

## Supplementary Information for

### **Divergent clonal evolution of castration resistant neuroendocrine prostate cancer**

Himisha Beltran<sup>1,2,3,#</sup>, Davide Prandi<sup>4,#</sup>, Juan Miguel Mosquera<sup>1,5</sup>, Matteo Benelli<sup>4</sup>, Loredana Puca<sup>1</sup>, Joanna Cyrta<sup>1</sup>, Clarisse Marotz<sup>1</sup>, Eugenia Giannopoulou<sup>6</sup>, Balabhadrapatruni V.S.K. Chakravarthi<sup>7</sup>, Sooryanarayana Varambally<sup>7</sup>, Scott A Tomlins<sup>8</sup>, David M Nanus<sup>2,3</sup>, Scott T Tagawa<sup>2,3</sup>, Eliezer M Van Allen<sup>9,10</sup>, Olivier Elemento<sup>1,6</sup>, Andrea Sboner<sup>1,11</sup>, Levi A Garraway<sup>9,10\*</sup>, Mark A Rubin<sup>1,3,5\*</sup>, Francesca Demichelis<sup>1,4,11\*</sup>

<sup>1</sup> Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine. New York, NY

<sup>2</sup> Department of Medicine, Division of Hematology and Medical Oncology, Weill Cornell Medicine. New York, NY

<sup>3</sup> Sandra and Edward Meyer Cancer Center at Weill Cornell Medicine. New York, NY

<sup>4</sup> Centre for Integrative Biology, University of Trento. Trento, Italy

<sup>5</sup> Department of Pathology and Laboratory Medicine. Weill Cornell Medicine. New York, NY

<sup>6</sup> Department of Physiology and Biophysics. Weill Cornell Medicine. New York, NY

<sup>7</sup> Department of Pathology, University of Alabama, Birmingham, AL

<sup>8</sup> Department of Pathology, University of Michigan. Ann Arbor, MI

<sup>9</sup> Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA

<sup>10</sup> The Broad Institute of MIT and Harvard, Boston, MA

<sup>11</sup> Institute for Computational Biomedicine, Weill Cornell Medicine. New York, NY

# co-first authorship; \* co-senior authorship

#### **This document includes:**

Supplementary Note

Figures Data List

Supplementary Figures 1 through 15

Supplementary Tables Legends

Supplementary References

## Supplementary Note

### CLONality Estimate in Tumors (CONET) 2.0

CLONET 2.0 extends on the framework presented in Prandi et al<sup>1</sup>. For each tumor/matched normal pair, the input includes a set *Seg* of genomic intervals and the read counts at germline heterozygous SNP loci (referred to as *informative* SNPs) from massively parallel sequencing data. Each genomic segment *S* in *Seg* is characterized by the Log2 ratio (*Log2R*) that represents the number of copies within *S* of the tumor genome with respect to the matched normal genome. In the context of sequencing, the *Log2R* value results from the ratio of mean coverage of *S* from the tumor DNA,  $cov_T$ , and from the normal DNA,  $cov_N$ . To account for possible different experimental mean coverages between the two experiments, the mean tumor ( $COV_T$ ) and mean normal ( $COV_N$ ) are considered to normalize the ratio. Formally, we have

$$Log2R = \log_2 \frac{cov_T/cov_N}{COV_T/COV_N} \quad 1$$

The *Log2R* distribution is indicative of the tumor genome state. In the presence of highly aberrant tumor genomes or of polyploidy, its correct interpretation is not straightforward. One example is reported in **Supplementary Fig. 14a**; the distribution peaks/modes might be interpreted from left to right as bi-allelic deletions (orange peak), mono-allelic deletions, copy number neutral segments, and amplifications (blue). **Supplementary Fig. 14b** shows the same signal upon ploidy correction. The signal shift is needed as the mean tumor  $COV_T$  does not necessary represent the mean coverage of a diploid genome; the adjustment is obtained as  $COV_T * (2/Pl)$ , where *Pl* is the tumor ploidy. By substituting the adjusted coverage in Equation 1, we obtain the corrected *Log2Rc* value

$$Log2Rc = \log_2 \frac{cov_T/cov_N}{COV_T * (2/Pl)/COV_N} = Log2R + \log_2 Pl/2 \quad 2$$

Corrections are implemented in CLONET based on the use of allelic fractions (AF) of the informative SNPs within a segment. The AF of a genomic position with coverage *cov* is defined as the ratio between the total number of reads supporting the reference base and *cov*. The empirical distribution of the AF of the informative SNPs in *S* results from the aggregate signal of the different cell populations in the tumor specimen. To de-convolve the underlying cell populations, we reason that the reads mapped with a segment *S* comprise reads that equally represent the parental chromosomes (copy number *neutral* reads) and reads from only one parental chromosome (*active* reads); we introduced *beta* as the proportion of neutral reads in *S*. Segments represented in the *beta* vs *Log2R* space facilitate the interpretation of complex copy number profiles. **Supplementary Fig. 14c** shows the copy number status of the tumor sample from **Supplementary Fig. 14b**; segments with copy number 4 (blue) form 3 well defined clusters with different combinations of the two alleles: dark blue segments correspond to two copies per

allele; mid blue segments to three copies on one allele and one copy of the other; light blue segments to four copies of one allele. Note that the orange peak, representative of mono-allelic deletions (**Supplementary Fig. 14b**), includes clonal mono-allelic deletion (orange cluster) and subclonal deletions (brown cluster). **CLONET 2.0, allele specific copy number**. The *beta* vs *Log2R* space fully represents allele specific copy number<sup>2</sup> and lesion clonality. However, intra-sample comparison might be counter-intuitive also due to different levels of tumor DNA purity. For the current study where multiple tumor samples are available for the same patient, this becomes highly relevant. In CLONET 2.0, we now introduced a space transformation from *beta* vs *Log2R* to *cnB* vs *cnA*, where *cnA* and *cnB* represent the copy number of allele A and B, respectively. **Supplementary Fig. 14d** shows the *cnA* vs *cnB* plot derived from **Supplementary Fig. 14c**. Subclonal segments, copy number neutral losses and copy number allelic imbalance events can be identified easily. To formalize the relation between *beta* vs *Log2R* and *cnB* vs *cnA*, we consider five variables, *beta*, *Log2R*, *cnA*, *cnB*, and *G*, where *G* is the global DNA admixture (1-purity). We use the superscripts *N* and *T* for normal and tumor cells within the same sample. The following four equations summarize the relations among these variables:

$$\left\{ \begin{array}{l} \text{Log2R} = \log_2 \frac{2G + (1 - G)(cnA^T + cnB^T)}{2} \\ \text{beta} = \frac{2 \min(cnA, cnB)}{2 \min(cnA, cnB) + |cnA - cnB|} \\ cnA = GcnA^N + (1 - G)cnA^T = G + (1 - G)cnA^T \\ cnB = GcnB^N + (1 - G)cnB^T = G + (1 - G)cnB^T \end{array} \right. \quad 3$$

Our purpose is to represent *beta* and *Log2R*, as a function of *cnA<sup>T</sup>* and *cnB<sup>T</sup>*. To solve Equation 3, we consider that:

- i. *G* is constant for a given tumor sample and is computed from raw *beta* and *Log2R* values of the sample selected segments<sup>1</sup>;
- ii. allele A and B represent the parental alleles. In the absence of parental genomes, we always consider A the allele with greater copy state with respect to allele B, that is  $cnA \geq cnB$ . Therefore  $\min(cnA, cnB) = cnB$  and  $|cnA - cnB| = cnA - cnB$ .

We can then conclude that:

$$\left\{ \begin{array}{l} cnA^T = \frac{(2 - \text{beta})(\text{beta} 2^{\text{Log2R}} - G) + 2G(1 - \text{beta})}{(1 - G)\text{beta}} \\ cnB^T = \frac{\text{beta} 2^{\text{Log2R}} - G}{1 - G} \end{array} \right. \quad 4$$

## Proof of Equation 4

Given

$$\left\{ \begin{array}{l} \text{Log}2R = \log_2 \frac{2G + (1-G)(cnA^T + cnB^T)}{2} \\ \text{beta} = \frac{2 \min(cnA, cnB)}{2 \min(cnA, cnB) + |cnA - cnB|} \\ cnA = GcnA^N + (1-G)cnA^T = G + (1-G)cnA^T \\ cnB = GcnB^N + (1-G)cnB^T = G + (1-G)cnB^T \end{array} \right.$$

As  $cnA \geq cnB$ , we obtain:

$$\left\{ \begin{array}{l} \text{Log}2R = \log_2 \frac{2G + (1-G)(cnA^T + cnB^T)}{2} \\ \text{beta} = \frac{cnB}{2 cnB + cnA - cnB} \\ cnA = G + (1-G)cnA^T \\ cnB = G + (1-G)cnB^T \end{array} \right.$$

As we are interested in determining  $cnA^T$  and  $cnB^T$ , we rewrite  $\text{beta}$  as

$$\left\{ \begin{array}{l} \text{Log}2R = \log_2 \frac{2G + (1-G) cnA^T + (1-G) cnB^T}{2} \\ \text{beta} = \frac{2G + 2(1-G)cnB^T}{2G + (1-G)cnB^T + (1-G)cnA^T} \end{array} \right.$$

By writing  $cnA^T$  as a function of  $\text{beta}$  and of  $cnB^T$

$$\begin{aligned} 2 G \text{beta} + (1-G) \text{beta} cnB^T + (1-G)\text{beta} cnA^T &= 2G + 2(1-G)cnB^T \\ (1-G) \text{beta} cnA^T &= 2 G + 2 (1-G) cnB^T - 2 G \text{beta} - (1-G) \text{beta} cnB^T \\ (1-G) \text{beta} cnA^T &= [2(1-G) - \text{beta} (1-G)] cnB^T + 2 G (1 - \text{beta}) \\ (1-G) \text{beta} cnA^T &= (1-G)(2 - \text{beta}) cnB^T + 2 G (1 - \text{beta}) \\ cnA^T &= \frac{(1-G)(2 - \text{beta}) cnB^T + 2 G (1 - \text{beta})}{(1-G) \text{beta}} \end{aligned}$$

By substituting the equation for  $cnA^T$  in the  $\text{Log}2R$ , we can express  $cnB^T$  as a function of  $\text{beta}$  and  $\text{Log}2R$

$$\begin{aligned} \text{Log}2R &= \log_2 \frac{2G + (1-G) \frac{(1-G)(2-\text{beta}) cnB^T + 2G(1-\text{beta})}{(1-G) \text{beta}} + (1-G) cnB^T}{2} \\ 2^{\text{Log}2R} &= \frac{2G \text{beta} + (1-G)(2 - \text{beta}) cnB^T + 2 G (1 - \text{beta}) + (1-G) \text{beta} cnB^T}{2} \\ 2 \text{beta} 2^{\text{Log}2R} &= 2G \text{beta} + (1-G)(2 - \text{beta}) cnB^T + 2 G (1 - \text{beta}) + (1-G) \text{beta} cnB^T \\ 2 \text{beta} 2^{\text{Log}2R} &= cnB^T [(1-G)(2 - \text{beta}) + (1-G) \text{beta}] + 2G \text{beta} + 2 G (1 - \text{beta}) \\ 2 \text{beta} 2^{\text{Log}2R} &= cnB^T [(1-G)(2 - \text{beta} + \text{beta})] + 2G(\text{beta} + 1 - \text{beta}) \\ 2 \text{beta} 2^{\text{Log}2R} &= cnB^T 2(1-G) + 2G \end{aligned}$$

$$cnB^T = \frac{\beta 2^{\log_2 R} - G}{(1 - G)}$$

Finally, we substitute  $cnB^T$  in  $cnA^T$ :

$$cnA^T = \frac{(1 - G)(2 - \beta) \frac{\beta 2^{\log_2 R} - G}{(1 - G)} + 2 G (1 - \beta)}{(1 - G) \beta}$$

$$cnA^T = \frac{(2 - \beta)(\beta 2^{\log_2 R} - G) + 2 G (1 - \beta)}{(1 - G) \beta}$$

**Differential expression analysis excluding bone metastases.** As CRPC-NE samples are primarily from soft tissue versus typical CRPC-Adeno being from bone, we checked whether gene expression from contaminating normal cells might affect transcriptome analysis. To this end, we re-run differential expression analysis between CRPC-NE and CRPC-Adeno excluding bone metastasis (n = 9). We found no significant difference with respect to the original analysis (**Supplementary Fig. 3**).

**Integrated NEPC scores from reduced gene sets.** Integrated NEPC score comprises 70 genes derived by differentially deleted/amplified, expressed and methylated genes analyses. **Supplementary Fig. 15** reports Ward's hierarchical clustering of normalized counts of 636 prostate samples and shows that selected genes correctly distinguish CRPC-NE samples. Heat map also shows that a number of genes are highly correlated. In order to understand if reducing the set of genes included in the Integrated NEPC score preserved the classification of samples, we tested the three following versions of Integrated NEPC score, each of which characterized by a different subset of genes.

ROC-selected genes. In this version, integrated NEPC score is defined by considering only genes selected by ROC curve analysis (see **Online Methods**), resulting in a total of 49 genes.

