Genomic alterations in the YAP/TAZ pathway are associated with stem cell-like castration-resistant prostate cancer

Authors: Marjorie L. Roskes[1,2,3,4*], Alexander Martinez-Fundichely[1,2,3,4*], Sandra Cohen[1], Metin Balaban[5], Chen Khuan Wong[6], Weiling Li[1,2,3,4], Tonatiuh A. Gonzalez[1,2,3,4], Anisha B. Tehim[7], Hao Xu[8], Shahd ElNaggar[9], Matthew Myers[5], Rohan Bareja[4], Princesca Dorsaint[4], Kathryn Gorski[4], Muhammad Asad[10], Majd Al Assaad[4,10], Brian D. Robinson[1,4,10], Michael Sigouros[4], Ethan Barnett[11], Jyothi Manohar[4], Scott Tagawa[1,4,12], David Nanus[1,4,12], Ana Molina[1,4,12], Jones T. Nauseef[1,4,12], Cora N. Sternberg[1,4,12], Juan Miguel Mosquera[1,4,10], Howard I. Scher[11], Andrea Sboner[1,4], Benjamin J. Raphael[5], Yu Chen[6,11,12,#], Ekta Khurana[1,2,3,4,#]

[1]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021, USA
[2]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA
[3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10021, USA
[4]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA
[5]Department of Computer Science, Princeton University, New Jersey, USA
[6]Human Oncology and Pathogenesis Program, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[7]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14850, USA
[8]School of Computer Science, McGill University, Montreal, QC, H3A 0G4, Canada
[9]Department of Biology, Barnard College, Columbia University, New York, NY 10027, USA
[10]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY 10065, USA
[11]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
[12]Department of Medicine, Weill Cornell Medicine and New York–Presbyterian Hospital, New York, NY 10065, USA

[*]These authors contributed equally to this work.
[#]Corresponding author. Email: cheny1@mskcc.org; ekk2003@med.cornell.edu

The authors declare no potential conflicts of interest.

Abstract

Castration-resistant prostate cancer (CRPC) is an aggressive disease exhibiting multiple epigenomic subtypes: androgen receptor-dependent CRPC-AR, and lineage plastic subtypes CRPC-SCL (stem cell-like), CRPC-WNT (Wnt-dependent), and CRPC-NE (neuroendocrine). By transcriptomic profiling of tissue, and whole-genome sequencing (WGS) of tissue and cell-free DNA (cfDNA) from 500 samples, we relate genomic variants with epigenomic state. We find lineage plasticity is associated with higher epigenomic and genomic heterogeneity. Samples with CRPC-SCL show higher chromosomal instability. We find DNA alterations, particularly chromosomal rearrangements, in the YAP/TAZ pathway associated with CRPC-SCL. For example, complex rearrangements on chromosome 4, which are supported by patient-matched 3D genome architecture data, decrease promoter interactions of *MOB1B*, a YAP/TAZ pathway inhibitor, with its enhancers. Together, the genomic variants in the pathway can predict CRPC-SCL with 79% accuracy. We show the utility of cfDNA WGS for joint inference of epigenomic state and genomic variants, which can guide patient stratification for clinical decisions.

Significance

This study reveals genomic variants associated with the presence of lineage-plastic CRPC stem cell-like state. We leverage the utility of minimally invasive cfDNA sequencing to obtain genomic and epigenomic insights about CRPC heterogeneity, which have implications for patient stratification for treatment decisions.

Introduction

In the treatment-naïve stage, the growth of prostate cancer is largely dependent on the androgen receptor (AR) pathway. As such, the first line of treatment for advanced prostate cancer is androgen deprivation therapy (ADT). Tumors often develop resistance to ADT and are termed castration-resistant prostate cancer (CRPC). Thus, often next generation AR signaling inhibitors (ARSIs)[1–4] are used in combination with ADT as a first-line therapy. ARSIs work to inhibit AR signaling by different mechanisms. For example, abiraterone inhibits steroidogenesis of adrenal androgens, while enzalutamide and apalutamide inhibit translocation of the AR dimer to the nucleus thereby blocking DNA and cofactor binding[5]. Selective pressure can result in tumors that acquire resistance to ARSIs, either by way of genomic alterations of AR such as amplification of the AR gene or alternative splicing of AR mRNA, or by circumventing dependence on AR altogether[5]. While some of these lineage plastic, AR-independent tumors transdifferentiate into a state with neuroendocrine histology, others remain in the non-AR, non-neuroendocrine state, often referred to as “double-negative” prostate cancer, retaining their adenocarcinoma histology[6,7].

By using chromatin and transcriptomic profiling, we identified four subtypes of CRPC: CRPC-AR which depends on the AR pathway, CRPC-SCL which overexpresses stem cell-like markers, CRPC-WNT which is dependent on the WNT signaling pathway, and CRPC-NE which constitutes the neuroendocrine tumors expressing neuroendocrine gene markers[8]. CRPC-WNT and CRPC-SCL tumors also exhibit enrichment of fibroblast growth factor receptor (FGFR) signaling, consistent with previous studies of double-negative tumors[6]. Moreover, CRPC-SCL tumors also exhibit overexpression of Janus kinase/signal transducer and activator of

transcription (JAK/STAT) signaling, which has been shown to drive lineage plasticity in CRPC[6,9]. Here, we refer to CRPC-SCL, CRPC-WNT, and CRPC-NE as lineage plastic subtypes since they show divergence from the CRPC-AR state.

The identification of these four subtypes in patient tumors is critical since it can offer therapeutic opportunities[8]. While recent efforts have found evidence of intra-patient CRPC subtype heterogeneity, they were limited by small patient cohorts and did not examine the extent of heterogeneity in the context of all these four subtypes[10,11]. The CRPC-SCL subtype in particular has not been characterized well by previous studies. While genomic events have been associated with CRPC-AR (alterations of the AR gene and enhancer), CRPC-NE (loss of RB1), and CRPC-WNT (alterations in the Wnt signaling pathway), no such events have been described for CRPC-SCL[8,12–17].

In this study, we seek to fill these gaps via a combined analysis of genomic events and the four epigenomic subtypes, defined by their gene expression and chromatin accessibility profiles. To do so, we annotate the samples with epigenomic subtypes by quantifying the fractional contribution of each using tissue RNA-seq and/or nucleosomal profiling using cell-free DNA (cfDNA) whole-genome sequencing (WGS). We then relate the genomic events with the epigenomic subtypes by analyzing matched WGS and RNA-seq in the tissue cohort, as well as WGS in the cfDNA cohort.

We find novel genomic alterations at the YAP/TAZ pathway genes that are associated with CRPC-SCL. Because genomic alterations can be difficult to interpret in the absence of other data types, we used matched Hi-C, which reveals three-dimensional (3D) genome architecture, to validate their presence and infer their functional impact[18]. Additionally, we find associations of epigenomic subtypes with genomic instability and heterogeneity, as computed from copy number alterations (CNAs) and structural variations (SVs).

Results

Overview of patient samples used in the study

To perform a combined analysis of genomic events and epigenomic subtypes, we annotate the samples by quantifying the fractional contribution of each CRPC subtype. This is done by using deconvolution approaches on tissue RNA-seq (n=455) and nucleosome profiling from cfDNA WGS (n=68)[8,10,14,15,19,20] (fig. 1A, 2A,B). We then relate the genomic events from WGS with the epigenomic subtypes. This is done by using the patient samples for whom both RNA-seq and WGS was performed on tissue (n=105), as well as the cfDNA whole-genomes[14,21]. Since both genomic variants and epigenomic subtypes are inferred from the same WGS data, the cfDNA whole-genomes in particular offer a unique opportunity to examine their relationship at the same tumoral state and timepoint[15]. Moreover, the use of cfDNA whole-genomes offers the opportunity to monitor the progression of tumors into different epigenomic subtypes with the accompanying genomic changes. Additionally, since prostate cancer genomic alterations are dominated by large SVs[22–25], the use of deep WGS is important to capture the whole spectrum of genomic events as they may relate to chromatin-based subtypes. Due to the limited availability of deep cfDNA WGS of lineage plastic subtypes (n=2 among n=35 patients[15]), we sequenced additional patient samples (n=20) at 40X coverage from Weill Cornell Medicine (WCM) and Memorial Sloan

Kettering (MSK) (fig. 1B). For 7 of these 20 samples, we also had RNA-seq from tissue obtained at the same timepoint as plasma, allowing comparison of molecular subtypes inferred from tissue RNA-seq versus cfDNA WGS. cfDNA whole-genomes from an additional control cohort of healthy participants (n=14) were sequenced to account for the signal from peripheral blood mononuclear cells (PBMCs) in estimating tumoral subtypes from cfDNA (table S1).

Inter- and intra- Sample CRPC subtype heterogeneity in different histological classes

Due to the concordance of the four molecular subtypes identified from chromatin (ATAC-seq) and transcriptomic (RNA-seq) profiling[8], we used their signature genes and accessible chromatin sites to infer their presence in patient samples from RNA-seq and cfDNA WGS, respectively. Signature expression genes for each subtype were defined as those showing subtype-specific expression in cell lines/organoids[8] (tables S2, S3). By comparing expression profiles of subtype-specific signature genes in patient samples to that in cell lines/organoids, we can estimate the fractional contribution of each CRPC subtype (fig. 2A, S1A, S2A) (table S4)[20]. For cfDNA, we leverage the fact that it is enriched for nucleosome-bound regions in tumor tissue[26]. The aggregate coverage profile across cfDNA samples from patient-derived xenografts (PDXs) was computed for subtype-specific accessible chromatin sites. To account for the signal from PBMCs inherent in cfDNA, we generated a panel of no-cancer controls (n=14) (table S5). By comparing coverage profiles in subtype-specific signature sites in patient samples to that of the PDX-derived signature, we can estimate the fractional contribution of each CRPC subtype (fig. 2B, S1B) (tables S6, S7)[10]. The strength of these deconvolution approaches is that they allow assignment of a single sample to a mixture of multiple subtypes.

To validate the assignment of samples to different subtypes, we checked the expression of the top, well-studied subtype-specific marker genes and pathways (table S8)[8]. As expected, we find that *AR* is upregulated in CRPC-AR samples[27] (fig. S3A). Interestingly, we find that *TP53* and *PTEN* are downregulated in lineage plastic samples, the mutational loss of which has been reported to be associated with lineage plasticity and aggressive disease in CRPC[8,17,28] (fig. S3B). Similarly, *RB1* is downregulated while *ASCL1*, *NEUROD1*, *SYP*, and genes in the NEPC pathway are upregulated in CRPC-NE[8,29] (fig S3C). CRPC-WNT samples show overexpression of genes in the WNT pathway[27] (fig. S3D), while CRPC-SCL samples show upregulation of many of the YAP/TAZ pathway target genes (fig. S3E)[8]. Furthermore, we find that CRPC-SCL patients show shorter survival on ARSIs than those without CRPC-SCL, consistent with their dependence on transcription factors (TFs) other than AR (fig. 2C)[8]. Furthermore, to confirm that non-tumor cells do not influence the CRPC subtype fraction estimation, we used single-cell RNA-seq (scRNA-seq) data from CRPC patients that allows annotation of each cell as tumoral or non-tumoral[30]. To compare with our approach, we created pseudo-bulk RNA-seq from only tumor cells and all cells. We then computed the fractional components estimated in pseudo-bulk from all cells versus that from tumor cells only and find high concordance, even in samples with tumor fraction as low as 0.13 (fig. 3A, B). This shows that the signature genes and the deconvolution approach used in our study allows accurate assessment of the subtypes from bulk RNA-seq.

Although samples are generally classified as adenocarcinoma, neuroendocrine, or mixed based on histology clinically, we find that molecularly they are, in fact, quite heterogeneous[31]. While most of the adenocarcinoma samples (96% for tissue and 94% for cfDNA) showed evidence of CRPC-AR as expected, a large majority show some fraction of other molecular subtypes (76%

178 for tissue and 60% for cfDNA). These include 33% with CRPC-SCL, 28% with CRPC-WNT, and
179 28% with CRPC-NE (fig. S2B, C). Similarly, while most neuroendocrine samples show evidence
180 of CRPC-NE (89% for tissue and 70% for cfDNA), many of them also show presence of other
181 subtypes. CRPC-SCL is the most common molecular subtype in adenocarcinoma samples after
182 CRPC-AR, and CRPC-WNT is the most common subtype in neuroendocrine samples after
183 CRPC-NE (fig. S2B, C).
184
185 We find that the samples with the presence of any of the lineage plastic subtypes tend to be
186 more heterogeneous: these samples most often have co-occurrence of multiple CRPC subtypes.
187 This phenomenon is especially true for the CRPC-SCL subtype, which co-occurs with other
188 subtypes in 99% of cases (fig. 3C, D). We compared the ratio of numbers of samples with mixed
189 subtypes to single subtype and call this the subtype mixture index. We find that CRPC-SCL,
190 CRPC-WNT, and CRPC-NE have significantly higher mixture index values than CRPC-AR
191 samples, which exhibit single subtype in 33% of the cases (fig. 3E). Although CRPC-NE has
192 significantly higher subtype mixture index than CRPC-AR, we find that the samples contributing
193 to the higher index are from the adenocarcinoma histology. In other words, when CRPC-NE
194 exists as the only subtype at the molecular level, it has neuroendocrine histology, while its
195 existence in adenocarcinoma samples is in the presence of other subtypes, mostly CRPC-AR
196 (Fisher's Exact Odds Ratio = 18.85, P = 1.21e-10) (fig. S2D). This trend is as expected and
197 shows that histology-based classification may miss the beginning of lineage plasticity, which is
198 captured by our molecular annotation. Thus, molecular annotations, when used together with
199 histological classifications, can have clinical relevance by capturing the transition to lineage
200 plasticity, and beginning of treatment resistance.

