrotransposon element
109 (37–39). Transecting SVs within the *LSAMP* and *PTPRD* genes were predominantly deletions.
110 Both of these genes are sites of deletion/rearrangement in cancer and have been reported to
111 function as tumor suppressors, though they have not been extensively studied within the context
112 of prostate cancer (40–43) (**Figure 1C**). Of note, although gene transecting events would be
113 predicted to disrupt individual genes, the most frequent transecting events identified via this
114 analysis were deletion events that span the adjacent *TMPRSS2* and *ERG* genes (observed in
115 37.8%), which actually produces an activating *TMPRSS2-ERG* fusion.

116 Duplication events that flank an intact gene could activate oncogenes, either by resulting in copy
117 number gain of the gene or by duplicating non-coding regulatory regions (13, 15). In our combined
118 cohort, we observed recurrent tandem duplication events with breakpoints located in the flanking
119 gene regions of several known prostate cancer oncogenes, including *AR* (35.7%), *FOXA1*
120 (16.8%), *MYC* (16.8%), and *CCND1* (14.0%), consistent with frequencies that have been
121 previously reported by us and others (10, 13, 15) (**Figure 1C**).

122 Certain prostate cancer driver genes were altered by multiple classes of structural alterations in
123 both the gene body and flanking regions (e.g., *AR*, *PTEN*), while others were predominantly
124 altered by a single alteration class (e.g., SNVs for *TP53*, intragenic translocations for *TTC28*, or
125 flanking tandem duplications for *MYC*) (**Figure 1C, Figure S1C**). Collectively, these results
126 catalog how diverse classes of rearrangements, both within genes and in intergenic regions, alter
127 prostate cancer genes across disease states.

128 Significantly recurrent breakpoint regions in the mCRPC genome are enriched within 129 enhancer regions and *AR* binding sites

130 Next, we sought to identify significantly recurrent breakpoint (SRB) regions across our combined
131 mCRPC cohort of 143 cases in a genome-wide, unbiased manner. We applied a Gamma-Poisson
132 regression approach to model the occurrences of SV breakpoints within 100 kB windows across
133 the cohort as previously described (44). Importantly, this model nominates significantly recurrent
134 breakpoint regions likely to function as cancer drivers by accounting for six different covariates,
135 including sequence features (e.g., GC-content and transposable elements), fragile sites,
136 heterochromatin regions, DNase I hypersensitivity sites (DHS), and replication timing (**Methods**),
137 which may increase specificity over prior studies that have accounted only for SV frequency or
138 for breakpoint density within a genomic window (10, 13, 15, 45).

139 We identified a total of 55 significantly recurrent breakpoint regions genome-wide across our
140 combined mCRPC cohort (Benjamini-Hochberg corrected, q-value ≤ 0.1, **Figure 2A**, **Table S2A**).
141 Thirty-six (65.5%) SRB regions were located within 1 Mb of 14 known prostate cancer driver
142 genes, including *AR* and its enhancer, *TMPRSS2/ERG*, *TP53*, *PTEN*, *FOXA1*, and *MYC*. For
143 these 14 driver genes, we did not observe significant differences in SV alteration frequencies
144 when comparing between pre-treatment (N=54) and post-progression (N=89) samples, except in
145 the case of *ERG, for* which the SV frequency was enriched in pre-treatment samples (Fisher's
146 exact test, $p = 0.0395$; all other genes had $p > 0.05$, **Figure S3B**). We also did not identify any

147 major differences in the alteration frequencies of prostate cancer genes in four patients who had
148 paired samples collected both before treatment with and after progression on an ARSi. (**Figure**
149 **S3A**).

150 We then sought to compare how SVs drive prostate cancer across disease states. For the
151 localized disease state, we utilized genome alteration calls from 278 primary localized prostate
152 cancer tumors from the PCAWG study (9, 33). Using Gamma-Poisson regression, we first
153 identified 47 SRB regions in localized prostate cancer tumors (**Figure S2B, Table S2B**). Six
154 prostate cancer genes (*TMPRSS2*, *ERG*, *TP53*, *PTEN*, *IL6ST*, *ELK4*) within mCRPC SRB regions
155 were also found within or in proximity (less than 1 Mb) to an SRB region in localized disease. By
156 contrast, four SRBs (three near *SEL1L3* and one near *PRKDC*) were unique to localized disease,
157 while 27 SRBs were unique to mCRPC with six genes nearby (*LSAMP*, *ETV1*, *MYC*, *PTPRD*,
158 *FOXA1*, *AR*). When comparing SV alteration frequencies for the 14 genes located within SRB
159 regions in either mCRPC or localized tumors, 12 genes were significantly more altered in mCRPC
160 samples, while *TMPRSS2* and *ERG* were significantly more altered in localized disease (Fisher's
161 exact test, $p < 0.05$ for all genes, **Figure 2B**). We repeated this comparison using a non-
162 overlapping cohort of localized prostate cancers profiled by WGS and found similar genes
163 enriched for SVs in either the localized or metastatic disease states (**Figure S2C-E**) (18). Thus,
164 localized prostate cancer and mCRPC have significantly different landscapes of recurrent SVs.

165 To explore the potential functional consequences of SVs in intergenic SRB regions, we
166 overlapped SV breakpoints with locations of H3K27ac marks specific to mCRPC (46). We
167 observed that intergenic SVs within SRB regions in the mCRPC cohort included gene flanking
168 events that were enriched at putative enhancer regions for *AR*, *MYC*, and *FOXA1*, which all had
169 frequent focal duplication events at sites marked by mCRPC-specific H3K27ac deposition (**Figure**
170 **2C, S2F**). Interestingly, an intragenic deletion SRB region was observed near the transcription
171 start site of *LSAMP*, also overlapping H3K27ac marks. *PTEN* had a high level of both gene

172 transecting and flanking deletions, leading to SV breakpoints that were spread more broadly
173 around the gene.

174 We also observed an enrichment of metastatic-specific *AR* binding sites (ARBS) compared to
175 localized primary ARBS within the 55 mCRPC SRB regions (**Figure 2D**, one-sided proportion test,
176 $p = 1.05 \times 10^{-8}$). This enrichment was not observed for localized primary SRB regions ($p = 0.22$).
177 These results highlight that SVs within mCRPC SRB regions may be capturing the genome-wide
178 footprint of activated *AR* signaling that occurs with castration resistance.

179 Refined landscape of ETS gene fusions from integrated analysis of the genome and 180 transcriptome

181 We applied gene fusion analysis by integrating both genome rearrangements and fusion RNA
182 transcript information from 127 samples with RNA-seq data (**Figure 1A**, **Table S2C**, **Methods**).
183 For gene fusions involving E26 transformation-specific (ETS) transcription factor gene family
184 members (*ERG*, *ETV1*, *ETV4* and *ETV5*), we detected 50 events supported by both DNA and
185 RNA evidence, 15 supported by only DNA evidence, and 10 supported by only RNA evidence
186 (**Figure 2E**, **Figure S2G**). Overall, 74 samples (51.7% of 143 samples) harbored a fusion event
187 of the *ETS* gene family, consistent with previous reports (47, 48) (**Figure 1B, Table S2C**).

188 Among the ETS fusions, *ERG* was most commonly involved with *TMPRSS2* as the fusion partner
189 (54 out of 57 cases, **Figure 2G**). Other common ETS fusion partners were *SLC45A3* (7 cases)
190 and lncRNA RP11-356O9.1 downstream of *FOXA1* (3 cases). *ETV1* had eight distinct fusion
191 partners, which is consistent with previous reports that *ETV1* is a promiscuous ETS fusion
192 member (49) (**Figure 2F**).

193 We observed that fusions of the ETS family members *ERG*, *ETV1*, *ETV4* and *ETV5* were mutually
194 exclusive, except for one sample which harbored fusions of both *ERG* and *ETV1* (**Figure S2G**).
195 In addition, gene fusion events were correlated with higher expression of the corresponding ETS

196 genes they involved (Wilcoxon rank-sum tests, $p < 0.05$ for all genes, **Figure 2E**). In the 38 cases
197 which did not show any evidence for an ETS fusion, we noted that presence of high-level
198 expression (z-score > 1) of ETS genes *ERG*, *ETV1*, *ETV4*, and *ETV5* were also mutually
199 exclusive (Fisher's exact test, $p = 0.480$ for *ETV4*, $p = 0.363$ for *ETV5*, **Figure S2G**). These may
200 represent cases of missed fusion calls, or cases in which ETS family members are
201 transcriptionally activated through non-genetic mechanisms.

202 Interestingly, we also observed 20 cases (14.0% of 143 cases) involving fusions between the ETS
203 family member *ELK4* and its upstream gene *SLC45A3*. While the *ELK4* locus was an SRB in our
204 analysis (**Figure 2A** and **Figure S2G**), manual inspection of individual samples revealed evidence
205 for a genomic event capable of producing an *ELK4* fusion in only 1 out of 20 cases (**Figure S2G**
206 **and data not shown).** In contrast, 19 other cases showed *ELK4* fusions on RNA-sequencing
207 alone, consistent with a mechanism of cis-splicing or transcriptional read-through events that may
208 perhaps be induced by local genomic alterations (50–52) (**Table S2C**). Importantly, although
209 *ELK4* fusions were significantly correlated with higher expression of *ELK4* (Wilcoxon rank-sum
210 test, $p = 7.91 \times 10^{-5}$, **Figure S2G**), these events were not mutually exclusive with fusions of other
211 ETS family members (Fisher's exact test, $p = 0.472$). Thus, the functional consequences of these
212 *ELK4* fusions and whether they contribute to prostate cancer pathogenesis in a manner similar to
213 other ETS fusions remains to be determined.

214 Classes of rearrangements driving *AR* signaling in mCRPC

215 Genomic alterations involving the *AR* locus play an important role in sustaining *AR* signaling in
216 mCPRC (13, 15, 26, 53). We sought to catalog the spectrum of diverse structural mechanisms
217 that underlie *AR* activation in mCRPC, and the relationship between them, in our combined
218 mCRPC cohort. To understand the relationship between different modes of somatic *AR* activation,
219 we first determined copy number at the *AR* gene body and its upstream enhancer and categorized
220 samples into distinct groups of: **(1)** co-amplification (N = 99, 69.2% of 143 cases); **(2)** selective

221 *AR* gene body amplification (N = 4, 2.8% of 143 cases); **(3)** selective *AR* enhancer gains (N = 17,
222 11.9% of 143 cases), and **(4)** lack of amplification for both (N = 23, 16.0% of 143 cases) (**Figure**
223 **3A-C, Table S3**). For the 122 samples with expression data available, we observed that *AR* gene
224 expression was higher in the co-amplification and selective enhancer categories compared to
225 samples with no amplification, after accounting for tumor purity and ploidy (ANCOVA/TukeyHSD
226 p-values 5.6×10^{-11} and 4.5×10^{-4}, respectively), but not for selective *AR* status (ANCOVA p = 0.098)
227 (**Figure 3B, Methods**). Interestingly, we observed that samples with selective enhancer
228 duplication exhibited similar *AR* expression levels to samples with co-amplification (ANCOVA, p
229 = 0.31), even though enhancer duplications involved lower-copy gains (mean 2.73, range 1.97 -
230 5.02) compared to co-amplified samples (mean 12.87, range 1.55 - 150.57) (**Figure 3A**). This is
231 consistent with previous results (15, 28) and suggests a mechanism whereby *AR* expression
232 levels are increased through even modest genomic expansion of enhancer elements.

233 We then systematically and manually curated the diverse mechanisms of rearrangements
234 activating *AR* signaling by analyzing patterns of SVs at the *AR* locus (**Figure 3C, Table S3,**
235 **Methods**). We observed a total of 62 samples (43.4% of 143 samples) with tandem duplication
236 SV events that spanned the enhancer with breakpoints located within 1 Mb, including 16 cases
237 (11.2% of 143 samples) with selective enhancer copy number amplification status (**Figure 3D**).
238 Thirty-two samples (22.4% of 143 samples) harbored intragenic rearrangements within *AR*, which
239 may have implications for the production of truncated, constitutively-active *AR* splice variants (31).
240 For example, in case 01115414-TA1, we observed a tandem duplication breakpoint selectively
241 amplifying exons 1-4 of AR, but not exons 5-8 of *AR*, which includes the ligand binding domain;
242 such an event could promote selective expression of a constitutively active truncated *AR* variant,
243 although RNA-Seq data was not available on this sample. (**Figure 3E**). In another case, DTB-
244 124-BL, our reanalysis confirmed that a focal deletion involving exons 1-4 resulted in the
245 expression of truncated *AR* variants as previously described (28) (**Figure S3C**). Interestingly, in

246 the 21 samples with selective *AR* enhancer or selective *AR* gene body copy number gain, none
247 harbored intragenic SV events in *AR*.