AR signaling genes excluded. This version comprises all genes excepting the 3 genes selected from AR signaling gene list (*AR*, *KLK3*, *NKX3-1*), resulting in a set of 67 genes.

After Feature Selection. In this version we applied a consistency-based Feature Selection strategy<sup>3</sup> to reduce the number of genes. By using the R-package FSelector version 0.20 (<http://cran.r-project.org/package=FSelector>), we obtained a list of 12 genes (*PROX1*, *ETV5*, *PCSK1*, *LRRC16B*, *GPX2*, *MYCN*, *SLC44A4*, *CATSPERB*, *EPN3*, *SLC25A37*, *LMAN1L*, *RGS10*). The performance of the original Integrated NEPC score (all 70 genes) and the three reduced gene sets variants are reported in **Supplementary Fig. 11**, that shows AR-signaling versus Integrated NEPC score across 5 different prostate cancer datasets (730 samples in total, see **Supplementary Table 10** for more details).

**Statistical evaluation of the CRPC-NE classifier.** We evaluated the CRPC-NE prediction accuracy of Integrated NEPC score, AR signaling, mRNA level of *AR* and *SPDEF* and a correlation-based classifier built by considering conventional neuroendocrine markers (i.e., *CGHA*, *SYP*, *NSE*, *CD56*; *KLK3* (*PSA*) was also included; we refer to this set as NE genes) by using the following statistical measure: Precision, Recall and p-values and Odd-Ratio (OR) from Fisher Exact Test. True and False events were defined as samples classified as CRPC-NE and not CRPC-NE (CRPC-Adeno, PCa, Benign), respectively. Positive and negative events were defined as follows.

- Samples with high Integrated NEPC score were defined as samples with Integrated NEPC score  $\geq 0.4$ .
- Samples with low AR signaling (as expected in CRPC-NE) were defined as samples with AR signaling  $\leq 0.25$ .
- Samples with low mRNA level of *AR* and *SPDEF* (as expected in CRPC-NE) were defined as samples with  $\log_2(\text{FPKM}+1) \leq 2$  for these genes.
- Samples with high NE genes score were defined as samples with NE genes score  $\geq 0.4$ .

The results of these statistical measures for different sets of data are reported in **Supplementary Table 14, Figure 4c and Supplementary Figure 12**. In each set of data, Integrated NEPC score shows the best performance for all the four statistical measures.

### Figures Data List

For each data figure in the manuscript, here we report where the data underlying the used graphical representations can be retrieved. Two main files are referred:

- Supplementary Tables
- Additional Figures data reported in Supplementary TableFig

### Figure 1

Panel **a**: For each study sample, **Supplementary TableFig 1a** lists sample site (column “Man annotation”) and pie chart class (columns “Pie chart annotation”).

Panel **b**: **Supplementary Table 15** includes AR Signaling value for each sample in the study.

Panel **c**: **Supplementary TableFig 1c** reports MutSig output for the study cohort. Genes included in **Fig. 1c** show q-value  $< 10\%$ . Sample based single nucleotide variants (SNVs) are reported in **Supplementary Table 3**. Ploidy and pathology tracks above SNVs map uses data from **Supplementary Table 11** (columns “Ploidy (CLONET)” and “Pathology Classification”, respectively). Finally, copy number data of genes *RB1*, *PTEN*, *CHD1*, *TP53*, and *BRCA2* is obtained intersecting segmented data adjusted by ploidy reported in **Supplementary Table 6** (column “Log2 Ratio ploidy adjusted”) with genes coordinates.

Panel **d**: SU2C sample data of AR SNVs is accessible from cBio portal. AR SNVs identified in this current study are reported in **Supplementary Table 3**.

Panel **e**: Ploidy adjusted Log<sub>2</sub> ratio (ratio=tumor/normal) of the genomic region spanning AR gene is reported in **Supplementary Table 6** (column “Log<sub>2</sub> Ratio ploidy adjusted”). Genomic regions where ploidy adjustment cannot be performed (value NA in **Supplementary Table 6**) have a Log<sub>2</sub> ratio corresponding to the mean of the adjacent 3’ and 5’ segments.

Panel **f**: The genomic profile is obtained by sampling the genome and counting the number of aberrant samples at each position across the study cohort. **Supplementary TableFig 1f** reports the chromosome and the genomic position (columns “chr” and “pos”), the number of CRPC-NE and CRPC-Adeno samples available at the considered genomic position (columns “n.sample\_CRPC-NE”, “n.sample\_CRPC-Adeno” columns), the number of deleted and amplified samples for each position and pathology class (columns “n.deleted\_CRPC-NE”, “n.deleted\_CRPC-Adeno”, “n.amplified\_CRPC-NE”, “n.amplified\_CRPC-Adeno”), and the fraction of deleted/amplified samples used in **Fig. 1f** (columns “perc.deleted\_CRPC-NE”, “perc.deleted\_CRPC-Adeno”, “perc.amplified\_CRPC-NE”, “perc.amplified\_CRPC-Adeno”). Log<sub>2</sub> ratio data adjusted by ploidy and purity used for genomic sampling is reported in **Supplementary Table 6** (column “Log<sub>2</sub> Ratio ploidy and purity adjusted”).

### **Figure 2**

Circos plot, phylogenetic trees and cnA vs cnB plots of **Fig. 2** uses the allele specific copy number data reported in **Supplementary TableFig 2**.

### **Figure 3**

Panel **a**: Dendrogram showing the hierarchical clustering on unselected methylation loci is obtained by the function “clusterSamples” of methylKit R package (<https://code.google.com/p/methylkit/>).

Panel **b**: Hyper- and Hypo methylated sites and corresponding annotation by BedTools (<https://github.com/arq5x/bedtools2>) on gene promoters, first exons and gene bodies are reported in **Supplementary Table 8**. Functional enrichment analysis was performed by ToppFun (<http://toppgene.cchmc.org/enrichment.jsp>) and results are reported in **Supplementary TableFig 3b**.

Panel **c**: Box plots refers to data in **Supplementary TableFig 3c** and report the SPDEF log<sub>2</sub>(FPKM+1) in the data used in “this Study” (left) and in SU2C 2015 (right).

Panel **d**: Bar plots report the log<sub>2</sub>(FPKM+1) medians (divided by the median of the log<sub>2</sub>(FPKM+1) obtained in benign samples) of under-expressed HOX genes in CRPC-NE vs. CRPC-Adeno, a selection of EZH2 target genes, DNA methyltransferases genes and EZH2 in this Cohort (this Study + WCMC 2011/14 + TCGA) and SU2C 2015 cohort. Data is reported in **Supplementary TableFig 3d**.

Panel **e**: **Supplementary TableFig 3e** reports cell viability in prostate adenocarcinoma cell lines (DU145, LNCaP) the neuroendocrine prostate cell line NCI-H660 assessed at 48 hours after treatment with escalating doses of the EZH2 inhibitor GSK343 (5, 7.5, 10uM).

#### **Figure 4**

Panel **a**: Weighted Venn diagram and pie chart are obtained by counting the number of deleted/amplified (DNA), differentially expressed (RNA), differentially methylated (Methylation) genes. Genes altered in multiple layers are categorized with the following rule: methylation overall and DNA over RNA. Complete data is reported in **Supplementary TableFig 4a**.

Panel **b**: Heat maps and annotation tracks show Integrated NEPC score analysis across 604 samples from four different RNA-Seq prostate cancer datasets (This Study, SU2C 2015, WCMC 2011/2014 and TCGA). Raw expression data is reported in **Supplementary TableFig 4b**.

Panel **c**: Precision-Recall plot for the different classifiers (different point symbols) and across the 4 RNA-seq datasets (different colors). Data to build the figure is reported in **Supplementary Table 14**.

Panel **d**: AR Signaling versus Integrated NEPC score across 5 prostate cancer datasets (This Study, SU2C 2015, WCMC 2011/2014, TCGA, Michigan 2012). Colors reflect distinct pathology classes. Data to build the figure is reported in **Supplementary Table 15**.



## **List of Supplementary Figures**

Supplementary Figure 1: Histology of Prostate Cancer categories used in the study.

Supplementary Figure 2: Comparison of molecular features between CRPC-Adeno and CRPC-NE.

Supplementary Figure 3: Comparative analysis for sites of metastasis.

Supplementary Figure 4: Characteristics of the six tumor samples with the highest number of non-synonymous SNVs.

Supplementary Figure 5: CYLD loss in CRPC-NE.

Supplementary Figure 6: Tumor evolution and clonality analysis for multi-sample cases.

Supplementary Figure 7: Unsupervised clustering of differentially expressed genes that are also differentially methylated in CRPC-NE versus CRPC-Adeno.

Supplementary Figure 8: Overview of differential methylation analysis of CRPC-NE versus CRPC-Adeno.

Supplementary Figure 9: Enzalutamide resistant LNCaP cells and SPDEF expression and methylation across cell lines.

Supplementary Figure 10: Integrated NEPC score analysis across 636 samples from four different prostate cancer datasets.

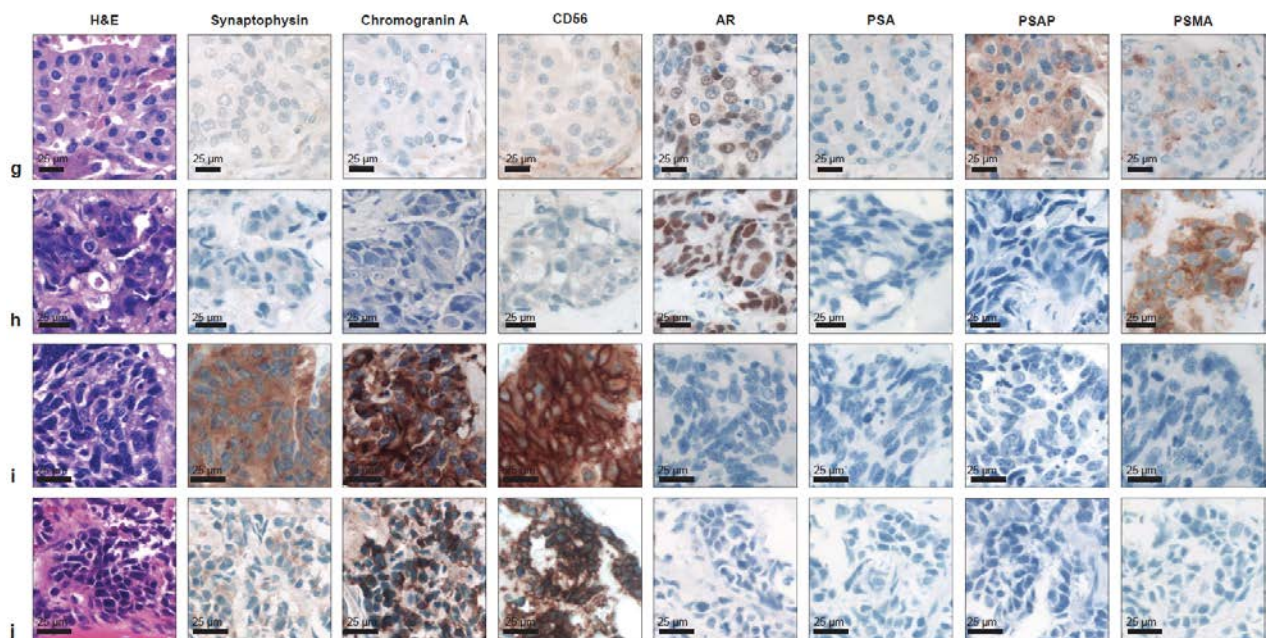
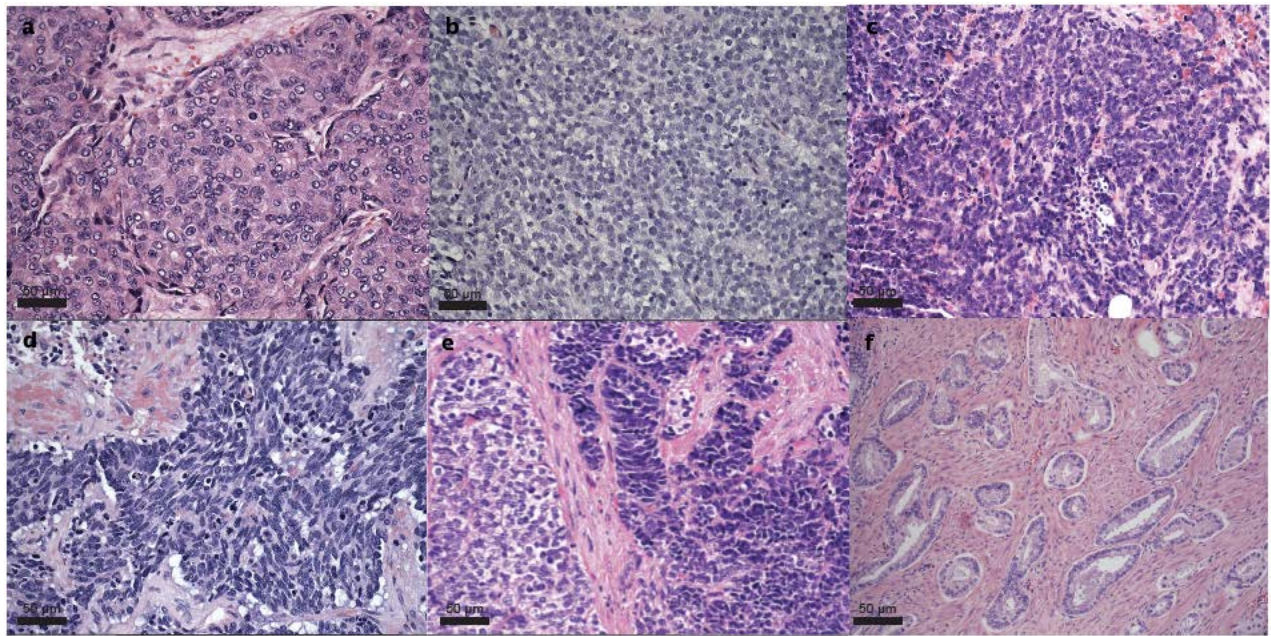
Supplementary Figure 11: AR signaling versus Integrated NEPC Score.

