Nucleosome patterns in circulating tumor DNA reveal transcriptional regulation of advanced prostate cancer phenotypes

Navonil De Sarkar[1,2,3,4*], Robert D. Patton[1,2,*], Anna-Lisa Doebley[1,2,5,6], Brian Hanratty[2], Mohamed Adil[1], Adam J. Kreitzman[1,2], Jay F. Sarthy[7], Minjeong Ko[1,2], Sandipan Brahma[7], Michael P. Meers[7], Derek H. Janssens[7], Lisa S. Ang[2], Ilsa M. Coleman[2], Arnab Bose[2], Ruth F. Dumpit[2], Jared M. Lucas[2], Talina A. Nunez[2], Holly M. Nguyen[8], Heather M. McClure[9], Colin C. Pritchard[10,11], Michael T. Schweizer[3,12], Colm Morrissey[8], Atish D. Choudhury[9,13], Sylvan C. Baca[9,13], Jacob E. Berchuck[9], Matthew L. Freedman[9,13], Kami Ahmad[6], Michael C. Haffner[2,3,10], R. Bruce Montgomery[12], Eva Corey[8], Steven Henikoff[7,14], Peter S. Nelson[2,3,8,11,12,†], Gavin Ha[1,2,11,15,†]

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Center, 1100 Fairview Ave. N, Seattle, WA 98109
[2] Division of Human Biology, Fred Hutchinson Cancer Center, 1100 Fairview Ave. N, Seattle, WA 98109
[3] Division of Clinical Research, Fred Hutchinson Cancer Center, 1100 Fairview Ave. N, Seattle, WA 98109
[4] Department of Pathology and Prostate Cancer Center of Excellence, Medical College of Wisconsin, 8701 W Watertown Plank Road, Milwaukee, WI 53226
[5] Molecular and Cellular Biology Graduate Program, University of Washington, 1959 NE Pacific St, Seattle WA 98195
[6] Medical Scientist Training Program, University of Washington, 1959 NE Pacific St, Seattle WA 98195
[7] Division of Basic Sciences, Fred Hutchinson Cancer Center, 1100 Fairview Ave. N, Seattle, WA 98109
[8] Department of Urology, University of Washington, 1959 NE Pacific St, Seattle, WA, 98195
[9] Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215
[10] Department of Laboratory Medicine and Pathology, University of Washington, 1959 NE Pacific St, Seattle, WA 98195
[11] Brotman Baty Institute for Precision Medicine, 1959 NE Pacific St, Seattle, WA, 98195
[12] Division of Oncology, Department of Medicine, University of Washington
[13] Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142
[14] Howard Hughes Medical Institute, USA
[15] Department of Genome Sciences, University of Washington, 1959 Pacific St, Seattle, WA, 98195
* These authors contributed equally.
† Co-senior authors.

RUNNING TITLE: Cancer phenotype classification using ctDNA

KEYWORDS: Circulating tumor DNA, liquid biopsies, castration-resistant prostate cancer, patient-derived xenografts, fragmentomics

ADDITIONAL INFORMATION

Financial support:

This research was supported by the Pacific Northwest Prostate Cancer SPORE grant (P50 CA097186) and Department of Defense Idea Development Award (W81XWH-21-1-0513). This work was also supported by the Brotman Baty Institute for Precision Medicine grants (to G.H., R.D.P., and N.D.S.), Prostate Cancer Foundation Young Investigator Awards (G.H. and N.D.S.), National Institutes of Health (K22 CA237746, R21 CA264383, and DP2 CA280624 to G.H.; R01 CA234715 and R01 CA266452 to P.S.N.; P01 CA163227 to P.S.N., E.C., and C.M.; R01 CA251555 to M.L.F.; K99 GM138920 to S.B.; K99 GM140251 to M.P.M.), Department of Defense (W81XWH-18-1-0406 to P.S.N.; W81XWH-17-1-0380 to N.D.S.; W81XWH-18-0756, W81XWH-18-1-0356, PC170510, PC170503P2, PC200262P to C.C.P.;

W81XWH-20-1-0118 to J.E.B.; W81XWH-20-1-0084 to A.B.). Support was also provided by the Prostate Cancer Foundation, the Institute for Prostate Cancer Research, V Foundation Scholar Grants (to G.H. and M.C.H.), Fund for Innovation in Cancer Informatics Major Grant (to G.H.); Doris Duke Charitable Foundation and Safeway Foundation (to M.C.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, the Cutler Family Fund for Prevention and Early Detection, and Claudia Adams Barr Program for Innovative Cancer Research (to M.L.F.); ASCO Young Investigator Award, Kure It Cancer Research Foundation, and PhRMA Foundation (to S.C.B.). This research was also supported in part by the NIH/NCI Cancer Center Support Grant (P30 CA015704) and Scientific Computing Infrastructure (ORIP Grant S10OD028685).

Correspondence:

Gavin Ha, Ph.D.
Fred Hutchinson Cancer Center
1100 Fairview Ave. N,
Seattle, WA 98109
206-667-2802
gha@fredhutch.org

Peter S. Nelson, M.D.
Fred Hutchinson Cancer Center
1100 Fairview Ave. N,
Seattle, WA 98109
206-667-3377
pnelson@fredhutch.org

Conflict of interest disclosure:

G.H., A-L.D., N.D.S., R.D.P., P.S.N.: These authors have filed pending patent applications on methodologies developed in this manuscript (PCT/US2002/024082 and provisional patent USSN: 63/353,331).

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

B.M.: Has institutional funding from Clovis, Janssen, Astellas, BeiGene, and AstraZeneca.

E.C.: Received research funding under institutional SRA from Janssen Research and Development, Bayer Pharmaceuticals, KronosBio, Forma Pharmaceutics Foghorn, Gilead, Sanofi, AbbVie, and GSK for work unrelated to the present study.

M.L.F.: Serves as a consultant to and has equity in Nuscan Diagnostics. This activity is outside of the scope of this manuscript. M.L.F. has a pending patent for detecting NEPC using DNA methylation.

M.T.S.: Paid consultant and/or received Honoria from Sanofi, AstraZeneca, PharmaIn and Resverlogix. He has received research funding to his institution from Zenith Epigenetics, Bristol Myers Squibb, Merck, Immunomedics, Janssen, AstraZeneca, Pfizer, Madison Vaccines, Hoffman-La Roche, Tmunity, SignalOne Bio and Ambrx, Inc.

All other authors declare no competing interests.

ABSTRACT

Advanced prostate cancers comprise distinct phenotypes, but tumor classification remains clinically challenging. Here, we harnessed circulating tumor DNA (ctDNA) to study tumor phenotypes by ascertaining nucleosome positioning patterns associated with transcription regulation. We sequenced plasma ctDNA whole genomes from patient-derived xenografts representing a spectrum of androgen receptor active (ARPC) and neuroendocrine (NEPC) prostate cancers. Nucleosome patterns associated with transcriptional activity were reflected in ctDNA at regions of genes, promoters, histone modifications, transcription factor binding, and accessible chromatin. We identified the activity of key phenotype-defining transcriptional regulators from ctDNA, including AR, ASCL1, HOXB13, HNF4G, and GATA2. To distinguish NEPC and ARPC in patient plasma samples, we developed prediction models that achieved accuracies of 97% for dominant phenotypes and 87% for mixed clinical phenotypes. While phenotype classification is typically assessed by immunohistochemistry or transcriptome profiling from tumor biopsies, we demonstrate that ctDNA provides comparable results with diagnostic advantages for precision oncology.

STATEMENT OF SIGNIFICANCE

This study provides insights into the dynamics of nucleosome positioning and gene regulation associated with cancer phenotypes that can be ascertained from ctDNA. New methods for classification in phenotype mixtures extend the utility of ctDNA beyond assessments of somatic DNA alterations with important implications for molecular classification and precision oncology.

INTRODUCTION

Metastatic castration-resistant prostate cancer (mCRPC) describes the stage in which the disease has developed resistance to androgen ablation therapies and is lethal (1). Androgen receptor signaling inhibitors (ARSI), designed for the treatment of CRPC, repress androgen receptor (AR) activity and improve survival, but these therapies eventually fail (2,3). Since the adoption of ARSI as standard-of-care for mCRPC, there has been a prominent increase in the frequency of treatment-resistant tumors with neuroendocrine (NE) differentiation and features of small-cell carcinomas (4–7). These aggressive tumors may develop through a resistance mechanism of trans-differentiation from AR-positive adenocarcinoma (ARPC) to NE prostate cancer (NEPC) that lack AR activity (4,7–10). Additional phenotypes can also arise based on expression of AR activity and NE genes, including AR-low prostate cancer (ARLPC) and double-negative prostate cancer (DNPC; AR-null/NE-null) (5,11–13). Distinguishing prostate cancer subtypes has clinical relevance in view of differential responses to therapeutics, but the need for a biopsy to diagnose tumor histology can be challenging: invasive procedures are expensive and accompanied by morbidity, a subset of tumors are not accessible to biopsy, and bone sites pose particular challenges with respect to sample quality (7,14).

Circulating tumor DNA (ctDNA) released from tumor cells into the blood as cell-free DNA (cfDNA) is a non-invasive “liquid biopsy” solution for accessing tumor molecular information. The analysis of ctDNA to detect mutation and copy-number alterations has served to classify genomic subtypes of CRPC tumors (4,15–21). However, the defining losses of *TP53* and *RB1* in NEPC do not always lead to NE trans-differentiation (7,22). Rather, ARPC and NEPC tumors are associated with distinct reprogramming of transcriptional regulation (8,9,23). Methylation analysis of cfDNA in mCRPC to profile the epigenome shows promise for distinguishing phenotypes, but requires specialized assays such as bisulfite conversion, enzymatic treatment, or immunoprecipitation (24–27).

The majority of cfDNA represents DNA protected by nucleosomes when released from dying cells into circulation, leading to DNA fragmentation that is reflective of the non-random enzymatic cleavage by nucleases (28,29). Emerging approaches to analyze cfDNA fragmentation patterns from plasma for studying cancer can be performed directly from standard whole genome sequencing (WGS) (30–35). cfDNA fragments primarily have a characteristic size of ~167 bp, consistent with protection by a single core nucleosome octamer and histone linkers, but the size distribution may vary between healthy individuals and cancer patients (36–

39). Recent studies have demonstrated that the nucleosome occupancy in cfDNA at the transcription start site (TSS) and transcription factor binding site (TFBS) can be used to infer gene expression and transcription factor (TF) activity from cfDNA (40–42). However, nucleosome positioning and spacing are dynamic in active and repressed gene regulation (43–45). A detailed understanding of the nucleosome patterns and accessible chromatin associated with transcriptional regulation in tumor phenotypes has not been fully explored in cfDNA.

The objective of this study is to determine if ctDNA could be used to accurately classify tumor phenotypes in men with mCRPC. A major challenge for ctDNA analysis is the low tumor content (tumor fraction) in patient plasma samples. By contrast, plasma from patient-derived xenograft (PDX) models may contain nearly pure human ctDNA after bioinformatic exclusion of mouse DNA reads (37,39,46). This provides a resource that is ideal for studying the properties of ctDNA, developing new analytical tools, and validating both genetic and phenotypic features by comparison to matching tumors. In this study, we performed WGS of ctDNA from mouse plasma across 24 CRPC PDX lines with diverse phenotypes. Applying newly developed computational methods, we comprehensively interrogated the nucleosome patterns in ctDNA across genes, regulatory loci, TFBSs, TSSs, and open chromatin sites to reveal transcriptional regulation associated with mCRPC phenotypes. Finally, we designed two probabilistic models to accurately classify treatment-resistant tumors into divergent phenotypes and to estimate the phenotype heterogeneity within a ctDNA sample. We then validated the performance of these models in 159 plasma samples from three mCRPC patient cohorts. Overall, these results highlight that transcriptional regulation of tumor phenotypes can be ascertained from ctDNA and has potential utility for diagnostic applications in cancer precision medicine.

RESULTS

Comprehensive resource of matched tumor and liquid biopsies from patient derived xenograft (PDX) models of advanced prostate cancer

To develop approaches for the accurate classification of mCRPC using ctDNA, we evaluated 26 models from the LuCaP PDX series of advanced prostate cancer with well-defined phenotypes determined by whole transcriptome RNAseq and immunohistochemical assays for protein expression (47). The models consisted of 18 classified as ARPC, two classified as AR-low and NE-negative prostate cancer (ARLPC), and six classified as NEPC (**Figure 1A**). For each PDX line, we pooled mouse plasma (1.9 – 3.0 mL) from four to eight mice (mean tumor volume range 393-1239 mm^3), extracted cfDNA, and performed deep whole genome sequencing (WGS; mean

38.4x coverage, range 21 – 85x) **(Methods, Supplementary Table S1)**. We used bioinformatic subtraction of mouse sequenced reads to obtain nearly pure human ctDNA data **(Methods)**. We observed that 25 lines had human ctDNA comprising more than 10% of the sample (mean 52.9%, range 10.6 – 96%) with NEPC samples having significantly higher human fractions (mean 85.1%, range 77.1 – 96%, two-tailed Mann-Whitney U test $p = 9.6 \times 10^{-4}$) (**Figure 1B, Supplementary Table S1**). After subsequent filtering by human ctDNA sequencing coverage, 24 PDX lines remained for further analysis (16 ARPC, 6 NEPC, 2 ARLPC; mean 20.5x, range 3.8 – 50.6x, **Supplementary Table S1**). In the matching tumors, we performed Cleavage Under Targets and Release using Nuclease (CUT&RUN) to profile H3K27ac, H3K4me1, and H3K27me3 histone post-translational modifications (PTMs) (48,49) (**Supplementary Fig. S1**). We hypothesized that nucleosome organization inferred from ctDNA reflects the transcriptional activity state regulated by histone PTMs (50).

To study transcriptional regulation in mCRPC phenotypes from ctDNA, we interrogated four different features: (i) local promoter coverage, (ii) nucleosome positioning, (iii) fragment size analysis, and (iv) composite TFBSs plus open chromatin sites analysis using the Griffin framework (51) (**Figure 1A, Methods**). First, we analyzed three different local regions within ctDNA: all gene promoters and gene bodies and sites of histone PTMs guided by CUT&RUN analysis. For each of the three local regions, we extracted features of nucleosome periodicity using a nucleosome phasing approach and computed the fragment size variability. For promoter regions, we also computed the coverage at the transcription start site (TSS). Next, we analyzed ctDNA at transcription factor binding sites (TFBSs) and open chromatin regions. For each transcription factor (TF), ctDNA coverage at TFBSs were aggregated into composite profiles representing the inferred activity (42,51). Similarly, features in the composite profiles of phenotype-specific open chromatin regions were extracted for analyzing the signatures of chromatin accessibility in ctDNA. Altogether, we assembled a multi-omic sequencing dataset from matching tumor and plasma for a total of 24 PDX lines, making this a unique molecular resource and platform for developing transcriptional regulation signatures of tumor phenotype prediction from ctDNA.

Characterizing transcriptional activity of AR and ASCL1 in PDX phenotypes through analysis of tumor histone modifications and ctDNA

We sought to further characterize the transcriptional activity in different tumor phenotypes by studying epigenetic regulation via histone PTMs. We identified broad peak regions for H3K4me1 (median of 17,643 regions, range 1,894 – 64,934), H3K27ac (median 7,093, range 1610 - 34,047), and H3K27me3 (median 8,737, range 2,024 - 42,495) in the tumors of the 24 PDX lines and an additional nine LuCaP PDX lines where only tumors were available (total of 25 ARPC, 2 ARLPC, and 6 NEPC) (**Methods**, **Supplementary Fig. S1**, **Supplementary Table S2**). Using unsupervised clustering and principal components analysis (PCA), we identified putative active regulatory regions of enhancers and promoters (H3K27ac, H3K4me1) and gene repressive heterochromatic marks (H3K27me3) that were specific to ARPC, ARLPC, and NEPC phenotypes (52) (**Supplementary Fig. S2A**).

AR and ASCL1 are two key differentially expressed TFs with known regulatory roles in ARPC and NEPC phenotypes, respectively (9,53–55). When inspecting AR binding sites in ARPC tumors, we observed increased signals from flanking nucleosomes with H3K27ac PTMs compared to the other phenotypes (area under mean peak profile of 18.46 vs. 15.08 in ARLPC and 10.63 in NEPC, **Figure 2A**, **Supplementary Fig. S2B**, **Methods**). We also observed the strongest signals at the nucleosome depleted region (NDR) in ARPC for H3K27ac (1.54 coverage decrease vs. 0.78 for ARLPC and 0.41 for NEPC). Conversely, in NEPC tumors, we observed stronger signals at nucleosomes with H3K27ac PTMs flanking ASCL1 binding sites (area under mean peak profile 62.65 vs. 29.18 for ARLPC and 10.83 for ARPC), and stronger NDR signals (2.26 coverage decrease vs. 0.19 for ARPC and 0.37 for ARLPC). We observed similar trends for H3K4me1 PTMs in the LuCaP PDX lines (**Supplementary Fig. S2C**).

We analyzed the ctDNA composite coverage profiles at 1,000 consensus TFBSs to evaluate nucleosome accessibility, where lower normalized central (±30 bp window) mean coverage across these sites suggests more nucleosome depletion (**Methods**). For AR TFBSs, we observed the strongest signal for nucleosome depletion in ARPC, as indicated by the lowest mean central coverage (average 0.64, n=16), compared to moderate signals for ARLPC (average 0.88, n=2), and weakest signals for NEPC (average 0.95, n=6) (**Figure 2B**). Conversely, the composite coverage profile at ASCL1 TFBSs showed the strongest nucleosome depletion for NEPC samples (mean central coverage 0.69) compared to ARLPC (0.86) and ARPC (0.88) (**Figure 2C**). These observations were consistent with the differential binding activity by AR and ASCL1 in their respective phenotypes from tumor tissue (**Figure 2A**). We confirmed the same differential binding activity trends when analyzing TFBSs identified from other primary tissue sources (9,56,57) (**Supplementary Fig. S3A-B**). We also noted that the composite TFBS coverage patterns in ctDNA resembled the NDR flanked by nucleosomes with H3K27ac and H3K4me1 modifications inferred by CUT&RUN (**Figure 2A**, **Supplementary Fig.**

S2B-C). Together, these results suggest that the nucleosome depletion in ctDNA at AR and ASCL1 binding sites represents active TF binding and regulatory activity in specific prostate PDX tumor phenotypes.

