TITLE

Patterns of Structural Variation Define Prostate Cancer Across Disease States

AUTHORS/AFFILIATIONS

Meng Zhou[1,2,9], Minjeong Ko[4,9], 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[4,7,10,*]

[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 Research 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

[9] These authors contributed equally

[10] Lead Contact

* Correspondence: mtaplin@partners.org (M-E.T.), matthew_meyerson@dfci.harvard.edu (M.M.), srinivas_viswanathan@dfci.harvard.edu (S.R.V.), gha@fredhutch.org (G.H.)

KEYWORDS

Prostate cancer; castration-resistant prostate cancer; androgen receptor; whole genome sequencing; structural variant; enhancer

SUMMARY

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 reflects the genomic landscape of this disease. We comprehensively 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 in mCRPC. 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 (Leinonen et al., 2011). 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 (de Bono et al., 2011; Scher et al., 2012), poly (ADP-ribose) polymerase (PARP) inhibitors in *BRCA1/2*-altered prostate cancers, and immune checkpoint inhibitors in cancers with microsatellite instability (Abida et al., 2019, 2020; de Bono et al., 2020; Pritchard et al., 2016).

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 (Baca et al., 2013; Campbell et al., 2020; van Dessel et al., 2019; Fraser et al., 2021; Glodzik et al., 2017; Hadi et al., 2020; Nik-Zainal et al., 2016; Quigley et al., 2018; Stephens et al., 2011; Viswanathan et al., 2018; Weinhold et al., 2014). These studies highlight the 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 (Mateo et al., 2020).

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 (Armenia et al., 2018; Grasso et al., 2012). A cardinal hallmark of CRPC is the reactivation of *AR* signaling in the face of maximal ADT (Chen et al., 2004; Yuan et al., 2014). 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 (Brand and Dehm, 2013; Céraline et al., 2004; Henzler et al., 2016) or of regulatory elements distal to the gene body (Quigley et al., 2018; Takeda et al., 2018; Viswanathan et al., 2018). 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.

In this study, we performed linked-read WGS on 36 mCRPC tumor-normal pairs. We combined these data with WGS and whole transcriptome sequencing (RNA-Seq) data from previously described localized and metastatic CRPC cohorts (Campbell et al., 2020; Li et al., 2020; Quigley et al., 2018; Viswanathan et al., 2018). We then established a harmonized workflow for the integrative genomic analysis of 278 localized and 143 metastatic CRPC samples, interrogated both hotspots and genome-wide patterns of structural variation, and evaluated their consequences.

84 # RESULTS

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

86 We performed linked-read whole genome sequencing on 36 biopsy specimens from 33 mCRPC
87 patients and matched blood normal controls. After quality control, 17 samples were excluded
88 based on insufficient tumor purity and contamination (**Methods, Figure 1A, Table S1**). We re-
89 analyzed a linked-read WGS dataset of 23 samples published previously (Viswanathan et al.,
90 2018), resulting in a total of 42 linked-read WGS samples from 38 patients with mean coverage
91 of 34X (range 21X - 54X) and 33X (range 25X - 45X) for tumor and normal samples, respectively
92 (**Table S1A**). The mean molecule length was 29 kB and 34 kB in tumor and normal samples,
93 respectively (**Table S1A**).

94 We further combined these data with 101 mCRPC samples sequenced with standard short-read
95 sequencing, published previously (Quigley et al., 2018). This resulted in the generation of a final
96 combined cohort of 143 tumor-normal pairs (**Figure 1A**). Fifty-four samples (37.8% of 143
97 samples) were collected at castration resistance, prior to receiving treatment of second-
98 generation androgen receptor signaling inhibitor (ARSi) such as abiraterone and/or enzalutamide
99 ("pre-treatment"), while the remaining 89 samples (62.2% of 143 samples) were collected at
100 progression ("post-treatment", **Figure 1B**, **Table S1B**). We analyzed the somatic single nucleotide
101 variant (SNVs), insertion-deletions (indels), copy number alterations (CNAs), and SVs in the
102 combined cohort and identified recurrent somatic alterations in each of these classes (**Figure 1A,**
103 **Methods**).

104 A total of 2,315,452 SNVs and indels were called, with a mean tumor mutation burden (TMB) of
105 2.82 mutations per million bases (Mb). We confirmed that known driver genes of prostate cancer
106 were enriched for non-synonymous mutations, including *TP53*, *RB1*, *PTEN*, *FOXA1*, *CDK12*, *AR*
107 and *SPOP* among known COSMIC Cancer Gene Census genes (dndscv, $q \leq 0.1$, **Table S1C and**
108 **S1D, Methods**). We detected an average of 272 (range 96-833) SV events per sample. Based
109 on breakpoint orientations, SV events were classified into deletions, inversions, tandem
110 duplications, inter-chromosomal translocations, and intra-chromosomal translocations, while
111 intra-chromosomal translocations were further divided into balanced and unbalanced events
112 based on copy number information (**Methods**). Chromoplexy was detected in 53 samples (37.1%
113 of 143 samples) while chromothripsis was detected in 37 samples (25.9%); these events were
114 not mutually exclusive (Fisher's exact test, log-odds=1.417, p-value=0.612). Ten cases (7.0%)
115 harbored a genome-wide tandem duplicator phenotype (TDP), all of which had *CDK12*
116 inactivating alterations, as recently reported (Viswanathan et al., 2018; Wu et al., 2018). We found
117 that TDP was mutually exclusive with ETS rearrangements (Fisher's exact test, log-odds
118 ratio=0.133, p=0.043) and chromothripsis (log-odds ratio=0.301, p-value=0.007), as previously
119 reported (van Dessel et al., 2019; Quigley et al., 2018; Viswanathan et al., 2018; Wu et al., 2018).

120 Analysis of CNA events across the genome revealed amplification and deletion peaks in the
121 regions of known prostate cancer genes (Armenia et al., 2018; van Dessel et al., 2019; Quigley
122 et al., 2018; Viswanathan et al., 2018). Many oncogenic drivers of mCRPC, such as *AR* and *MYC*,
123 are within peaks of amplification across the cohort, while tumor suppressors such as *PTEN, TP53,*
124 and *KMT2C* were found within deletion peaks (**Figure S1C**, **Table S1E and S1F**).

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

126 Structural variants may either activate or inactivate gene function, depending on the location of
127 the breakpoints and the specific class of SV. We analyzed the impact of SVs across our cohort,
128 distinguishing between those with predicted inactivating ("gene transecting events") and
129 activating ("gene flanking events") effects (**Figure 1C, Figure S1C, Table S1G and S1H**).

Frequent gene transecting alterations were observed at the *TTC28* (37.1% of 143 samples), *LSAMP* (31.5%), and *PTPRD* (23.8%) loci, which have not been extensively studied in prostate cancer. Rearrangements involving *TTC28* were predominantly inter-chromosomal translocations between the gene body and various non-recurrent partner loci (**Figure S2B**). This likely represents retrotransposon activity, given that the *TTC28* locus harbors an active L1 retrotransposon element (Pitkänen et al., 2014; Pradhan et al., 2017; Tubio et al., 2014). Transecting SVs within the *LSAMP* and *PTPRD* genes were predominantly deletions. Both of these genes are sites of deletion/rearrangement in cancer and have been reported to function as tumor suppressors, though they have not been extensively studied within the context of prostate cancer (Chen et al., 2003; Kresse et al., 2009; Kühn et al., 2012; Veeriah et al., 2009) (**Figure 1C**). Of note, although gene transecting events would be predicted to disrupt individual genes, the most frequent transecting events identified via this analysis were deletion events that span the adjacent *TMPRSS2* and *ERG* genes (observed in 37.8%), which actually produces an activating*TMPRSS2-ERG* fusion.

Duplication events that flank an intact gene could activate oncogenes, either by resulting in copy number gain of the gene or by duplicating non-coding regulatory regions (Quigley et al., 2018; Viswanathan et al., 2018). Indeed, we observed recurrent tandem duplication events with breakpoints located in the flanking gene regions of several known prostate cancer oncogenes, including *AR* (35.7%), *FOXA1* (16.8%), *MYC* (16.8%), and *CCND1* (14.0%) (**Figure 1C**).

Certain prostate cancer driver genes were altered by multiple classes of structural alterations in both the gene body and flanking regions (e.g., *AR*, *PTEN*), while others were predominantly altered by a single alteration class (e.g., SNVs for *TP53*, intragenic translocations for *TTC28*, or flanking tandem duplications for *MYC*) (**Figure 1C, Figure S1C**). Collectively, these results demonstrate that prostate cancer is associated with diverse classes of rearrangements, both within genes and in intergenic regions.

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

Next, we sought to identify significantly recurrent breakpoint (SRB) regions across our combined mCRPC cohort of 143 cases in a genome-wide, unbiased manner. We applied a Gamma-Poisson regression approach to model the occurrences of SV breakpoints within 100 kB windows across the cohort as previously described (Imielinski et al., 2017). Importantly, this model nominates significantly recurrent breakpoint regions likely to function as cancer drivers by accounting for six different covariates, including sequence features (e.g., GC-content and transposable elements), fragile sites, heterochromatin regions, DNase I hypersensitivity sites (DHS), and replication timing (**Methods**).

We identified a total of 55 significantly recurrent breakpoint regions genome-wide across our combined mCRPC cohort (Benjamini-Hochberg corrected, q-value ≤ 0.1, **Figure 2A**, **Table S2A**). Thirty-six (65.5%) SRB regions were located within 1 Mb of 14 known prostate cancer driver genes, including *AR* and its enhancer, *TMPRSS2/ERG*, *TP53*, *PTEN*, *FOXA1*, and *MYC*. For these 14 driver genes, we did not observe significant differences in SV alteration frequencies when comparing between pre-treatment (N=54) and post-progression (N=89) samples, except in the case of *ERG, for* which the SV frequency was enriched in pre-treatment samples (Fisher's exact test, $p = 0.0395$; all other genes had $p > 0.05$, **Figure S3B**). We also did not identify any major differences in the alteration frequencies of prostate cancer genes in four patients who had paired samples collected both before treatment with and after progression on an ARSi. (**Figure S3A**).

176 We then sought to compare how SVs drive prostate cancer across disease states. For the
177 localized disease state, we utilized genome alteration calls from 278 primary localized prostate
178 cancer tumors from the PCAWG study (Campbell et al., 2020; Li et al., 2020). Using Gamma-
179 Poisson regression, we first identified 47 SRB regions in localized prostate cancer tumors (**Figure**
180 **S2A, Table S2B**). Six prostate cancer genes (*TMPRSS2*, *ERG*, *TP53*, *PTEN*, *IL6ST*, *ELK4*) within
181 mCRPC SRB regions were also found within or in proximity (less than 1 Mb) to an SRB region in
182 localized disease. By contrast, four SRBs (three near *SEL1L3* and one near *PRKDC*) were unique
183 to localized disease, while 27 SRBs were unique to mCRPC with six genes nearby (*LSAMP*,
184 *ETV1*, *MYC*, *PTPRD*, *FOXA1*, *AR*). When comparing SV alteration frequencies for the 14 genes
185 located within SRB regions in either mCRPC or localized tumors, 12 genes were significantly
186 more altered in mCRPC samples, while *TMPRSS2* and *ERG* were significantly more altered in
187 localized disease (Fisher's exact test, $p < 0.05$ for all genes, **Figure 2B**). Thus, localized prostate
188 cancer and mCRPC have significantly different landscapes of recurrent SVs.

189 To explore the potential functional consequences of SVs in intergenic SRB regions, we
190 overlapped SV breakpoints with locations of H3K27ac marks specific to mCRPC (Pomerantz et
191 al., 2020). We observed that intergenic SVs within SRB regions in the mCRPC cohort included
192 gene flanking events that were enriched at putative enhancer regions for *AR*, *MYC*, and *FOXA1*,
193 which all had frequent focal duplication events at sites marked by mCRPC-specific H3K27ac
194 deposition (**Figure 2C**). Interestingly, an intragenic deletion SRB region was observed near the
195 transcription start site of *LSAMP*, also overlapping H3K27ac marks. *PTEN* had a high level of
196 both gene transecting and flanking deletions, leading to SV breakpoints that were spread more
197 broadly around the gene.

198 We also observed an enrichment of metastatic-specific *AR* binding sites (ARBS) compared to
199 localized primary ARBS within the 55 mCRPC SRB regions (**Figure 2D**, one-sided proportion test,
200 $p = 1.05 \times 10^{-8}$). This enrichment was not observed for localized primary SRB regions ($p = 0.22$).
201 These results highlight that SVs within mCRPC SRB regions may be capturing the genome-wide
202 footprint of activated *AR* signaling that occurs with castration resistance.

203 Refined landscape of ETS gene fusions from integrated analysis of the genome and 204 transcriptome

205 We applied gene fusion analysis by integrating both genome rearrangements and fusion RNA
206 transcript information from 127 samples with RNA-seq data (**Figure 1A**, **Table S2C**, **Methods**).
207 For gene fusions involving E26 transformation-specific (ETS) transcription factor gene family
208 members (*ERG*, *ETV1*, *ETV4* and *ETV5*), we detected 50 events supported by both DNA and
209 RNA evidence, 15 supported by only DNA evidence, and 10 supported by only RNA evidence
210 (**Figure 2E**, **Figure S2D**). Overall, 74 samples (51.7% of 143 samples) harbored a fusion event
211 of the *ETS* gene family, consistent with previous reports (Tomlins et al., 2005, 2007) (**Figure 1B,**
212 **Table S2C**).