201 **Enrichment of subtypes at different metastatic sites**

202 Formation of metastases at distant sites is a key feature of cancer and cancer-related death[32].
203 In order for a metastasis to form, a tumor cell has to invade the blood stream, survive intense
204 combat with the immune system, exit the bloodstream, and colonize at a distant site[33]. Different
205 tumor cells are more prone to seeding in specific distant sites, a phenomenon known as organ
206 tropism, which is largely influenced by the compatibility of the site-specific microenvironment to
207 the tumor cell[33]. Moreover, the tendencies of tumor cells to metastasize to specific sites can be
208 subtype-specific within a given cancer type[34]. While bone is the most common site of prostate
209 cancer metastasis, other sites are also frequently found. In light of this, we next looked for
210 metastatic site enrichment in tissue RNA-seq data across CRPC subtypes to determine if
211 specific CRPC subtypes are more likely to originate from a particular site. To avoid double
212 counting, the samples were annotated as belonging to a subtype if they showed it as their major
213 subtype (>50%). Consistent with previous studies, we find that CRPC-NE samples are more
214 likely to be from liver (Fisher's Exact Odds Ratio = 5.45, P = 6e-6, fig. 4) and soft tissue (Fisher's
215 Exact Odds Ratio = 3.55, P = 7.12e-3, fig. 4)[35,36]. While most samples from all sites are CRPC-
216 AR (fig. S4), as expected, we do find enrichment of this subtype at lymph node (Fisher's Exact
217 Odds Ratio = 3.37, P = 4.97e-7, fig. 4), which is consistent with previous reports[37].
218
219 We find that CRPC-SCL samples are more likely to originate from bone compared to other
220 tissues (Fisher's Exact Odds Ratio = 6.03, P = 2e-6, fig. 4). This is consistent with previous
221 reports that PC3 (a CRPC-SCL cell line) colonizes to bone in mice[38], and evidence of tumor cells
222 with AP-1 signaling (that is characteristic of CRPC-SCL) in scRNA-seq from bone metastases[39].
223 A mechanistic study to examine how this occurs, found that protein regulator of cytokinesis 1

(PRC1), which has been implicated in promoting stemness in cancer, is elevated in prostate cancer cells with stem cell traits and drives colonization to the bone[38]. Moreover, it is well known that bone provides a unique niche for stem cells, in part because of its hypoxic microenvironment that promotes commitment to stemness[40]. Thus, we hypothesize that it is the bone microenvironment that makes it especially hospitable to CRPC-SCL, which is characterized by stem cell-like features.

Genomic alterations of the YAP/TAZ pathway associated with epigenomic subtypes

We identified genome-wide copy number alterations (CNAs), structural variants (SVs), and single nucleotide variants (SNVs) from WGS data from cfDNA and tissue. When matched WGS from cfDNA and tissue was available, we observe a high concordance of both type and breakpoints (for SVs) between called variants, showing the reliability of calling genomic alterations from cfDNA (fig. S5). Moreover, we observe multiple events that are consistent with previous studies. Samples with CRPC-AR are enriched for amplifications of the *AR* gene (fold change = 2; binomial p value = 2.55e-5; fig. 5B, S6B); those with CRPC-NE are enriched for biallelic inactivation of *RB1*[13,17,28] (fold change = 5.21; binomial p value = 1.41e-02, fig. S6C); and *CTNNB1* (β-catenin) mutations are enriched in CRPC-WNT samples (6.5% for CRPC-WNT vs. zero for others; binomial p value = 0; fig. S6D) (tables S9, S10, S11, S12, S13).

The chromatin accessibility landscape of CRPC-SCL has been shown to be enriched for binding sites of the AP-1 and TEAD families of TFs, which work together to promote the oncogenic growth of CRPC-SCL tumors[8]. Activation of the TEAD family is dependent on the YAP/TAZ pathway (fig. 5A). At baseline, the YAP/TAZ pathway is turned off by proteins such as TAOK1/2/3, NF2, WWC1, FRMD6, SAV1, STK3/4, MOB1A/B, and LATS1/2 which work to phosphorylate, and thereby deactivate YAP1 and TAZ (an alias for WWTR1). In the absence of this deactivation mechanism, YAP1 and TAZ are in their active form, and translocate to the nucleus to bind to and activate TEAD, which in turn likely works with AP-1 to regulate the transition to CRPC-SCL[8,41]. Furthermore, YAP/TAZ target genes are upregulated in the context of CRPC-SCL[8] (fig. S3E).

So far there have been no reports of CRPC-SCL-specific genomic alterations. YAP/TAZ pathway alterations have been reported in many other cancers, and are most prevalent in mesothelioma, kidney renal papillary cell carcinoma, and cervical squamous cell carcinoma[41]. In light of this, we sought to leverage our unique dataset to describe the genomic alterations observed in CRPC. Indeed, we find alterations of certain YAP/TAZ pathway genes enriched in samples with CRPC-SCL (fig. 5B, 5C). These include complex rearrangements of chromosome 4 spanning the YAP/TAZ pathway inhibitor *MOB1B* (fold change = 5.89, binomial p value = 1.10e-05, figs. 4B,C, S6E), and stopgain mutations in another inhibitor, *LATS1* (7.4% for CRPC-SCL vs. zero for others, binomial p value = 0, figs. 5B,C). As opposed to pathway inhibitors, we might expect activators to not be disabled in CRPC-SCL samples, and indeed we observe a depletion of a complex rearrangement of chromosome 3 spanning the YAP/TAZ pathway activator *TAZ* in these samples (fold change = 2.55, binomial p value = 4.60e-02, fig. 5B,C, S6E). In our analysis, complex rearrangements are a type of SVs defined as large events that span the entire gene or multiple genes. They are predicted to be either end-to-end telomere translocations or pericentric inversions of the region and are supported by discordant reads at the breakpoints upon manual

269 inspection[41] (fig. S7). We hypothesize that these events may have functional implications in
270 regulating the genes, and thereby, the YAP/TAZ pathway.
271
272 A major advantage of using cfDNA is that it can be used for serial monitoring made feasible by
273 its simple and minimally invasive manner of collection. Similar to the previous observation of
274 emergence of biallelic inactivation of *RB1* accompanying the emergence of CRPC-NE in a
275 patient (AE-180, fig. 5B)[15], we are able to track the co-evolution of *YAP/TAZ* pathway alterations
276 with CRPC subtype using cfDNA. The sample AE-129 began with presence of CRPC-SCL but
277 showed reduction of this subtype at progression on ARSIs accompanied with a complex
278 rearrangement at chromosome 3 spanning the *TAZ* gene, potentially inactivating the YAP/TAZ
279 pathway (fig. 5B,D). Since most prostate tumors begin as AR dependent, this example where
280 we see reduction of CRPC-SCL and increase of CRPC-AR potentially also points to the highly
281 dynamic nature of lineage plasticity where patients may shift back and forth between subtypes
282 and highlights the power of using cfDNA to detect such transitions.
283

284 **Hi-C data validates complex rearrangements and shows they reduce access of *MOB1B*
285 *and TAZ* to their cis-regulatory elements**

286 Large-scale chromosomal alterations have been implicated in many cancers. Loss of
287 heterozygosity of chr8p frequently occurs in breast and prostate cancers[42,43]. Amplification of
288 chr20q is associated with many cancer types such as pancreatic, gastric, colon, ovarian,
289 cervical, and breast cancers[44]. Arm level deletion of chr13q, which includes the *RB1* gene, is
290 associated with neuroendocrine prostate cancer[45]. The CRPC-SCL-enriched/depleted complex
291 rearrangements identified in this study also represent large-scale chromosomal
292 rearrangements. For example, the complex rearrangements spanning *MOB1B*, ranging from
293 9Mb to 190Mb, cover ~48% of the chromosome on average (fig. 6B).
294
295 Such complex SVs are characteristic of prostate cancer genomes[14,22,46], however it is
296 generally hard to confidently identify true positive events using short-read sequencing alone.
297 Thus, we leveraged the matched Hi-C and tissue RNA-seq data (n=55) to evaluate 3D
298 chromosome conformation of each sample together with its molecular state[18,47]. In the Hi-C
299 contact map, topologically associated domains (TADs) manifest as large genomic regions with
300 elevated internal 3D conformation interactions, whereas SVs manifest as interactions isolated
301 to the breakpoints and their immediate surrounding regions only[18]. Thus, Hi-C data can be
302 used to validate the SVs, besides inferring their impact on chromatin conformation[48].
303
304 We first sought evidence for the complex rearrangements on chromosome 4 spanning *MOB1B*
305 and enriched in CRPC-SCL samples from matched Hi-C data. Towards this end, we looked at
306 the contact strength in the Hi-C map at the locus that stores the interaction between the two
307 breakpoints of the complex rearrangement, referred to as the breakpoint junction (fig. 6A).
308 Assuming the event is true, we expect increased interactions at the breakpoint junction relative
309 to surrounding regions, since regions that are far away from each other in the reference
310 genome are now close by in the tumor genome due to the rearrangement. Indeed, we observe
311 higher contact strength at the breakpoint junction and a reduction of contact strength in its
312 surrounding bins, whereas randomly selected control regions of the same length show no such
313 increase (fig. 6A). Similarly, the complex rearrangements enriched in non-CRPC-SCL samples
314 on chromosome 3 spanning *TAZ* are also supported by increased interactions at breakpoint

junctions in those samples (fig. S8A). Furthermore, concordant breakpoints identified separately in matched tissue and cfDNA provide further support for these rearrangements (fig. 6B,C).

We next investigated a potential functional role of the rearrangements in regulating CRPC-SCL. We expected that genes affected by complex rearrangements may have decreased access to their cis-regulatory elements such as enhancers since *MOB1B* shows lower expression in CRPC-SCL samples relative to non-CRPC-SCL (fig. S3E). We analyzed 0.5 Mb region spanning *MOB1B* and find that the strength of 3D contacts, averaged across samples, is lower in CRPC-SCL samples with the rearrangements when compared to those without it (fig. 6B,C,D). By overlaying H3K27ac and ATAC-seq marks with Hi-C contacts, we find that CRPC-SCL samples with the rearrangement show reduced number and contact strength for enhancers connected to *MOB1B* promoter (fig. 6E,F). These trends were reproduced for *TAZ* in non-CRPC-SCL samples (fig. S8). These findings suggest that the complex rearrangements may be associated with *MOB1B* and *TAZ* downregulation at the transcriptional level by reducing promoter access to their enhancers, in addition to their regulation by phosphorylation at the proteomic levels[41]. Even though the breakpoints of rearrangements are far from *MOB1B* and *TAZ* promoters and enhancers, previous studies have shown that large SVs may have a long-distance effect through their impact on the 3D spatial architecture of the chromatin[49]. In line with this, we expect that these rearrangements may also be associated with disruption of other TADs on the chromosome besides the ones surrounding the genes in question. Indeed, by using samples without presence of CRPC-SCL or complex rearrangements as reference, we find that CRPC-SCL samples with the rearrangements have overall fewer conserved TADs than the samples without the rearrangements (fig. S8B).

Logistic regression model to predict CRPC-SCL using genomic alterations

Besides the *MOB1B*, *TAZ*, and *LATS1* variants discussed above, we checked if overall the genomic alterations in the YAP/TAZ pathway are subtype-specific enough to enable identification of CRPC-SCL subtype. Towards this end, we trained a logistic regression model to classify samples containing CPRC-SCL based on alteration events in the YAP/TAZ pathway. The input to the method is the presence or absence of genomic alterations for each gene of the pathway and the output is the probability that the sample exhibits presence of CRPC-SCL subtype. Using Monte Carlo Cross Validation and L1 regularization with a manual parameter search to find the optimal number of features and lambda value for each fold, we classified samples as CRPC-SCL with an accuracy of 0.79 and area under the receiver operating characteristic curve (AUROC) of 0.78, averaged across folds (fig. 5E). These results show that these genomic events are predictive of CRPC-SCL.

CRPC subtypes and genomic heterogeneity and instability

Chromosomal instability and heterogeneity, measured by loss or rearrangement of chromosomes has been reported as a hallmark of cancer and is associated with high disease burden, treatment failure, metastatic progression, and death[50]. Chromosomal instability can be described numerically (i.e. the number of copies of whole chromosomes) or structurally (i.e. the fraction of each chromosome altered)[51,52]. Here, we quantify chromosomal instability as the fraction of the genome altered (FGAlt) by a copy number or structural variant event. We observe

360 that CRPC-SCL samples are enriched for very high FGAlt (fig. 7A,7B). It is possible that high
361 FGAlt may partly contribute to the shorter survival of patients with CRPC-SCL, as it is also
362 associated with aggressive disease in castration-sensitive prostate cancer[53] (fig. 2C).
363
364 We have shown that samples with lineage plastic subtypes tend to be more heterogeneous for
365 the presence of multiple subtypes (fig. 3). Because many genomic events are early cancer
366 events and are associated with different molecular states, we expect genomic and epigenomic
367 heterogeneity to relate. Epigenomic heterogeneity refers to the diversity of CRPC subtype
368 composition within a given sample. Genomic heterogeneity refers to the diversity in tumor clone
369 composition within a given sample. Genomic clones are estimated for each sample by HATCHet,
370 the clone-specific copy number caller we used to estimate genome-wide CNAs[54]. CNAs are
371 especially useful in identifying genomic clones because they occur at large scales, often
372 affecting considerable genomic regions, and can be called in an allele-specific manner. We used
373 clone-specific CNAs to construct a genomic heterogeneity score to account for the number of
374 tumor clones, as well as the similarity between them. The score is defined as $n \times (1 - JI)$, where
375 n is the number of tumor clones, and JI is the Jaccard index between clones. We find that lineage
376 plastic samples also show higher genomic heterogeneity scores, which is positively correlated
377 with the number of epigenomic subtypes (Spearman R = 0.22, P = 2.37e-02, fig. 7A,C). Because
378 just the number of subtypes may not fully capture the epigenomic heterogeneity, we assigned
379 an intra-sample epigenomic heterogeneity score to each sample that measures the diversity of
380 the CRPC fraction composition. The score is defined by the normalized Shannon entropy of the
381 subtype fractions, such that a sample entirely comprised of one CRPC subtype would have a
382 score of 0, and a sample comprised of equal parts of all four CRPC subtypes would have a score
383 of 1. Indeed, we observe that genomic heterogeneity is positively correlated with epigenomic
384 heterogeneity (Spearman R = 0.21, P = 3.36e-02).