Nucleosome patterns at gene promoters inferred from ctDNA are consistent with transcriptional activity for phenotype-specific genes

We selected 47 genes comprising 12 ARPC and 35 NEPC lineage markers established previously (4,5,58,59) and confirmed their phenotype associations by RNA-Seq from the PDX tumors (**Figure 2D**, **Supplementary Table S3**, **Methods**). To assess the activity of these genes from ctDNA, we analyzed the ctDNA fragment size in TSSs (± 1 kb window) and gene bodies and found that the differential size variability between phenotypes was positively correlated with relative expression (Spearman's $r = 0.844$, $p = 9.4 \times 10^{-14}$, **Figure 2E**, **Supplementary Fig. S4**, **Supplementary Table S2**, **Methods**). However, closer inspection of ctDNA coverage patterns at promoters revealed consistent nucleosome organization for transcription activity and repression (40,60–62) (**Figure 2D**). Therefore, we grouped the genes based on differential signals in H3K27me3 histone PTMs, which are linked with polycomb repressive complex mediated regulation and chromatin compaction (63).

For 25 genes without differential H3K27me3 peaks (Group 1), including AR, KLK3 and ASCL1, we observed nucleosome depletion at the TSS consistent with presence of active PTMs, such as for AR (mean coverage 0.47, n=16) in ARPC and ASCL1 (0.30, n=6) in NEPC samples (**Figure 2F, Supplementary Fig. S5**). By contrast, we observed increased coverage at the TSS of AR (1.08) in NEPC and ASCL1 (0.42) in ARPC, which supports nucleosome depletion in the absence of PTMs and inactive transcription. For 22 genes with differential H3K27me3 peaks (Group 2), including INSM1, CHGB and SRRM4, we observed relatively consistent increase in nucleosome occupancy and phasing in the TSS as well as in the gene body for ~50% of the genes (**Figure 2G, Supplementary Fig. S6**). The neuronal signaling genes in this group, such as UNC13A and INSM1, had reduced signals for the stable nucleosome dyad position, consistent with the heterogeneous ('fuzzy') nucleosome patterns described for actively transcribed genes (44,64). Interestingly, while UNC13A was active in NEPC tumors, we did not detect H3K27ac nor H3K4me1 PTM marks in the regulatory loci of this gene (**Supplementary Fig. S7A-B**). These results illustrate that ctDNA analysis can reveal patterns that are consistent with different modalities of transcriptional regulation by histone modifications for key genes that define prostate cancer phenotypes.

Phasing analysis in ctDNA reveals nucleosome periodicity associated with transcriptional activity between CRPC phenotypes

Regions of inactive or repressed transcription are expected to have stably bound nucleosomes, resulting in more periodic phasing in the gene body (61,65,66). Conversely, actively transcribed regions may exhibit overall disordered phasing in the gene body due to fast nucleosome turnover, resulting in relatively aperiodic patterns with highly varied protection from nucleases along the gene (67). To systematically quantify inter-nucleosomal spacing and predict nucleosome phasing, we developed TritonNP, a tool utilizing Fourier transforms and band-pass filters on GC-corrected ctDNA coverage to isolate frequency components corresponding to phased nucleosomes (**Figure 3A**, **Supplementary Fig. S8A-B**, **Methods**). This approach allows for calling phased nucleosome dyad positions to generate an average inter-nucleosome distance from the originating cells, encapsulating potential heterogeneity in nucleosome occupancy and stability. In PDX ctDNA, we observed a larger mean phased-nucleosome distance across 17,946 genes in the ARPC lines compared to the NEPC lines (median 291.1 bp vs. 282.6 bp, p = 0.027; two-tailed Mann-Whitney U test, **Figure 3B**). The phased nucleosome distance was also negatively correlated with the mean cell cycle progression (CCP) score (Spearman's rho = -0.563, p = 0.006, **Figure 3C**, **Methods**). These results suggest that increased nucleosome periodicity in NEPC ctDNA may reflect the condensed chromatin in hyperchromatic nuclei of NE cells (14), and the phasing analysis may have potential utility for assessing tumor proliferation and aggressiveness (68).

To model the relationship between nucleosome phasing and transcriptional activity more directly, we further extracted the frequency components corresponding to the inter-dyad distances of "stable" nucleosomes (180 – 210 bp) and a "baseline" component (150 – 180 bp) for normalization between samples of differing depths (69). We then computed the ratio of the mean frequency amplitudes between these components, which we designated the nucleosome phasing score (NPS), where a higher score corresponded to more ordered nucleosome phasing and repressed transcription. As an example, HOXB13, which is transcriptionally inactive in NEPC, had higher overall GC-corrected coverage (mean 56.85 depth) and a phased nucleosome distance of 249 bp with a 1.93 NPS in the LuCaP 93 NEPC PDX (**Figure 3A**). By contrast, HOXB13 is actively transcribed in ARPC and had lower coverage (mean 13.54 depth) and a more disordered phased-nucleosome distance of 332 bp with a 1.63 NPS in the LuCaP 136 ARPC PDX. When assessing the 47-prostate cancer phenotype marker genes, we observed that the mean NPS for the 35 NE genes was lower in NEPC lines compared to ARPC (median NPS 1.95 vs. 2.21, $p = 0.134$; two-tailed Mann-Whitney U test, **Figure 3D**); although this was not statistically significant, it was consistent with their active transcription. Conversely, the mean NPS for the 12 AR-regulated genes was lower in ARPC lines compared to NEPC (median NPS 1.82 vs 2.13, $p = 0.070$; two-tailed Mann-Whitney U test). In particular, 26 (74%) of the NE genes had lower NPS in NEPC compared to ARPC ($\log_2$ fold-change [ARPC:NEPC] > 0), including seven genes (ASCL1, CHGB, CHRNB2, GRP, MYCL, XKR7, NEUROD1) that were statistically significant ($p < 0.05$); ten (83%) of the AR-regulated genes had lower NPS in ARPC ($\log_2$ fold-change < 0), with TMPRSS2 being statistically significant (**Figure 3E**, **Supplementary Table S3**). These results illustrate that the NPS captured signals distinguishing key lineage-specific gene markers.

Inferred TF activity from analysis of nucleosome accessibility at TFBSs in ctDNA confirms key regulators of tumor phenotypes

To characterize the regulation of prostate tumor phenotype lineages, we considered nucleosome accessibility at TFBSs in PDX ctDNA for 338 TFs from the Gene Transcription Regulation Database (GTRD)(70) (**Methods**). First, we identified 108 TFs out of the 338 that were differentially expressed between ARPC and NEPC PDX tumors by RNAseq (**Supplementary Fig. S9, Supplementary Table S3, Methods**). Through unsupervised hierarchical clustering of composite TFBS central coverage values for these 108 TFs, we observed distinct groups of TFs in PDX ctDNA (**Figure 3F**). Of these 108 TFs, 38 had significantly different accessibility in ctDNA between ARPC and NEPC phenotypes (two tailed Mann-Whitney U test, Benjamini-Hochberg adjusted $p < 0.05$, **Supplementary Table S3**). Most of these TFs (27/38 [71%]) had differential inferred accessibility in ctDNA that was consistent with their up-regulation in the same phenotype by tumor mRNA expression, although some TFs (11/38, [29%]) did not show this trend (**Figure 3F**, **Supplementary Fig. S10**). A comparison of TFBS between paralogous TFs revealed that the binding sites used in the analysis had limited overlap (median 18.3%, range 0-81.2%), suggesting that many of the TFs may have some independent inferred accessibility (**Supplementary Fig. S11, Supplementary Table S3**). For paralogs with high TFBS overlap (≥ 19%), such as AR, NR3C1 and PGR, we noted only a subset of TFs were expressed in one phenotype.

REST had the largest difference in accessibility as supported by a decrease in coverage within ARPC models compared to NEPC ($\log_2$ fold-change -0.77, adjusted $p = 5.7 \times 10^{-4}$, **Supplementary Fig. S12A**, **Supplementary Table S3**). FOXA1, and GRHL2 binding sites were significantly more accessible in ARPC (and ARLPC) samples compared to NEPC ($\log_2$ fold-change < -0.57, adjusted $p < 1.3 \times 10^{-3}$). AR, HOXB13, and NKX3-1 had higher accessibility in ARPC compared to NEPC ($\log_2$ fold-change < -0.37, adjusted $p < 1.3 \times 10^{-3}$), but with only moderate accessibility in ARLPC, as expected. We also observed a group of TFs that followed a similar trend, including nuclear hormone receptors (NR2F2, RARG), pioneer factor GATA2, and nuclear factors HNF4G and HNF1A ($\log_2$ fold-change < -0.10, adjusted $p < 0.027$, **Supplementary Fig. S12A**).

For factors that had higher accessibility in NEPC models compared to ARPC and ARLPC, ASCL1 had the largest TFBS coverage difference ($\log_2$ fold-change 0.36, adjusted $p = 5.7 \times 10^{-4}$, **Figure 2C**, **Figure 3F)**. Other TFs, including RUNX1, BCL11B, POU3F2, NEUROG2, and SOX2 also had sites with higher accessibility in NEPC ($\log_2$ fold-change > 0.06, adjusted $p < 0.048$, **Supplementary Fig. S12B**), although the difference was modest. Other notable factors such as MYC and ETS transcription family genes (ETV4, ETV5, ETS1, ETV1) had high accessibility across all phenotypes, while NEUROD1, RUNX3, and TP63 sites were inaccessible in nearly all samples. Furthermore, we considered restricting the analysis to 20 TFs with TFBSs that were observed in prostatic tissue and cell lines and were also differentially expressed in the PDX tumors by RNAseq (**Methods**). However, while hierarchical clustering distinguished PDX tumor phenotypes, key NEPC-defining markers, such as ASCL1, were omitted from this analysis as ChIP-seq for many NEPC-defining markers had not been performed on prostate lineage samples in GTRD (**Supplementary Fig. S13**). Overall, we identified the accessibility of known prostate cancer regulators, including ASCL1, HNF4G, HNF1A, GATA2 and SOX2 (71–73), that have not been shown before from ctDNA analysis in these tumor phenotypes.

Phenotype-specific open chromatin regions (ATAC-Seq) in PDX tumor tissue are reflected in ctDNA profiles of nucleosome accessibility

Nucleosome profiling from cfDNA sequencing analysis has shown agreement with overall chromatin accessibility in tumor tissue (38,42,74); however, its application for distinguishing tumor phenotypes has been limited. We hypothesized that due to lack of protection from nucleases, regions of open chromatin would be under-represented in ctDNA assays. We investigated the use of ATAC-Seq data from tumor tissue for 10 LuCaP PDX lines (5 ARPC and 5 NEPC) to inform phenotype-related differences in chromatin accessibility (9). We defined an initial set of 28,765 ARPC and 21,963 NEPC differential consensus open chromatin regions which we further restricted to those that overlapped TFBSs for 338 TFs, resulting in 15,879 ARPC and 11,692 NEPC sites (**Methods**, **Figure 4A**). For ARPC-specific open chromatin sites, we observed decreased overall composite site coverage (+/- 1 kb window) and central coverage (+/- 30 bp) in the ctDNA for ARPC PDX lines (mean central coverage 0.75, n=16) compared to NEPC lines (mean 0.96, n=6) and cfDNA from healthy human donors (mean 0.97, n=14) (**Figure 4B**, **Supplementary Table S3, Methods**). Conversely, for NEPC-specific open chromatin sites, coverage was decreased in ctDNA for NEPC lines (mean 0.89) compared to ARPC lines (mean 1.01) and healthy donors cfDNA (mean 1.00) (**Figure 4C**, **Supplementary Table S3**). Coverage patterns were discernable between phenotypes for as few as 100 sites, suggesting that even a smaller subset of open chromatin regions may still be informative (**Supplementary Fig. S14A-B**). These results confirmed that tumor tissue chromatin accessibility can be corroborated in ctDNA and that ARPC and NEPC phenotypes have distinct ctDNA coverage profiles at these sites.

Comprehensive evaluation of ctDNA features across genomic contexts for CRPC phenotype classification

To assess the utility of ctDNA nucleosome profiling for informing prostate cancer phenotype classification, we systematically evaluated four groups of global genome-wide ctDNA features: phasing, fragment sizes, local coverage profiling, and composite site coverage profiling (**Figure 1A**). From principal components analysis (PCA), we observed distinct feature signals between ARPC and NEPC phenotypes for composite TFBS coverage of TFs, NPS of the 47 phenotype marker genes, and fragment size variability at global sites of PTMs (**Figure 4D**, **Supplementary Fig. S15A, Supplementary Table S4**, **Methods**). In addition to these features, we also included previously reported features, including short-long fragment ratio and local coverage patterns at the TSS (max wave height between -120bp to 195bp) (30,41) (**Methods**).

We then quantitatively evaluated all combinations of coverage, phasing, and fragment size features for different genomic contexts to investigate their potential to classify ARPC and NEPC phenotypes. For each feature set, we conducted 100 iterations of stratified cross-validation using a supervised machine learning classifier (XGBoost) on ctDNA samples from the ARPC and NEPC models and computed the area under the receiver operating characteristic curve (AUC) (**Methods**). First, we evaluated an established set of 10 genes associated with AR activity (5,12). We observed that the phased nucleosome distance at H3K27ac sites and the central coverage at TSSs had moderate predictive performance (AUC 0.88) (**Supplementary**

Fig. S15B, **Supplementary Table S4**). For the set of 47 phenotype markers, the NPS of gene bodies was most predictive (AUC 0.98) (**Supplementary Fig. S15C, Supplementary Table S4)**. When considering all PTM sites, promoters, genes, TFs, and open chromatin regions, the best performing features included mean fragment size at H3K4me1 sites (n=9,750, AUC 1.0) and promoter TSSs (n=17,946, AUC 1.0), and both open chromatin composite site features (AUC 1.0) (**Figure 4E**, **Supplementary Table S4**).

Accurate classification of ARPC and NEPC phenotypes from patient plasma using a probabilistic model informed by PDX ctDNA analysis

An important consideration and challenge in analyzing plasma from patients is the presence of cfDNA released by hematopoietic cells, which leads to a lower ctDNA fraction (i.e., tumor fraction). Furthermore, the small patient cohorts with available tumor phenotype information make supervised machine learning approaches suboptimal. Therefore, we developed ctdPheno, a probabilistic model to classify ARPC and NEPC from an individual plasma sample, accounting for the tumor fraction (**Figure 5A, Methods**). We focused on the phenotype-specific open chromatin composite site features and used the PDX plasma ctDNA signals (27,571 total sites, **Figure 4B-C**, **Supplementary Table S3**) to inform the model. The model produces a normalized prediction score that represents the estimated signature of ARPC (lower values) and NEPC (higher values). We applied this method to benchmarking datasets generated by simulating varying tumor fractions and sequencing coverages using five ARPC and NEPC PDX ctDNA samples each, and healthy donor plasma cfDNA (**Supplementary Figure 15D**, **Methods**). We achieved a 1.0 AUC at 25X coverage down to 0.01 tumor fraction, 1.0 AUC at 1X down to 0.2 tumor fraction, and 1.0 AUC at 0.2x coverage at 0.3 tumor fraction, suggesting a possible upper-bound performance for classifying samples with lower tumor fraction in plasma (**Figure 5B**, **Supplementary Table S4**).

To test the performance of ctdPheno on patient samples, we analyzed a published dataset of ultra-low-pass whole genome sequencing (ULP-WGS) of plasma cfDNA (mean coverage 0.52X, range 0.28-0.92X) from 101 mCRPC patients comprising 80 adenocarcinoma (ARPC) and 21 NEPC samples (DFCI cohort I) (25). Using ctdPheno, which was unsupervised and used parameters informed only by the PDX analysis, we achieved an overall AUC of 0.96 (**Figure 5C**, **Supplementary Table S5**). The performance was 0.97 AUC and 0.76 AUC when considering samples with high (≥ 0.1) and low (< 0.1) tumor fraction, respectively, and 0.83 AUC when using only 2000 sites for analysis (**Supplementary Fig. S16A-B**). We identified an optimal overall performance at 97.5% specificity (ARPC) and 90.4% sensitivity (NEPC) which corresponded to the prediction score of 0.3314 (**Figure 5C**). These results were concordant (92.1%) with phenotype classification by cfDNA methylation on the same plasma samples (**Supplementary Fig. S16C**, **Supplementary Table S5**). In another published dataset of 11 mCRPC samples from 6 patients who had high PSA, treatment with ARSI, or both (DFCI cohort II) (75,76), the model correctly classified patients as ARPC in 8 (73%) ULP-WGS (~0.1x) samples when using the optimal score cutoff (**Supplementary Fig. S16D**, **Supplementary Table S5).**

Next, we analyzed 61 clinical plasma samples from 31 CRPC patients with ARPC, NEPC, and mixed phenotypes that are representative of typical clinical histories (UW Cohort, **Supplementary Table S5**). We performed ULP-WGS of cfDNA and selected 47 samples (26 ARPC, 5 NEPC, and 16 mixed phenotype) from 27 patients based on having greater than 3% estimated tumor fraction (**Supplementary Table S5**, **Methods**). For the 26 samples with ARPC clinical phenotype, ctdPheno correctly classified 22 (85%) samples with ARPC-dominant clinical phenotype and all five (100%) samples with NEPC-dominant clinical phenotype using the score cutoff of 0.3314 (**Figure 5D**). For the remaining 16 samples with clinical histories or tumor histologies that reflected mixed phenotypes such as a tumor with AR-positive adenocarcinoma intermixed with NEPC, the classification results were variable (**Figure 5D**, **Supplementary Table S5, Supplementary Fig. S17**). Overall, we achieved an accuracy of 87% for ULP-WGS data of ctDNA samples with dominant clinical phenotypes, but the variable predictions for mixed phenotype samples underscore the complexities associated with tumor heterogeneity in the setting of metastatic disease.