248 We also examined the landscape of complex rearrangement mechanisms involving *AR*; these
249 mechanisms involve multiple SV events and copy number patterns, including chromothripsis,
250 extrachromosomal DNA (ecDNA), chromoplexy, and breakage-fusion-bridge cycle (BFB)
251 (**Methods**). Chromothripsis of a region or the entire X chromosome involving the *AR* locus was
252 detected in 5 samples, all of which had co-amplification of *AR* and enhancer, suggesting that
253 following repair after catastrophic DNA shattering the *AR* locus was retained or further amplified
254 (**Figure 3F, Figure 3G**). Thirteen samples (9.1% of 143 samples) showed very high levels of *AR*
255 and enhancer copy number, suggesting the possibility of their presence on extrachromosomal
256 elements (ecDNA, **Figure 3H**). In 40 samples (28.0% of 143 samples), the most frequent complex
257 rearrangement mechanism, BFB, led to *AR* locus amplification, including instances following
258 chromothripsis (14, 54) (**Figure 3G**). Overall, we noted that complex rearrangement events, which
259 frequently co-occurred, were significantly enriched in samples with co-amplification of *AR* and
260 enhancer compared to those with selective enhancer copy number gain status (Fisher's exact
261 test, $p = 1.52 \times 10^{-4}$).

262 **Distinct signatures of structural rearrangement patterns in mCRPC**

263 To systematically characterize genome-wide structural rearrangement patterns in mCRPC, we
264 performed rearrangement signature analysis using SV breakpoint features, non-negative matrix
265 factorization, and known reference signatures (12, 55) (**Methods**). First, we derived signatures
266 *de novo*, which identified eight signatures: six that matched reference signatures (RefSigs) also
267 observed in localized prostate cancer (> 0.91 cosine similarity), one that matched an ovarian
268 cancer RefSig.R14 associated with large segment (100 kB-10 Mb) TDP (0.96 cosine similarity),
269 and one that was likely an artifact specific to linked-read sequencing (**Figure S4A-C, Table S4A**
270 **and 4B**). Therefore, we excluded the linked-read data and focused on standard WGS data from

271 101 mCRPC cases for further SV signature analysis. We fit these samples to the nine known
272 RefSigs from localized prostate cancer (R1-4, R6a-b, R8-9, R15) and the one (R14) from ovarian
273 cancer (**Figure S4A**, **Table S4C**). Overall, eight of the RefSigs were detected across our cohort
274 (R1-2, R4, R6a-b, R9, R14-15). Notably absent in mCRPC were RefSig.R8 (short, 1-10 kB
275 inversions) and RefSig.R3, which is associated with germline *BRCA1* mutations and short (1-100
276 kB) tandem duplications (11, 12, 55, 56) (**Figure S4D**). By contrast, we observed increased
277 prevalence of some signatures in mCRPC compared to localized disease, including RefSig.R2
278 (large SV classes, abundant translocations; 97% vs. 60%), RefSig.R4 (clustered translocation
279 events; 37% vs. 27%), and RefSig.R15 (large deletions and inversions, 48% vs. 37%) (**Figure**
280 **S4D**).

281 To investigate whether molecular subtypes in mCRPC can be grouped based on SV patterns, we
282 applied hierarchical clustering on the exposure of the eight fitted signatures and identified nine
283 distinct SV clusters (**Figure 4, Table S4C**). We observed that samples in SV Cluster 1 were
284 composed of non-clustered translocation events and were significantly enriched for the presence
285 of chromoplexy (χ^2 test, FDR corrected, q = 0.12). SV Cluster 3 was characterized by many short
286 deletions and was significantly enriched for *BRCA2* mutations ($q = 5.01 \times 10^{-4}$). SV Cluster 5 was
287 significantly enriched for *SPOP* mutations (q = 0.02), with no instances of ETS gene family fusion
288 (q=0.06), consistent with previous reports (57). SV Cluster 6 had the highest prevalence of *TP53*
289 mutation (q = 0.02), while SV Cluster 7 samples harbored the TDP associated with *CDK12*
290 inactivation ($q = 3.52 \times 10^{-11}$) as well as enrichment for *CCND1* gains (q = 0.02), consistent with
291 previous reports (35, 58). The remaining clusters did not have enrichment for any alterations in
292 known driver genes; however, distinct SV patterns were still evident in SV Cluster 4 (non-clustered
293 tandem duplications), 8, and 9 (increased clustered SV events of various classes).

294 While SV Clusters 3, 5 and 6 had significant enrichment of mutations in *BRCA2*, *SPOP,* and *TP53*,
295 respectively, not all samples within each cluster harbored these mutations. Intriguingly, we further

296 noted that clinical outcomes showed significantly better stratification when using SV Clusters 3,
297 5, and 6 for outcome stratification compared to using the associated mutation status itself (**Figure**
298 **S4D-E**). Specifically, SV Cluster 5 had significantly better overall survival than SV Clusters 3 and
299 6 (log-rank test, p=0.01), while the sample group with *SPOP* mutations did not have significantly
300 greater survival compared to the sample groups with *BRCA2* and *TP53* mutations (log-rank test,
301 p=0.45) in this cohort. Together, these results indicate the analysis of genome-wide patterns of
302 rearrangements may provide a way to further refine molecular subtypes in mCRPC.

303 DISCUSSION

304 We present a large-scale and comprehensive integrative genomic analysis of both localized
305 prostate cancer and mCRPC, with a focus on how structural variation drives each of these
306 clinically distinct disease states. The size of our cohort as well as our harmonized analysis pipeline
307 enable a sharper view of the genetic alterations that drive prostate cancer across its natural history
308 as compared with prior studies, which have involved either smaller cohorts or been limited to a
309 single disease state (9, 13, 15, 59).

310 In contrast to somatic SNVs/indels and CNAs that occur within coding regions, the functional and
311 clinical significance of alterations within noncoding regions has often been more challenging to
312 interpret, as localized variations in mutability may result in the nomination of certain recurrently
313 mutated sites that do not necessarily drive cancer (11, 12, 44). This issue is even more complex
314 for SVs, in which different classes of SVs spanning the same loci would be predicted to have
315 distinct functional consequences. Our study addresses the former issue by identifying genomic
316 hotspots of structural variation with rigorous correction for covariates including nucleotide
317 composition, replication timing, sensitivity to DNA breaks, repetitive elements, and chromatin
318 state. We address the latter issue by careful curation of SV classes to distinguish those that are
319 likely to be activating versus inactivating (**Figures 1B and 3; Methods**).

320 Our approach has produced several insights into the recurrent rearrangements that drive prostate
321 cancer. First, several top hotpots of rearrangement genome-wide lie in noncoding regions outside
322 the boundaries of known prostate cancer genes, as previously reported (10, 13, 15). In many
323 cases, such as for *AR*, *MYC*, and *FOXA1*, these hotspots overlap with active chromatin marks
324 and likely represent distal regulatory regions for neighboring prostate cancer genes, as shown by
325 our analyses overlapping SVs with ChIP-Seq on mCRPC specimens (46) (**Figure 2**). These data
326 are intriguing in light of the observation that a majority of prostate cancer germline susceptibility
327 loci are in noncoding regions (60). Second, the loci altered by rearrangements differ across
328 prostate cancer disease states (**Figure 2B**). For example, *TMPRSS2-ERG* rearrangements are
329 enriched in localized prostate cancer versus mCRPC, while alterations in *AR, FOXA1, MYC*, and
330 *LSAMP* are more frequent in mCRPC than in localized disease. Third, certain driver genes are
331 enriched for alteration by SVs as compared to other mutagenic processes. For example, *PTEN*
332 inactivation frequently occurs via gene transecting SV events, while *TP53* inactivation is primarily
333 caused by SNVs (**Figure 1C and Figure S1**).

334 Our systematic genomic discovery efforts confirm the primacy of *AR* as a target of somatic
335 alteration in hormone-refractory mCRPC (13, 15, 26, 53, 61). We have precisely catalogued the
336 diverse genomic mechanisms leading to *AR* activation across our large aggregate cohort and find
337 that different alteration mechanisms are associated with differing levels of *AR* amplification.
338 Whether the precise mechanism by which *AR* is altered in a given patient is associated with
339 differences in response to *AR* pathway inhibition warrants further investigation in clinically
340 annotated cohorts. High levels of *AR* signaling in mCRPC may also underlie the patterns of
341 structural variation seen in this disease state. Strikingly, we found that *AR* binding sites
342 overlapped several of the top SV hotspots in mCRPC (**Figure 2D**), consistent with the notion that
343 androgen signaling may induce DNA double-strand breaks that resolve as rearrangements (62).

344 In addition to alterations in highly validated prostate cancer genes, we identified highly recurrent
345 rearrangements near or involving genes that have not been extensively studied in prostate cancer
346 in multiple cohorts, such as *LSAMP, PTPRD*, and *TTC28*. *LSAMP* encodes a cell-surface
347 glycoprotein and has a possible tumor suppressor role in several cancers (40–42); notably,
348 deletions near the *LSAMP* locus have been shown in one report to be enriched in African
349 American men with prostate cancer (63). *PTPRD*, a receptor protein tyrosine kinase, has been
350 previously identified as a target of inactivating alteration in glioblastoma (43). We observed
351 frequent SVs near the *TTC28* locus, which encodes an L1 retrotransposon element, specifically
352 in mCRPC (**Figure 1C)**. L1 retrotranspositions originating from *TTC28* have been reported
353 previously in colorectal cancer (37–39); our results raise the intriguing possibility that they may
354 also be frequent in prostate cancer, and may be activated by the pressure of hormonal therapy.
355 Interestingly, we also observed SRBs near *ELK4* along with a relatively high frequency of
356 *SLC45A3-ELK4* chimeric transcripts, although it was not clear how the rearrangements at this
357 locus produced the chimeric transcripts in most cases. Whether this fusion functions similarly to
358 or in a distinct mode from other ETS fusions is an exciting area for future study.

359 Our study also extends beyond the analysis of SVs at individual loci to molecularly subclassify
360 prostate cancers based on their genome-wide signatures of structural variation. Sample clustering
361 based on SV signature exposure defines distinct molecular subtypes of prostate cancer and may
362 find utility alongside signatures of single base substitution and copy number to more precisely
363 define tumor subtypes (55, 64–67). In the mCRPC cohort, we identified 9 molecular subtypes
364 based on SV signature, and several clusters had clear associated genomic alterations including
365 chromoplexy (cluster 1), *BRCA2* alterations (cluster 3), *SPOP* alterations (cluster 5), *TP53*
366 alterations (cluster 6) and *CDK12*/*CCND1* alterations (cluster 7). Future studies with larger WGS
367 and RNA-Seq cohorts may identify associated alterations or transcriptional signatures in the
368 remaining clusters. Notably, unsupervised clustering identified samples with clear SV signatures

369 but without detectable associated mutations in genes or pathways that plausibly contribute to the
370 genomic alterations (**Figure 4**). Moreover, clinical outcomes were more separated by SV
371 signature cluster than by alterations of the mutations associated with those clusters (**Figure S4D-**
372 **E**).

373 In sum, these results highlight the dynamic complexity of rearrangements in prostate cancer
374 across disease states and provide insights into new mechanisms of oncogenesis that can be
375 functionally prioritized in future studies. More broadly, our work underscores the key role of large-
376 scale WGS studies in the derivation of a comprehensive molecular taxonomy of prostate cancer.

377 # METHODS

378 ## Data and code availability

379 - Whole genome sequencing data have been deposited at dbGaP under accession number
380 phs001577 and access is available upon request.
381 - All original code has been deposited at GitHub and is publicly available as of the date of
382 publication. Links are provided in the key resources table.
383 - Any additional information required to reanalyze the data reported in this paper is available
384 from the lead contact upon request.

