doi:10.1038/nature11125

Supplementary Results

The mutational landscape of CRPC by whole exome sequencing

In total, we generated 25,525,520,145 bases, with an average 116-fold coverage of each targeted base per tissue sample, and 91.78% of annotated targeted bases with sufficient coverage to call somatic mutations (**Supplementary Tables 2&3**). A total of 3,875 high confidence protein-altering somatic mutations (**see Methods**) were identified in 3,044 genes (out of ~19,365 targeted coding genes) among the 61 tumours, including 3,169 missense, 203 nonsense, 68 splice site mutations, and 435 indels (**Supplementary Table 4**). Neutral mutations were also identified, including 2,179 intronic and 1,225 synonymous (**Supplementary Table 5**). Coverage rates for all genes per sample are provided in **Supplementary Table 6**. Confirmation as somatic by Sanger sequencing of candidate point mutations (219/227, 96%) and indels (16/16, 100%), confirmed the stringency of our somatic mutation calling algorithm (**Supplementary Table 4 and Supplementary Fig. 1**). The estimated average tumour content for CRPC and localized prostate cancer samples was 68% (range 40%-100%) and 56% (range 35%-77%), respectively (p =0.04) (**Supplementary Fig. 2)**.

Of the 3,875 identified non-synonymous somatic mutations, only 54 somatic SNVs are present in COSMIC, including, but not limited to, one each in *SPOP, ARID1A,* and *KRAS* (G12V), two in *TTN*, three each in *APC, CTNNB1,* and *RB1* and 23 in *TP53* (see **Methods**). The average number of mutations per tumour was 46.6 over an average of 28.7 Mb of annotated targeted bases in each exome with sufficient coverage to call somatic mutations (range 13-100 somatic mutations per sample, **Supplementary Fig. 3**), excluding three samples with outlier number of mutations: WA56 (169 mutations), WA48 (238 mutations) and WA16 (731 mutations).

Rare CRPC xenografts with outlier number of mutations were observed by Kumar *et al.* [1], in one case likely due to a mutation in the mismatch repair gene *MSH6* previously associated with Lynch syndrome. In our cohort, WA16 harbored a somatic, focal homozygous deletion in the mismatch repair gene *MSH2*, while WA48 harbored a somatic homozygous deletion of a ~2MB region on chr 13 harboring *BRCA2* (**Supplementary Fig. 4**).

In our cohort, the mutation rate for localized prostate cancers (0.93/Mb) was consistent with the rate observed in the whole genome sequencing of seven localized prostate tumours (0.9/Mb)[2] and with the low reported rates in other targeted studies of localized prostate cancer (0.33 and 0.31/Mb[3,4]). The mutation rate for heavily treated CRPC (2.00/Mb) was only two-fold higher than that of the localized tumours.

Mutational spectrum of castrate resistant prostate cancer

Based on the low mutation rate, the metastatic prostate cancer mutation signature likely does not reflect exposure to tobacco carcinogens, UV light or mutagenic alkylating chemotherapy[5], consistent with lack of etiologic associations with prostate cancer. The metastatic prostate cancer mutation signature was enriched for C to T transitions at 5'-CG base pairs (30.5% of nonsynonymous mutations) (**Supplementary Fig. 5**), similar to the mutational spectrum of ovarian clear cell carcinoma identified by exome sequencing[6], and gastric[5], colorectal[5,7,8] and pancreatic adenocarcinoma[9], and glioblastoma multiforme[10]. Unlike breast[5,8], lung and ovarian carcinoma, and melanoma[5], the prostate cancer mutation signature is not enriched for C:G>G:C changes at 5'-TC base pairs. The localized prostate cancer mutation spectrum was almost identical to the spectrum for metastatic prostate cancer ($R^2 = 0.974$), indicating that heavy treatment does not substantially alter the types of mutations arising in prostate cancer with C to T transitions at 5'-CG being the dominant type of mutation (27.9% of nonsynonymous mutations) in localized and metastatic prostate cancer,

Sequencing of different foci confirms the monoclonal origin of lethal CRPC

Previously, we and others have used multiple lines of evidence, including the clonality of ETS gene fusions and copy number profiles to demonstrate the monoclonal origin of lethal CRPC[11-13]. To confirm these findings at the mutational level, we profiled three foci (bladder [WA43-44], celiac lymph node [WA43-27], and right lung [WA43-71]) from a 52 year old man who died of CRPC 5 years after initial treatment with radical prostatectomy, which demonstrated high-grade (Gleason score 9), organ confined disease with focally positive margins, and subsequent treatment with anti-androgen therapy, external beam radiation to the tumour bed, and numerous chemotherapeutics. As shown in **Supplementary Figure 6**, we identified 59 mutations in the bladder, 55 in the celiac lymph node and 47 in the right lung focus; 37 mutations were present in all three foci, including mutations in *TP53* and *PIK3C2A*, consistent with monoclonal origin.

Comparison of nonsynonymous mutations to previously published prostate cancer genomes and exomes

We compared our nonsynonomous mutations to nonsuynonmous mutations observed in prostate cancer genomes and exomes, reported by Berger *et al.*[2] and Kumar *et al.* [1], respectively. Berger *et al*. recently reported the genomes of seven localized prostate cancers [2], and 26 genes harbored nonsynonymous mutations in both studies, representing significant overlap (26 overlapping genes out of 2,485 genes harboring nonsynonymous mutations in this study, [excluding WA43-27, WA43-71, and WA16] and 105 genes harboring nonsynonymous mutations in Berger *et al.*, out of 19,365 total genes sequenced, Fisher's exact test, p = 0.0006). Both studies identified mutations effecting the same residue (F133) in *SPOP*, (**Fig. 1, Supplementary Table 4**), which has been identified in a prostate cancer sample previously [3]. Similarly, *CHD1* harbored splice site mutations in a single sample in both studies (**Fig. 1, Supplementary Table 4**). Kumar *et al.* recently reported putative somatic mutations from 23 prostate cancer exomes from unmatched xenograft samples (derived from 16 metastatic samples and three high-grade localized cancers)[1], and 18 genes harbored recurrent mutations in both studies, representing significant overlap (18 overlapping genes out of 396 genes with recurrent mutations in this study and 131 genes with recurrent mutations in Kumar *et al.* out of 19,365 total genes, Fisher's exact text, p = 2E-10).

Exome copy number profiling and aCGH

Exome sequencing data can be used to identify somatic copy number alterations [14], and we applied this methodology to all profiled CRPC and localized prostate cancer samples (see

Methods and **Supplementary Fig. 7**). As shown in **Supplementary Fig. 8**, we identified recurrent aberrations previously associated with prostate cancer development and progression, including broad losses of 1p, 8p and 6q, and gains of 1q, 3q, 7q and 8q, and deletions between *TMPRSS2* and *ERG* (in cases with *TMPRSS2:ERG* fusions through deletion)[12,13,15,16].

Gene expression profiling identified over-expression of DLX1 in prostate cancer and CRPC

Matched aCGH and gene expression profiling was performed on 3 localized prostate cancers and 31 metastatic CRPCs subjected to exome sequencing, as well as an additional 28 benign prostate tissues, 56 localized prostate cancers and 4 CRPCs (**Supplementary Table 10**). Generated profiles were uploaded into Oncomine (www.oncomine.com) for automated data processing, analysis and visualization. Global gene expression profiles for benign prostate tissue, localized prostate cancer and CRPC were similar to previous studies (analyses available in Oncomine), although we identified *DLX1*, a gene not monitored in most previous microarray studies, to be the most differentially expressed gene between benign prostate tissue and localized prostate cancer (**Supplementary Fig. 10a**, fold change 22.4, p = 7.2E-27), with *AMACR* (fold change 13.1, p = 4.57E-24), which is currently used diagnostically (by immunohistochemistry) as a prostate cancer biomarker, being the second most differentially expressed gene. We confirmed the differential expression of *DLX1* by qPCR in prostate cancer (both localized and CRPC, n = 62, median 418) compared to benign prostate tissue *(n* =10, median 1.0, Mann Whitney test p< 0.0001) (**Supplementary Fig. 10b**). We also confirmed the over-expression of DLX1 by western blotting in both localized and CRPC compared to benign tissue (**Supplementary Fig. 10c**).

Integration of exome sequencing with transcriptome sequencing of prostate cancer cell lines

As transcriptome sequencing has also been used to discover recurrent mutations in cancer[17,18], we analyzed the transcriptome of 11 prostate cancer cell lines (primarily CRPC, **Supplementary Table 11**), sequenced using the Illumina GAIIx platform, comprising 22,731,390,482 bases, and identified an average of 5,905 known coding polymorphisms and 1,031 novel protein-altering variants (756 point mutations and 275 indels) per sample (**Supplementary Table 12**). Given the lack of normal genomic DNA from these cell lines, germline and somatic variants cannot be distinguished. Thus, we only considered variants fulfilling one of three high stringency filters as likely somatic mutations: 1) deleterious variants affecting a gene harboring a somatic mutation in our study (**Supplementary Table 13**), 2) variants affecting the same nucleotide as a somatic mutation in our study (**Supplementary Table 14**), or 3) variants affecting the same nucleotide as a confirmed somatic variant in COSMIC (**Supplementary Table 15**).

This integrative approach identified additional variants in *TP53*, *AR* and *APC,* supporting the utility of the analysis. A *TP53* R248W variant, present in WA10 and previously reported as somatic[19], was identified in the VCaP cell line, while previously reported P223L and V274F somatic variants were identified in DU-145[16], with a V274G variant present in WA37. A confirmed somatic *TP53* variant R175H was identified in both WA30 and LAPC-4, consistent with previous reports (**Supplementary Table 15**)[19]. Finally, a Y234H confirmed somatic variant (predicted to be probably damaging) was also present in C4-2B (**Supplementary Table 15**). This approach also identified additional mutations in *AR*, including additional T878A mutations, which has been reported as frequently mutated in CRPC [20], in LNCaP (and its derivative C4-2B) and MDA-PCa-2B (**Supplementary Table 14**). MDA-PCa-2B also harbored the previously reported somatic mutation L702H[21], while 22RV1 (and its parental line CWR22) harbored a previously confirmed somatic H875Y variant[22] (**Supplementary Table 14**). Finally, WA40 and WA52 harbored a nonsense mutation (E1576*) and a frameshifting indel, respectively, in *APC*, while MDA-PCa-2B harbored a missense variant (K1454E) (**Supplementary Table 15**) previously confirmed as a somatic mutation in urothelial carcinoma[23].

Importantly, integrating transcriptome sequencing data also identified recurrent variants in genes not previously identified as being mutated in prostate cancer, including *STAG2*, *MLL3*, *CNOT1*, *FAM123B* (*WTX*) and *FOXA1* (**Supplementary Tables 13-15**). WA32 harbored a R370W somatic mutation in *STAG2*, and a R370G variant was identified by transcriptome sequencing in LNCaP; mutations in *STAG2* have recently been identified as causing aneuploidy across cancer types[24]. WA56 and WA50 harbored a frameshifting indel and a probably damaging C4432R mutation, respectively, in *MLL3*, while MDA-PCa-2B harbored a N4685fs indel. Similarly, frameshifting indels were identified in *MLL5* in both WA57 and DU-145. *CNOT1*, which harbored mutations in three samples from our exome sequencing and one in Berger *et al*.'s dataset, also had a frame shifting indel in LAPC-4 (F128fs). A confirmed S548F somatic variant in *FAM123B* (*WTX*) was identified in T12 and a one bp indel was identified in LNCaP during Sanger sequencing validation efforts. Finally, T12 harbored a somatic 2 bp indel in *FOXA1* (S453fs), and transcriptome sequencing identified A340fs and P358fs frame shifting indels in DU-145 and LAPC-4, respectively.

Significantly mutated genes and pathways in prostate cancer

As described in the main text, we identified three genes as significantly mutated that did not have a previously described role in prostate cancer: *MLL2*, *CDK12*, and *OR5L1*. *MLL2* encodes a H3K4-specific histone methyltransferase[25] that is recurrently mutated in multiple cancers including diffuse large B-cell lymphoma[26], urothelial carcinoma[27] and medulloblastoma [28], and is a direct coactivator of the estrogen receptor[29]. *CDK12*, which encodes a transcription elongation-associated C-terminal repeat domain (CTD) kinase[30], was recently identified as one of nine significantly mutated genes in ovarian serous carcinoma[31], and silencing of *CDK12* has previously been shown to cause resistance to tamoxifen and estrogen deprivation in ER-dependent breast cancer models [32], suggesting a potential role in endocrine resistance in CRPC. *OR5L1* is an olfactory gene that exhibits a higher than average mutation rate as a result of its late replication, arguing against a role in cancer[33].

Previously, through single gene and focused panel resequencing studies of prostate cancer, mutations in candidates including *AR*, *TP53*, *CHEK2*, *KLF6*, *EPHB2*, *ZFHX3* (*ATBF1*), *NCOA2*, *PLXNB1, SPTA1,* and *SPOP* have been reported[2,16,34-40]. Through our exome sequencing, we identified recurrent mutations in several genes previously reported to be recurrently mutated in prostate cancer, including *AR, TP53,* and *ZFHX3* (each of which was significantly mutated), as well as *SPOP*; however no mutations were identified in *CHEK2*, *KLF6* or *NCOA2* (previously reported to be mutated in prostate cancer). Importantly, 61 (100%), 51 (84%) and 60 (98%) of the 61 samples had at least 70% of bases with sufficient coverage to call somatic mutations for *CHEK2*, KLF6 and *NCOA2*, respectively (**Supplementary Table 6),** suggesting that the lack of identified mutations is unlikely to be due to inadequate sequencing, and instead suggests that mutations in these genes may be rare, present in a small population of tumour cells or negatively selected for in CRPC.

We also identified 88 significantly mutated canonical pathways out of 880 considered (**see Methods & Supplementary Table 17**), including 49 with substantial contributions from the nine significantly mutated genes. For example, we identified the 'WNT signaling' KEGG pathway to be significantly mutated (57 somatic mutations, 38 samples, *q*-value = 1E-6). Half of these mutations occurred in genes other than *TP53* and *APC*, including three missense mutations in *CTNNB1* and a splice site mutation in *MYC*. Additionally, WA57 harbored concurrent nonsense mutation (W509*) and high-level copy loss in *SMAD4*, a gene which has recently been described as controlling lethal metastasis in CRPC[41].

We next utilized the matched somatic point mutation and exome copy number data to identify altered subnetworks in a large protein-protein interaction network using HOTNET[42]. This analysis identified 14 known KEGG pathways or protein complexes (**Supplementary Table 18**) as significantly mutated in CRPC, including a *PTEN* interaction network, which was altered in 81% of samples (**Supplementary Fig. 11**). While 48% of CRPC samples have *PTEN* mutations, 33% of CRPC samples have mutations in a protein that directly interacts with *PTEN*, indicating an even broader role for *PTEN* in prostate cancer pathogenesis and suggesting that mutational status of numerous genes may be required for stratification of therapies targeting the PTEN pathway. For example, we identified a probably damaging R215W mutation in WA57 of *MAGI2*, which encodes a PTEN interacting protein and was reported as recurrently deregulated by rearrangements by Berger *et al*[2]. Similarly, while most members of the *PTEN* interacting protein network were altered as a result of copy number changes, two genes exhibited recurrent somatic point mutations: *MAGI3* and *HDAC11* (each mutated in 4% of CRPC samples, **Supplementary Table 4**), suggesting potential roles in prostate cancer progression.

In addition to those significantly mutated genes and pathways, we also identified recurrent mutations in intriguing candidates and pathways. For example, we identified three CRPC samples with mutations in *FRY* (R100C in WA32, I1480T in WA56, S2510N in WA57), the homologue of the Drosophila gene *Furry* that encodes a microtubule binding protein required for precise chromosome alignment[43]. Mutations in *FRY* may promote chromosomal instability in CRPC or result from selection during treatment with docetaxel (a microtubule binding agent), a standard therapy for men with CRPC. In addition, we identified a *KRAS* G12V mutation in WA42 (ETS$^-$), consistent with our previous reports of rare RAF and RAS family aberrations in ETS$^-$ prostate cancers [44,45].

Identification of potential drivers by combined copy number and mutation analysis

To identify potential drivers, we considered genes with recurrent high-level gains or losses present in peaks of global copy number change, and compared results to mutated genes (**Supplementary Fig. 12**). Using this approach, *AR* had the maximum copy number sum (57), with 25 samples showing high-level copy number gain (all CRPC). Likewise *PTEN* had the minimum copy number sum (-64), with 25 samples showing high-level copy number loss. Both genes also harbored recurrent somatic mutations (**Figure 1**), supporting the validity of this approach.

Focal deletions and somatic mutations in *CHD1* define a novel ETS$^-$ prostate cancer subtype

In our cohort, three CRPCs (WA7, WA19 and WA10), all of which were ETS$^-$, showed focal high-level copy loss of *CHD1*. Additionally, we identified a single *CHD1* mutation (e28+1 splice site mutation) in WA27 (ETS$^-$). One additional ETS$^-$ localized prostate cancer (T93) and two ETS$^+$ CRPCs (WA12 and WA60), showed focal single copy loss involving *CHD1* (**Fig 2a**). Finally, by aCGH of our matched cohort, we confirmed focal deletions in WA7, WA19 and WA10, and identified three additional localized prostate cancers with focal, high copy number loss of *CHD1* (**Supplementary Fig. 13a**).

To further explore the association of *CHD1* focal deletion/mutation and ETS status in prostate cancer, we analyzed the association between *CHD1* and ETS status in prostate cancer using three prostate cancer aCGH studies (totaling 331 additional cancers) using Oncomine Powertools (http://powertools.oncomine.com). Importantly, each study showed a peak of copy number loss on 5q21, and in each study, all cancers with focal deletions of *CHD1* were ETS$^-$ (15 of 331 total, 4%) (**Supplementary Table 19**, **Supplementary Fig. 14**). For example, in the Taylor *et al*. study with 218 prostate cancers[16], we identified 9 with focal deletions of *CHD1*, all of which were ETS$^-$ (**Supplementary Figure 14a**). Thus, in total, we identified 25 of 450 prostate cancers as *CHD1*$^-$ in DNA based studies, 23 of which were ETS$^-$ (two sided Fisher's exact test, p=0.0002).

Finally, we explored the association of *CHD1* and ETS status by gene expression profiling using an additional 9 microarray studies (totaling 504 prostate cancers) available on Oncomine (www.oncomine.com). We identified 25 of 504 (5%) prostate cancers with outlier under-expression of *CHD1*, all of which were ETS$^-$ as assessed by lack of outlier over-expression of *ERG*, *ETV1*, *ETV4* or *ETV5* (**Supplementary Table 19**, p<0.0001, two sided Fisher's exact test), with the Glinsky *et al*. and Lapionte *et al*. datasets [46,47] shown in **Supplementary Figure 14d**. Thus, in total, across 13 DNA and RNA based studies, we identified 50 of 954 (5.2%) prostate cancers as being *CHD1*$^-$, 48 of which (96%) were ETS$^-$

($p<0.0001$, two sided Fisher's exact test, **Fig. 2b**). Of note, *CHD1*$^{-}$prostate cancers show some overlap with *SPINK1*$^{+}$ cancers, suggesting that these are not mutually exclusive classes of ETS^{-} tumours (**Supplementary Table 19**).

While our study was in preparation, Liu *et al*. and Huang *et al*. reported that *CHD1* is frequently deleted in prostate cancer (exclusively in ETS^{-} cancers in Liu *et al*.'s cohort) and has tumour suppressor properties, confirming our observations[48,49]. Additionally, we also identified other tumours with focal deletions involving other genes at 5q21, including *PJA2* (high-level copy loss in T65 and T53, and Y505C in WA53) suggesting the existence of other potential drivers at 5q21 (**Supplementary Fig. 13a,c&d**). Hence, in summary, our integrated analysis identifies deletion or mutation of *CHD1* as defining a novel subtype (*CHD1*$^{-}$) of ETS^{-} prostate cancer.