Quantifying ARPC and NEPC phenotype heterogeneity within individual patient plasma ctDNA

Phenotype heterogeneity may arise in the clinical setting, particularly when trans-differentiation can lead to a mixture of ARPC and NEPC cells or lesions. To account for and predict phenotype mixtures within a patient ctDNA sample, we developed Keraon, an analytical model that estimates the proportions of phenotypes from WGS using the same ctDNA features as ctdPheno (**Figure 5E, Methods**). First, we evaluated Keraon using a benchmark dataset generated for simulating varying tumor fractions and proportions of ARPC-NEPC mixtures at 25x coverage using PDX ctDNA and healthy donor cfDNA data (**Figure 5F, Methods**). In 810 simulated phenotype mixtures, we observed the estimated total NEPC fraction was consistent with expected proportions (Pearson's r=0.884) with a mean absolute error (MAE) of 0.028, highlighting the method's potential for accurate estimation of emergent phenotypes in mixed histology samples (**Figure 5G, Supplementary Table S5**). Next, we evaluated Keraon for classifying NEPC in DFCI Cohort I and observed the highest performance (0.96 AUC) using all 27,571 open chromatin sites, with decreased performance (0.84 AUC) when using only 2,000 sites (**Supplementary Fig. S16D**). Applying Keraon to analyze DFCI Cohort II, we correctly estimated dominant ARPC with undetectable NEPC phenotype in 10 (91%) samples with WGS (mean coverage, 27x) (**Supplementary Fig. 18**, **Supplementary Table S5**).

We performed deeper WGS (22.13x mean coverage, range 15.15x – 31.79x) for the UW Cohort ctDNA samples and applied Keraon to classify the presence of NEPC and to estimate the proportions of ARPC and NEPC phenotypes (**Figure 5H**). Keraon correctly estimated the dominant phenotype (≥ 0.5 relative phenotype fraction) in 25 of 26 (96%) samples with ARPC clinical phenotype and in 5 of 5 (100%) NEPC samples. For 10 samples with presence of ARPC and NEPC phenotypes reported in the clinical histories, Keraon correctly detected both phenotypes in nine samples (NEPC fraction ≥ 0.028, ARPC fraction ≥ 0.06). In two samples with ARPC-DNPC phenotypes, one was estimated to be ARPC-dominant (0.20 fraction), and in three samples with NEPC-DNPC phenotypes, all three were estimated as being NEPC-dominant (≥ 0.028 fraction). In 14 (82%) out of 17 patients with multiple plasma collected, the predicted phenotypes were consistent across all ctDNA samples. Overall, we observed an accuracy of 97% for correctly classifying ARPC and NEPC dominant phenotypes and 87% for estimating NEPC fractions in samples with admixed clinical phenotypes from ctDNA.

DISCUSSION

The development of minimally invasive blood-based assays of ctDNA to define tumor subtypes has dramatically changed the landscape of clinical oncology. To date, the majority of these assays characterize genomic alterations in oncogenes such as *EGFR* or tumor suppressors, such as *BRCA2*, that inform outlier responses to specific therapeutics. However, tumor classification determined by gene expression analyses, such as the PAM50 subtyping of breast carcinoma and the transcript-based classification of urothelial cancers is also informative of clinical trajectories. Consequently, the ability to characterize tumor phenotype using blood-based assays has the potential to add relevant information for guiding treatment allocation.

In the present study, we analyzed multiple features of DNA to infer the activity of gene expression programs corresponding to distinct prostate cancer phenotypes. A key component of the work that allowed for the development of optimized methods and the identification of the most informative ctDNA features was the use of PDX models. The sequencing of mouse plasma provided a unique opportunity to comprehensively interrogate the epigenetic nucleosome patterns in ctDNA from well-characterized tumor models. We developed and applied computational methodologies to evaluate a multitude of ctDNA features, each of which were associated with transcriptional regulation across CRPC tumor phenotypes. The use of PDX mouse plasma overcomes the challenge of low ctDNA content or incomplete knowledge of the tumor when studying patient samples. Using features learned from the PDX ctDNA, we developed models to accurately classify ARPC and NEPC and to estimate their proportions in phenotypically heterogenous samples from patient plasma in three clinical cohorts. While these data were focused on ARPC and NEPC phenotypes, the approaches may serve as a framework for the use of ctDNA to subtype malignancies arising in other organ sites based on distinctive gene expression programs.

The analysis of the LuCaP PDX ctDNA sequencing data confirmed the activity of key regulators between ARPC and NEPC phenotypes, including a set of 47 established differentially expressed genes that associate with cell lineage. While gene expression inference from ctDNA has been shown in proof-of-concept studies (34,41), the PDX ctDNA allowed for a detailed dissection of nucleosome organization associated with transcriptional activity of individual genes that define the tumor phenotypes. Previous analytical approaches have profiled nucleosome occupancy from cfDNA (38,74). However, our assessment of nucleosome stability by means of the Nucleosome Phasing Score is the first to capture the highly variable spacing, positioning, and turnover of the nucleosome arrays associated with transcription and tumor aggressiveness (43,67,68,77).

In addition to the existing molecular profiling available for these models, we now provide characterization of histone PTMs in LuCaP PDX tumors using CUT&RUN. At regions with these PTMs on histone tails, we observed expected nucleosome patterns inferred in ctDNA that were consistent with active or repressed gene transcription. To our knowledge, this is the first time that ctDNA analysis has been performed in the context of histone PTMs and will provide a blueprint to develop new approaches for studying additional epigenetic alterations using PDX plasma.

While the regulation of key factors such as AR, HOXB13, NKX-3.1, FOXA1, and REST has been shown from ctDNA in CRPC (35,42), we report the differential activity of other key factors in CRPC from ctDNA analysis. This included nuclear factors HNF4G and HNF1A, and pioneering factor GATA2, which are associated with prostate adenocarcinoma (ARPC) (71,73,78). ASCL1 is a pioneer TF with roles in neuronal differentiation and was recently described to be active during NE trans-differentiation and in NEPC (9,55). To our knowledge, this study is the first to demonstrate ASCL1 binding site accessibility and provide a detailed characterization of its transcriptional activity in NEPC from plasma ctDNA.

We show an expansive analysis of TFBSs for 338 factors in each plasma sample without the need for chromatin immunoprecipitation or other epigenetic assays. However, we did not find a significant difference in accessibility for 70 out of the 108 TFs in ctDNA, which may be consistent with TF activity not necessarily being correlated with its own expression level (79). On the other hand, the accessibility of TFBSs may not necessarily indicate true TF activity as other co-bound TFs or co-activators/co-repressors influence gene regulation. Moreover, our analyses were based on TFBSs obtained from public databases, including for a limited number of prostate-specific TFs; however, expanded phenotype-specific TF cistrome data may improve this approach.

We applied state-of-the-art computational approaches built on existing and new concepts of ctDNA data analysis to extract tumor-specific features, including the representation of nucleosome phasing, periodicity, and spacing associated with transcriptional activity. Other approaches have also considered regions, such as TSSs, TFBSs, and DNase hypersensitivity sites (33,38,41,42); however, after a systematic evaluation, we found that ctDNA features in open chromatin sites derived from ATAC-Seq of PDX tissue (9) provided the highest performance for distinguishing CRPC phenotypes. We presented ctdPheno, which is a probabilistic model that classifies ARPC and NEPC from ULP-WGS data, and Keraon, an analytical model which estimates the proportion of ARPC and NEPC from WGS of patient plasma. Both models are unsupervised and utilize a statistical framework informed directly by parameters from the LuCaP PDX ctDNA analysis. These models do not require training on patient samples but do require tumor fraction estimates (ichorCNA (80)) and in the case of ctdPheno a prediction score cutoff determined from DFCI cohort I. Both frameworks can also be extended to model additional phenotypes. Insights from additional datasets such as single-cell nucleosome and accessibility profiling (81,82) of PDX tumors and clinical samples may improve the resolution for ctDNA analysis. While we observed optimal performance analyzing all open chromatin sites, a smaller subset was still informative which may be useful when considering targeted assays for clinical applications.

Applying the prediction models to patient datasets with definitive clinical phenotypes yielded high performance even when using low depth of coverage sequencing. In particular, our performance for the DFCI cohort I was also consistent with the reported phenotype classification results using ctDNA methylation in the same patients (25). Similarly, in the UW cohort, samples with well-defined clinical phenotypes had near-perfect concordance from WGS data. We established the lower limits of phenotype classification performance to be at 8% tumor fraction for ctdPheno (ULP-WGS) and 3% for Keraon (WGS). These results support a strategy whereby ULP-WGS is performed for screening using ctdPheno, along with clinical assessments, and followed-up with standard WGS for more accurate and comprehensive phenotype characterization using Keraon. While this framework may have limited performance for low (<3%) ctDNA levels, it may be optimal at initial assessment of metastatic disease and at tumor progression on therapy, which is when the clinical decision points are most critical.

Tumor heterogeneity and co-existence of different molecular phenotypes are common in mCRPC where treatment-induced phenotypic plasticity may vary within and between tumors in an individual patient. In real data simulations and patients with cases of mixed clinical phenotypes, Keraon accurately detected the contributions of mixed phenotypes with a detection limit of 2.8% NEPC, providing the first approach to directly quantify phenotype proportions and heterogeneity from ctDNA. In this study, estimation of phenotype heterogeneity using Keraon required standard depths of WGS. Larger studies with comprehensive assessment of the tumor histologies will be needed for evaluating these models as potential biomarkers of treatment response.

In summary, this study illustrates that analysis of ctDNA from PDX mouse plasma at scale can facilitate a detailed investigation of tumor regulation. These results, together with the suite of computational methods presented here, highlight the utility of ctDNA for surveying transcriptional regulation of tumor phenotypes and its potential diagnostic applications in cancer precision medicine.

ACKNOWLEDGEMENTS

We thank the many patients and their families for their altruistic contributions to this study. We thank the Fred Hutchinson Cancer Center Genomics Shared Resources Core members, the Institute for Prostate Cancer Research clinicians and staff that support the University of

Washington rapid autopsy program and the PDX program. We thank Patricia Galipeau and members of the Ha and Nelson Laboratories for critically reading this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: N.D.S., R.D.P., G.H., P.S.N.

Methodology: R.D.P., N.D.S., A-L.D., P.S.N., G.H.

Software: R.D.P., A-L.D., N.D.S., B.H., A.J.K., G.H.

Formal Analysis: R.D.P., N.D.S., A-L.D., B.H., A.J.K., M.K., M.A., I.M.C., G.H.

Investigation: N.D.S., R.D.P., A-L.D., B.H., J.F.S., J.M.L., A.B., G.H.

Resources: N.D.S., J.S., S.B., M.P.M., D.H.J., L.S.A., R.F.D., T.A.N., H.M.M., S.C.B., J.E.B., M.L.F., C.M., H.M.N., E.C., S.H., P.S.N., G.H.

Data Curation: N.D.S., R.D.P., A-L.D., M.P.M., S.B., D.H.J., E.C., C.M., A.D.C., M.C.H., P.S.N., G.H.

Writing – Original Draft: R.D.P, N.D.S, P.S.N., G.H.

Writing – Review & Editing: N.D.S, R.D.P., A-L.D., C.C.P., C.M., A.D.C., M.T.S., R.B.M., M.C.H., E.C., K.A., S.H., P.S.N., G.H.

Visualization: R.D.P., N.D.S., B.H., M.K., M.C.H., P.S.N., G.H.

Supervision: P.S.N., G.H.

Funding Acquisition: G.H., P.S.N.

MATERIALS AND METHODS

PDX mouse models

The establishment and characterization of the LuCaP PDX models were described previously (83). PDXs were propagated in vivo in male NOD-scid IL2R-gamma-null (NSG) mice (cat#005557). The collection of tumors for the establishment of PDX lines was approved by the University of Washington Human Subjects Division IRB (IRB #2341). PDX lines were evaluated using histopathology by at least two expert pathologists, and histological phenotypic subtype annotations were orthogonally validated based on transcriptome-derived signature marker expression scores to define phenotypes (4,5,22): adenocarcinoma AR-positive (ARPC), neuroendocrine positive (NEPC), and AR-low, neuroendocrine negative (ARLPC). Resected PDX tumors (300-800 mm^3) were divided into ~50mg to ~100mg pieces and stored at -80°C. Animal studies were approved by the Fred Hutchinson Cancer Center (FHCC) IACUC (protocol 1618) and performed in accordance with the NIH guidelines. For the current study, blood was collected by cardiac puncture from animals bearing PDX tumors (measurable size 300-1400 mm^3).

Human subjects

UW cohort: Blood samples were collected from men with metastatic castration resistant prostate cancer at the University of Washington (collected under University of Washington Human Subjects Division IRB protocol number CC6932 between years 2014-2021). Patients in this study have provided written informed consent for research participation. In this study, 61 plasma samples from 31 patients were analyzed. After initial ultra-low pass whole genome sequencing (ULP-WGS) analysis, 47 plasma samples from 27 patients with sufficient tumor fraction (> 3%, based on initial ichorCNA analysis using GRCh37 genome build) and three additional samples not meeting the threshold but with clear AR amplification seen in manual curation (FH0243_E_1_A, FH0345_E_1_A, FH0482_E_1_A) were retained for further high depth of coverage whole genome sequencing (WGS) analysis. All samples were de-identified prior to ctDNA analysis and we employed a double blinded approach for evaluating clinical phenotype predictions.

DFCI cohort I: Plasma was collected from men diagnosed with mCRPC and treated at the Dana-Farber Cancer Institute (DFCI), Brigham and Women's Hospital, or Weill Cornell Medicine (WCM) between April 2003 and August 2021. All patients provided written informed consent for research participation and genomic analysis of their biospecimen and blood. The use of samples was approved by the DFCI IRB (#01-045 and 09-171) and WCM (1305013903) IRBs. The ULP-WGS data at mean coverage 0.5x (range 0.3x – 0.9x) for 101 patients were published previously (25).

DFCI cohort II: Plasma samples in this cohort were collected from men diagnosed with mCRPC and treated at the Dana-Farber Cancer Institute (DFCI). All patients provided written informed consent for blood collection and the analysis of their clinical and genetic data for research purposes (DFCI Protocol # 01-045 and 11-104). WGS data at mean coverage 27x (range 11x – 44x) (75), and ULP-WGS data at mean coverage 0.13x (range 0.07x – 0.18x) (76,80) were downloaded from dbGAP accession phs001417. Eleven samples from six patients had matching WGS and ULP-WGS with paired-end reads, necessary for analysis by Griffin. Prostate specific antigen (PSA, ng/mL) values and treatment at the time of the blood draw were previously published (76). The six patients were treated for adenocarcinoma using Abiraterone, Enzalutamide, or Bicalutamide, or the patients had detectable levels of PSA.

Healthy donor plasma cfDNA WGS data used in this study were obtained from previously published studies. Two samples (HD45 and HD46, both male) with coverage of 13x and 15x, respectively, were accessed from dbGAP under accession phs001417 (75,80). These donors were consented under DFCI protocol IRB (# 03-022). Thirteen healthy donor plasma cfDNA WGS data (12 male: NPH002, 03, 06, 07, 12, 18, 23, 26, 33, 34, 35, 36; 1 female (used in admixtures): NPH004) with coverages between 13.5x – 27.6x were obtained from the European Phenome Archive (EGA) under accession EGAD00001005343 (42).

PDX plasma processing

Blood samples were collected from NSG mice bearing subcutaneous PDX tumors at the time of sacrifice. The PDX lines were maintained at vivaria in the University of Washington and FHCC. The blood was processed following methods described for human plasma DNA processing for subsequent DNA isolation. Blood was collected in Sarstedt Micro sample tube K3 EDTA tubes and processed within 4 hours. All blood samples were sequentially double spun, first at 2500g for 10 minutes followed by a 16000g centrifugation of the plasma fraction for 10 minutes at room temperature. For each PDX line, 4-8 mouse plasma samples were pooled. Processed plasma samples were preserved in clean, screw-capped cryo-microfuge tubes and stored at -80°C prior to cfDNA isolation.

Cell-free DNA isolation

The QIAamp Circulating Nucleic Acid Kit was used to isolate cfDNA from PDX mouse-derived plasma using the recommended protocol. The pooled plasma samples from 4-8 mice for each PDX line contained 1.9 to 3 mL total plasma volume for each line. The filter retention-based cfDNA kit method does not implement any fragment size class enrichment. Isolated cfDNA was quantified using the Qubit dsDNA HS assay (Invitrogen) and the cfDNA fragment size profiles were analyzed using TapeStation HS D5000 and HS D1000 assays (Agilent).

Cell-free DNA library preparation and sequencing

For LuCaP PDX mouse plasma samples, NGS libraries were prepared with 50ng input cfDNA. Illumina NGS sequencing libraries were prepared with the KAPA hyperprep kit, adopting nine cycles of amplification, and purified using lab standardized SPRI beads. We used KAPA UDI dual indexed library adapters. Library concentrations were balanced and pooled for multiplexing and sequenced using the Illumina HiSeq 2500 at the Fred Hutch Genomics Shared Resources (200 cycles) and Illumina NovaSeq platform at the Broad Institute Genomics Platform Walkup-Seq Services using S4 flow cells (300 cycles). To match with Illumina HiSeq 2500 data, truncated 200 cycles FASTQ files were generated (100 bp paired end reads).

Clinical patient plasma samples collected at University of Washington (UW cohort) were submitted to the Broad Institute Blood Biopsy Services. Briefly, cfDNA was extracted from 2 mL double-spun plasma and ultra-low-pass whole genome sequencing (ULP-WGS) to approximately 0.2x coverage was performed. The ichorCNA pipeline was used to estimate tumor DNA content (i.e., tumor fraction, see below). Forty-seven samples (from 31 patients) had either ≥ 5% tumor fraction or ≥ 2% tumor fraction with AR amplification observed in ichorCNA and were subsequently sequenced to deeper WGS coverage (~20x).

Cell-free DNA sequencing analysis and mouse subtraction

All cfDNA sequencing data used in this study were realigned to the hg38 (GRCh38) human reference genome (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz). FASTQ files were realigned using BWA (v0.7.17) mem (84). The complete alignment pipeline including configuration settings may be access at https://github.com/GavinHaLab/fastq_to_bam_paired_snakemake.