385 ## Sequence data processing for linked-read genome sequencing data

386 Data processing of the linked-read genome sequencing data include high molecular weight DNA
387 preparation and sequencing library construction followed protocols as previously described (15).
388 DNA was extracted from tumor samples using the MagAttract HMW DNA Kit (QIAGEN), and then
389 quantified using Quant-it Picogreen assay kit (Thermo Fisher) on a Varioskan Flash Microplate
390 Reader (Thermo Fisher). For germline samples, pre-extracted DNA was size-selected on the
391 PippinHT platform (Sage Science) and then quantified using the Quant-it Picogreen assay kit
392 (Thermo Fisher) on a Varioskan Flash Microplate Reader (Thermo Fisher). Libraries were
393 constructed using the 10X Chromium protocol (10X Genomics), with the fragment sizes
394 determined using the DNA 1000 Kit and 2100 BioAnalyzer (Agilent Technologies) and quantified
395 using qPCR (KAPA Library Quantification Kit, Kapa Biosystems). WGS libraries were sequenced
396 using the Illumina HiSeqX platform. The Long Ranger v2.2.2 pipeline (10X Genomics) was used
397 for aligning sequence reads to the human genome hg38 (GRCh38).

398 Samples were excluded from the analysis based on having tumor purity less than 15% estimated
399 by TitanCNA or based on cross-individual contamination indicated by SNP fingerprinting. A total
400 of 17 samples with linked-read data was excluded (**Table S1J**).

401 ## List of known prostate cancer driver genes

402 For analyses limited to established prostate cancer driver genes, a curated list of 159 known
403 prostate cancer driver genes was assembled from several prior studies (10, 13, 15, 24). The list
404 of genes are provided in **Table S1**.

405 ## Somatic mutation analysis

406 ### *Somatic mutation detection*

407 Somatic mutation calls for samples based on linked-read sequencing were generated by Mutect2
408 from the Genome Analysis Toolkit (GATK) (68). Default parameters were used on individual pairs
409 of tumor and normal samples following the standard GATK pipeline. A panel of normals based on
410 all normal samples was used to filter out germline variants. The SNV calls were further processed
411 using the modified version of LoLoPicker (69) as described previously (15). The panel of normals
412 for LoLoPicker was generated from 52 normal samples based on linked-read sequencing. The
413 final SNV call set was composed of the common variants called by both Mutect2 and LoLoPicker.
414 Somatic indels for linked-read samples were called by Strelka (70). All parameters were default
415 except the following modifications: sindelNoise = 0.000001, minTier1Mapq = 20. Somatic
416 mutation calls for the 101 WGS samples based on short-read sequencing including SNV and
417 indels based on Strelka were obtained from a prior study (13). All variants were further annotated
418 using annovar with “table_annovar.pl” to functionally annotate genetic variants. The parameter -
419 neargene was set to 5000 to define the promoter region as 5 kB upstream of the transcription
420 start site of a protein coding gene.

421 ### *Analysis of significantly mutated genes*

422 R package dndscv (71) was used to identify significantly mutated genes. For driver discovery on
423 GRCh38, a precomputed database corresponding to human genome GRCh38.p12 was
424 downloaded and used as the reference database. A global q-value ≤ 0.1 was applied to identify

425 statistically significant (novel) driver genes. To reduce false positives and increase the signal to
426 noise ratio, we only considered mutations in Cancer Gene Census genes (v81) (72).

427 Copy-number analysis of linked-read WGS and short-read WGS data

428 *Copy-number calls*

429 The ploidy and purity corrected copy-number of all mCRPC samples in this study was analyzed
430 by TitanCNA (73) and ichorCNA (74), with different pipeline settings. For WCDT samples, the
431 snakemake workflow for Illumina sequencing was applied with the following parameters modified:
432 ichorCNA_normal: c(0.25, 0.5, 0.75); ichorCNA_ploidy: c(2,3,4); ichorCNA_includeHOMD: TRUE;
433 ichorCNA_minMapScore: 0.75; ichorCNA_maxFracGenomeSubclone: 0.5;
434 ichorCNA_maxFracCNASubclone: 0.7; TitanCNA_maxNumClonalClusters: 3;
435 TitanCNA_maxPloidy: 4. The workflow is available at
436 https://github.com/GavinHaLab/TitanCNA_SV_WGS.

437 For linked-read data samples, a Snakemake workflow for 10X Genomics whole genome
438 sequencing data was used with the following parameters modified:
439 TitanCNA_maxNumClonalClusters: 3; TitanCNA_maxPloidy: 4. TitanCNA solutions were
440 generated for number of clonal clusters from 1 to 3 and ploidy initializations from 2 to 4. Optimal
441 solutions were selected as described, with manual inspection to confirm tumor ploidy and clonal
442 cluster selection (15); solutions are provided in **Table S1J**. The workflow can be accessed at
443 https://github.com/GavinHaLab/TitanCNA_10X_snakemake. The final copy-number call-set is
444 included in **Table S1I**.

445 *Recurrent somatic copy-number alteration*

446 GISTIC 2.0 was used to detect regions with recurrent CNA in mCRPC samples. For input, all copy
447 numbers (logR_Copy_Number from TITAN output) were converted to log2 copy ratio using the
448 median logR copy number from genome-wide (separately for autosomes and X chromosome) as

449 denominator. We set corrected logR copy number to -1.5 for segments where corrected log R
450 copy number below -1.5 and set values to 0 if copy neutral. GISTIC2.0 was run with the following
451 parameters: td 0.5; ta 0.1; genegistic 0; maxseg 5000; js 4; cap 1.5; broad 1; brlen 0.75; conf 0.99;
452 qvt 0.25; armpeel 1; rx 0; gcm mean; do_gene_gistic 1; savegene 1; scent median. Wide peaks
453 detected by GISTIC2 were re-annotated based on overlapping genomic coordinates, using
454 prostate cancer driver genes.

455 **Structural variant analysis**

456 *Structural variant detection in linked-read and short-read whole genome sequencing data*

457 For each tumor-normal pair of samples with linked-read genome sequencing data, three variant
458 callers were used to detect structural variants: SvABA (75), GROC-SVS (76), Long Ranger
459 version 2.2.2 (https://support.10xgenomics.com/genome-
460 exome/software/pipelines/latest/using/wgs).

461 The SvABA analysis was performed using default tumor-normal paired settings. Re-analysis of
462 low confidence (based on evidence from discordant and split reads) events filtered by SvABA was
463 performed to 'rescue' SVs using linked-read barcode overlap between pairs of breakpoints within
464 a given SV event, as previously described (15). Only SV events having span of 1.5 times the
465 mean molecule length in the library were considered for rescue. We further rescued low
466 confidence intra-chromosomal SV events with span > 50 kB filtered by SvABA if at least one of
467 the breakpoint pair was within 100 kB of a CNA boundary or (2) if both breakpoints were each
468 within 1 Mb of the boundaries for the overlapping CNA event and the length of the SV overlaps
469 this CNA event by > 75%. Inter-chromosomal translocation SV events filtered by SvABA are
470 rescued if both breakpoints were within 100 kB of CNA boundaries.

471 GROC-SVS analysis was performed using two-sample (tumor-normal paired) mode or three-
472 sample (pre-treatment, post-progression, normal) mode when applicable. SV events were

473 retained if all following conditions were satisfied: (1) $p < 1x10^{-10}$, (2) minimum barcode overlap ≥
474 2 on the same haplotype, (3) no more than 1 barcode overlap between different haplotypes, (4)
475 FILTER value reported by the software was within this set {"PASS", "NOLONGFRAGS",
476 "NEARBYSNVS", or "NEARBYSNVS; NOLONGFRAGS"}, and (5) classified as somatic.

477 Long Ranger analysis generated SV calls for tumor and normal samples, independently. For each
478 tumor-normal pair, both large SVs ("large_sv_calls.bedpe") and deletions ("dels.vcf") were
479 combined for individual samples. Somatic tumor SVs were determined as events that were not
480 found in the matched normal sample based on the left breakpoints in tumor and normal being
481 within 1 kB and the right breakpoints in tumor and normal samples being within 1 kB. Only SV
482 events with FILTER values within this set {"PASS", "LOCAL_ASM", "SV", "CNV, SV"} and intra-
483 chromosomal events with span ≥ 100 kB were considered. SV events were only retained if both
484 breakpoints of an SV event were within 500 kB the boundaries of an overlapping CNA event and
485 the length of SV overlaps this CNA event by > 75%.

486 SV events from these three callers were then combined by taking the union of the filtered events
487 from. Intersecting events between 2 or more call-sets were determined if both breakpoints of one
488 event were located within 5 kB from both breakpoints of the event detected by the other tool. Then
489 the details of this event were retained based the priority ordered by SvABA, GROC-SVS, Long
490 Ranger. Long Ranger SV events were further filtered out if they were not intersecting events
491 detected by at least one other tool. SV events with span less than 1 kB were excluded from
492 downstream analyses.

493 An SV panel of normals (PoN) was generated using germline events from SvABA and Long
494 Ranger calls. There are two components to this panel: (1) frequency of germline events at exact
495 breakpoint locations (SVpon.bkpt) and (2) frequency of germline event breakpoint overlapping
496 within tiled windows of 1 kB (SVpon.blackListBins). The PoN was used to filter events in the

497 combined SV call-set when an SV has at least one breakpoint with SVpon.bkpt ≥ 2 and
498 overlapping bin with SVpon.blackListBins ≥ 100.

499 The workflow for SV analysis from linked-read sequencing data can be accessed at
500 https://github.com/GavinHaLab/SV_10X_analysis. Manual curation of filtered SV events in the *AR*
501 locus was performed and rescued events were labeled "Manual". The final SV call-set is included
502 in **Table S1K**.

503 For samples based on short-read WGS, SvABA was used in tumor-normal paired mode for SV
504 detection with default parameters. Intra-chromosomal SV events with span > 1 kB were retained.
505 The SvABA workflow can be accessed at https://github.com/GavinHaLab/TitanCNA_SV_WGS

506 *Classification of structural variants in mCRPC*

507 SV types were annotated based on orientations of breakpoints and bin-level copy-number around
508 breakpoints. The orientation of one breakpoint was defined based on the fragment of DNA
509 molecule being connected to the altered molecule. If the connected fragment was to the 5'-end of
510 the breakpoint, *i.e.*, "upstream" or "left" to the breakpoint, then the orientation was annotated as
511 forward or "+"; on the contrary, if the connected fragment was located to the 3'-end of the
512 breakpoint, the orientation was annotated as reverse or "-". The copy-number near each
513 breakpoint was evaluated using 10 kB bins. For one SV event, copy-number values of the bins
514 located to the upstream and downstream of breakpoint 1 were denoted as c_1^{up} and c_1^{down},
515 respectively; similarly, the copy-number values for breakpoint 2 were denoted as c_2^{up} and c_2^{down}.
516 In addition, then mean copy-number c^{mean} of the 10 kB bins between the two breakpoints of one
517 SV event and the number of bins s were also considered during SV classification. Intra-
518 chromosomal SV events, *i.e.*, both breakpoints were located on the same chromosome, were
519 classified to the list of SV types below following the corresponding classification criteria.

520 • Deletion. Events having the orientation combination (reverse, forward) and length
521 between 10 kB and 1 Mb were classified as deletions. The copy-number values of
522 breakpoints should satisfy $c_1^{up} > c_1^{down}$ or $c_2^{up} < c_2^{down}$, and $c_1^{up} > c^{mean}$ or $c_2^{down} > c^{mean}$, and
523 $s \leq 5$. In addition, events overlapping copy-number deletion or LOH segments were also
524 considered as deletions.
525 • Tandem duplication. Events having the orientation combination (forward, reverse) and
526 length between 10 kB and 1 Mb were classified as tandem duplications. The copy-number
527 values of breakpoints should satisfy $c_1^{up} < c_1^{down}$ or $c_2^{up} > c_2^{down}$, and $c_1^{up} < c^{mean}$ or $c_2^{down} <$
528 c^{mean}, and $s \leq 5$. In addition, events overlapping copy-number gain or copy neutral LOH
529 segments were also considered as tandem duplications.
530 • Inversion. Events having the orientation combination (forward, forward) or (reverse,
531 reverse) and length between 10 kB and 5 Mb were classified as inversions. Furthermore,
532 inversion events shorter than 30 kB with unequal copy-numbers around either breakpoint
533 were classified as fold-back inversions.
534 • Balanced rearrangement (balanced). Events having the orientation combination same to
535 inversion (forward, forward) or (reverse, reverse), but length larger than 5 Mb were
536 classified as balanced events. The copy-number values of breakpoints should satisfy c_1^{up}
537 $= c_1^{down}$ and $c_2^{up} = c_2^{down}$, or $c_1^{up} = c^{mean}$ and $c_2^{down} = c^{mean}$.
538 • Unbalanced rearrangement (unbalanced). Intra-chromosomal events which did not fulfill
539 any of the above criteria and having length larger than 10 kB were classified as
540 unbalanced events.
541 • All SV events with two breakpoints located on different chromosomes were classified as
542 translocations.