ETS2 is both deleted and mutated in prostate cancer

Our copy number profiling data generated here demonstrates that multiple *ERG* rearrangement positive CRPCs show focal deletions extending telomeric from *ERG* (**Supplementary Figure 15a**), consistent with previous observations in the LuCap35 xenograft and the NCI-H660 prostate cancer cell line (small cell *ERG*$^{+}$)[15,50]. Importantly, WA31 (*ERG*$^{+}$ through insertion) shows a focal, high copy number loss of *ETS2*, and our gene expression data demonstrates decreased *ETS2* expression in localized cancer and CRPC, with the lowest expression in WA31 (**Supplementary Fig. 15b**).

Hence, to investigate the functional consequences of *ETS2* disruption, we generated VCaP cells (a prostate cancer cell line that endogenously expresses *TMPRSS2:ERG*) that stably over-express wild type *ETS2* (VCaP *ETS2* wt), *ETS2* R437C (VCaP ETS2 R437C) or *LACZ* as control (VCaP *LACZ*) (**Supplementary Fig. 15c**). As shown in **Figure 2d** (left), VCaP *ETS2* wt showed decreased cell migration (in a Boyden chamber migration assay) compared to VCaP *LACZ* (0.6 fold, *p*=1.0E-5, two-sided t-test), while VCaP *ETS2* R437C showed increased migration compared to VCaP *LACZ* (1.2 fold *p*=0.03) and VCaP *ETS2* wt (2.0 fold, *p*=7.1E-7). Effects on cell invasion were even more pronounced (**Fig. 2d,** middle), with expression of *ETS2* wt significantly decreasing invasion compared to *LACZ* (0.4 fold, *p*=2.1E-5), while expression of *ETS2* R437C resulted in significantly increased invasion (1.7 fold, *p*=0.006). Lastly, as shown in **Figure 2d** (right), while VCaP *ETS2* R437C showed only minimally increased cell proliferation compared to VCaP *LACZ* (1.07 fold, *p*=0.004), VCaP *ETS2* wt showed markedly decreased cell proliferation compared to VCaP *LACZ* (0.65 fold, *p*=8.2E-9). As ETS genes involved in gene fusions have been shown to dramatically impact cell invasion[51,52], *ETS2* may directly compete with other ETS transcription factors for binding to target genes and further investigation will be needed to clarify the role of *ETS2* in ETS$^+$ and ETS$^-$ prostate cancers. These results are consistent with distinct ETS genes having oncogenic and potential tumour suppressive roles in prostate cancer[53].

Identification of chromatin/histone modifying genes mutated in CRPC that interact with the androgen receptor

In addition to *MLL2* and *CHD1*, our integrated analysis identified mutations and copy number aberrations in multiple other genes involved in chromatin/histone modification (**Fig. 1**)., including *MLL2*, which was the 7th ranked significantly mutated gene in our data set. The MLL genes (*MLL*, *MLL2* and others) encode histone methyltransferases that function in multi-protein complexes that mediate H3K4 methylation required for epigenetic transcriptional activation[25]. In addition to *MLL2*, we identified a frame preserving indel in *MLL* (Q1815fp in WA28) and deleterious mutations in *MLL3* (R1742fs in WA18 and F4463fs in WA56) and *MLL5* (E1397fs in WA57). In total, 10 of 58 (17.2%) of all samples harbored mutations in an MLL gene. Additionally, while the MLL proteins possess catalytic activity through a SET domain, MLL and MLL2 function as part of a multi-protein complex that includes ASH2L, RBBP5, WDR5 and MEN1 (menin)—all of which harbor varying levels of aberration in CRPC (see below and **Fig. 3**).

Additional deregulated epigenetic modifiers identified in our analysis included the polycomb group gene *ASXL2* which was the 17th ranked significantly mutated gene in our data set (p=3.4E-4) and was mutated in 4 samples, with 3 samples harboring nonsense mutations (Y1163* in WA31, Q1104* in WA56 and Q172* in WA23) (**Figure 1**). We also identified single samples with nonsense mutations in *ASXL1* (P749fs in WA52) and *ASXL3* (L2240V and R2248* in WA22). Interestingly, *ASXL1* is recurrently mutated in myeloid disorders, predominantly through frameshift mutations in the last exon[54], the same exon affected by the P749fs mutation observed in WA52. Similarly, although *UTX* (*KDM6A*), which encodes a histone H3K27 demethylase that complexes with MLL3[25], is located in a broad region of copy number gain on chr X, it is located at a local copy number minimum, and two samples (WA28 and WA40) show focal high copy loss (**Fig. 1**). *UTX* has been shown to be mutated in a number of cancers including renal carcinoma and urothelial carcinoma[25,55,56]. This was of interest, as we previously showed that the histone H3K27 methyltransferase *EZH2* is overexpressed in the majority of CRPCs[57].

Identification and characterization of recurrent mutations in *FOXA1*

As described in the main text, we identified a somatic 2 bp insertion in *FOXA1* (S453fs) in the localized prostate cancer sample T12, and 340fs and P358fs indels in DU-145 and LAPC-4 (identified by transcriptome sequencing), respectively. Screening 101 localized and 46 CRPCs (including foci from all CRPC samples subjected to exome sequencing) identified somatic mutations of *FOXA1* in 4 localized prostate cancers and 1 CRPC (total 5 of 147, 3.4%) prostate cancers. Mutations identified in localized cancers included the S453 insertion identified in T12 in the exome sequencing, G87R in T68, L388M in T70 and L455M in T18086, while we identified a F400I mutation in WA40, a small cell CRPC, which was from a different metastatic focus from that used for exome sequencing.

Previously, exploring the role of *FOXA1* in androgen signaling, Wang *et al*. recently reported that down-regulation of *FOXA1* (by siRNA) in LNCaP cells triggers dramatic reprogramming of the hormonal response and enhances entrance to S phase, and decreased expression of *FOXA1* is associated with poor outcome in CRPC[58]. In contrast, Gerhardt *et al*. reported that *FOXA1* is over-expressed in CRPC and siRNA knockdown of *FOXA1* results in decreased growth of LNCaP cells[59].

Thus, we hypothesized that mutations in *FOXA1* may affect proliferation and/or AR signaling, and generated stable LNCaP cells expressing empty vector (LNCaP vector), wild type *FOXA1* (*FOXA1* wt) and the five *FOXA1* mutants observed in our clinical samples as N-terminal FLAG fusions. Western blot and QPCR analyses confirmed equivalent levels of expression of each *FOXA1* construct (**Fig. 4b** and **Supplementary Fig. 17a**). The S453fs insertion allele encodes a protein with a predicted molecular weight 49 kDa, similar to wild type FOXA1 (49.2kDa).

In LNCaP cells grown in the presence of 10nM DHT, all *FOXA1* mutants, as well as *FOXA1* wt, showed significantly increased cell proliferation compared to LNCaP vector (p=0.006 for *FOXA1* F400I, p<0.001 for all comparisons to LNCaP vector), while only *FOXA1* L388M showed significantly increased growth compared to FOXA1 wt (p=0.005). Expression of *FOXA1* wt or mutants had no significant effect on LNCaP proliferation in the absence of androgen (**Supplementary Fig. 17b**).

Given the role of *FOXA1* as a cofactor for AR signaling, and the reported ability of *FOXA1* to repress portions of the AR program as well as enhance AR transcription[60,61], we performed gene expression profiling from LNCaP vector, *FOXA1* wt and *FOXA1* mutant cells stimulated with vehicle or 10nM DHT for 48 hours. Focusing on the AR mediated program, we identified 352 probes showing ≥ 2 fold over-expression and 262 probes showing ≤ -2 fold under-expression upon DHT stimulation in LNCaP vector cells (**Fig. 4d**). We observed generalized repression of AR signaling in LNCaP *FOXA1* wt and *FOXA1* mutant cells, with 81% of these DHT stimulated probes in LNCaP vector cells showing <1.5 (for over-expressed probes) or >-1.5 fold change (for under-expressed probes) in LNCaP *FOXA1* wt cells. In contrast, only 6% of probes showed enhanced expression in LNCaP FOXA1 wt cells (>2 or <-2 fold change). Similar effects were observed in *FOXA1* mutant cell lines with an average of 59% repressed probes (range 43-73%) vs. 23% enhanced probes (range 5-39%). Of note, the stimulation of *KLK2*, *KLK3* (*PSA*) and *NKX3-1* were not significantly repressed by *FOXA1* wt or *FOXA1* mutants.

Based on the effects of *FOXA1* wt and *FOXA1* mutants on proliferation, we also generated LNCaP cells stably expressing 3xHA-N-terminally tagged *FOXA1* wt, *FOXA1* S453fs, or *LACZ* (as control) through a different lentivirus construct. These cells were used for soft agar colony forming assays, and as shown in **Figure 17c**, both *FOXA1* wt and *FOXA1* S453fs formed significantly more colonies than *LACZ* cells ($p<0.05$ for each) in the presence of 1nM of the synthetic androgen R1881.Finally, parental LNCaP, LNCaP *FOXA1* wt and LNCaP *FOXA1* S453fs cells were used in xenograft experiments. As shown in **Figure 4e**, by 20 days, both LNCaP *FOXA1* wt and *FOXA1* S453fs cells formed significantly larger tumours than parental LNCaP cells. Taken together, we identified mutations in the AR collaborating factor *FOXA1*, which occur in both untreated localized prostate cancer and CRPC, and promote cell growth and repress AR signaling, with similar effects to over-expression of wild type *FOXA1*. Our results are consistent with Gerhardt *et al*., who showed that siRNA knockdown of *FOXA1* resulted in decreased growth of LNCaP cells[59], although additional experiments will be needed to fully characterize the effects of *FOXA1* mutations on androgen signaling and prostate cancer growth and development.

Supplementary Discussion

Future studies are needed to understand the temporal development of aberrations identified in this study, to delineate their occurrence during prostate cancer development and progression (including therapy and resistance mechanisms). Importantly, our results provide insight into the ability of sequencing based studies to identify potential therapeutic targets in patients with advanced cancers[62]. For example, 25 of 48 (52%) of patients with CRPC in our cohort have high-level *AR* amplification and continued activation of AR signaling is common in CRPC (e.g., 17 of 21 [81%] CRPCs with *TMPRSS2:ERG* fusions still have outlier over-expression of the androgen-regulated transcript), suggesting therapeutic benefit with novel anti-androgens. Likewise, 24 of 48 (50%) of CRPC harbored high-level copy number aberrations or mutation of *PTEN* and/or *PI3KCA*, and aberrations in the PTEN interacting network occurred in 83% of CRPCs, supporting the continued investigation of PI3K inhibitors in this population[63]. Similarly, identification of rare lesions such as high-level somatic copy loss of *BRCA2* (**Supplementary Fig. 4**), high-level copy gain of *CDK4* (**Supplementary Fig. 7b**), or potentially activating mutations of *RET* (R873W in WA47) may provide rationale for targeted clinical trials or treatment with novel agents.

Methods

Tissue samples and cell lines

Prostate tissues were from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program[64], both of which are part of the University of Michigan Prostate Cancer Specialized Program of Research Excellence (SPORE) Tissue Core. All samples were collected with informed consent of the patients and previous University of Michigan Institutional Review Board approval.

All CPRC specimens (WA2-WA60) and paired normal tissues were obtained at rapid autopsy from men who died of lethal castrate resistant metastatic disease. Our rapid autopsy protocol has been described in detail previously[64]. Briefly, at the time of autopsy, portions of all cancerous tissue grossly identified, as well as uninvolved organs, were processed by routine formalin fixation and paraffin embedding (FFPE). Corresponding samples were also snap frozen in OCT or chunks. Hematoxylin and eosin (H&E) stained FFPE and frozen sections were reviewed by study pathologists (R.M., L.P.K. and SAT) to identify blocks with highest tumour content (tumour) or lack of cancer or high grade prostatic intraepithelial neoplasia (normal). For each frozen block used, a level was taken for H&E staining, consecutive 3 x 10um sections were cut for DNA isolation, a level was taken for H&E staining, consecutive 3 x 10um sections were cut for RNA isolation, and a final level was taken for H&E staining. All H&E stained levels were reviewed to confirm tumour/normal content before DNA/RNA isolation.

All prostatectomy specimens (N1-N29 and T1-T97) were obtained from treatment naïve men at the time of prostatectomy, where fresh tissue was obtained as part of our standardized procurement protocol. Intervening sections not embedded for routine histological assessment from prostatectomies were quartered and snap frozen in OCT. Evaluation of frozen sections to identify blocks with the highest tumour content (tumour) or lack of cancer and unremarkable morphology (normal) and RNA/DNA isolation were performed as described above.

The immortalized prostate cancer cell lines 22Rv1, C4-2B, CWR22, DU-145, LAPC-4, LNCaP, MDA-PCa-2B, NCI-H660, PC3, VCaP and WPE1-NB26 (**Supplementary Table 11**) were obtained from the American Type Culture Collection (Manassas, VA). PC3, DU-145, LNCaP, 22Rv1, and CRW22 cells were grown in RPMI 1640 (Invitrogen) and supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. VCaP cells were grown in DMEM (Invitrogen) and supplemented with 10% fetal bovine serum (FBS) with 1% penicillin-streptomycin. NCI-H660 cells were grown in RPMI 1640 supplemented with 0.005 mg/ml insulin, 0.01 mg/ml transferrin, 30 nM sodium selenite, 10 nM hydrocortisone, 10 nM beta-estradiol, 5% FBS and an extra 2 mM of L-glutamine (for a final concentration of 4 mM). MDA-PCa-2B cells were grown in F-12K medium (Invitrogen) supplemented with 20% FBS, 25 ng/ml cholera toxin, 10ng/ml EGF, 0.005 mM phosphoethanolamine, 100 pg/ml hydrocortisone, 45 nM selenious acid, and 0.005 mg/ml insulin. LAPC-4 cells were grown in Iscove's media (Invitrogen) supplemented with 10% FBS and 1 nM R1881. C4-2B cells were grown in 80% DMEM supplemented with 20% F12, 5% FBS, 3 g/L NaCo3, 5ug/ml insulin, 13.6 pg/ml triiodothyonine, 5ug/ml transferrin, 0.25ug/ml biotin, and 25 ug/ml adenine. WPE1-NB26 cells were grown in Keratinocyte Serum Free Medium (Invitrogen) and supplemented with bovine pituitary extract (BPE, 0.05mg/ml) and human recombinant epidermal growth factor (EGF, 5ng/ml). Androgen treated LNCaP and VCaP cell line samples were also generated for transcriptome analysis, using cells grown in androgen-depleted media lacking phenol red and supplemented with 10% charcoal-stripped serum and 1% penicillin-streptomycin. After 48 hours, cells were treated with 5nM methyltrienolone (R1881, NEN Life Science Products) or an equivalent volume of ethanol. Cells were harvested for RNA isolation at 6, 24, and 48 hours post-treatment.

High molecular weight genomic DNA (gDNA) Isolation

gDNA from frozen tissue specimens was isolated using the Qiagen DNeasy Blood & Tissue Kit according to the manufacturer's instructions. Briefly, cell or tissue lysates were incubated at 56^{o} C in the presence of proteinase K and SDS, purified on silica membrane-based mini-columns, and eluted in buffer AE (10 mM Tris-HCl, 0.5 mM EDTA pH 9.0).

Generation of Exome-capture libraries

Exome libraries of matched pairs of tumour / normal genomic DNAs (**Supplementary Table 1**) were generated using the Illumina Paired-End Genomic DNA Sample Prep Kit, following the manufacturers' instructions. 3 ug of each genomic DNA was sheared using a Covaris S2 to a peak target size of 250 bp. Fragmented DNA was concentrated using AMPure XP beads (Beckman Coulter), and DNA ends were repaired using T4 DNA polymerase, Klenow polymerase, and T4 polynucleotide kinase. 3' A-tailing with exo-minus Klenow polymerase was followed by ligation of Illumina paired-end adapters to the genomic DNA fragments. The adapter-ligated libraries were electrophoresed on 3% Nusieve 3:1 (Lonza) agarose gels and fragments between 300 to 350 bp were recovered using QIAEX II gel extraction reagents (Qiagen). Recovered DNA was then amplified using Illumina PE1.0 and PE2.0 primers for 9 cycles. The amplified libraries were purified using AMPure XP beads and the DNA concentration was determined using a Nanodrop spectrophotometer. 1 mg of the libraries were hybridized to the Agilent biotinylated SureSelect Capture Library at 65°C for 72 hr or to the

Roche EZ Exome capture library at 47°C for 72 hr following the manuufacturer's protocol. The targeted exon fragments were captured on Dynal M-280 streptavidin beads (Invitrogen), washed, eluted, and enriched by amplification with the Illumina PE1.0 / PE2.0 primers for 8 additional cycles. After purification of the PCR products with AMPure XP beads, the quality and quantity of the resulting exome libraries were analyzed using an Agilent Bioanalyzer.

Somatic point mutation identification by exome capture sequencing

All captured DNA libraries were sequenced with the Illumina GAII Genome Analyzer or the Illumina HiSeq in paired end mode, yielding 80 base pairs from the final library fragments. The reads that passed the chastity filter of Illumina BaseCall software were used for subsequent analysis. Next, matepairs were pooled and then mapped as single reads to the reference human genome (NCBI build 36.1, hg18), excluding unordered sequence and alternate haplotypes, using Bowtie[65], keeping unique best hits, and allowing up to two mismatched bases per read. Reads in the tumour that mapped to another location in the genome with three mismatches were excluded from further consideration. Likely PCR duplicates were removed by removing reads that have the same match interval on the genomic sequence. Individual basecalls with Phred quality less than Q20 were excluded from further consideration.

A mismatched base (SNV) was identified as a somatic mutation only when 1) it had six reads of support (this cut-off was selected based on Sanger validation rates in T12 [**Supplementary Fig. 1**]), 2) it was in at least 10% of the coverage at that position in the tumour, 3) it was observed on both strands, 4) there was 8X coverage in the matched normal, and 5) it did not occur in the matched normal sample in more than two reads and 2% of the coverage (to ensure that somatic variants are not filtered out due to tumour contamination in the normal, we retained variants present in 2-4% of the coverage in the matched normal, if they were in at least 20% of the coverage in the tumour). SNVs were excluded from further consideration as somatic mutations if 1) they did not fall within 50 bases of a target region, 2) they occurred in any two matched normal samples in at least two reads and 2% of the coverage, or 3) they occurred in another tumour and its matched normal sample in two reads and 4% of the coverage.

Identification of coding indels in exome capture data

The methodology for identifying indels in exome capture data was adapted from [66] with minor modifications. Reads for which Bowtie was unsuccessful in identifying an ungapped alignment were converted to fasta format and mapped to the target regions, padded by 200 bases on either side, with cross_match (v0.990329, http://www.phrap.org), using parameters –gap_ext -1 –bandwidth 10 –minmatch 20 –maxmatch 24. Output options were –tags –discrep_lists –alignments. Alignments with an indel were then filtered for those that: 1) had a score at least 40 more than the next best alignment, 2) mapped at least 75 bases of the read, and 3) had two or fewer substitutions in addition to the indel. Reads from filtered alignments that mapped to the negative strand were then reverse-complemented and, together with the rest of the filtered reads, remapped with cross_match using the same parameters (to reduce ambiguity in called indel positions due to different read orientations). After the second mapping, alignments were re-filtered using criteria 1-3. Reads that had redundant start sites were removed as likely PCR duplicates, after which the number of reads mapping to either the reference or the non-reference allele was counted for each. An indel was called if there were at least six non-reference allele reads making up at least 10% of all reads at that genomic position. Indels were reported with respect to genomic coordinates. For insertions, the position reported is the last base before the insertion. For deletions, the position reported is the first deleted base. Indel somatic mutation candidates were excluded from further consideration if 1) they did not occur on both strands, 2) they did not fall within 50 bases of a target region, 3) there wasn't 8X coverage in the matched normal at that position, 4) they occurred in the matched normal sample in more than 2 reads and 4% of the coverage, 5) they occurred in any two matched normal samples, or 6) they occurred in any single matched normal sample in more than 2 reads.

Annotation

We annotated the resulting somatic mutations using CCDS transcripts wherever possible. If no CCDS transcript was available, we used the coding regions of RefSeq transcripts. HUGO gene names were used. The impact of coding nonsynonymous amino acid substitutions on the structure and function of a protein was assessed using PolyPhen-2[67]. We also assessed whether the somatic variant was previously reported in dbSNP or COSMIC v56[68].