385 **Discussion**

386 In this study we annotate the epigenomic subtypes of CRPC patient samples using nucleosomal
387 profiling (cfDNA) and transcriptomic (tissue) data. We find that the histologic classifications of
388 adenocarcinoma and neuroendocrine do not capture the intra- and inter-sample subtype
389 heterogeneity, potentially missing important molecular information relevant for therapeutic
390 decisions. Overall, we find that higher subtype heterogeneity is also associated with higher
391 genomic heterogeneity, and CRPC-SCL samples tend to show higher genomic instability as
392 measured by FGAlt. Enrichment of CRPC-SCL tumors in bone biopsies highlights the
393 importance of studying the tumor immune microenvironment of this subtype in future since bone
394 is the most common site of prostate cancer metastasis. We validate our subtype fraction
395 estimations with scRNA-seq and confirmed that non-tumor cells do not confound our
396 calculations[8,30,39]. By relating the genomic events from WGS with the epigenomic subtypes, we
397 find novel genomic alterations at the YAP/TAZ pathway genes that are associated with the
398 CRPC-SCL subtype. The breakpoints of these rearrangements are validated by Hi-C reads.
399 Indeed, by combining all putative functionally disrupting events in the YAP/TAZ pathway, we are
400 able to predict the presence of CRPC-SCL in a given sample with a high accuracy of 0.79 using
401 a logistic regression model. Previously, few genomic alterations such as *RB1* loss were known
402 to be associated with lineage plasticity in CRPC, and our study adds more genomic events to
403 this list. Interestingly, we find the regression model for CRPC-SCL prediction trained on cfDNA
404 WGS samples outperforms the one trained with tissue samples (RNA-seq and WGS), potentially

405 due to higher fraction of cfDNA samples that exhibit high fraction of CRPC-SCL in cfDNA in our
406 cohort.
407
408 It is well known that alterations affecting non-coding regions of the genome (e.g. enhancers) can
409 play a major role in cancer[55]. In prostate cancer, *AR* enhancer amplifications and fusion of the
410 5' *TMPRSS2* UTR to *ERG* are well studied[14,56]. Large complex rearrangements that span the
411 entire gene or multiple genes have generally been harder to interpret. The availability of matched
412 Hi-C data from patient samples with RNA-seq and WGS allowed us to analyze the enhancer—
413 promoter contacts in samples with versus without rearrangements. Our findings point to potential
414 dysregulation of *MOB1B* by rearrangements on chromosome 4 associated with presence of
415 CRPC-SCL, and of *TAZ* on chromosome 3 associated with absence of CRPC-SCL. While Hi-C
416 data provides support for putative functional roles of these SVs, future studies will be needed to
417 establish the mechanism of their action of gene disruption as well as their role in up- or down-
418 regulation of the YAP/TAZ pathway and lineage plasticity.
419
420 We have shown that both tissue RNA-seq and cfDNA WGS can be used to infer the molecular
421 CRPC subtypes and generally provide consistent results. This is especially impactful since
422 cfDNA is minimally invasive and can be used for serial monitoring. Interestingly, unlike previous
423 observations for genomic heterogeneity where cfDNA exhibits more heterogeneity than tissue,
424 we generally did not observe this trend for epigenomic subtype heterogeneity (fig. S9A)[15]. This
425 observation is also supported by another work using spatial gene expression profiling in which
426 the authors found concordance in gene expression and phenotype classification between
427 metastases from the same individual, albeit with exceptions[11]. However, this observation of
428 absence of higher epigenomic heterogeneity in cfDNA relative to tissue could also be partly due
429 to the limitations of using cfDNA for mixed subtypes inference. Thus, we estimated the limit of
430 detection of the current computational approach to infer CRPC subtype fractions from cfDNA.
431 Under the assumption that each metastatic site sheds cfDNA from molecular subtypes in
432 proportion to their presence in tissue, we find that a subtype fraction of ≤ 0.2 in tissue cannot be
433 detected in cfDNA (fig. S9). These metrics can serve as a guide for the clinical use of cfDNA
434 WGS for molecular subtype inference and guide the further development of computational
435 approaches for subtype inference. Overall, our study provides a framework to move beyond
436 histological classifications of patient tumors to epigenetic subtyping using a clinically feasible
437 assay of cfDNA WGS and shows that it can be used beyond biomarker for discovery of novel
438 genomic alterations associated with lineage plasticity.

439 **Methods**

440 **<u>Cohort Overview</u>**

441 We sequenced cfDNA from 20 patients with CRPC. By combining this with published data, we
442 have amassed a cohort of 488 patients with metastatic CRPC. cfDNA and tissue were collected
443 from 68 and 455 patients, respectively, 23 of which were time matched (fig. 1A) (table
444 S1)[8,14,15,18,19,47].
445
446 We analyzed cfDNA WGS from 95 cfDNA samples from 68 patients. 61 samples were from a
447 previously published study[15], which were enriched for adenocarcinoma patients with only 2
448 patients showing neuroendocrine histology. Thus, we sequenced 20 patient samples from WCM

449 and MSK, out of which 8 showed neuroendocrine histology. For some of these patients, we also
450 generated RNA-seq (n=8) and WGS (n=5) data from matched tissue. Additionally, we
451 sequenced 14 samples from patients with no cancer, as controls. cfDNA samples in Herberts et
452 al. were collected from patients at clinically relevant timepoints, such that patients had either 1
453 (n = 14), 2 (n = 13), or 3 (n = 7) cfDNA samples in total[15]. Tumor fraction ranged from 0.11 to
454 0.92 across samples (mean = 0.46, median = 0.45)[57]. Mean coverage achieved across samples
455 ranged from 8.64X to 182.39X (mean = 87.57X, median = 102.17X) (fig. 1B) (table S1). 32,444
456 structural variants (SVs) were identified across all samples, with a mean of 421.35 SVs per
457 sample (range: 17 – 1,086). 2,165,361 single nucleotide variants (SNVs) were identified across
458 all samples, with a mean of 5,217 non-intergenic, annotated SNVs per sample (range: 15 –
459 59,738).
460
461 Tissue RNA-seq, WGS, and Hi-C datasets from 482, 105, and 55 samples, respectively were
462 analyzed [8,14,18,19,21]. 514, 71, and 327 samples from patients with adenocarcinoma,
463 neuroendocrine, and poorly differentiated/inadequate for diagnosis/not available histologies,
464 respectively. Out of these 123 samples were from WCM and 789 were from other published
465 studies[8,14,18,19,47]. Tissue samples were collected from various metastatic sites including soft
466 tissue (n = 71), lymph node (n = 336), bone (n = 263), bladder (n = 2), prostate (n = 19), lung (n
467 = 15), brain (n = 9), liver (n = 122), skull base (n = 1), ureter (n = 1), adrenal (n = 3), mediastinum
468 (n = 1), spine (n = 1), epidural thoracic (n = 1), and intercostal (n = 1) (table S1).
469

470 **<u>cfDNA WGS</u>**

471 *Patient samples from WCM and MSK*

472 Fresh blood or frozen plasma samples from consented prostate cancer patients were collected
473 under Weill Cornell Medicine Institutional Review Board approval 20-04021962 prior to sample
474 acquisition. Patient blood was collected in Streck tubes and spun at 1600xg for 15 min at 25°C.
475 The plasma was decanted from the top fraction of the Streck tube and cfDNA was isolated
476 immediately if possible, or frozen at -80°C. Plasma cfDNA was isolated using the NeoGeneStar
477 cfDNA variable volume kit according to manufacturer's instructions. For germline DNA, we took
478 the top 2ml of the remaining red blood cell component of the sample from the Streck tube, added
479 4ml RPMI-1640, mixed gently, underlayered with 3ml Histopaque 1.077, spun 400xg for 30 min.,
480 room temperature, with no break, and isolated the peripheral blood mononuclear cells (pbmcs)
481 from the interface. DNA from the pbmcs was isolated using the Qiagen DNeasy Blood and
482 Tissue Kit according to manufacturer's instructions. All samples were quantitated on Qubit HS
483 and frozen at -80°C. Library preparation and WGS were performed at the New York Genome
484 Center using the Kapa Hyper PCR+ cfDNA on the Illumina NovaSeq with 150 paired-end reads
485 and coverages of 0.1, 10, 40 or 80X. Library quality was verified using Agilent Bioanalyzer 2100
486 and Qubit.
487
488 The raw data was aligned to hg38 using BWA mem (version 0.7.15) with parameters -Y -K
489 100000000 -t 8. Base quality scores were adjusted using GATK ApplyBQSR (version 4.1.0.0).

Publicly available patient samples

Raw data was downloaded from the European Genome-Phenome Archive (EGAS00001005783). Fastqs were aligned to hg38 using BWA mem (version 0.7.17) with parameters -M -t 20.

Publicly available PDX samples

Raw data was downloaded from NCBI BioProject (PRJNA900550)[10]. Fastqs were aligned to hg38 and mouse reads were removed using the same pipeline described in De Sarkar and Patton et al[10].

ATAC-seq

Human PBMCs

Fresh PBMCs were collected from a healthy control WCM sample and processed according to the Omni-ATAC protocol in Corces et. al[58]. The partially tagmented samples were sent to the WCM Epigenomics Core for the remainder of the library prep (Illumina) and P2 100 cycles run on NextSeq2000.

Fastqs were aligned to the hg38 reference using BWA mem (version 0.7.15) with parameters -t 6 -M -a -h 100 -L 50,50 and -O 7. Open chromatin peaks were called using genrich (version v0.6.1, https://github.com/jsh58/Genrich) with parameters -j -d 50 -r -e chrM -p 0.01 -q 0.05 -a 20 -g 20 -l 100 -v. IDR (version v2.0.4.2, https://github.com/nboley/idr) was used to get consensus peaks across technical replicates with parameters --output-file-type narrowPeak --rank p.value --plot --use-best-multisummit-IDR --verbose.

Publicly available mouse PDX samples

Publicly available LuCaP ATAC-seq data was downloaded from GEO (GSE156292)[10]. Fastqs were aligned to hg38 using bowtie2 (version 2.4.1) with parameters --wrapper basic-0 -k 10 -t -p 16. Open chromatin peaks were called using genrich (version 0.6.1) with parameters -t -j -v -r -q 0.05 -l 100 and -E to exclude gaps, patches, centromeres, alternative haplotypes, centromeres, and black-listed regions.

RNA-Seq

Publicly available tissue RNA-seq was downloaded directly from Tang et al.[8], https://quigleylab.ucsf.edu/data[14,47], and https://github.com/cBioPortal/datahub/raw/refs/heads/master/public/prad_su2c_2019[19].

Tumor fraction

ichorCNA (version 0.3.2, https://github.com/broadinstitute/ichorCNA) with default parameters was used to estimate tumor fraction from WGS cfDNA[57].

Copy number alterations

cfDNA WGS and matched tissue WGS

HATCHet was used to call copy number alterations from WGS cfDNA and matched tissue (when available)[54]. This applied to the in-house (WCM and MSK) and Herberts, Annala, and Sipola et al[15] cohorts.

530 *Tissue WGS*

531 Copy number alterations were called using CopyCat and downloaded directly from
532 https://quigleylab.s3.us-west-
533 2.amazonaws.com/datasets/2018_04_15_matrix_CN_integer_symbol_copycat.txt.zip. This
534 applied to the SU2C (West Coast) cohort[14].

535 **Structural variant calling**

536 *cfDNA WGS*

537 SVs from publicly available data were retrieved from the authors[15]. To achieve symmetry, we
538 followed the protocol of Herberts et al for our in-house samples (relaxing certain thresholds to
539 match sequencing depth), using Breakfast (version 0.6.0, https://github.com/annalam/breakfast)
540 with parameters --max-frag-len = 1000 and --anchor-len = 30, requiring at least 5 unique
541 junction-spanning reads[15]. The following thresholds were used to further filter the data: at least
542 4 mutant reads, at least 4 reads of total depth, VAF ≥ 0.1, average mapping quality of at least
543 20, and average sidedness (distance to read end) of at least 15.

544 *Tissue WGS*

545 SVs from publicly available data were called using Lumpy and were downloaded directly from
546 https://s3-us-west-2.amazonaws.com/feng.genomics/WGBS/secondary_data.zip[18]. SVs found
547 in more than 4 samples (breakpoint within 10 bps) in the cohort were considered non-unique
548 and removed. SVs that were found in either of two databases of germline variants (dbVar[59] and
549 InvFEST[60]) were removed (breakpoints within 100 bps).