Supplementary Figure 12: Rapid Autopsy Cases from the University of Michigan Cohort with elevated NEPC score.

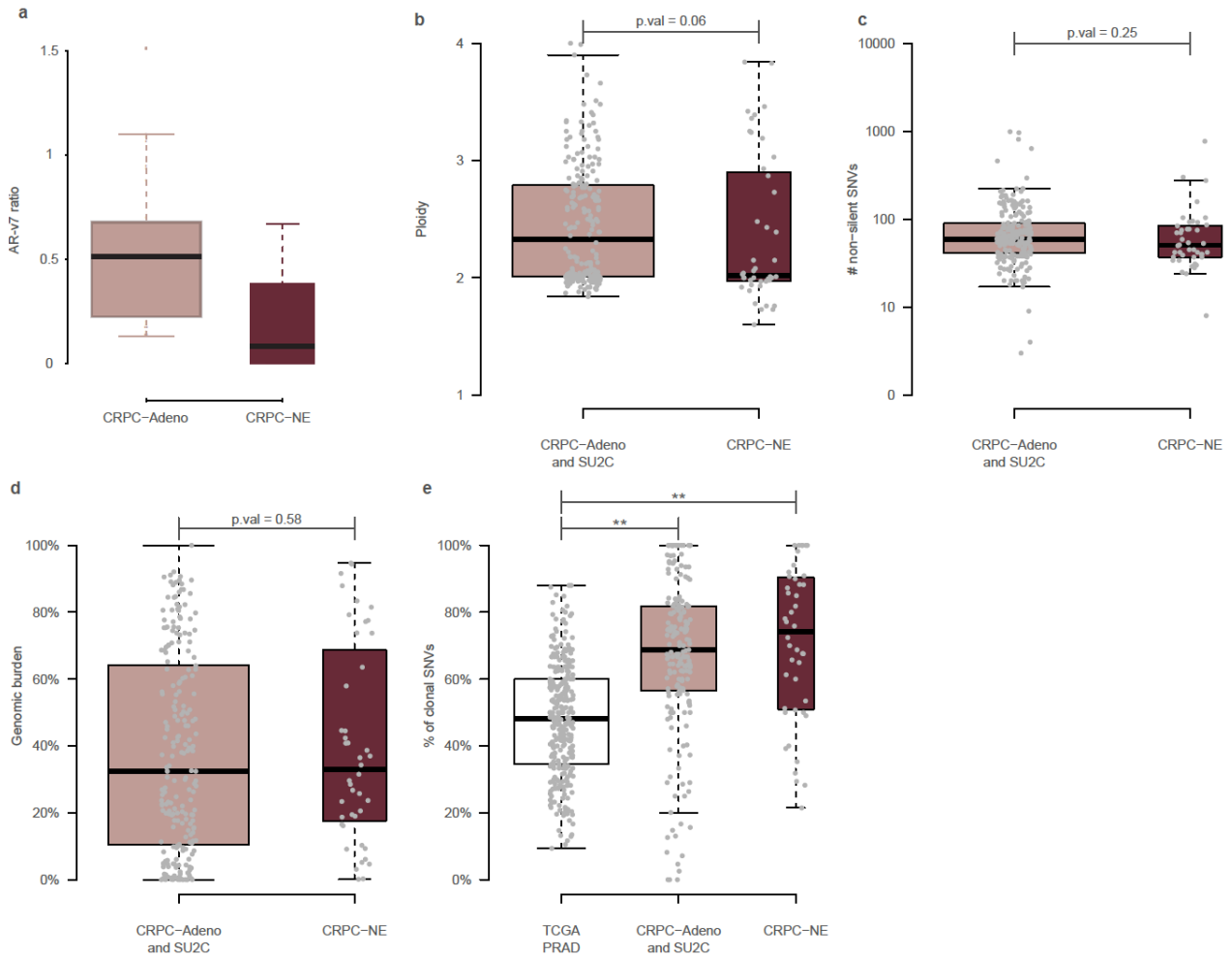
Supplementary Figure 13: Enzalutamide resistant cell line transitions toward CRPC-NE - like phenotype.

Supplementary Figure 14: Schematic of CLONET 2.0 new features.

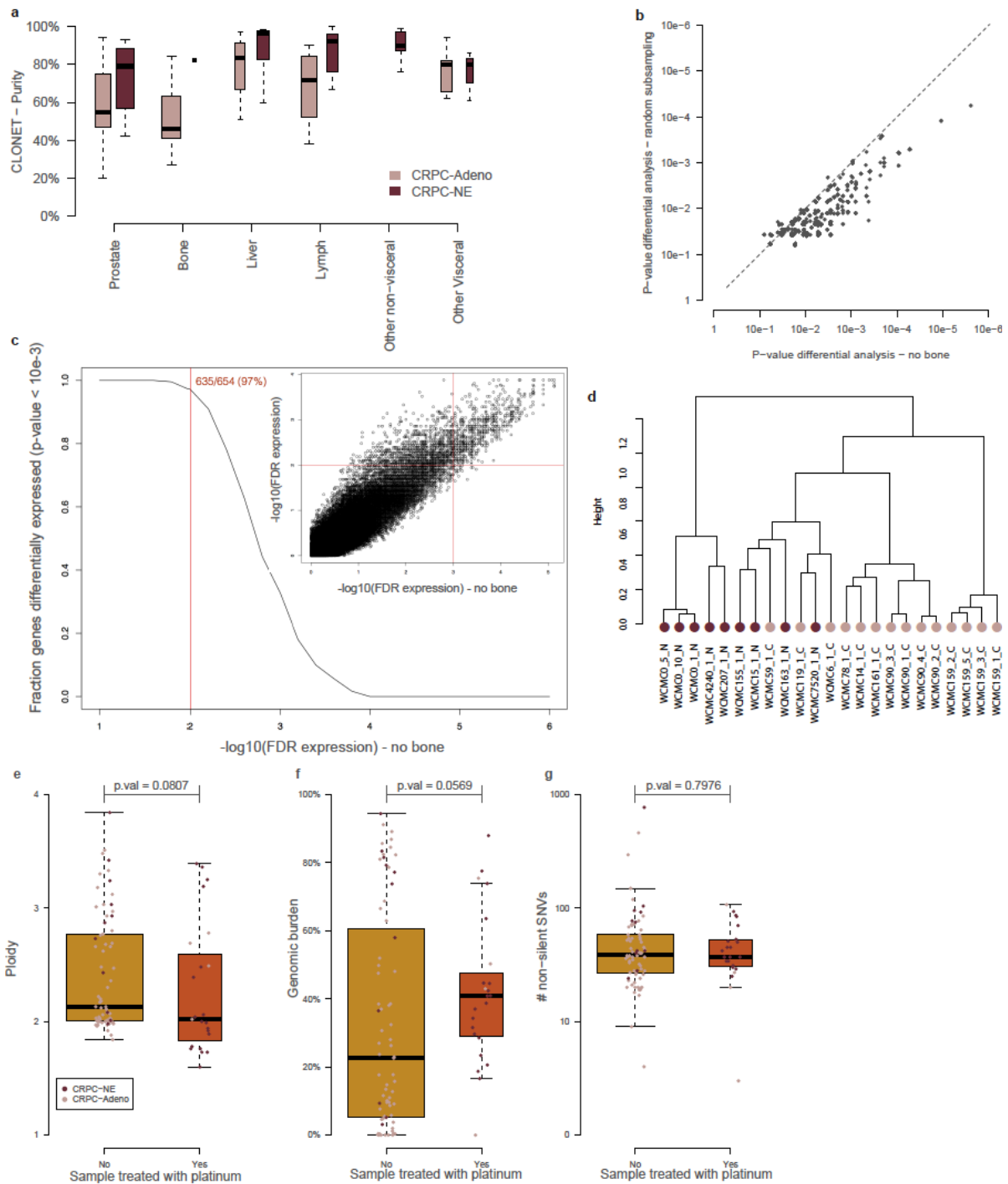
Supplementary Figure 15: AR signaling score and Integrated NEPC score genes.



**Supplementary Figure 1: Histology of Prostate Cancer categories used in the study.** Representative H&E images of (a) Usual prostate adenocarcinoma *without* neuroendocrine differentiation, (b) Usual prostate adenocarcinoma *with* neuroendocrine differentiation (as demonstrated by immunohistochemistry, not shown), (c) Small cell carcinoma, (d) Large cell neuroendocrine carcinoma, (e) Mixed small cell carcinoma (right) - adenocarcinoma (left), and (f) Conventional primary prostate adenocarcinoma (Gleason grade 6). Histologic evaluation (a-e) followed the criteria of the Proposed 2015 WHO Morphologic Classification of Prostate Cancer with Neuroendocrine Differentiation<sup>4</sup>. Representative Hematoxylin & Eosin (H&E) images and immunohistochemical markers in CRPC-Adeno and CRPC-NE liver and bone metastases: (g)= liver metastasis from patient with CRPC-Adeno (WCMC90); (h)= bone metastasis from patient with CRPC-Adeno (WCMC131); (i)= liver metastasis from patient with CRPC-NE (WCMC0); (j)= bone metastasis from patient with CRPC-NE (WCMC154). AR= androgen receptor; PSA= prostate specific antigen; PSAP = prostate acid phosphatase; PSMA= prostate specific membrane antigen.



**Supplementary Figure 2: Comparison of molecular features between CRPC-Adeno and CRPC-NE.** (a) ARv7 versus wild type AR ratio differences between CRPC-Adeno and CRPC-NE. The ARv7 splice variant was observed in both CRPC-Adeno and CRPC-NE tumors, and as overall AR expression was lower in CRPC-NE, ARv7:wild type AR ratio was significantly decreased in CRPC-NE compared to CRPC-Adeno ( $p=0.0025$ , Wilcoxon test). Comparison of the genomic characteristics of the CRPC-Adeno and CRPC-NE for ploidy (b), the number of non-silent SNVs (c) and the genomic burden (d) of CRPC-NE and CRPC-Adeno. To gain statistical power, CRPC-Adeno samples from our study cohort were integrated with samples from SU2C<sup>5</sup>. The boxplots highlight the similarity between the two classes of advance prostate cancer when considering basic genomic features. (e) The plot reports percentage of clonal SNVs in our study plus SU2C cohort compared with 483 untreated localized prostate cancers. The localized samples show a significant low percentage of clonal SNVs with respect to advanced tumors ( $P < 10^{-7}$ ). All distributions are compared by means of Wilcoxon test.

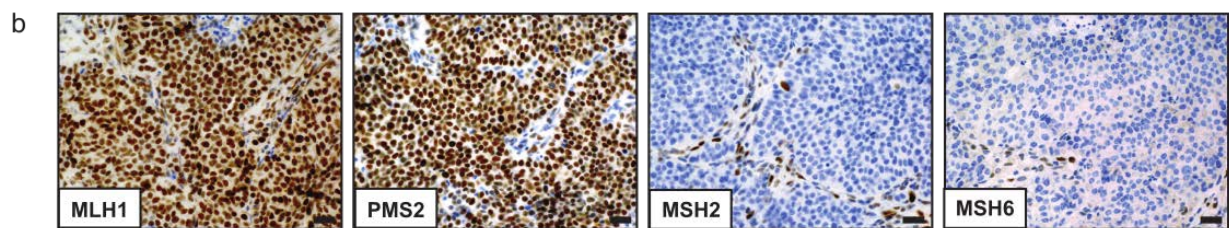


**Supplementary Figure 3: Comparative analysis for sites of metastasis.** Impact on tumor purity and molecular features (a) tumor purity between bone samples and other metastatic biopsy sites within our cohort showing no significant differences; (b) copy number differential analysis showed no significant differences when performed excluding the bone metastases; (c) transcriptome analysis showed no significant differences when performed excluding the bone metastases including in the correlation between original (y-axis) and no-bone-met (x-axis) analysis (inset); (d) methylation analysis showed no significant differences when performed excluding the bone metastases and the unsupervised hierarchical clustering of methylation sites recapitulates the segregation of CRPC-NE samples. Comparison of the genomic characteristics of pre and post platinum samples with respect to the distribution of the ploidy (e) the number of non-silent SNVs (f) and the genomic burden (g) of sample treated with platinum against and untreated cases. The two groups of samples show comparable distribution when evaluating genomic features. All distributions are compared by means of Wilcoxon test.