213 Among the ETS fusions, *ERG* was most commonly involved with *TMPRSS2* as the fusion partner
214 (54 out of 57 cases, **Figure 2G**). Other common ETS fusion partners were *SLC45A3* (7 cases)
215 and lncRNA RP11-356O9.1 downstream of *FOXA1* (3 cases). *ETV1* had eight distinct fusion
216 partners, which is consistent with previous reports that *ETV1* is a promiscuous ETS fusion
217 member (Kumar-Sinha et al., 2015) (**Figure 2F**).

218 We observed that fusions of the ETS family members *ERG*, *ETV1*, *ETV4* and *ETV5* were mutually
219 exclusive, except for one sample which harbored fusions of both *ERG* and *ETV1* (**Figure S2D**).
220 In addition, gene fusion events were correlated with higher expression of the corresponding ETS
221 genes they involved (Wilcoxon rank-sum tests, $p < 0.05$ for all genes, **Figure 2E**). In the 38 cases
222 which did not show any evidence for an ETS fusion, we noted that presence of high-level

223 expression (z-score > 1) of ETS genes *ERG*, *ETV1*, *ETV4*, and *ETV5* were also mutually
224 exclusive (Fisher's exact test, p = 0.480 for *ETV4*, p = 0.363 for *ETV5*, **Figure S2D**). These may
225 represent cases of missed fusion calls, or cases in which ETS family members are
226 transcriptionally activated through non-genetic mechanisms.

227 Interestingly, we also observed 20 cases (14.0% of 143 cases) involving fusions between the ETS
228 family member *ELK4* and its upstream gene *SLC45A3*. While the *ELK4* locus was an SRB in our
229 analysis (**Figure 2A** and **Figure S2B**), manual inspection of individual samples revealed evidence
230 for a genomic event capable of producing an *ELK4* fusion in only 1 out of 20 cases (**Figure S2D**
231 **and data not shown).** In contrast, 19 other cases showed *ELK4* fusions on RNA-sequencing
232 alone, consistent with a mechanism of cis-splicing or transcriptional read-through events that may
233 perhaps be induced by local genomic alterations (Qin et al., 2017; Rickman et al., 2009; Zhang
234 et al., 2012) (**Table S2C**). Importantly, although *ELK4* fusions were significantly correlated with
235 higher expression of *ELK4* (Wilcoxon rank-sum test, p = 7.91×10^{-5}, **Figure S2D**), these events
236 were not mutually exclusive with fusions of other ETS family members (Fisher's exact test, p =
237 0.472). Thus, the functional consequences of these *ELK4* fusions and whether they contribute to
238 prostate cancer pathogenesis in a manner similar to other ETS fusions remains to be determined.

239 **Diverse and complex rearrangements driving *AR* signaling in mCRPC**

240 Genomic alterations involving the *AR* locus play an important role in sustaining *AR* signaling in
241 mCPRC (Chen et al., 2004; Quigley et al., 2018; Visakorpi et al., 1995; Viswanathan et al., 2018).
242 However, the complete spectrum of diverse structural mechanisms that underlie *AR* activation in
243 mCRPC has not been fully characterized. To understand the relationship between different modes
244 of somatic *AR* activation, we determined copy number at the *AR* gene body and its upstream
245 enhancer and categorized samples into distinct groups of: **(1)** co-amplification (N = 99, 69.2% of
246 143 cases); **(2)** selective *AR* gene body amplification (N = 4, 2.8% of 143 cases); **(3)** selective *AR*
247 enhancer gains (N = 17, 11.9% of 143 cases), and **(4)** lack of amplification for both (N = 23, 16.0%
248 of 143 cases) (**Figure 3A-C, Table S3**). For the 122 samples with expression data available, we
249 observed that *AR* gene expression was higher in the co-amplification and selective enhancer
250 categories compared to samples with no amplification, after accounting for tumor purity and ploidy
251 (ANCOVA/TukeyHSD p-values 5.6×10^{-11} and 4.5×10^{-4}, respectively), but not for selective *AR*
252 status (ANCOVA p = 0.098) (**Figure 3B, Methods**). Interestingly, we observed that samples with
253 selective enhancer duplication exhibited similar *AR* expression levels to samples with co-
254 amplification (ANCOVA, p = 0.31), even though enhancer duplications involved lower-copy gains
255 (mean 2.73, range 1.97 - 5.02) compared to co-amplified samples (mean 12.87, range 1.55 -
256 150.57) (**Figure 3A**). This is consistent with previous results (Viswanathan et al., 2018) and
257 suggests a mechanism whereby *AR* expression levels are increased through even modest
258 genomic expansion of enhancer elements.

259 We then systematically and manually curated the diverse mechanisms of rearrangements
260 activating *AR* signaling by analyzing patterns of SVs at the *AR* locus (**Figure 3C, Table S3,**
261 **Methods**). We observed a total of 62 samples (43.4% of 143 samples) with tandem duplication
262 SV events that spanned the enhancer with breakpoints located within 1 Mb, including 16 cases
263 (11.2% of 143 samples) with selective enhancer copy number amplification status (**Figure 3D**).
264 Thirty-two samples (22.4% of 143 samples) harbored intragenic rearrangements within *AR*, which
265 may have implications for the production of truncated, constitutively-active *AR* splice variants
266 (Henzler et al., 2016). For example, in case DTB-124-BL, we observed a focal intragenic deletion
267 spanning exons 4-8 of *AR*, which includes the ligand binding domain, resulting in the expression
268 of truncated *AR* variants (Kanayama et al., 2021) (**Figure 3E**, **Figure S3C**). Interestingly, in the
269 21 samples with selective *AR* enhancer or selective *AR* gene body copy number gain, none
270 harbored intragenic SV events in *AR*.

271 We also examined the landscape of complex rearrangement mechanisms involving *AR*; these
272 mechanisms involve multiple SV events and copy number patterns, including chromothripsis,
273 extrachromosomal DNA (ecDNA), chromoplexy, and breakage-fusion-bridge cycle (BFB)
274 (**Methods**). Chromothripsis of a region or the entire X chromosome involving the *AR* locus was
275 detected in 5 samples, all of which had co-amplification of *AR* and enhancer, suggesting that
276 following repair after catastrophic DNA shattering the *AR* locus was retained or further amplified
277 (**Figure 3F, Figure 3G**). Thirteen samples (9.1% of 143 samples) showed very high levels of *AR*
278 and enhancer copy number, suggesting the possibility of their presence on extrachromosomal
279 elements (ecDNA, **Figure 3H**). In 40 samples (28.0% of 143 samples), the most frequent complex
280 rearrangement mechanism, BFB, led to *AR* locus amplification, including instances following
281 chromothripsis (Stephens et al., 2011; Umbreit et al., 2020) (**Figure 3G**). Overall, we noted that
282 complex rearrangement events, which frequently co-occurred, were significantly enriched in
283 samples with co-amplification of *AR* and enhancer compared to those with selective enhancer
284 copy number gain status (Fisher's exact test, $p = 1.52 \times 10^{-4}$).

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

286 To systematically characterize genome-wide structural rearrangement patterns in mCRPC, we
287 performed rearrangement signature analysis using SV breakpoint features, non-negative matrix
288 factorization, and known reference signatures (Degasperi et al., 2020; Nik-Zainal et al., 2016)
289 (**Methods**). First, we derived signatures *de novo*, which identified eight signatures: six that
290 matched reference signatures (RefSigs) also observed in localized prostate cancer (> 0.91 cosine
291 similarity), one that matched an ovarian cancer RefSig.R14 associated with large segment (100
292 kB-10 Mb) TDP (0.96 cosine similarity), and one that was likely an artifact specific to linked-read
293 sequencing (**Figure S4A-C, Table S4A and 4B**). Therefore, we excluded the linked-read data
294 and focused on standard WGS data from 101 mCRPC cases for further SV signature analysis.
295 We fit these samples to the nine known RefSigs from localized prostate cancer (R1-4, R6a-b, R8-
296 9, R15) and the one (R14) from ovarian cancer (**Figure S4A**, **Table S4C**). Overall, eight of the
297 RefSigs were detected across our cohort (R1-2, R4, R6a-b, R9, R14-15). Notably absent in
298 mCRPC were RefSig.R8 (short, 1-10 kB inversions) and RefSig.R3, which is associated with
299 germline *BRCA1* mutations and short (1-100 kB) tandem duplications (Degasperi et al., 2020;
300 Glodzik et al., 2017; Nik-Zainal et al., 2016; Willis et al., 2017) (**Figure S4D**). By contrast, we
301 observed increased prevalence of some signatures in mCRPC compared to localized disease,
302 including RefSig.R2 (large SV classes, abundant translocations; 97% vs. 60%), RefSig.R4
303 (clustered translocation events; 37% vs. 27%), and RefSig.R15 (large deletions and inversions,
304 48% vs. 37%) (**Figure S4D**).

305 To investigate whether molecular subtypes in mCRPC can be grouped based on SV patterns, we
306 applied hierarchical clustering on the exposure of the eight fitted signatures and identified nine
307 distinct SV clusters (**Figure 4, Table S4C**). We observed that samples in SV Cluster 1 were
308 composed of non-clustered translocation events and were significantly enriched for the presence
309 of chromoplexy (χ^2 test, FDR corrected, q = 0.12). SV Cluster 3 was characterized by many short
310 deletions and was significantly enriched for *BRCA2* mutations ($q = 5.01 \times 10^{-4}$). SV Cluster 5 was
311 significantly enriched for *SPOP* mutations (q = 0.02), with no instances of ETS gene family fusion
312 (q=0.06), consistent with previous reports (Barbieri et al., 2012). SV Cluster 6 had the highest
313 prevalence of *TP53* mutation (q = 0.02), while SV Cluster 7 samples harbored the TDP associated
314 with *CDK12* inactivation ($q = 3.52 \times 10^{-11}$) as well as enrichment for *CCND1* gains (q = 0.02),
315 consistent with previous reports (Nguyen et al., 2020; Wu et al., 2018). The remaining clusters
316 did not have enrichment for any alterations in known driver genes; however, distinct SV patterns
317 were still evident in SV Cluster 4 (non-clustered tandem duplications), 8, and 9 (increased
318 clustered SV events of various classes).

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

DISCUSSION

We present a large-scale and comprehensive integrative genomic analysis of both localized prostate cancer and mCRPC, with a focus on how structural variation drives each of these clinically distinct disease states. The size of our cohort as well as our harmonized analysis pipeline enable a sharper view of the genetic alterations that drive prostate cancer across its natural history as compared with prior studies, which have involved either smaller cohorts or been limited to a single disease state (Campbell et al., 2020; Cancer Genome Atlas Research Network, 2015; Quigley et al., 2018; Viswanathan et al., 2018).

In contrast to somatic SNVs/indels and CNAs that occur within coding regions, the functional and clinical significance of alterations within noncoding regions has often been more challenging to interpret, as localized variations in mutability may result in the nomination of certain recurrently mutated sites that do not necessarily drive cancer (Glodzik et al., 2017; Imielinski et al., 2017; Nik-Zainal et al., 2016). This issue is even more complex for SVs, in which different classes of SVs spanning the same loci would be predicted to have distinct functional consequences. Our study addresses the former issue by identifying genomic hotspots of structural variation with rigorous correction for covariates including nucleotide composition, replication timing, sensitivity to DNA breaks, repetitive elements, and chromatin state. We address the latter issue by careful curation of SV classes to distinguish those that are likely to be activating versus inactivating (**Figures 1B and 3; Methods**).

Our approach has produced several insights into the recurrent rearrangements that drive prostate cancer. First, several top hotpots of rearrangement genome-wide lie in noncoding regions outside the boundaries of known prostate cancer genes. In many cases, such as for *AR*, *MYC*, and *FOXA1*, these hotspots overlap with active chromatin marks and likely represent distal regulatory regions for neighboring prostate cancer genes (**Figure 2**). These data are intriguing in light of the observation that a majority of prostate cancer germline susceptibility loci are in noncoding regions (Giambartolomei et al., 2021). Second, the loci altered by rearrangements differ across prostate cancer disease states (**Figure 2B**). For example, *TMPRSS2-ERG* rearrangements are enriched in localized prostate cancer versus mCRPC, while alterations in *AR, FOXA1, MYC*, and *LSAMP* are more frequent in mCRPC than in localized disease. Third, certain driver genes are enriched for alteration by SVs as compared to other mutagenic processes. For example, *PTEN* inactivation frequently occurs via gene transecting SV events, while *TP53* inactivation is primarily caused by SNVs (**Figure 1C and Figure S1**).