550 **Single nucleotide variant calling**

551 *cfDNA WGS*

552 SNVs from publicly available data were retrieved from the authors[15]. To achieve symmetry, we
553 followed the protocol of Herberts, Annala, Sipola et al for our in-house samples, using mutato
554 (version 0.9.0, https://github.com/annalam/mutato) with arguments --alt-reads = 5 --alt-frac =
555 0.05. Only SNVs with at least 1 supporting read with average mapping quality ≥ 10 and within
556 15 bp of the read end, VAF ≥ 0.1, and a population allele frequency $< 0.5\%$ (from GNOMAD,
557 version 3.0) were retained[15]. The functional impact of the SNVs was annotated using ANNOVAR
558 (version 20191024) with parameters hg38 regGene gene-based annotation[61].

559 *Tissue WGS*

560 SNVs from publicly available data were called using Strelka and were downloaded directly from
561 https://quigleylab.s3.us-west-
562 2.amazonaws.com/datasets/2018_04_15_WCDT_somatic_vcf.tar[14,18]. The functional impact of
563 the SNVs was annotated using ANNOVAR (version 20191024) with parameters hg38 regGene
564 gene-based annotation[61].

565 **CRPC subtype-specific ATAC-seq peaks**

566 ATAC-seq peaks in CRPC cell lines and organoids were called using genrich (version 0.6.1)
567 with parameters -t -d 50 -e chrM -j -v -r -a 20 -g 20 -q 0.05 -l 100 and -E to exclude gaps, patches,
568 centromeres, alternative haplotypes, centromeres, and black-listed regions. IDR (version 2.0.2)
569 was used to get consensus peaks across replicates with parameters --output-file-type
570 narrowPeak --rank p.value --verbose. Differentially accessible ATAC-seq peaks (DAAPs)

generated with macs2 peaks were downloaded from Tang, Xu, Wang, & Wong et al[8]. For every macs2 peak of CRPC subtype A, if there was at least one cell line or organoid of subtype A with an overlapping genrich peak and there were no cell lines or organoids of any other subtype with an overlapping genrich peak, the subtype specific peak was defined by a start (min(start of all genrich peaks of cell lines/organoids of subtype A that overlap the macs2 DAAP)), end (max(end of all genrich peaks of cell lines/organoids of subtype A that overlap the macs2 DAAP)), and summit (int(median(summits of all genrich peaks of cell lines/organoids of subtype A that overlap the macs2 DAAP))). Any peak that overlapped a genrich ATAC-seq peak of our in-house no cancer control sample (ATAC-seq Human PBMCs above) was removed. Overall, there were 11,658 CRPC-AR sites, 13,959 CRPC-SCL sites, 13,732 CRPC-WNT sites, and 11,642 CRPC-NE sites (table S14).

CRPC subtype fraction estimation

cfDNA WGS

We performed deconvolution using Keraon[10]. The signature matrix was built by averaging the WGS cfDNA mean coverage across PDXs of the same CRPC subtype in CRPC subtype-specific accessible chromatin sites of that subtype (table S5). Because we had no CRPC-WNT PDXs, we used a CRPC-WNT patient for the signature and excluded it from other analyses. Features used were central coverage (30 bp window centered at the summit) and window coverage (990 bp window centered at the summit). Signatures were also computed from 14 healthy control samples and were subtracted out of the PDX signature (tables S5,S6). The mixture matrix was the WGS cfDNA mean coverage of each sample in CRPC subtype-specific accessible chromatin sites (fig. S1B).

Tissue RNA-seq

Deconvolution was performed using EPIC[20]. The signature matrix was built based on averaging the expressions (Transcripts Per Million, TPM) across cell lines and organoids of the same CRPC subtype in CRPC signature genes[8] (tables S2,S3). The mixture matrix was the gene expression values (TPM) of each sample for CRPC signature genes[8]. The mixture matrix is then regressed back onto the signature matrix to estimate fractional contributions of each CRPC subtype to each patient's expression profile (fig. S1A).

scRNA-seq analysis

Single-cell RNA-seq count matrices derived from human metastatic CRPC biopsy samples were downloaded directly from GEO (GSE210358)[9]. To perform pseudo-bulking and conversion to TPM, gene counts were summed across all cells in each sample (only common genes across samples were retained). Gene lengths were acquired from a biomaRt query to use in TPM conversion[62]. For tumor-only pseudo-bulking, the same approach was used across only tumor cells (GSM6428952)[9]. To annotate these cells, the authors used a hierarchical strategy wherein broad categories of epithelial, immune, and stromal cells were identified using expression of marker genes for these categories. Expression of genes that are known to be upregulated in prostate cancer are used to identify tumor cells. Cell type assignments were confirmed by showing inferred CNVs were higher in tumor cells relative to non-tumor cells[30].

612 **Metastatic site enrichment**

613 For each CRPC subtype, we performed a two-tailed Fisher's exact test to determine if the odds
614 of samples from that subtype coming from a particular metastatic site versus others were
615 statistically significant. This analysis was performed on the tissue dataset and was limited to
616 samples with CRPC subtype fraction ≥ 0.5.

617 **Genomic event enrichment**

618 For each comparison, we performed a one-tailed binomial test to determine if the probability of
619 samples from a given group having the alteration was significantly different from the probability
620 of other samples having the alteration. We had segment-level copy number calls for cfDNA and
621 gene-level scalar copy number calls for tissue. For AR amplifications, we checked for
622 amplifications (CN>3 in cfDNA, CN>3 in tissue – values chosen based on CN distribution) in
623 CRPC-AR samples versus others. For biallelic inactivation of RB1, we checked for homozygous
624 deletions (CN≤1.2 in cfDNA, CN≤0.5 in tissue – values chosen based on CN distribution) or a
625 combination of at least two of the following events: heterozygous deletions (1.2≤CN<1.75 in
626 cfDNA, 0.5≤CN<1.6 in tissue – values chosen based on CN distribution), inactivation by SV (i.e.
627 SV breakpoint intersecting gene), stopgain, and splice mutations. For CTNNB1 mutations, they
628 were putative driver missense mutations. Only complex rearrangements that were significant in
629 both cfDNA and tissue were reported.

630 **CRPC-SCL prediction using genomic variants**

631 We trained a logistic regression model to classify samples as CRPC-SCL or not, using a 2:1
632 train/test split. We performed this 10 times using different seeds each time. We performed L1
633 regularization with a manual parameter search to find the optimal number of features and lambda
634 value, testing c values within an np.logspace(-4,4,20) for each train/test split. The feature matrix
635 had samples in the rows and YAP/TAZ pathway genes in the columns and was populated by 1
636 if the sample had events deemed putatively functional (a complex rearrangement, missense
637 (putative driver) mutation, or stopgain mutation of the gene), and 0 otherwise. The target classes
638 were 0 if the sample was not CRPC-SCL and 1 if it was. We checked that including other
639 genomic events did not significantly change the performance of the model.
640
641 We also trained models to classify the other CRPC subtypes, but CRPC-SCL performed the best
642 by far. We hypothesize that classification of CRPC-NE (AUC=0.61, ACC=0.65) and CRPC-WNT
643 (AUC=0.64, ACC=0.64) is better than random because many of the samples also contain CRPC-
644 SCL.

645 **Hi-C analysis**

646 Normalized contact maps were downloaded from GEO (GSE249494)[18]. We leverage the
647 information obtained from the Hi-C experiment in two main tasks. First, the predicted breakpoints
648 were validated using Hi-C interactions by searching the nonlinear contact map. We compare the
649 contact strength at the breakpoint junction to its surrounding regions. We define a search space
650 of 10 Hi-C bins (100 Kb) on each side of the breakpoint junctions. For each sample with the
651 complex rearrangement and the Hi-C contact map available, we calculated the contact values
652 in the search space, aggregated across samples. Then, we evaluate the trend of the aggregated
653 contact strength in the search space by fitting a curve using a smoothing window of two Hi-C bins
654 (20 Kb). As a control, we evaluate the same trends in random regions to show that nonlinear
655 interactions at the same distance are only seen due to breakpoints (fig. 6A). Secondly, we

656 assessed the functional effect of the rearrangements on the Hi-C interactions at genomic regions
657 associated with the CRPC-SCL subtype transition and the impact on enhancer-promoter
658 interactions. We evaluated the average strength per contact within 500 Kb TADs harboring the
659 YAP/TAZ inhibitor, MOB1B, using CRPC-SCL samples with and without the rearrangement (fig.
660 6C,D). We reproduced these trends in YAP/TAZ activator, TAZ, in Non-CRPC-SCL samples (fig.
661 S8). Enhancer regulatory regions were then defined by the overlap of in-house generated ATAC-
662 seq peaks[8] and DU145 H3K27ac marks downloaded from
663 https://figshare.com/articles/dataset/H3K27ac_Signal_Track_for_DU145/12442796. Thus, we
664 further verified the specific change in the average strength of enhancer-promoter connections
665 between SCL samples with and without the rearrangements (fig. 6E,F, S6C).

666 **Epigenomic and genomic heterogeneity scores**

667 The epigenomic heterogeneity score is the Shannon entropy of the subtype fractions, normalized
668 by the number of possible CRPC subtypes (i.e. 4) to restrict the range from 0 to 1 (table S15).
669
670 The genomic heterogeneity score is $\#\ genomic\ clones\ \times (1 - Jaccard\ index\ between\ clones)$
671 where the Jaccard index was 1 for samples with only one clone (table S15).

672 **Fraction of genome altered**

673 Fraction of genome altered is the sum of CNA segment lengths with abnormal copy number in
674 at least one clone intersected with SVs (excluding interchromosomal translocations) divided by
675 the length of the genome (table S16). We find that high FGAlt (>0.85) is enriched in CRPC-SCL.
676 This trend is the same across thresholds (fig. S10). We find that CNAs are the largest contributor
677 to FGAlt.

678 **Survival analysis**

679 We generated Kaplan-Meier curves using progression free survival in units of days as the
680 measure of time. Every event was counted as progression. We performed a one-tailed log rank
681 test to determine if the curves were statistically different from one another (fig. 2C).

682 **Limit of detection**

683 We compared CRPC subtype fractions in tissue RNA-seq and cfDNA WGS in patients for whom
684 we have matched data (fig. S9A). We performed in-silico simulations to estimate our limit of
685 detection with respect to the fraction of CRPC subtype present, the tumor fraction of the sample,
686 and the average sequencing coverage of the sample.

687 *Subtype fraction*

688 To estimate the lowest fraction of CRPC subtype detectable from cfDNA, we chose patient AE-
689 180 as a model. At baseline, we estimated he was 100% CRPC-AR (tumor fraction = 0.4718,
690 mean coverage = 111.1X). At progression2, however, he was 100% CRPC-NE (tumor fraction
691 = 0.614, mean coverage = 69.67X). By mixing specific amounts of AE-180-Baseline with AE-
692 180-Progression2 and a healthy control, we were able to create a sample with a target CRPC-
693 AR fraction, while maintaining a fixed coverage (40X) and tumor fraction (0.4718), the tumor
694 fraction of AE-180-Baseline without mixing. For each target CRPC-AR fraction, we repeated the
695 experiment 100 times, each with a different seed, generating a distribution of estimated CRPC-
696 AR fractions across experiments. By assuming the target CRPC-AR fraction as ground truth, we
697 evaluated our statistical and systematic errors in estimating subtype fractions. The statistical

error was never more than 2.1% across target CRPC-AR fractions (fig. S9B). When the CRPC subtype fraction is less than 0.3, we do not detect it at all. When it is close to this limit, our systematic error is high, and we underestimate the contribution of the small subtype.

We compute the fractions of AE-180-Baseline, AE-180-Progression2, and control required for mixing to achieve the desired target CRPC-AR and CRPC-NE fraction as follows: We define three samples S_1, S_2, and S_3 with CRPC-AR fraction F_1, F_2, and F_3, tumor fraction T_1, T_2, and T_3 and coverage C_1, C_2, and C_3 such that $T_1 \neq T_2 \neq T_3$ and $C_1 \neq C_2 \neq C_3$. Let S_3 be the healthy sample such that $T_3 = 0$. Our target mixed sample is S_4, such that $T_4 = \frac{a_1T_1C_1+a_2T_2C_2+a_3T_3C_3}{a_1C_1+a_2C_2+a_3C_3}$ and $C_4 = a_1C_1 + a_2C_2 + a_3C_3$, for some constants a_1, a_2, and a_3. This relies on the assumption that every sample is made up of some (potentially zero) tumor and healthy component such that $S_i = S_i^T + S_i^H$, $T_i = T_i^T + T_i^H$, and $C_i = C_i^T + C_i^H$. Then $T_4 = \frac{a_1T_1C_1+a_2T_2C_2+a_3T_3C_3}{a_1C_1+a_2C_2+a_3C_3} = \frac{C_4^T}{C_4}$. And $a_1C_1 + a_2C_2 + a_3C_3 = C_4$, $a_1T_4C_1 + a_2T_4C_2 + a_3T_4C_3 = a_1T_1C_1 + a_2T_2C_2 + a_3T_3C_3$, and $C_1(T_4 - T_1)a_1 + C_2(T_4 - T_2)a_2 + C_3(T_4 - T_3)a_3 = 0$. Also, $(a_1T_1C_1 + a_2T_2C_2 + a_3T_3C_3)F_4 - (a_1T_1C_1F_1 + a_2T_2C_2F_2 + a_3T_3C_3F_3) = 0$, and $(a_1T_1C_1)(F_4 - F_1) + (a_2T_2C_2)(F_4 - F_2) + (a_3T_3C_3)(F_4 - F_3) = 0$. Thus, we solve the following linear equation to solve for fractions a_1, a_2, and a_3:

$$\begin{bmatrix} C_1 & C_2 & C_3 \\ C_1(T_4 - T_1) & C_2(T_4 - T_2) & C_3(T_4 - T_3) \\ T_1C_1(F_4 - F_1) & T_2C_2(F_4 - F_2) & T_3C_3(F_4 - F_3) \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} C_4 \\ 0 \\ 0 \end{bmatrix}.$$

Statistical error was estimated by the standard deviation of CRPC-AR fraction across 100 runs of a given target (CRPC-AR, CRPC-NE) fraction pair. Systematic error was estimated by the difference between the mean CRPC-AR fraction across 100 runs of a given target (CRPC-AR, CRPC-NE) pair and the actual CRPC-AR fraction of the sample without mixing.