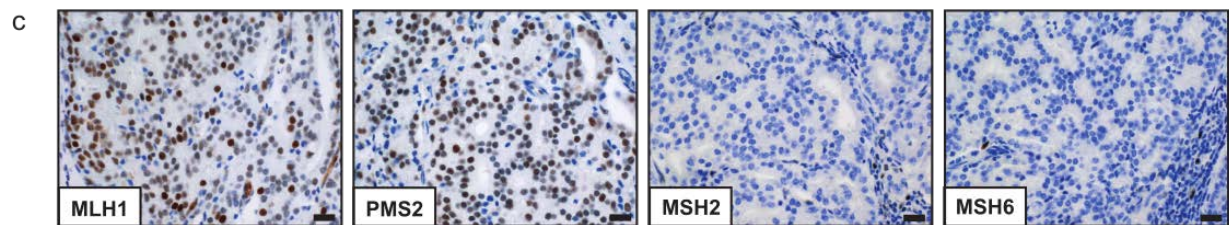
a

| Sample ID     | Pathology | Non-synonymous SNV number | CNA burden | Alterations in MMR genes |                             | Protein expression by IHC |      |        |        |
|---------------|-----------|---------------------------|------------|--------------------------|-----------------------------|---------------------------|------|--------|--------|
|               |           |                           |            | Gene name                | Alteration(s)               | MLH1                      | PMS2 | MSH2   | MSH6   |
| WCMC10362_2_C | CRPC      | 293                       | 9.63%      | PMS2                     | p.L83*, p.M136V and p.S229F | +                         | +    | +      | +      |
| WCMC21016_2_C | CRPC      | 147                       | 89.53%*    | PMS2                     | p.Q30P                      | +                         | +    | +      | +      |
| WCMC10_1_N    | CRPC      | 768                       | < 5%       | MSH6                     | p.R1095H                    | NA                        | NA   | NA     | NA     |
| WCMC131_1_C   | CRPC      | 458                       | NA         | -                        | none found                  | +                         | +    | +      | +      |
| WCMC21010_1_N | NEPC      | 299                       | < 1%       | MSH2                     | heterozygous deletion       | +                         | +    | absent | absent |
| WCMC21014_2_N | NEPC      | 274                       | < 1%       | MSH2                     | heterozygous deletion       | +                         | +    | absent | absent |

WCMC21010\_1\_N

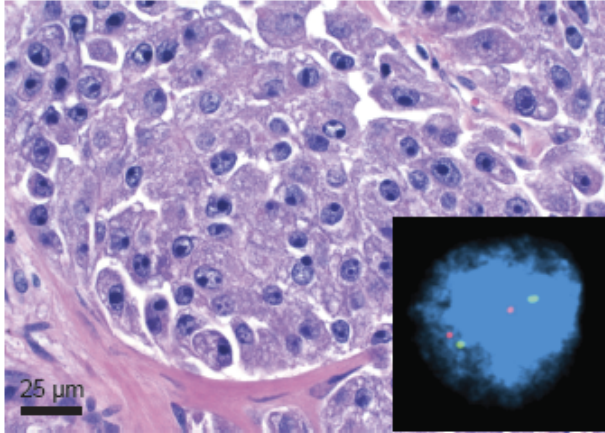


WCMC21014\_2\_N

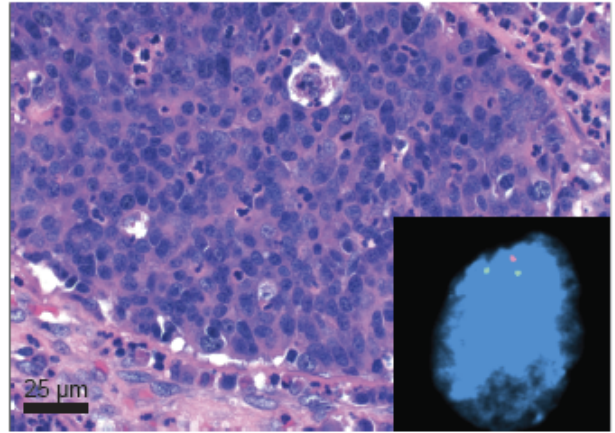


**Supplementary Figure 4: Characteristics of the six tumor samples with the highest number of non-synonymous SNVs.** (a) The four genes *MLH1*, *PMS2*, *MSH2* and *MSH6* were considered as "MMR genes". In case WCMC21016\_2\_C, polyploidy accounts for a high CNA burden. Immunohistochemistry for each of the four MMR proteins in samples WCMC21010\_1\_N (b) and WCMC21014\_2\_N (c) demonstrates a combined loss of nuclear MSH2 and MSH6 expression in tumor cells. Conserved expression in endothelial and stromal cells served as internal control. No loss of MLH1 or PMS2 expression is observed (scale bar, 100µm). (SNV - single nucleotide variation; MMR - mismatch repair; CNA - copy number alterations; IHC - immunohistochemistry; "+" - positive nuclear staining; NA - not available)

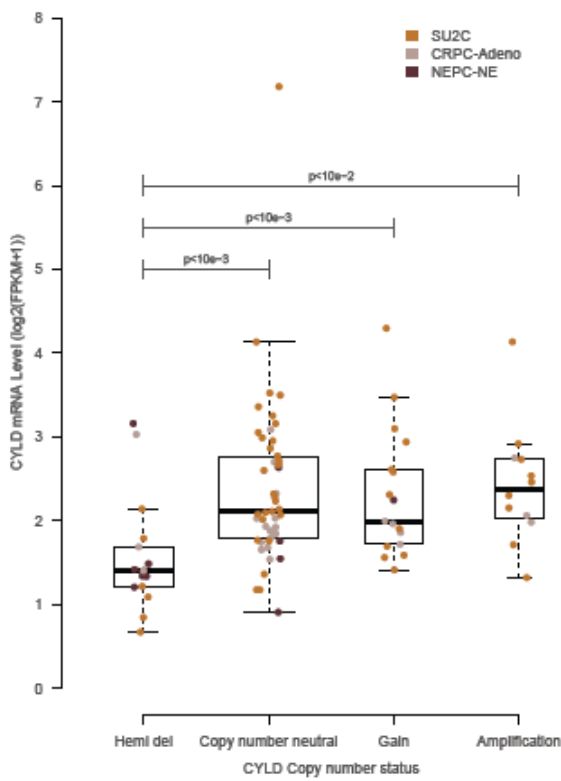
**a** WCMC159 (CRPC-Adeno)  
CYLD wt



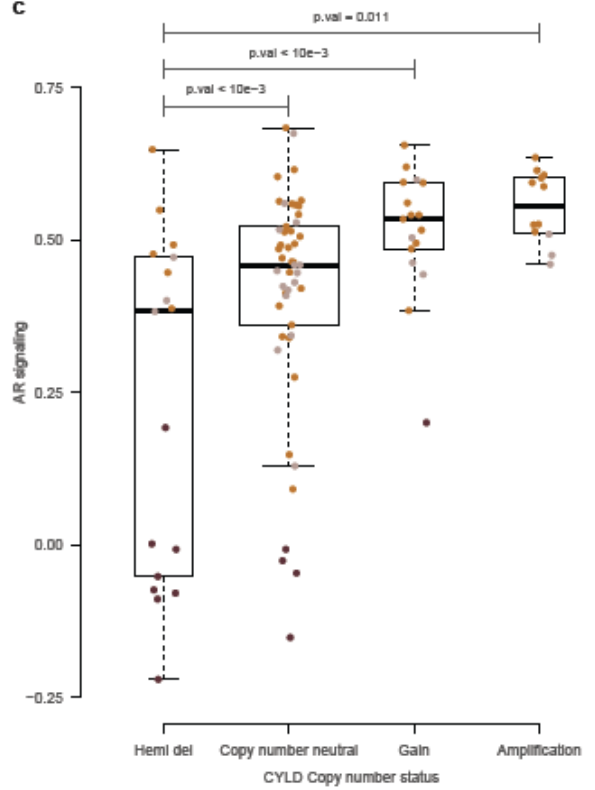
WCMC12 (CRPC-NE)  
CYLD heterozygous deletion