543 *ICGC/TCGA PCAWG localized prostate cancer structural variants*

544 We obtained localized prostate cancer structural variant calls from ICGC Data Portal release 28
545 (https://dcc.icgc.org/releases/PCAWG/consensus_sv). In this consensus SV file, each SV event
546 was predicted by at least two variant callers. Samples that were classified as prostate
547 adenocarcinoma (PRAD) and early onset prostate cancer (EOPC) were selected. A total of 278
548 samples successfully lifted over to genome build GRCh38. To maximize consistency with mCRPC
549 datasets, we used only the PCAWG consensus SVs that included "SNOWMAN" as one of the
550 tools. Note that "SNOWMAN" was the previous name for SvABA. Intrachromosomal SV events
551 shorter than 10 kB were excluded.

552 We obtained ICGC early-onset PC (EOPC) and late-onset (LOPC) structural variant calls from
553 Gerhauser et al, 2018 (18) (https://data.mendeley.com/datasets/6gtrrxrn2c/1). This dataset
554 includes a total of 253 samples, consisting of 206 EOPC and 47 LOPC. We used the SV calls
555 generated in the original study using DELLY2 (Rausch et al., 2012) PCAWG analysis workflow
556 (https://github.com/ICGC-TCGA-PanCancer/pcawg_delly_workflow). SV coordinates were lifted
557 over to GRCh38.

558 **Tandem duplicator phenotype**

559 For all samples in the combined cohort, the TDP status was predicted using copy-number and
560 SV by counting the number of copy-number segments overlapping with tandem duplication SV
561 events, *i.e.,* gain segments. A sample was considered as TDP if it has more than 300, or 90 gain
562 segments for samples based on linked-read sequencing and short-read sequencing, respectively.
563 The number of segments with gain and median length SV are reported in **Table S1L**.

564 **Chromothripsis analysis**

565 Chromothripsis events were detected by ShatterSeek R package (77). Structural variants calls by
566 SvABA and copy-number calls by TitanCNA were used as input data (excluding Y chromosome).

567 In the input, consecutive segments were joined as one if they had the same copy-number value
568 and centromere regions were filtered out.

569 Manual inspection was performed for reported chromothripsis-like events after adapting criteria
570 thresholds. For samples based on short-read sequencing, confidence classification criteria were
571 refined from the ShatterSeek documentation. Following criteria were used for high confidence
572 calls: total number of intra-chromosomal structural variants events involved in the event ≥ 10; max
573 number of oscillating CN segments (two states) ≥ 10; satisfying either the chromosomal
574 enrichment or the exponential distribution of breakpoints test ($p \leq 0.05$). For samples based on
575 linked-read sequencing, we filtered these calls based on a weighted score that is primarily
576 determined by the number of SVs in a cluster, with less weight given to CN oscillations. In this
577 analysis, events with a score over 0.8 were considered as high confidence and all other events
578 were excluded. The score is defined based on the following terms (ranges from 0 to 1).

579 - Weight 0.6 if total number of intra-chromosomal structural variants events involved in the
580 event ≥ 10.
581 - Weight 0.2 for max number of oscillating CN segments (two states) ≥ 7 or max number of
582 oscillating CN segments (three states) ≥ 14.
583 - Weight 0.1 for passing chromosomal enrichment test by ShatterSeek.
584 - Weight 0.1 for passing exponential distribution of breakpoints test.

585 **Chromoplexy analysis**

586 ChainFinder was used to detect chromoplexy events (8). Ten samples that were considered as
587 TDP (01115374-TA2, 01115202-TC2, 01115248-TA3, 01115503-TC2, 01115257-TA4,
588 01115284-TA9, 01115414-TA1, DTB-063-BL, DTB-183-BL, DTB-214-BL) were excluded from
589 this analysis. In addition, four samples that were found to cause numeric instabilities of
590 ChainFinder were also excluded (DTB-023-BL, DTB-102-PRO, DTB-111-PRO, DTB-151-BL).

591 The SV calls of remaining samples were further filtered to exclude those that were located within
592 5 Mb from chromosomal ends or overlapping chromothripsis regions. For copy-number input,
593 segments that were determined as copy neutral by TitanCNA were set to have log copy-ratio of
594 0. Copy-ratio of the other segments were computed from copy-number values generated by
595 TitanCNA divided by 2 for autosomes or 1 for X chromosome. Log copy-ratio values less than -
596 1.5 were set to -1.5. The output of ChainFinder was used for determining chromoplexy status of
597 individual samples. A chromoplexy event was defined as a chain including at least 5
598 rearrangement events and involving more than 2 different chromosomes. Samples having at least
599 2 such events were considered positive for chromoplexy status.

600 **ChIP-seq data analysis**

601 ChIP-seq data used in this study were downloaded from Gene Expression Omnibus (GEO) (78,
602 79) and the Sequence Read Archive (SRA) (1). Short reads were mapped to the human genome
603 GRCh38 (hg38) using bwa (80). Because read lengths were less than 50bp, the bwa aln
604 command with default parameters was used for mapping. MACS2 (81) was used to identify peaks
605 from mapped ChIP-seq data. For histone modification marks, MACS2 callpeak command was
606 applied with --nomodel --broad –extsize 146. For CTCF data, MACS2 callpeak command was
607 used with --nomodel --extsize 200. Below is the list of ChIP-seq datasets involved in this analysis.

608 - H3K4me3, H3K27me3 and CTCF (GSE38685) (82).
609 - H3K36me3 and H3K9me3 (GSE98732) (83).
610 - H3K4me1 and H3K27ac (GSE73785) (84).

611 For *AR* binding site (ARBS), the peak files were downloaded from two different datasets and
612 converted to hg38 coordinates. For primary prostate cancer, ARBS data were downloaded from
613 GSE70079 (Pomerantz et al., 2015). The union of all tumor sample peaks was used. For mCRPC,
614 met-specfic ARBS data were obtained from a previous study (46).

615 ## Identification of SRB regions

616 ### *Masking the human genome based on mappability*

617 The human genome was divided into 100 kB non-overlapping bins for detection of significantly
618 recurrent breakpoint regions (SRB). A low-mappability mask was generated for the hg38 genome
619 to screen out out regions that are difficult for variant calling based on short-read sequencing. We
620 adopted procedures from a previous study (86) to construct a mask corresponding to regions with
621 low mappability in the human genome. Below is a list of masked regions included in the low-
622 mappability mask.

623 - Composition mask. This set of masked regions includes regions with low sequence
624 complexity detected by mdust, regions with long homopolymers detected by seqtk,
625 satellite regions annotated by RepeatMasker (87), and low complexity regions annotated
626 by RepeatMasker.
627 - Mappability mask. This mask was based on mappability of *k*-mers in the human genome
628 hg38. The value *k* was set to 75 which is half of the read length of WGS data in this study.
629 Each base in the genome was assigned a mappability level, based on the mapping
630 ambiguity of all 75-mers overlapping this specific base. See below for the list of mappability
631 levels. Regions with mappability level 0 and 1 were included in the low-mappability mask.
632 - Level 0: all 75-mers overlapping this base could not be mapped to the genome
633 uniquely.
634 - Level 1: more than 50% of overlapping 75-mers are not uniquely mapped.
635 - Level 2: more than 50% of overlapping 75-mers are uniquely mapped with 1-
636 mismatch hits.
637 - Level 3: more than 50% of overlapping 75-mers are uniquely mapped without 1-
638 mismatch hits.

639 In addition, we used GATK CallableLoci to mark regions with high confidence of variant detection
640 based on coverage. Together, the intersection of unmasked regions and callable loci were defined
641 as the eligible territories for SRB detection. The 100 kB bins with less than 75% overlap with
642 eligible territories were excluded from the analysis.

643 *Generating covariates for regression analysis*

644 To accurately model the genomic features of mCRPC, we incorporated the following covariates.

645 • Nucleotide composition, including GC content, CpG fraction and TpC fraction per 10 kB
646 non-overlapping bin in the genome.
647 • Replication timing of LNCaP (data obtained from ENCODE under accession
648 ENCFF995YGM, lifted over from hg19 to hg38) (88, 89).
649 • DNase I hypersensitive sites (data obtained from ENCODE under accession
650 ENCFF434GSJ, lifted over to hg38).
651 • Repeats annotated by RepeatMasker, including LINE, SINE, LTR, DNA transposon and
652 simple repeats.
653 • Heterochromatin regions inferred by ChromHMM (90) with the 18-state model parameters
654 from the Roadmap Epigenomics Project (91), based LNCaP ChIP-seq data of H3K4me1,
655 H3K4me3, H3K4ac H3K27me3, H3K36me3 and H3K9me3.
656 • Common fragile sites downloaded from HGNC biomart (92).

657 *SRB detection*

658 Structural variants from the final call set were used for statistical enrichment of recurrent
659 breakpoints within 100 kB bins using a Gamma-Poisson regression implemented in the package,
660 fish.hook (44). Breakpoints of SVs were treated independently. The Benjamini-Hochberg
661 procedure was used for multiple testing correction and bins with q-value ≤ 0.1 were determined
662 to be significant. The distances of individual known driver genes to those significant bins were

663 evaluated based on the shortest genomic distance between the gene and bin boundaries,
664 regardless of gene orientations.

665 **Annotation of gene alteration status**

666 *Gene alteration by copy-number*

667 Copy-number segments were excluded if their cellular fraction was lower than 0.8, except for
668 those which were determined as copy neutral or copy-number greater than 4. The gene
669 annotation was based on known protein coding genes from GenCode release 30 (GRCh38.p12)
670 (93). For each gene, its copy-number was assigned to the copy-number value and LOH status of
671 the segment that has the largest overlap with it. The gene-level copy-number was normalized
672 based on ploidy of the corresponding sample, with autosomal genes normalized by the inferred
673 ploidy rounded to nearest integer, and X-linked genes normalized by half such value. Then the
674 copy-number status of each gene was categorized based on the following criteria.

675 - Amplification. Normalized gene-level copy-number is greater than or equal to 2.5.
676 - Gain. Normalized gene-level copy-number is between 2 and 2.5.
677 - Homozygous deletion. Normalized gene-level copy-number is 0.
678 - Deletion with LOH. Normalized gene-level copy-number is between 0 and 1, and LOH
679 status was found.
680 - Copy neutral LOH. Normalized gene-level copy-number is 1 and LOH status was found.

681 *Gene alteration by structural variant*

682 Gene coordinates were based on ENSEMBL v33 of hg38 (94). Gene body region of one gene
683 was defined as the widest region of all known isoforms collapsed. Gene flanking region was
684 defined as the corresponding two 1 Mb regions next to the gene body region on 5'-end and 3'-
685 end, respectively.

686 Gene alteration status by genome rearrangements was defined based on the breakpoints and
687 directions of involving structural variant events. A gene in one WGS sample (gene-sample pair)
688 was considered having gene transecting events if any breakpoints of SV events were located
689 within the gene body region. If the gene transecting status did not apply, then this gene-sample
690 pair was examined for gene flanking status if the breakpoints of any intra-chromosomal SV events,
691 including tandem duplications, deletions, and inversions, were located within the gene flanking
692 regions. Additionally, translocation events including intra-chromosomal balanced and unbalanced
693 events which spanned over 10 Mb, and inter-chromosomal translocation events were considered
694 altering the gene flanking regions if any of their breakpoints was in the gene flanking region, and
695 the direction of the SV was going towards the gene body region. The alteration status of
696 rearrangements for each gene-sample pair was exclusive between gene transecting and gene
697 flanking, with the former being prioritized in report.

698 *AR alteration analysis*

699 Copy-number of the *AR* gene (chrX:67,544,623-67,730,619) and the *AR* enhancer region
700 (chrX:66,895,000-66,910,000) were each computed as the mean corrected total copy-number
701 across the 10 kB bins overlapping each region. The copy-number was further normalized by
702 sample ploidy as previously described. Amplification status of *AR* was determined by comparing
703 the log2 fold-change *FC* of enhancer-level over gene-level copy-number. Four distinct groups
704 were defined based on copy-number and *FC* as below.