Tumor fraction and coverage

To estimate our ability to accurately detect CRPC subtype fraction from cfDNA with respect to coverage and tumor fraction, we chose sample AE-092-Baseline (coverage = 17.18X, tumor fraction = 0.78), who had a CRPC-SCL fraction of 0.43, as a model. We chose this sample because of its high tumor fraction and relatively high CRPC-SCL fraction. By downsampling and mixing specific amounts of AE-092-Baseline with a healthy control, we were able to create a sample with a target coverage and tumor fraction combination. For each target coverage, tumor fraction pair, we repeated the experiment 100 times, each with a different seed, generating a distribution of estimated CRPC-SCL fractions across experiments. By assuming the CRPC-SCL fraction without downsampling or mixing – i.e. 0.43 – as ground truth, we evaluated our statistical and systematic errors in estimating subtype fractions. Statistical error stems from having fixed, finite coverage and can be estimated by the standard deviation of CRPC-SCL fraction estimates across the 100 experiments. Although the statistical error goes up with lower coverage and tumor fraction, it remains below 5% for coverage down to 3X and tumor fraction down to 0.1. Using these experiments as a template, we estimated the statistical errors on each of our patient samples using their coverage and tumor fraction. On average, the statistical error of the cohort was 1.3% (fig. S9C). We find that the systematic error is mostly independent of coverage but increases with lower tumor fraction and remains around 6% or below for tumor fraction down to 0.1. Using these experiments as a template, we estimated the systematic errors on each of our patient samples using their coverage and tumor fraction. On average, the systematic error was

740 1.5% (fig. S9D). Calculations were performed similarly to estimating limit of detection of subtype
741 fraction above.

742 **Data Availability**

743 Processed data is provided in supplemental tables. Raw data will be made available after
744 publication.

745 **Acknowledgements**

746 We thank the Khurana and Chen lab members for their input and discussion. We are grateful
747 to the Englander Institute for Precision Medicine for WCM CRPC patient samples, including
748 Alexandria Norton, Christina Federov, John Mills, and John Otilano for their facilitation. We
749 also thank the WCM Genomics Core and Epigenomics Core facilities.

750 **Funding**

751 S.EN. thanks the National Institutes of Health (R25CA233208) for their support. M.R. thanks
752 the National Institutes of Health (T32 GM083937) for their support**.** T.A.G. was supported by a
753 Medical Scientist Training Program grant from the National Institute of General Medical
754 Sciences of the National Institutes of Health under award number**:** T32GM152349 to the Weill
755 Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program. E.K. thanks the National
756 Institutes of Health (R01CA218668 and P50CA211024), the Department of Defense (HT9425-
757 23-1-0074), the Ann and William Bresnan Foundation, and the WorldQuant Foundation for
758 their support.

759 **Author Contributions**

760 M.R., A.M.F., and E.K. conceived of and designed the project. M.R., A.M.F., and E.K. wrote
761 the manuscript with the help of all authors. M.R. and A.M.F. performed the majority of the data
762 analysis. S.C. performed cfDNA sample preparation for WGS. M.M., M.B. and B.J.R.
763 redesigned HATCHet for cfDNA WGS. W.L., A.B.T., T.A.G., S.EN., R.B., P.D., and K.G.
764 helped with data analysis. J.M.M., B.D.R., M.A., and M.AA. performed pathological review.
765 M.S., J.M., S.T., D.N., A.M., J.N., C.N.S., J.M., and A.S. managed and coordinated WCM data
766 and sample delivery. E.B., H.I.S., and Y.C., managed and coordinated MSK data and sample
767 delivery. B.J.R. and Y.C. helped supervise the study. E.K. supervised the study.

768 **References**

769 1. Chi, K. N. *et al.* Apalutamide for metastatic, castration-sensitive prostate cancer. *New*
770 *England Journal of Medicine* **381**, (2019).
771 2. Fizazi, K. *et al.* Abiraterone plus Prednisone in Metastatic, Castration-Sensitive Prostate
772 Cancer. *New England Journal of Medicine* **377**, 352–360 (2017).
773 3. Davis, I. D. *et al.* Enzalutamide with Standard First-Line Therapy in Metastatic Prostate
774 Cancer. *New England Journal of Medicine* **381**, 121–131 (2019).
775 4. James, N. D. *et al.* Abiraterone for Prostate Cancer Not Previously Treated with
776 Hormone Therapy. *New England Journal of Medicine* **377**, 338–351 (2017).

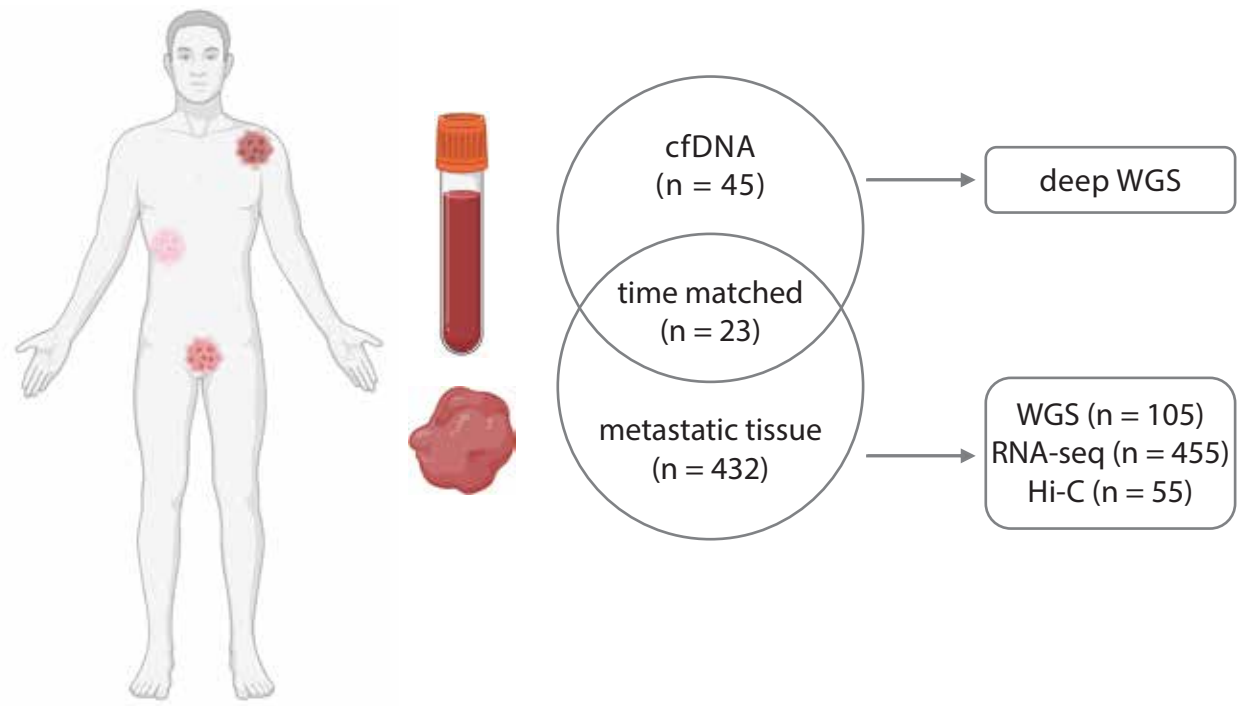
5. Watson, P. A., Arora, V. K. & Sawyers, C. L. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. *Nat Rev Cancer* **15**, 701–711 (2015).
6. Bluemn, E. G. *et al.* Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. *Cancer Cell* **32**, 474-489.e6 (2017).
7. Labrecque, M. P. *et al.* Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. *Journal of Clinical Investigation* **129**, 4492–4505 (2019).
8. Tang, F. *et al.* Chromatin profiles classify castration-resistant prostate cancers suggesting therapeutic targets. *Science (1979)* **376**, (2022).
9. Chan, J. M. *et al.* Lineage plasticity in prostate cancer depends on JAK/STAT inflammatory signaling. *Science (1979)* **377**, (2022).
10. De Sarkar, N. *et al.* Nucleosome Patterns in Circulating Tumor DNA Reveal Transcriptional Regulation of Advanced Prostate Cancer Phenotypes. *Cancer Discov* **13**, 632–653 (2023).
11. Brady, L. *et al.* Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. *Nat Commun* **12**, (2021).
12. Visakorpi, T. *et al.* In vivo amplification of the androgen receptor gene and progression of human prostate cancer. *Nat Genet* **9**, (1995).
13. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447.e19 (2018).
14. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758-769.e9 (2018).
15. Herberts, C. *et al.* Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. *Nature* **608**, 199–208 (2022).
16. Zhou, Z. *et al.* Synergy of p53 and Rb deficiency in a conditional mouse model for metastatic prostate cancer. *Cancer Res* **66**, 7889–7898 (2006).
17. Yu Ku, S. *et al.* Rb1 and Trp53 cooperate to suppress prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. *Science (1979)* **355**, 78–83 (2017).
18. Zhao, S. G. *et al.* Integrated analyses highlight interactions between the three-dimensional genome and DNA, RNA and epigenomic alterations in metastatic prostate cancer. *Nat Genet* **56**, (2024).
19. Abida, W. *et al.* Genomic correlates of clinical outcome in advanced prostate cancer. *PNAS* **116**, (2019).
20. Racle, J. ; G. D. EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data. in *Bioinformatics for Cancer Immunotherapy: Methods and Protocols* (ed. Boegel, S.) 233–248 (Springer US, 2020). doi:10.1007/978-1-0716-0327-7_17.
21. Zhao, S. G. *et al.* The DNA methylation landscape of advanced prostate cancer. *Nat Genet* **52**, 778–789 (2020).
22. Zhou, M. *et al.* Patterns of structural variation define prostate cancer across disease states. *JCI Insight* **7**, (2022).
23. van Dessel, L. F. *et al.* The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nat Commun* **10**, (2019).
24. Khani, F. *et al.* Evolution of structural rearrangements in prostate cancer intracranial metastases. *NPJ Precis Oncol* **7**, (2023).

25. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
26. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).
27. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
28. Mu, P. *et al.* SOX2 promotes lineage plasticity and antiandrogen resistance in TP53-and RB1-deficient prostate cancer. *Science (1979)* **355**, 84–88 (2017).
29. Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* **22**, 298–305 (2016).
30. Zaidi, S. *et al.* Single-cell analysis of treatment-resistant prostate cancer: Implications of cell state changes for cell surface antigen–targeted therapies. *Proc Natl Acad Sci U S A* **121**, (2024).
31. Beltran, H. *et al.* The role of lineage plasticity in prostate cancer therapy resistance. *Clinical Cancer Research* **25**, 6916–6924 (2019).
32. Steeg, P. S. Tumor metastasis: Mechanistic insights and clinical challenges. *Nat Med* **12**, 895–904 (2006).
33. Horak, C. E. & Steeg, P. S. Metastasis gets site specific. *Cancer Cell* **2**, (2005).
34. Chen, W., Hoffmann, A. D., Liu, H. & Liu, X. Organotropism: new insights into molecular mechanisms of breast cancer metastasis. *NPJ Precis Oncol* **2**, (2018).
35. Aggarwal, R., Zhang, T., Small, E. J. & Armstrong, A. J. Neuroendocrine Prostate Cancer: Subtypes, Biology, and Clinical Outcomes. *J Natl Compr Canc Netw* **12**, (2014).
36. Sipola, J. *et al.* Plasma cell-free DNA chromatin 1 immunoprecipitation profiling depicts phenotypic and 2 clinical heterogeneity in advanced prostate cancer. *Cancer Res* (2024).
37. Datta, K., Muders, M., Zhang, H. & Tindall, D. J. Mechanism of lymph node metastasis in prostate cancer. *Future Oncology* **6**, 823–836 (2010).
38. Su, W. *et al.* The Polycomb Repressor Complex 1 Drives Double-Negative Prostate Cancer Metastasis by Coordinating Stemness and Immune Suppression. *Cancer Cell* **36**, 139-155.e10 (2019).
39. Kfoury, Y. *et al.* Human prostate cancer bone metastases have an actionable immunosuppressive microenvironment. *Cancer Cell* **39**, 1464-1478.e8 (2021).
40. Mohyeldin, A., Garzón-Muvdi, T. & Quiñones-Hinojosa, A. Oxygen in stem cell biology: A critical component of the stem cell niche. *Cell Stem Cell* **7**, 150–161 (2010).
41. Wang, Y. *et al.* Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. *Cell Rep* **25**, 1304-1317.e5 (2018).
42. Anbazhagan, R., Fujii, H. & Gabrielson, E. *Allelic Loss of Chromosomal Arm 8p in Breast Cancer Progression*. *American Journal of Pathology* vol. 152 (1998).
43. Kluth, M. *et al.* Deletion of 8p is an independent prognostic parameter in prostate cancer. *Oncotarget* **8**, 379–392 (2017).
44. Kimura, Y. *et al.* Genetic alterations in 102 primary gastric cancers by comparative genomic hybridization: Gain of 20q and loss of 18q are associated with tumor progression. *Modern Pathology* **17**, 1328–1337 (2004).
45. Melamed, J., Einhorn, J. M. & Ittmann2, M. M. Allelic loss on chromosome 13q in human prostate carcinoma. *Clin Cancer Res* **3**, 1867–1872 (1997).