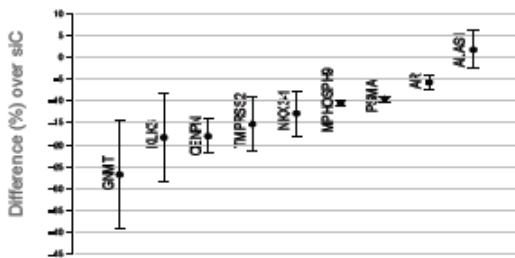
**b**



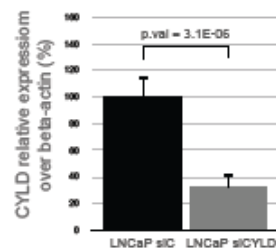
**c**



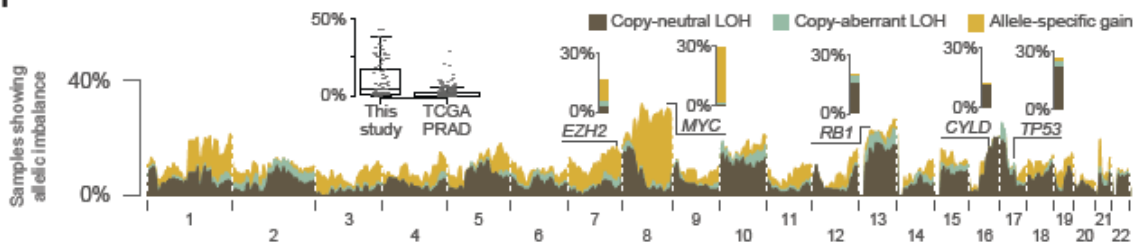
**d**



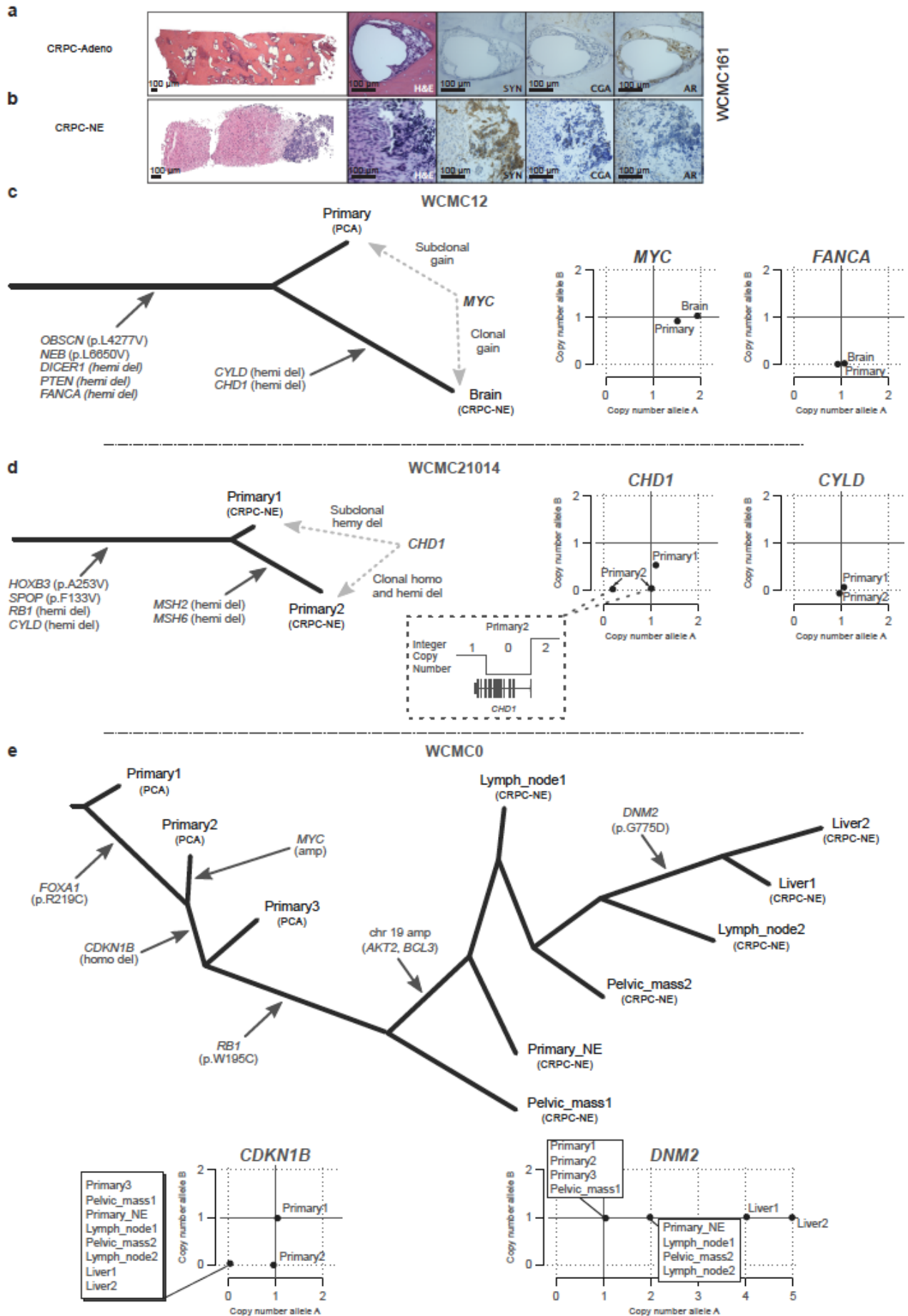
**e**



**f**

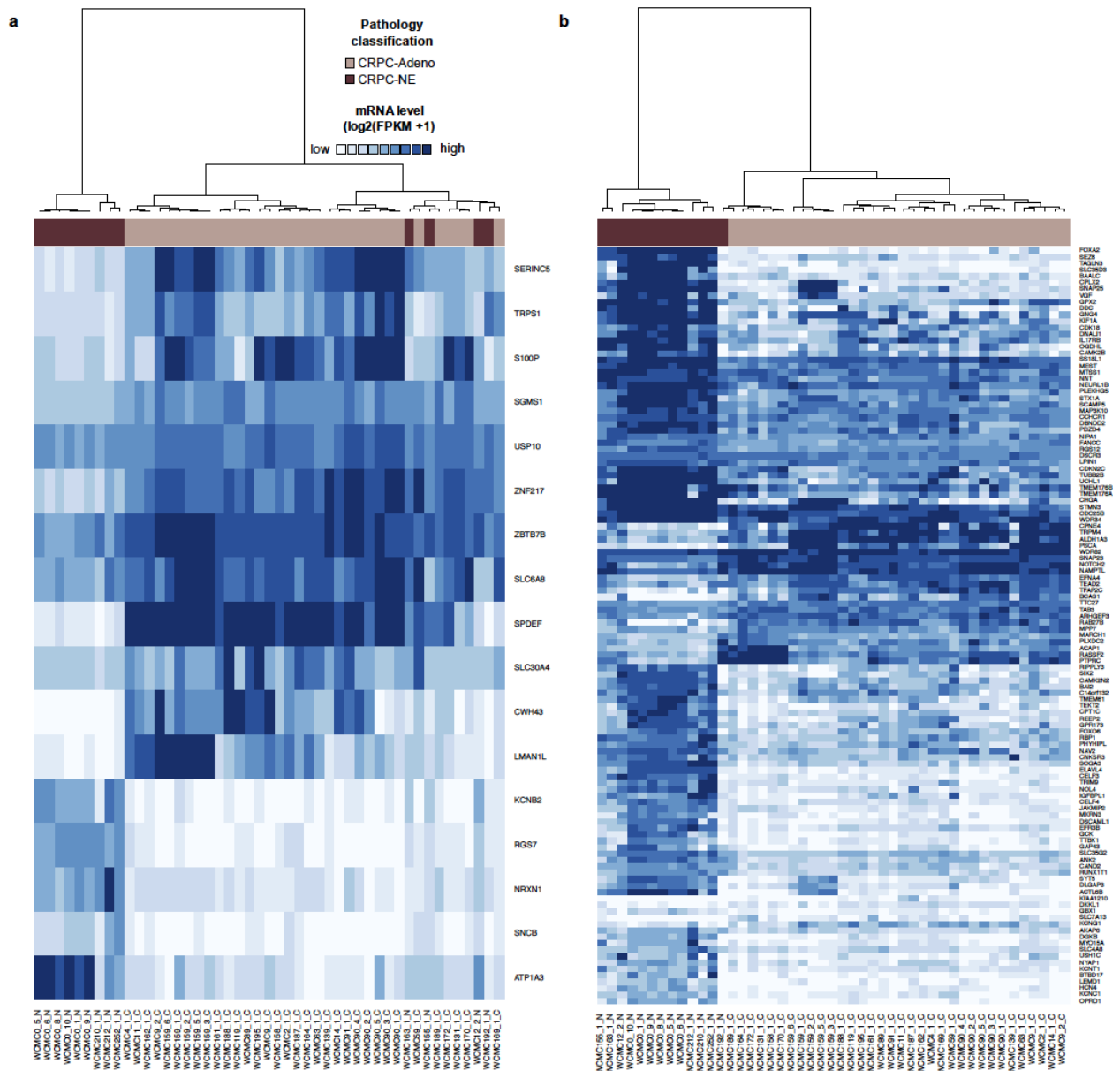


**Supplementary Figure 5: *CYLD* loss in CRPC-NE.** (a) *CYLD* fluorescence *in situ* hybridization (FISH). To assess *CYLD* deletion in tissues, we developed a dual-color FISH assay consisting of a locus specific probe (red) plus reference probe (green). At least 100 nuclei were evaluated per tissue section. Representative FISH images showing both wild type (wt) and *CYLD* deletion are shown above. (b) The boxplots compare the copy number state of *CYLD* against its transcription level (b) and AR signaling (c). The figures report the data for our cohort and SU2C<sup>5</sup>. Log<sub>2</sub> ratio (Log<sub>2</sub>R) values have been discretized in four states: hemizygous deletions (Hemi del) have Log<sub>2</sub>R adjusted by ploidy and purity (Log<sub>2</sub>R.adj) less than -0.4; Copy number neutral includes those samples with Log<sub>2</sub>R.adj between -0.4 and 0.4; Gain and Amplifications include samples with three or four copies and more than five copies of *CYLD*, respectively. Hemizygous deletion of *CYLD* leads to a significant decrease of *CYLD* transcript level (b) and a global reduction of AR signaling (c). All distributions are compared by means of Wilcoxon test. (d) To assess the impact of *CYLD* loss on AR signaling *in vitro*, LNCaP cells were transfected with siRNAs targeting *CYLD* or non-targeting (siC). RNA was extracted 48 hours post transfection and mRNA expression levels of AR signaling genes determined by nanostring. The plot shows the effect of siCYLD on the expression of a subset of androgen target genes (two independent experiments). 7 of 23 AR signaling genes were downregulated after siCYLD transfection including *KLK3*, *TMPRSS2*, *NKX3-1*, *PSMA*, *CENPN*, *GNMT*, *MPHOSPH9*. Values represent the difference of expression over siC. *ALAS1* is used as reference gene. (e) A reverse transcription followed by a quantitative PCR was performed. Data are presented after normalization with a housekeeping gene (B-actin) and are reported in a % scale (representative qPCR). Error bars correspond to standard deviation. This experiment shows the efficiency of siRNA targeting *CYLD*. P-value is calculated using t-test statistics. (f) Genome-wide frequencies of copy number allelic imbalance stratified by copy-neutral loss (brown), loss of heterozygosity gain (green), and allele specific gain (yellow). Significant increase in allelic imbalance in this advanced prostate cancer cohort compared to clinically localized tumors ( $P < 10^{-10}$ , Wilcoxon test). Highlighted genes *RB1*, *CYLD*, and *TP53* favor loss of heterozygosity while *EZH2* and *MYC* are characterized by gains with allele specific copy number.