Calculation of somatic mutation rates

The somatic mutation rate was calculated as described[2]. We identified a base as "covered", if there was at least 14X total coverage after PCR duplicate removal in the tumour and 8X total coverage after PCR duplicate removal in the matched normal sample. We considered only mutations called at covered annotated targeted positions; the total number of covered annotated targeted positions ranged from 22.3-30.4 Mb per sample, with 74.4 - 94.3% of annotated targeted positions covered per sample. Because this calculation does not take into consideration the sensitivity of the somatic mutation calling method or tumour purity, it may underestimate the actual mutation rate for the sample.

Tumour Content Estimation

Tumour content was estimated for each cancer sample by fitting a binomial mixture model with two components to the set of most likely SNV candidates on 2-copy genomic regions. The set of candidates used for estimation consisted of coding variants that (1) exhibited at least >=3 variant fragments in the cancer sample, (2) exhibited zero variant fragments in the matched benign sample with at least 16 fragments of coverage, (3) were not present in dbSNP, (4) were within a targeted exon or within 100 base pairs of a targeted exon, (5) were not in homopolymer runs of four or more bases, and (6) exhibited no evidence of amplification or deletion.

In order to filter out regions of possible amplification or deletion, we used exon coverage ratios to infer copy number changes, following the approach of [14]. Resulting SNV candidates were not used for estimation of tumour content if the segmented log-ratio exceeded 0.25 in absolute value. Candidates on the X and Y chromosomes were also eliminated because they were unlikely to exist in 2-copy genomic regions.

Using this set of candidates, we fit a binomial mixture model with two components using the R package flexmix, version 2.2-8[69]. One component consisted of SNV candidates with very low variant fractions, presumably resulting from recurrent sequencing errors and other artifacts. The other component, consisting of the likely set of true SNVs, was informative of tumour content in the cancer sample. Specifically, under the assumption that most or all of the observed SNV candidates in this component are heterozygous SNVs, we expect the estimated binomial proportion of this component to represent one-half of the proportion of tumour cells in the sample. Thus, the estimated binomial proportion as obtained from the mixture model was doubled to obtain an estimate of tumour content in each sample.

Determination of significantly mutated genes and pathways

The determination of significantly mutated genes and pathways was done as described[2,70] using methodology based on that of Getz *et al*.[71] and Ding *et al*.[72]. Before doing the calculations, we selected one of the three samples derived from distinct metastatic sites from the same individual (WA43) for inclusion in the sample set in order to ensure that the requirement of independence was met for the set of considered mutations. We selected WA43-44, because it contained all of the recurrent somatic mutations that occurred in WA43-27 or WA43-71, along with additional recurrent mutations not contained in the other two. We also excluded hyper-mutated sample WA16. In this approach, significantly mutated genes are identified based on the observed number of mutations for each sequence context-based mutation class (CpG, other C:G, A:T, and indels), the sample-specific and class-specific background mutation rates, and the number of covered bases per gene (**Supplementary Table 6)**. Before calculating the background mutation rate, we excluded genes that have been reported in the literature as having recurrent somatic mutations in prostate cancer: *AR*, *TP53*, *CHEK2*, *KLF6*, *EPHB2*, *ZFHX3*, *NCOA2*, *PLXNB1*, *SPTA1*, and *SPOP* [2,16,34-40]. The resulting background mutation rate for localized prostate cancer samples was 5.03/MB for CpG, 0.71/Mb for other C:G, 0.39/Mb A:T and 0.10/Mb indels. The resulting background mutation rate for metastatic prostate cancer samples was 8.45/MB for CpG, 1.80/Mb for other C:G, 0.95/Mb A:T and 0.21/Mb indels. For each gene, we calculated the probability of obtaining the observed set of mutations (or a more extreme one) given the observed background mutation rates. *P*-values are converted to *q*-values using the

Benjamini-Hochberg procedure for controlling False Discovery Rate (FDR).

We repeated this analysis to consider significantly mutated pathways, considering a list of 880 gene sets corresponding to the set of canonical pathways used in Gene Set Enrichment Analysis (GSEA). For this analysis, we tabulated the number of mutations and the number of covered bases in all component genes of each gene set and the total number of covered bases in the set. As in the single-gene analysis, the mutation counts were broken down into the context-based mutation classes (CpG, other C:G, and A:T), and then the *P*-value and subsequent *q*-value were calculated.

Sanger sequencing to validate somatic point mutations and indels

Various genomic locations nominated for somatic point mutations and indels were amplified from whole genome amplified DNA[73] from corresponding matched normal-tumour tissue pairs or cell lines. Briefly, fifty ng of input genomic DNA was subjected to fragmentation, library preparation and amplification steps using Genomeplex-Complete Whole Genome Amplification Kit (Sigma-Aldrich) according to manufacturer's instructions. The final whole genome amplified DNA was purified by AMPure XP beads (Beckman-Coulter) and quantified by a Nanodrop spectrophotometer (Thermo Scientific). Fifty ng of DNA were used as template in PCR amplifications with HotStar Taq DNA polymerase (Qiagen) with the suggested initial denaturation and cycling conditions. Primer sequences were as described[9] and are based on human hg18, March 2006 assembly. Primers for *FOXA1* can be found in **Supplementary Table 20.** The PCR products were subjected to Sanger sequencing by the University of Michigan DNA Sequencing Core after treatment with ExoSAP-IT (GE Healthcare) and sequences were analyzed using MacVecotr software (MacVector).

Exome Copy Number Analysis

Copy number aberrations were quantified and reported for each gene as the segmented normalized log2-transformed exon coverage ratios between each tumour sample and its matched normal (**Supplementary Table 7**) as described[14].

We used sample-specific cutoffs, based on estimated tumour content, to define regions of gain and loss, as follows. For a sample with tumour percentage P, genomic regions with N copies in cancer cells and 2 copies in normal cells would be predicted to give log-ratios centered at log2(N*P + 2*(1-P))-1. For each sample, using its estimated tumour content we computed the predicted locations of these N-copy peaks in the distribution of log-ratios, and chose cutoffs to fall between these predicted peaks.

To define high-level gains (i.e., greater than 3 copies), we computed the weighted average of the 3-copy and 4-copy predicted peaks with weights 0.25 and 0.75, respectively. Similarly, to define low-level gains (i.e., greater than 2 copies), we computed the weighted average of the 2-copy and 3-copy predicted peaks, using the same weights. These weighted averages were used as cut-offs to define high-level gain and low-level gain, respectively. Next, the negatives of the cutoffs for high-level gain and low-level gain were used as the cut-offs for high-level loss (two-copy loss) and low-level loss (single-copy loss), respectively. Histograms of the distributions of segmented log2 copy number ratios were then examined (**Supplementary Fig. 7**) and cutoffs adjusted manually in cases in which this algorithmic approach appeared to misclassify large numbers of genomic regions (due to lower tumour content, multiple clones, severe aneuploidy, etc. All cutoffs with estimated tumour percentages are given in **Supplementary Table 8**. The resulting copy number alterations were reported for all sixty one prostate cancer tumours with +2 representing high-level (> 1copy) gain, +1 representing 1 copy gain, 0 representing no change, -1 representing 1 copy loss, and -2 representing high-level (>1 copy) loss (**Supplementary Table 9**). To identify potential drivers, we summed all called copy number alterations across all samples and identified genes with the maximum number of high level gains or losses occurring in peaks summed copy number gain or loss, respectively. For all analyses and visualizations, WA43-27 and WA43-71 were excluded.

Identification of Significantly Mutated Protein-Protein Interaction Subnetworks

We used HotNet[42] to find subnetworks of a large protein-protein interaction network containing a significant number of mutations and copy number alterations (CNAs). The input to HotNet is a dataset of matched somatic mutations and copy number alterations for a set of tumour samples. The output of HotNet is a list of subnetworks, each containing at least *n* genes. HotNet employs a two-stage statistical test to assess the significance of the output. In the first stage the *p*-value for the number of subnetworks in the list is computed. In the second stage the false discovery rate (FDR) of the *list* of subnetworks is estimated. At the end, the significance of each individual subnetwork in the list is assessed by comparison to known pathways and protein complexes.

We analyzed the combined somatic mutations (**Supplementary Table 4**) and CNAs (generated from exome data [**Supplementary Table 9**]), considering only high-level (> 1 copy) gains and two-copy losses, for the 47 metastatic samples (we did not include hyper-mutated sample WA16 and only included one of the three metastatic sites from the same individual, WA43-44). We removed CNAs for which the sign of aberration was not consistent in at least 90% of the altered samples. We used the interaction network derived from the Human Protein

Reference Database (HPRD)[74]. For the statistical test, we generated random aberrations as follows. We simulated mutations using the estimated background mutation rate (1.97×10^{-6}). We simulated CNAs from the observed distribution of CNA lengths, permuting their positions. In this way artifacts resulting from functionally related genes that are both neighbors on the network and close enough on the genome (and thus affected by the same CNA) are minimized. We also discarded subnetworks reported by HotNet that contain 3 or more genes in the same CNA in one or more samples. Moreover, for subnetworks with two genes g1, g2 in the same CNA in 1 or more samples, we removed the genes that were not reported when alterations in either g1 or g2 are removed.

Using the approach above, HotNet identifies 28 candidate subnetworks containing at least 10 genes ($p < 0.01$) with FDR = 0.32. A total of 24 subnetworks remained after CNA filtering. We then compared those 24 subnetworks with known pathways in the KEGG database and with protein complexes from PINdb[75]. Of the 24 subnetworks, 14 had statistically significant ($p<0.05$ after Bonferroni correction) overlap with at least one KEGG pathway or protein complex (**Supplementary Table 18**).

RNA isolation and cDNA synthesis

Total RNA was isolated from frozen prostate tissue samples (for gene expression analysis and qPCR) and cell lines (for transcriptome sequencing, qPCR/expression profiling from cell lines) using Trizol (or Qizol [Qiagen]) and an RNeasy Kit (Invitrogen) with DNase I digestion according to the manufacturer’s instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies). cDNA was synthesized from total RNA using

Superscript III (Invitrogen) and random primers (Invitrogen).

Transcriptome library preparation and sequencing

Next generation RNA sequencing was performed on 11 prostate cell lines according to Illumina's protocol using 2ug of total RNA. RNA integrity was measured using an Agilent 2100 Bioanalyzer, and only samples with a RIN score >7.0 were advanced for library generation. PolyA+ RNA was selected for using Sera-Mag oligo(dT) beads (Thermo Scientific) and fragmented with the Ambion Fragmentation Reagents kit (Ambion, Austin, TX). cDNA synthesis, end-repair, A-base addition, and ligation of the Illumina PCR adaptors (single read or paired-end where appropriate) were performed according to Illumina's protocol. Libraries were then size-selected for 250-300 bp cDNA fragments on a 3.5% agarose gel and PCR-amplified using Phusion DNA polymerase (Finnzymes) for 15 –18 PCR cycles. PCR products were then purified on a 2% agarose gel and gel-extracted. Library quality was credentialed by assaying each library on an Agilent 2100 Bioanalyzer forproduct size and concentration. Libraries were sequenced as 36-45mers on an Illumina Genome Analyzer I or Genome Analyzer II flowcell according to Illumina's protocol. All single read samples were sequenced on a Genome Analyzer I, and all paired-end samples were sequenced on a Genome Analyzer II.

Somatic point mutation identification in transcriptome sequence data

Transcriptome short reads were trimmed to remove the first two bases and as many bases as necessary to ensure the read length was less than 40bp. Trimmed short read sequences were mapped to the reference human genome (NCBI build 36.1, hg18), excluding unordered sequence and alternate haplotypes, and the 2008 Illumina splice junction set using Bowtie[65] in single read mode keeping unique best hits and allowing up to two mismatched bases. Matepairs from paired end runs were pooled and treated as single reads. Likely PCR duplicates were removed by removing reads that have the same match interval on the genomic sequence or an exon junction. Individual basecalls with Phred quality less than Q20 were excluded from further consideration. A mismatched base (SNV) was identified as a candidate somatic mutation when it had three reads of support and was in at least 10% of the coverage at that position in the tumour. Less stringent criteria were applied for nominating candidate somatic mutations in the transcriptome as compared to the exome capture data, since only variants in the transcriptome recurrent to known somatic mutations were further considered (see below). SNVs were excluded from further consideration as recurrent somatic mutations if 1) they occurred in any two matched normal exomes in at least two reads and 2% of the coverage, or 2) they occurred in another tumour exome and its matched normal exome in two reads and 4% of the coverage.

Identification of coding indels in transcriptome data

The methodology for identifying indels in transcriptome data was adapted from [66]. Reads for which Bowtie was unsuccessful in identifying an ungapped alignment were converted to fasta format and mapped to the set of full-length CCDS transcripts, padded by 32 genomic bases on either side, with cross_match (v0.990329, http://www.phrap.org), using parameters –gap_ext -1 –bandwidth 10 –minscore 24 –minmatch 16 –maxmatch 24. Output options were –tags –discrep_lists –alignments. Alignments with an indel were then filtered for those that: 1) had a score at least 20 more than the next best alignment; and 2) had two or fewer substitutions in addition to the indel. Reads from filtered alignments that mapped to the negative strand were then reverse-complemented and, together with the rest of the filtered reads, remapped with cross_match using the same parameters (to reduce ambiguity in called indel positions due to different read orientations). After the second mapping, alignments were re-filtered using criteria 1) and 2). Reads that had redundant start sites were removed as likely PCR duplicates, after which the number of reads mapping to either the reference or the non-reference allele were counted for each. An indel was called if there were at least four non-reference allele reads making up at least 10% of all reads at that transcript position. Indels were reported with respect to genomic coordinates. For insertions, the position reported is the last base before the insertion. For deletions, the position reported is the first deleted base. Indel somatic mutation candidates were excluded from further consideration if they were present in dbSNP132, or if they occurred in a single read in any two matched normal exome samples or in a single matched normal exome sample with two or more reads. Identified indel variants are given in **Supplementary Table 13**.

Identification of transcriptome somatic SNVs recurrent to known somatic variants

The somatic mutations identified in the exome data in this study (excluding the eight that did not validate by Sanger sequencing) were combined with the confirmed somatic variants in COSMIC v56 to yield a comprehensive somatic mutation dataset. A transcriptome SNV was considered recurrent to a known somatic variant, if it resulted in the same nucleotide change, amino acid change, or if it disrupted the same amino acid. Identified variants recurrent to our exome data are given in **Supplementary Table 14**, and those recurrent to somatic variants in COSMIC are given in **Supplementary Table 15**.

Array comparative genomic hybridization (aCGH)

aCGH of 28 benign prostate tissues, 59 localized prostate cancers (including 56 not subjected to exome sequencing) and 35 CRPCs (including 4 not subjected to exome sequencing, see **Supplementary Table 10**) was performed using gDNA on Agilent's 105K or 244K aCGH microarrays (Human Genome CGH 105K or 244K Oligo Microarray) using Agilent's standard Direct Method protocol and Wash Procedure B. Briefly, 1.5 - 3 ug of gDNA from prostate specimens (isolated as above) was restriction digested with AluI and RsaI, labeled with Cy-5 (test channel), purified using Microcon YM-30 columns and hybridized with an equal amount of Cy-3 (reference channel) labeled Human Male Genomic DNA (Promega) for 40 hours at 65^{o} C. Post-hybridization wash was performed with acetonitrile wash and Agilent Stabilization and Drying Solution wash according to the manufacturer's instructions. Scanning was performed on an Agilent scanner Model G2505B (5 micron scan with software v7.0), and data was extracted using Agilent Feature Extraction software v9.5 using protocol CGH-v4_95_Feb07.

For data analysis, probes on all arrays were limited to those on the 105K array. Log(2) ratios for each probe were determined as rProcessedSignal/gProcessedSignal. To remove copy number variants, all probes with log(2) values >1 or <-1 in any of the 28 benign prostate samples were excluded. The final dataset (consisting of localized prostate cancer and castrate resistant metastatic samples) was uploaded into a custom instance of Oncomine (www.oncomine.com) for automated copy number analysis. In Oncomine, circular binary segmentation was performed on the dataset using the DNACopy package (v1.18) available via the Bioconductor package. Agilent Probe IDs are mapped to segments and reporter values are used to generate segment values (mean of reporters). Resulting segments are mapped to hg18 (NCBI 36.1) RefSeq coordinates (UCSC refGene) as provided by UCSC (UCSC refGene, July 2009, hg18, NCBI 36.1, March

2006) and segment values are assigned to each gene. Copy number profiles were visualized using Oncomine Power Tools.

Gene Expression Microarray Analysis

Gene expression microarray analysis of the same prostate tissue samples subjected to aCGH (**Supplementary Table 10**) was performed using Agilent Whole Human 44k element arrays (1x44k or 4x44k format) as described[76]. RNA from indicated prostate samples were labeled with Cy-5 (test channel) and hybridized against Cy-3 (reference channel) labeled pooled benign prostate RNA (Clontech). Arrays were scanned using an Agilent Model G2505B scanner, and data was extracted using Agilent Feature Extraction software. Control probes were removed from all arrays and the LogRatio for all probes, which were used for subsequent analysis, were converted to log(2). Although the 1x44k and 4x44k arrays have the same probes, the 4x44k arrays have 10 replicates of some probes. Thus, to generate a final data set, the median value of replicated probes was used for 4x44k arrays. The final data set (including benign prostate, localized prostate cancer and CRPC) was uploaded into a custom instance of Oncomine for automated analysis. In Oncomine, the dataset was median centered (per array) prior to indicated analyses. Copy number and gene expression data is also available from GEO (GSE35988).

ETS/RAF/CHD1 status

ETS/RAF gene fusion status for all samples was assigned based on expression of *TMPRSS2:ERG* by qPCR[77], outlier expression and/or rearrangement of *ERG*, *ETV1*, *ETV4* or *ETV5* by FISH[11,76-78], *RAF* family member rearrangement by transcriptome sequencing and subsequent qPCR and FISH validation[44], presence of deletion between *TMPRSS2* and *ERG* by aCGH, or ERG protein expression by immunohistochemistry[79]. *CHD1*$^-$ status was determined by examination of exome copy number profiles (or aCGH profiles) for all samples, and those with focal deletions involving *CHD1* (without a larger focal deletion within 10 MB) or nonsynonymous mutations in *CHD1* were considered *CHD1*$^-$. Assessment of ETS status in aCGH profiling studies in Oncomine was performed as follows, and samples in each study with focal deletions (log2 ratio <-0.23 or -0.24) or high level focal deletions arising in background deletions were considered *CHD1*$^-$. For the Demichelis *et al*. study [15], ETS$^+$ samples were those identified by the authors as harboring *TMPRSS2:ERG* gene fusions (available in sample data in Oncomine). For the Taylor *et al*. study[16], samples with specific deletions between *TMPRSS2* and *ERG*, or those with outlier expression in matched gene expression data (also available in Oncomine) of *ERG*, *ETV1*, *ETV4* or *ETV5*, were considered ETS$^+$. For the TCGA study, samples with specific deletions between *TMPRSS2* and *ERG* were considered ETS$^+$. For evaluation of ETS/*CHD1* status from gene expression profiling studies, 9 prostate cancer profiling studies [46,47,80-85] (and the International Genomics Consortium's expO dataset) were accessed in Oncomine. In each study, samples with outlier over-expression of *ERG*, *ETV1*, *ETV4* or *ETV5* were considered ETS$^+$, samples with *CHD1* outlier under-expression were considered *CHD1*$^-$ and samples with outlier over-expression of *SPINK1* were considered *SPINK1*$^+$ (see **Supplementary Table 19**).