Our systematic genomic discovery efforts again highlight the primacy of *AR* as a target of somatic alteration in hormone-refractory mCRPC. We have precisely catalogued the diverse genomic mechanisms leading to *AR* activation across our large cohort and find that different alteration mechanisms are associated with differing levels of *AR* amplification. Whether the precise mechanism by which *AR* is altered in a given patient is associated with differences in response to *AR* pathway inhibition warrants further investigation in clinically annotated cohorts. High levels

366 of *AR* signaling in mCRPC may also underlie the patterns of structural variation seen in this
367 disease state. Strikingly, we found that *AR* binding sites overlapped several of the top SV hotspots
368 in mCRPC (**Figure 2D**), consistent with the notion that androgen signaling may induce DNA
369 double-strand breaks that resolve as rearrangements (Haffner et al., 2010).

370 In addition to alterations in highly validated prostate cancer genes, we identified highly recurrent
371 rearrangements near or involving genes that have not been extensively studied in prostate cancer,
372 such as *LSAMP, PTPRD*, and *TTC28*. *LSAMP* encodes a cell-surface glycoprotein and has a
373 possible tumor suppressor role in several cancers (Chen et al., 2003; Kresse et al., 2009; Kühn
374 et al., 2012); notably, deletions near the *LSAMP* locus have been shown in one report to be
375 enriched in African American men with prostate cancer (Petrovics et al., 2015). *PTPRD*, a
376 receptor protein tyrosine kinase, has been previously identified as a target of inactivating
377 alteration in glioblastoma (Veeriah et al., 2009). We observed frequent SVs near the *TTC28* locus,
378 which encodes an L1 retrotransposon element, specifically in mCRPC (**Figure 1C)**. L1
379 retrotranspositions originating from *TTC28* have been reported previously in colorectal cancer
380 (Pitkänen et al., 2014; Pradhan et al., 2017; Tubio et al., 2014); our results raise the intriguing
381 possibility that they may also be frequent in prostate cancer, and may be activated by the pressure
382 of hormonal therapy. Interestingly, we also observed SRBs near *ELK4* along with a relatively high
383 frequency of *SLC45A3-ELK4* chimeric transcripts, although it was not clear how the
384 rearrangements at this locus produced the chimeric transcripts in most cases. Whether this fusion
385 functions similarly to or in a distinct mode from other ETS fusions is an exciting area for future
386 study.

387 Our study also extends beyond the analysis of SVs at individual loci to molecularly subclassify
388 prostate cancers based on their genome-wide signatures of structural variation. Sample clustering
389 based on SV signature exposure defines distinct molecular subtypes of prostate cancer and may
390 find utility alongside signatures of single base substitution and copy number to more precisely
391 define tumor subtypes (Alexandrov et al., 2013, 2020; Degasperi et al., 2020; Macintyre et al.,
392 2018; Wang et al., 2021). In the mCRPC cohort, we identified 9 molecular subtypes based on SV
393 signature, and several clusters had clear associated genomic alterations including chromoplexy
394 (cluster 1), *BRCA2* alterations (cluster 3), *SPOP* alterations (cluster 5), *TP53* alterations (cluster
395 6) and *CDK12*/*CCND1* alterations (cluster 7). Future studies with larger WGS cohorts may identify
396 associated alterations in the remaining clusters. Notably, unsupervised clustering identified
397 samples with clear SV signatures but without detectable associated mutations in genes or
398 pathways that plausibly contribute to the genomic alterations (**Figure 4**). Moreover, clinical
399 outcomes were more separated by SV signature cluster than by alterations of the mutations
400 associated with those clusters (**Figure S4D-E**).

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

405 ACKNOWLEDGEMENTS

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

409 This work was supported by the National Institutes of Health (K22 CA237746 to G.H.; P01
410 CA163227 and R01 CA234715 to P.S.N.; R01 GM107427, R01 CA193910, and R01 CA251555
411 to M.L.F.; R35 CA197568 to M.M.), Department of Defense Prostate Cancer Research Program
412 (Physician Research Award W81XWH-17-1-0358 to S.R.V.; W81XWH-19-1-0565 and W81XWH-

21-1-0234 to M.L.F.; PC200262 to P.S.N.), PCF Young Investigator Awards (G.H. and S.R.V.), PCF-Movember Challenge Award (to E.M.V.), Brotman Baty Institute for Precision Medicine (to G.H.), the Fund for Innovation in Cancer Informatics Major Grant (to G.H.), the V Foundation Scholar Grant (to G.H.), Wong Family Award in Translational Oncology and Dana-Farber Cancer Institute Medical Oncology grant (to A.D.C.), H.L. Snyder Medical Research Foundation and the Cutler Family Fund for Prevention and Early Detection (to M.L.F.), and the Pan-Mass Challenge team IMAGINE (to M-E.T.), American Cancer Society Research Professor (M.M.).

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

AUTHOR CONTRIBUTIONS

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

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

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

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.

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

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

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.

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

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

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

DECLARATION OF INTERESTS

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

E.M.V.: Advisory/Consulting: Tango Therapeutics, Genome Medical, Invitae, Enara Bio, Janssen, Manifold Bio, Monte Rosa; Research support: Novartis, BMS; Equity: Tango Therapeutics, Genome Medical, Syapse, Enara Bio, Manifold Bio, Microsoft, Monte Rosa; Travel reimbursement: Roche/Genentech; Patents: Institutional patents filed on chromatin mutations and immunotherapy response, and methods for clinical interpretation; intermittent legal consulting on patents for Foaley & Hoag

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

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

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

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

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

All other authors declare no competing interests.

454 ## MAIN FIGURE LEGENDS

455 ### Figure 1. Study design and genomic landscape of mCRPC.

456 **(A)** Workflow of study and data analysis. Tumor specimens (grey) from both primary prostate
457 cancer and mCRPC were included in this study. Linked-read and short-read whole-genome
458 sequencing (WGS) and RNA-sequencing datasets were either generated for this study or
459 reanalyzed from prior studies (Quigley et al., 2018; Viswanathan et al., 2018). A pooled dataset
460 of 143 mCRPC samples with WGS data was used in this study after curation (**Methods**). Genomic
461 alteration call-sets for 278 primary localized prostate cancer samples were obtained from
462 ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) (Campbell et al., 2020; Li et al.,
463 2020). For 125 mCRPC samples, RNA-seq was used. The overview of the genomic alteration
464 and characterization analysis is shown.

465 **(B)** Clinical annotations and somatic alterations for 143 patient samples in the pooled mCRPC
466 cohort. Samples are ordered by treatment type; the four patients with pre-treatment and post-
467 progression pairs are placed at the right. (Top) Clinical and sample information and genomic
468 pattern classifications. (Middle) Distribution of genomic rearrangement types in individual samples.
469 (Bottom) Mutational burden for SNVs and indels computed as number of mutations per mega-
470 base pair (Mb). Y-axis shown in logarithmic scale. Threshold lines indicates mutational burden at
471 2.5 and 5 mutations per Mb.

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

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

484 **(A)** Analysis of significantly recurrent breakpoint (SRB) identified regions of rearrangement
485 hotspots, genome-wide, using a Gamma-Poisson regression model. Each dot corresponds to a
486 100 kB bin (n=26,663 total bins). Statistically significant SRB bins with FDR (Benjamini-Hochberg)
487 q-value ≤ 0.1 (n=55) are colored based on the distance to the nearest known prostate cancer
488 driver gene, within 1 Mb. The driver genes within 1 Mb of the SRB bins are labeled. A square
489 bracket is used for genes spanning multiple bins. Bins with q-value > 0.1 were not significant
490 (grey).

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

496 **(C)** Patterns of rearrangements at the loci of driver genes identified at SRB regions in mCRPC
497 cohort of 143 tumors. Cumulative counts of intra-chromosomal SV events (tandem duplications
498 "TandemDup", deletions, and inversions) were computed based on the breakpoints and span of
499 the events. Histone H3 lysine 27 acetylation (H3K27ac) and *AR* binding sites (ARBS) specific to

500 mCRPC were obtained from a previous study (Pomerantz et al., 2020). Inter-chromosomal
501 translocations are not shown. Genome coordinates based on hg38 build.

502 **(D)** Overlap of *AR* binding sites (ARBS) within SRB hotspots of mCRPC (55 regions) and primary
503 localized prostate (47 regions) cohorts. Metastatic-specific and primary localized-specific ARBS
504 were obtained from previous studies (Pomerantz et al., 2015, 2020). χ^2 test of independence p-
505 values are shown.

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

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

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

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

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

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

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

534 **(D-H)** Complex and simple rearrangement patterns involving the *AR* locus, including focal
535 duplication events on *AR* enhancer **(D)**, intragenic deletion event leading to loss of ligand binding
536 domain of *AR* **(E)**, chromosomal level chromothripsis events involving *AR* and enhancer **(F)**, arm-
537 level chromothripsis coinciding with *AR* amplification by break-fusion-break cycle **(G)**, extra-
538 chromosomal DNA amplicon including *AR* and enhancer **(H)**. *AR* gene boundary (green) and its
539 enhancer (yellow) are shown; concave arcs, intra-chromosomal SV events; convex arcs, inter-
540 chromosomal SV events. Copy number values represent 10 kB bins and have been tumor purity
541 corrected.

542 **Figure 4. Clustering of mCRPC SV signatures**

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

549

550 # RESOURCE AVAILABILITY

551 ## Lead contact

552 Further information and requests for resources and reagents should be directed to and will be
553 fulfilled by the lead contact, Gavin Ha (gha@fredhutch.org).

554 ## Materials availability

555 This study did not generate new unique reagents.

556 ## Data and code availability

557 • Whole genome sequencing data have been deposited at dbGaP under accession number
558 phs001577 and access is available upon request.
559 • All original code has been deposited at GitHub and is publicly available as of the date of
560 publication. Links are provided in the key resources table.
561 • Any additional information required to reanalyze the data reported in this paper is available
562 from the lead contact upon request.

563 # EXPERIMENTAL MODEL AND SUBJECT DETAILS

564 ## Human subjects

565 For tumor biopsies profiled via linked-read sequencing, samples were collected from individuals
566 with mCRPC who provided informed consent on institutional IRB-reviewed protocols, as
567 previously described (Viswanathan et al., 2018). Uniformly reanalyzed data were generated as
568 described in the respective studies (Campbell et al., 2020; Quigley et al., 2018).

569 # METHOD DETAILS

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

571 Data processing of the linked-read genome sequencing data include high molecular weight DNA
572 preparation and sequencing library construction followed protocols as previously described
573 (Viswanathan et al., 2018). DNA was extracted from tumor samples using the MagAttract HMW
574 DNA Kit (QIAGEN), and then quantified using Quant-it Picogreen assay kit (Thermo Fisher) on a
575 Varioskan Flash Microplate Reader (Thermo Fisher). For germline samples, pre-extracted DNA
576 was size-selected on the PippinHT platform (Sage Science) and then quantified using the Quant-
577 it Picogreen assay kit (Thermo Fisher) on a Varioskan Flash Microplate Reader (Thermo Fisher).
578 Libraries were constructed using the 10X Chromium protocol (10X Genomics), with the fragment
579 sizes determined using the DNA 1000 Kit and 2100 BioAnalyzer (Agilent Technologies) and
580 quantified using qPCR (KAPA Library Quantification Kit, Kapa Biosystems). WGS libraries were
581 sequenced using the Illumina HiSeqX platform. The Long Ranger v2.2.2 pipeline (10X Genomics)
582 was used for aligning sequence reads to the human genome hg38 (GRCh38).

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

586 # QUANTIFICATION AND STATISTICAL ANALYSIS

587 ## List of known prostate cancer driver genes

588 For analyses limited to established prostate cancer driver genes, a curated list of 159 known
589 prostate cancer driver genes was assembled from several prior studies (Armenia et al., 2018; van
590 Dessel et al., 2019; Quigley et al., 2018; Viswanathan et al., 2018). The list of genes are provided
591 in **Table S1**.

592 ## Somatic mutation analysis

593 *Somatic mutation detection*

594 Somatic mutation calls for samples based on linked-read sequencing were generated by Mutect2
595 from the Genome Analysis Toolkit (GATK) (Van der Auwera and O'Connor, 2020). Default
596 parameters were used on individual pairs of tumor and normal samples following the standard
597 GATK pipeline. A panel of normals based on all normal samples was used to filter out germline
598 variants. The SNV calls were further processed using the modified version of LoLoPicker (Carrot-
599 Zhang and Majewski, 2017) as described previously (Viswanathan et al., 2018). The panel of
600 normals for LoLoPicker was generated from 52 normal samples based on linked-read sequencing.
601 The final SNV call set was composed of the common variants called by both Mutect2 and
602 LoLoPicker. Somatic indels for linked-read samples were called by Strelka (Saunders et al., 2012).
603 All parameters were default except the following modifications: sindelNoise = 0.000001,
604 minTier1Mapq = 20. Somatic mutation calls for the 101 WGS samples based on short-read
605 sequencing including SNV and indels based on Strelka were obtained from a prior study (Quigley
606 et al., 2018). All variants were further annotated using annovar with "table_annovar.pl" to
607 functionally annotate genetic variants. The paramter -neargene was set to 5000 to define the
608 promoter region as 5 kB upstream of the transcription start site of a protein coding gene.

