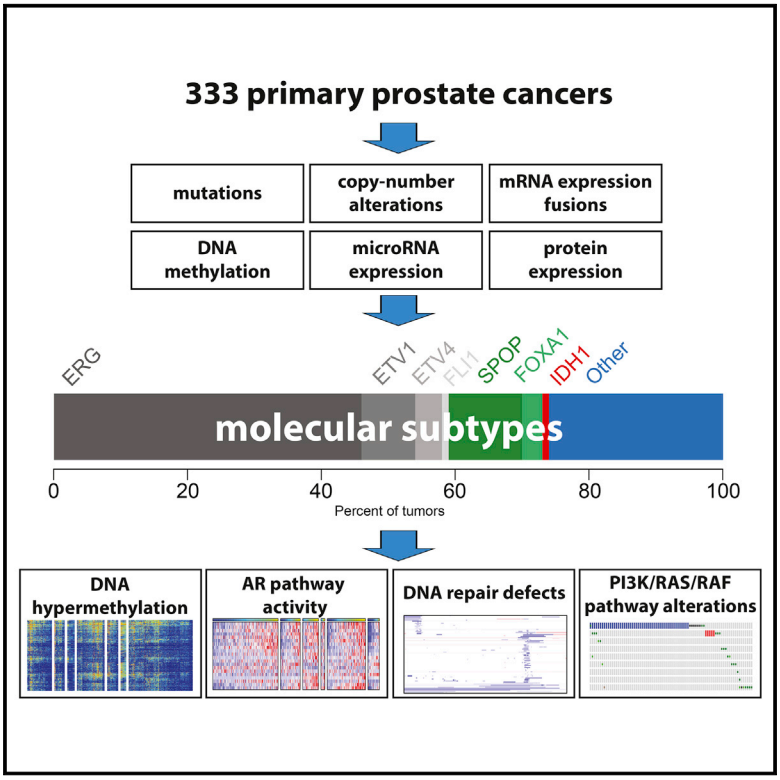
Cell

Resource

The Molecular Taxonomy of Primary Prostate Cancer

Graphical Abstract



Highlights

- Comprehensive molecular analysis of 333 primary prostate carcinomas
- Seven subtypes defined by ETS fusions or mutations in *SPOP*, *FOXA1*, and *IDH1*
- Substantial epigenetic heterogeneity, including a hypermethylated *IDH1* mutant subset
- Presumed actionable lesions in the PI3K, MAPK, and DNA repair pathways

Authors

The Cancer Genome Atlas Research Network

Correspondence

schultz@cbio.mskcc.org (N.S.), massimo_loda@dfci.harvard.edu (M.L.), sander.research@gmail.com (C.S.)

In Brief

Molecular analysis of 333 primary prostate carcinomas reveals substantial heterogeneity and major subtypes among patients, as well as potentially actionable lesions valuable for clinical management of the disease.

The Cancer Genome Atlas Research Network, 2015, Cell *163*, 1011–1025
November 5, 2015
http://dx.doi.org/10.1016/j.cell.2015.10.025

CellPress

Resource

Cell

The Molecular Taxonomy of Primary Prostate Cancer

The Cancer Genome Atlas Research Network[1,*]
[1]The Cancer Genome Atlas Program Office, National Cancer Institute at NIH, 31 Center Drive, Building 31, Suite 3A20, Bethesda, MD 20892, USA
*Correspondence: schultz@cbio.mskcc.org (N.S.), massimo_loda@dfci.harvard.edu (M.L.), sander.research@gmail.com (C.S.)
http://dx.doi.org/10.1016/j.cell.2015.10.025

SUMMARY

There is substantial heterogeneity among primary prostate cancers, evident in the spectrum of molecular abnormalities and its variable clinical course. As part of The Cancer Genome Atlas (TCGA), we present a comprehensive molecular analysis of 333 primary prostate carcinomas. Our results revealed a molecular taxonomy in which 74% of these tumors fell into one of seven subtypes defined by specific gene fusions (*ERG*, *ETV1/4*, and *FLI1*) or mutations (*SPOP*, *FOXA1*, and *IDH1*). Epigenetic profiles showed substantial heterogeneity, including an *IDH1* mutant subset with a methylator phenotype. Androgen receptor (AR) activity varied widely and in a subtype-specific manner, with *SPOP* and *FOXA1* mutant tumors having the highest levels of AR-induced transcripts. 25% of the prostate cancers had a presumed actionable lesion in the PI3K or MAPK signaling pathways, and DNA repair genes were inactivated in 19%. Our analysis reveals molecular heterogeneity among primary prostate cancers, as well as potentially actionable molecular defects.