ETS2

Full length wild type *ETS2* with N-terminal HA-tag was PCR amplified and cloned into pCR8/GW/TOPO vector (Invitrogen). *ETS2* R437C was generated using the Quick change-mutagenesis kit (Stratagene). *ETS2* wildtype and R437C were transferred into pLenti4-V5 DEST vector (Invitrogen). After confirmation of the insert sequence, lentiviruses were generated by the University of Michigan Vector Core. VCaP cells were infected and stably expressing *ETS2* wild type, *ETS2* R437C mutant and lacZ control were generated by selection with Zeocin (Invitrogen). *ETS2* expression was confirmed by qPCR for *ETS2* expression and western blotting with anti-HA antibody as above. For proliferation assays, 50,000 cells (n=*4*) were plated per well in 24-well poly-lysine coated plates, and cells were harvested and counted at the indicated time points by Coulter counter (Beckman Coulter, Fullerton, CA). For in vitro migration and invasion, $2.0x10^5$ cells (migration n=8; invasion n=12) were placed in the top chamber with a noncoated membrane or Matrigel coated membrane, respectively (24-well insert; pore size 8µm; BD Biosciences). In both the assays, cells were plated in medium without serum, and medium supplemented with 10% serum was used as a chemoattractant in the lower chamber. Cells were incubated for 48hr and cells that did not migrate or invade through the pores were gently removed with a cotton swab. Cells on the lower surface of the membrane were stained with crystal violet and counted.

AR interaction with histone/chromatin remodelers

VCaP cells were lysed in Triton X-100 lysis buffer (20mM MOPS, pH 7.0, 2mM EGTA, 5mM EDTA, 30mM sodium fluoride, 60mM β-glycerophosphate, 20mM sodium pyrophosphate, 1mM sodium orthovanadate, 1% Triton X-100, 1mM DTT, protease inhibitor cocktail (Roche, #14309200)). Cell lysates (0.5-1.0mg) were then pre-cleaned with protein A/G agarose beads (Santa Cruz, # sc-2003) by incubation for 1 hour with shaking at 4°C followed by centrifugation at 2000 rpm for 3 minutes. Antibody coupling reactions were performed according to the Dynabeads Antibody Coupling Kit (Invitrogen, Cat# 143.11D). Briefly, 10mg Dynabeads M-270 were washed with buffer and mixed with primary antibody as indicated. Reactions were then incubated on a roller at 37°C overnight (16-24 hours), washed with buffer and resuspended to a final concentration of 10mg antibody coupled beads/mL. Lysates were then incubated overnight with the coupled antibodies as indicated. The mixture was then incubated with shaking at 4°C for another 4 hours or overnight prior to washing the lysate-bead precipitate (centrifugation at 2000 rpm for 3 minutes) 4 times in Triton X-100 lysis buffer. Beads were finally precipitated by centrifugation, resuspended in 25µL of 2x loading buffer and boiled at 80°C for 10 minutes for separation of proteins and beads.

Samples were then analyzed by SDS-PAGE and transferred onto Polyvinylidene Difluoride membrane (GE Healthcare, Piscataway, NJ). The membrane was then incubated in blocking buffer [Tris-buffered saline, 0.1% Tween (TBS-T), 5% nonfat dry milk] for 1 hours at room temperature with the following: anti-ASH2L rabbit polyclonal (1:4000 in blocking buffer, Bethyl lab # A300-489A), anti-MLL mouse monoclonal (1:1000 in blocking buffer, Millipore# 05-765), anti-AR rabbit polyclonal (1:1000 in blocking buffer, Millipore Cat #06-680), anti-FOXA1 mouse monoclonal (1:2000 in blocking buffer, Abcam Cat #ab23738), anti-UTX mouse monoclonal (1:1000 in blocking buffer, Abcam # ab91231), anti-MLL2-Rabbit polyclonal (1:2000 in blocking buffer, Bethyl lab Cat # A300-113A), anti-ASXL2 Rabbit polyclonal (1:2000 in blocking buffer, Abcam Cat # ab69420), anti-CHD1 (1:4000 in blocking buffer, Bethyl lab Cat# A310-411A) and anti-ERG (1:1000 in blocking buffer, Epitomics Cat # EPR 3864. Following washes with TBS-T, the blot was incubated with horseradish peroxidase-conjugated secondary antibody and the signals visualized by enhanced chemiluminescence system as described by the manufacturer (GE Healthcare).

Knockdown of *ASH2L* or *MLL* in VCaP cells was accomplished by RNA interference using commercially available siRNA duplexes for *ASH2L* (Dharmacon, Cat# J-019831-05 [AAAGAUGGCUAUCGGUAUA] and J-019831-08 [GAGACAGAAUCAUCUAAUG]) and *MLL* (Dharmacon, Cat# J-009914-05 [CAAUUGACCUCUUCUGUUA] and J-009914-08 [GAUCAAAGGCGAAGUGGUU]). Transfections were performed with OptiMEM (Invitrogen) and Oligofectamine (Invitrogen) as previously described[86]. For evaluation of effect on androgen signaling, cells were first hormone starved and treated with indicated siRNAs against *ASH2L* or *MLL*. After 48 hours, cells were treated with 1nM R1881 for 3, 6 and 24 hrs for qPCR prior to RNA isolation. qPCR was performed essentially as described using Power SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems 7300 Real Time PCR system for quantification of *ASH2L* and *MLL* knockdown and *PSA* expression [76]. Primer sequences are in **Supplementary Table 20**.

FOXA1

FOXA1 wildtype and *FOXA1* mutants (S453fs, G87R, L388M, L455M and F400I) were cloned and inserted into pCDH (System Biosciences), which has been modified to express an N-terminal FLAG tag and puromycin resistance. Lentiviruses were generated in 293FT cells using the ViraPower Lentiviral Expression System (Invitrogen). LNCaP cells were infected with the generated viruses (or empty control virus) and stable pooled populations were selected with puromycin. Expression was confirmed by western blotting with anti-FLAG antibody (Sigma) or qPCR for *FOXA1* expression as above, and *FOXA1* primers are in **Supplementary Table 20**. qPCR experiments were performed in triplicate, and *FOXA1* expression was normalized to *GAPDH*. For proliferation, cells were starved for 24 hours in phenol red-free media media with

1% charcoal-dextran stripped serum, and grown in media with 1% charcoal-dextran stripped serum +/- 10nM DHT. Relative cell numbers were measured in quadruplicate by WST-1 assays at indicated time points following the manufacturer's instructions (Roche).

Gene expression microarray analysis was performed as above, using LnCAP cells expressing empty vector, *FOXA1* wild type, or *FOXA1* mutants as just described. Cells were starved in 1% charcoal stripped media for 24 hours and treated with 10 nM DHT or vehicle for 48 hours. RNA was isolated using Qiazol. All samples were hybridized against vehicle stimulated vector control in duplicate. Probes not passing filtering in both duplicate hybridizations were excluded from further analysis, and remaining probes were averaged. For each set of *FOXA1* wildtype or mutant hybridization (duplicates of DHT and vehicle treated), DHT vs. vehicle stimulated ratios for each probe passing filtering on all four arrays were computed. Probes were filtered to include only those with average LogRatio (converted to log base 2) of >1 or <-1 in the DHT vs. vehicle stimulated pair. Clustering of probes using centroid-linkage clustering was performed using Cluster 3.0 and heatmaps were generated using JavaTreeview.

In parallel, *FOXA1* wildtype and *FOXA1* mutant (S453fs; resulting from chr14:37130381insCC observed in T12) ORFs were generated by gene synthesis (Blue Heron) and cloned into the pLL_IRES_GFP lentival vector. Lentiviruses (and pLL_IRES_GFP expressing *LACZ* as control) were generated by the University of Michigan Vector Core. LNCaP cells were transduced in the presence of 4μg/mL polybrene (Sigma). After 72 hours, GFP+ cells were sorted at the University of Michigan flow cytometry core. Cells were genotyped to confirm identify. GFP fluorescence was monitored every other day. Soft agar colony forming assays were performed as described[51], except colonies were counted and photographed without staining.

For xenograft experiments, four week-old male SCID C.B17 mice were procured from a mice breeding colony at University of Michigan (maintained by K.J.P.). Mice were anesthetized using a cocktail of xylazine (80 mg/kg IP) and ketamine (10 mg/kg IP) for chemical restraint. Indicated LNCaP cells (2 million cells per implantation site) as above (or parental LNCaP cells) were suspended in 100ul of 1X PBA with 20% high concentration Matrigel (BD Biosciences). Cells were implanted subcutaneously on both sides into the flank region. Tumour growth was monitored bi-weekly by using digital calipers in LNCaP *FOXA1* wildtype (n=9), LNCaP *FOXA1* S453fs (n=10) and parental LNCaP (n=6) groups. Tumour volumes were calculated using the formula (π/6) (L×W2), where L= length of tumour and W= width. All procedures involving mice were approved by the University Committee on Use and Care of Animals (UCUCA) at the University of Michigan and conform to their relevant regulatory standards.

DLX1

For DLX1 immunoblotting, prostate tissues were homogenized in NP40 lysis buffer containing 50 mm Tris-HCl (pH 7.4), 1% NP40 (Sigma), and complete proteinase inhibitor mixture (Roche). Western blotting with ten micrograms of each protein extract was performed as above. Transferred membrane was incubated for 1 h in blocking buffer and over-night with anti-DLX1 rabbit polyclonal antibody (PTG laboratory, #13046-1-AP, 1:1000 dilution). After washing three times with TBS-T buffer, the membrane was incubated with horseradish peroxidase-linked donkey anti-rabbit IgG antibody (GE Healthcare, 1:5,000) for 1 h at room temperature prior to visualization by enhanced chemiluminescence (GE Healthcare). To monitor equal loading, the membrane was re-probed with anti-β-Actin mouse monoclonal antibody (1:30,000 dilution; Sigma, # A5316).

qPCR was performed on 10 benign prostate tissues (included in gene-expression profiling), 55 localized prostate cancers (including 32 samples subjected to gene-expression profiling) and 7 CRPCs (including 6 samples subjected to gene-expression profiling) as above. The amount of *DLX1* in each sample was normalized to the average of *GAPDH* and *HMBS* for each sample. Primers for *DLX1* are given in **Supplementary Table 20**; *GAPDH* and *HMBS* primers were as described[87]. All oligonucleotide primers were synthesized by Integrated DNA Technologies.

Supplementary References:

1 Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc Natl Acad Sci U S A* **108**, 17087-17092 (2011).

2 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).

3 Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-873 (2010).

4 Tomlins, S. A. *et al.* ETS Gene Fusions in Prostate Cancer: From Discovery to Daily Clinical Practice. *Eur Urol* **56**, 275-286 (2009).

5 Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).

6 Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231 (2010).

7 Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274 (2006).

8 Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-1113 (2007).

9 Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-1806 (2008).

10 Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-1812 (2008).

11 Mehra, R. *et al.* Characterization of TMPRSS2-ETS gene aberrations in androgen-independent metastatic prostate cancer. *Cancer Res* **68**, 3584-3590 (2008).

12 Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**, 559-565 (2009).

13 Holcomb, I. N. *et al.* Comparative analyses of chromosome alterations in soft-tissue metastases within and across patients with castration-resistant prostate cancer. *Cancer Res* **69**, 7793-7802 (2009).

14 Lonigro, R. J. *et al.* Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. *Neoplasia* **13**, 1019-1025 (2011).

15 Demichelis, F. *et al.* Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer* **48**, 366-380 (2009).

16 Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11-22 (2010).

17 Shah, S. P. *et al.* Mutation of FOXL2 in granulosa-cell tumours of the ovary. *N Engl J Med* **360**, 2719-2729 (2009).

18 Wiegand, K. C. *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med* **363**, 1532-1543 (2010).

19 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).

20 Gaddipati, J. P. *et al.* Frequent detection of codon 877 mutation in the androgen receptor gene in advanced prostate cancers. *Cancer Res* **54**, 2861-2864 (1994).

21 Zhao, X. Y. *et al.* Glucocorticoids can promote androgen-independent growth of prostate cancer cells through a mutated androgen receptor. *Nat Med* **6**, 703-706 (2000).

22 Tan, J. *et al.* Dehydroepiandrosterone activates mutant androgen receptors expressed in the androgen-dependent human prostate cancer xenograft CWR22 and LNCaP cells. *Mol Endocrinol* **11**, 450-459 (1997).
23 Kastritis, E. *et al.* Somatic mutations of adenomatous polyposis coli gene and nuclear b-catenin accumulation have prognostic significance in invasive urothelial carcinomas: evidence for Wnt pathway implication. *Int J Cancer* **124**, 103-108 (2009).
24 Solomon, D. A. *et al.* Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science* **333**, 1039-1043 (2011).
25 Varier, R. A. & Timmers, H. T. Histone lysine methylation and demethylation pathways in cancer. *Biochim Biophys Acta* **1815**, 75-89 (2011).
26 Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298-303 (2011).
27 Gui, Y. *et al.* Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet* **43**, 875-878 (2011).
28 Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435-439 (2011).
29 Mo, R., Rao, S. M. & Zhu, Y. J. Identification of the MLL2 complex as a coactivator for estrogen receptor alpha. *J Biol Chem* **281**, 15714-15720 (2006).
30 Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24**, 2303-2316 (2010).
31 Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
32 Iorns, E., Martens-de Kemp, S. R., Lord, C. J. & Ashworth, A. CRK7 modifies the MAPK pathway and influences the response to endocrine therapy. *Carcinogenesis* **30**, 1696-1701 (2009).
33 Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393-395 (2009).
34 Yamaoka, M., Hara, T. & Kusaka, M. Overcoming persistent dependency on androgen signaling after progression to castration-resistant prostate cancer. *Clin Cancer Res* **16**, 4319-4324 (2010).
35 Huusko, P. *et al.* Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer. *Nat Genet* **36**, 979-983 (2004).
36 Sun, X. *et al.* Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nat Genet* **37**, 407-412 (2005).
37 Dong, J. T. Prevalent mutations in prostate cancer. *J Cell Biochem* **97**, 433-447 (2006).
38 Agell, L. *et al.* KLF6 and TP53 mutations are a rare event in prostate cancer: distinguishing between Taq polymerase artifacts and true mutations. *Mod Pathol* **21**, 1470-1478 (2008).
39 Narla, G. *et al.* KLF6, a candidate tumour suppressor gene mutated in prostate cancer. *Science* **294**, 2563-2566 (2001).
40 Wong, O. G. *et al.* Plexin-B1 mutations in prostate cancer. *Proc Natl Acad Sci U S A* **104**, 19040-19045 (2007).
41 Ding, Z. *et al.* SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. *Nature* **470**, 269-273 (2011).
42 Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* **18**, 507-522 (2011).

43 Chiba, S., Ikeda, M., Katsunuma, K., Ohashi, K. & Mizuno, K. MST2- and Furry-mediated activation of NDR1 kinase is critical for precise alignment of mitotic chromosomes. *Curr Biol* **19**, 675-681 (2009).

44 Palanisamy, N. *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat Med* **16**, 793-798 (2010).

45 Wang, X. S. *et al.* Characterization of KRAS Rearrangements in Metastatic Prostate Cancer. *Cancer Discov* **1**, 35-43 (2011).

46 Glinsky, G. V., Glinskii, A. B., Stephenson, A. J., Hoffman, R. M. & Gerald, W. L. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* **113**, 913-923 (2004).

47 Lapointe, J. *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816 (2004).

48 Liu, W. *et al.* Identification of novel CHD1-associated collaborative alterations of genomic structure and functional assessment of CHD1 in prostate cancer. *Oncogene* (2011).

49 Huang, S. *et al.* Recurrent deletion of CHD1 in prostate cancer with relevance to cell invasiveness. *Oncogene* (2011).

50 Mertz, K. D. *et al.* Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model. *Neoplasia* **9**, 200-206 (2007).

51 Tomlins, S. A. *et al.* Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia* **10**, 177-188 (2008).

52 Hollenhorst, P. C. *et al.* Oncogenic ETS proteins mimic activated RAS/MAPK signaling in prostate cells. *Genes Dev* **25**, 2147-2157 (2011).

53 Turner, D. P. & Watson, D. K. ETS transcription factors: oncogenes and tumour suppressor genes as therapeutic targets for prostate cancer. *Expert Rev Anticancer Ther* **8**, 33-42 (2008).

54 Gelsi-Boyer, V. *et al.* Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol* **145**, 788-800 (2009).

55 Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-363 (2010).

56 van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat Genet* **41**, 521-523 (2009).

57 Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629 (2002).

58 Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* (2011).

59 Gerhardt, J. *et al.* FOXA1 Promotes Tumour Progression in Prostate Cancer and Represents a Novel Hallmark of Castrate-Resistant Prostate Cancer. *Am J Pathol* (2011).

60 Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394 (2011).

61 Sahu, B. *et al.* Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *Embo J* **30**, 3962-3976 (2011).

62 Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* **3**, 111ra121 (2011).

63 Sarker, D., Reid, A. H., Yap, T. A. & de Bono, J. S. Targeting the PI3K/AKT pathway for the treatment of prostate cancer. *Clin Cancer Res* **15**, 4799-4805 (2009).

64 Rubin, M. A. *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* **6**, 1038-1045 (2000).

65 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

66 Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

67 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).

68 Forbes, S. A. *et al.* COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* **38**, D652-657 (2010).

69 Grun, B. & Leisch, F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software* **28**, 1-35 (2008).

70 Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467-472 (2011).

71 Getz, G. *et al.* Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).

72 Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075 (2008).

73 Kim, J. H. *et al.* Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* **67**, 8229-8239 (2007).

74 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772 (2009).

75 Luc, P. V. & Tempst, P. PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics* **20**, 1413-1415 (2004).

76 Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599 (2007).

77 Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).

78 Helgeson, B. E. *et al.* Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. *Cancer Res* **68**, 73-80 (2008).

79 Park, K. *et al.* Antibody-based detection of ERG rearrangement-positive prostate cancer. *Neoplasia* **12**, 590-598 (2010).

80 LaTulippe, E. *et al.* Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* **62**, 4499-4506 (2002).

81 Liu, P. *et al.* Sex-determining region Y box 4 is a transforming oncogene in human prostate cancer cells. *Cancer Res* **66**, 4011-4019 (2006).

82 Tamura, K. *et al.* Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. *Cancer Res* **67**, 5117-5125 (2007).

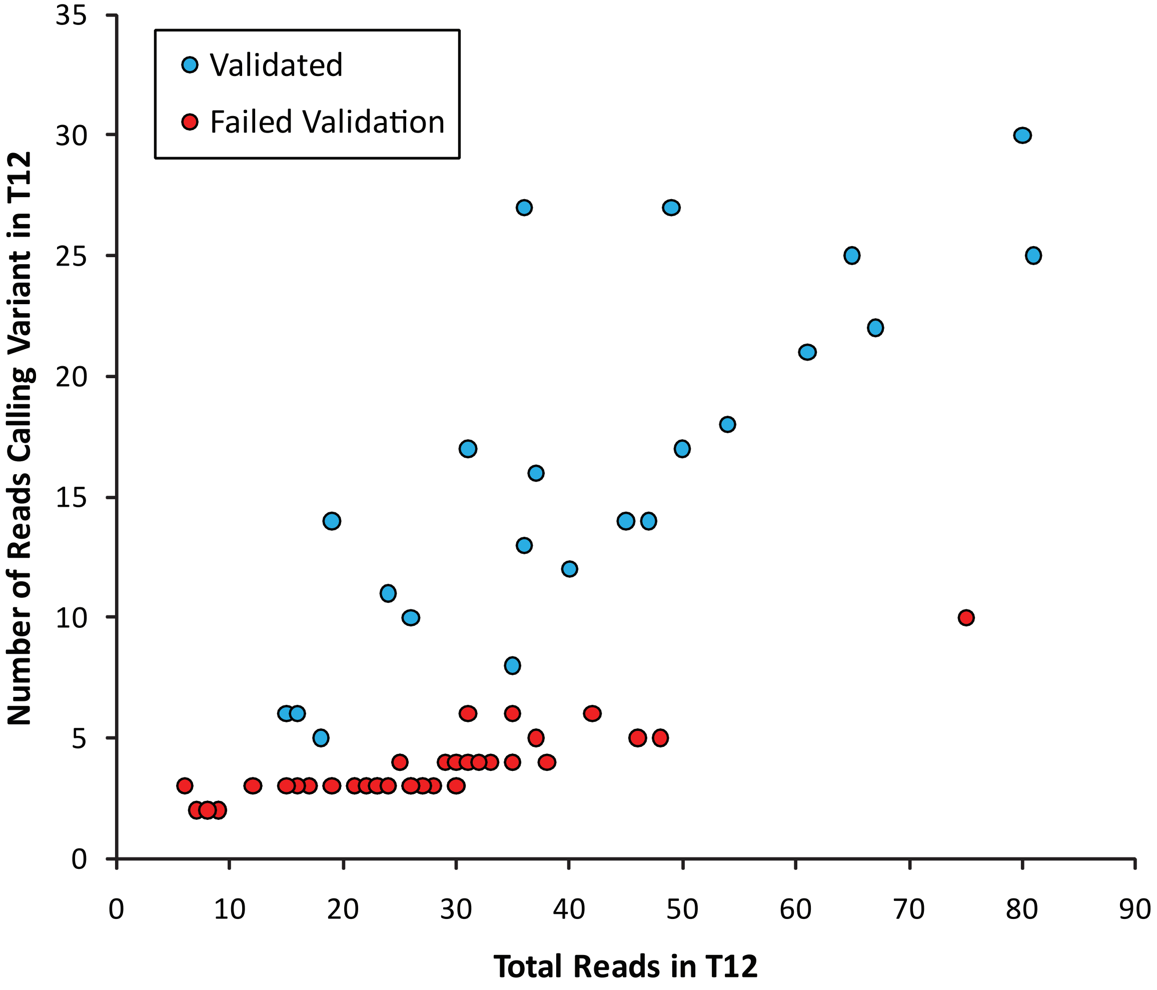
83 Wallace, T. A. *et al.* Tumour immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res* **68**, 927-936 (2008).