869 46. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature*
870 **470**, 214–220 (2011).
871 47. Lundberg, A. *et al.* The Genomic and Epigenomic Landscape of Double-Negative
872 Metastatic Prostate Cancer. *Cancer Res* **83**, 2763–2774 (2023).
873 48. Song, F., Xu, J., Dixon, J. & Yue, F. Analysis of Hi-C Data for Discovery of Structural
874 Variations in Cancer. in *Hi-C Data Analysis; Methods and Protocols* (eds. Bicciato, S. &
875 Ferrari, F.) 143–161 (Humana, 2021).
876 49. Shanta, O. *et al.* The effects of common structural variants on 3D chromatin structure.
877 *BMC Genomics* **21**, (2020).
878 50. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**,
879 646–674 (2011).
880 51. Vargas-Rondón, N., Villegas, V. E. & Rondón-Lagos, M. The role of chromosomal
881 instability in cancer and therapeutic responses. *Cancers (Basel)* **10**, (2018).
882 52. Bayani, J. *et al.* Genomic mechanisms and measurement of structural and numerical
883 instability in cancer cells. *Semin Cancer Biol* **17**, 5–18 (2007).
884 53. Stopsack, K. H. *et al.* Oncogenic genomic alterations, clinical phenotypes, and outcomes
885 in metastatic castration-sensitive prostate cancer. *Clinical Cancer Research* **26**, 3230–
886 3238 (2020).
887 54. Zaccaria, S. & Raphael, B. J. Accurate quantification of copy-number aberrations and
888 whole-genome duplications in multi-sample tumor sequencing data. *Nat Commun* **11**,
889 (2020).
890 55. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**,
891 93–108 (2016).
892 56. Tomlins, S. A. *et al.* Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes
893 in Prostate Cancer. *Science (1979)* **310**, (2005).
894 57. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals
895 high concordance with metastatic tumors. *Nat Commun* **8**, (2017).
896 58. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables
897 interrogation of frozen tissues. *Nat Methods* **14**, 959–962 (2017).
898 59. Phan, L. *et al.* dbVar structural variant cluster set for data analysis and variant
899 comparison. *F1000Res* **5**, 673 (2016).
900 60. Martínez-Fundichely, A. *et al.* InvFEST, a database integrating information of
901 polymorphic inversions in the human genome. *Nucleic Acids Res* **42**, (2014).
902 61. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants
903 from high-throughput sequencing data. *Nucleic Acids Res* **38**, (2010).
904 62. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res* **52**, D891–D899 (2024).

905

A



B

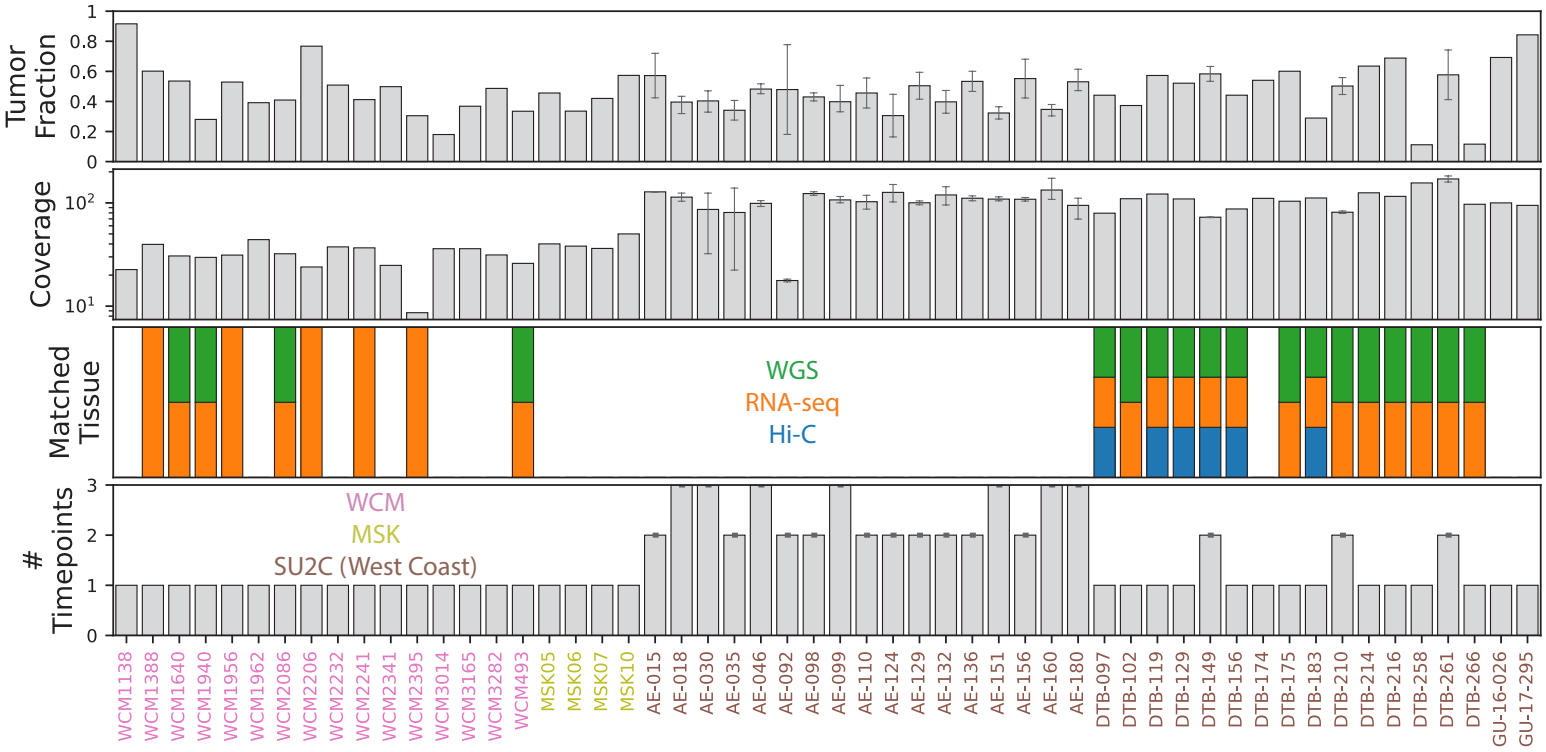


Fig. 1: Cohort overview. (A) Number of WGS cfDNA samples from plasma and WGS, RNA-seq, and Hi-C samples from tissue. Some patients have >1 sample. (B) per-sample overview of data availability for cfDNA samples.

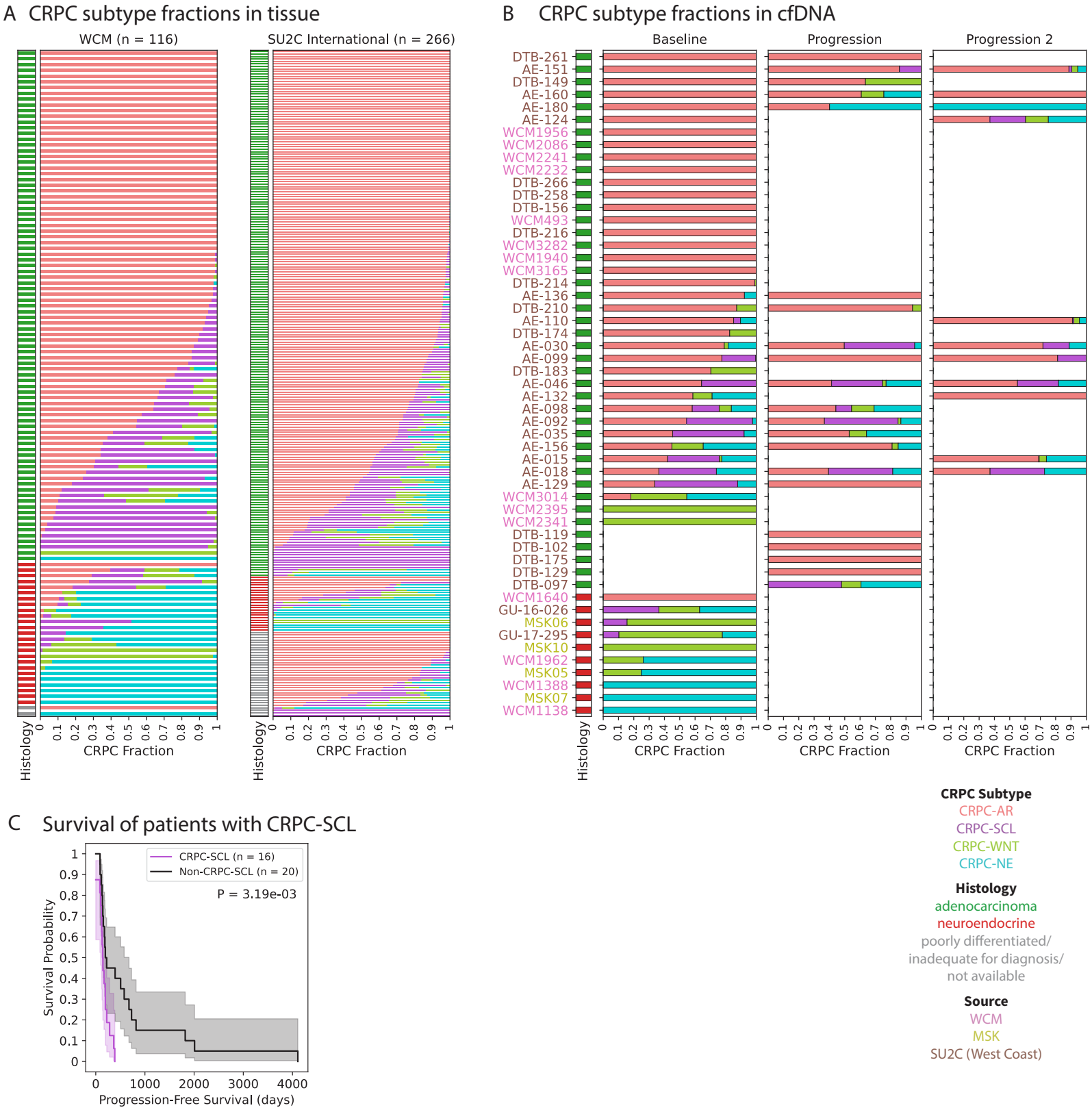


Fig. 2: CRPC subtype components in cfDNA and tissue. (A-B) CRPC fractions from deconvolution in tissue (A) and cfDNA (B) cohorts. (C) Kaplan Meier curve showing shorter progression free survival in patients with CRPC-SCL.