INTRODUCTION

Prostate cancer is the second most common cancer in men and the fourth most common tumor type worldwide (Ferlay et al., 2013). It is estimated that, in 2015, prostate cancer will be diagnosed in 220,800 men in the United States alone and that 27,540 will die of their disease (Siegel et al., 2015). Multiple genetic and demographic factors, including age, family history, genetic susceptibility, and race, contribute to the high incidence of prostate cancer (Al Olama et al., 2014).

In the current era of prostate-specific antigen (PSA) screening, nearly 90% of prostate cancers are clinically localized at the time of their diagnosis (Penney et al., 2013). The clinical behavior of localized prostate cancer is highly variable—while some men will have aggressive cancer leading to metastasis and death from the disease, many others will have indolent cancers that are cured with initial therapy or may be safely observed. Multiple risk stratification systems have been developed, combining the best currently available clinical and pathological parameters (such as Gleason score, PSA levels, and clinical and pathological staging); however, these tools still do not adequately predict outcome (Cooperberg et al., 2009; D'Amico et al., 1998; Kattan et al., 1998). Further risk stratification using molecular features could potentially help distinguish indolent from aggressive prostate cancer.

Molecular and genetic profiles are increasingly being used to subtype cancers of all types and to guide selection of more precisely targeted therapeutic interventions. Several recent studies have explored the molecular basis of primary prostate cancer and have identified multiple recurrent genomic alterations that include mutations, DNA copy-number changes, rearrangements, and gene fusions (Baca et al., 2013; Barbieri et al., 2012; Berger et al., 2011, Lapointe et al., 2007; Pflueger et al., 2011; Taylor et al., 2010; Tomlins et al., 2007; Wang et al., 2011). The most common alterations in prostate cancer genomes are fusions of androgen-regulated promoters with *ERG* and other members of the E26 transformation-specific (ETS) family of transcription factors. In particular, the *TMPRSS2-ERG* fusion is the most common molecular alteration in prostate cancer (Tomlins et al., 2005), being found in between 40% and 50% of prostate tumor foci, translating to more than 100,000 cases annually in the United States (Tomlins et al., 2009). Nevertheless, among treated prostate cancers, and despite extensive study, affected individuals with fusion-bearing tumors do not appear to have a significantly different prognosis following prostatectomy than those without (Gopalan et al., 2009; Pettersson et al., 2012). Prostate cancers also have varying degrees of DNA copy-number alteration; indolent and low-Gleason tumors have few alterations, whereas more aggressive primary and metastatic tumors have extensive burdens of copy-number alteration genome wide (Taylor et al., 2010; Hieronymus et al., 2014; Lalonde et al., 2014). In contrast, somatic point mutations are less common in prostate cancer than in most other solid tumors. The most frequently mutated genes in primary prostate cancers are *SPOP*, *TP53*, *FOXA1*, and *PTEN* (Barbieri et al., 2012). Only recently has the spectrum of epigenetic changes in prostate cancer genomes been explored (Börno et al., 2012; Friedlander et al., 2012; Kim et al., 2011; Kobayashi et al., 2011; Mahapatra et al., 2012).

Importantly, no studies have comprehensively integrated diverse omics data types to assess the robustness of previously defined subtypes and potentially prognostic alterations. Here, to gain further insight into the molecular-genetic heterogeneity of primary prostate cancer and to establish a molecular taxonomy of the disease for future diagnostic, prognostic, and therapeutic stratification, the TCGA Network has comprehensively characterized 333 primary prostate cancers using seven genomic platforms. This analysis reveals novel molecular features that provide a better understanding of this disease and suggest potential therapeutic strategies.

Table 1. Cohort Characteristics

Clinical Feature	
Age	61 (43–76)
Pre-operative PSA	7.4 (1.6–87.0)
Gleason Score	
3+3	65
3+4	102
4+3	78
≥ 8	88
Tumor Cellularity (pathology)	
<20%	7
21–40%	40
41–60%	84
61–80%	115
81–100%	87
Pathologic Stage	
pT2a/b	18
pT2c	111
pT3a	110
pT3b	82
pT4	6
PSA Recurrence	
Yes	33
No[a]	248
Not available	47
Margin Status	
Positive	69
Negative	193
Not available	71
Ethnicity	
Caucasian	270
African descent	43
Asian	8
Not available	12

[a]Either no evidence of recurrence or insufficient follow-up.