609 *Analysis of significantly mutated genes*

610 R package dndscv (Martincorena et al., 2017) was used to identify significantly mutated genes.
611 For driver discovery on GRCh38, a precomputed database corresponding to human genome
612 GRCh38.p12 was downloaded and used as the reference database. A global q-value ≤ 0.1 was
613 applied to identify statistically significant (novel) driver genes. To reduce false positives and
614 increase the signal to noise ratio, we only considered mutations in Cancer Gene Census genes
615 (v81) (Tate et al., 2019).

616 ## Copy-number analysis of linked-read WGS and short-read WGS data

617 *Copy-number calls*

618 The ploidy and purity corrected copy-number of all mCRPC samples in this study was analyzed
619 by TitanCNA (Ha et al., 2014) and ichorCNA (Adalsteinsson et al., 2017), with different pipeline
620 settings. For WCDT samples, the snakemake workflow for Illumina sequencing was applied with
621 the following parameters modified: ichorCNA_normal: c(0.25, 0.5, 0.75); ichorCNA_ploidy:
622 c(2,3,4); ichorCNA_includeHOMD: TRUE; ichorCNA_minMapScore: 0.75;
623 ichorCNA_maxFracGenomeSubclone: 0.5; ichorCNA_maxFracCNASubclone: 0.7;
624 TitanCNA_maxNumClonalClusters: 3; TitanCNA_maxPloidy: 4. The workflow is available at
625 https://github.com/GavinHaLab/TitanCNA_SV_WGS.

626 For linked-read data samples, a Snakemake workflow for 10X Genomics whole genome
627 sequencing data was used with the following parameters modified:
628 TitanCNA_maxNumClonalClusters: 3; TitanCNA_maxPloidy: 4. TitanCNA solutions were
629 generated for number of clonal clusters from 1 to 3 and ploidy initializations from 2 to 4. Optimal

solutions were selected as described, with manual inspection to confirm tumor ploidy and clonal cluster selection (Viswanathan et al., 2018); solutions are provided in **Table S1J**. The workflow can be accessed at https://github.com/GavinHaLab/TitanCNA_10X_snakemake. The final copy-number call-set is included in **Table S1I**.

Recurrent somatic copy-number alteration

GISTIC 2.0 was used to detect regions with recurrent CNA in mCRPC samples. For input, all copy numbers (logR_Copy_Number from TITAN output) were converted to log2 copy ratio using the median logR copy number from genome-wide (separately for autosomes and X chromosome) as denominator. We set corrected logR copy number to -1.5 for segments where corrected log R copy number below -1.5 and set values to 0 if copy neutral. GISTIC2.0 was run with the following parameters: td 0.5; ta 0.1; genegistic 0; maxseg 5000; js 4; cap 1.5; broad 1; brlen 0.75; conf 0.99; qvt 0.25; armpeel 1; rx 0; gcm mean; do_gene_gistic 1; savegene 1; scent median. Wide peaks detected by GISTIC2 were re-annotated based on overlapping genomic coordinates, using prostate cancer driver genes.

Structural variant analysis

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

For each tumor-normal pair of samples with linked-read genome sequencing data, three variant callers were used to detect structural variants: SvABA (Wala et al., 2018), GROC-SVS (Spies et al., 2017), Long Ranger version 2.2.2 (https://support.10xgenomics.com/genome-exome/software/pipelines/latest/using/wgs).

The SvABA analysis was performed using default tumor-normal paired settings. Re-analysis of low confidence (based on evidence from discordant and split reads) events filtered by SvABA was performed to 'rescue' SVs using linked-read barcode overlap between pairs of breakpoints within a given SV event, as previously described (Viswanathan et al., 2018). Only SV events having span of 1.5 times the mean molecule length in the library were considered for rescue. We further rescued low confidence intra-chromosomal SV events with span > 50 kB filtered by SvABA if at least one of the breakpoint pair was within 100 kB of a CNA boundary or (2) if both breakpoints were each within 1 Mb of the boundaries for the overlapping CNA event and the length of the SV overlaps this CNA event by > 75%. Inter-chromosomal translocation SV events filtered by SvABA are rescued if both breakpoints were within 100 kB of CNA boundaries.

GROC-SVS analysis was performed using two-sample (tumor-normal paired) mode or three-sample (pre-treatment, post-progression, normal) mode when applicable. SV events were retained if all following conditions were satisfied: (1) $p < 1x10^{-10}$, (2) minimum barcode overlap ≥ 2 on the same haplotype, (3) no more than 1 barcode overlap between different haplotypes, (4) FILTER value reported by the software was within this set {"PASS", "NOLONGFRAGS", "NEARBYSNVS", or "NEARBYSNVS; NOLONGFRAGS"}, and (5) classified as somatic.

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

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

An SV panel of normals (PoN) was generated using germline events from SvABA and Long Ranger calls. There are two components to this panel: (1) frequency of germline events at exact breakpoint locations (SVpon.bkpt) and (2) frequency of germline event breakpoint overlapping within tiled windows of 1 kB (SVpon.blackListBins). The PoN was used to filter events in the combined SV call-set when an SV has at least one breakpoint with SVpon.bkpt ≥ 2 and overlapping bin with SVpon.blackListBins ≥ 100.

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

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

Classification of structural variants in mCRPC

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

- Deletion. Events having the orientation combination (reverse, forward) and length between 10 kB and 1 Mb were classified as deletions. The copy-number 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 $c_2^{down} > c^{mean}$, and $s \leq 5$. In addition, events overlapping copy-number deletion or LOH segments were also considered as deletions.
- Tandem duplication. Events having the orientation combination (forward, reverse) and length between 10 kB and 1 Mb were classified as tandem duplications. The copy-number 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 $c_2^{down} < c^{mean}$, and $s \leq 5$. In addition, events overlapping copy-number gain or copy neutral LOH segments were also considered as tandem duplications.
- Inversion. Events having the orientation combination (forward, forward) or (reverse, reverse) and length between 10 kB and 5 Mb were classified as inversions. Furthermore, inversion events shorter than 30 kB with unequal copy-numbers around either breakpoint were classified as fold-back inversions.

- Balanced rearrangement (balanced). Events having the orientation combination same to inversion (forward, forward) or (reverse, reverse), but length larger than 5 Mb were classified as balanced events. The copy-number values of breakpoints should satisfy $c_1^{up} = c_1^{down}$ and $c_2^{up} = c_2^{down}$, or $c_1^{up} = c^{mean}$ and $c_2^{down} = c^{mean}$.
- Unbalanced rearrangement (unbalanced). Intra-chromosomal events which did not fulfill any of the above criteria and having length larger than 10 kB were classified as unbalanced events.
- All SV events with two breakpoints located on different chromosomes were classified as translocations.

ICGC/TCGA PCAWG localized prostate cancer structural variants

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

Tandem duplicator phenotype

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

Chromothripsis analysis

Chromothripsis events were detected by ShatterSeek R package (Cortés-Ciriano et al., 2020). Structural variants calls by SvABA and copy-number calls by TitanCNA were used as input data (excluding Y chromosome). In the input, consecutive segments were joined as one if they had the same copy-number value and centromere regions were filtered out.

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

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

768 ## Chromoplexy analysis

769 ChainFinder was used to detect chromoplexy events (Baca et al., 2013). Ten samples that were
770 considered as TDP (01115374-TA2, 01115202-TC2, 01115248-TA3, 01115503-TC2, 01115257-
771 TA4, 01115284-TA9, 01115414-TA1, DTB-063-BL, DTB-183-BL, DTB-214-BL) were excluded
772 from this analysis. In addition, four samples that were found to cause numeric instabilities of
773 ChainFinder were also excluded (DTB-023-BL, DTB-102-PRO, DTB-111-PRO, DTB-151-BL).
774 The SV calls of remaining samples were further filtered to exclude those that were located within
775 5 Mb from chromosomal ends or overlapping chromothripsis regions. For copy-number input,
776 segments that were determined as copy neutral by TitanCNA were set to have log copy-ratio of
777 0. Copy-ratio of the other segments were computed from copy-number values generated by
778 TitanCNA divided by 2 for autosomes or 1 for X chromosome. Log copy-ratio values less than -
779 1.5 were set to -1.5. The output of ChainFinder was used for determining chromoplexy status of
780 individual samples. A chromoplexy event was defined as a chain including at least 5
781 rearrangement events and involving more than 2 different chromosomes. Samples having at least
782 2 such events were considered positive for chromoplexy status.

783 ## ChIP-seq data analysis

784 ChIP-seq data used in this study were downloaded from Gene Expression Omnibus (GEO)
785 (Barrett et al., 2013) and the Sequence Read Archive (SRA) (Leinonen et al., 2011). Short reads
786 were mapped to the human genome GRCh38 (hg38) using bwa (Li and Durbin, 2009). Because
787 read lengths were less than 50bp, the bwa aln command with default parameters was used for
788 mapping. MACS2 (Zhang et al., 2008) was used to identify peaks from mapped ChIP-seq data.
789 For histone modification marks, MACS2 callpeak command was applied with --nomodel --broad
790 –extsize 146. For CTCF data, MACS2 callpeak command was used with --nomodel --extsize 200.
791 Below is the list of ChIP-seq datasets involved in this analysis.

792 - H3K4me3, H3K27me3 and CTCF (GSE38685) (Bert et al., 2013).
793 - H3K36me3 and H3K9me3 (GSE98732) (Du et al., 2019).
794 - H3K4me1 and H3K27ac (GSE73785) (Taberlay et al., 2016).

795 For *AR* binding site (ARBS), the peak files were downloaded from two different datasets and
796 converted to hg38 coordinates. For primary prostate cancer, ARBS data were downloaded from
797 GSE70079 (Pomerantz et al., 2015). The union of all tumor sample peaks was used. For mCRPC,
798 met-specfic ARBS data were obtained from a previous study (Pomerantz et al., 2020).

799 ## Identification of SRB regions

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

801 The human genome was divided into 100 kB non-overlapping bins for detection of significantly
802 recurrent breakpoint regions (SRB). A low-mappability mask was generated for the hg38 genome
803 to screen out out regions that are difficult for variant calling based on short-read sequencing. We
804 adopted procedures from a previous study (Mallick et al., 2016) to construct a mask
805 corresponding to regions with low mappability in the human genome. The unmasked regions were
806 defined as the eligible territories for SRB detection. The 100 kB bins with less than 75% overlap
807 with eligible territories were excluded from the analysis. Below is a list of masked regions included
808 in the low-mappability mask.

809 - Composition mask. This set of masked regions includes regions with low sequence
810 complexity detected by mdust, regions with long homopolymers detected by seqtk,
811 satellite regions annotated by RepeatMasker (Smit, AFA, Hubley, R & Green, P, 2013),
812 and low complexity regions annotated by RepeatMasker.

- Mappability mask. This mask was based on mappability of *k*-mers in the human genome hg38. The value *k* was set to 75 which is half of the read length of WGS data in this study. Each base in the genome was assigned a mappability level, based on the mapping ambiguity of all 75-mers overlapping this specific base. See below for the list of mappability levels.
    - Level 0: all 75-mers overlapping this base could not be mapped to the genome uniquely.
    - Level 1: more than 50% of overlapping 75-mers are not uniquely mapped.
    - Level 2: more than 50% of overlapping 75-mers are uniquely mapped with 1-mismatch hits.
    - Level 3: more than 50% of overlapping 75-mers are uniquely mapped without 1-mismatch hits.

Regions with mappability level 0 and 1 were included in the low-mappability mask.

Generating covariates for regression analysis

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

- Nucleotide composition, including GC content, CpG fraction and TpC fraction per 10 kB non-overlapping bin in the genome.
- Replication timing of LNCaP (data obtained from ENCODE under accession ENCFF995YGM, lifted over from hg19 to hg38) (Davis et al., 2018; ENCODE Project Consortium, 2012).
- DNase I hypersensitive sites (data obtained from ENCODE under accession ENCFF434GSJ, lifted over to hg38).
- Repeats annotated by RepeatMasker, including LINE, SINE, LTR, DNA transposon and simple repeats.
- Heterochromatin regions inferred by ChromHMM (Ernst and Kellis, 2012) with the 18-state model parameters from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015), based LNCaP ChIP-seq data of H3K4me1, H3K4me3, H3K4ac H3K27me3, H3K36me3 and H3K9me3.
- Common fragile sites downloaded from HGNC biomart (Tweedie et al., 2021).

SRB detection

Structural variants from the final call set were used for statistical enrichment of recurrent breakpoints within 100 kB bins using a Gamma-Poisson regression implemented in the package, fish.hook (Imielinski et al., 2017). Breakpoints of SVs were treated independently. The Benjamini-Hochberg procedure was used for multiple testing correction and bins with q-value ≤ 0.1 were determined to be significant. The distances of individual known driver genes to those significant bins were evaluated based on the shortest genomic distance between the gene and bin boundaries, regardless of gene orientations.