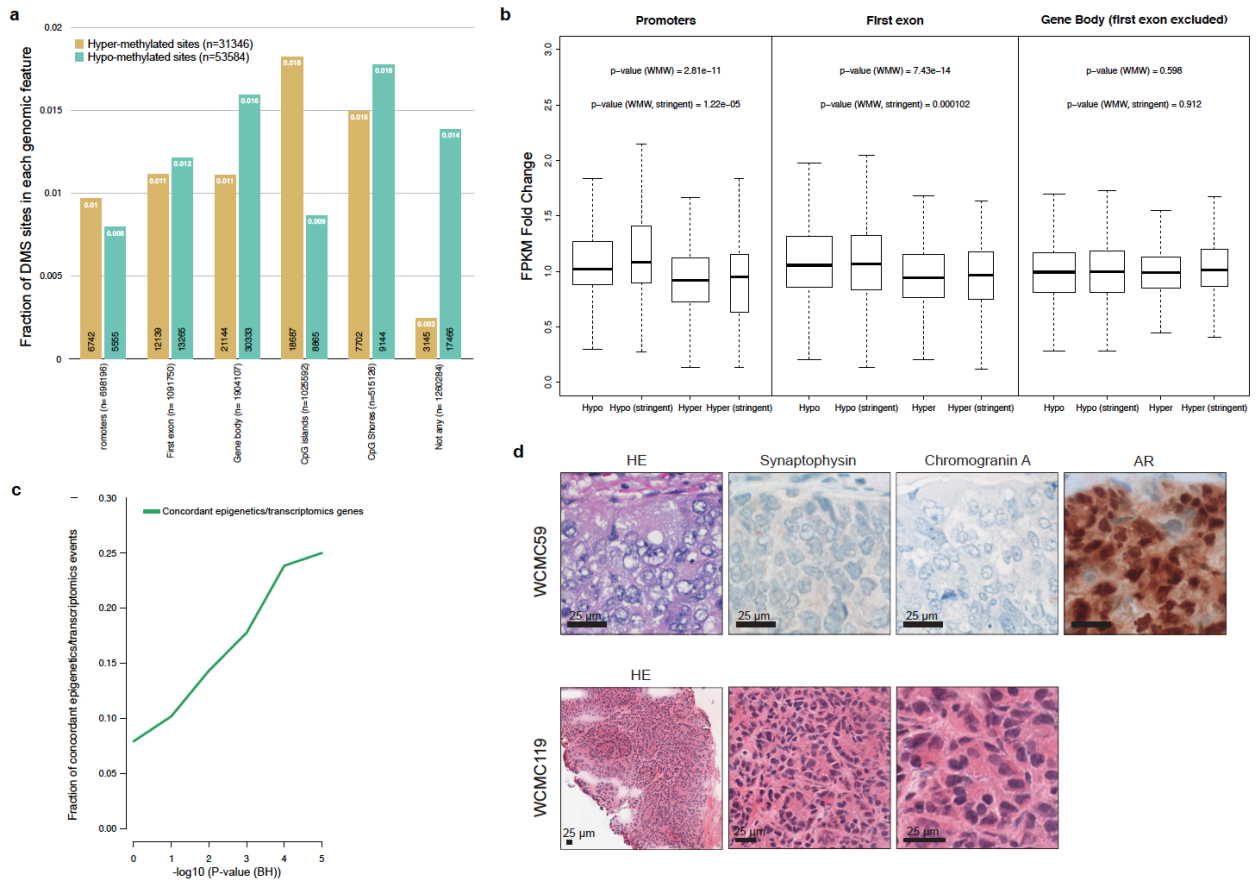




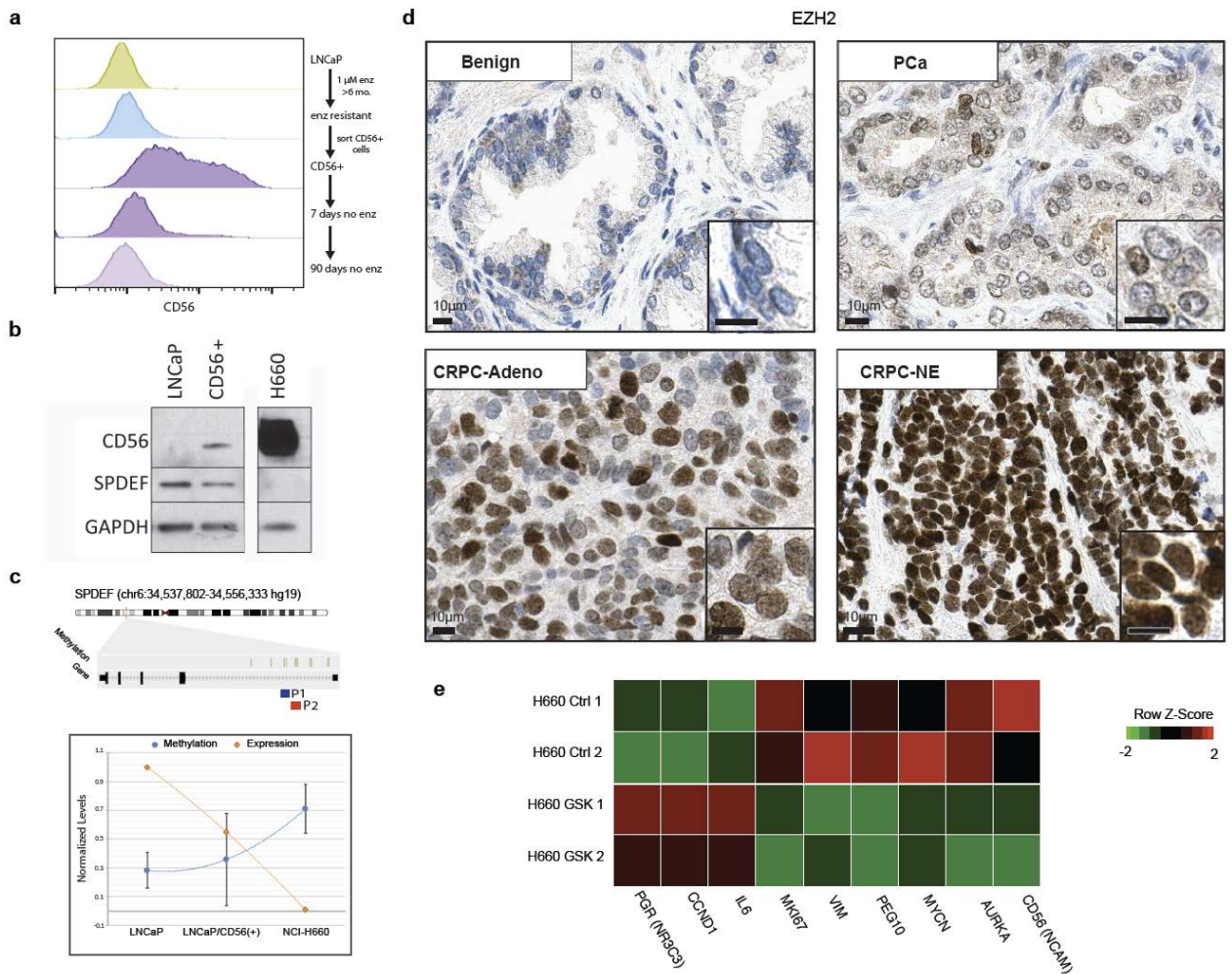
**Supplementary Figure 6: Tumor evolution and clonality analysis for multi-sample cases.** (a) CRPC-Adeno bone biopsy (time 43 months) with negative immunohistochemical staining for the neuroendocrine markers synaptophysin and chromogranin and positive for AR protein expression and (b) CRPC-NE liver biopsy (46 months) from WCMC161 (Fig. 2c). Tumor morphology is consistent with small cell carcinoma with positive immunohistochemical staining for the neuroendocrine markers synaptophysin and chromogranin and negative for AR protein expression. Phylogenetic reconstruction of multi-sample patients WCMC12 (c), WCMC21014 (d), and WCMC0 (e) is based on parsimony tree ratchet method<sup>6</sup> applied to allele specific copy number data. (c) WES of WCMC12 patient's primary adenocarcinoma is compared to a resected CRCP-NE brain metastasis, which developed 2 years after diagnosis. Certain alterations, such as *FANCA* deletion (right inset), are conserved during progression while other key drivers such as *MYC* gain (left inset) occurred later. (d) Genomic analysis of adenocarcinoma and neuroendocrine carcinoma foci within WCMC21014 patient's primary tumor points to an adenocarcinoma cell of origin of CRPC-NE and rules against independent or parallel evolution. Insets show that *CYLD* hemizygous deletion (right) is conserved through tumor evolution while *CHD1* homozygous deletion (left) is acquired later. (e) Multiple tumor foci (n=4) within WCMC0 patient's primary tumor (adenocarcinoma) were compared with multiple sites (n=7) of metastases (CRPC-NE) obtained at the time death (rapid autopsy). In this case, one primary is most similar to the metastases, suggesting this was the lethal clone of origin. Similar to the other cases, certain alterations are maintained with progression (such as *CDKN1B* loss, left inset) with later acquisition of other clonal alterations (e.g., *DNM2*, right inset), favoring a branched divergent clonal evolution model of CRPC-NE progression. Overall, the data from these additional cases provide further support and collectively seem to rule out parallel evolution and support divergent clonal evolution of CRPC-NE from an adenocarcinoma precursor.



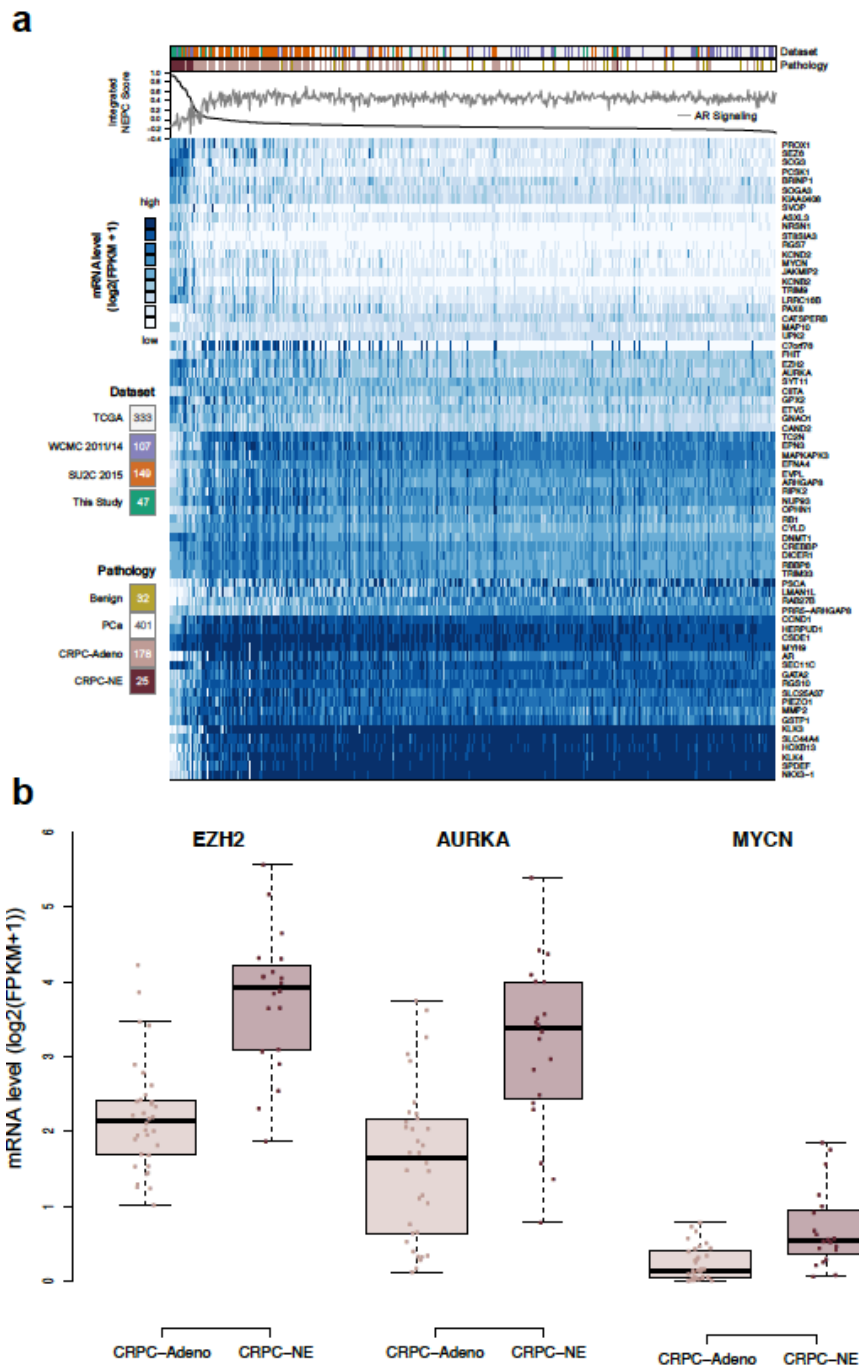
Supplementary Figure 7: Unsupervised clustering of differentially expressed genes that are also differentially methylated in CRPC-NE versus CRPC-Adeno. Heat maps generated by hierarchical clustering of the normalized FPKM values of genes that are differentially expressed ( $p$ -value (BH)  $< 10^{-3}$ ) and (a) hypo-methylated and (b) hyper-methylated across 47 RNA-seq samples. Pearson's correlation and Euclidean distance were used as distance measures for samples (columns) and genes (rows), respectively. Pathology classification (CRPC-NE, CRPC-Adeno) is indicated in the annotation track above the heat map. A quantile approach has been used to define the color-key data ranges. Of note, low AR gene expression in CRPC-NE was not explained by methylation changes in our dataset, and there was hypo-methylation with over-expression of the neuroendocrine-associated gene *CGHA* (chromogranin A) in CRPC-NE.



**Supplementary Figure 8: Overview of differential methylation analysis of CRPC-NE versus CRPC-Adeno.** (a) Column bars reporting the fraction of Differential Methylated Sites (DMS) divided by genomic feature annotation; within gene promoters, first exons, gene bodies, CpG Island, CpG shores and none of the previous (not any) with respect to the total number of sequenced sites. Absolute numbers are showed at the base of each column both for hypo- and hyper- methylated events. The total number of unique sequenced sites per genomic feature is reported in parentheses after the corresponding genomic feature name. All the differences between hypo- and hyper- methylated events within each category are statistically significant (Fisher Exact Test,  $p$ -value  $< 10^{-100}$ ) (b) The box plots show the association between differential methylation status and gene expression (fold change CRPC-NE versus CRPC-Adeno), stratified by genomic features: gene promoters, first exons and gene body (excluding first exon). The term “stringent” refers to DMSs with  $\text{methylKit.meth\_diff} > 40$  (see **Supplemental Table 8**). As expected<sup>7</sup>, methylation of first exons and gene promoters are associated with gene expression ( $p$ -values reported over the box plots). (c) Concordance analysis between epigenetic and transcriptomic status. Fraction of concordant epigenetic/transcriptomic events falling into different bins of adjusted  $p$ -values (BH) from differential expression analysis. Concordant epigenetic/transcriptomic events are defined as genes hyper-methylated or hypo-methylated in their promoter/first exon and under- or over- expressed. (d) Pathology of the cases of CRPC-Adeno that clustered with CRPC-NE based on unselected methylation sites (shown in **Fig. 3a**). Despite aggressive clinical features, morphologic and immunohistochemical profiling studies are consistent with CRPC-Adeno with negative staining for the neuroendocrine markers synaptophysin and chromogranin and positive for AR protein expression.

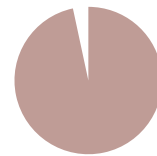
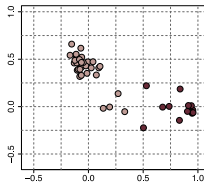


**Supplementary Figure 9: Enzalutamide resistant LNCaP cells and SPDEF expression and methylation across cell lines.** (a) AR-dependent cell line LNCaP was grown in the presence of 1 $\mu$ M enzalutamide for >6 months. FACS sorting is shown which was used to obtain a population of CD56 expressing, enzalutamide resistant cells (CD56+). Upon removal of enzalutamide (7 days, 90 days), the cells revert back and no longer express CD56. (b) Western blot of LNCaP, CD56+ and H660 cell lysate for CD56, SPDEF, and GAPDH (c) SPDEF methylation was assessed by a targeted methyl-PCR-based (MassArray) approach and the SPDEF gene is depicted with areas of promoter methylation and areas for P1 and P2 primers selection for MassArray. Both methylation and mRNA expression by nanostring were assessed across the cell lines. The NCI-H660 has higher methylation and lower expression of SPDEF compared to the adenocarcinoma (LNCaP). Enzalutamide resistant LNCaP –CD56 cells lie at an intermediate range between LNCaP and NCI-H660. (d) Representative images of EZH2 immunohistochemistry. Progressive increase in EZH2 protein expression across benign prostate, clinically localized hormone naïve prostate cancer (PCa), CRPC-Adeno, and CRPC-NE. (e) mRNA expression changes in the neuroendocrine prostate cell line NCI-H660 after 48 hours of drug treatment with the EZH2 inhibitor GSK343 (10 $\mu$ M). There is significant downregulation of the neuroendocrine associated genes CD56, AURKA, MYCN, and PEG10, the epithelial mesenchymal transition gene VIM (vimentin), and the proliferation marker MK167 (Ki67). IL6, CCND1, PGR are included as positive controls.