RESULTS

Cohort and Platforms

The cohort of primary prostate cancers analyzed resulted from extensive pathologic, analytical, and quality control review, yielding 333 tumors from 425 available cases. Images of frozen tissue were evaluated by multiple expert genitourinary pathologists, and cases were excluded if no tumor cells were identifiable in the sample or if there was evidence of significant RNA degradation (Figure S1; Supplemental Experimental Procedures). For the subset of cases reviewed by two pathologists, tumor cellularity estimates were within 20% of each other in 71% of cases. In total, 78% of Gleason scores were concordant within one grade of the secondary pattern (Supplemental Experimental Procedures). Moreover, due to the challenge of acquiring primary prostate cancer specimens of high tumor cellularity, we also performed a multi-platform analysis of tumor content, estimating tumor purity with analytical approaches utilizing both DNA (Carter et al., 2012; Prandi et al., 2014) and RNA (Quon et al., 2013; Ahn et al., 2013) sequencing data. The molecular and pathologic estimates are presented in Table S1A and Figure S1. The clinical and pathological characteristics of the final cohort are presented in Table 1. The average follow-up time following radical prostatectomy was just under 2 years, which precluded outcomes analysis due to the long natural history of primary prostate cancer.

We characterized isolated biomolecules from these 333 tumor samples using four platforms: whole-exome sequencing for somatic mutations, array-based methods for profiling both somatic copy-number changes and DNA methylation, and mRNA sequencing. We also performed microRNA (miRNA) sequencing on 330 of these samples, reverse-phase protein array (RPPA) on 152 samples, and low-pass and high-pass whole-genome sequencing (WGS) on 100 and 19 tumor/normal pairs, respectively (Supplemental Experimental Procedures). For 19 samples, non-malignant adjacent prostate samples were also examined for DNA methylation and RNA/miRNA expression analyses.

The Molecular Taxonomy of Primary Prostate Cancer

Previous studies indicate that many genomically distinct subsets of prostate cancer exist. These are driven in some cases by frequent events, such as androgen-regulated fusions of *ERG* and other ETS family members, or recurrent *SPOP* mutations and, in other cases, by less common genomic aberrations. Given the comprehensive nature of our data, we sought to unify these disparate findings to establish a molecular taxonomy of primary disease that integrates results from somatic mutations, gene fusions, somatic copy-number alterations (SCNA), gene expression, and DNA methylation. We first performed unsupervised clustering of data from each molecular platform, as well as integrative clustering using iCluster (Shen et al., 2009) (Figures S2, S3, S4, S5, S6, and S7). These analyses uncovered both known and novel associations, with 74% of all tumors being assignable to one of seven molecular classes based on distinct oncogenic drivers: fusions involving (1) *ERG*, (2) *ETV1*, (3) *ETV4*, or (4) *FLI1* (46%, 8%, 4%, and 1%, respectively); mutations in (5) *SPOP* or (6) *FOXA1*; or (7) *IDH1* mutations (11%, 3%, and 1%, respectively) (Figures 1 and S2 and Table S1A).

In total, 53% of tumors were found to have ETS family gene fusions (*ERG*, *ETV1*, *ETV4*, and *FLI1*) after analysis with two complementary algorithms (Sboner et al., 2010; Wang et al., 2010) (see the Experimental Procedures). While *TMPRSS2* was the most frequent fusion partner in all ETS fusions, we identified fusions with other previously described androgen-regulated 5′ partner genes, including *SLC45A3* and *NDRG1* (Table S1E). We also identified several tumors that overexpressed full-length ETS transcripts that were mutually exclusive with ETS fusions (12 *ETV1* high tumors, 6 *ETV4*, and 2 *FLI1*) (Table S1E). ETS overexpression in these cases could possibly be mediated via epigenetic mechanisms or cryptic translocations of the entire gene locus to a transcriptionally active neighborhood. In the one case with elevated ETV1 full-length expression studied by whole-genome sequencing, we identified a cryptic genomic