705 - Co-amplification. Ploidy normalized copy-number values of both *AR* gene body and
706 enhancer are greater than 1.5.
707 - Selective *AR* amplification. $FC < -\log2(1.5)$ and enhancer copy-number is less than 1.5.
708 - Selective enhancer copy gain. $FC > \log2(1.5)$ and *AR* gene body copy-number is less than
709 1.5.

710 • Lack of amplification for both. All other cases were considered as no amplification for both
711 regions.

712 ANCOVA test was used to test if different patterns of *AR* amplification have an impact on *AR*
713 expression. See Statistics section.

714 **Gene expression**

715 TPM values for a subset of the samples based on linked-read sequencing were obtained from
716 cBioportal (95, 96). For samples based on short-read sequencing the TPM values were obtained
717 from a previous study (13).

718 **Gene fusion analysis**

719 Fusion status of the main members of the ETS family, including *ERG*, *ETV1*, *ETV4*, *ETV5* and
720 *ELK4* was analyzed. Determination of gene fusion status was based on both DNA and RNA levels.
721 For DNA, structural variants transecting gene body regions were used. SV events were
722 considered supporting gene fusion only if they satisfy the following criteria: (1) the breakpoints of
723 this event must be located within the ETS gene and another protein coding gene, respectively; (2)
724 the orientation of the breakpoint located within the ETS gene must be pointing towards the coding
725 sequence of ETS domain. For RNA, arriba was used to detect fusion transcripts from RNA-seq
726 data (97). The fusion status was only confirmed if all following conditions were satisfied: (1) the
727 complete ETS domain was included in the fusion product; (2) detection confidence reported by
728 arriba is “high”; (3) coding sequence in the fusion transcript was in sense orientation and no out-
729 of-frame shifts.

730 **SV signature analysis**

731 *Signature extraction and clustering*

732 *De novo* signature extraction was performed on all SV events called by SvABA of the combined
733 cohort using signature.tools.lib (55) with the recommended settings of 20 bootstraps, 200 repeats,
734 the clustering with matching algorithm, the KLD objective function, and RTOL = 0.001. The
735 exposure of one signature in one sample is defined as the median activity of the signature within
736 the sample across all bootstraps. For clustering, the reference signature exposure values for each
737 sample based on short-read sequencing were normalized such that the sum of exposure values
738 per sample is 1, and the normalized exposure values for each signature were mean-centered
739 across all samples. A Euclidean distance matrix was computed and then samples were clustered
740 with the Ward.D2 algorithm using R's hclust function. We chose the number of clusters to be k =
741 9 based on dendrogram using cutree function in R.

742 **STATISTICS**

743 *Association of AR locus amplification status and AR expression*

744 ANCOVA test was used to test if different patterns of *AR* amplification have an impact on *AR*
745 expression. Batch corrected log10(TPM+1) values using ComBat from sva R package (v3.34.0)
746 were used for *AR* expression level. We fit the ANCOVA model using *AR* expression as the
747 response variable, *AR* amplification status as the predictor variable, and ploidy, purity as
748 covariates. The function Anova in the car package (v3.0-5) was used with Type III sum of squares
749 for the model. Post hoc analysis was performed to determine the specific differences among four
750 different *AR* amplification status. The function glht was used within the multcomp package (v1.4-
751 11) in R to perform Tukey's Test for multiple comparisons.

752

753 *Enrichment of alterations in SV clusters*

754 All 9 identified SV clusters were analyzed for enrichment of alterations. To make the analysis
755 unbiased by SV signature, we limited our search to alteration types that were orthogonal to
756 rearrangements, which include SNV, copy-number gain and copy-number loss. We performed
757 hypothesis testing on each driver-alteration pair, and also on chromoplexy and chromothripsis.
758 For each SV cluster, a χ^2 test was performed for each driver gene alteration status, with samples
759 within group being tested against samples belonging to all 8 other SV clusters. Multiple testing
760 adjustment based on Benjamini-Hochberg FDR was performed to compute q-values. Alteration
761 categories with q-values less than 0.25 were determined as enriched in the corresponding SV
762 cluster.

763

764 *Survival analysis*

765 Survival data was obtained from (98). Survival analyses were conducted using the Kaplan-Meier
766 method with log-rank testing for significance. The function survfit from survival R package was
767 used to perform the analysis.

768

769 **STUDY APPROVAL**

770 For tumor biopsies profiled via linked-read sequencing, samples were collected from individuals
771 with mCRPC who provided informed consent on institutional IRB-reviewed protocols, as
772 previously described (15). Uniformly reanalyzed data were generated as described in the
773 respective studies (9, 13, 18).

774

775 ## AUTHOR CONTRIBUTIONS

776 **Conceptualization:** M-E.T, M.M., S.R.V., G.H.

777 **Methodology:** M.Z., M.K., S.R.V., G.H.

778 **Software:** M.Z., M.K., A.C.H., G.H.

779 **Formal Analysis:** M.Z., M.K., A.C.H., K.L., Y.L., M.R., W.H., J.C-Z, S.R.V., G.H.

780 **Data Curation:** M.Z., M.K., A.C.H., Z.Z., S.R.V., G.H.

781 **Writing – Original Draft:** M.Z., M.M., G.H., S.R.V.

782 **Writing – Review & Editing:** M.Z., R.B., E.M.V., A.D.C. P.S.N., M.L.F., M-E.T., M.M., G.H., S.R.V.

783 **Visualization:** M.Z., M.K., A.C.H., S.R.V., G.H.

784 **Supervision:** M-E.T., M.M., S.R.V., G.H.

785 **Funding Acquisition:** S.R.V., G.H., M.M.

786

787 ## CONFLICTS OF INTERESTS

788 A.D.C.: Honoraria: OncLive, Bayer, Targeted Oncology, Aptitude Health, Journal of Clinical
789 Pathways, Cancer Network; Consulting: Blackstone; Advisory Board: Clovis, Dendreon, Bayer,
790 Eli Lilly, AstraZeneca, Astellas, Blue Earth; Research Funding: Bayer

791 E.M.V.: Advisory/Consulting: Tango Therapeutics, Genome Medical, Invitae, Enara Bio, Janssen,
792 Manifold Bio, Monte Rosa; Research support: Novartis, BMS; Equity: Tango Therapeutics,
793 Genome Medical, Syapse, Enara Bio, Manifold Bio, Microsoft, Monte Rosa; Travel reimbursement:
794 Roche/Genentech; Patents: Institutional patents filed on chromatin mutations and immunotherapy

795 response, and methods for clinical interpretation; intermittent legal consulting on patents for
796 Foaley & Hoag

797 M-E.T.: Advisory boards: Janssen, Pfizer, Astra Zeneca, Bayer

798 M.L.F.: Served as a consultant to and has equity in Nuscan Diagnostics. This activity is outside
799 of the scope of this manuscript.

800 M.M.: Consultant for Bayer, Interline and Isobl; an inventor of patents licensed to LabCorp and
801 Bayer; and receives research funding from Bayer, Janssen, and Ono Pharmaceuticals.

802 P.S.N.: Served as a consultant to Bristol Myers Squibb, Janssen, and Pfizer in work unrelated to
803 the present study.

804 S.R.V.: Consulting (current or previous 3 years), MPM Capital and Vida Ventures; spouse is an
805 employee of and holds equity in Kojin Therapeutics.

806 All other authors declare no competing interests.

807

808 ACKNOWLEDGEMENTS

809 We thank the many patients and their families for their generosity in contributing to this study. We
810 also thank the Prostate Cancer Foundation (PCF) and Stand Up 2 Cancer (SU2C) International
811 Prostate Cancer Dream Team for contributions to specimen acquisition.

812 This work was supported by the National Institutes of Health (K22 CA237746 to G.H.; P01
813 CA163227 and R01 CA234715 to P.S.N.; R01 GM107427, R01 CA193910, and R01 CA251555
814 to M.L.F.; R35 CA197568 to M.M.), Department of Defense Prostate Cancer Research Program
815 (Physician Research Award W81XWH-17-1-0358 to S.R.V.; W81XWH-19-1-0565 and W81XWH-
816 21-1-0234 to M.L.F.; PC200262 to P.S.N.), PCF Young Investigator Awards (G.H. and S.R.V.),
817 PCF-Movember Challenge Award (to E.M.V.), Brotman Baty Institute for Precision Medicine (to

818 G.H.), the Fund for Innovation in Cancer Informatics Major Grant (to G.H.), the V Foundation
819 Scholar Grant (to G.H.), Wong Family Award in Translational Oncology and Dana-Farber Cancer
820 Institute Medical Oncology grant (to A.D.C.), H.L. Snyder Medical Research Foundation and the
821 Cutler Family Fund for Prevention and Early Detection (to M.L.F.), and the Pan-Mass Challenge
822 team IMAGINE (to M-E.T.), American Cancer Society Research Professor (M.M.).

823 This research was also supported in part by the NIH/NCI Cancer Center Support Grant P30
824 CA015704, Pacific Northwest Prostate Cancer SPORE (P50 CA097186), and Scientific
825 Computing Infrastructure (ORIP Grant S10OD028685).

826

827 ## REFERENCES

828 1. Leinonen R, et al. The sequence read archive. *Nucleic Acids Res* 2011;39(Database
829 issue):D19-21.

830 2. de Bono JS, et al. Abiraterone and Increased Survival in Metastatic Prostate Cancer. *N Engl*
831 *J Med* 2011;364(21):1995–2005.

832 3. Scher HI, et al. Increased Survival with Enzalutamide in Prostate Cancer after
833 Chemotherapy. *New England Journal of Medicine* 2012;367(13):1187–1197.

834 4. Abida W, et al. Analysis of the Prevalence of Microsatellite Instability in Prostate Cancer and
835 Response to Immune Checkpoint Blockade. *JAMA Oncol* 2019;5(4):471–478.

836 5. Abida W, et al. Rucaparib in Men With Metastatic Castration-Resistant Prostate Cancer
837 Harboring a BRCA1 or BRCA2 Gene Alteration. *J Clin Oncol* 2020;38(32):3763–3772.

838 6. de Bono J, et al. Olaparib for Metastatic Castration-Resistant Prostate Cancer. *N Engl J Med*
839 2020;382(22):2091–2102.

840 7. Pritchard CC, et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate
841 Cancer. *New England Journal of Medicine* 2016;375(5):443–453.

842 8. Baca SC, et al. Punctuated evolution of prostate cancer genomes. *Cell* 2013;153(3):666–
843 677.

844 9. Campbell PJ, et al. Pan-cancer analysis of whole genomes. *Nature* 2020;578(7793):82–93.

845 10. van Dessel LF, et al. The genomic landscape of metastatic castration-resistant prostate
846 cancers reveals multiple distinct genotypes with potential clinical impact. *Nat Commun*
847 2019;10(1):5251.

848 11. Glodzik D, et al. A somatic-mutational process recurrently duplicates germline susceptibility
849 loci and tissue-specific super-enhancers in breast cancers. *Nat Genet* 2017;49(3):341–348.

850 12. Nik-Zainal S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome
851 sequences. *Nature* 2016;534(7605):47–54.

852 13. Quigley DA, et al. Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer.
853 *Cell* 2018;174(3):758-769.e9.

854 14. Stephens PJ, et al. Massive genomic rearrangement acquired in a single catastrophic event
855 during cancer development. *Cell* 2011;144(1):27–40.

856 15. Viswanathan SR, et al. Structural Alterations Driving Castration-Resistant Prostate Cancer
857 Revealed by Linked-Read Genome Sequencing. *Cell* 2018;174(2):433-447.e19.

858 16. Weinhold N, et al. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat*
859 *Genet* 2014;46(11):1160–1165.

860 17. Woodcock DJ, et al. Prostate cancer evolution from multilineage primary to single lineage
861 metastases with implications for liquid biopsy. *Nat Commun* 2020;11(1):5070.

862 18. Gerhauser C, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular
863 Risk Markers and Clinical Trajectories. *Cancer Cell* 2018;34(6):996-1011.e8.

864 19. Espiritu SMG, et al. The Evolutionary Landscape of Localized Prostate Cancers Drives
865 Clinical Aggression. *Cell* 2018;173(4):1003-1013.e15.

866 20. Wedge DC, et al. Sequencing of prostate cancers identifies new cancer genes, routes of
867 progression and drug targets. *Nat Genet* 2018;50(5):682–692.

868 21. Fraser M, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature*
869 2017;541(7637):359–364.

870 22. Gundem G, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*
871 2015;520(7547):353–357.