For PDX ctDNA whole-genome sequence data, we performed mouse genome subtraction following the protocol described previously (85), wherein reads were aligned using BWA mem to a concatenated reference consisting of both human (hg38) and mouse (mm10, GRCm38.p6, http://igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Mus_musculus/NCBI/GRCm38/Mus_musculus_NCBI_GRCm38.tar.gz) reference genomes. Read pairs where both reads aligned to the human reference genome were retained and all other read pairs were removed. Then, remaining reads were re-aligned to the human-only reference. Finally, the GATK best practices workflow was applied to each sample (86); the complete mouse subtraction pipeline used in this study, including tool versions and parameters, can be accessed at https://github.com/GavinHaLab/PDX_mouseSubtraction. Following mouse subtraction samples with < 3X depth were removed for downstream analysis.

Cell cycle progression (CCP) score calculation

The 31-gene cell cycle proliferation (CCP) signature (68) was computed from RNAseq data using GSVA (87). The single-sample enrichment scores were calculated with default parameters using genome-wide log2 FPKM values as input for the 31 genes.

Differential mRNA expression analysis

RNA isolation of 102 tumors from 46 LuCaP PDX samples was performed as described previously (11). RNA concentration, purity, and integrity was assessed by NanoDrop (Thermo Fisher Scientific Inc) and Agilent TapeStation and RNA RIN >=8 was retained for library preparation. RNA-Seq libraries were constructed from 1 ug of total RNA using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit according to the manufacturer's protocol. Barcoded libraries were pooled and sequenced by Illumina NovaSeq 6000 or Illumina HiSeq 2500 generating 50 bp paired end reads. Sequencing reads were mapped to the hg38 human reference genome and mm10 mouse reference genomes using STAR.v2.7.3a (88). All subsequent analyses were performed in R-4.1.0. Sequences aligning to the mouse genome and therefor derived from potential contamination with mouse tissue were removed from the analysis using XenofilteR (v1.6) (89). Gene level abundance was quantitated using the R package GenomicAlignments v1.32.0 summarizeOverlaps function using mode=IntersectionStrict, restricted to primary aligned reads. We used refSeq gene annotations for transcriptome analysis. Transcript abundances (FPKM) were input to edgeR v3.38.1 (90), filtered for a minimum expression level using the filterByExpr function with default parameters, and then limma v3.52.1 voom was used for differential expression analysis of NEPC vs. ARPC and ARLPC vs. ARPC. We then filtered the results using a list of 1,635 human transcription factors published previously

(91), which resulted in 514 genes with FDR<0.05 and $\log_2$ fold change > 1.58. Out of these 514, deregulation of gene expression for 404 transcription factor genes delineated ARPC from NEPC.

Cleavage Under Targets & Release Using Nuclease (CUT&RUN)

CUT&RUN is an antibody targeted enzyme tethering chromatin profiling assay in which controlled cleavage by micrococcal nuclease releases specific protein-DNA complexes into the supernatant for paired-end DNA sequencing analysis. We performed CUT&RUN assays for three histone modifications, H3K27ac, H3K4me1, and H3K27me3, according to published protocols (48). We performed CUT&RUN on LuCaP PDX tumors using ~75mg flash-frozen tissue pieces.

Paired-end (50 bp) sequencing was performed and reads were aligned using bowtie2 v2.4.2 (92) to the hg38 human reference assembly. Aligned reads were processed as described in the SEACR protocol (https://github.com/FredHutch/SEACR#preparing-input-bedgraph-files). Peaks were called using SEACR version 1.3 (49) using “stringent” settings and with reference to paired IgG controls. BigWig files were prepared using bamCoverage in deepTools 3.5.0 (93). Genomewide peak heatmap, targeted heatmap, and respective profiles were plotted using deepTools v3.5.0. bigwig formatted files for each phenotype were obtained using the mean function in wiggletools 1.2.8. and deepTools computeMatrix. Phenotype-specific informative region coordinates were obtained from diffBind v3.5.0, and the top 10,000 most significant regions (all with FDR < 0.05) differentially open between ARPC and NEPC lines were used for downstream feature analyses (see Gene body and promoter region selection for additional subsetting criteria applied on a feature-by-feature basis). For heatmaps and profiles the plotHeatmap function was used. We utilized the “Peak Center” option to derive desired heatmaps. These steps were all performed for H3K27ac, H3K4me1 and H3K27me3 antibodies. Scaled heatmap profiles’ area under the curve (AUC; ±1.5kbp) and peak height at the profile center were estimated using deepStats v0.4 (https://zenodo.org/record/3668336) (comparable profiles are scaled to 10 units).

Differential histone post-translational modification (PTM) analysis

Differential PTM analysis was performed with the Diffbind version 2.16.0 package (94) in R-4.0.1 using standard parameters (https://bioconductor.riken.jp/packages/3.0/bioc/html/DiffBind .html). ARPC, NEPC and ARLPC samples were grouped by histopathological and transcriptome signature defined phenotypes described in the “PDX mouse models” section (**Supplementary Table S2**). Samples were loaded with the dba function, reads counted with the dba.count function, and contrast specified as phenotype with dba.contrast and a minimum members of 2. Differential peak sites were computed with the dba.analyze function with default settings. Differential peak binding of NEPC and ARLPC was computed against ARPC samples. Unique binding sites in NEPC and ARLPC were catalogued using bedtools v2.29.2 (95). Intergroup differentially bound peaks were annotated using ChIPseeker 1.28.3 (96) and TxDb.Hsapiens.UCSC.hg38.knownGene 3.2.2 in R 4.1.0.

ATAC-Seq analysis

ATAC-Seq sequence data for 15 tumor samples from 10 PDX lines were published previously (9). These lines included LuCaP PDX lines with ARPC (23.1, 77, 78, 81, 96) and NEPC (three replicates of 173.1, two replicates each of 49, 93, 145.1, and one replicate of 145.2) phenotypes. Paired end reads were aligned using bowtie2 v2.4.2 (92) to the UCSC hg38 human reference assembly with the "very-sensitive" "-k 10" settings. Peaks were called using Genrich version 0.6.1 (https://github.com/jsh58/Genrich). Differential binding analysis was performed using Diffbind version 3.5.0 package in R version 4.1.0. ENCODE blacklisted regions were excluded using hg38-blacklist.v2 (97) (https://github.com/Boyle-Lab/Blacklist). Phenotype specific regions were isolated by first selecting for positive fold change open chromatin enrichment and then using Intervene 0.6.5 (98) where regions were considered overlapping if they shared at least 1 bp with another phenotype. Regions with FDR adjusted p-values < 0.05 were then subset to those overlapping the 3,380,000 established TFBSs (338 TFs x 10,000 binding sites, see Griffin analysis for site selection) by at least 1 bp using BedTools v2.30.0 Intersect. Only regions that overlapped an established TFBS from those lists were retained. For analyses restricted to 10,000, 1,000, or 100 sites, sites were ranked and chosen by adjusted p-value.

Nucleosome profiling of ctDNA

Griffin is a method for profiling nucleosome protection and accessibility on predefined genomic loci (51). For this study, Griffin (v0.1.0) was used and can be found on GitHub (https://github.com/adoebley/Griffin/releases/tag/v0.1.0). The analysis was performed as follows: First, GC bias was quantified for each sample using an approach described previously (99). Briefly, for each possible fragment length and GC content, the number of reads in a bam file and the number of genomic positions with that specific length and GC content were counted. The GC bias for each fragment length and GC content was calculated by dividing the number of observed reads by the number of observed genomic positions for that fragment length and GC content. The GC bias for all possible GC contents at a given fragment length was then normalized to a mean bias of 1. GC biases were then smoothed by taking the median of values for fragments with similar lengths and GC contents (k nearest neighbors smoothing) to generate smoothed GC bias values.

After GC correction, nucleosome profiling was performed in each sample. For each mappable site of interest, fragments aligning to the region ± 5000 bp from the site were fetched from the bam file. Fragments were filtered to remove duplicates and low-quality alignments (<20 mapping quality) and by fragment length. Nucleosome size fragments (140-250 bp) were retained and used in all down-stream Griffin analyses. Fragments were then GC corrected by assigning each fragment a weight of 1/GC_bias for that given fragment length and GC content. The fragment midpoint was identified and the number of weighted fragment midpoints in 15bp bins across the site were counted. For composite sites, all sites of a given type (such as all sites for a given transcription factor) were summed together to generate a single coverage profile. Individual or composite coverage profiles were normalized to a mean coverage of 1 in the ± 5000bp region surrounding the site. Finally, sites were smoothed using a Savitsky-Golay filter with a window length of 165bp and a polynomial order of 3. The window ± 1000 bp around the site was retained for plotting and feature extraction when plotting sites, shading illustrates the 95% confidence interval within sample groups. Features extracted from individual or composite sites included:

a) “mean central coverage,” the mean coverage between -30 to 30 bp relative to the site center,

b) “mean window coverage,” the mean coverage between -990 to 990 bp relative to the site center, and

c) “max wave height,” the absolute difference between the minimum coverage within the window from -120 to 30 bp and maximum coverage in the window from 31 to 195 bp relative to the TSS.

Transcription factor binding site (TFBS) selection from GTRD

TFBS identified using ChIP-seq were downloaded from the GTRD database version 19.10 (https://gtrd.biouml.org/downloads/19.10/chip-seq/Homo%20sapiens_meta_clusters.interval.gz). This database contains binding sites (meta-clusters) that were observed in one or more ChIP seq experiment. Low mappability sites were excluded by examining the mean mappability score in a window around each site (+/- 5000 bp). Mappability information (hg38 Umap multi-read mappability for 50bp reads) was obtained from UCSC genome browser (100) (https://hgdownload.soe.ucsc.edu/gbdb/hg38/hoffmanMappability/k50.Umap.MultiTrackMappability.bw). Highly mappable sites (>0.95 mean mappability) were retained for further analysis. 338 TFs were selected for analysis using three criteria: (i) TF was contained in GTRD, (ii) had at least 10,000 highly mappable binding sites on autosomes (chr1-22) in GTRD, (iii) TF was present in the CIS-BP database (101) (CIS-BP v2.00 downloaded from http://cisbp.ccbr.utoronto.ca/bulk.php) and had a known binding motif ('TF_status' is not N). Unless otherwise noted, analyses utilized the top 1,000 TFBSs ranked by the highest 'peak.count' across all experiments as computed by GTRD (70). In addition, in the case of AR and ASCL1 we also compared the top 1,000 vs the top 10,000 sites chosen with the same 'peak.count' criterion.

After intersecting these 338 TFs with the 404 differentially expressed TFs identified through RNA-Seq 108 remained. On both the 108 and prostate-specific 41 TFs (described below) we performed unsupervised hierarchical clustering of central window mean values (see Griffin analysis). Hierarchical clustering was performed using the Ward.D2 method with Euclidean distance and complete linkage settings; the groupings were determined using cutree_cols=2 for columns (LuCaP CRPC phenotypes) and cutree_rows=13 for rows (TFs) on the dendrograms.

To generate a prostate lineage-specific TF set, we first merged GTRD metadata (file; http://gtrd.biouml.org:8888/downloads/current/metadata/ChIP-seq.metadata.txt & http://gtrd.biouml.org:8888/downloads/current/metadata/cell_types_and_tissues.metadata.txt). We identified human prostate lineage-specific experiments by restricting the "species" field to "Homo sapiens" and the "title" (tissue or cell type) field by performing a string match of the following {"Prostate", "prostate", "LNCaP", "DU145", "PrEC"}. This resulted in a list of 1,086 prostate lineage ChIP seq experiments. Then, we selected metapeaks from the "Homo_sapiens_meta_clusters.interval" file which had been observed in at least one of the prostate lineage experiments using the "exp.set" field. This resulted in a set of 82 TFs. We then filtered the peaks by mappability and kept only highly mappable peaks (as described above). We excluded any TF that wasn't in the initial set of 338 TFs (this removed ChIP targets that weren't true TFs, lacked a known binding site, or didn't have 10,000 total autosomal peaks in GTRD). Of the remaining TFs, we analyzed those with 1,000 highly mappable peaks on autosomes in prostate lineage experiments, resulting in 41 TFs. 20 out of 41 of these TFs overlapped the list of 108 differentially expressed TFs by RNAseq of the PDX tumors. Note that the top 1000 sites for each of the 41 TFs were different than in same TFs of the 338 set because sites must meet the criteria of being derived from at least one experiment involving prostate tissue or cell lines.

Transcription factor binding site (TFBS) selection from other sources

For AR we further considered 17,619 sites identified through ChIP-seq by Pomerantz et al. (56) (which overlapped 10.9% of the GTRD top 1,000 using bedtools), 41,633 sites identified by Severson et al. (57) across four metastatic tumors (which overlapped 99.4% of the GTRD top 1,000). For ASCL1 we obtained 11,124 ChIP-seq sites from Cejas et al. (9) (which overlapped 60.9% of the GTRD 1,000). All of these site lists were lifted over from genome build GRCh37 to GRCh38. No mappability filtering was applied so that all possible sites from these prostate experiments and studies were considered.

Phenotype-lineage specific gene marker selection

We selected 47 genes comprising 12 ARPC and 35 NEPC lineage markers established previously (4,5,58,59) and confirmed by differential expression analysis from PDX tumor RNA-Seq data (**Supplementary Table S3**). In tissues, AR and NE activities were measured on lineage determinant signature gene's mRNA expression (GSVA score)(87). The 47 selected gene list comprises the majority of these signature sets of genes defining mPC characteristic phenotypes or phenotypic activities.

Gene body and promoter region selection

For individual gene body and promoter analyses Ensembl BioMart v104 (hg38) (102) was used to directly retrieve protein coding transcript start (TSS) and end (TES) coordinates. For promoter region analysis the window ±1000 bp relative to the TSS was considered. For gene body analysis, the region between the TSS and TES was considered. In the case of genes with multiple transcripts, analyses were limited to the longest transcript resulting in 19,336 regions. In downstream analysis of LuCaP PDX cfDNA, if any lines did not meet specific criteria in a region (including differentially open histone modification regions) that feature/region combination was excluded from analysis, leading to a variable lower number of regions considered based on the feature. These criteria included requiring at least 10 total fragments in a region for all Fragment size analysis (see below) and a non-zero number of "short" and "long" fragments for the short-long ratio; short-long ratios less than 0.01 or greater than 10.0 were also excluded as outliers. For Phasing analysis (see below) we also excluded amplitude components and thus NPS where individual components were 0, or where the ratio was less than 0.01 or greater than 10.0, indicative of insufficient coverage. In the case of mean phased nucleosome distance, if no peaks were identified or the value in a region exceeded 500 (indicative of highly irregular/sparse pileups also from low coverage) those regions were also excluded. Any region with no coverage in a line was excluded from all analyses. This resulted in gene lists that differed in numbers between genomic contexts and feature types.

Cell-free DNA fragment size analysis

Fragments were first filtered to remove duplicates and low-quality alignments (<20 mapping quality) and by fragment length (15-500 bp). In individual genomic loci/windows, we computed the fragment short-long ratio (FSLR) as the ratio of short (15 - 120 bp) to long (140 - 250 bp) fragments. We also calculated the mean, median absolute deviation (MAD: $median(|X_i - median(X)|)$), and coefficient of variation (CV: $\frac{\sigma}{\mu}$ where σ = standard deviation, μ = mean) of the fragment length distribution for each selected window. The fragment size analysis code and implementation used in this study can be accessed at https://github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/FragmentAnalysis.

Nucleosome phasing analysis (TritonNP)

Fragments were first filtered to remove duplicates and low-quality alignments (<20 mapping quality) and by fragment length (nucleosome-sized: 140-250 bp). Next, we performed fragment-level GC bias correction utilizing the same pre-processing method defined in Griffin. A band-pass filter was then applied to the corrected coverage in each region of interest by taking the Fast Fourier Transform (FFT) (scipy.fft v1.8.0) (103) and removing high-frequency components corresponding to frequency components < 146 bp before reconstructing the signal. This cutoff was chosen to ensure that periodic fit signal for downstream evaluation must come from the minimum possible inter-nucleosome distance, thus excluding peak pileups that would not indicate an overall trend in nucleosome phasing. Local peak calling was then done on the smoothed signal to infer average inter nucleosome distance or “phased nucleosome distance” by finding maxima directly. To quantify clarity of overall phasing we took the average frequency amplitude in two bands corresponding to stably bound, well-phased nucleosomes (180-210 bp) and a baseline (150-180 bp), with the former measuring the strength of typically aligned nucleosomes and the latter giving a measure of the underlying signal strength not coming from either high frequency noise or low frequency shifts in total coverage. The ratio of these two amplitude averages forms the Nucleosome Phasing Score (NPS). Because peak locations are assumed to be independent of copy number alterations or depth, and the NPS by virtue of being a ratio divides out any confounding DNA/depth variation between sites, both features are taken as agnostic of CNAs or variable depth. Code and implementation of the method can be found at https://github.com/denniepatton/TritonNP.

ctDNA tumor-normal admixtures and benchmarking

Admixtures for evaluating benchmarking performance were constructed using 5 ARPC (LuCaP 35, 35CR, 58, 92, 136CR) and 5 NEPC (LuCaP 49, 93, 145.2, 173.1, 208.4) lines mixed to 1%, 5%, 10%, 20%, and 30% tumor fraction with a single healthy donor plasma line (NPH023, EGAD00001005343) for use in binary classification (50 admixes), and in mixtures of 1%, 3%, 5%, 10%, 20%, and 30% tumor fraction at ARPC:NEPC ratios, of 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, and 1.0 in all possible combinations (810 admixes) for mixture model evaluation. All admixes were mixed at ~25X mean coverage, assuming 100% tumor fraction in post-mouse subtracted PDX sequencing data. After extracting chromosomal DNA (chr1-22, X, Y) with SAMtools v1.14 (104) and removing duplicates with Picard (https://broadinstitute.github.io/picard/), SAMtools was used to merge BAM files. To evaluate the ultra-low pass WGS performance, admixtures were then down-sampled using SAMtools to the number of reads corresponding to 1X or 0.2X. During unsupervised benchmarking of each admixture, the healthy donor and the LuCaP line used in the admixture were excluded from the generation of feature distributions to ensure the model would not learn from the lines being interrogated. The admixture pipeline used in this study can be accessed at https://github.com/GavinHaLab/Admixtures_snakemake.