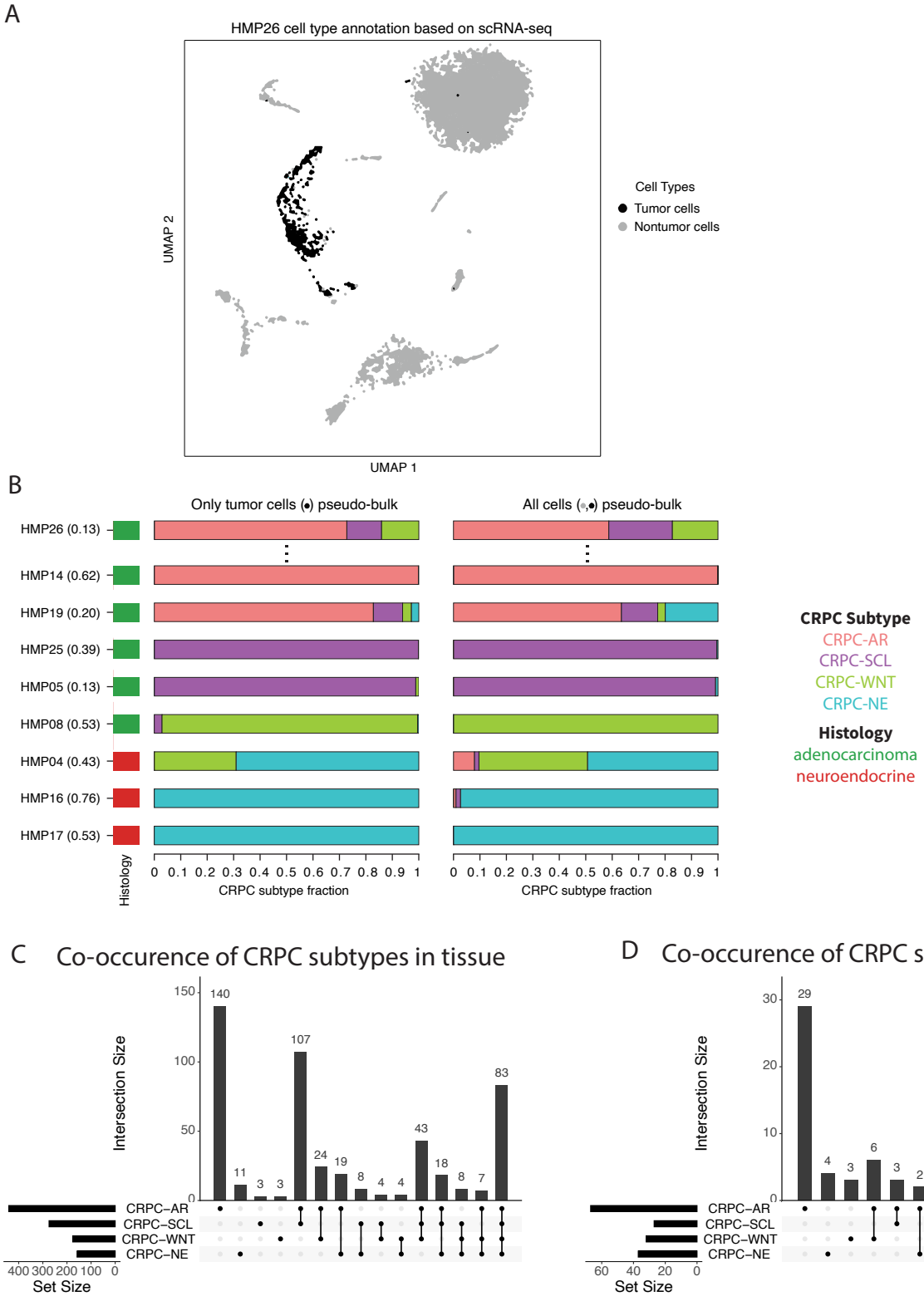


Fig. 3: Heterogeneity of CRPC subtypes. (A) Cell type annotation based on scRNA-seq in example sample. (B) CRPC fractions from deconvolution in tumor cells pseudobulk and all cells pseudobulk validates non-tumor cells do not impact subtype inference. (C-D) Upset plots showing number of samples with each CRPC subtype combination in tissue (C) and cfDNA (D). (E) ratio of number of samples with subtype mixtures to pure samples across subtypes.

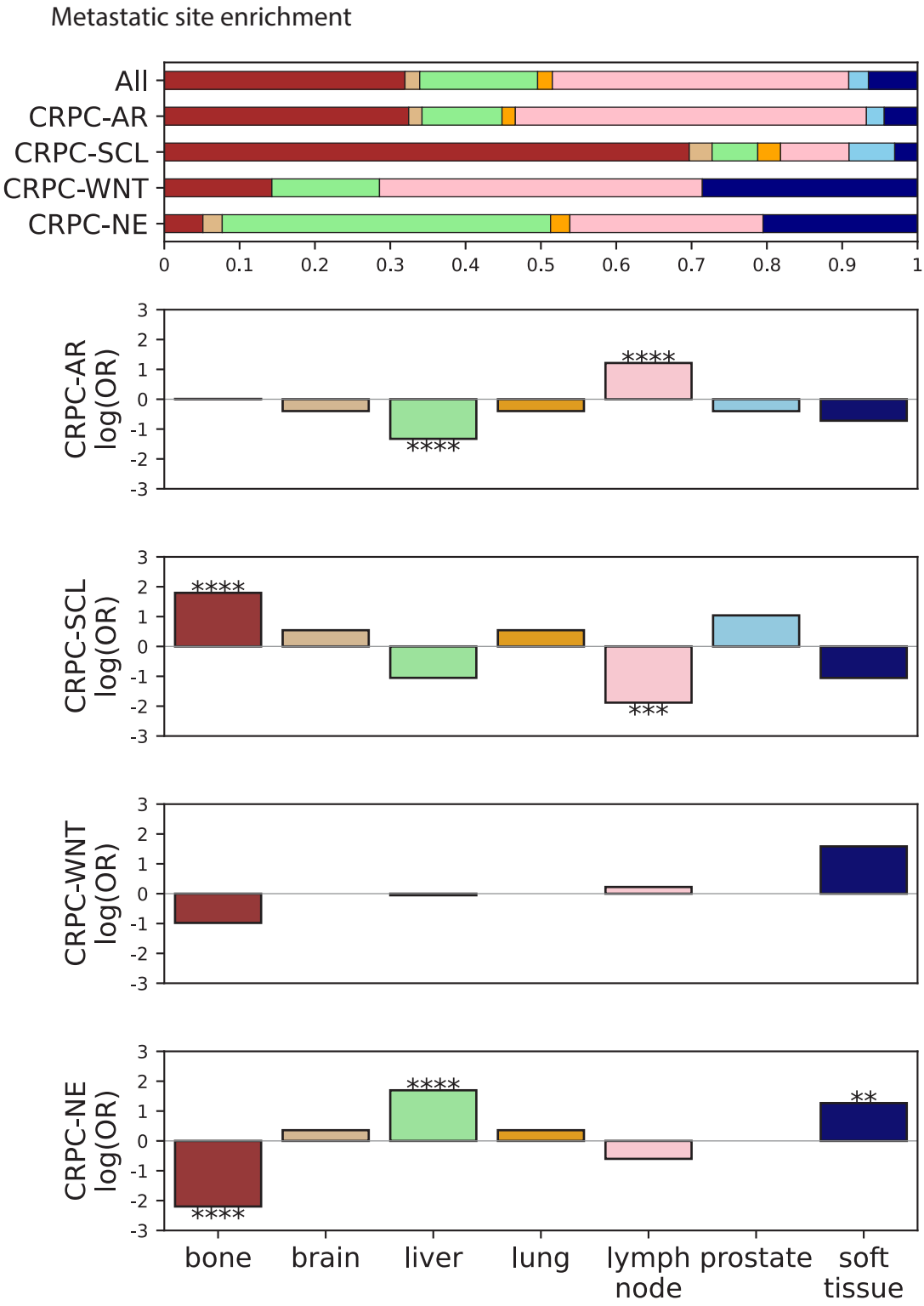
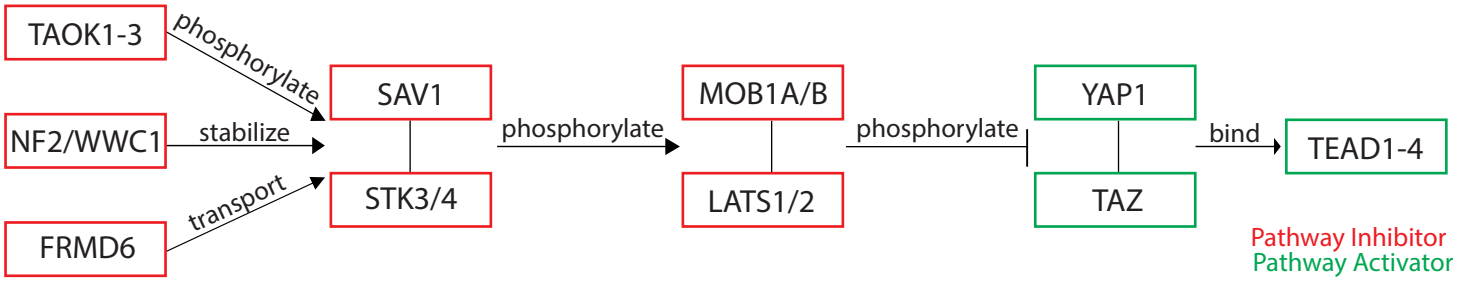
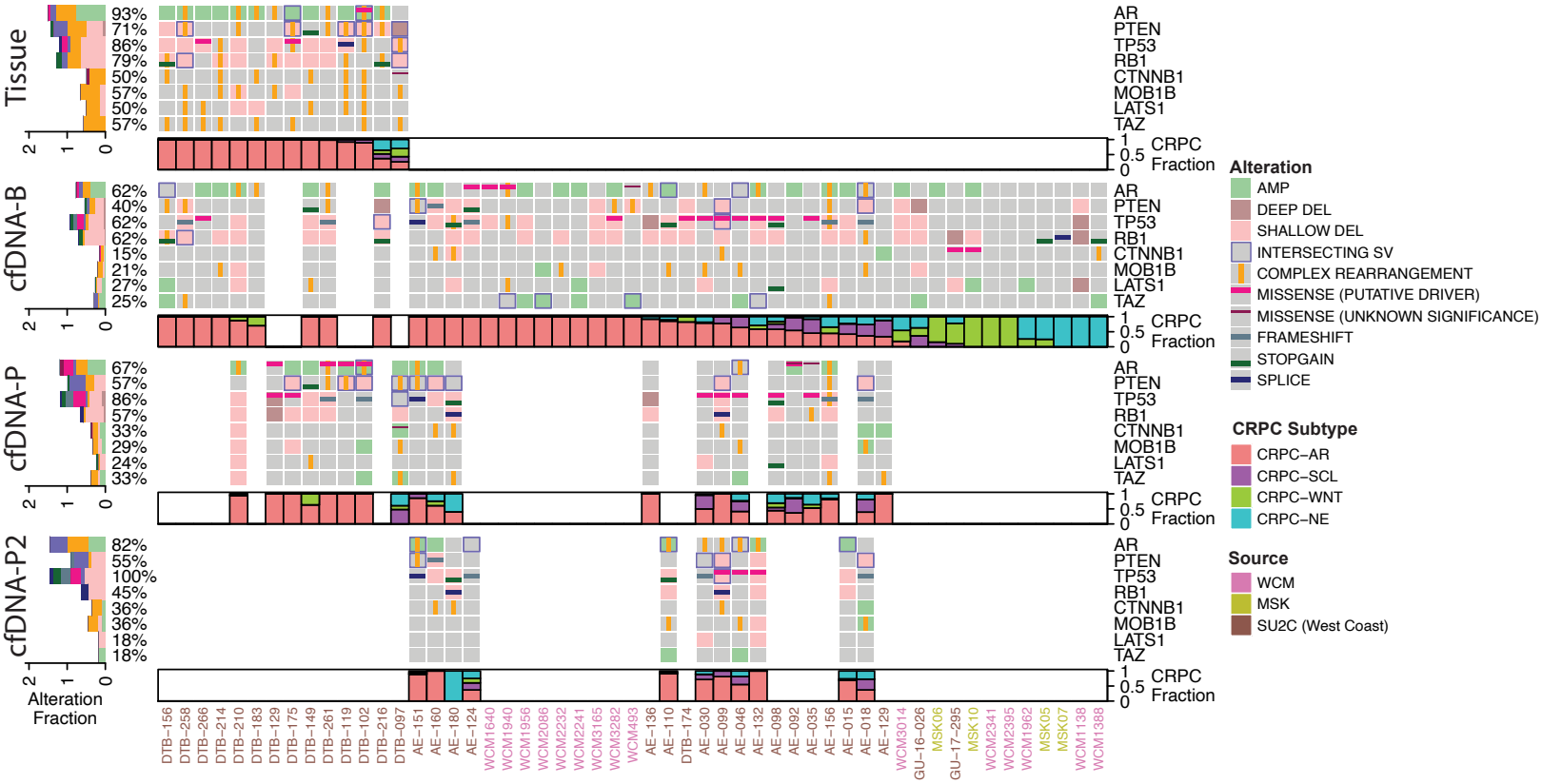


Fig. 4: Metastatic site enrichment. Fractions of samples (across all subtypes, and for within each subtype for subtype fraction ≥0.5) from each metastatic site (top), log (natural) odds ratio (OR) of a sample coming from each metastatic site across subtype fractions (bottom).