Annotation of gene alteration status

Gene alteration by copy-number

Copy-number segments were excluded if their cellular fraction was lower than 0.8, except for those which were determined as copy neutral or copy-number greater than 4. The gene annotation was based on known protein coding genes from GenCode release 30 (GRCh38.p12) (Frankish et al., 2019). For each gene, its copy-number was assigned to the copy-number value and LOH status of the segment that has the largest overlap with it. The gene-level copy-number was normalized based on ploidy of the corresponding sample, with autosomal genes normalized

by the inferred ploidy rounded to nearest integer, and X-linked genes normalized by half such value. Then the copy-number status of each gene was categorized based on the following criteria.

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

Gene alteration by structural variant

Gene coordinates were based on ENSEMBL v33 of hg38 (Howe et al., 2021). Gene body region of one gene was defined as the widest region of all known isoforms collapsed. Gene flanking region was defined as the corresponding two 1 Mb regions next to the gene body region on 5'-end and 3'-end, respectively.

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

AR alteration analysis

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

- Co-amplification. Ploidy normalized copy-number values of both *AR* gene body and enhancer are greater than 1.5.
- Selective *AR* amplification. $FC < -\log2(1.5)$ and enhancer copy-number is less than 1.5.
- Selective enhancer copy gain. $FC > \log2(1.5)$ and *AR* gene body copy-number is less than 1.5.
- Lack of amplification for both. All other cases were considered as no amplification for both regions.

ANCOVA test was used to test if different patterns of *AR* amplification have an impact on *AR* expression. Batch corrected log10(TPM+1) values using ComBat from sva R package (v3.34.0) were used for *AR* expression level. We fit the ANCOVA model using *AR* expression as the response variable, *AR* amplification status as the predictor variable, and ploidy, purity as covariates. The function Anova in the car package (v3.0-5) was used with Type III sum of squares for the model. Post hoc analysis was performed to determine the specific differences among four

903 different *AR* amplification status. The function glht was used within the multcomp package (v1.4-
904 11) in R to perform Tukey's Test for multiple comparisons.

905 **Gene expression**

906 TPM values for a subset of the samples based on linked-read sequencing were obtained from
907 cBioportal (Cerami et al., 2012; Gao et al., 2013). For samples based on short-read sequencing
908 the TPM values were obtained from a previous study (Quigley et al., 2018).

909 **Gene fusion analysis**

910 Fusion status of the main members of the ETS family, including *ERG*, *ETV1*, *ETV4*, *ETV5* and
911 *ELK4* was analyzed. Determination of gene fusion status was based on both DNA and RNA levels.
912 For DNA, structural variants transecting gene body regions were used. SV events were
913 considered supporting gene fusion only if they satisfy the following criteria: (1) the breakpoints of
914 this event must be located within the ETS gene and another protein coding gene, respectively; (2)
915 the orientation of the breakpoint located within the ETS gene must be pointing towards the coding
916 sequence of ETS domain. For RNA, arriba was used to detect fusion transcripts from RNA-seq
917 data (Uhrig et al., 2021). The fusion status was only confirmed if all following conditions were
918 satisfied: (1) the complete ETS domain was included in the fusion product; (2) detection
919 confidence reported by arriba is "high"; (3) coding sequence in the fusion transcript was in sense
920 orientation and no out-of-frame shifts.

921 **SV signature analysis**

922 *Signature extraction and clustering*

923 *De novo* signature extraction was performed on all SV events called by SvABA of the combined
924 cohort using signature.tools.lib (Degasperi et al., 2020) with the recommended settings of 20
925 bootstraps, 200 repeats, the clustering with matching algorithm, the KLD objective function, and
926 RTOL = 0.001. The exposure of one signature in one sample is defined as the median activity of
927 the signature within the sample across all bootstraps. For clustering, the reference signature
928 exposure values for each sample based on short-read sequencing were normalized such that the
929 sum of exposure values per sample is 1, and the normalized exposure values for each signature
930 were mean-centered across all samples. A Euclidean distance matrix was computed and then
931 samples were clustered with the Ward.D2 algorithm using R's hclust function. We chose the
932 number of clusters to be k = 9 based on dendrogram using cutree function in R.

933 *Enrichment of alterations in SV clusters*

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

943 *Survival analysis*

944 Survival data was obtained from (Chen et al., 2019). Survival analyses were conducted using the
945 Kaplan-Meier method with log-rank testing for significance. The function survfit from survival R
946 package was used to perform the analysis.

947 ## SUPPLEMENTAL FIGURE LEGENDS

948 **Figure S1. Recurrent CNA and alteration profiles of most frequently altered genes, related**
949 **to Figure 1.**

950 **(A)** Recurrent copy number gain events in the genome. The frequencies of copy number gain are
951 plotted in red according to their genomic coordinates. Regions with significantly recurring CNA
952 are colored in black. Known driver genes that are within those regions are labeled.

953 **(B)** Recurrent copy number loss events in the genome. The frequencies of copy number loss are
954 plotted in blue with y-axis inverted.

955 **(C)** Alteration profiles of known prostate cancer driver genes. Alterations are categorized into CNA,
956 SNV and SV, with SV being further divided into gene transecting (SV tr.) and gene flanking (SV
957 fl.). The percentages of samples carrying corresponding alterations are shown as stacked bars.
958 All known prostate cancer driver genes were considered and the top 16 genes with overall
959 alteration frequencies above 30% are shown.

960 **Figure S2. Recurrent SV in localized prostate cancer and landscape of ETS fusion in**
961 **mCRPC, related to Figure 2.**

962 **(A)** SRBs detected in the cohort of localized prostate cancers. The criteria for coloring and labeling
963 are the same as Figure 2.

964 **(B)** Translocation events originating from *TTC28*. In the circos plot, *TTC28* is labeled with a
965 vertical bar at the 22q12.1 locus. Translocation events which have breakpoints located within 5
966 kB to the 3'-end of the L1 retrotransposon are visualized as blue arcs.

967 **(C)** Schematics of cumulative counts from intra-chromosomal SV events. Individual SV events
968 are indicated by a grey arc, and colored crosses correspond to breakpoints of each event.

969 **(D)** Expression and fusion status for main genes of the ETS family. The expression values were
970 normalized from TPM to z-score within each gene. Grey boxes indicate expression data are not
971 available. For fusion status, color indicates the data type which was used to call fusion.

972 **Figure S3. Comparison of genomic alterations in disease states, related to Figure 3.**

973 **(A)** Alteration status of paired samples from the same patients before and after treatment. Known
974 prostate driver genes with alteration status in any of the included samples are shown.

975 **(B)** Comparison of rearrangement frequency in different disease states of mCRPC. The known
976 prostate cancer driver genes that were located within 1 Mb to any SRB region are included.

977 **(C)** *AR* splice variants in sample DTB-124-BL. The expression values of all known *AR* exons,
978 including both canonical and cryptic ones, are shown in the top panel. In the bottom panel, the
979 number of reads covering the junction sites of two exons are indicated by weighted arcs.

980 **Figure S4. Signature analysis of SV events, related to Figure 4.**

981 **(A)** Workflow of SV signature analysis. Samples involved in this analysis are described in green
982 boxes. Details of relevant signatures are shown in blue boxes. The steps for obtaining the final 9
983 SV clusters are indicated by numbers.

984 **(B)** Signature exposure based on de novo SV signatures. The exposure values of each sample
985 were normalized such that the sample-wise sum is 1. Samples are ordered alphabetically based
986 on names. The sequencing technology used for each sample is labeled at the bottom.

(C) Similarity between de novo SV signatures and reference signatures. Pairwise cosine similarity between de novo and reference SV signatures is shown.

(D) Comparison of reference signature (RefSig) prevalence between mCRPC and localized prostate cancer. The prevalence value for a signature in mCRPC was computed based on samples harboring at least 5% signature exposure. Localized prostate cancer prevalence values were obtained from signal.mutationalsignatures.com (Degasperi et al., 2020) computed from 199 PCAWG samples.

(E) Kaplan Meier curve of prediction using mutation class of key marker genes. Samples were grouped based on the mutation status of the corresponding marker gene.

(F) Kaplan Meier curve of prediction using SV cluster information. Samples were grouped based on their assignments of the corresponding SV cluster.

SUPPLEMENTAL TABLE LEGENDS

Table S1. Sequencing, clinical and alteration information of all samples involved in this study. Related to Figures 1, 2, 4, S1-3.

(A) Sequencing metrics of all samples based on linked-read sequencing.

(B) Clinical properties and key genomics metrics of the cohort.

(C) Somatic mutation status of 159 prostate cancer drivers in the cohort. Sample and genes were sorted alphabetically. Genes with no detected mutations were left blank.

(D) Significantly mutated genes ($q \leq 0.1$) detected by dN/dS algorithm.

(E) Somatic copy-number alteration status of 159 prostate cancer drivers in the cohort.

(F) Recurrent copy-number alteration peaks detected by GISTIC.

(G) Gene transecting rearrangements of 159 prostate cancer drivers in the cohort. Types of rearrangement events were included.

(H) Gene flanking rearrangements of prostate cancer drivers in the cohort.

(I) TITAN copy number segments for all samples. Columns with "Corrected_*" were used for analysis in this study.

(J) TITAN optimal solutions selected for all samples.

(K) Structural variant calls for all samples. For samples with linked-read data ("CRPC10X"), union set of detected calls from SvABA, GROC-SVS, and Long Ranger are indicated. `SV.Filter` indicate SV events after filtering. `support` contain evidence from various callers; manual curation of events is indicated here. `CN_overlap_type` contain the final SV classification after annotation with copy number information.

(L) TDP status, copy number gain event counts, and median tandem duplication lengths for all samples.

Table S2. Significantly recurrent breakpoint regions and ETS fusion. Related to Figure 2, S2 and S3.

(A) Significantly recurrent breakpoints (SRB) regions ($q \leq 0.1$) in the mCRPC cohort of 143 samples.

(B) Significantly recurrent breakpoints (SRB) regions ($q \leq 0.1$) in the localized prostate cancer cohort of 278 samples.

(C) Fusion status of the ETS family genes. Gene expression was normalized to z-score for each gene. Genes with no detected fusion events or available expression data were left blank.

Table S3. *AR* alteration patterns in the mCRPC cohort. Related to Figure 3 and S3.

Table S4. SV signature in the mCRPC cohort. Related to Figure 4 and S4.

(A) Matrix of cosine similarity with rows representing reference signatures and columns representing *de novo* signatures.

(B) Exposure of all 8 *de novo* signatures in the cohort. Values were not normalized.

(C) Exposure of 8 chosen reference signatures in the cohort. Values were not normalized.

1037 ## REFERENCES

1038 Abida, W., Cheng, M.L., Armenia, J., Middha, S., Autio, K.A., Vargas, H.A., Rathkopf, D., Morris,
1039 M.J., Danila, D.C., Slovin, S.F., et al. (2019). Analysis of the Prevalence of Microsatellite Instability
1040 in Prostate Cancer and Response to Immune Checkpoint Blockade. JAMA Oncol *5*, 471–478.

1041 Abida, W., Patnaik, A., Campbell, D., Shapiro, J., Bryce, A.H., McDermott, R., Sautois, B.,
1042 Vogelzang, N.J., Bambury, R.M., Voog, E., et al. (2020). Rucaparib in Men With Metastatic
1043 Castration-Resistant Prostate Cancer Harboring a BRCA1 or BRCA2 Gene Alteration. J Clin
1044 Oncol *38*, 3763–3772.

1045 Adalsteinsson, V.A., Ha, G., Freeman, S.S., Choudhury, A.D., Stover, D.G., Parsons, H.A.,
1046 Gydush, G., Reed, S.C., Rotem, D., Rhoades, J., et al. (2017). Scalable whole-exome sequencing
1047 of cell-free DNA reveals high concordance with metastatic tumors. Nat. Commun. *8*, 1324.

1048 Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V.,
1049 Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational
1050 processes in human cancer. Nature *500*, 415–421.

1051 Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A.,
1052 Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational
1053 signatures in human cancer. Nature *578*, 94–101.

1054 Armenia, J., Wankowicz, S.A.M., Liu, D., Gao, J., Kundra, R., Reznik, E., Chatila, W.K.,
1055 Chakravarty, D., Han, G.C., Coleman, I., et al. (2018). The long tail of oncogenic drivers in
1056 prostate cancer. Nature Genetics *50*, 645–651.

1057 Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K.,
1058 Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate
1059 cancer genomes. Cell *153*, 666–677.

1060 Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.-P., White,
1061 T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies
1062 recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet *44*, 685–689.

1063 Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A.,
1064 Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional
1065 genomics data sets--update. Nucleic Acids Res *41*, D991-995.

1066 Bert, S.A., Robinson, M.D., Strbenac, D., Statham, A.L., Song, J.Z., Hulf, T., Sutherland, R.L.,
1067 Coolen, M.W., Stirzaker, C., and Clark, S.J. (2013). Regional activation of the cancer genome by
1068 long-range epigenetic remodeling. Cancer Cell *23*, 9–22.

1069 de Bono, J., Mateo, J., Fizazi, K., Saad, F., Shore, N., Sandhu, S., Chi, K.N., Sartor, O., Agarwal,
1070 N., Olmos, D., et al. (2020). Olaparib for Metastatic Castration-Resistant Prostate Cancer. N Engl
1071 J Med *382*, 2091–2102.