84 Welsh, J. B. *et al.* Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* **61**, 5974-5978 (2001).

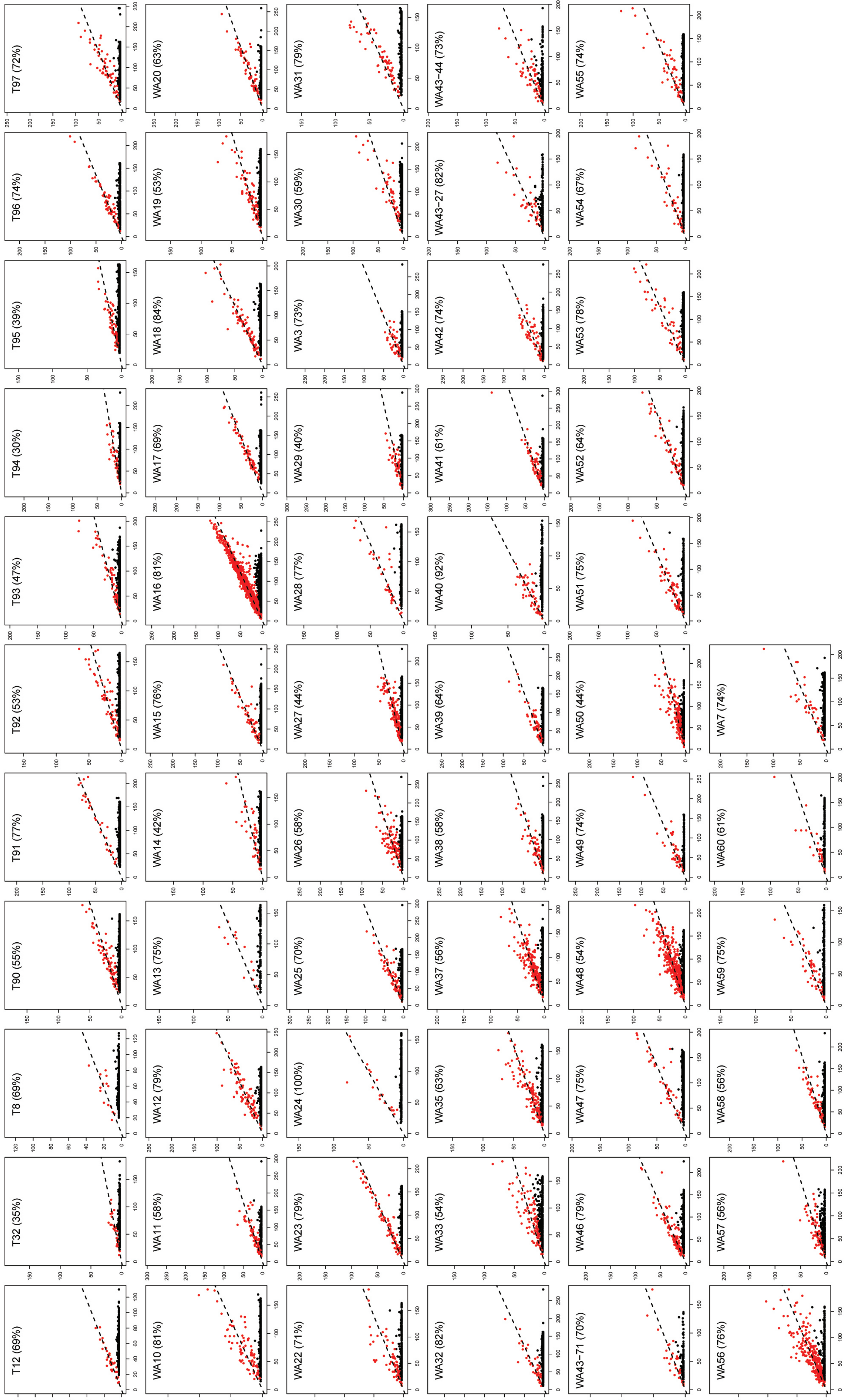
85 Yu, Y. P. *et al.* Gene expression alterations in prostate cancer predicting tumour aggression and preceding development of malignancy. *J Clin Oncol* **22**, 2790-2799 (2004).

86 Varambally, S. *et al.* Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science* **322**, 1695-1699 (2008).

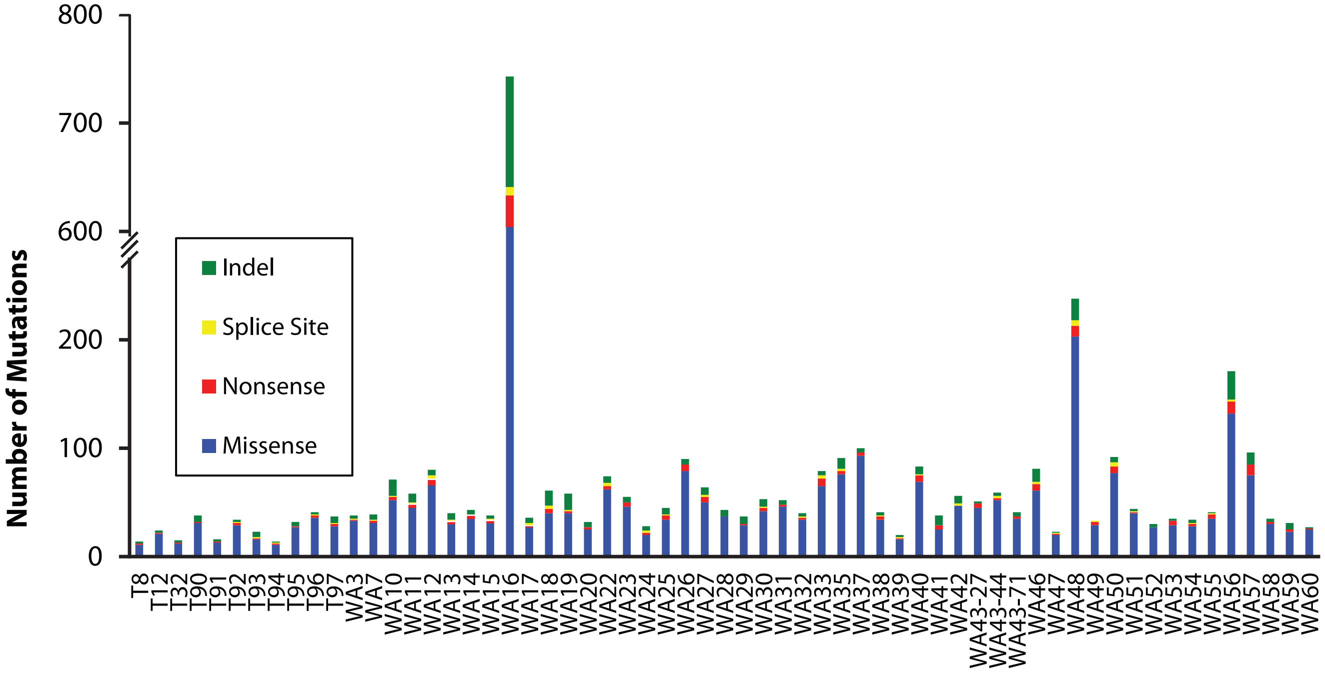
87 Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034 (2002).



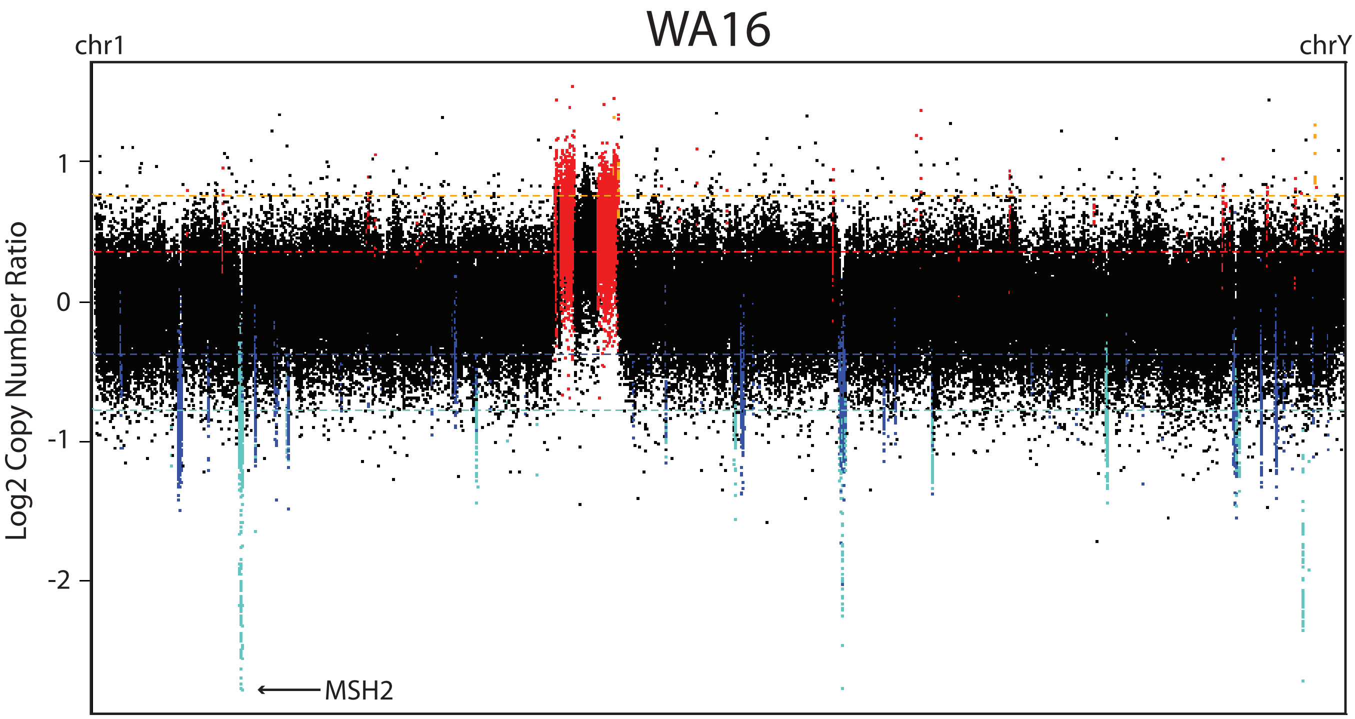
Supplementary Figure 1. **Somatic mutation validation as a function of the number of reads calling the variant and the total number of reads**. Variants validated (blue) or failing validation (red) as somatic mutations in T12 by Sanger sequencing are indicated.

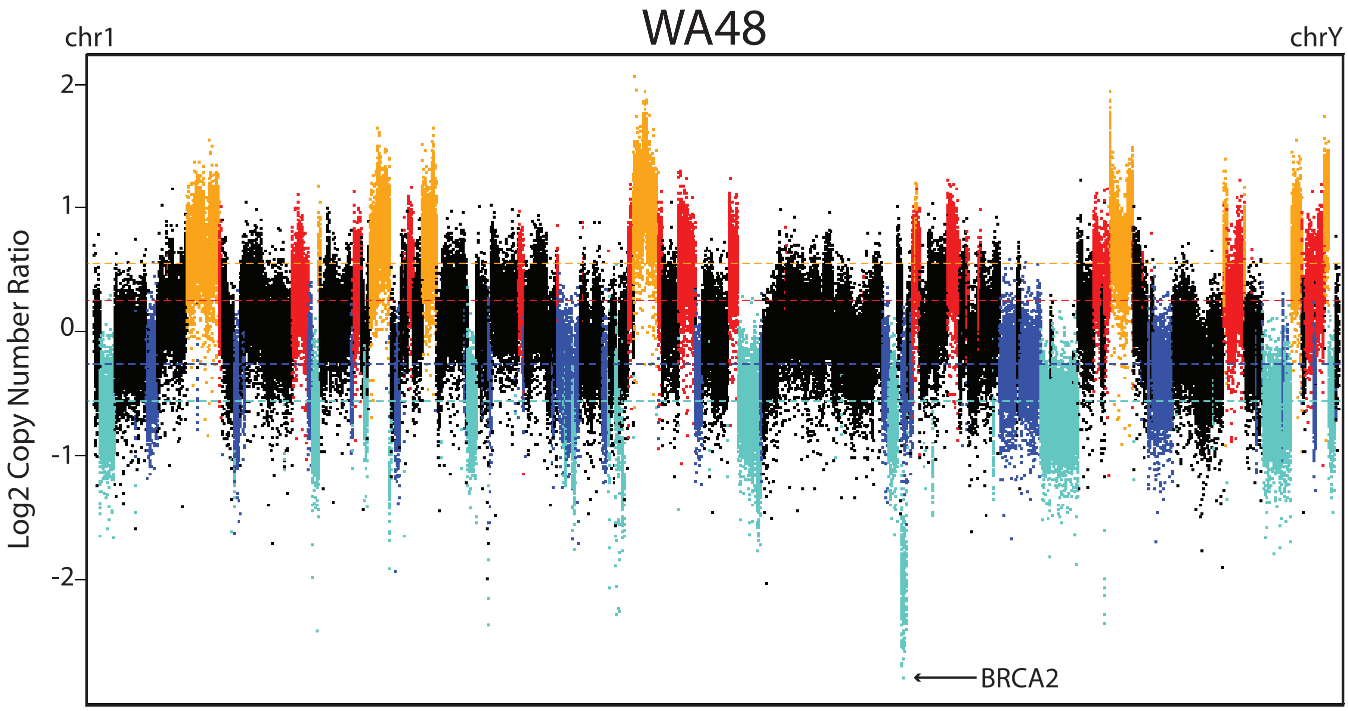

T97 (72%)
T96 (74%)
T95 (39%)
T94 (30%)
T93 (47%)
T92 (53%)
T91 (77%)
T90 (55%)
T8 (69%)
T32 (35%)
T12 (69%)
WA20 (63%)
WA19 (53%)
WA18 (84%)
WA17 (69%)
WA16 (81%)
WA15 (76%)
WA14 (42%)
WA13 (75%)
WA12 (79%)
WA11 (58%)
WA10 (81%)
WA31 (79%)
WA30 (59%)
WA3 (73%)
WA29 (40%)
WA28 (77%)
WA27 (44%)
WA26 (58%)
WA25 (70%)
WA24 (100%)
WA23 (79%)
WA22 (71%)
WA43−44 (73%)
WA43−27 (82%)
WA42 (74%)
WA41 (61%)
WA40 (92%)
WA39 (64%)
WA38 (58%)
WA37 (56%)
WA35 (63%)
WA33 (54%)
WA32 (82%)
WA55 (74%)
WA54 (67%)
WA53 (78%)
WA52 (64%)
WA51 (75%)
WA50 (44%)
WA49 (74%)
WA48 (54%)
WA47 (75%)
WA46 (79%)
WA43−71 (70%)
WA7 (74%)
WA60 (61%)
WA59 (75%)
WA58 (56%)
WA57 (56%)
WA56 (76%)

Supplementary Figure 2. **Tumor content estimates across prostate cancer samples.** For each sample, a binomial mixture model was applied to a set of candidate somatic mutations classifying them either as likely SNVs (red) or likely sequencing errors (black) based on the fraction of the reads calling the variant. The tumor content was estimated as twice variant fraction (the slope of the fitted dotted red line) under the assumption that most of the somatic mutations considered are clonal, heterozygous, and in 2-copy number genomic regions.

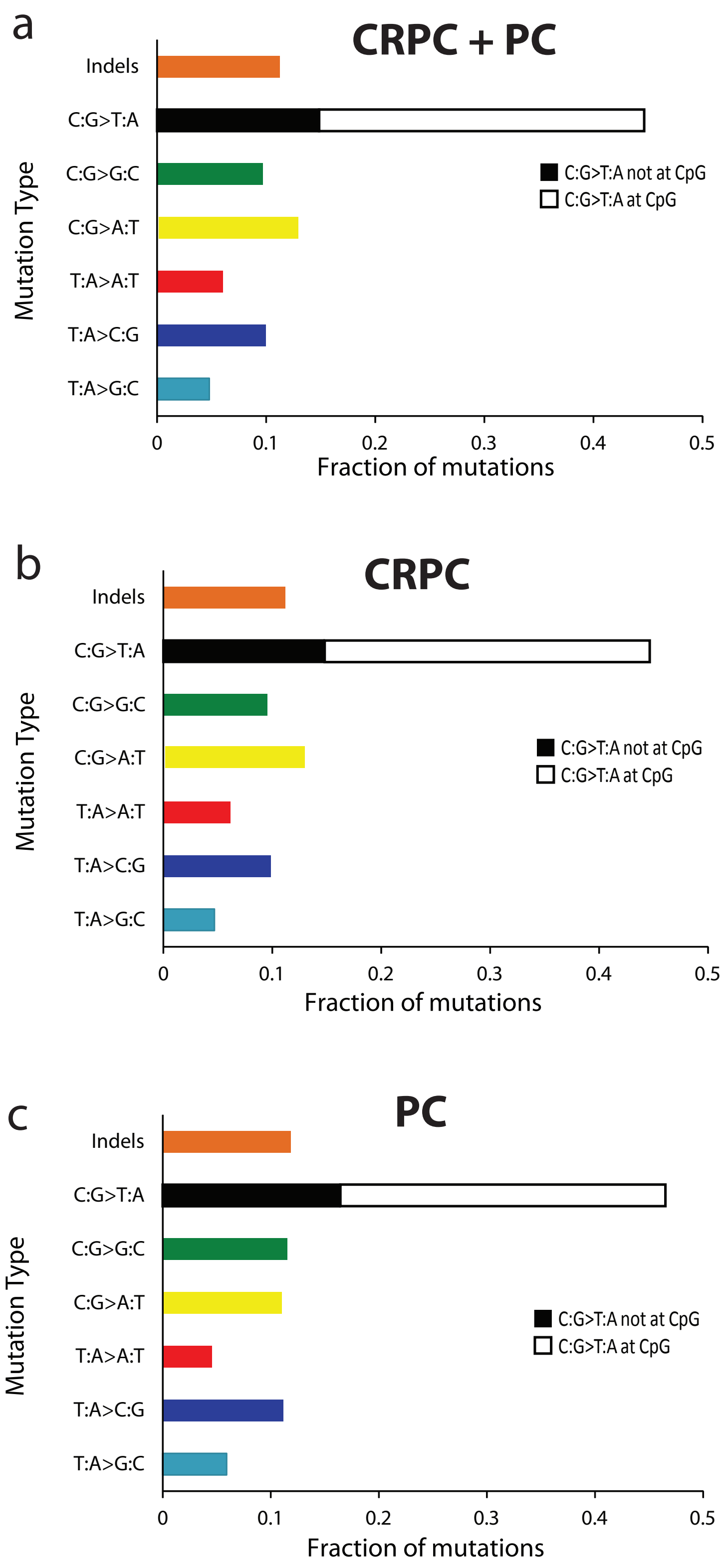


Supplementary Figure 3. Mutational burden of castrate resistant metastatic prostate cancer (CRPC). Exomes of 50 CRPC (WA3-WA60; three foci from WA43) and 11 high grade localized prostate cancers (T8-T97) were sequenced for determination of somatic mutations and copy number alterations. The number of nonsynonomous somatic mutations, including missense (blue), nonsense (red), splice site mutations (yellow) and indels (green) for each sample are shown.