Supervised binary classification of ARPC and NEPC

Binary classification of ARPC and NEPC subtypes using individual region and feature combinations was conducted using XGBoost v1.4.2 ‘XGBClassifier’ implemented in Python with default parameters. Features included NPS and Mean Phased Nucleosome Distance (see Phasing analysis) in histone modification regions, promoters, and gene bodies; fragment size mean, short-long ratio, and coefficient of variation (see Fragment size analysis) in histone modification regions, promoters, and gene bodies; central and window coverage (see Griffin analysis) in promoters, composite TFBSs, and composite differentially open chromatin regions identified through ATAC-Seq; and Max Wave Height (See Griffin analysis) in promoters. We applied stratified 6-fold cross-validation where two ARPC samples and one NEPC sample were held out in each fold. This was repeated 100 times and performance was computed using area under the receiver operating characteristic (ROC) curve (AUC) and 95% confidence intervals for each individual feature and region combination. Code and implementation of the method can be found at https://github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/SupervisedLearning.

Tumor fraction estimation

Tumor fractions from patient plasma samples were assessed using ichorCNA (80) with binSize 1,000,000 bp and both GRCh37 and GRCh38 reference genomes. Default tumor fraction estimates reported by ichorCNA were used. See https://github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/ichorCNA_configuration for complete configuration settings.

Phenotype class prediction model (ctdPheno)

We developed a probabilistic model to classify the mCRPC phenotype (ARPC or NEPC) in an individual patient plasma ctDNA sample. This is a generative mixture model that is unsupervised—it does not train on the patient cohort of interest. However, the model accepts the pre-estimated tumor fraction from ichorCNA for the given patient ctDNA sample, as well as the pre-computed ctDNA features values from the LuCaP PDX ctDNA and healthy donor ctDNA as prior information. For each patient ctDNA sample, it fits specific feature values against the pure PDX LuCaP models, shifted towards healthy based on the estimated tumor fraction. The expected feature value (mean μ and standard deviation σ) from each phenotype *k* for feature i were taken from the mean of LuCaP PDX samples ($\mu_{i,k}$) or taken from the mean of a panel of normals *H* ($\mu_{i,H}$, male only, n = 14; see Healthy Donor cohort). Assuming a Gaussian distribution, feature values were shifted such that the shifted $\mu'_{i,k}$, $\sigma'_{i,k}$took the form:

$$\mu'_{i,k} = \alpha\mu_{i,k} + (1 - \alpha)\mu_{i,H}$$

$$\sigma'_{i,k} = \sqrt{\alpha\sigma_{i,k}^2 + (1 - \alpha)\sigma_{i,H}^2}$$

where α is the tumor fraction estimate for each test sample. In the final model, four features were used: composite open chromatin regions (central and window mean coverage) for specific phenotypes (ARPC and NEPC) identified from the LuCaP PDX ATAC-Seq analysis using Griffin (see Griffin analysis). For each feature *i*, we then found the probability that the observed sample came from a mixture of the tumor-fraction-corrected Gaussian distributions, where θ is the NEPC mixture weight:

$$p_i(x|\theta) = \theta p(x|k = NEPC) + (1 - \theta)p(x \mid k = ARPC)$$

The θ parameter is estimated by maximizing the joint log-likelihood L for a given patient sample:

$$\theta' = \underset{\theta}{\operatorname{argmax}}[L(x|\ \theta)]$$

$$where\ L(x|\ \theta) = \sum_i \ln[p_i(x|\theta)]$$

θ has range [0,1], where higher values indicate an increased probability of the sample having a NEPC phenotype and was used as the NEPC prediction score metric. Code and implementation of the method can be found at https://github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/ctdPheno.

Phenotype heterogeneity prediction and quantification (Keraon)

We developed an analytical model to directly estimate the contributing fractions of ctDNA from different mCRPC phenotypes (ARPC and NEPC) in individual patient plasma ctDNA samples. Like ctdPheno this model is unsupervised and does not require training on the patient cohort of interest. However, the model accepts the pre-estimated tumor fraction from ichorCNA for the given patient ctDNA sample, as well as the pre-computed ctDNA features values from the LuCaP PDX ctDNA and healthy donor ctDNA as prior information (see Class phenotype prediction model).

As a pre-processing step, the model first computes the mean vector μ_i and covariance matrix Σ_i for each anchor class *i* in K, under the assumption that each subtype (including healthy) fits a multivariate Gaussian distribution. Based on model constraints, K - 1 non-correlated features fully specify the system, and so for ARPC:NEPC:Healthy (K = 3) fraction estimation we limited analyses to sets of two features of interest (F = 2).

Next, for each sample defined by some location in feature space ***v*** and estimated tumor fraction *t* we first performed a change of basis to translate the sample's location from feature space to class space, where each (not necessarily orthogonal) axis defined a single phenotype, and the origin represented pure healthy. If F = K -1, this was accomplished by solving the determined, linear matrix equation for the shifted basis components **X**:

$$\boldsymbol{BX = S}$$

Where **B** = [$\mu_{i \neq HD}$- μ_{HD}] is the matrix defining all basis vectors from the healthy mean anchor to each phenotype mean anchor, and **S** is the vector from the healthy mean anchor to the sample of interest, **S** = ***v*** - μ_{HD}. If the system is overdetermined (F > K – 1), least squares was used to estimate the approximate solution. This step allows us to learn where in the class space the sample lies, which determined how estimates were evaluated:

1. Anchor Space: if all basis components are positive then the sample lies within the volume of order K – 1 which has vertices defined by the class means. The relative ratio of basis component magnitudes in the direction of each class are corrected by estimated tumor fraction directly: $BC_{i \neq HD} = \frac{X_i}{\sum X} t$
2. Contra Space: if all basis components are negative then the sample lies within the volume of order K – 1 which forms a reflection of that formed by the class vertices about healthy. Component fractions for each basis are computed to capture the inverse distance from the healthy anchor, such that $BC_{i \neq HD} = \frac{X_i+1}{\sum(X+1)} t$
3. Extra Space: if some basis components are positive but others are negative, the sample lies in some space outside of the anchor or contra space. In this case only positive contributions are considered, such that $BC_{i \neq HD} = \frac{X_i}{\sum X} t$ for all *i* such that $X_i > 0$.

The tumor fraction normalized basis component estimates BC have range [0,1], where values directly correspond to the total fraction of each class in the sample.

Code and implementation of the method can be found at https://github.com/denniepatton/Keraon.

Analysis and classification of clinical patient samples

After establishing feature distributions using the LuCaP PDX lines and normal panel as described above, both models were applied to three clinical patient cohorts (see Human subjects for cohort information).

Binary class prediction: Initial scoring using ctdPheno was run on DFCI cohort I, consisting of 101 ULP-WGS samples with paired-end reads. Tumor fraction estimates predicted by ichorCNA and tumor phenotype classifications were obtained from the original study (25). A prediction score threshold of 0.3314 for calling NEPC was chosen because it offered an optimal performance for sensitivity (90%) and specificity (97.5%), where sensitivity is the true positive rate for identifying NEPC samples $\left(\frac{TP}{TP+FN}\right)$ and specificity is the true negative rate for identifying ARPC samples $\left(\frac{TN}{TN+FP}\right)$. Alternative thresholds maximizing sensitivity and specificity were 0.1077, at which 95% sensitivity was achieved with a lower specificity of 93.8%, and 0.3769 with a lower sensitivity of 81.0% but higher specificity of 98.8%. To compare these predictions with cfDNA methylation (cfMeDIP-seq) classification on the same plasma samples in DFCI cohort I, the concordance was computed between the ctdPheno NEPC prediction score and the cfMeDIP NEPC score obtained from the original study using a 0.15 threshold (25).

We then validated the model on two cohorts, beginning with the already published DFCI cohort II (75,76,80). We restricted our analysis to eleven samples from six patients with matched ULP-WGS and WGS data with paired-end reads. Tumor fraction estimates from ichorCNA were obtained from the original study (80). All samples were considered adenocarcinoma (ARPC) based on clinical histories (see Human subjects). The scoring threshold of 0.3314, determined from DFCI cohort I was used for phenotype classification.

For the *UW cohort*, consisting of 47 samples from 27 patients (average 22.13X depth of coverage sequencing), ichorCNA was used to estimate sample tumor fractions as described above (GRCh38), while clinical phenotype was determined from clinical histories and expert chart review. We evaluated model performance on matched ULP-WGS and WGS data for unambiguous clinical phenotypes of ARPC and NEPC. The chosen scoring threshold of 0.3314 was used, and the fraction of correctly predicted ARPC (n=26) and NEPC (n=5) was computed. The remaining 16 samples with mixed histologies were not evaluated for performance in ctdPheno.

Phenotype prediction and proportion estimation: Keraon does not require de novo threshold selection, and so all clinical cohorts were treated as validation sets. Based on the Mean Absolute Error (MAE) of 2.8% for estimating NEPC fraction garnered in the heterogenous mixture benchmarking, this value was chosen as the minimum NEPC fraction threshold for calling presence of NEPC in WGS cohorts. The same tumor fraction estimates used by ctdPheno in ULP were utilized by Keraon, with standard classification conducted on pure clinical phenotypes. The 16 samples with mixed phenotypes in the UW cohort were evaluated both qualitatively and based on the 2.8% threshold in the absence of quantifiable burden estimates from histories.

STATISTICAL ANALYSIS

Quantification of and statistical approaches for high-throughput sequencing data analysis are described in the methods above. When non-parametric distributions (not normally distributed) of numerical values of a particular parameter in a population were compared (using boxplots or in tables), the two-tailed Mann-Whitney U test (also known as the Wilcoxon Rank Sum test; scipy.stats.mannwhitneyu, (103) was used to test if any two distributions being compared were significantly different, with Benjamini-Hochberg (statsmodels.stats.multitest.fdrcorrection, https://www.statsmodels.org) correction applied in multiple testing scenarios. All boxplots represent the median with a centerline, interquartile range (IQR) with a box, and first quartile – 1.5 IQR and third quartile + 1.5 IQR with whiskers. PCA was conducted in Python (sklearn.decomposition.PCA; https://scikit-learn.org)

DATA AVAILABILITY

The LuCaP patient derived xenograft (PDX) plasma ctDNA sequencing data generated in this study can be accessed under NCBI BioProject accession PRJN900550 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA90055). The processed patient plasma data can be accessed at https://github.com/GavinHaLab/CRPCSubtypingPaper/tree/main/Data. The raw sequencing data generated for the UW cohort are not publicly available because patients did not consent to genomic data sharing but are available upon reasonable request from the corresponding author. This paper also analyzes existing, publicly available data, including LuCaP PDX RNA-Seq (GSE199596) and ATAC-Seq data (GSE156292). The CUT&RUN processed data can be accessed at https://github.com/nielOnav/LuCaP_nucleosome_profile. Published data for DFCI Cohort I was obtained from the authors (25) after establishing a data use agreement with the Dana-Farber Cancer Institute.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

REFERENCES

1. Karantanos T, Corn PG, Thompson TC. Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. Oncogene. Nature Publishing Group; 2013;32:5501–11.

2. Ryan CJ, Smith MR, de Bono JS, Molina A, Logothetis CJ, de Souza P, et al. Abiraterone in Metastatic Prostate Cancer without Previous Chemotherapy. N Engl J Med. Massachusetts Medical Society; 2013;368:138–48.

3. Scher HI, Fizazi K, Saad F, Taplin M-E, Sternberg CN, Miller K, et al. Increased Survival with Enzalutamide in Prostate Cancer after Chemotherapy. Cabot RC, Harris NL, Rosenberg ES, Shepard J-AO, Cort AM, Ebeling SH, et al., editors. New England Journal of Medicine. 2012;367:1187–97.

4. Beltran H, Prandi D, Mosquera JM, Benelli M, Puca L, Cyrta J, et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. Nature Medicine. Nature Publishing Group; 2016;22:298–305.

5. Bluemn EG, Coleman IM, Lucas JM, Coleman RT, Hernandez-Lopez S, Tharakan R, et al. Androgen Receptor Pathway-Independent Prostate Cancer Is Sustained through FGF Signaling. Cancer cell. Elsevier; 2017;32:474-489.e6.

6. Conteduca V, Oromendia C, Eng KW, Bareja R, Sigouros M, Molina A, et al. Clinical features of neuroendocrine prostate cancer. European Journal of Cancer. 2019;121:7–18.

7. Aggarwal R, Huang J, Alumkal JJ, Zhang L, Feng FY, Thomas GV, et al. Clinical and Genomic Characterization of Treatment-Emergent Small-Cell Neuroendocrine Prostate Cancer: A Multi-institutional Prospective Study. JCO. American Society of Clinical Oncology; 2018;36:2492–503.

8. Baca SC, Takeda DY, Seo J-H, Hwang J, Ku SY, Arafeh R, et al. Reprogramming of the FOXA1 cistrome in treatment-emergent neuroendocrine prostate cancer. Nat Commun. Nature Publishing Group; 2021;12:1979.

9. Cejas P, Xie Y, Font-Tello A, Lim K, Syamala S, Qiu X, et al. Subtype heterogeneity and epigenetic convergence in neuroendocrine prostate cancer. Nat Commun. 2021;12:5775.

10. Spetsieris N, Boukovala M, Patsakis G, Alafis I, Efstathiou E. Neuroendocrine and Aggressive-Variant Prostate Cancer. Cancers. Multidisciplinary Digital Publishing Institute; 2020;12:3792.

11. Labrecque MP, Coleman IM, Brown LG, True LD, Kollath L, Lakely B, et al. Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer. J Clin Invest. American Society for Clinical Investigation; 2019;129:4492–505.

12. Labrecque MP, Alumkal JJ, Coleman IM, Nelson PS, Morrissey C. The heterogeneity of prostate cancers lacking AR activity will require diverse treatment approaches. Endocrine-Related Cancer. Bioscientifica Ltd; 2021;28:T51–66.

13. Liu Y, Horn JL, Banda K, Goodman AZ, Lim Y, Jana S, et al. The androgen receptor regulates a druggable translational regulon in advanced prostate cancer. Science Translational Medicine. American Association for the Advancement of Science; 2019;11:eaaw4993.

14. Epstein JI, Amin MB, Beltran H, Lotan TL, Mosquera J-M, Reuter VE, et al. Proposed Morphologic Classification of Prostate Cancer With Neuroendocrine Differentiation. The American Journal of Surgical Pathology. 2014;38:756–67.

15. Annala M, Taavitsainen S, Khalaf DJ, Vandekerkhove G, Beja K, Sipola J, et al. Evolution of Castration-Resistant Prostate Cancer in ctDNA during Sequential Androgen Receptor Pathway Inhibition. Clinical Cancer Research. 2021;27:4610–23.

16. Aparicio AM, Shen L, Tapia ELN, Lu J-F, Chen H-C, Zhang J, et al. Combined Tumor Suppressor Defects Characterize Clinically Defined Aggressive Variant Prostate Cancers. Clinical Cancer Research. 2016;22:1520–30.

17. Carreira S, Romanel A, Goodall J, Grist E, Ferraldeschi R, Miranda S, et al. Tumor clone dynamics in lethal prostate cancer. Science translational medicine. 2014;6:254ra125.

18. Du M, Tian Y, Tan W, Wang L, Wang L, Kilari D, et al. Plasma cell-free DNA-based predictors of response to abiraterone acetate/prednisone and prognostic factors in metastatic castration-resistant prostate cancer. Prostate Cancer Prostatic Dis. Nature Publishing Group; 2020;23:705–13.

19. Sumanasuriya S, Seed G, Parr H, Christova R, Pope L, Bertan C, et al. Elucidating Prostate Cancer Behaviour During Treatment via Low-pass Whole-genome Sequencing of Circulating Tumour DNA. European Urology. 2021;80:243–53.

20. Ulz P, Belic J, Graf R, Auer M, Lafer I, Fischereder K, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. Nat Commun. Institute of Human Genetics, Medical University of Graz, A-8010 Graz, Austria. Department of Urology, Medical University of Graz, A-8036 Graz, Austria. Department of Internal Medicine I, Hospital Barmherzige Schwestern Linz, A-4020 Linz, Austria. Departme; 2016;7:12008.

21. Wyatt AW, Annala M, Aggarwal R, Beja K, Feng F, Youngren J, et al. Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. JNCI: Journal of the National Cancer Institute. Oxford University Press; 2018;110:78–86.

22. Nyquist MD, Corella A, Coleman I, De Sarkar N, Kaipainen A, Ha G, et al. Combined TP53 and RB1 Loss Promotes Prostate Cancer Resistance to a Spectrum of Therapeutics and Confers Vulnerability to Replication Stress. Cell Reports. 2020;31:107669.

23. Berger A, Brady NJ, Bareja R, Robinson B, Conteduca V, Augello MA, et al. N-Myc–mediated epigenetic reprogramming drives lineage plasticity in advanced prostate cancer. J Clin Invest. American Society for Clinical Investigation; 2019;129:3924–40.

24. Beltran H, Romanel A, Conteduca V, Casiraghi N, Sigouros M, Franceschini GM, et al. Circulating tumor DNA profile recognizes transformation to castration-resistant

neuroendocrine prostate cancer. J Clin Invest. American Society for Clinical Investigation; 2020;130:1653–68.

25. Berchuck JE, Baca SC, McClure HM, Korthauer K, Tsai HK, Nuzzo PV, et al. Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clinical Cancer Research. 2022;28:928–38.

26. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. Nature Publishing Group; 2018;563:579–83.

27. Wu A, Cremaschi P, Wetterskog D, Conteduca V, Franceschini GM, Kleftogiannis D, et al. Genome-wide plasma DNA methylation features of metastatic prostate cancer. J Clin Invest. American Society for Clinical Investigation; 2020;130:1991–2000.

28. Heitzer E, Auinger L, Speicher MR. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. Trends in Molecular Medicine. Elsevier Ltd; 2020;26:519–28.

29. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science [Internet]. American Association for the Advancement of Science; 2021 [cited 2021 Apr 12];372. Available from: https://science.sciencemag.org/content/372/6538/eaaw3616

30. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. Nature Publishing Group; 2019;570:385–9.

31. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. Cancer Discov. American Association for Cancer Research; 2020;10:664–73.

32. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nat Commun. 2021;12:5060.

33. Peneder P, Stütz AM, Surdez D, Krumbholz M, Semper S, Chicard M, et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. Nat Commun. Nature Publishing Group; 2021;12:3230.

34. Zhu G, Guo YA, Ho D, Poon P, Poh ZW, Wong PM, et al. Tissue-specific cell-free DNA degradation quantifies circulating tumor DNA burden. Nature Communications. Nature Publishing Group; 2021;12:2229.

35. Herberts C, Annala M, Sipola J, Ng SWS, Chen XE, Nurminen A, et al. Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. Nature. 2022;608:199–208.

36. Jiang P, Chan CWM, Chan KCA, Cheng SH, Wong J, Wong VW-S, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proceedings of the National Academy of Sciences of the United States of America. 2015;112:E1317-25.

37. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine. 2018;10:eaat4921.

38. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. Elsevier Inc.; 2016;164:57–68.

39. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, et al. Fragment Length of Circulating Tumor DNA. PLOS Genet. 2016;12:426–37.

40. Ramachandran S, Ahmad K, Henikoff S. Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates. Molecular Cell. Cell Press; 2017;68:1038-1053.e4.

41. Ulz P, Thallinger GG, Auer M, Graf R, Kashofer K, Jahn SW, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nature Genetics. Nature Publishing Group; 2016;48:1273–8.

42. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature Communications. 2019;10:4666.

43. Brahma S, Henikoff S. Epigenome Regulation by Dynamic Nucleosome Unwrapping. Trends in Biochemical Sciences. Elsevier; 2020;45:13–26.

44. Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat Rev Mol Cell Biol. 2017;18:548–62.

45. Yen K, Vinayachandran V, Batta K, Koerber RT, Pugh BF. Genome-wide Nucleosome Specificity and Directionality of Chromatin Remodelers. Cell. 2012;149:1461–73.

46. Rao S, Han AL, Zukowski A, Kopin E, Sartorius CA, Kabos P, et al. Transcription factor–nucleosome dynamics from plasma cfDNA identifies ER-driven states in breast cancer. Science Advances. American Association for the Advancement of Science; 2022;8:eabm4358.

47. Nguyen HM, Vessella RL, Morrissey C, Brown LG, Coleman IM, Higano CS, et al. LuCaP Prostate Cancer Patient-Derived Xenografts Reflect the Molecular Heterogeneity of Advanced Disease an--d Serve as Models for Evaluating Cancer Therapeutics. The Prostate. John Wiley & Sons, Ltd; 2017;77:654–71.

48. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Reinberg D, editor. eLife. eLife Sciences Publications, Ltd; 2017;6:e21856.

49. Meers MP, Tenenbaum D, Henikoff S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenetics & Chromatin. 2019;12:42.

50. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011;12:7–18.

51. Doebley A-L, Ko M, Liao H, Cruikshank AE, Kikawa C, Santos K, et al. Griffin: Framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. medRxiv. 2021;2021.08.31.21262867.

52. Soares LM, He PC, Chun Y, Suh H, Kim T, Buratowski S. Determinants of Histone H3K4 Methylation Patterns. Molecular Cell. 2017;68:773-785.e6.

53. Brady NJ, Bagadion AM, Singh R, Conteduca V, Van Emmenis L, Arceci E, et al. Temporal evolution of cellular heterogeneity during the progression to advanced AR-negative prostate cancer. Nat Commun. Nature Publishing Group; 2021;12:3372.

54. Wang YA, Sfakianos J, Tewari AK, Cordon-cardo C, Kyprianou N. Molecular tracing of prostate cancer lethality. Oncogene. Nature Publishing Group; 2020;39:7225–38.

55. Rapa I, Ceppi P, Bollito E, Rosas R, Cappia S, Bacillo E, et al. Human ASH1 expression in prostate cancer with neuroendocrine differentiation. Mod Pathol. Nature Publishing Group; 2008;21:700–7.

56. Pomerantz MM, Qiu X, Zhu Y, Takeda DY, Pan W, Baca SC, et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nat Genet. 2020;52:790–9.

57. Severson TM, Zhu Y, De Marzo AM, Jones T, Simons JW, Nelson WG, et al. Epigenetic and transcriptional analysis reveals a core transcriptional program conserved in clonal prostate cancer metastases. Molecular Oncology. 2021;15:1942–55.

58. Labrecque MP, Brown LG, Coleman IM, Lakely B, Brady NJ, Lee JK, et al. RNA Splicing Factors SRRM3 and SRRM4 Distinguish Molecular Phenotypes of Castration-Resistant Neuroendocrine Prostate Cancer. Cancer Research. 2021;81:4736–50.

59. Tsai HK, Lehrer J, Alshalalfa M, Erho N, Davicioni E, Lotan TL. Gene expression signatures of neuroendocrine prostate cancer and primary small cell prostatic carcinoma. BMC Cancer. 2017;17:759.

60. Jiang Z, Zhang B. On the role of transcription in positioning nucleosomes. PLOS Computational Biology. Public Library of Science; 2021;17:e1008556.

61. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics. Nature Publishing Group; 2019;20:207–20.

62. Oruba A, Saccani S, van Essen D. Role of cell-type specific nucleosome positioning in inducible activation of mammalian promoters. Nat Commun. 2020;11:1075.

63. Guo Y, Zhao S, Wang GG. Polycomb Gene Silencing Mechanisms: PRC2 Chromatin Targeting, H3K27me3 “Readout”, and Phase Separation-Based Compaction. Trends in Genetics. Elsevier; 2021;37:547–65.

64. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009;10:161–72.

65. Saxton DS, Rine J. Nucleosome Positioning Regulates the Establishment, Stability, and Inheritance of Heterochromatin in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences; 2020;117:27493–501.

66. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature. Nature Publishing Group; 2011;474:516–20.

67. Deal RB, Henikoff JG, Henikoff S. Genome-Wide Kinetics of Nucleosome Turnover Determined by Metabolic Labeling of Histones. Science. American Association for the Advancement of Science; 2010;328:1161–4.

68. Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. The Lancet Oncology. 2011;12:245–55.

69. Chereji RV, Bryson TD, Henikoff S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. Genome Biology. 2019;20:198.

70. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res. Oxford Academic; 2019;47:D100–5.

71. Arora VK, Schenkein E, Murali R, Subudhi SK, Wongvipat J, Balbas MD, et al. Glucocorticoid Receptor Confers Resistance to Antiandrogens by Bypassing Androgen Receptor Blockade. Cell. Elsevier; 2013;155:1309–22.

72. Mu P, Zhang Z, Benelli M, Karthaus WR, Hoover E, Chen C-C, et al. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. Science. American Association for the Advancement of Science; 2017;355:84–8.

73. Shukla S, Cyrta J, Murphy DA, Walczak EG, Ran L, Agrawal P, et al. Aberrant Activation of a Gastrointestinal Transcriptional Circuit in Prostate Cancer Mediates Castration Resistance. Cancer Cell. Elsevier; 2017;32:792-806.e7.

74. Sun K, Jiang P, Cheng SH, Cheng THT, Wong J, Wong VWS, et al. Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. Genome research. Cold Spring Harbor Laboratory Press; 2019;29:418–27.

75. Viswanathan SR, Ha G, Hoff AM, Wala JA, Carrot-Zhang J, Whelan CW, et al. Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. Cell. Elsevier; 2018;174:433-447.e19.

76. Choudhury AD, Werner L, Francini E, Wei XX, Ha G, Freeman SS, et al. Tumor fraction in cell-free DNA as a biomarker in prostate cancer. JCI Insight [Internet]. American Society for Clinical Investigation; 2018 [cited 2019 Mar 1];3. Available from: https://insight.jci.org/articles/view/122109

77. Klein DC, Hainer SJ. Genomic methods in profiling DNA accessibility and factor localization. Chromosome Res. 2020;28:69–85.

78. Chaytor L, Simcock M, Nakjang S, Heath R, Walker L, Robson C, et al. The Pioneering Role of GATA2 in Androgen Receptor Variant Regulation Is Controlled by Bromodomain and Extraterminal Proteins in Castrate-Resistant Prostate Cancer. Mol Cancer Res. American Association for Cancer Research; 2019;17:1264–78.

79. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362.

80. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications. 2017;8.

81. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun. Nature Publishing Group; 2021;12:1337.

82. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okur HS, Feroze AH, Emerson SN, et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat Biotechnol. 2021;39:819–24.

83. Lam H-M, Nguyen HM, Corey E. Generation of Prostate Cancer Patient-Derived Xenografts to Investigate Mechanisms of Novel Treatments and Treatment Resistance. In: Culig Z, editor. Prostate Cancer: Methods and Protocols [Internet]. New York, NY: Springer; 2018 [cited 2022 Mar 22]. page 1–27. Available from: https://doi.org/10.1007/978-1-4939-7845-8_1

84. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] [Internet]. 2013 [cited 2022 Mar 22]; Available from: http://arxiv.org/abs/1303.3997

85. Jo S-Y, Kim E, Kim S. Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis. Genome Biology. 2019;20:231.

86. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. Nature Publishing Group; 2011;43:491–8.

87. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14:7.

88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

89. Kluin RJC, Kemper K, Kuilman T, de Ruiter JR, Iyer V, Forment JV, et al. XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. BMC Bioinformatics. 2018;19:366.

90. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

91. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.

92. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 2019;35:421–32.

93. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research. 2016;44:W160–5.

94. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481:389–93.

95. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

96. Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31:2382–3.

97. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. Nature Publishing Group; 2019;9:9354.

98. Khan A, Mathelier A. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. BMC Bioinformatics. 2017;18:287.

99. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research. 2012;40:e72–e72.

100. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Research. 2018;46:e120.

101. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.

102. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Research. 2021;49:D884–91.

103. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. Nature Publishing Group; 2020;17:261–72.

104. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

Figure 1. Characterizing advanced prostate cancer through matched tumor and liquid biopsies from PDX models

(A) (**top**) Both blood and tissue samples were taken from 26 patient-derived xenograft (PDX) mouse models with tumors originating from metastatic castration-resistant prostate cancer (mCRPC) with AR-positive adenocarcinoma (ARPC), neuroendocrine (NEPC), AR-low neuroendocrine-negative (ARLPC) phenotypes. Cell-free DNA (cfDNA) was extracted from pooled plasma collected from 4-8 mice and whole genome sequencing (WGS) was performed. Following bioinformatic mouse read subtraction, pure human circulating tumor DNA (ctDNA) reads remained. From PDX tissue, ATAC-Seq and CUT&RUN (targeting H3K27ac, H3K4me1, and H3K27me3) data were generated. (**middle**) Four distinct ctDNA features were analyzed at five genomic region types using Griffin (51) or nucleosome phasing methods developed in this study (**Methods**). (**bottom**, **left**) Overview of PDX ctDNA features profiled to characterize the mCRPC pathways, transcriptional regulation, and nucleosome positioning. ctDNA features were evaluated for phenotype classification. (**bottom, right**) Phenotype classification using probabilistic and analytical models that accounted for ctDNA tumor content and were informed by PDX features were applied to 159 samples in three patient cohorts.

(B) PDX phenotypes and mouse plasma sequencing. Inclusion status based on final mean depth after mouse read subtraction (< 3x coverage were excluded; red dotted line). Phenotype status, including 6 NEPC, 18 ARPC (2 excluded), and 2 ARLPC. Average depth of coverage before and after mouse subtraction (mean coverage 20.5x; dotted line). Percentage of the cfDNA sample that contains human ctDNA after mouse read subtraction.

Figure 2. Analysis of tumor histone modifications and ctDNA reveals nucleosome patterns consistent with transcriptional regulation in CRPC phenotype-specific genes

(A) H3K27ac peak signals between ARLPC, ARPC, and NEPC PDX tumor phenotypes at 10,000 AR binding sites (left) and at ASCL1 binding sites (right). Binding sites were selected from the GTRD (70) (**Methods**).

(B-C) Composite coverage profiles at 1,000 AR (**B**) and ASCL1 (**C**) binding sites in ctDNA analyzed using Griffin for 140-250 bp fragments (Methods). Coverage profile means (lines) and 95% confidence interval computed using 1000 bootstraps for a subset of sites (shading) are shown. The region ±150 bp is indicated with vertical dotted line and yellow shading.

(D) Heatmap of log$_2$ fold change in 47 key genes up and down regulated between ARPC and NEPC established through RNA-Seq (**left**) grouped by the type of histone modification which dictates translation levels: Group 1 shows genes activity attributable to H3K27ac or H3K4me1 PTM marks in the gene promoters or putative distal enhancers and lacking H3K27me3 heterochromatic mark in the gene body; Group 2 features gene body spanning H3K27me3 repression marks. Central columns show differential peak intensity for each of the assayed histone modifications, separated by whether they appear upstream or in the promoter or the body of each gene. On the right the log$_2$ fold change between ARPC and NEPC lines' cfDNA fragment size coefficient of variation (CV) is shown for TSS+/- 1KB windows and respective gene bodies.

(E) Comparison of the log$_2$ fold change (ARPC/NEPC) of mean mRNA expression vs mean coefficient of variation (CV) in the 47 phenotypic lineage marker genes' promoter regions.

(F) (**top**) Illustrations of expected ctDNA coverage profiles for Group 1 genes with and without H3K27ac or H3K4me1 modification leading to active and inactive transcription, respectively. (**bottom**) ±1000 bp surrounding the promoter region for AR and ASCL1 in ARPC and NEPC. Shown are coverage profile means (lines) and 95% confidence interval computed using 1000 bootstraps for a subset of sites (shading). Decreased coverage is reflective of increased nucleosome accessibility and thus increased transcription. Dotted line and yellow shading highlight the transcription start site (TSS) at -230 bp and +170 bp.

(G) Illustration of expected ctDNA coverage profiles for Group 2 genes with repressed transcription caused by H3K27me3 modifications in the gene body. Neuronal gene UNC13A has increased nucleosome phasing in ctDNA of ARPC samples compared to NEPC.

Figure 3. Phasing analysis in ctDNA recapitulates nucleosome stability and trends in transcriptional activity between CRPC phenotypes

(A) Illustration of nucleosome phasing analysis using TritonNP for HOXB13, which is expressed in ARPC but not NEPC. Fourier transform and a band-pass filter-based smoothing method was used to identify phased peaks (grey dotted lines). Frequency components corresponding to nucleosome dyads (wavelength > 146 bp) are shown in purple. The mean inter-nucleosome distance was computed from all peaks in the gene body: lower values represent more periodic and stable nucleosomes. Nucleosome Phasing Score (NPS) is defined as the ratio of the mean amplitudes between frequency components 180-210 bp ("stable", green curve) and 150-180 bp ("baseline", red curve).

(B) Boxplot of mean phased-nucleosome distance in 17,946 gene bodies per ctDNA sample for ARPC and NEPC PDX lines. Two-tailed Mann-Whitney U test p-value shown.

(C) Comparison of the mean phased-nucleosome distance and the mean cell-cycle progression (CCP) score (estimated from RNA-Seq) for 16 ARPC and 6 NEPC PDX lines.

(D) Boxplot of NPS in gene bodies of 47 phenotype-defining genes (35 NE-regulated and 12 AR-regulated) between ARPC and NEPC lines. Two-tailed Mann-Whitney U test p-values shown.

(E) Volcano plot of NPS $\log_2$-fold-change (ARPC/NEPC) in the 47 phenotype-defining genes. Genes with significantly higher NPS scores (solid-colored dots (two-tailed Mann-Whitney U test, Benjamini-Hochberg adjusted FDR at $p < 0.05$) and non-significant genes (open circle) are shown.

(F) Hierarchical clustering of the normalized composite central mean coverage at TFBSs from the Griffin analysis of ctDNA for 108 TFs in LuCaP PDX lines of ARPC (n=16), NEPC (n=6), and ARLPC (n=2) phenotypes. This list of TFs was initially selected as having differential expression between ARPC and NEPC from LuCaP PDX RNA-Seq analysis. Heatmap colors indicate increased accessibility (low values; yellow, orange, red) and decreased accessibility (higher values; black) in ctDNA. TFs with increased accessibility in NEPC samples ($\log_2$-fold-change > 0.05, Mann-Whitney U test $p < 0.05$) are indicated with red bars; increased accessibility in ARPC (log2-fold-change < -0.05, $p < 0.05$) are indicated with blue bars. Text color indicates relative expression between ARPC and NEPC PDX tumors by RNAseq shown for TFs with significant differential accessibility.

Figure 4. Comprehensive evaluation of ctDNA features throughout the genome for CRPC phenotype classification in PDX models

(A) Volcano plot of $\log_2$-fold change of ATAC-Seq peak intensity between 5 ARPC and 5 NEPC lines; the dotted line demarcates sites by q-value < 0.05.

(B-C) Composite coverage profiles at open chromatin sites specific to ARPC (**B**) and NEPC (**C**) PDX tumors analyzed by Griffin. Sites from (A) were filtered for overlap with known TFBSs in 338 factors from GTRD (70). Coverage profile means (lines) and 95% confidence interval with 1000 bootstraps (shading) are shown. The region ±150 bp is indicated with vertical dotted line and yellow shading.

(D) PCAs of ctDNA features demonstrates grouping between ARPC and NEPC phenotypes: (**left**) Composite central coverage of TFBSs significant for 74 TFs with differential accessibility out of 338 factors between ARPC and NEPC (**Supplementary Table S4**). (**center**) NPS in the gene bodies of the 47 phenotype defining genes. (**right**) Fragment size variability (coefficient of variation) at H3K4me1 histone modification sites (n=9,750).

(E) Performance of classifying ARPC vs NEPC PDX from ctDNA using supervised machine learning (XGBoost) in various region types (all genes, TFBSs, and open regions, **Methods**). Area under the receiver operating characteristic curve (AUC) with 95% confidence interval (100 repeats of stratified cross validation) is shown for performance of all feature types.