1072 de Bono, J.S., Logothetis, C.J., Molina, A., Fizazi, K., North, S., Chu, L., Chi, K.N., Jones, R.J.,
1073 Goodman, O.B., Saad, F., et al. (2011). Abiraterone and Increased Survival in Metastatic Prostate
1074 Cancer. N Engl J Med *364*, 1995–2005.

1075 Brand, L.J., and Dehm, S.M. (2013). Androgen Receptor Gene Rearrangements: New
1076 Perspectives on Prostate Cancer Progression. Curr Drug Targets *14*, 441–449.

1077 Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-
1078 Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes.
1079 Nature *578*, 82–93.

1080 Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate
1081 Cancer. Cell *163*, 1011–1025.

1082 Carrot-Zhang, J., and Majewski, J. (2017). LoLoPicker: detecting low allelic-fraction variants from
1083 low-quality cancer samples. Oncotarget *8*, 37032.

1084 Céraline, J., Cruchant, M.D., Erdmann, E., Erbs, P., Kurtz, J.-E., Duclos, B., Jacqmin, D., Chopin,
1085 D., and Bergerat, J.-P. (2004). Constitutive activation of the androgen receptor by a point mutation
1086 in the hinge region: a new mechanism for androgen-independent growth in prostate cancer. Int.
1087 J. Cancer *108*, 152–157.

1088 Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne,
1089 C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An Open Platform
1090 for Exploring Multidimensional Cancer Genomics Data: Figure 1. Cancer Discovery *2*, 401–404.

1091 Chen, C.D., Welsbie, D.S., Tran, C., Baek, S.H., Chen, R., Vessella, R., Rosenfeld, M.G., and
1092 Sawyers, C.L. (2004). Molecular determinants of resistance to antiandrogen therapy. Nature
1093 Medicine *10*, 33–39.

1094 Chen, J., Lui, W.-O., Vos, M.D., Clark, G.J., Takahashi, M., Schoumans, J., Khoo, S.K., Petillo,
1095 D., Lavery, T., Sugimura, J., et al. (2003). The t(1;3) breakpoint-spanning genes LSAMP and
1096 NORE1 are involved in clear cell renal cell carcinomas. Cancer Cell *4*, 405–413.

1097 Chen, W.S., Aggarwal, R., Zhang, L., Zhao, S.G., Thomas, G.V., Beer, T.M., Quigley, D.A., Foye,
1098 A., Playdle, D., Huang, J., et al. (2019). Genomic Drivers of Poor Prognosis and Enzalutamide
1099 Resistance in Metastatic Castration-resistant Prostate Cancer. Eur Urol *76*, 562–571.

1100 Cortés-Ciriano, I., Lee, J.J.-K., Xi, R., Jain, D., Jung, Y.L., Yang, L., Gordenin, D., Klimczak, L.J.,
1101 Zhang, C.-Z., Pellman, D.S., et al. (2020). Comprehensive analysis of chromothripsis in 2,658
1102 human cancers using whole-genome sequencing. Nat Genet *52*, 331–341.

1103 Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain,
1104 K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements
1105 (ENCODE): data portal update. Nucleic Acids Res *46*, D794–D801.

1106 Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., Glodzik, D., Morganella, S.,
1107 Nanda, A.S., Badja, C., Koh, G., et al. (2020). A practical framework and online tool for mutational
1108 signature analyses show inter-tissue variation and driver dependencies. Nat Cancer *1*, 249–263.

1109 van Dessel, L.F., van Riet, J., Smits, M., Zhu, Y., Hamberg, P., van der Heijden, M.S., Bergman,
1110 A.M., van Oort, I.M., de Wit, R., Voest, E.E., et al. (2019). The genomic landscape of metastatic
1111 castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical
1112 impact. Nat Commun *10*, 5251.

1113 Du, Q., Bert, S.A., Armstrong, N.J., Caldon, C.E., Song, J.Z., Nair, S.S., Gould, C.M., Luu, P.-L.,
1114 Peters, T., Khoury, A., et al. (2019). Replication timing and epigenome remodelling are associated
1115 with the nature of chromosomal rearrangements in cancer. Nat Commun *10*, 416.

1116 ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human
1117 genome. Nature *489*, 57–74.

1118 Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and
1119 characterization. Nat Methods *9*, 215–216.

1120 Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M.,
1121 Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human
1122 and mouse genomes. Nucleic Acids Res *47*, D766–D773.

1123 Fraser, M., Livingstone, J., Wrana, J.L., Finelli, A., He, H.H., van der Kwast, T., Zlotta, A.R.,
1124 Bristow, R.G., and Boutros, P.C. (2021). Somatic driver mutation prevalence in 1844 prostate
1125 cancers identifies ZNRF3 loss as a predictor of metastatic relapse. Nat Commun *12*, 6248.

1126 Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A.,
1127 Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical
1128 profiles using the cBioPortal. Sci Signal *6*, pl1.

1129 Giambartolomei, C., Seo, J.-H., Schwarz, T., Freund, M.K., Johnson, R.D., Spisak, S., Baca, S.C.,
1130 Gusev, A., Mancuso, N., Pasaniuc, B., et al. (2021). H3K27ac HiChIP in prostate cell lines
1131 identifies risk genes for prostate cancer susceptibility. The American Journal of Human Genetics
1132 *108*, 2284–2300.

1133 Glodzik, D., Morganella, S., Davies, H., Simpson, P.T., Li, Y., Zou, X., Diez-Perez, J., Staaf, J.,
1134 Alexandrov, L.B., Smid, M., et al. (2017). A somatic-mutational process recurrently duplicates
1135 germline susceptibility loci and tissue-specific super-enhancers in breast cancers. Nat Genet *49*,
1136 341–348.

1137 Grasso, C.S., Wu, Y.-M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J.,
1138 Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-
1139 resistant prostate cancer. Nature *487*, 239–243.

1140 Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L.M., Melnyk, N., McPherson, A.,
1141 Bashashati, A., Laks, E., et al. (2014). TITAN: inference of copy number architectures in clonal
1142 cell populations from tumor whole-genome sequence data. Genome Res *24*, 1881–1893.

1143 Hadi, K., Yao, X., Behr, J.M., Deshpande, A., Xanthopoulakis, C., Tian, H., Kudman, S., Rosiene,
1144 J., Darmofal, M., DeRose, J., et al. (2020). Distinct Classes of Complex Structural Variation
1145 Uncovered across Thousands of Cancer Genome Graphs. Cell *183*, 197-210.e32.

1146 Haffner, M.C., Aryee, M.J., Toubaji, A., Esopi, D.M., Albadine, R., Gurel, B., Isaacs, W.B., Bova,
1147 G.S., Liu, W., Xu, J., et al. (2010). Androgen-induced TOP2B-mediated double-strand breaks and
1148 prostate cancer gene rearrangements. Nat Genet *42*, 668–675.

1149 Henzler, C., Li, Y., Yang, R., McBride, T., Ho, Y., Sprenger, C., Liu, G., Coleman, I., Lakely, B.,
1150 Li, R., et al. (2016). Truncation and constitutive activation of the androgen receptor by diverse
1151 genomic rearrangements in prostate cancer. Nat Commun *7*, 13668.

1152 Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M.,
1153 Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res *49*, D884–D891.

1154 Imielinski, M., Guo, G., and Meyerson, M. (2017). Insertions and Deletions Target Lineage-
1155 Defining Genes in Human Cancers. Cell *168*, 460-472.e14.

1156 Kanayama, M., Lu, C., Luo, J., and Antonarakis, E.S. (2021). AR Splicing Variants and Resistance
1157 to AR Targeting Agents. Cancers (Basel) *13*, 2563.

1158 Kresse, S.H., Ohnstad, H.O., Paulsen, E.B., Bjerkehagen, B., Szuhai, K., Serra, M., Schaefer, K.-
1159 L., Myklebost, O., and Meza-Zepeda, L.A. (2009). LSAMP, a novel candidate tumor suppressor
1160 gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes
1161 Chromosomes Cancer *48*, 679–693.

1162 Kühn, M.W.M., Radtke, I., Bullinger, L., Goorha, S., Cheng, J., Edelmann, J., Gohlke, J., Su, X.,
1163 Paschka, P., Pounds, S., et al. (2012). High-resolution genomic profiling of adult and pediatric
1164 core-binding factor acute myeloid leukemia reveals new recurrent genomic alterations. Blood *119*,
1165 e67.

1166 Kumar-Sinha, C., Kalyana-Sundaram, S., and Chinnaiyan, A.M. (2015). Landscape of gene
1167 fusions in epithelial cancers: seq and ye shall find. Genome Medicine *7*, 129.

1168 Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database
1169 Collaboration (2011). The sequence read archive. Nucleic Acids Res *39*, D19-21.

1170 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
1171 transform. Bioinformatics *25*, 1754–1760.

1172 Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak,
1173 S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer
1174 genomes. Nature *578*, 112–121.

1175 Macintyre, G., Goranova, T.E., De Silva, D., Ennis, D., Piskorz, A.M., Eldridge, M., Sie, D.,
1176 Lewsley, L.-A., Hanif, A., Wilson, C., et al. (2018). Copy number signatures and mutational
1177 processes in ovarian carcinoma. Nat Genet *50*, 1262–1270.

1178 Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N.,
1179 Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes
1180 from 142 diverse populations. Nature *538*, 201–206.

1181 Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H.,
1182 Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic
1183 Tissues. Cell *171*, 1029-1041.e21.

1184 Mateo, J., McKay, R., Abida, W., Aggarwal, R., Alumkal, J., Alva, A., Feng, F., Gao, X., Graff, J.,
1185 Hussain, M., et al. (2020). Accelerating precision medicine in metastatic prostate cancer. Nat
1186 Cancer *1*, 1041–1053.

1187 Nguyen, B., Mota, J.M., Nandakumar, S., Stopsack, K.H., Weg, E., Rathkopf, D., Morris, M.J.,
1188 Scher, H.I., Kantoff, P.W., Gopalan, A., et al. (2020). Pan-cancer Analysis of CDK12 Alterations

1189 Identifies a Subset of Prostate Cancers with Distinct Genomic and Clinical Characteristics.
1190 European Urology *78*, 671–679.

1191 Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I.,
1192 Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560
1193 breast cancer whole-genome sequences. Nature *534*, 47–54.

1194 Petrovics, G., Li, H., Stümpel, T., Tan, S.-H., Young, D., Katta, S., Li, Q., Ying, K., Klocke, B.,
1195 Ravindranath, L., et al. (2015). A novel genomic alteration of LSAMP associates with aggressive
1196 prostate cancer in African American men. EBioMedicine *2*, 1957.

1197 Pitkänen, E., Cajuso, T., Katainen, R., Kaasinen, E., Välimäki, N., Palin, K., Taipale, J., Aaltonen,
1198 L.A., and Kilpivaara, O. (2014). Frequent L1 retrotranspositions originating from TTC28 in
1199 colorectal cancer. Oncotarget *5*, 853–859.

1200 Pomerantz, M.M., Li, F., Takeda, D.Y., Lenci, R., Chonkar, A., Chabot, M., Cejas, P., Vazquez,
1201 F., Cook, J., Shivdasani, R.A., et al. (2015). The androgen receptor cistrome is extensively
1202 reprogrammed in human prostate tumorigenesis. Nat Genet *47*, 1346–1351.

1203 Pomerantz, M.M., Qiu, X., Zhu, Y., Takeda, D.Y., Pan, W., Baca, S.C., Gusev, A., Korthauer, K.D.,
1204 Severson, T.M., Ha, G., et al. (2020). Prostate cancer reactivates developmental epigenomic
1205 programs during metastatic progression. Nat. Genet.

1206 Pradhan, B., Cajuso, T., Katainen, R., Sulo, P., Tanskanen, T., Kilpivaara, O., Pitkänen, E.,
1207 Aaltonen, L.A., Kauppi, L., and Palin, K. (2017). Detection of subclonal L1 transductions in
1208 colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. Sci Rep *7*, 14521.

1209 Pritchard, C.C., Mateo, J., Walsh, M.F., De Sarkar, N., Abida, W., Beltran, H., Garofalo, A., Gulati,
1210 R., Carreira, S., Eeles, R., et al. (2016). Inherited DNA-Repair Gene Mutations in Men with
1211 Metastatic Prostate Cancer. New England Journal of Medicine *375*, 443–453.

1212 Qin, F., Zhang, Y., Liu, J., and Li, H. (2017). SLC45A3-ELK4 functions as a long non-coding
1213 chimeric RNA. Cancer Lett *404*, 53–61.

1214 Quigley, D.A., Dang, H.X., Zhao, S.G., Lloyd, P., Aggarwal, R., Alumkal, J.J., Foye, A., Kothari,
1215 V., Perry, M.D., Bailey, A.M., et al. (2018). Genomic Hallmarks and Structural Variation in
1216 Metastatic Prostate Cancer. Cell *174*, 758-769.e9.