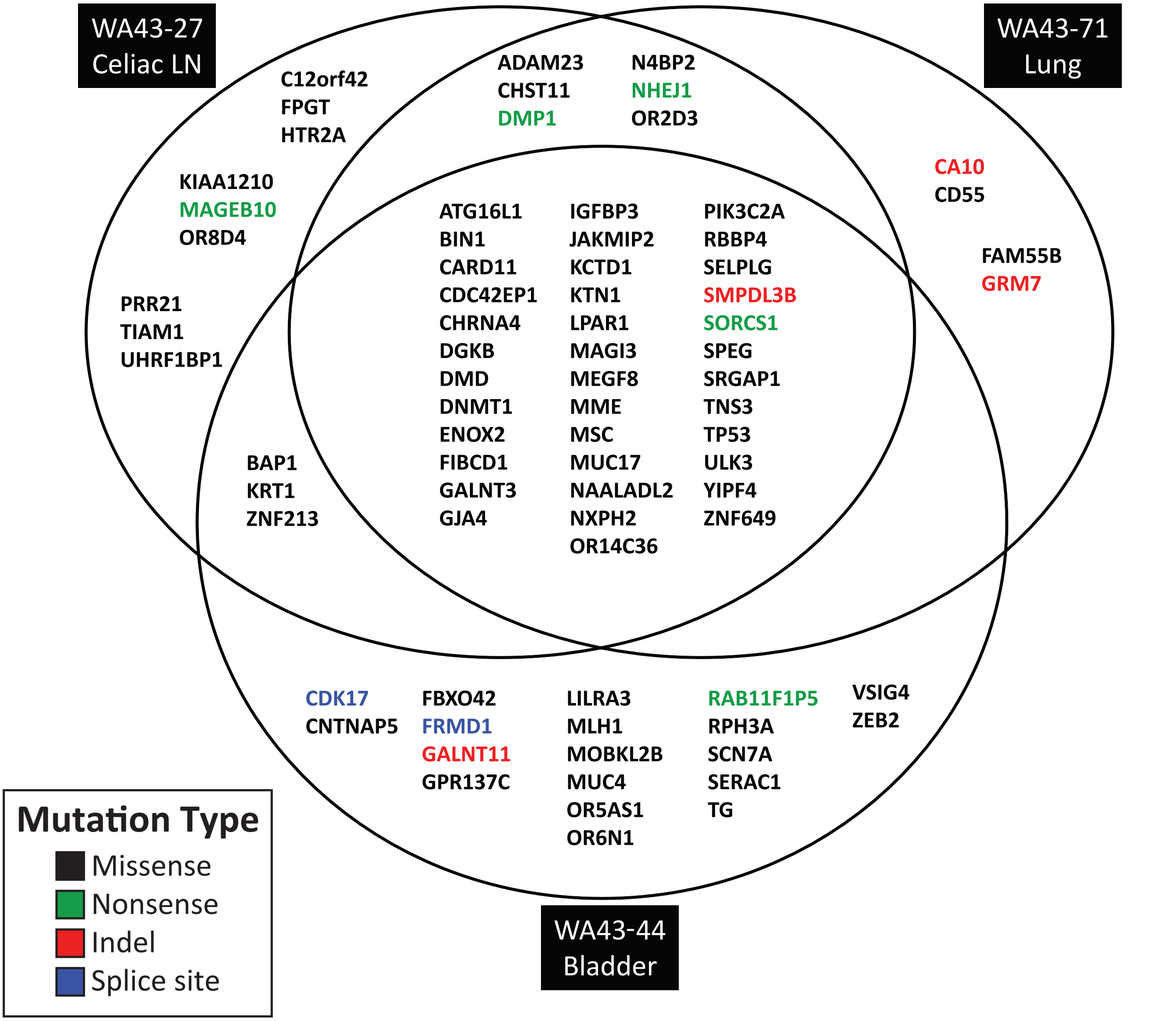




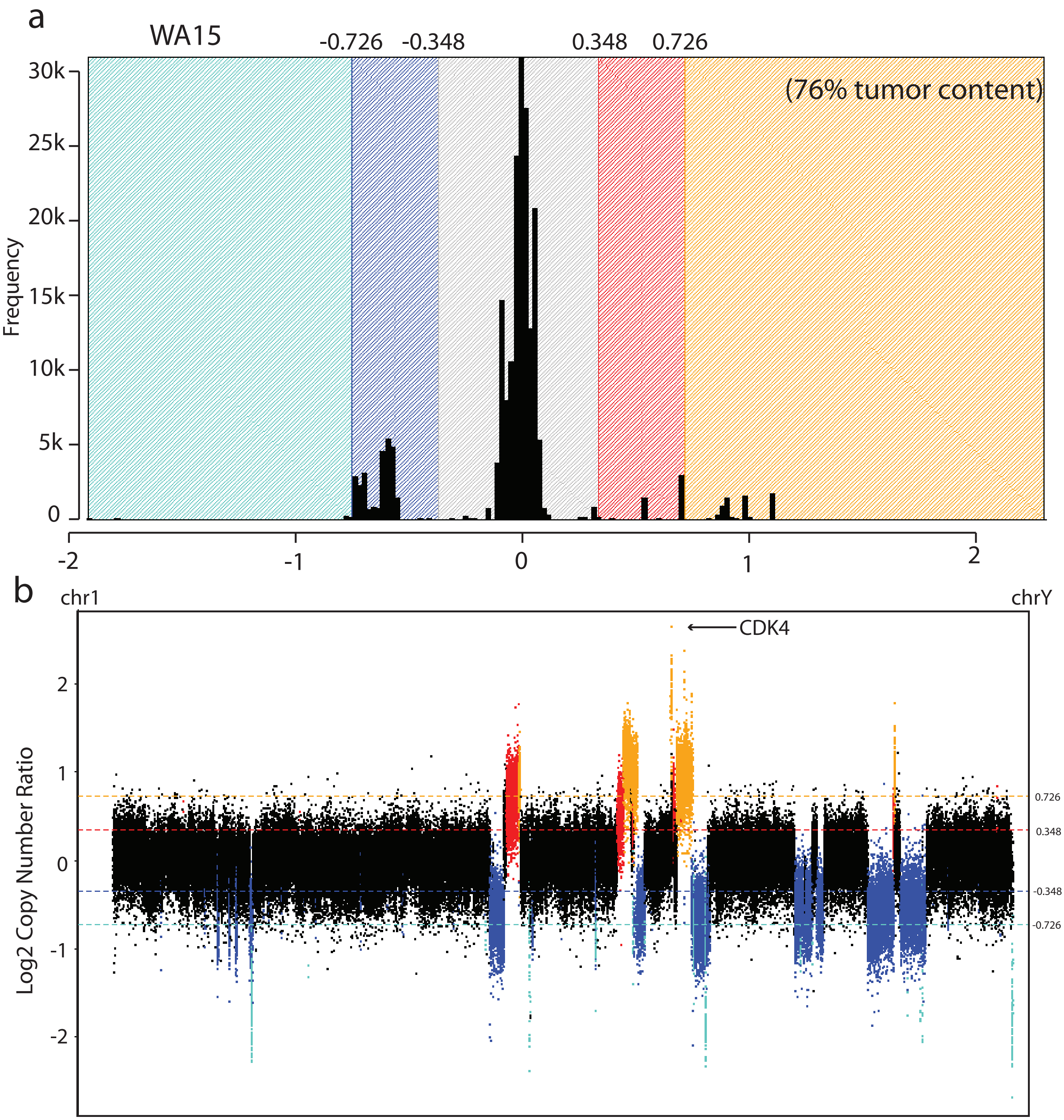
Supplementary Figure 4. Deletion of genes involved in DNA repair in hypermutated CRPC samples. Genome wide copy number plots by exome sequencing for two hypermutated CRPC samples (WA16 and WA48). For each sample, the Log2 copy number ratio between tumor and matched normal is plotted for each targeted exon and then ordered by genomic location. Genes with 1 copy gain/loss are indicated by red and blue points, respectively, and those with > 1 copy gain/loss are indicated by orange and cyan points, respectively. The location of focal high level deletions of *MSH2* in WA16 and *BRCA2* in WA48 are indicated.



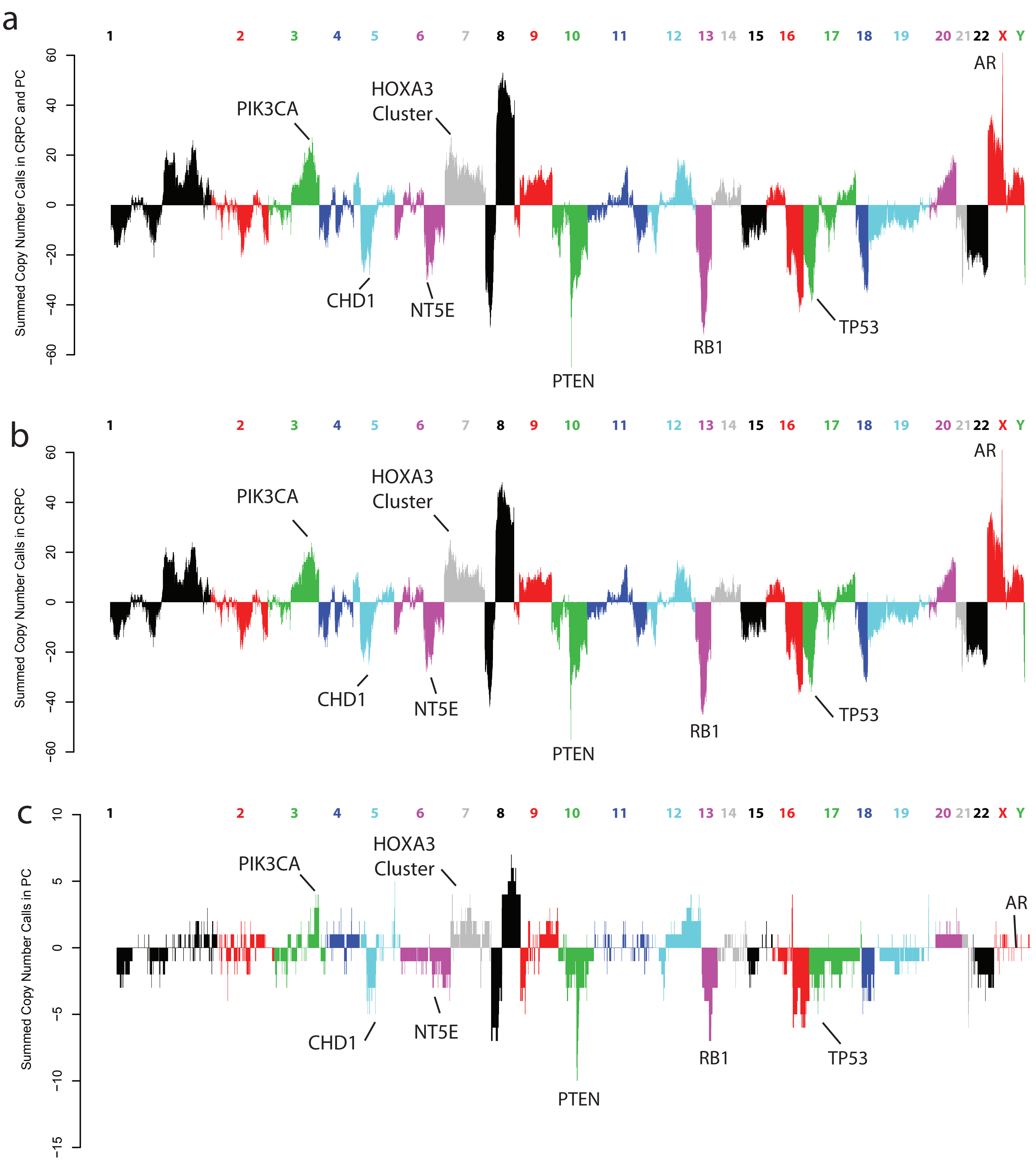
Supplementary Figure 5. Mutation spectrum of prostate cancer. The percentage of coding somatic mutations for each of the six classes of base substitutions and indels are shown for **a**) both castrate resistant prostate cancer (CPRC) and localized prostate cancer (PC), **b**) just CRPC, and **c**) just PC. C:G>T:A substitutions have been divided into those at CpG dinucleotides (white) and those not at CpG dinucleotides (black).



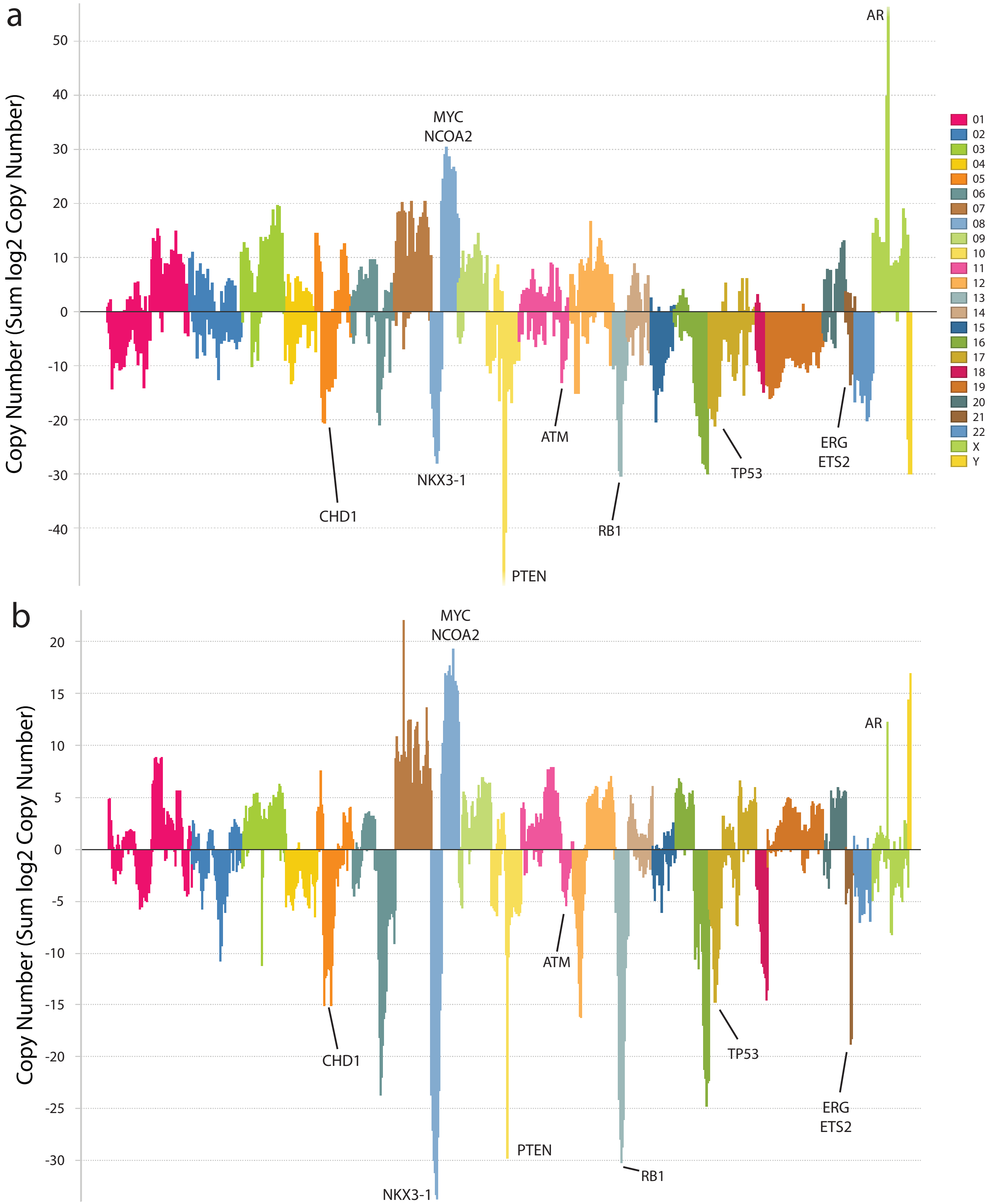
Supplementary Figure 6. Somatic mutations in three different metastatic foci from the same patient confirm the monoclonal origin of lethal metastatic castrate resistant prostate cancer. Venn diagram displaying somatic mutations, including missense (black), nonsense (green), indels (red), and splice site (blue), identified in the celiac lymph node metastatic site (WA43-27), the lung metastatic site (WA43-71), and the bladder local extension/metastatic site (WA43-44).



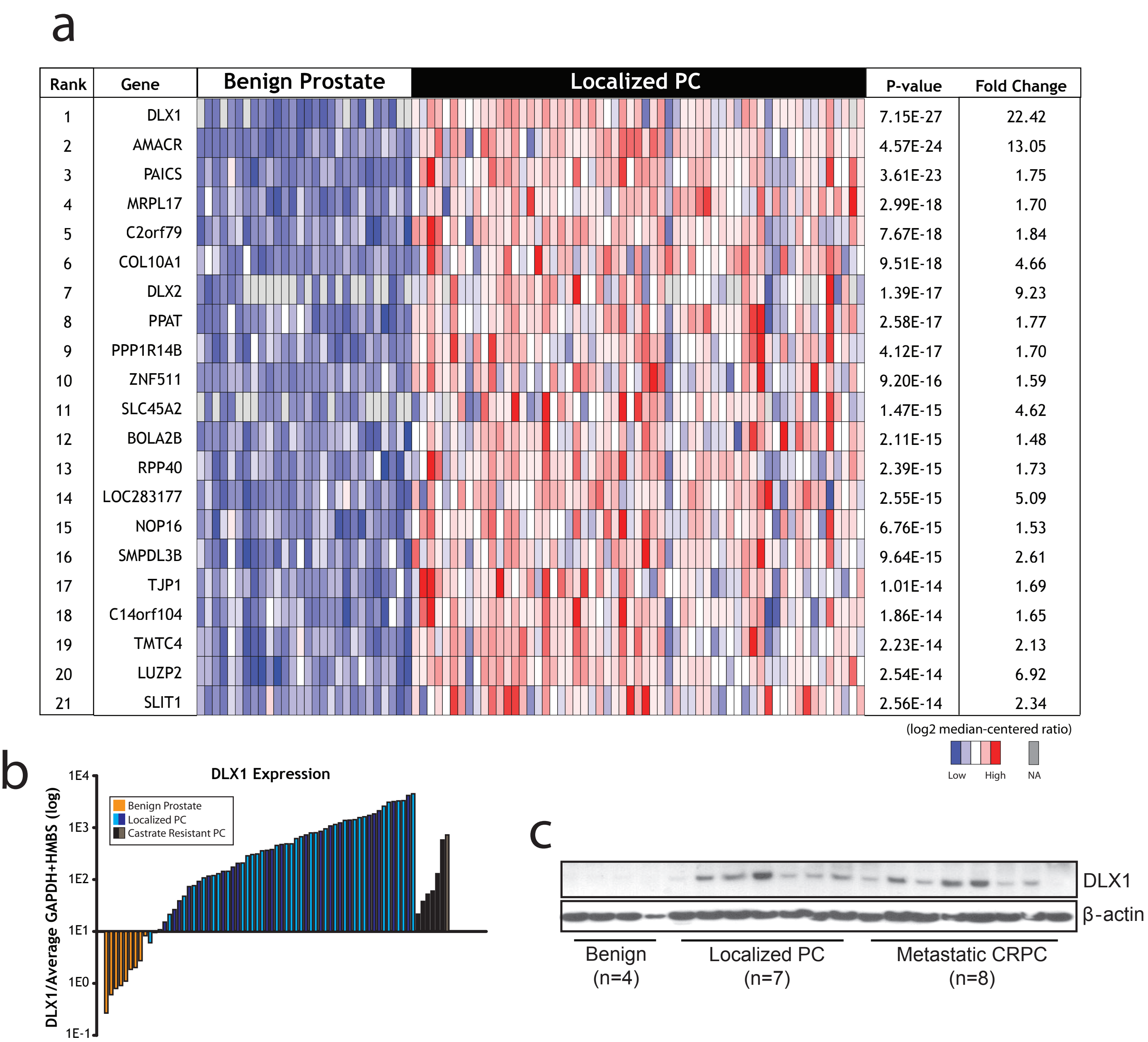
Supplementary Figure 7. Genome wide copy number analysis by exome sequencing and identification of 1 copy and >1 copy gains/losses. For every sample, segmented normalized log2-transformed exon coverage ratios between each tumor sample and its matched normal were computed. Sample-specific cutoffs were generated, based on estimated tumor content, to define regions of 1 copy gain and loss, and >1 copy gain/loss. Manual adjustment of cutoffs was performed for samples where the algorithmic approach appeared to misclassify large numbers of genomic regions. **a**. Distribution histogram of all Log2 copy number ratios (tumor to normal) for each targeted exon in WA15. The automated cutoffs for no change (gray), 1 copy loss (blue) and gain (red), and > 1 copy loss (cyan) and gain (orange) are indicated. Manual adjustment was not performed for this sample. **b**. Genome wide copy number aberrations for WA15. The Log2 copy number ratio (tumor to normal) for each targeted exon in WA15, ordered by genomic location is shown. Using the cutoffs in **a**, genes with called 1 copy gain/loss are indicated by red and blue points, respectively, and those with >1 copy gain/loss are indicated by orange and cyan points, respectively. The location of a focal amplification including *CDK4* on chr12 is indicated.