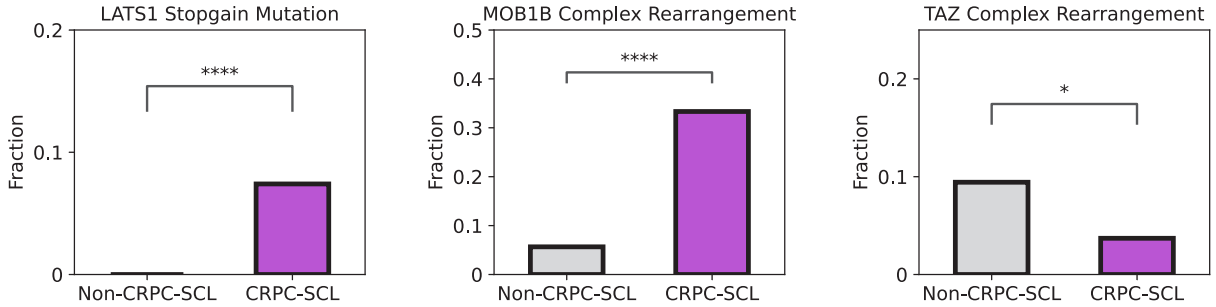
A YAP/TAZ Pathway



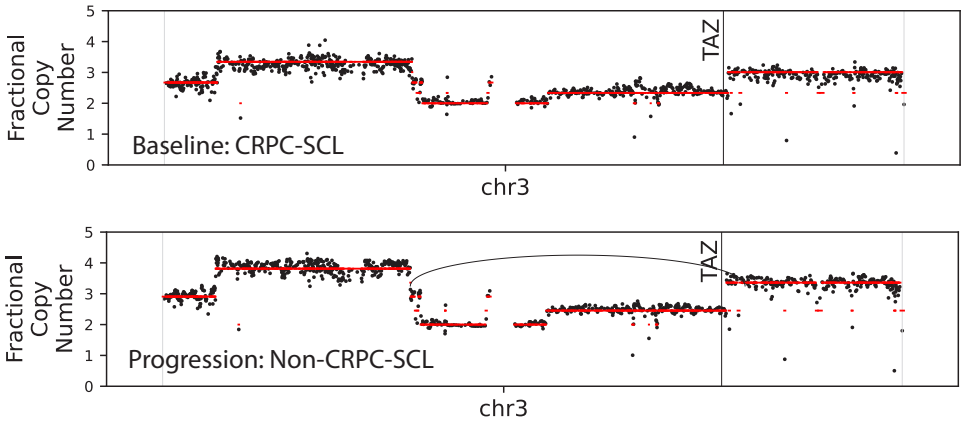
B OncoPrint



C Enrichment in CRPC-SCL



D Disappearance of CRPC-SCL upon emergence of rearrangement spanning TAZ



E CRPC-SCL classification ROC

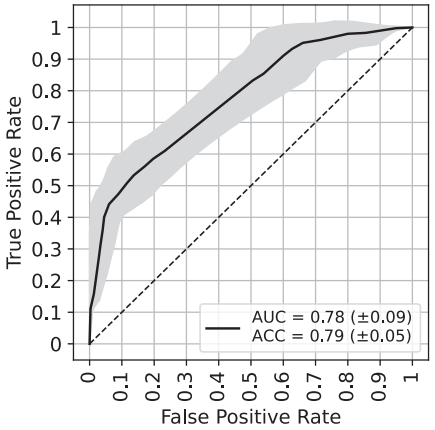


Fig. 5: YAP/TAZ pathway alterations. (A) YAP/TAZ pathway. (B) Oncoprint shows the genomic alterations for each patient in tissue and cfDNA timepoints (where available, Baseline (B), Progression (P), and Progression2 (P2)) for various genes of interest, each row is followed by the CRPC fraction of each sample from deconvolution. (C) LATS1 stopgain mutations in CRPC-SCL, complex rearrangement of chromosomes spanning MOB1B and TAZ in CRPC-SCL. (D) AE-129 signal track along chr3 across timepoints. Fractional copy number (FCN) along chr3, TAZ gene locus marked by vertical line, location of complex rearrangement shown as curved line at Progression. (E) receiver operating characteristic (ROC) curve of logistic regression model to classify samples as CRPC-SCL.

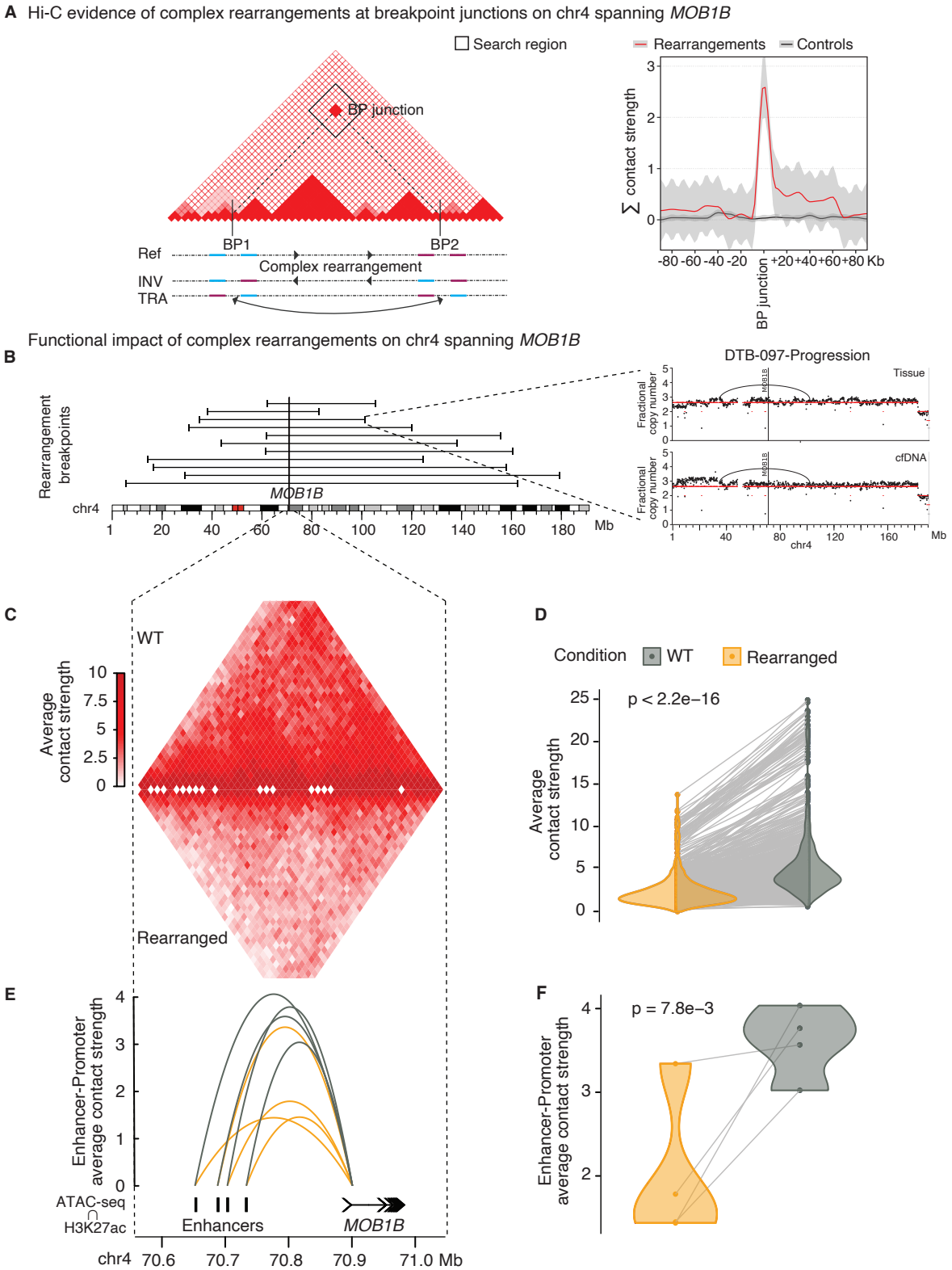
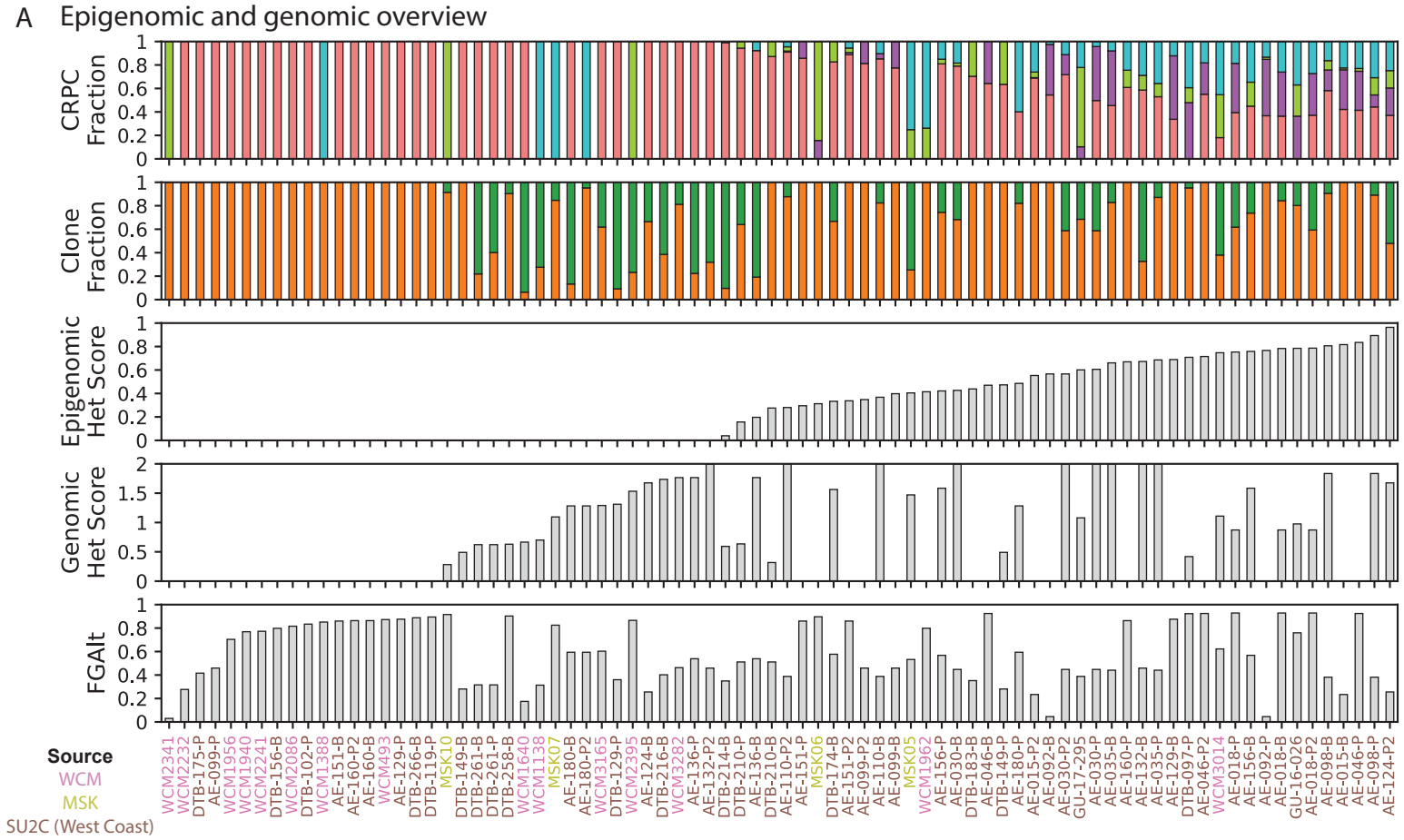
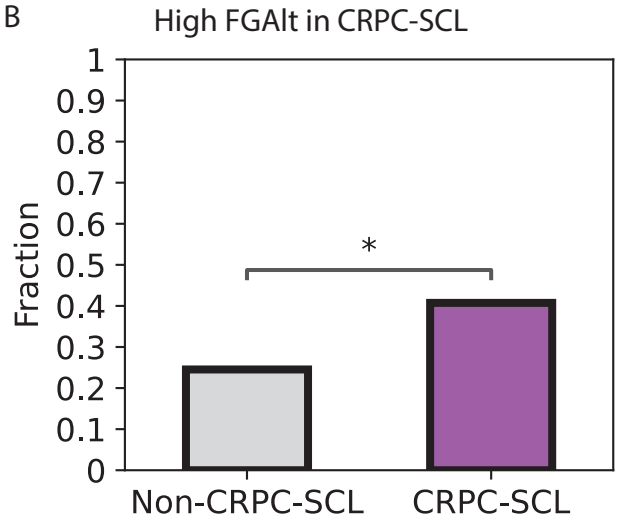


Fig. 6: Hi-C data validates structural variants and shows complex rearrangements reduce access of *MOB1B* to its regulatory elements. (A) schematic of Hi-C contact map and indices storing the connection strength of regions surrounding example inversion (INV) and translocation (TRA) breakpoints (i.e. breakpoint (BP) junction) (left) and contact strength increase in search region relative to negative control regions support evidence of rearrangement (right). (B) breakpoint locations of complex rearrangements spanning *MOB1B* in CRPC-SCL samples (left) and example of concordant breakpoints in cfDNA and matched tissue in one sample (right). (C) contact maps in TAD surrounding *MOB1B* in CRPC-SCL samples without (wildtype (WT), top) and with (rearranged, bottom) the rearrangement. (D) Strength of the contacts averaged across samples is lower in CRPC-SCL samples with the rearrangement (rearranged) compared to those without it (WT). (E) Average interaction strength between *MOB1B* promoter and its enhancers in CRPC-SCL samples with (rearranged) and without (WT) the rearrangement. (F). Strength of the enhancer-promoter interactions averaged across samples is lower in CRPC-SCL samples with the rearrangement (rearranged) compared to those without it (WT).





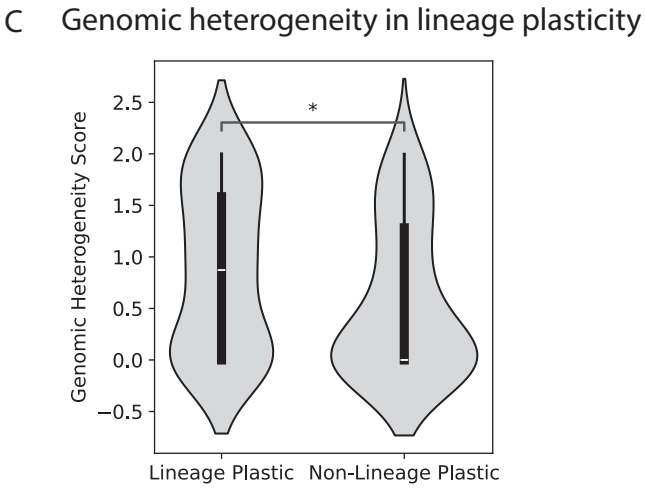


Fig. 7: Genomic heterogeneity and instability and CRPC subtypes. (A) Per-sample molecular and genomic overview: CRPC fraction from deconvolution (top row), tumor clone fraction composition (second row), molecular heterogeneity score (third row), genomic heterogeneity score (fourth row), and fraction of genome altered (last row). (B) High FGAlt enrichment in CRPC-SCL samples. (C) Genomic heterogeneity score distribution in lineage plastic and non-lineage plastic samples.