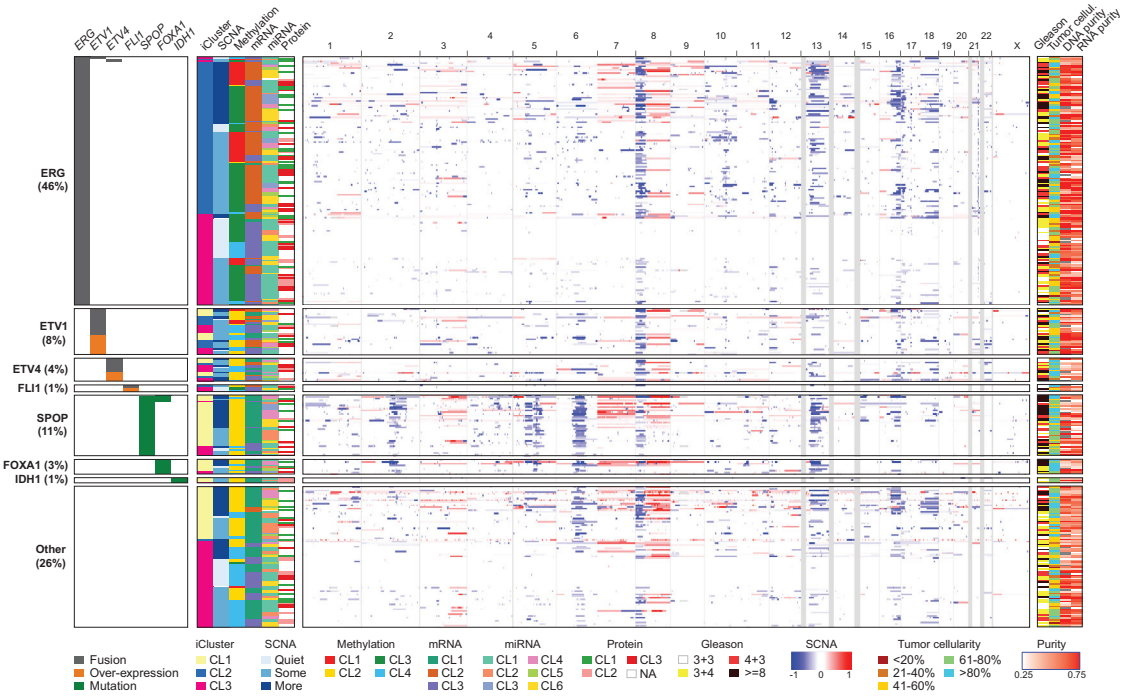


Figure 1. The Molecular Taxonomy of Primary Prostate Cancer

Comprehensive molecular profiling of 333 primary prostate cancer samples revealed seven genomically distinct subtypes, defined (top to bottom) by *ERG* fusions (46%), *ETV1*/*ETV4*/*FLI1* fusions or overexpression (8%, 4%, 1%, respectively), or by *SPOP* (11%), *FOXA1* (3%), and *IDH1* (1%) mutations. A subset of these subtypes was correlated with clusters computationally derived from the individual characterization platforms (somatic copy-number alterations, methylation, mRNA, microRNA, and protein levels from reverse phase protein arrays). The heatmap shows DNA copy-number for all cases, with chromosomes shown from left to right. Regions of loss are indicated by shades of blue, and gains are indicated by shades of red.

See also Figures S1, S2, S3, S4, S5, S6, and S7 and Tables S1A, S1B, S1E, and S2.

rearrangement 3′ of the ETV1 locus with a region on chromosome 14 near the *MIPOL1* gene adjacent to *FOXA1*. This event is similar to previously described *ETV1* translocations in LNCaP and MDA-PCa2b cell lines and in patient samples (Tomlins et al., 2007; Gasi et al., 2011). Overall, while fusions in the four genes were mostly mutually exclusive, three tumors showed evidence for fusions involving more than one of these genes (Table S1E). Given that histologically defined single tumor foci have been shown to be rarely composed of different ETS fusion-positive clones (Cooper et al., 2015; Kunju et al., 2014; Pflueger et al., 2011), it is likely these cases reflect convergent phenotypic evolution in clonally heterogeneous tumors. Tumors defined by *SPOP* mutations were mutually exclusive with all ETS fusion-positive cases, though four of the *SPOP* mutant tumors also possessed *FOXA1* mutations. In all four of these tumors, both the *SPOP* and *FOXA1* mutations were clonal, indicating that they are present in the same tumor cells.

Beyond the class-defining lesions, there were multiple patterns of both known and novel concurrent alterations in key prostate cancer genes. The former included the preponderance of *PTEN* deletions in *ERG* fusion-positive cases (Taylor et al., 2010). Similarly, SPOP mutations have previously been found to occur in ~10% of clinically localized prostate cancers, were mutually exclusive of tumors defined by ETS rearrangements, and may designate a distinct molecular class of disease based primarily on distinctive SCNA profiles (including deletion of *CHD1*, 6q, and 2q) (Barbieri et al., 2012; Blattner et al., 2014). Beyond reaffirming these known patterns, our taxonomy revealed new relationships and subtypes. Specifically, the *SPOP* mutant/*CHD1*-deleted subset of prostate cancers had notable molecular features, including elevated levels of DNA methylation, homogeneous gene expression patterns, and frequent overexpression of *SPINK1* mRNA, supporting *SPOP* mutation as a key feature in the molecular taxonomy of prostate cancer. Interestingly, mRNA, copy-number, and methylation profiles were similar in tumors with *FOXA1* mutations and those with *SPOP* mutations. Furthermore, we identified a new genomically distinct subtype of prostate cancer defined by hotspot mutations in IDH1, described in greater depth below.