Supplementary Figure 8. Comparison of copy number aberrations identified by exome sequencing in castrate resistant prostate cancer (CRPC) and localized prostate cancer. Exomes of 50 CRPC (WA3-WA60; three foci from WA43) and 11 high-grade untreated localized prostate cancers (T8-T97) were sequenced for determination of somatic mutations and copy number alterations. Genome wide copy number analysis of each sample was performed using exome sequencing. For all genes, the sum of somatic copy number calls (+/-1: one copy gain or loss, respectively; +/-2: high level copy gain/loss, respectively) across **a**) all profiled samples, **b**) only CRPC samples or **c**) only localized prostate cancers was plotted and ordered by genome location (WA43-24 and -71 are excluded from **a** and **b**). Genes in peaks of copy changes are indicated.

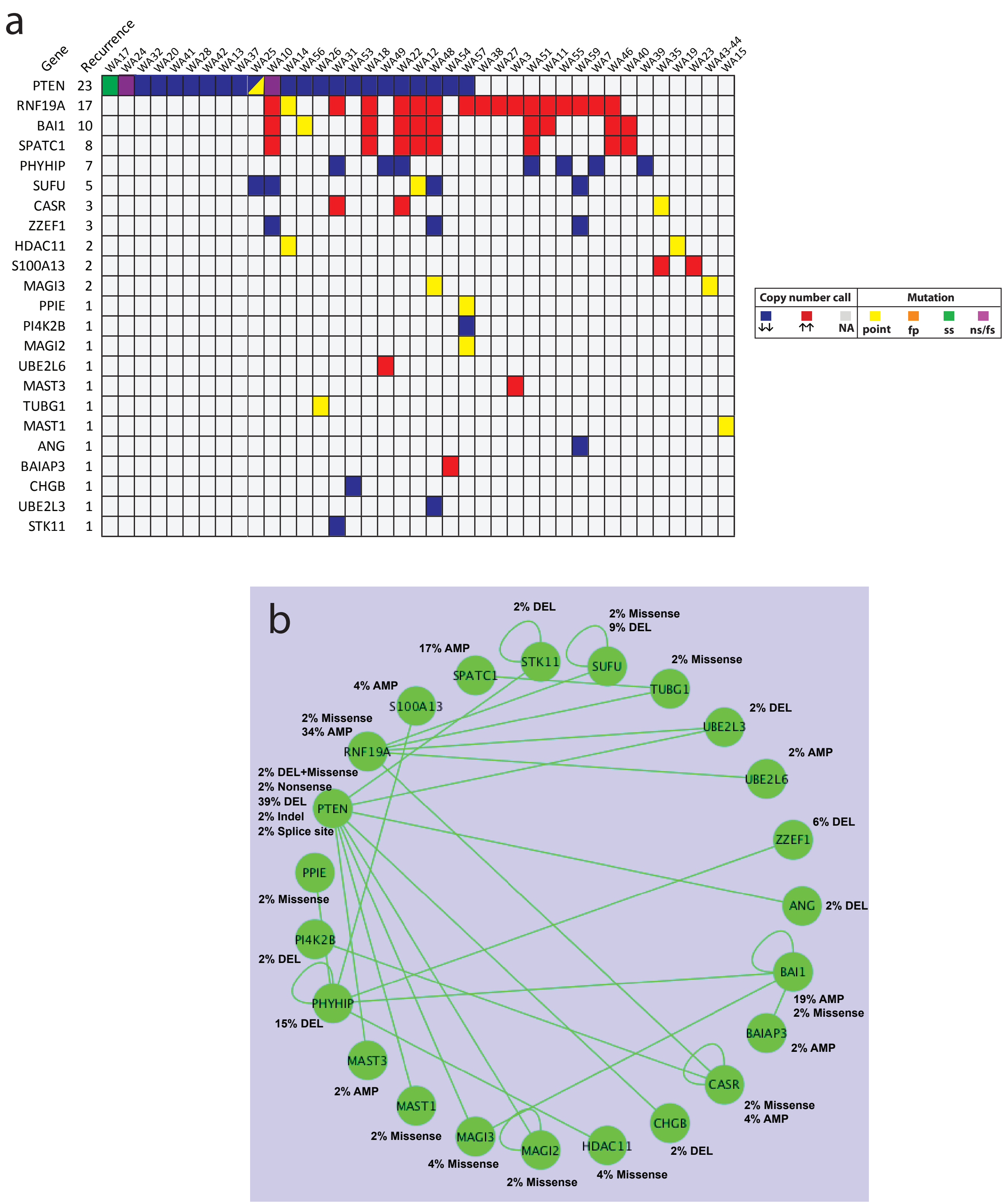


Supplementary Figure 9. Comparison of copy number profiling studies of prostate cancer. a. Our aCGH profiling of localized prostate cancer (PC, *n*=59) and CRPC (*n*=35) was uploaded into Oncomine for analysis and visualization. The sum of the log2 copy number for all genes (chromosomes indicated by color scale on right) from all samples is plotted. **b**. As in **a**, except the overall sum of log2 copy numbers from three individual prostate cancer profiling studies available in Oncomine (Demichelis *et al.*[15], *n*= 49, localized PC; Taylor *et al.*[16], *n*= 218, localized and hormone treated localized PC and metastatic PC; and TCGA, *n*= 64, localized PC) are plotted. Genes present in areas of copy number gains/losses are indicated.

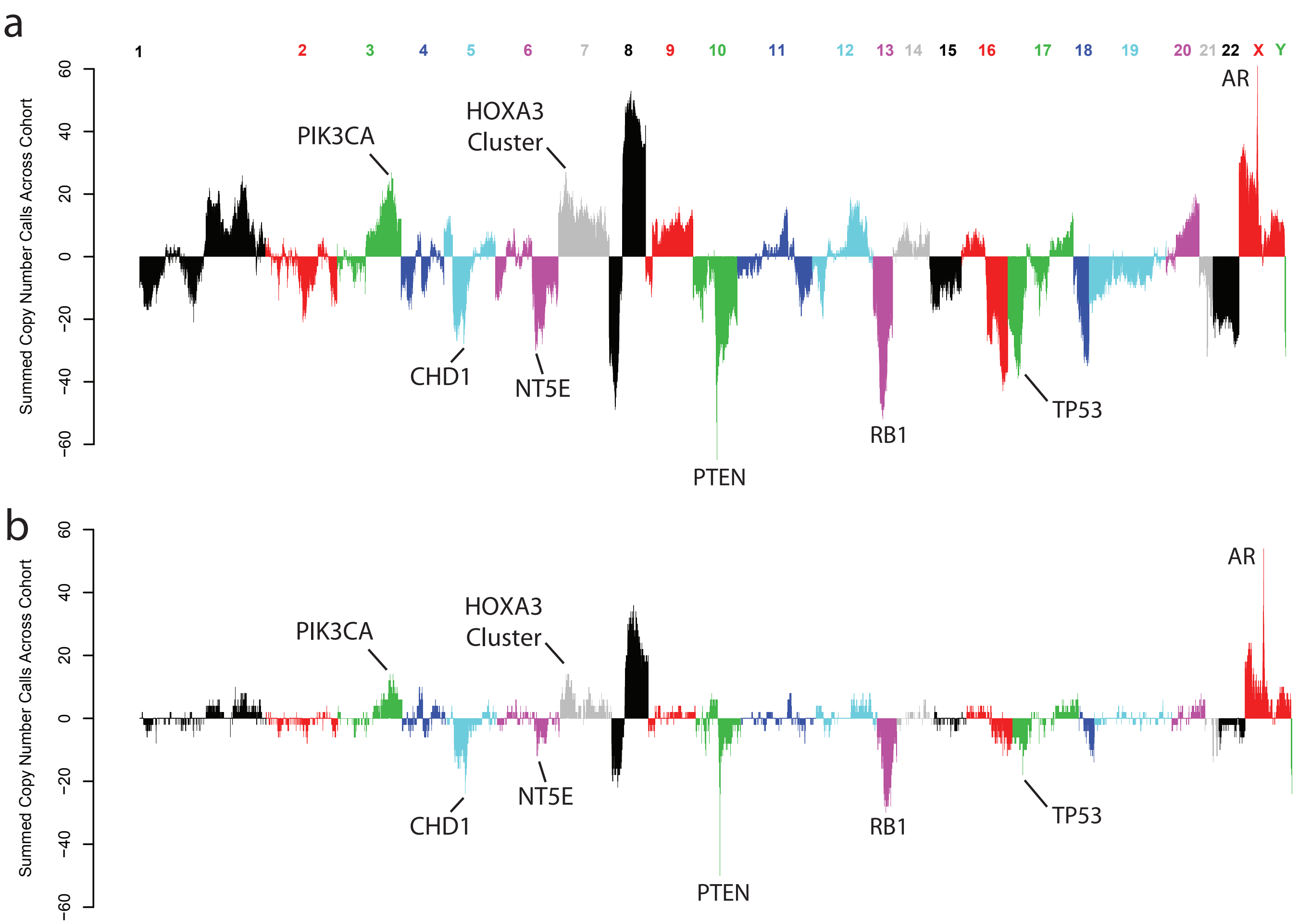


Rank	Gene	P-value	Fold Change
1	DLX1	7.15E-27	22.42
2	AMACR	4.57E-24	13.05
3	PAICS	3.61E-23	1.75
4	MRPL17	2.99E-18	1.70
5	C2orf79	7.67E-18	1.84
6	COL10A1	9.51E-18	4.66
7	DLX2	1.39E-17	9.23
8	PPAT	2.58E-17	1.77
9	PPP1R14B	4.12E-17	1.70
10	ZNF511	9.20E-16	1.59
11	SLC45A2	1.47E-15	4.62
12	BOLA2B	2.11E-15	1.48
13	RPP40	2.39E-15	1.73
14	LOC283177	2.55E-15	5.09
15	NOP16	6.76E-15	1.53
16	SMPDL3B	9.64E-15	2.61
17	TJP1	1.01E-14	1.69
18	C14orf104	1.86E-14	1.65
19	TMTC4	2.23E-14	2.13
20	LUZP2	2.54E-14	6.92
21	SLIT1	2.56E-14	2.34

Supplementary Figure 10. Differential expression of *DLX1* between benign prostate tissue and localized prostate cancer. a. Gene expression profiles from benign prostate tissues (*n*=28), localized prostate cancer (PC, *n*=59), and CRPC (*n*=35, not shown), including samples subjected to exome sequencing, were loaded into Oncomine for automated analysis. The signature of genes most over-expressed in localized PC compared to benign prostate tissue is shown. Genes are ranked according to *P*-values from *t*-tests between the two groups and the fold change between the means of the two groups is given. *Z*-score normalization of values for heatmap visualization is used, with the median value for each gene indicated in white, and the largest changes in the positive and negative direction indicated in bright red and blue, respectively. Gray indicates probes not passing filtering. **b**. *DLX1* expression was measured by qPCR in 10 benign prostate tissues (orange, all included in gene expression profiling), 55 localized PCs (samples included or not included in gene expression profiling indicated in cyan and dark blue, respectively) and 7 metastatic CRPCs (samples included or not included in gene expression profiling indicated in black and gray, respectively). *DLX1* expression in each sample was normalized to the average amount of *GAPDH* and *HMBS*. **c**. Expression of DLX1 by western blotting in 4 benign prostate tissues, 7 localized prostate cancers and 8 metastatic CRPCs. β-actin was used as loading control.



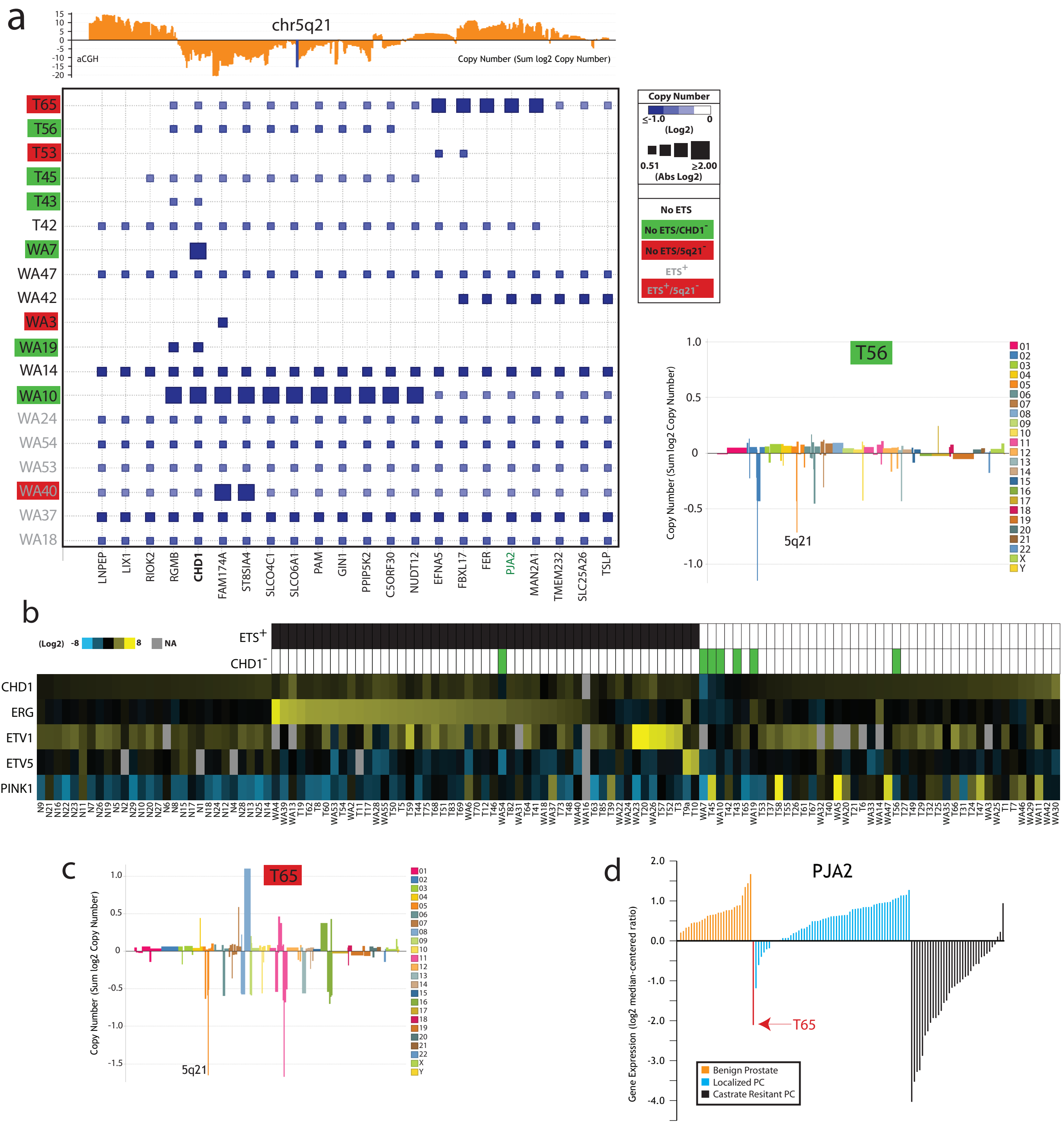
Supplementary Figure 11. Significantly mutated *PTEN* protein-interaction subnetwork. a. Matrix indicating the mutations observed in each sample and gene in the *PTEN* subnetwork, according to the legend. **b.** Network graph showing the interactions (edges) between proteins (nodes) and indicating the percentage of samples with mutations affecting each protein, classified by type: indel, amplification (AMP), copy number loss (DEL), missense, nonsense and splice site.