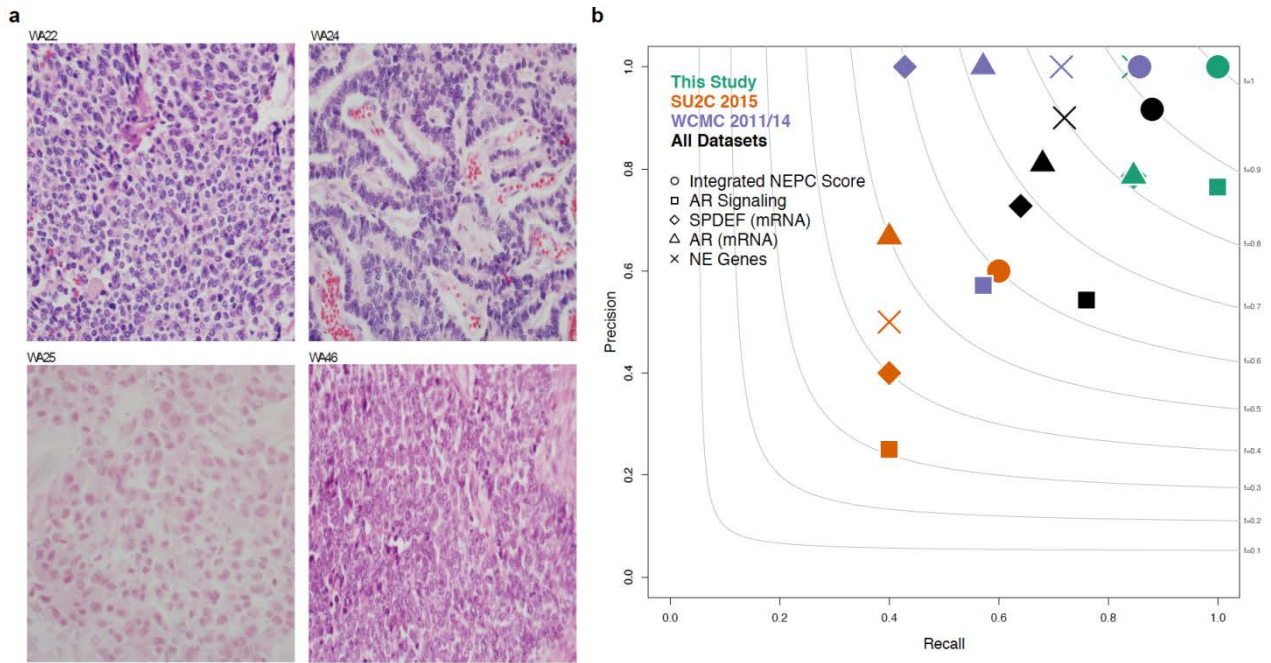


**Supplementary Figure 10: (a)** Integrated NEPC score analysis across 636 samples from four different prostate cancer datasets. Integrated NEPC score analysis across 636 RNA-seq samples from four different prostate cancer datasets (This Study, SU2C 2015, WCMC 2011/14 and TCGA). Samples are ordered by decreasing values of Integrated NEPC score. Top, annotation tracks report original dataset and pathology classification. Middle, plot reports Integrated NEPC score (black line) and AR signaling (grey line) across samples. Bottom, heat map of normalized FPKMs for the 70 high priority genes (in rows) across the samples (in columns). The number of samples for each pathology classification is reported inside the square symbols of the corresponding legend. **(b)** mRNA levels of *EZH2*, *AURKA* and *MYCN* in CRPC-NE and CRPC-Adeno samples. Box plots showing mRNA levels estimated by normalized FPKMs in CRPC-NE and CRPC-Adeno for three genes: *EZH2* (left), *AURKA* (middle) and *MYCN* (right). Data are from 54 samples of 2 different datasets (This Study, WCMC 2011/14), corresponding to 20 CRPC-NE and 34 CRPC-Adeno. P-values of Mann-Whitney-Wilcoxon test are  $<10^{-7}$ ,  $<10^{-5}$ , and  $<10^{-4}$  for *EZH2*, *AURKA*, and *MYCN*, respectively.

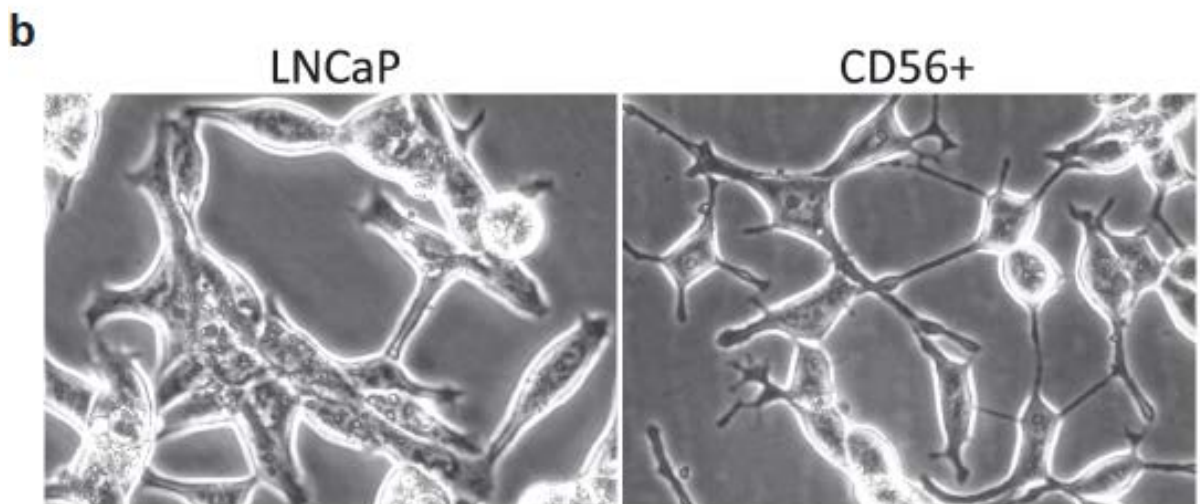
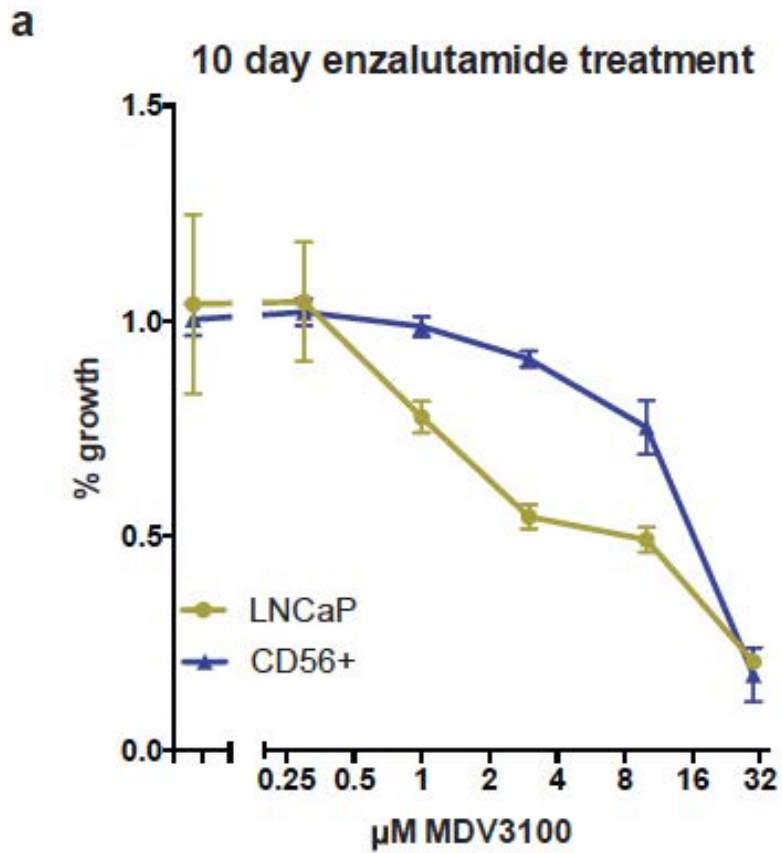
This Study (Discovery dataset)



**Supplementary Figure 11: AR signaling versus Integrated NEPC Score.** (a) AR signaling versus Integrated NEPC score across 730 RNA-seq/microarray samples from five independent prostate cancer datasets (This Study, TCGA, WCMC 2011/14, SU2C 2015, Michigan 2012) using transcriptome data as proxy. Each row refers to different definition of Integrated NEPC Score (see **Supplementary Notes**). For Michigan 2012 dataset, Integrated NEPC Score was calculated on available data (see Integrated NEPC Score section in **Online Methods**). (b) Pie charts showing sample size and pathology classification for the 730 samples across the five datasets.

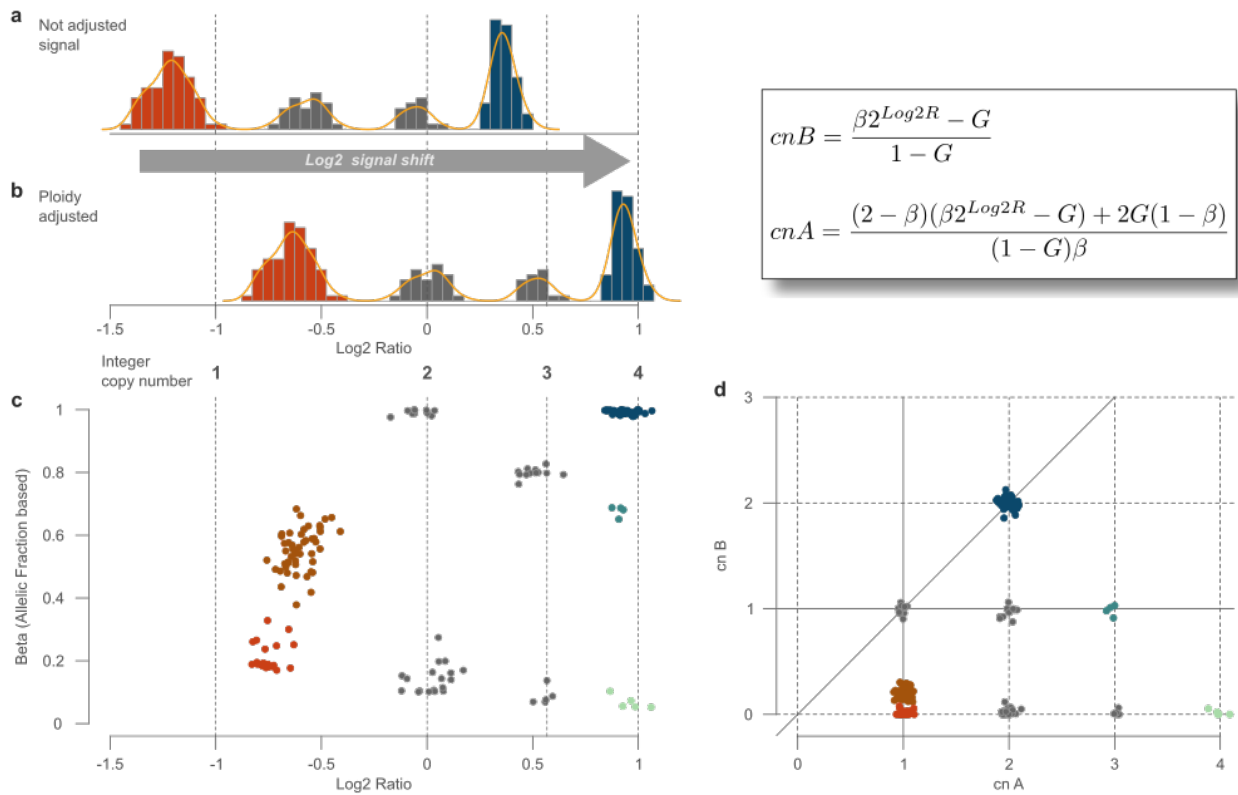


**Supplementary Figure 12: Rapid Autopsy Cases from the University of Michigan Cohort with elevated NEPC score.** (a) WA22 (liver metastasis) and WA24 (liver metastasis) have characteristics of small cell carcinoma. WA46 (lung metastasis) demonstrates morphologic features of prostate carcinoma with neuroendocrine differentiation (CRPC-NE). By report, case WA25 corresponds to a high-grade carcinoma with neuroendocrine differentiation in the prostate (shown), which metastasizes to the soft tissue and lymph node had a squamous-like morphology (not shown). (b) Accuracy of Integrated NEPC scores and other classifiers in the prediction of CRPC-NE samples. Conventionally used NEPC markers are included. Prediction accuracy of CRPC-NE samples by precision and recall statistics for Integrated NEPC Score (circles), AR signaling (squares), mRNA level of *SPDEF* (diamonds) and *AR* (triangles), and the neuroendocrine marker genes *CGHA*, *SYP*, *NSE*, *CD56* (crosses) in the following RNA-seq datasets: This Study (green), SU2C 2015 (orange), WCMC 2011/14 (violet) and all of these RNA-seq datasets (black). Grey curves represent F-measure levels, defined as the harmonic mean of precision and recall. Due to the absence of CRPC-NE samples (positive events), TCGA data were not reported here. In each dataset, Integrated NEPC score outperforms the other classifiers in terms of both precision and recall.