1217 Rickman, D.S., Pflueger, D., Moss, B., VanDoren, V.E., Chen, C.X., de la Taille, A., Kuefer, R.,
1218 Tewari, A.K., Setlur, S.R., Demichelis, F., et al. (2009). SLC45A3-ELK4 is a novel and frequent
1219 erythroblast transformation-specific fusion transcript in prostate cancer. Cancer Res *69*, 2734–
1220 2738.

1221 Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A.,
1222 Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of
1223 111 reference human epigenomes. Nature *518*, 317–330.

1224 Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012).
1225 Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.
1226 Bioinformatics *28*, 1811–1817.

1227 Scher, H.I., Fizazi, K., Saad, F., Taplin, M.-E., Sternberg, C.N., Miller, K., de Wit, R., Mulders, P.,
1228 Chi, K.N., Shore, N.D., et al. (2012). Increased Survival with Enzalutamide in Prostate Cancer
1229 after Chemotherapy. New England Journal of Medicine *367*, 1187–1197.

1230 Smit, AFA, Hubley, R & Green, P (2013). RepeatMasker Open-4.0.

1231 Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B.,
1232 Batzoglou, S., and Sidow, A. (2017). Genome-wide reconstruction of complex structural variants
1233 using read clouds. Nature Methods.

1234 Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D.,
1235 Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in
1236 a single catastrophic event during cancer development. Cell *144*, 27–40.

1237 Taberlay, P.C., Achinger-Kawecka, J., Lun, A.T.L., Buske, F.A., Sabir, K., Gould, C.M., Zotenko,
1238 E., Bert, S.A., Giles, K.A., Bauer, D.C., et al. (2016). Three-dimensional disorganization of the
1239 cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome
1240 Res *26*, 719–731.

1241 Takeda, D.Y., Spisák, S., Seo, J.-H., Bell, C., O'Connor, E., Korthauer, K., Ribli, D., Csabai, I.,
1242 Solymosi, N., Szállási, Z., et al. (2018). A Somatically Acquired Enhancer of the Androgen
1243 Receptor Is a Noncoding Driver in Advanced Prostate Cancer. Cell *174*, 422-432.e13.

1244 Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole,
1245 C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In
1246 Cancer. Nucleic Acids Res *47*, D941–D947.

1247 Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally,
1248 S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS
1249 transcription factor genes in prostate cancer. Science *310*, 644–648.

1250 Tomlins, S.A., Laxman, B., Dhanasekaran, S.M., Helgeson, B.E., Cao, X., Morris, D.S., Menon,
1251 A., Jing, X., Cao, Q., Han, B., et al. (2007). Distinct classes of chromosomal rearrangements
1252 create oncogenic ETS gene fusions in prostate cancer. Nature *448*, 595–599.

1253 Tubio, J.M.C., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas,
1254 C.P., Zamora, J., Raine, K., et al. (2014). Mobile DNA in cancer. Extensive transduction of
1255 nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

1256 Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, R.L., Yates, B., and Bruford, E.A. (2021).
1257 Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res *49*, D939–D946.

1258 Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U.H.,
1259 Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene
1260 fusions from RNA sequencing data. Genome Res *31*, 448–460.

1261 Umbreit, N.T., Zhang, C.-Z., Lynch, L.D., Blaine, L.J., Cheng, A.M., Tourdot, R., Sun, L.,
1262 Almubarak, H.F., Judge, K., Mitchell, T.J., et al. (2020). Mechanisms generating cancer genome
1263 complexity from a single cell division error. Science *368*, eaba0712.

1264 Van der Auwera, G.A., and O'Connor, B. (2020). Genomics in the Cloud.

1265 Veeriah, S., Brennan, C., Meng, S., Singh, B., Fagin, J.A., Solit, D.B., Paty, P.B., Rohle, D.,
1266 Vivanco, I., Chmielecki, J., et al. (2009). The tyrosine phosphatase PTPRD is a tumor suppressor
1267 that is frequently inactivated and mutated in glioblastoma and other human cancers. PNAS *106*,
1268 9435–9440.

1269 Visakorpi, T., Hyytinen, E., Koivisto, P., Tanner, M., Keinänen, R., Palmberg, C., Palotie, A.,
1270 Tammela, T., Isola, J., and Kallioniemi, O.P. (1995). In vivo amplification of the androgen receptor
1271 gene and progression of human prostate cancer. Nat. Genet. *9*, 401–406.

1272 Viswanathan, S.R., Ha, G., Hoff, A.M., Wala, J.A., Carrot-Zhang, J., Whelan, C.W., Haradhvala,
1273 N.J., Freeman, S.S., Reed, S.C., Rhoades, J., et al. (2018). Structural Alterations Driving
1274 Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. Cell *174*,
1275 433-447.e19.

1276 Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C.,
1277 Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection
1278 of structural variants and indels by local assembly. Genome Res *28*, 581–591.

1279 Wang, S., Li, H., Song, M., Tao, Z., Wu, T., He, Z., Zhao, X., Wu, K., and Liu, X.-S. (2021). Copy
1280 number signature analysis tool and its application in prostate cancer reveals distinct mutational
1281 processes and clinical outcomes. PLoS Genet *17*, e1009557.

1282 Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis
1283 of noncoding regulatory mutations in cancer. Nat Genet *46*, 1160–1165.

1284 Willis, N.A., Frock, R.L., Menghi, F., Duffey, E.E., Panday, A., Camacho, V., Hasty, E.P., Liu, E.T.,
1285 Alt, F.W., and Scully, R. (2017). Mechanism of tandem duplication formation in BRCA1-mutant
1286 cells. Nature.

1287 Wu, Y.-M., Cieślik, M., Lonigro, R.J., Vats, P., Reimers, M.A., Cao, X., Ning, Y., Wang, L., Kunju,
1288 L.P., de Sarkar, N., et al. (2018). Inactivation of CDK12 Delineates a Distinct Immunogenic Class
1289 of Advanced Prostate Cancer. Cell *173*, 1770-1782.e14.

1290 Yuan, X., Cai, C., Chen, S., Chen, S., Yu, Z., and Balk, S.P. (2014). Androgen receptor functions
1291 in castration-resistant prostate cancer and mechanisms of resistance to new agents targeting the
1292 androgen axis. Oncogene *33*, 2815–2825.

1293 Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C.,
1294 Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS).
1295 Genome Biol *9*, R137.

1296 Zhang, Y., Gong, M., Yuan, H., Park, H.G., Frierson, H.F., and Li, H. (2012). Chimeric transcript
1297 generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation. Cancer
1298 Discov *2*, 598–607.

1299 Gene Expression Omnibus: NCBI gene expression and hybridization array data repository -
1300 PubMed.

1301

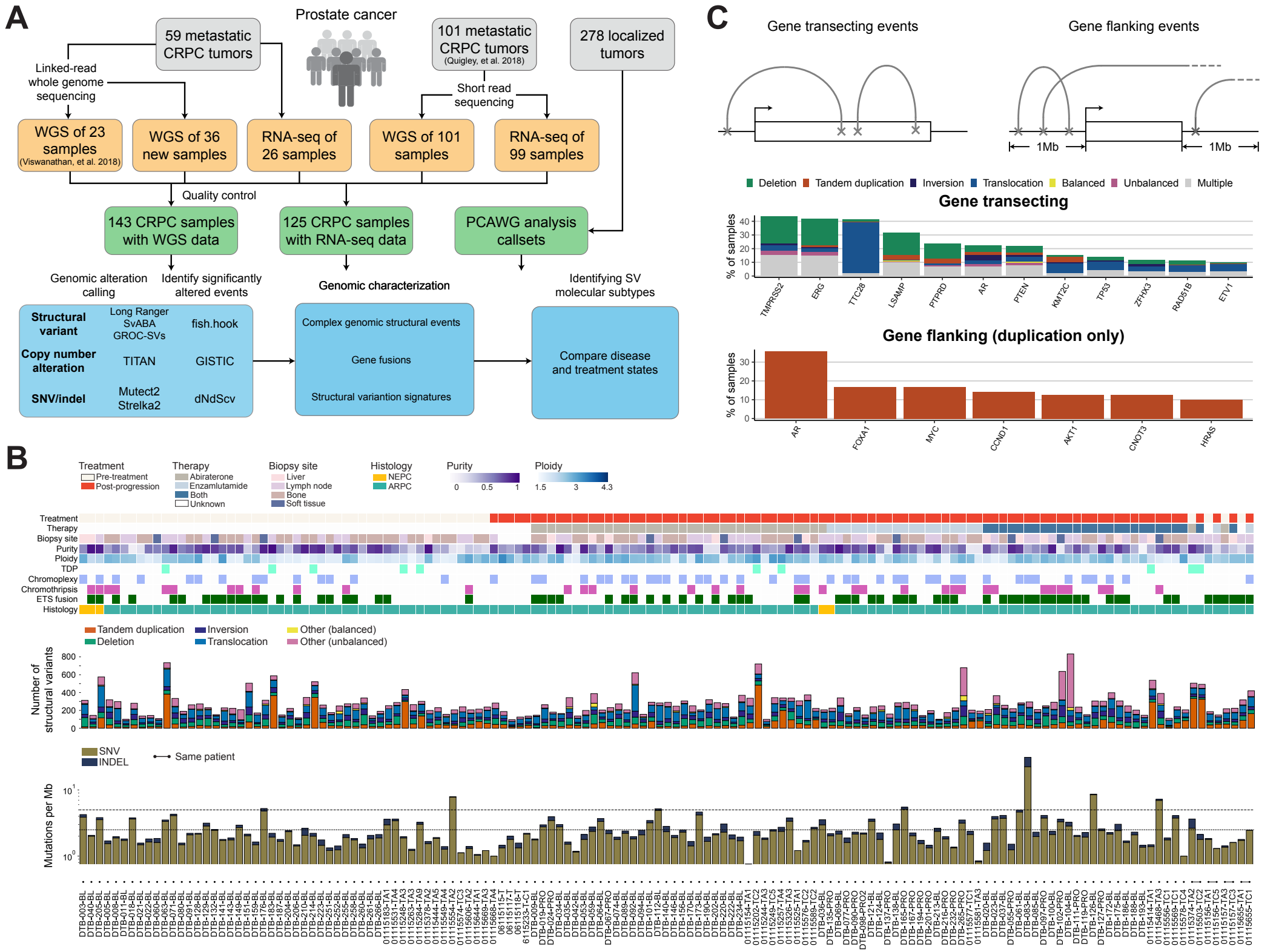
A
Prostate cancer
59 metastatic CRPC tumors
101 metastatic CRPC tumors (Quigley, et al. 2018)
278 localized tumors
Linked-read whole genome sequencing
Short read sequencing
WGS of 23 samples (Viswanathan, et al. 2018)
WGS of 36 new samples
RNA-seq of 26 samples
WGS of 101 samples
RNA-seq of 99 samples
Quality control
143 CRPC samples with WGS data
125 CRPC samples with RNA-seq data
PCAWG analysis callsets
Genomic alteration calling
Identify significantly altered events
Genomic characterization
Identifying SV molecular subtypes
Structural variant
Long Ranger SvABA GROC-SVs
fish.hook
Copy number alteration
TITAN
GISTIC
SNV/indel
Mutect2 Strelka2
dNdScv
Complex genomic structural events
Gene fusions
Structural variantion signatures
Compare disease and treatment states
C
Gene transecting events
Gene flanking events
1Mb
1Mb
Deletion
Tandem duplication
Inversion
Translocation
Balanced
Unbalanced
Multiple
Gene transecting
% of samples
TMPRSS2
ERG
TTC28
LSAMP
PTPRD
AR
PTEN
KMT2C
TP53
ZFHX3
RAD51B
ETV1
Gene flanking (duplication only)
% of samples
AR
FOXA1
MYC
CCND1
AKT1
CNOT3
HRAS
B
Treatment
Pre-treatment
Post-progression
Therapy
Abiraterone
Enzamlutamide
Both
Unknown
Biopsy site
Liver
Lymph node
Bone
Soft tissue
Histology
NEPC
ARPC
Purity
Ploidy
Treatment
Therapy
Biopsy site
Purity
Ploidy
TDP
Chromoplexy
Chromothripsis
ETS fusion
Histology
Tandem duplication
Inversion
Other (balanced)
Deletion
Translocation
Other (unbalanced)
Number of structural variants
SNV
INDEL
Same patient
Mutations per Mb