872 23. Mateo J, et al. Accelerating precision medicine in metastatic prostate cancer. *Nat Cancer*
873 2020;1(11):1041–1053.

874 24. Armenia J, et al. The long tail of oncogenic drivers in prostate cancer. *Nature Genetics*
875 2018;50(5):645–651.

876 25. Grasso CS, et al. The mutational landscape of lethal castration-resistant prostate cancer.
877 *Nature* 2012;487(7406):239–243.

878 26. Chen CD, et al. Molecular determinants of resistance to antiandrogen therapy. *Nature*
879 *Medicine* 2004;10(1):33–39.

880 27. Yuan X, et al. Androgen receptor functions in castration-resistant prostate cancer and
881 mechanisms of resistance to new agents targeting the androgen axis. *Oncogene*
882 2014;33(22):2815–2825.

883 28. Li Y, et al. Diverse AR Gene Rearrangements Mediate Resistance to Androgen Receptor
884 Inhibitors in Metastatic Prostate Cancer. *Clinical Cancer Research* 2020;26(8):1965–1976.

885 29. Brand LJ, Dehm SM. Androgen Receptor Gene Rearrangements: New Perspectives on
886 Prostate Cancer Progression. *Curr Drug Targets* 2013;14(4):441–449.

887 30. Céraline J, et al. Constitutive activation of the androgen receptor by a point mutation in the
888 hinge region: a new mechanism for androgen-independent growth in prostate cancer. *Int. J.*
889 *Cancer* 2004;108(1):152–157.

890 31. Henzler C, et al. Truncation and constitutive activation of the androgen receptor by diverse
891 genomic rearrangements in prostate cancer. *Nat Commun* 2016;7:13668.

892 32. Takeda DY, et al. A Somatically Acquired Enhancer of the Androgen Receptor Is a
893 Noncoding Driver in Advanced Prostate Cancer. *Cell* 2018;174(2):422-432.e13.

894 33. Li Y, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*
895 2020;578(7793):112–121.

896 34. Sailer V, et al. Bone Biopsy Protocol for Advanced Prostate Cancer in the Era of Precision
897 Medicine. *Cancer* 2018;124(5):1008–1015.

898 35. Wu Y-M, et al. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced
899 Prostate Cancer. *Cell* 2018;173(7):1770-1782.e14.

900 36. Robinson D, et al. Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*
901 2015;162(2):454.

902 37. Pitkänen E, et al. Frequent L1 retrotranspositions originating from TTC28 in colorectal
903 cancer. *Oncotarget* 2014;5(3):853–859.

904 38. Pradhan B, et al. Detection of subclonal L1 transductions in colorectal cancer by long-
905 distance inverse-PCR and Nanopore sequencing. *Sci Rep* 2017;7(1):14521.

906 39. Tubio JMC, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA
907 mediated by L1 retrotransposition in cancer genomes. *Science* 2014;345(6196):1251343.

908 40. Chen J, et al. The t(1;3) breakpoint-spanning genes LSAMP and NORE1 are involved in
909 clear cell renal cell carcinomas. *Cancer Cell* 2003;4(5):405–413.

910 41. Kresse SH, et al. LSAMP, a novel candidate tumor suppressor gene in human
911 osteosarcomas, identified by array comparative genomic hybridization. *Genes*
912 *Chromosomes Cancer* 2009;48(8):679–693.

913 42. Kühn MWM, et al. High-resolution genomic profiling of adult and pediatric core-binding factor
914 acute myeloid leukemia reveals new recurrent genomic alterations. *Blood* 2012;119(10):e67.

915 43. Veeriah S, et al. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently
916 inactivated and mutated in glioblastoma and other human cancers. *PNAS*
917 2009;106(23):9435–9440.

918 44. Imielinski M, Guo G, Meyerson M. Insertions and Deletions Target Lineage-Defining Genes
919 in Human Cancers. *Cell* 2017;168(3):460-472.e14.

920 45. Eteleeb AM, et al. SV-HotSpot: detection and visualization of hotspots targeted by structural
921 variants associated with gene expression. *Sci Rep* 2020;10(1):15890.

922 46. Pomerantz MM, et al. Prostate cancer reactivates developmental epigenomic programs
923 during metastatic progression. *Nat. Genet.* [published online ahead of print: July 20, 2020];
924 doi:10.1038/s41588-020-0664-8

925 47. Tomlins SA, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in
926 prostate cancer. *Science* 2005;310(5748):644–648.

927 48. Tomlins SA, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS
928 gene fusions in prostate cancer. *Nature* 2007;448(7153):595–599.

929 49. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM. Landscape of gene fusions in
930 epithelial cancers: seq and ye shall find. *Genome Medicine* 2015;7(1):129.

931 50. Qin F, et al. SLC45A3-ELK4 functions as a long non-coding chimeric RNA. *Cancer Lett*
932 2017;404:53–61.

933 51. Rickman DS, et al. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-
934 specific fusion transcript in prostate cancer. *Cancer Res* 2009;69(7):2734–2738.

935 52. Zhang Y, et al. Chimeric transcript generated by cis-splicing of adjacent genes regulates
936 prostate cancer cell proliferation. *Cancer Discov* 2012;2(7):598–607.

937 53. Visakorpi T, et al. In vivo amplification of the androgen receptor gene and progression of
938 human prostate cancer. *Nat. Genet.* 1995;9(4):401–406.

939 54. Umbreit NT, et al. Mechanisms generating cancer genome complexity from a single cell
940 division error. *Science* 2020;368(6488):eaba0712.

941 55. Degasperi A, et al. A practical framework and online tool for mutational signature analyses
942 show inter-tissue variation and driver dependencies. *Nat Cancer* 2020;1(2):249–263.

943 56. Willis NA, et al. Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature*
944 2017;

945 57. Barbieri CE, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12
946 mutations in prostate cancer. *Nat Genet* 2012;44(6):685–689.

947 58. Nguyen B, et al. Pan-cancer Analysis of CDK12 Alterations Identifies a Subset of Prostate
948 Cancers with Distinct Genomic and Clinical Characteristics. *European Urology*
949 2020;78(5):671–679.

950 59. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate
951 Cancer. *Cell* 2015;163(4):1011–1025.

952 60. Giambartolomei C, et al. H3K27ac HiChIP in prostate cell lines identifies risk genes for
953 prostate cancer susceptibility. *The American Journal of Human Genetics*
954 2021;108(12):2284–2300.

955 61. Taplin ME, et al. Mutation of the androgen-receptor gene in metastatic androgen-
956 independent prostate cancer. *N. Engl. J. Med.* 1995;332(21):1393–1398.

957 62. Haffner MC, et al. Androgen-induced TOP2B-mediated double-strand breaks and prostate
958 cancer gene rearrangements. *Nat Genet* 2010;42(8):668–675.

959 63. Petrovics G, et al. A novel genomic alteration of LSAMP associates with aggressive prostate
960 cancer in African American men. *EBioMedicine* 2015;2(12):1957.

961 64. Alexandrov LB, et al. Signatures of mutational processes in human cancer. *Nature*
962 2013;500(7463):415–421.

963 65. Alexandrov LB, et al. The repertoire of mutational signatures in human cancer. *Nature*
964 2020;578(7793):94–101.

965 66. Macintyre G, et al. Copy number signatures and mutational processes in ovarian carcinoma.
966 *Nat Genet* 2018;50(9):1262–1270.

967 67. Wang S, et al. Copy number signature analysis tool and its application in prostate cancer
968 reveals distinct mutational processes and clinical outcomes. *PLoS Genet*
969 2021;17(5):e1009557.

970 68. Van der Auwera GA, O'Connor B. Genomics in the
971 Cloud2020;https://www.oreilly.com/library/view/genomics-in-the/9781491975183/. cited
972 December 10, 2021

973 69. Carrot-Zhang J, Majewski J. LoLoPicker: detecting low allelic-fraction variants from low-
974 quality cancer samples. *Oncotarget* 2017;8(23):37032.

975 70. Saunders CT, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-
976 normal sample pairs. *Bioinformatics* 2012;28(14):1811–1817.

977 71. Martincorena I, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*
978 2017;171(5):1029-1041.e21.

979 72. Tate JG, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*
980 2019;47(D1):D941–D947.

981 73. Ha G, et al. TITAN: inference of copy number architectures in clonal cell populations from
982 tumor whole-genome sequence data. *Genome Res* 2014;24(11):1881–1893.

983 74. Adalsteinsson VA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high
984 concordance with metastatic tumors. *Nat. Commun.* 2017;8(1):1324.

985 75. Wala JA, et al. SvABA: genome-wide detection of structural variants and indels by local
986 assembly. *Genome Res* 2018;28(4):581–591.

987 76. Spies N, et al. Genome-wide reconstruction of complex structural variants using read clouds.
988 *Nature Methods* 2017;

989 77. Cortés-Ciriano I, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers
990 using whole-genome sequencing. *Nat Genet* 2020;52(3):331–341.

991 78. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids*
992 *Res* 2013;41(Database issue):D991-995.

993 79. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository -
994 PubMedhttps://pubmed.ncbi.nlm.nih.gov/11752295/. cited January 6, 2022

995 80. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
996 *Bioinformatics* 2009;25(14):1754–1760.

997 81. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9(9):R137.

998 82. Bert SA, et al. Regional activation of the cancer genome by long-range epigenetic
999 remodeling. *Cancer Cell* 2013;23(1):9–22.

1000 83. Du Q, et al. Replication timing and epigenome remodelling are associated with the nature of
1001 chromosomal rearrangements in cancer. *Nat Commun* 2019;10(1):416.

1002 84. Taberlay PC, et al. Three-dimensional disorganization of the cancer genome occurs
1003 coincident with long-range genetic and epigenetic alterations. *Genome Res* 2016;26(6):719–
1004 731.

1005 85. Pomerantz MM, et al. The androgen receptor cistrome is extensively reprogrammed in
1006 human prostate tumorigenesis. *Nat Genet* 2015;47(11):1346–1351.

1007 86. Mallick S, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse
1008 populations. *Nature* 2016;538(7624):201–206.

1009 87. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0.*. 2013:

1010 88. Davis CA, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic*
1011 *Acids Res* 2018;46(D1):D794–D801.

1012 89. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human
1013 genome. *Nature* 2012;489(7414):57–74.

1014 90. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization.
1015 *Nat Methods* 2012;9(3):215–216.

1016 91. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human
1017 epigenomes. *Nature* 2015;518(7539):317–330.

1018 92. Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids*
1019 *Res* 2021;49(D1):D939–D946.

1020 93. Frankish A, et al. GENCODE reference annotation for the human and mouse genomes.
1021 *Nucleic Acids Res* 2019;47(D1):D766–D773.

1022 94. Howe KL, et al. Ensembl 2021. *Nucleic Acids Res* 2021;49(D1):D884–D891.

1023 95. Cerami E, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring
1024 Multidimensional Cancer Genomics Data: Figure 1.. *Cancer Discovery* 2012;2(5):401–404.

1025 96. Gao J, et al. Integrative analysis of complex cancer genomics and clinical profiles using the
1026 cBioPortal. *Sci Signal* 2013;6(269):pl1.

1027 97. Uhrig S, et al. Accurate and efficient detection of gene fusions from RNA sequencing data.
1028 *Genome Res* 2021;31(3):448–460.

1029 98. Chen WS, et al. Genomic Drivers of Poor Prognosis and Enzalutamide Resistance in
1030 Metastatic Castration-resistant Prostate Cancer. *Eur Urol* 2019;76(5):562–571.