Figure 5. Accurate classification and estimation of prostate cancer in patient plasma samples

(A) Schematic illustration of the ctdPheno classification method. Griffin-derived features and ichorCNA tumor fraction estimates from patient plasma samples are combined in a probabilistic framework informed by PDX models to predict the presence of NEPC.

(B) Performance for classification on admixtures samples using ctdPheno. Five ctDNA admixtures were generated for each phenotype from PDX lines, each at various sequencing coverages and tumor fractions. In total, 125 admixtures were evaluated. The mean AUC across the 5 admixtures is shown for each configuration.

(C) Receiver operating characteristic (ROC) curve for 101 mCRPC patients (DFCI cohort I) with ultra-low-pass WGS (ULP-WGS) data. The optimal performance of 90.4% sensitivity (for predicting NEPC) and 97.5% specificity (for predicting ARPC) corresponding to a prediction score cutoff of 0.3314 is indicated with horizontal and vertical dotted lines, respectively.

(D) Prediction scores from ctdPheno for 47 ULP-WGS plasma samples with clinical phenotypes comprising 26 ARPC (blue), 5 NEPC (red), and 16 mixed or ambiguous phenotypes (purple, triangles), including double-negative prostate cancer (DNPC; grey). The 0.3314 score cutoff threshold (dotted line) was used for classifying NEPC and ARPC. Tumor fractions were estimated by ichorCNA from WGS data.

(E) Schematic illustration of the Keraon mixture estimation method. Griffin-derived features from PDX lines and healthy donors define a known feature space, which is transformed based on Griffin features and ichorCNA tumor fraction estimates for each patient plasma sample. Based on the patient's location in the transformed phenotype space, fractions of each phenotype are inferred directly.

(F) Illustration of mixture simulations. 5 ARPC and 5 NEPC PDX samples were combined in the ratios shown with a single healthy donor at the tumor fractions shown, for a total of 810 mixed phenotype samples at 25x for evaluating mixture proportions with Keraon.

(G) Boxplot of predicted total NEPC fraction in 810 simulated mixed-phenotype samples using Keraon, Pearson's r = 0.884. Median absolute error (MAE) was computed as the median absolute difference between estimated and expected NEPC fraction across all samples.

(H) Fractional phenotype estimates for 47 WGS plasma samples with clinical phenotypes comprising 26 ARPC (blue), 5 NEPC (red), and 16 mixed or ambiguous phenotypes (purple, triangles), including double-negative prostate cancer (DNPC; grey). The 2.8% NEPC fraction threshold indicates the predicted presence of NEPC (dotted line).

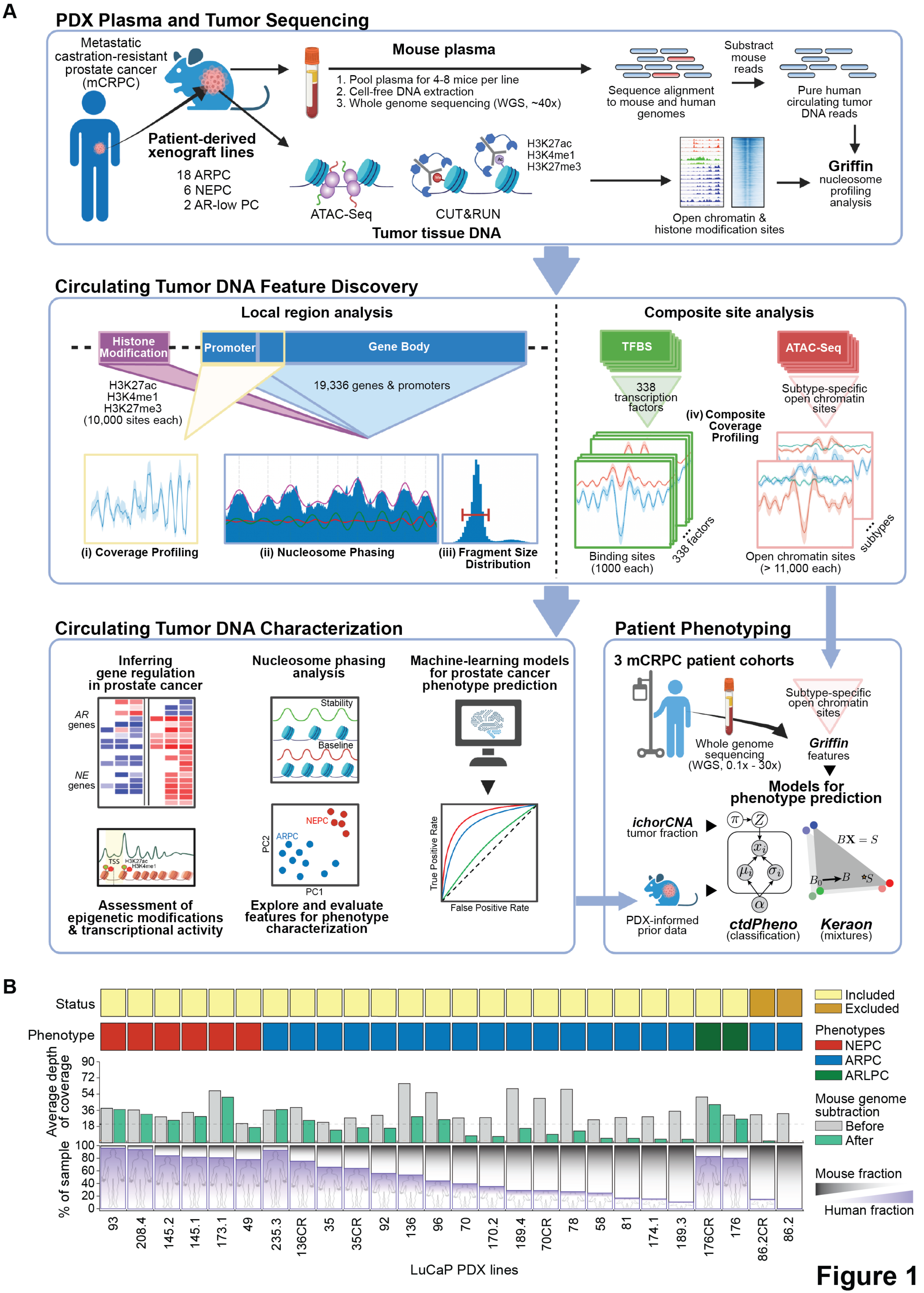
A
PDX Plasma and Tumor Sequencing
Metastatic castration-resistant prostate cancer (mCRPC)
Patient-derived xenograft lines
18 ARPC
6 NEPC
2 AR-low PC
Mouse plasma
1. Pool plasma for 4-8 mice per line
2. Cell-free DNA extraction
3. Whole genome sequencing (WGS, ~40x)
Sequence alignment to mouse and human genomes
Substract mouse reads
Pure human circulating tumor DNA reads
ATAC-Seq
CUT&RUN
H3K27ac
H3K4me1
H3K27me3
Tumor tissue DNA
Open chromatin & histone modification sites
Griffin nucleosome profiling analysis
Circulating Tumor DNA Feature Discovery
Local region analysis
Histone Modification
Promoter
Gene Body
H3K27ac
H3K4me1
H3K27me3
(10,000 sites each)
19,336 genes & promoters
(i) Coverage Profiling
(ii) Nucleosome Phasing
(iii) Fragment Size Distribution
Composite site analysis
TFBS
338 transcription factors
ATAC-Seq
Subtype-specific open chromatin sites
(iv) Composite Coverage Profiling
Binding sites (1000 each)
338 factors
Open chromatin sites (> 11,000 each)
subtypes
Circulating Tumor DNA Characterization
Inferring gene regulation in prostate cancer
AR genes
NE genes
TSS
H3K27ac
H3K4me1
Assessment of epigenetic modifications & transcriptional activity
Nucleosome phasing analysis
Stability
Baseline
NEPC
ARPC
PC2
PC1
Explore and evaluate features for phenotype characterization
Machine-learning models for prostate cancer phenotype prediction
True Positive Rate
False Positive Rate
Patient Phenotyping
3 mCRPC patient cohorts
Whole genome sequencing (WGS, 0.1x - 30x)
Subtype-specific open chromatin sites
Griffin features
Models for phenotype prediction
ichorCNA tumor fraction
PDX-informed prior data
π
Z
x_i
μ_i
σ_i
α
ctdPheno (classification)
$B\mathbf{X} = S$
B_0
B
S
Keraon (mixtures)
B
Status
Phenotype
Average depth of coverage
90
72
54
36
18
% of sample
100
80
60
40
20
0
93
208.4
145.2
145.1
173.1
49
235.3
136CR
35
35CR
92
136
96
70
170.2
189.4
70CR
78
58
81
174.1
189.3
176CR
176
86.2CR
86.2
LuCaP PDX lines
Included
Excluded
Phenotypes
NEPC
ARPC
ARLPC
Mouse genome subtraction
Before
After
Mouse fraction
Human fraction

Figure 1

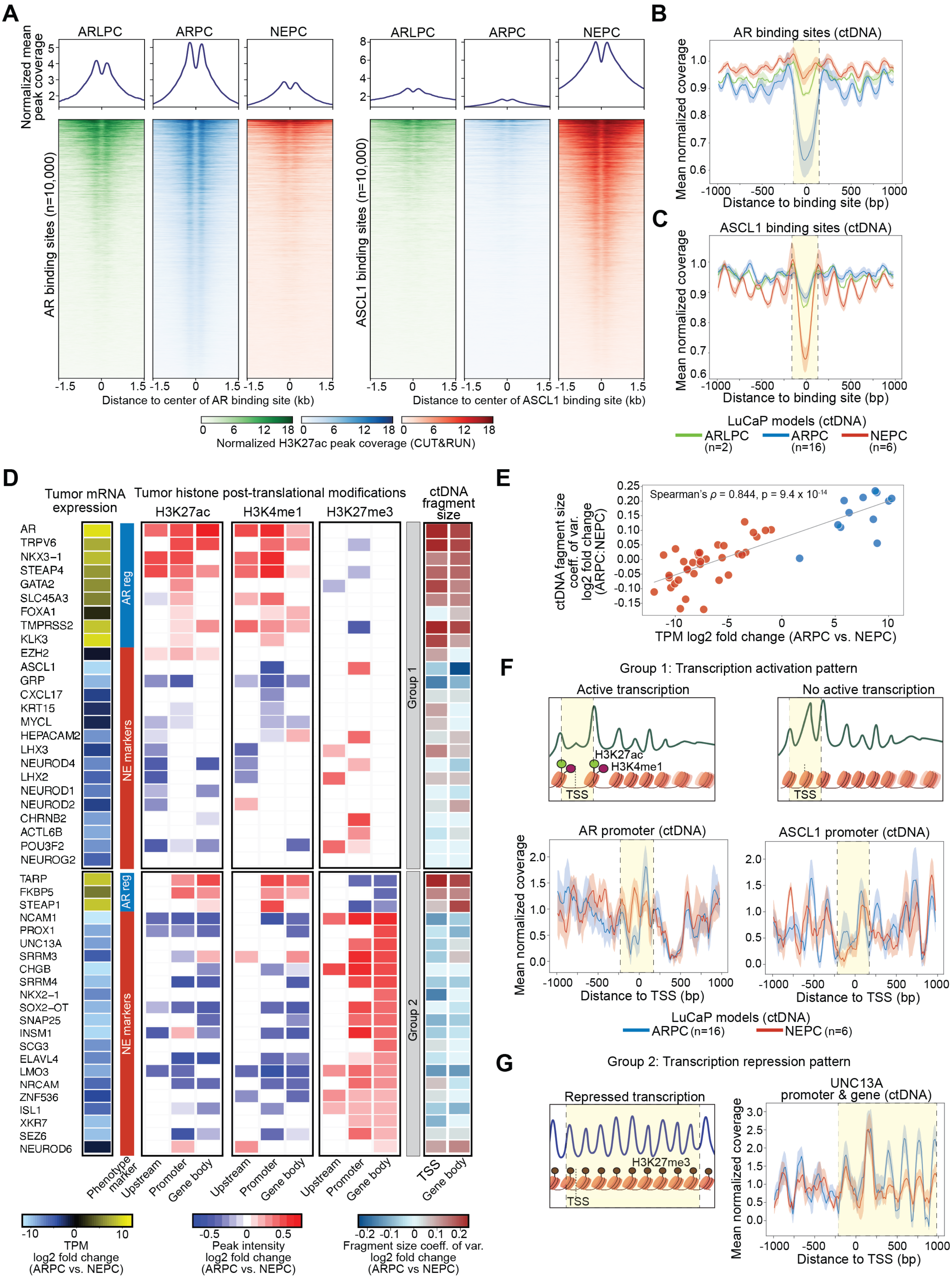
A
ARLPC
ARPC
NEPC
ARLPC
ARPC
NEPC
Normalized mean peak coverage
5
4
3
2
8
6
4
2
AR binding sites (n=10,000)
ASCL1 binding sites (n=10,000)
-1.5
0
1.5
Distance to center of AR binding site (kb)
Distance to center of ASCL1 binding site (kb)
0
6
12
18
Normalized H3K27ac peak coverage (CUT&RUN)
B
AR binding sites (ctDNA)
Mean normalized coverage
1.0
0.9
0.8
0.7
0.6
-1000
-500
0
500
1000
Distance to binding site (bp)
C
ASCL1 binding sites (ctDNA)
Mean normalized coverage
Distance to binding site (bp)
LuCaP models (ctDNA)
ARLPC (n=2)
ARPC (n=16)
NEPC (n=6)
D
Tumor mRNA expression
Tumor histone post-translational modifications
ctDNA fragment size
H3K27ac
H3K4me1
H3K27me3
AR
TRPV6
NKX3−1
STEAP4
GATA2
SLC45A3
FOXA1
TMPRSS2
KLK3
EZH2
ASCL1
GRP
CXCL17
KRT15
MYCL
HEPACAM2
LHX3
NEUROD4
LHX2
NEUROD1
NEUROD2
CHRNB2
ACTL6B
POU3F2
NEUROG2
TARP
FKBP5
STEAP1
NCAM1
PROX1
UNC13A
SRRM3
CHGB
SRRM4
NKX2−1
SOX2−OT
SNAP25
INSM1
SCG3
ELAVL4
LMO3
NRCAM
ZNF536
ISL1
XKR7
SEZ6
NEUROD6
AR reg
NE markers
Group 1
Group 2
Phenotype marker
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
Upstream
Promoter
Gene body
TSS
Gene body
-10
0
10
TPM log2 fold change (ARPC vs. NEPC)
-0.5
0
0.5
Peak intensity log2 fold change (ARPC vs. NEPC)
-0.2
-0.1
0
0.1
0.2
Fragment size coeff. of var. log2 fold change (ARPC vs NEPC)
E
Spearman's ρ = 0.844, p = 9.4 x 10-14
ctDNA fagment size coeff. of var. log2 fold change (ARPC:NEPC)
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-10
-5
0
5
10
TPM log2 fold change (ARPC vs. NEPC)
F
Group 1: Transcription activation pattern
Active transcription
No active transcription
H3K27ac
H3K4me1
TSS
AR promoter (ctDNA)
ASCL1 promoter (ctDNA)
Mean normalized coverage
2.0
1.5
1.0
0.5
0.0
Distance to TSS (bp)
LuCaP models (ctDNA)
ARPC (n=16)
NEPC (n=6)
G
Group 2: Transcription repression pattern
Repressed transcription
H3K27me3
TSS
UNC13A promoter & gene (ctDNA)
Mean normalized coverage
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Distance to TSS (bp)