Figure 1

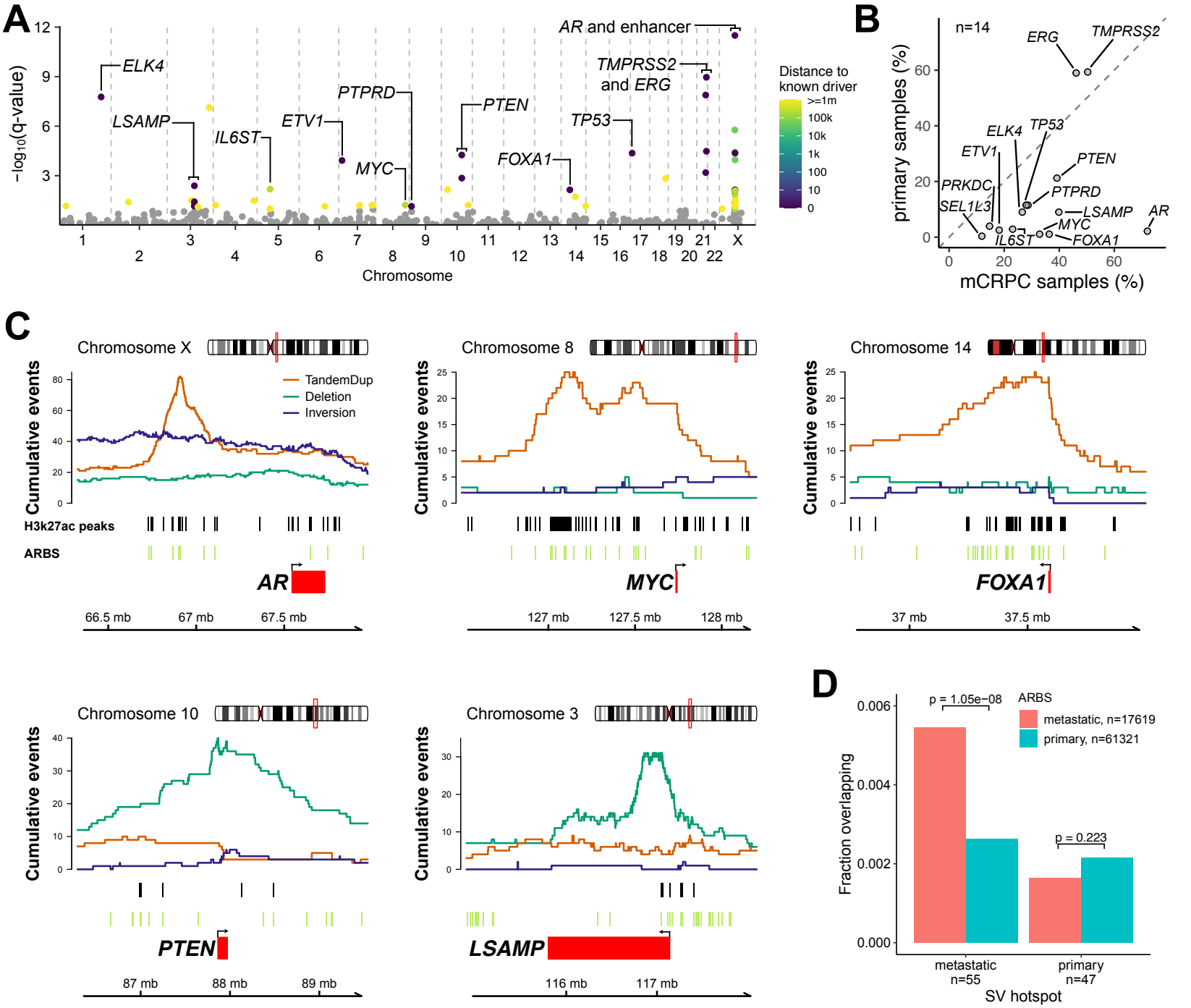
A
−log10(q-value)
Chromosome
ELK4
LSAMP
IL6ST
ETV1
PTPRD
MYC
PTEN
FOXA1
TP53
TMPRSS2 and ERG
AR and enhancer
Distance to known driver
>=1m
100k
10k
1k
100
10
0
B
n=14
primary samples (%)
mCRPC samples (%)
ERG
TMPRSS2
ELK4
TP53
ETV1
PTEN
PRKDC
PTPRD
SEL1L3
LSAMP
AR
MYC
IL6ST
FOXA1
C
Chromosome X
Chromosome 8
Chromosome 14
Chromosome 10
Chromosome 3
Cumulative events
TandemDup
Deletion
Inversion
H3k27ac peaks
ARBS
AR
MYC
FOXA1
PTEN
LSAMP
66.5 mb
67 mb
67.5 mb
127 mb
127.5 mb
128 mb
37 mb
37.5 mb
87 mb
88 mb
89 mb
116 mb
117 mb
D
p = 1.05e−08
p = 0.223
ARBS
metastatic, n=17619
primary, n=61321
Fraction overlapping
metastatic n=55
primary n=47
SV hotspot
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
13.9 mb
13.95 mb
14 mb
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

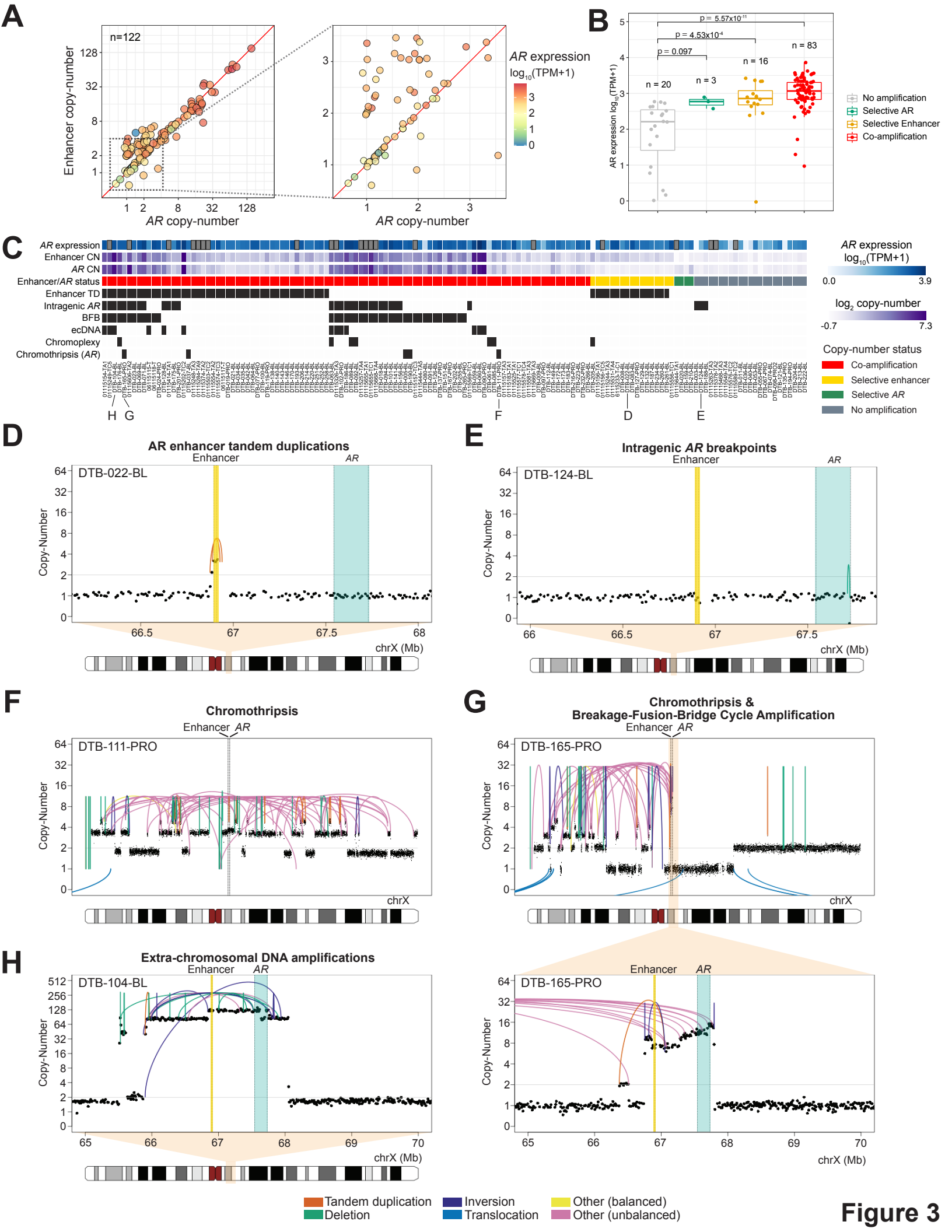
A
n=122
Enhancer copy-number
AR copy-number
AR expression
log10(TPM+1)
B
p = 5.57x10-11
p = 4.53x10-4
p = 0.097
n = 20
n = 3
n = 16
n = 83
AR expression log10(TPM+1)
No amplification
Selective AR
Selective Enhancer
Co-amplification
C
AR expression
Enhancer CN
AR CN
Enhancer/AR status
Enhancer TD
Intragenic 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
F
D
E
D
AR enhancer tandem duplications
Enhancer
AR
DTB-022-BL
Copy-Number
chrX (Mb)
E
Intragenic AR breakpoints
Enhancer
AR
DTB-124-BL
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

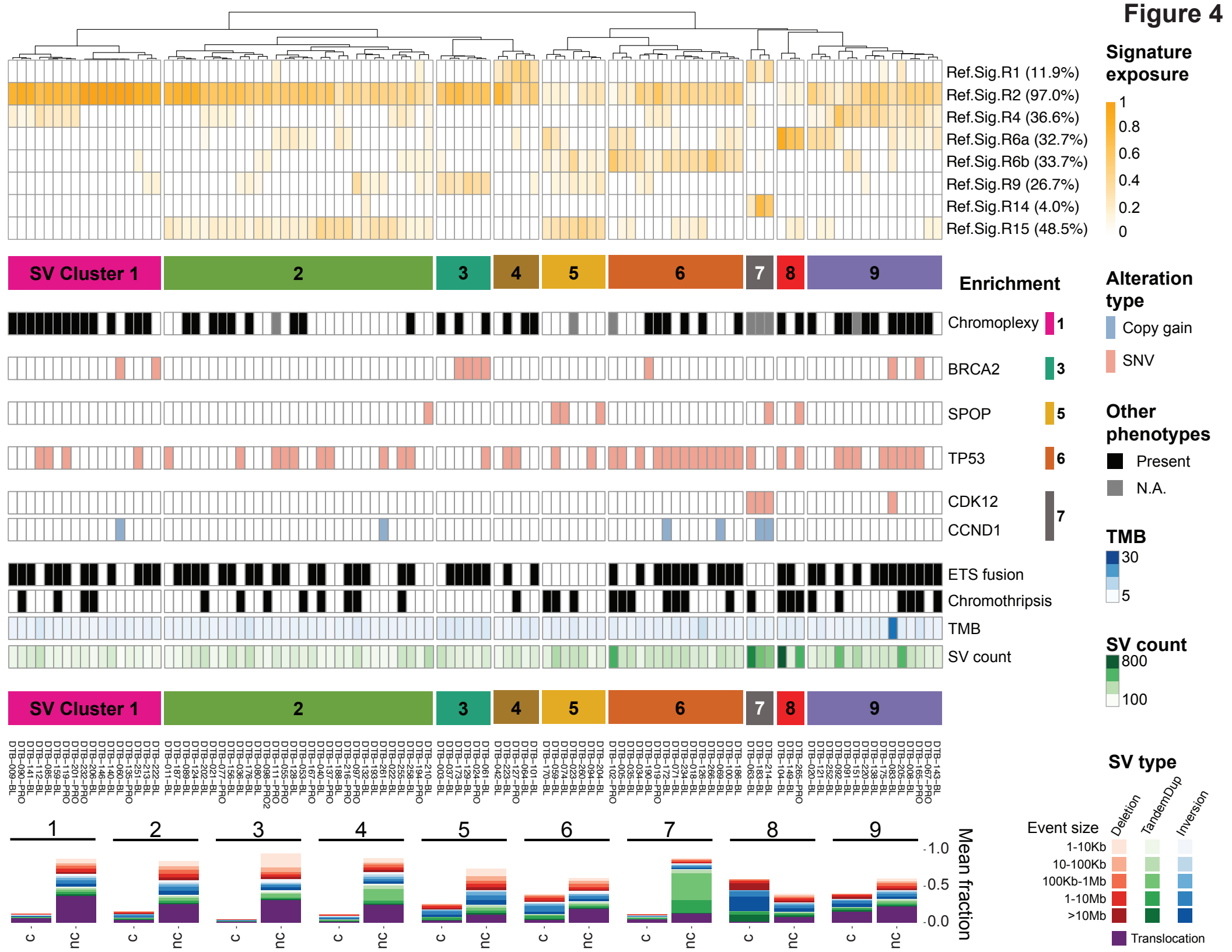
Figure 4
Signature exposure
1
0.8
0.6
0.4
0.2
0
Ref.Sig.R1 (11.9%)
Ref.Sig.R2 (97.0%)
Ref.Sig.R4 (36.6%)
Ref.Sig.R6a (32.7%)
Ref.Sig.R6b (33.7%)
Ref.Sig.R9 (26.7%)
Ref.Sig.R14 (4.0%)
Ref.Sig.R15 (48.5%)
SV Cluster 1
2
3
4
5
6
7
8
9
Enrichment
Chromoplexy 1
BRCA2 3
SPOP 5
TP53 6
CDK12
CCND1
7
ETS fusion
Chromothripsis
TMB
SV count
Alteration type
Copy gain
SNV
Other phenotypes
Present
N.A.
TMB
30
5
SV count
800
100
SV type
Event size
Deletion
TandemDup
Inversion
1-10Kb
10-100Kb
100Kb-1Mb
1-10Mb
>10Mb
Translocation
Mean fraction
1.0
0.5
0.0
c
nc