1031

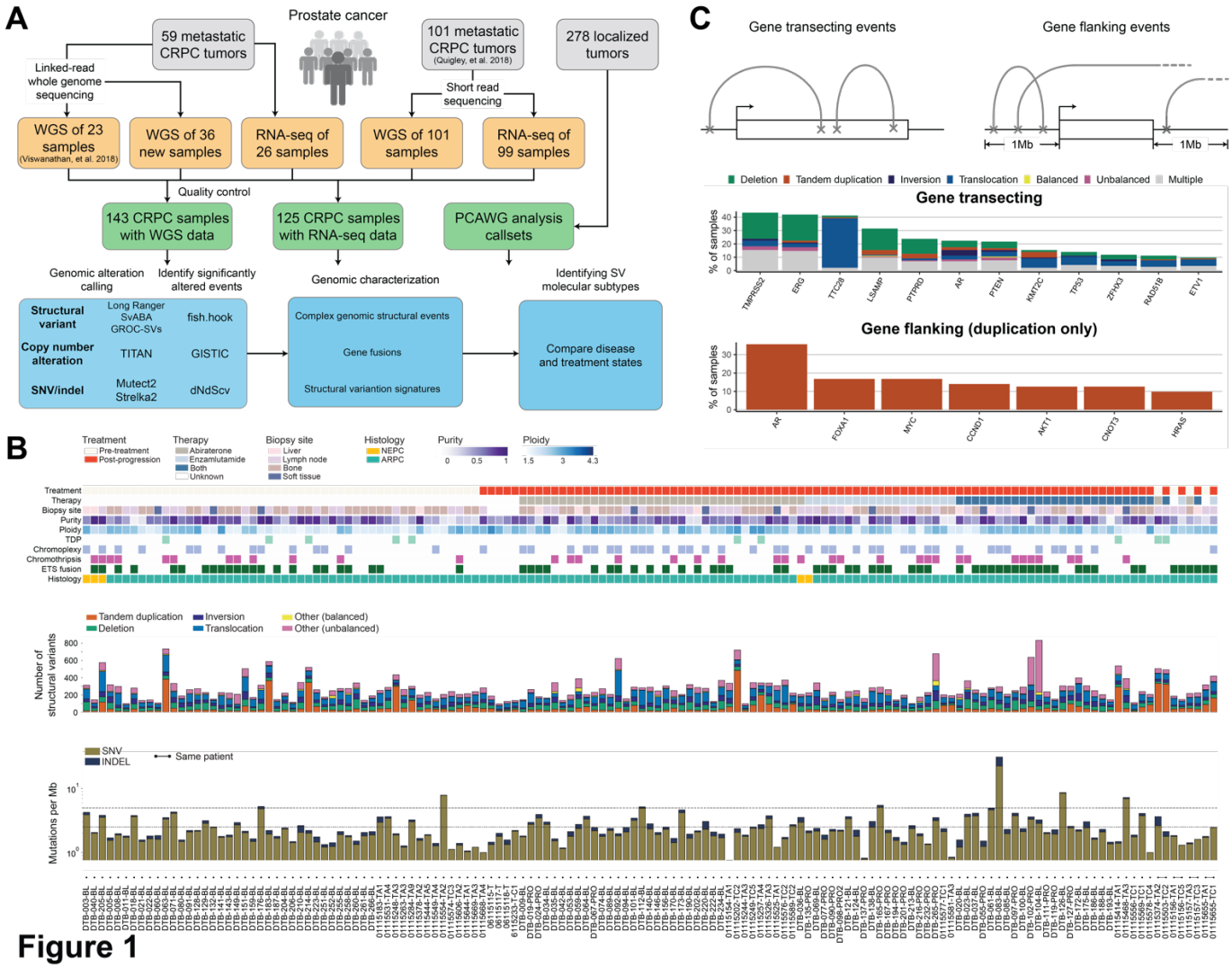

Figure 1

Figure 1. Study design and genomic landscape of mCRPC.

(A) Workflow of study and data analysis. Tumor specimens (grey) from both primary prostate cancer and mCRPC were included in this study. Linked-read and short-read whole-genome sequencing (WGS) and RNA-sequencing datasets were either generated for this study or reanalyzed from prior studies (13, 15). A pooled dataset of 143 mCRPC samples with WGS data was used in this study after curation (**Methods**). Genomic alteration call-sets for 278 primary localized prostate cancer samples were obtained from ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) (9, 33). For 125 mCRPC samples, RNA-seq was used. The overview of the genomic alteration and characterization analysis is shown.

(B) Clinical annotations and somatic alterations for 143 patient samples in the pooled mCRPC cohort. Samples are ordered by treatment type; the four patients with pre-treatment and post-progression pairs are placed at the right. (Top) Clinical and sample information and genomic pattern classifications. (Middle) Distribution of genomic rearrangement types in individual samples. (Bottom) Mutational burden for SNVs and indels computed as number of mutations per mega-

base pair (Mb). Y-axis shown in logarithmic scale. Threshold lines indicates mutational burden at 2.5 and 5 mutations per Mb.

(C) Genomic rearrangement alteration profiles of key mCRPC genes. (Top) Events were categorized into gene transecting and gene flanking events (**Methods**). Gene transecting: if any of its breakpoints was located within the gene body region. Gene flanking: rearrangements which were not gene transecting and had breakpoints located within 1 Mb of either transcription start site or termination site of the gene. Only 159 genes reported and known to be involved in prostate cancer were considered in this analysis (**Table S1G and S1H**). (Middle) Frequency and distribution of rearrangement types for gene transecting events; genes with ≥ 10% frequency are shown. Gene transecting events were prioritized over flanking events during annotation. The category "Multiple" represents gene-sample pairs carrying more than one type of rearrangement event. (Bottom) Frequency of gene flanking events by tandem duplication; genes with ≥ 10% are shown.

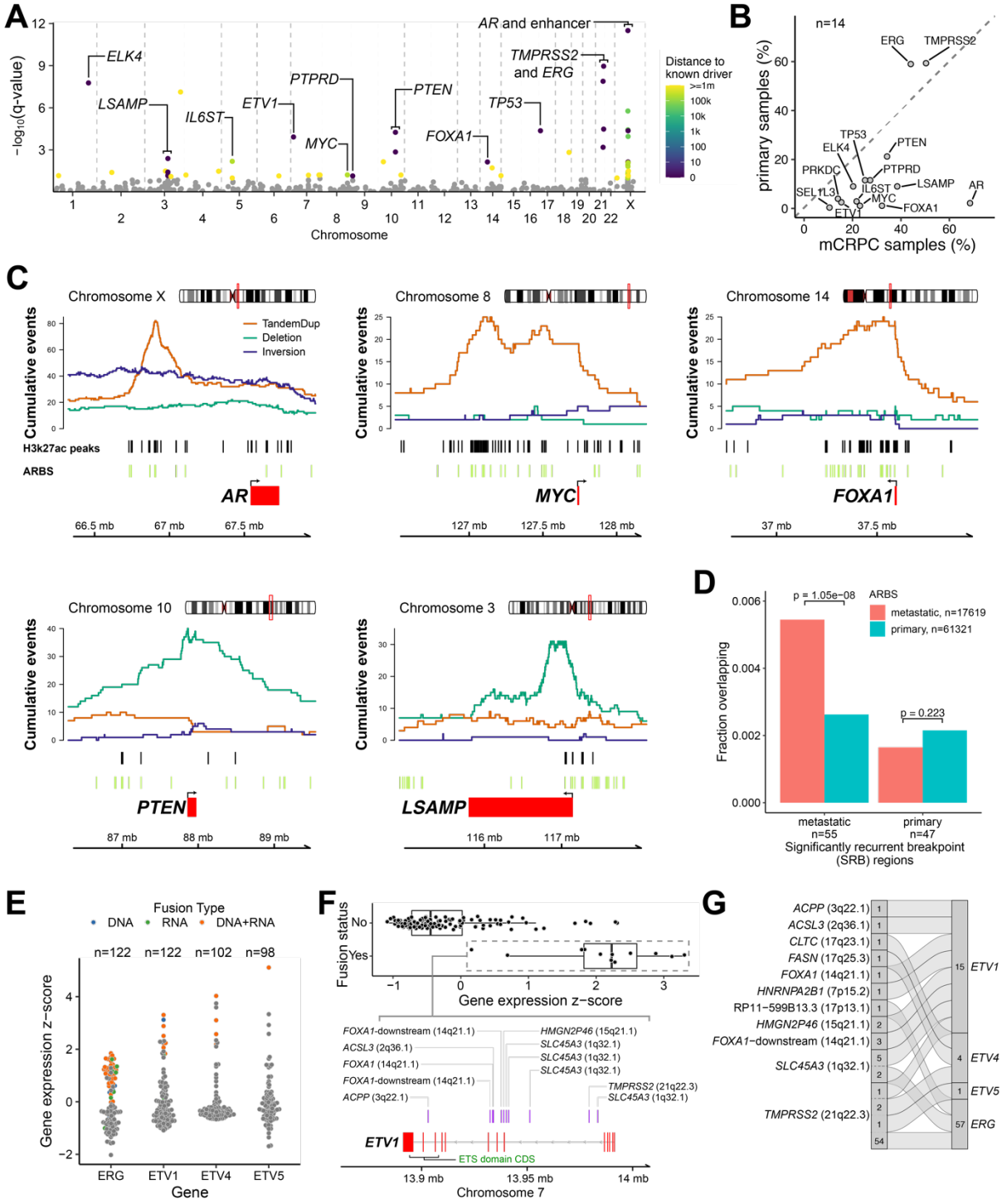
A
−log10(q-value)
ELK4
LSAMP
IL6ST
ETV1
PTPRD
MYC
PTEN
FOXA1
TP53
TMPRSS2 and ERG
AR and enhancer
Distance to known driver
Chromosome
B
n=14
primary samples (%)
mCRPC samples (%)
ERG
TMPRSS2
TP53
ELK4
PTEN
PRKDC
PTPRD
SEL1L3
IL6ST
LSAMP
MYC
AR
ETV1
FOXA1
C
Chromosome X
Chromosome 8
Chromosome 14
Chromosome 10
Chromosome 3
Cumulative events
TandemDup
Deletion
Inversion
H3k27ac peaks
ARBS
AR
MYC
FOXA1
PTEN
LSAMP
D
ARBS
metastatic, n=17619
primary, n=61321
p = 1.05e−08
p = 0.223
Fraction overlapping
metastatic n=55
primary n=47
Significantly recurrent breakpoint (SRB) regions
E
Fusion Type
DNA
RNA
DNA+RNA
n=122
n=122
n=102
n=98
Gene expression z−score
ERG
ETV1
ETV4
ETV5
Gene
F
Fusion status
No
Yes
Gene expression z−score
FOXA1-downstream (14q21.1)
ACSL3 (2q36.1)
FOXA1 (14q21.1)
FOXA1-downstream (14q21.1)
ACPP (3q22.1)
HMGN2P46 (15q21.1)
SLC45A3 (1q32.1)
SLC45A3 (1q32.1)
SLC45A3 (1q32.1)
TMPRSS2 (21q22.3)
SLC45A3 (1q32.1)
ETV1
ETS domain CDS
Chromosome 7
G
ACPP (3q22.1)
ACSL3 (2q36.1)
CLTC (17q23.1)
FASN (17q25.3)
FOXA1 (14q21.1)
HNRNPA2B1 (7p15.2)
RP11−599B13.3 (17p13.1)
HMGN2P46 (15q21.1)
FOXA1−downstream (14q21.1)
SLC45A3 (1q32.1)
TMPRSS2 (21q22.3)
ETV1
ETV4
ETV5
ERG

Figure 2

Figure 2. Genome-wide analysis of genomic rearrangements in mCRPC.

(A) Analysis of significantly recurrent breakpoint (SRB) identified regions of rearrangement hotspots, genome-wide, using a Gamma-Poisson regression model. Each dot corresponds to a 100 kB bin (n=26,663 total bins). Statistically significant SRB bins with FDR (Benjamini-Hochberg) q-value ≤ 0.1 (n=55) are colored based on the distance to the nearest known prostate cancer driver gene, within 1 Mb. The driver genes within 1 Mb of the SRB bins are labeled. A square bracket is used for genes spanning multiple bins. Bins with q-value > 0.1 were not significant (grey).

(B) Comparison of SV alteration frequency in mCRPC versus primary localized prostate cancer. The union set of genes (n=14) within 1 Mb of SRB hotspot regions in mCRPC and localized prostate cancer cohorts was included in the comparison. The frequencies represent total gene transecting and flanking SV events. All labeled genes were significantly enriched in either mCRPC or primary localized tumors (Fisher's test, p-value < 0.05).

(C) Patterns of rearrangements at the loci of driver genes identified at SRB regions in mCRPC cohort of 143 tumors. Cumulative counts of intra-chromosomal SV events (tandem duplications "TandemDup", deletions, and inversions) were computed based on the breakpoints and span of the events. Histone H3 lysine 27 acetylation (H3K27ac) and *AR* binding sites (ARBS) specific to mCRPC were obtained from a previous study (46). Inter-chromosomal translocations are not shown. Genome coordinates based on hg38 build.

(D) Overlap of *AR* binding sites (ARBS) within SRB hotspots of mCRPC (55 regions) and primary localized prostate (47 regions) cohorts. Metastatic-specific and primary localized-specific ARBS were obtained from previous studies (46, 85). χ^2 test of independence p-values are shown.