Figure 2

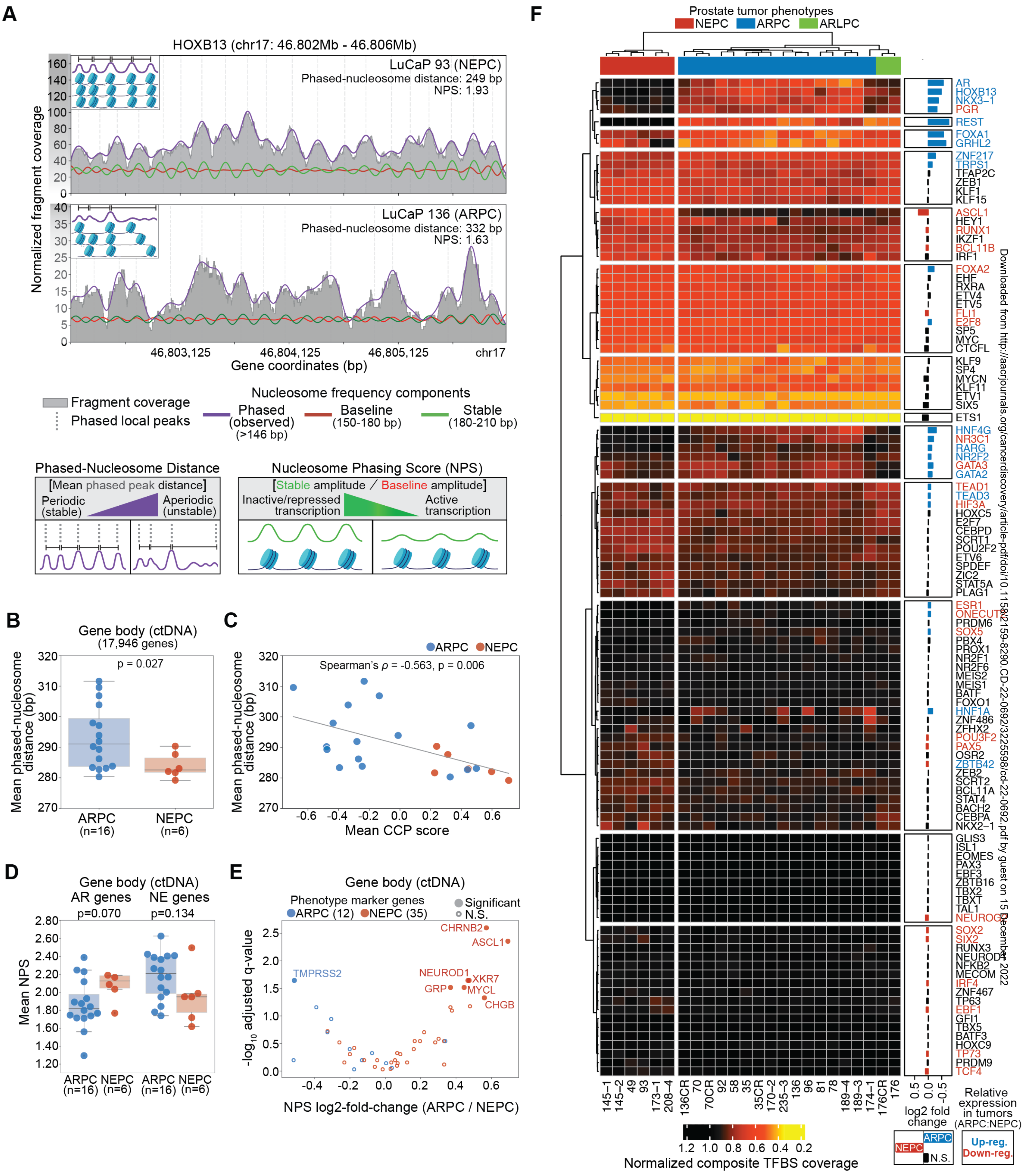
A
HOXB13 (chr17: 46.802Mb - 46.806Mb)
LuCaP 93 (NEPC)
Phased-nucleosome distance: 249 bp
NPS: 1.93
LuCaP 136 (ARPC)
Phased-nucleosome distance: 332 bp
NPS: 1.63
Normalized fragment coverage
46,803,125
46,804,125
46,805,125
chr17
Gene coordinates (bp)
Fragment coverage
Phased local peaks
Nucleosome frequency components
Phased (observed) (>146 bp)
Baseline (150-180 bp)
Stable (180-210 bp)
Phased-Nucleosome Distance
[Mean phased peak distance]
Periodic (stable)
Aperiodic (unstable)
Nucleosome Phasing Score (NPS)
[Stable amplitude / Baseline amplitude]
Inactive/repressed transcription
Active transcription
B
Gene body (ctDNA)
(17,946 genes)
p = 0.027
Mean phased-nucleosome distance (bp)
ARPC (n=16)
NEPC (n=6)
C
ARPC
NEPC
Spearman's ρ = -0.563, p = 0.006
Mean phased-nucleosome distance (bp)
Mean CCP score
D
Gene body (ctDNA)
AR genes p=0.070
NE genes p=0.134
Mean NPS
ARPC (n=16)
NEPC (n=6)
ARPC (n=16)
NEPC (n=6)
E
Gene body (ctDNA)
Phenotype marker genes
ARPC (12)
NEPC (35)
Significant
N.S.
-log10 adjusted q-value
TMPRSS2
CHRNB2
ASCL1
NEUROD1
XKR7
GRP
MYCL
CHGB
NPS log2-fold-change (ARPC / NEPC)
F
Prostate tumor phenotypes
NEPC
ARPC
ARLPC
AR
HOXB13
NKX3−1
PGR
REST
FOXA1
GRHL2
ZNF217
TRPS1
TFAP2C
ZEB1
KLF1
KLF15
ASCL1
HEY1
RUNX1
IKZF1
BCL11B
IRF1
FOXA2
EHF
RXRA
ETV4
ETV5
FLI1
E2F8
SP5
MYC
CTCFL
KLF9
SP4
MYCN
KLF11
ETV1
SIX5
ETS1
HNF4G
NR3C1
RARG
NR2F2
GATA3
GATA2
TEAD1
TEAD3
HIF3A
HOXC5
E2F7
CEBPD
SCRT1
POU2F2
ETV6
SPDEF
ZIC2
STAT5A
PLAG1
ESR1
ONECUT2
PRDM6
SOX5
PBX4
PROX1
NR2F1
NR2F6
MEIS2
MEIS1
BATF
FOXO1
HNF1A
ZNF486
ZFHX2
POU3F2
PAX5
OSR2
ZBTB42
ZEB2
SCRT2
BCL11A
STAT4
BACH2
CEBPA
NKX2−1
GLIS3
ISL1
EOMES
PAX3
EBF3
ZBTB16
TBX2
TBXT
TAL1
NEUROG2
SOX2
SIX2
RUNX3
NEUROD1
NFKB2
MECOM
IRF4
ZNF467
TP63
EBF1
GFI1
TBX5
BATF3
HOXC9
TP73
PRDM9
TCF4
145−1
145−2
49
93
173−1
208−4
136CR
70
70CR
92
58
35
35CR
170−2
235−3
136
96
81
78
189−4
189−3
174−1
176CR
176
Normalized composite TFBS coverage
log2 fold change
NEPC
ARPC
N.S.
Relative expression in tumors (ARPC:NEPC)
Up-reg.
Down-reg.

Figure 3

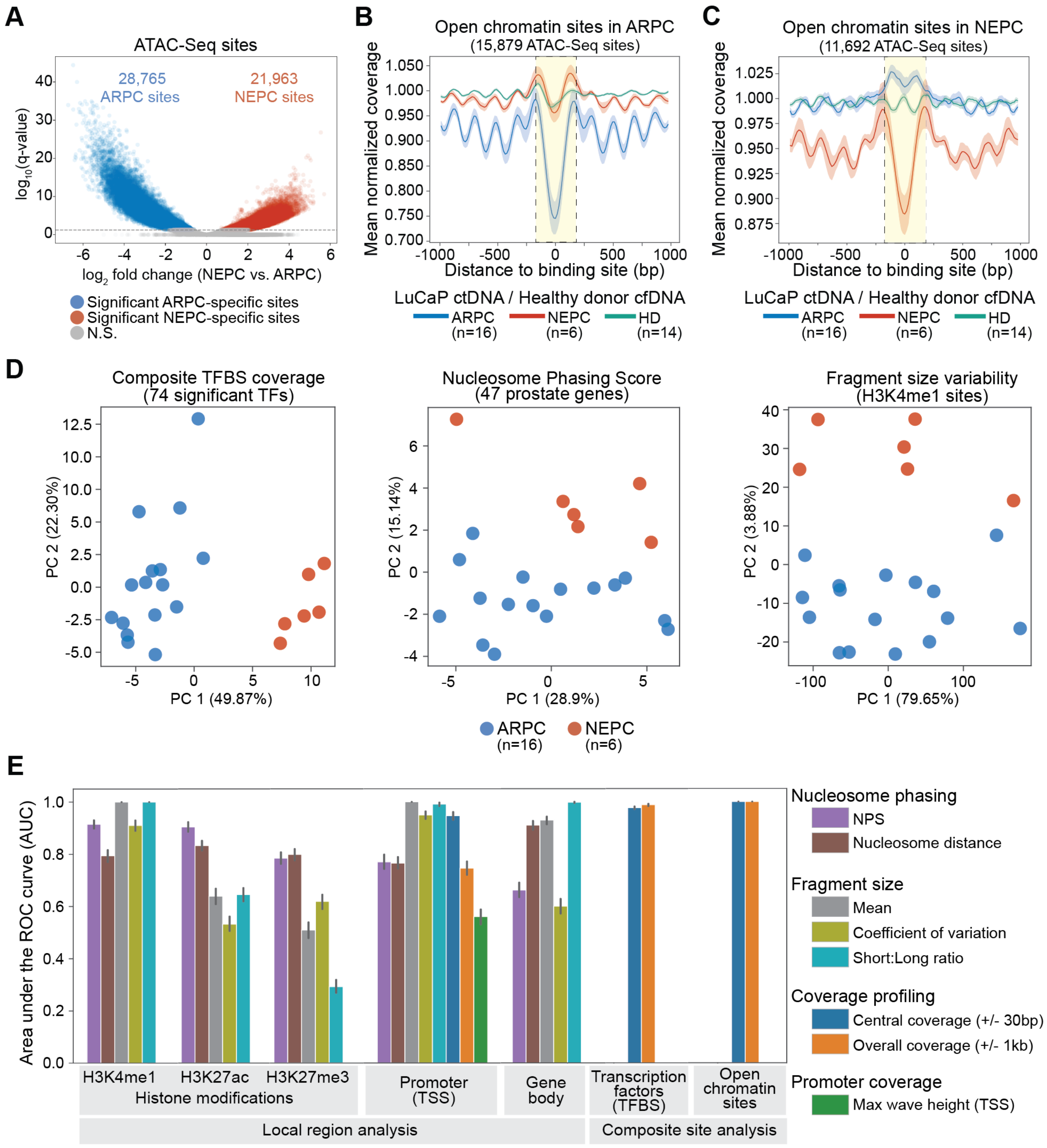
A
ATAC-Seq sites
28,765
ARPC sites
21,963
NEPC sites
$\log_{10}$(q-value)
$\log_2$ fold change (NEPC vs. ARPC)
Significant ARPC-specific sites
Significant NEPC-specific sites
N.S.
B
Open chromatin sites in ARPC
(15,879 ATAC-Seq sites)
Mean normalized coverage
Distance to binding site (bp)
LuCaP ctDNA / Healthy donor cfDNA
ARPC (n=16)
NEPC (n=6)
HD (n=14)
C
Open chromatin sites in NEPC
(11,692 ATAC-Seq sites)
Mean normalized coverage
Distance to binding site (bp)
LuCaP ctDNA / Healthy donor cfDNA
ARPC (n=16)
NEPC (n=6)
HD (n=14)
D
Composite TFBS coverage
(74 significant TFs)
PC 2 (22.30%)
PC 1 (49.87%)
Nucleosome Phasing Score
(47 prostate genes)
PC 2 (15.14%)
PC 1 (28.9%)
Fragment size variability
(H3K4me1 sites)
PC 2 (3.88%)
PC 1 (79.65%)
ARPC (n=16)
NEPC (n=6)
E
Area under the ROC curve (AUC)
H3K4me1
H3K27ac
H3K27me3
Histone modifications
Promoter (TSS)
Gene body
Transcription factors (TFBS)
Open chromatin sites
Local region analysis
Composite site analysis
Nucleosome phasing
NPS
Nucleosome distance
Fragment size
Mean
Coefficient of variation
Short:Long ratio
Coverage profiling
Central coverage (+/- 30bp)
Overall coverage (+/- 1kb)
Promoter coverage
Max wave height (TSS)

Figure 4

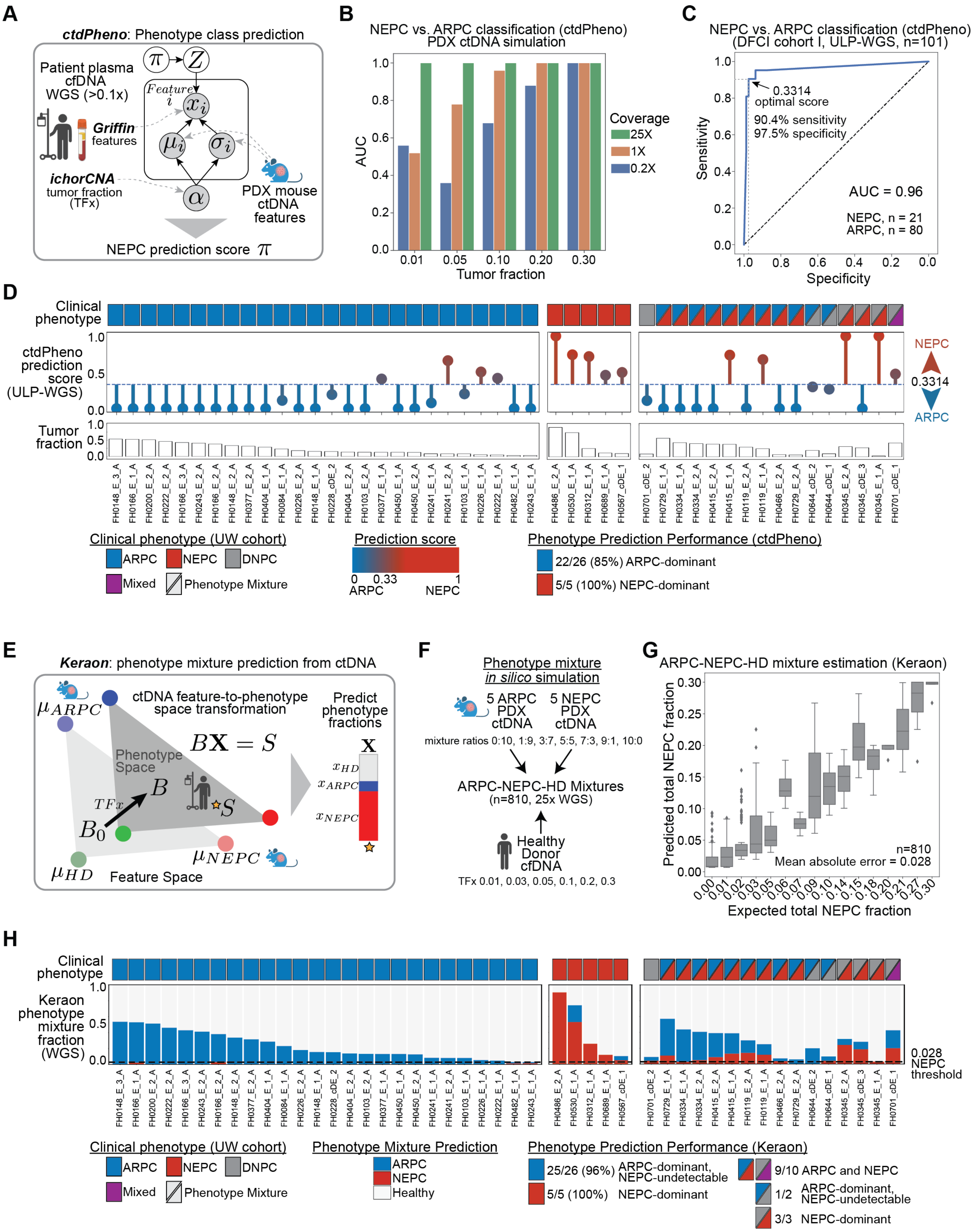
A
ctdPheno: Phenotype class prediction
Patient plasma cfDNA WGS (>0.1x)
Griffin features
ichorCNA tumor fraction (TFx)
PDX mouse ctDNA features
π
Z
Feature i
x_i
μ_i
σ_i
α
NEPC prediction score π
B
NEPC vs. ARPC classification (ctdPheno)
PDX ctDNA simulation
AUC
1.0
0.8
0.6
0.4
0.2
0.0
0.01
0.05
0.10
0.20
0.30
Tumor fraction
Coverage
25X
1X
0.2X
C
NEPC vs. ARPC classification (ctdPheno)
(DFCI cohort I, ULP-WGS, n=101)
Sensitivity
Specificity
0.3314 optimal score
90.4% sensitivity
97.5% specificity
AUC = 0.96
NEPC, n = 21
ARPC, n = 80
D
Clinical phenotype
ctdPheno prediction score (ULP-WGS)
Tumor fraction
NEPC
0.3314
ARPC
FH0148_E_3_A
FH0166_E_1_A
FH0200_E_2_A
FH0222_E_2_A
FH0166_E_3_A
FH0243_E_2_A
FH0166_E_2_A
FH0148_E_2_A
FH0377_E_2_A
FH0404_E_1_A
FH0084_E_1_A
FH0226_E_2_A
FH0148_E_1_A
FH0228_cDE_2
FH0404_E_2_A
FH0103_E_2_A
FH0377_E_1_A
FH0450_E_1_A
FH0450_E_2_A
FH0241_E_1_A
FH0241_E_2_A
FH0103_E_1_A
FH0226_E_1_A
FH0222_E_1_A
FH0482_E_1_A
FH0243_E_1_A
FH0486_E_2_A
FH0530_E_1_A
FH0312_E_1_A
FH0689_E_1_A
FH0567_cDE_1
FH0701_cDE_2
FH0729_E_1_A
FH0334_E_1_A
FH0334_E_2_A
FH0415_E_2_A
FH0415_E_1_A
FH0119_E_2_A
FH0119_E_1_A
FH0466_E_2_A
FH0729_E_2_A
FH0644_cDE_2
FH0644_cDE_1
FH0345_E_2_A
FH0345_cDE_3
FH0345_E_1_A
FH0701_cDE_1
Clinical phenotype (UW cohort)
ARPC
NEPC
DNPC
Mixed
Phenotype Mixture
Prediction score
0 0.33 1
ARPC NEPC
Phenotype Prediction Performance (ctdPheno)
22/26 (85%) ARPC-dominant
5/5 (100%) NEPC-dominant
E
Keraon: phenotype mixture prediction from ctDNA
μ_{ARPC}
ctDNA feature-to-phenotype space transformation
Predict phenotype fractions
$B\mathbf{X} = S$
Phenotype Space
B
TFx
B_0
S
μ_{NEPC}
μ_{HD}
Feature Space
$\mathbf{X}$
x_{HD}
x_{ARPC}
x_{NEPC}
F
Phenotype mixture in silico simulation
5 ARPC PDX ctDNA
5 NEPC PDX ctDNA
mixture ratios 0:10, 1:9, 3:7, 5:5, 7:3, 9:1, 10:0
ARPC-NEPC-HD Mixtures (n=810, 25x WGS)
Healthy Donor cfDNA
TFx 0.01, 0.03, 0.05, 0.1, 0.2, 0.3
G
ARPC-NEPC-HD mixture estimation (Keraon)
Predicted total NEPC fraction
0.30
0.25
0.20
0.15
0.10
0.05
0.00
n=810
Mean absolute error = 0.028
0.00 0.01 0.02 0.03 0.05 0.06 0.07 0.09 0.10 0.14 0.15 0.18 0.20 0.21 0.27 0.30
Expected total NEPC fraction
H
Clinical phenotype
Keraon phenotype mixture fraction (WGS)
1.0
0.5
0.0
0.028 NEPC threshold
Clinical phenotype (UW cohort)
ARPC
NEPC
DNPC
Mixed
Phenotype Mixture
Phenotype Mixture Prediction
ARPC
NEPC
Healthy
Phenotype Prediction Performance (Keraon)
25/26 (96%) ARPC-dominant, NEPC-undetectable
5/5 (100%) NEPC-dominant
9/10 ARPC and NEPC
1/2 ARPC-dominant, NEPC-undetectable
3/3 NEPC-dominant

Figure 5