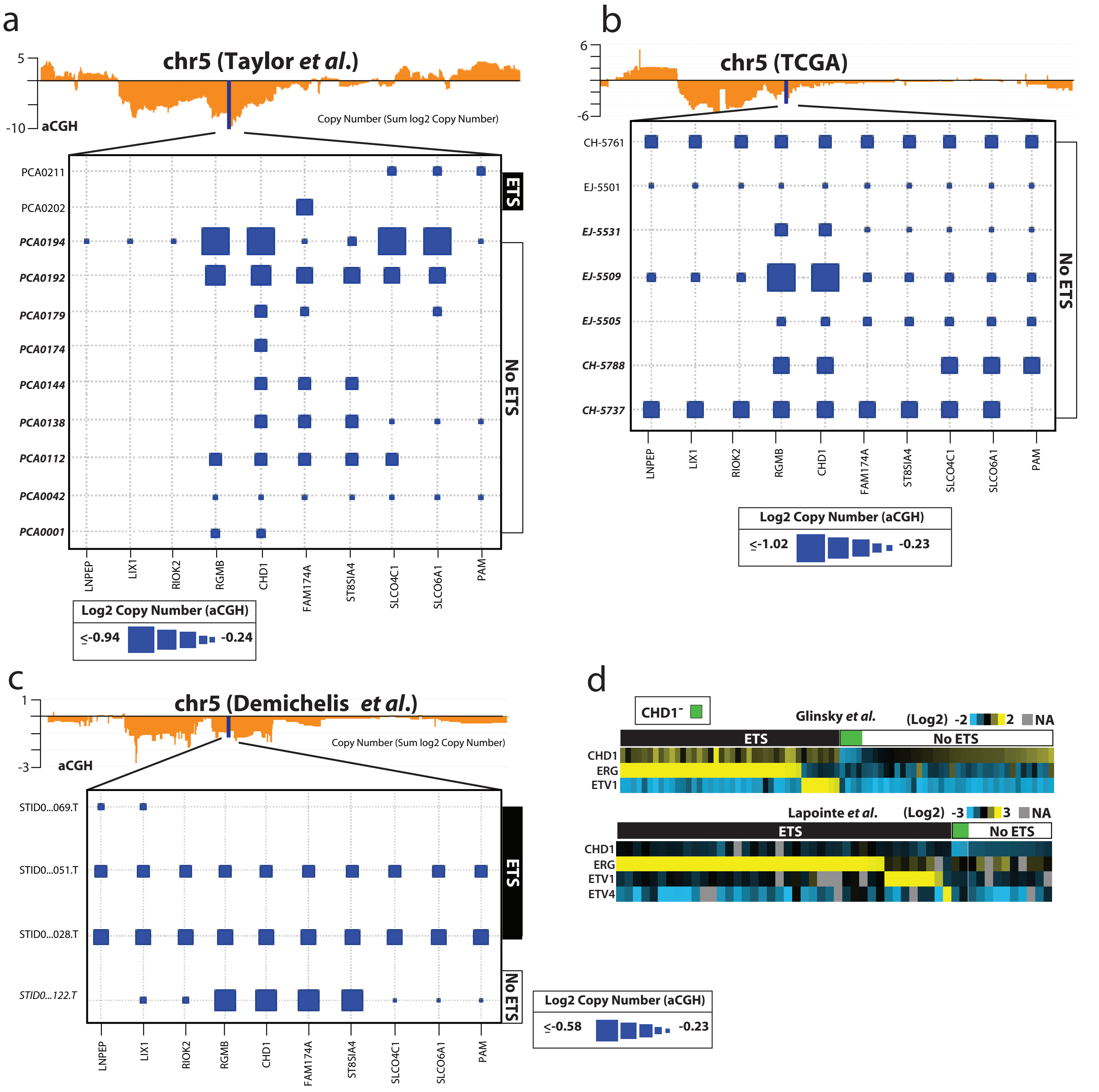
c

Gene	Location	Copy number call sum	>1 Copy gain	1 Copy gain	1 Copy loss	>1 Copy loss
PIK3CA	chr3:180,399,307-180,434,846	25	6	14	1	0
CHD1	chr5:98,219,983-98,289,990	-25	0	7	10	11
NT5E	chr6:86,216,576-86,260,441	-30	0	3	21	6
HOXA Cluster	chr7:27,100,583-27,206,221	27	7	16	3	0
PTEN	chr10:89,613,174-89,718,512	-64	0	0	14	25
RB1	chr13:47,776,049-47,952,208	-49	0	1	24	13
TP53	chr17:7,513,651-7,520,637	-39	0	0	21	9
AR	chrX:66,681,713-66,860,408	57	25	7	0	0

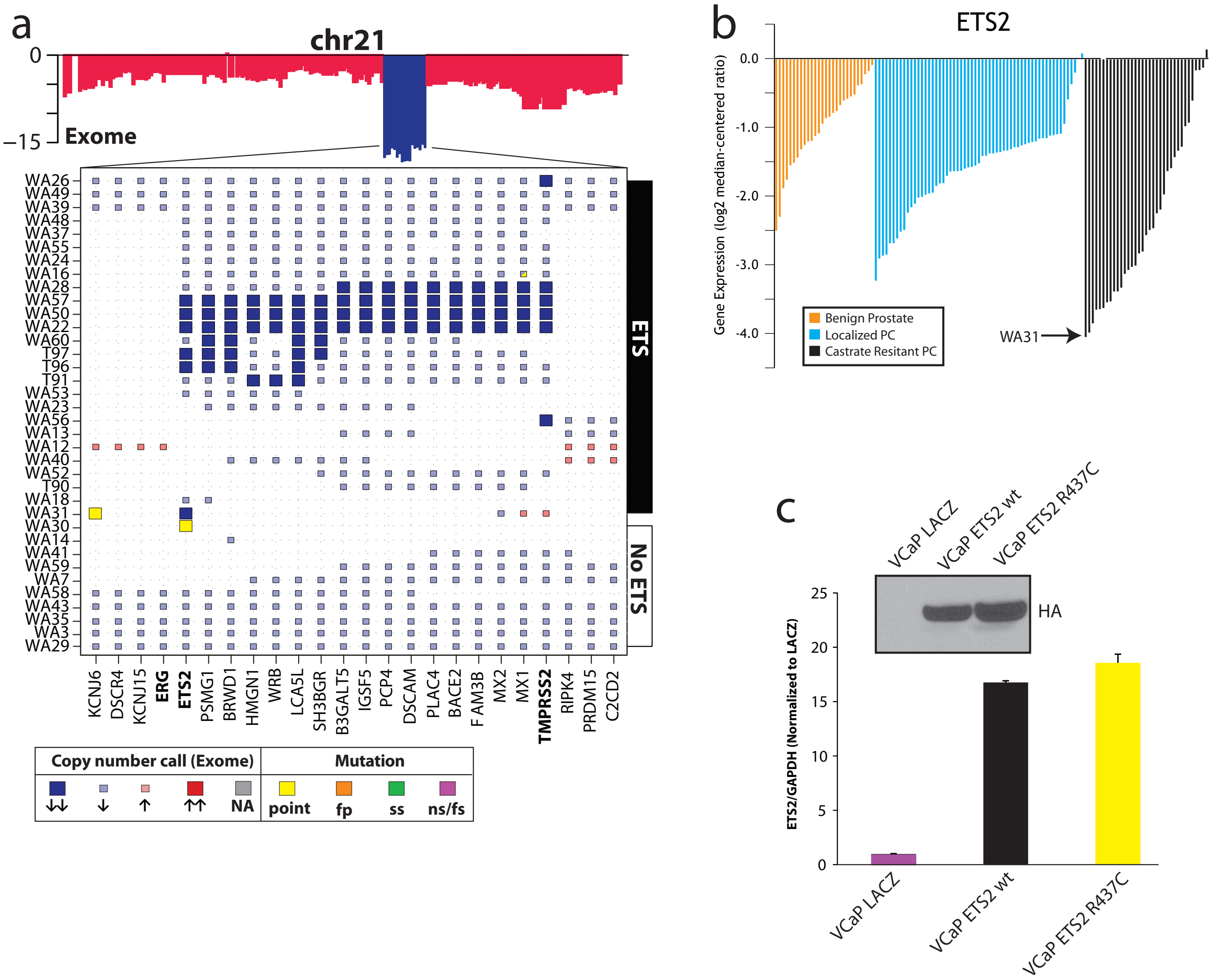
Supplementary Figure 12. Identification of high level, focal copy number aberrations in prostate cancer. **a**. Genome wide copy number analysis of each sample was performed using exome sequencing. For all genes, the sum of copy number calls (+/-1: one copy gain or loss, respectively; +/-2: high level copy gain/loss, respectively) across all samples is plotted and ordered by genome location. **b**. As in **a**, but only the sum of high level copy gains/losses (+/-2) is plotted. **c**. Table showing genes with maximum of high level copy number aberrations. For each gene, the sum of copy number calls, and the number of samples with 1 copy gain/loss or >1 copy gain/loss are indicated. Of the three profiled WA43 samples, only WA43-44 was used for copy number calls/sums.



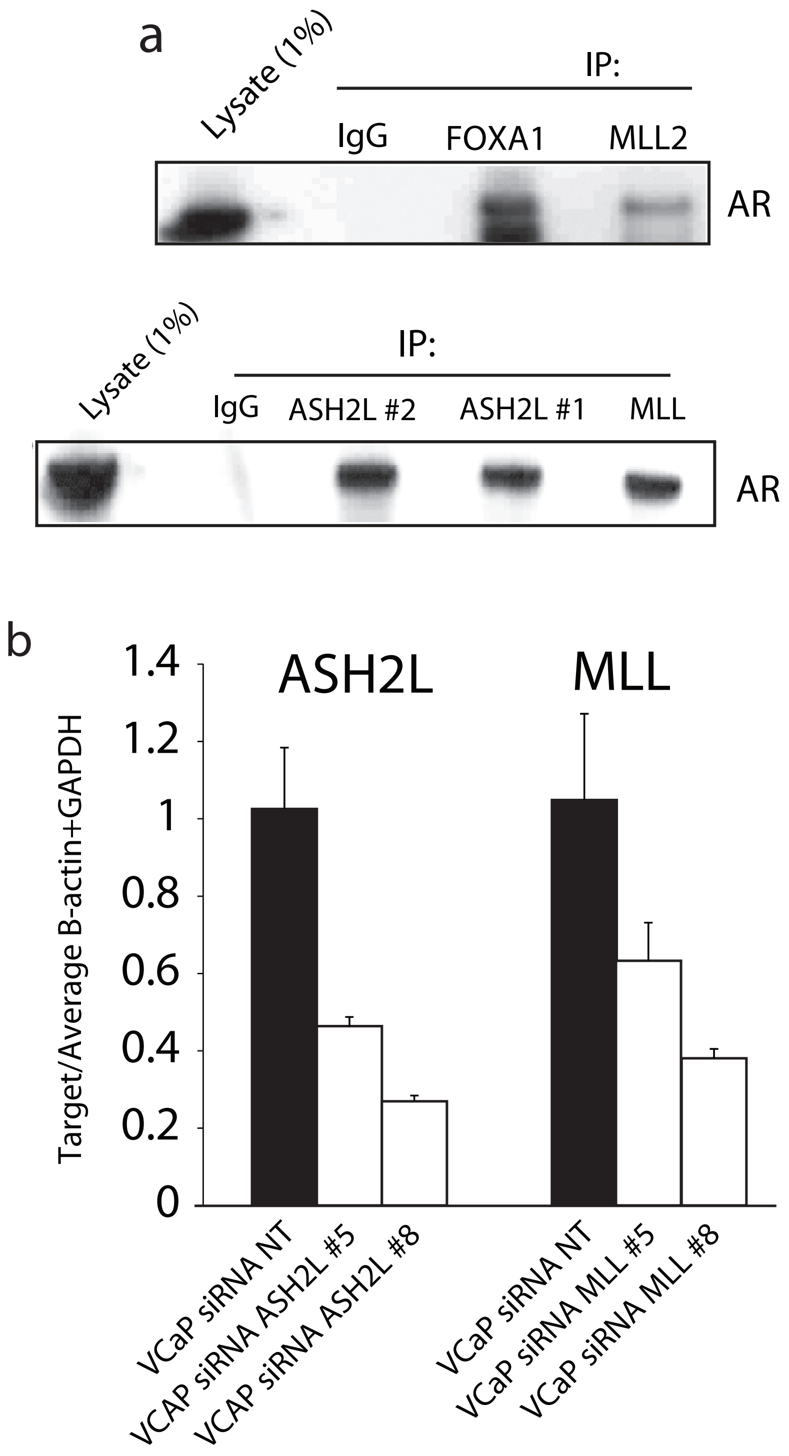
Supplementary Figure 13. Deregulation of genes at 5q21, including *CHD1*, confirmed by matched aCGH and gene expression profiling. Genome wide copy number analysis of high-grade localized prostate cancer and castrate resistant prostate cancer by exome sequencing identified a peak of copy number loss on chr 5q21 centered on *CHD1*. A subset of samples used for exome sequencing, and additional benign prostate tissue (N1-29) and localized prostate cancers were used for matched aCGH and gene expression profiling. **a**. Genome wide analysis by aCGH identified a similar peak of copy number loss on 5q21 (upper panel, sum log2 copy number across all samples plotted) centered on *CHD1*. The expanded view is as in **Figure 2a**, except the area (absolute Log2 ratio) and color intensity (Log 2 ratio; copy number loss in blue) of each box are proportional to binned copy number for that gene according to the legend. ETS$^-$ and ETS$^+$ samples are indicated in black or gray type, respectively. Samples with focal deletions of *CHD1* (*CHD1*$^-$) or other genes within 5q21 (5q21$^-$) by aCGH are indicated with green or red background, respectively, according to the legend. The adjoining plot shows the genome wide copy number plot for T56, which harbors a focal, high level deletion on 5q21 including *CHD1*. **b.** Co-expression of *CHD1* and ETS family members. Heatmap of *CHD1*, ETS genes (*ERG*, *ETV1*, *ETV5*) and *SPINK1* gene expression. Samples are stratified by benign prostate tissue and prostate cancer (including localized and CRPC). ETS and *CHD1* status was determined, with black and green indicating *ETS*$^+$ and *CHD1*$^-$, respectively. **c**. Genome wide copy number plot for T65, which shows focal, high level deletion of 5q21, including *PJA2*, but not *CHD1*. **d.** Expression of *PJA2* stratified by benign prostate tissues (orange), localized prostate cancers (cyan) and CRPCs (black). T65 is indicated in red.



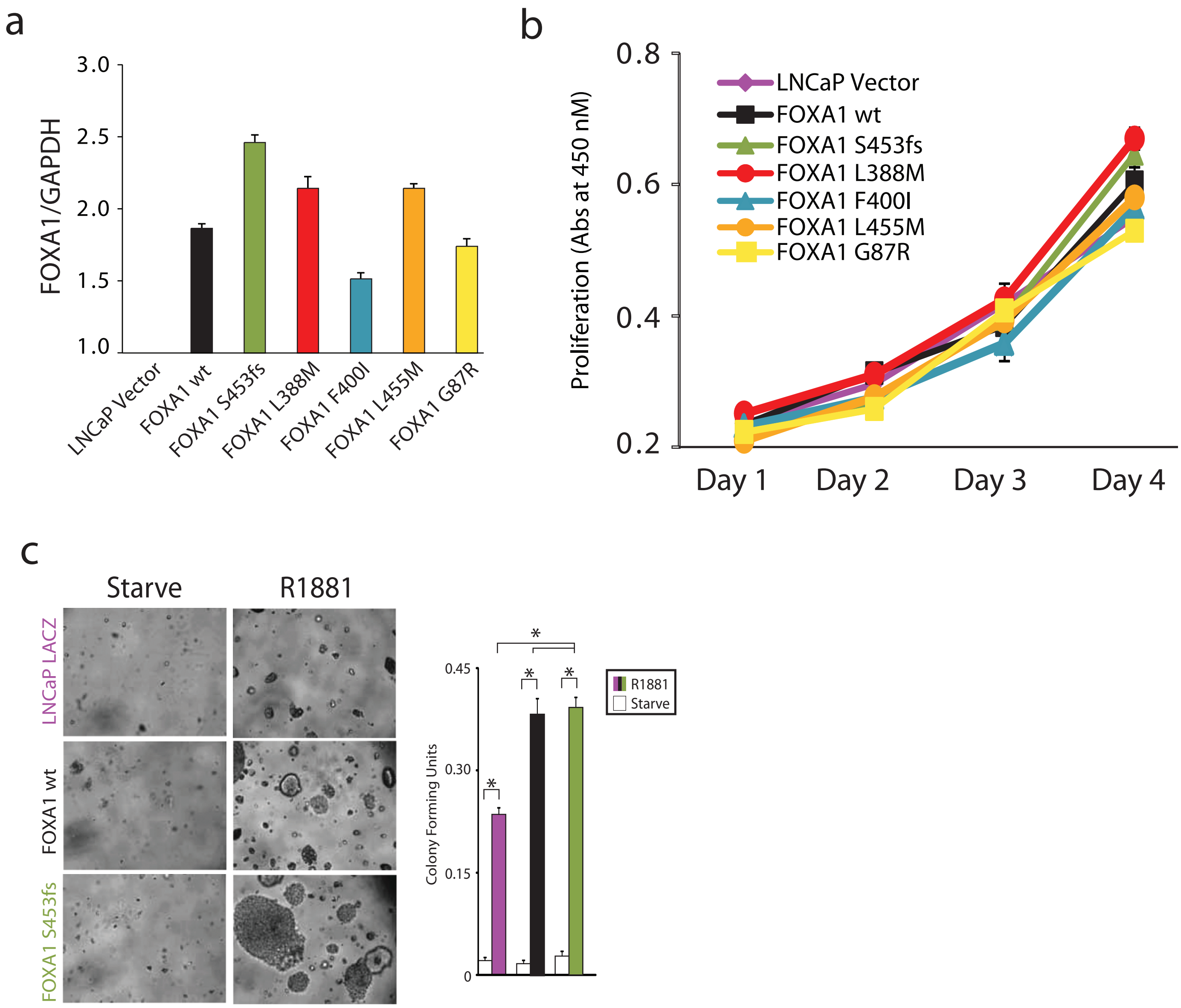
Supplementary Figure 14. ***CHD1*** **deregulation deletion in ETS fusion negative prostate cancer.** Prostate cancer copy number profiling studies (by aCGH) from **a**) Taylor et al.16, **b**) The Cancer Genome Atlas (TCGA) and **c**) Demichelis *et al.*15 were accessed at Oncomine. The summed log2 copy number profile for chr 5 for each study is shown in the upper panels. The expanded views show individual samples as rows, with indicated genes represented by boxes. Samples are stratified by ETS fusion status and those with focal deletions of *CHD1* are shown in bold. The area and color intensity (Log 2 ratio; copy number loss in blue) of each box is proportional to binned copy number for that gene according to the legend. Only samples with at least one gene in the region meeting the indicated Log2 ratio cutoffs (according to the legend) are shown, and missing boxes indicate that gene did not meet the cutoff. **d.** Co-expression of *CHD1* and ETS family members was analyzed in 9 prostate cancer gene expression studies available in Oncomine. Heatmaps of gene expression data from the Lapointe *et al*. and Glinsky *et al*. studies are shown with ETS and *CHD1* expression and annotation.



Supplementary Figure 15. *ETS2* aberrations in exome sequenced samples, and expression in prostate tissue samples and cell lines utilized for in vitro assays. a. Genome wide copy number analysis identified a peak of copy number loss on chr 21, consistent with *TMPRSS2:ERG* fusions through deletion (upper panel, blue bar). The expanded view shows individual samples as rows, with indicated genes represented by boxes. The area and size of each box indicates the copy number call (see legend). Only samples with at least one gene in the region with a called copy number gain/loss are shown, and missing boxes indicate that gene has no called copy number gain/loss. Mutations in *ETS2* are indicated according to the legend. **b.** Gene expression profiles from benign prostate tissues (n=28, orange), localized prostate cancer (PC, *n*=59, cyan), and metastatic castrate resistant prostate cancer (CRPC, *n*=35, black), including samples subjected to exome sequencing, were loaded into Oncomine for automated analysis. Expression of ETS2 is shown, including for sample WA31, which harbors a focal, high copy loss of *ETS2*. **c**. VCaP prostate cancer cells (*ERG*$^{+}$) stably expressing wild type (wt) *ETS2* (black) or *ETS2* R437C (yellow) with N-terminal HA tag, or *LACZ* as control (purple), were generated using lentiviruses (see **Fig. 2**). qPCR for *ETS2* expression was performed for each stable line, and the amount of *ETS2* was normalized to *GAPDH*. Normalized *ETS2* expression is plotted relative to *LACZ* control. Mean of normalized *ETS2* expression + S.E. (*n*=3) are plotted. The inset shows western blotting of the same samples with anti-HA to confirm protein expression.



Supplementary Figure 16. Confirmation of interaction between MLL components and androgen receptor (AR), and siRNA knockdown of *ASH2L* and *MLL*. a.). **a**. Reverse immunoprecipitation using anti-FOXA1 (positive control), an antibody against MLL2, two anti-ASH2L antibodies, an antibody against MLL, or IgG control, with Western blotting for androgen receptor (AR). 1% whole lysate was used as control. **b**. VCaP cells were treated with siRNAs against *ASH2L* or *MLL* (or non-targeting as control). qPCR for *ASH2L* or *MLL* expression (normalized to the average of *ACTB* and *GAPDH*) confirmed knockdown (*n*=3, + S.E.), prior to androgen stimulation experiments (see **Fig. 3b**).



Supplementary Figure 17. Expression of *FOXA1* mutants, proliferation in the absence of androgen and soft agar colony growth. a. Wild type *FOXA1* (wt, black) and *FOXA1* mutants observed in clinical samples were cloned and expressed in LNCaP cells as N-terminal FLAG fusions (empty vector, purple, used as control) through lentiviral infection (see **Fig. 4**). qPCR for *FOXA1* expression was performed for each stable line, and the amount of *FOXA1* was normalized to *GAPDH*. Normalized *FOXA1* expression is plotted relative to vector control. Mean of normalized *FOXA1* expression + S.E. (*n*=3) are plotted. **b.** Cell proliferation in 1% charcoal-dextran stripped serum was measured by WST-1 colorimetric assay (absorbance at 450nM) at the indicated time points. **c**. Soft agar colony forming assays using LNCaP cells stably expressing *LACZ* (control, purple), or N-terminally HA-tagged *FOXA1* wild type (wt, black) or *FOXA1* S453fs (green) generated through lentiviral infection. Representative photographs and quantification of colonies formed in the absence (white) or presence of 1nm R1881 are shown. For **b** and **c**, mean + S.E. (*n*=3) are plotted; * indicates *p*=0.05 from two tailed t-test.