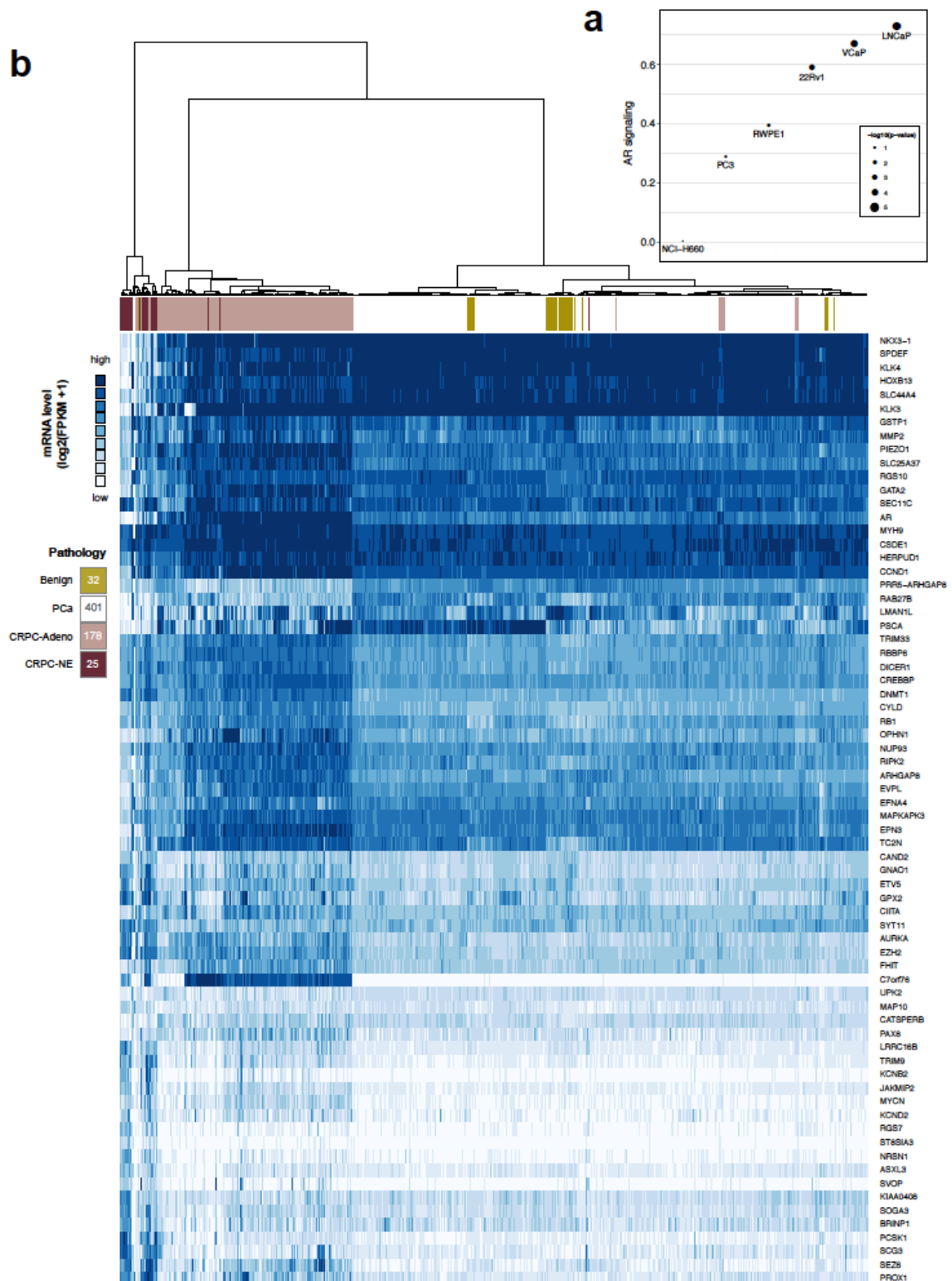


**Supplementary Figure 13: Enzalutamide resistant cell line transitions toward CRPC-NE -like phenotype.** (a) 10 day enzalutamide challenge of LNCaP and CD56+ LNCaP cell lines analyzed with CellTiter-Blue viability assay. Error bars report standard error of the mean of quadruplicate reads and results are representative of 3 independent assays. (b) 20x light microscope image of LNCaP and CD56+ cell line. The CD56+ enzalutamide resistant cells acquire neuronal/neurite like morphologic characteristics.





**Supplementary Figure 14: Schematic of CLONET 2.0 new features.** Different ways of interpreting copy number data. (a) *Log2 ratio* histogram of a tumor sample *T*. (b) Ploidy adjusted *Log2 ratio* of sample *T* as predicted by CLONET. (c) *Beta vs Log2 Ratio* representation of the segments of *T*. Clonal (orange) and subclonal (brown) deletions contribute to the orange peak of (b). Dark, mid, and light blue clusters completely characterizes copy number 4 segments in (b). (d) Allele specific copy number profile of *T* as predicted by the formulas in the inset. Code color associates segments in (c) and (d) representation.



**Supplementary Figure 15:** (a) AR signaling score and Integrated NEPC score genes. AR signaling score was computed as described in the methods. The most AR responsive cell lines (VCaP and LNCaP) have the highest score, whereas androgen independent cell lines, e.g. NCI-H660, have low scores. (b) Unsupervised clustering of prostate samples based on 70 high priority genes. Heat map generated by hierarchical clustering of the normalized FPKM values of the 70 high priority genes (**Supplementary Table 9**) across 636 RNA-seq samples from four different prostate cancer datasets (This Study, SU2C 2015, WCMC 2011/14 and TCGA). The number of samples for each pathology classification is reported inside the square symbols of the legend. A quantile approach has been used to define the color-key data ranges.

## **Supplementary Tables Legends**

**Supplementary Table 1: Baseline clinical and pathologic features.** Age, serum markers, and prior therapies were available for 48 CRPC-Adeno patients and 20 CRPC-NE patients. Median values or n (% or range) are reported above for age (years), serum PSA (ng/ml), serum chromogranin (upper limit of normal = 95 ng/ml), serum neuron specific enolase (NSE) (upper limit of normal= 8.9ug/L). Pathology subclass was determined by consensus criteria for all the cases): A (adenocarcinoma), B (adenocarcinoma with neuroendocrine differentiation), C (small cell carcinoma), D (large cell carcinoma), E (mixed small cell carcinoma-adenocarcinoma). ERG fusion considered positive if detected by at least one of the following methods: FISH, WES, IHC, RNA-seq, Nanostring.

**Supplementary Table 2: mRNA expression levels of AR signaling genes in CRPC-Adeno and CRPC-NE.** The expression levels (FPKM) of the genes used as “AR signaling” signature computed via RSEQtools as described in **Online Methods**. Sample annotation is provided in **Supplementary Table 11**.

**Supplementary Table 3: Somatic single nucleotide variants in protein coding regions.** Non-silent single nucleotide variants (SNV) in 114 Exome sequenced advanced prostate tumors. Standard information about SNV provided by our pipeline (MuTect + Oncotator) is enriched with CLONET specific columns: “Segment start”, “Segment end”, and “Segment Log2 ratio” report for each SNV *pm* the start position, the end position and *Log2 ratio* adjusted by ploidy and purity of the genomic segment containing *pm*; “VAF corrected” reports the variant allelic fraction (VAF) corrected per ploidy and purity; “VAF clonality threshold” reports the threshold of the VAF computed by CLONET to call clonality; “Clonality status” defines if a SNV is clonal or subclonal.

**Supplementary Table 4: Differential analysis of single nucleotide variants in protein coding regions.** The table reports aggregated data from single nucleotide variants (SNVs) reported in **Supplementary Table 3**. For each gene, the table details absolute number of mutated specimens in all cases (“# All Specimens”), CRPC-NE cases (“#CRPC-NE Specimens”), CRPC-Adeno cases (“#CRPC-Adeno Specimens”), as well as relative frequencies (“% All Specimens”, “%CRPC-NE Specimens”, and “%CRPC-Adeno Specimens”, respectively). Then, the same data aggregated by patients is reported in columns “# All Patients”, “#CRPC-NE Patients”, “#CRPC-NE Patients”, “% All Patients”, “%CRPC-NE Patients”, and “%CRPC-NE Patients”. Finally, p-value and associated FDR correction of the proportion test between “#CRPC-NE Patients” and “#CRPC-NE Patients” are reported in columns “p-val Binomial test” and “FDR”, respectively.

**Supplementary Table 5: Canonical gene fusions.** ERG fusion as determined by deletions between TMPRSS2 and ERG on chr21 detected by WES, ERG FISH breakapart assay, Erg IHC, and/or fusion transcript level by Nanostring or RNAseq (FusionSeq) since not all assays were available for each sample. Other canonical gene fusions listed were determined by FusionSeq. \*ETV1: This case with ETV1 overexpression harbors a putative inter-chromosomal fusion between DGKB, a gene located next to ETV1, and FOXA1. This could reflect a cryptic insertion of the ETV1 gene into the FOXA1 gene, similarly to a mechanism previously described<sup>8</sup>.

**Supplementary Table 6: Somatic copy-number alterations in 114 Exome-sequenced advanced prostate tumors.** The table accounts for the genomic segment identified in each

sample of our cohort prior and after ploidy and purity adjustment. Each line corresponds to a genomic segment and reports the identification of the sample (“Specimen ID”), the “Chromosome”, the “Start Position”, and the “End Position”, the “Allelic Fraction” of the informative SNPs, the *Log2 Ratio* uncorrected (“Log2 Ratio”), adjusted by ploidy (“Log2 Ratio ploidy adjusted”) and adjusted for both ploidy and purity (“Log2 Ratio ploidy and purity adjusted”).

**Supplementary Table 7: Differential analysis of somatic copy number and gene expression.** The table reports copy number status and expression levels of a set of approximately 19K RefSeq genes. Information about each gene is in columns A to J. Columns K and L specify mean expression levels in CRPC-Adeno and CRPC-NE samples, respectively. Column M distinguishes between valid (enough) and not valid (too low) expression levels. Columns N to S report the number of copy number events in CRPC-NE and CRPC-Adeno samples; neutral events refer to genes that lie in a segment with copy number 2. Columns T to Y report p-values and associated FDR for differential expression, deletion and amplification. Finally, columns Z to AC report frequencies of deletions and amplifications in pathological classes.

**Supplementary Table 8: Differentially methylated sites and genomic features annotation.** List of hypo- and hyper-methylated sites identified by methylKit and their annotation to the following genomic features: Promoters, First Exons, Gene Bodies, CpG Islands and CpG Shores. “methylKit.pvalue”, “methylKit.qvalue”, “methylKit.meth\_diff” refer to the statistics estimated by methylKit.

**Supplementary Table 9: Integrated CRPC-NE score high priority genes.** Summary table of the 70 genes used to calculate the Integrated NEPC score. Inclusion criteria refer to the following priority rule: Methylation overall and DNA over RNA. In addition, “a-priori” refers to EZH2, MYCN, AURKA and RB1 previously described as associated with CRPC-NE phenotype. Details about gene status across each layer (DNA, RNA, Methylation) are reported. P-values (CRPC-NE vs CRPC-Adeno) are calculated using Wilcoxon Mann Whitney test by considering the mRNA levels ( $\log_2(\text{FPKM}+1)$ ) of 20 CRPC-NE and 105 CRPC-Adeno samples.

**Supplementary Table 10: Expression datasets and samples classification.** Summary of expression datasets used to test Integrated NEPC score in correctly classifying CRPC-NE samples. For each dataset, the data type (Microarray or RNA-seq), the total number of samples (“Samples”) and the number of samples for each pathology classification (Benign, PCa, CRPC-Adeno, CRPC-NE) are reported.

**Supplementary Table 11: Sample annotation file.** This includes Patient and Sample Identification numbers for each case and control, with corresponding Pathologic Classification, Tumor Purity, Ploidy, Genomic Burden, Number of coding somatic mutations, and BAM ID.

**Supplementary Table 12: Exome sequencing statistic summaries.** Statistics from exome sequencing experimental data are reported for each sample.

**Supplementary Table 13: RNA sequencing statistic summaries.** The table covers basic statistics of the RNA sequencing samples in this study reporting for each sample its id (“Specimen ID”), the id of the patient (“Patient ID”), the number of sequenced and mapped reads (“Total reads” and “Mapped reads”, respectively), and the fraction of sequenced reads that were mapped (“% mapped reads”).

**Supplementary Table 14: CRPC-NE prediction statistics in RNA-seq datasets.** CRPC-NE prediction accuracy of Integrated NEPC score, AR signaling and mRNA level of *AR* and *SPDEF* by using Precision, Recall and p-values and Odd-Ratio (OR) from Fisher Exact Test as statistical measures. Statistics are computed as described in **Supplementary Notes**. Only RNA-seq datasets were considered for this analysis.

**Supplementary Table 15: Integrated NEPC Score and AR Signaling in prostate datasets.** Integrated NEPC score and AR signaling across 730 samples from five different prostate expression datasets (This Study, SU2C 2015, WCMC 2011/14, TCGA, Michigan 2012). The Pathology classification is reported for each sample. Sample size for each dataset is reported in **Supplementary Table 10**.

## References

- 1 Prandi, D. *et al.* Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome biology* **15**, 439, doi:10.1186/s13059-014-0439-6 (2014).
- 2 Romanel, A., Lago, S., Prandi, D., Sboner, A. & Demichelis, F. ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med Genomics* **8**, 9, doi:10.1186/s12920-015-0084-2  
10.1186/s12920-015-0084-2 [pii] (2015).
- 3 Manoranjan Dash, H. L. *Consistency-based search in feature selection*. Vol. 151(1-2) 155-176 (Elsevier, 2003).
- 4 Epstein, J. I. *et al.* Proposed morphologic classification of prostate cancer with neuroendocrine differentiation. *The American journal of surgical pathology* **38**, 756-767, doi:10.1097/PAS.0000000000000208 (2014).
- 5 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228, doi:S0092-8674(15)00548-6 [pii]  
10.1016/j.cell.2015.05.001 (2015).
- 6 Nixon, K. C. The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. *Cladistics* **15**, 407-414 (1999).
- 7 Brenet, F. *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One* **6**, e14524, doi:10.1371/journal.pone.0014524 (2011).
- 8 Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101, doi:nature07638 [pii]  
10.1038/nature07638 (2009).