(E) Fusion status and expression of selected genes in ETS transcription factor gene family in the mCRPC cohort with WGS and RNA-seq data. Fusion type was defined as the data evidence that supported the event: DNA-only, corresponds to WGS; RNA-only, corresponds to RNA-seq; DNA+RNA, corresponds to support from both WGS and RNA-seq. Each dot represents a tumor sample and is colored based on fusion type of each sample; grey indicates no evidence of fusion event. Data shown for samples with available expression data for the specific ETS gene. Gene expression values of full-length transcripts are z-score normalized.

(F) Fusion profile of *ETV1*. DNA rearrangement breakpoints supporting the fusion (purple bars) are indicated with the corresponding fusion partners. Exons of the ETS domain (red) are indicated. Genome coordinates based on hg38 build.

(G) Summary of fusion partners for selected genes in ETS transcription factor gene family in mCRPC cohort. Fusion events and partners are indicated by flow connections. Total counts of individual fusion events and partners across the cohort are shown.

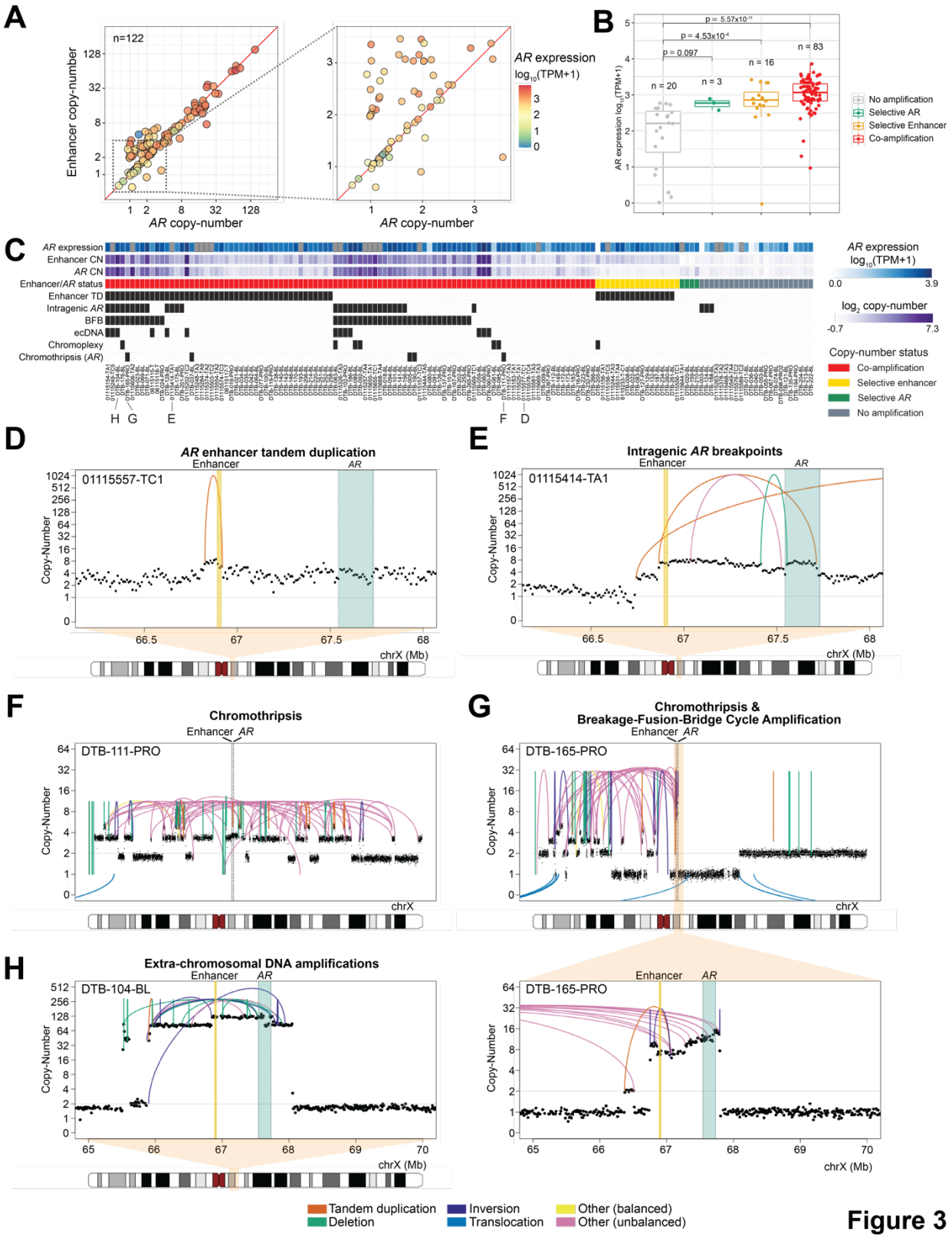
A
n=122
Enhancer copy-number
AR copy-number
AR expression
log10(TPM+1)
B
p = 5.57x10-11
p = 4.53x10-4
p = 0.097
n = 20
n = 3
n = 16
n = 83
AR expression log10(TPM+1)
No amplification
Selective AR
Selective Enhancer
Co-amplification
C
AR expression
Enhancer CN
AR CN
Enhancer/AR status
Enhancer TD
Intragenic AR
BFB
ecDNA
Chromoplexy
Chromothripsis (AR)
AR expression
log10(TPM+1)
0.0
3.9
log2 copy-number
-0.7
7.3
Copy-number status
Co-amplification
Selective enhancer
Selective AR
No amplification
H G E
F D
D
AR enhancer tandem duplication
Enhancer
AR
01115557-TC1
Copy-Number
chrX (Mb)
E
Intragenic AR breakpoints
Enhancer
AR
01115414-TA1
Copy-Number
chrX (Mb)
F
Chromothripsis
Enhancer AR
DTB-111-PRO
Copy-Number
chrX
G
Chromothripsis &
Breakage-Fusion-Bridge Cycle Amplification
Enhancer AR
DTB-165-PRO
Copy-Number
chrX
H
Extra-chromosomal DNA amplifications
Enhancer
AR
DTB-104-BL
Copy-Number
chrX (Mb)
Enhancer
AR
DTB-165-PRO
Copy-Number
chrX (Mb)
Tandem duplication
Deletion
Inversion
Translocation
Other (balanced)
Other (unbalanced)

Figure 3

Figure 3. Modes of *AR* activation in mCRPC.

(A) Copy number of *AR* gene and its enhancer (~624 kB upstream) for mCRPC cohort samples after adjustment by tumor purity and sample ploidy normalization. Data shown for samples with available *AR* gene expression data. (Left) Copy number of *AR* and its enhancer are shown in $\log_2$ scale, colored based on *AR* gene expression level (transcripts per million, TPM). (Right) Excerpt of figure highlighting *AR* expression for samples with lower copy number values.

(B) *AR* expression for *AR* locus copy number status for 122 samples with available *AR* gene expression data. ANCOVA test was performed to account for tumor purity and ploidy as covariates. TukeyHSD p-values for pair-wise comparisons between groups with *AR* locus amplification status and groups with no amplification.

(C) Patterns of rearrangements involving the *AR* locus in 143 mCRPC samples. Presence of specific alteration events and complex rearrangements (black) were predicted automatically and manually curated. *AR* gene expression shown (blue shades) for same samples in (B); samples with no available expression data are indicated in grey. Representative examples of each category are presented in (D) to (H).

(D-H) Complex and simple rearrangement patterns involving the *AR* locus, including focal duplication events on *AR* enhancer **(D)**, intragenic amplification event leading to a breakpoint within *AR* between exons 4 and 5 **(E)**, chromosomal level chromothripsis events involving *AR* and enhancer **(F)**, arm-level chromothripsis coinciding with *AR* amplification by break-fusion-break cycle **(G)**, extra-chromosomal DNA amplicon including *AR* and enhancer **(H)**. *AR* gene boundary (green) and its enhancer (yellow) are shown; concave arcs, intra-chromosomal SV events; convex arcs, inter-chromosomal SV events. Copy number values represent 10 kB bins and have been tumor purity corrected.

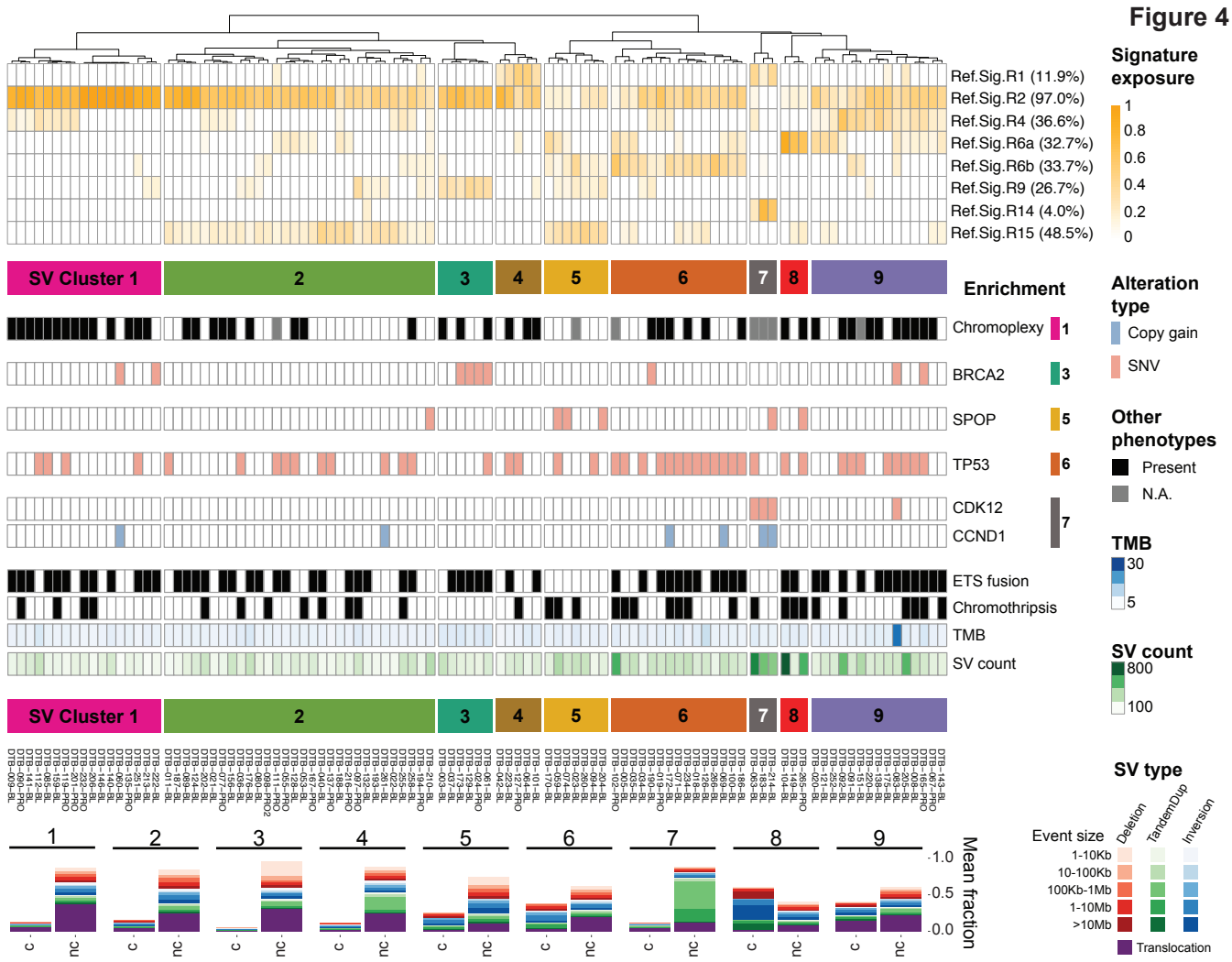

Figure 4. Clustering of mCRPC SV signatures

SV signature analysis and hierarchical clustering identifies nine distinct molecular groups. (Top) Dendrogram of the clustering of SV signature exposure. The prevalence of each signature was computed based on having ≥ 0.05 exposure (proportion of SVs). (Middle) Enrichment of altered prostate cancer drivers. Enriched alterations in Cluster 1, 3, 5, 6, and 7 are shown based on statistical significance by χ^2 test. (Bottom) Composition of SV types and sizes for each SV cluster, separated by non-clustered (nc) and clustered (c) SV events.