Despite this detailed molecular taxonomy of primary prostate cancers, 26% of all tumors studied appeared to be driven by still-occult molecular abnormalities or by one or more frequent alterations that co-occur with the genomically defined classes. Some of these tumors showed a high burden of copy-number alterations or DNA hypermethylation. Enrichment analysis indicated that this subset of tumors was enriched for mutations in *TP53*, *KDM6A*, and *KMT2D*; deletions of chromosomes 6 and 16; and amplifications of chromosomes 8 (spanning *MYC*) and 11 (*CCND1*) (Table S2). To characterize this group further, we performed whole-genome sequencing of 19 tumor specimens and

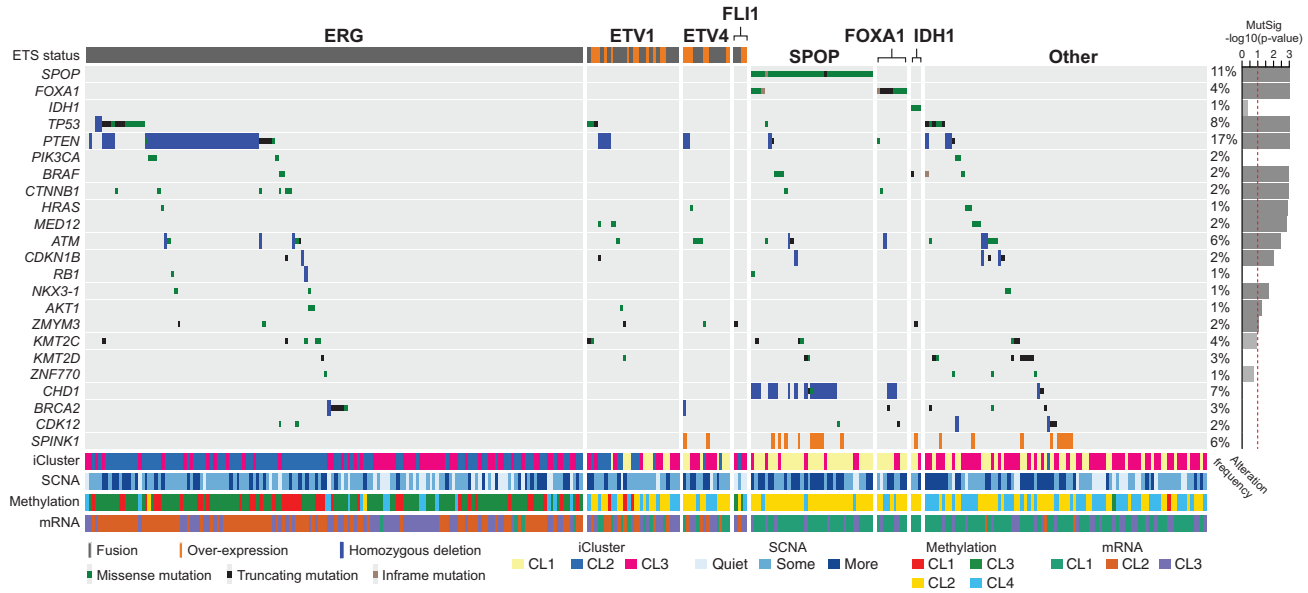


Figure 2. Recurrent Alterations in Primary Prostate Cancer

The spectrum and type of recurrent alterations and genes (mutations, fusions, deletions, and overexpression) in the cohort are shown (left to right) grouped by the molecular subtypes defined in Figure 1. On the right, the statistical significance of individual mutant genes (MutSig q value) is shown. Mutations in *IDH1*, *PIK3CA*, *RB1*, *KMT2D*, *CHD1*, *BRCA2*, and *CDK12* are also shown, despite their not being statistically significant. *SPINK1* overexpression is shown for reference. See also Tables S1B, S1C, S1D, and S1E.

their matched normal tissues, a subset of which had high tumor cellularity but still lacked DNA copy-number alterations or any known or presumed driver lesions. Interestingly, no occult driver abnormalities or highly recurrent regulatory mutations were identified, such as the *TERT* promoter mutation common to many other tumor types (Khurana et al., 2013). Therefore, a significant (up to 26%) subset of primary prostate cancers of both good and poor clinical prognosis (including those with Gleason scores of >8) is driven by as-yet-unexplained molecular alterations.

mRNA clusters were tightly correlated with ETS fusion status, where mRNA cluster 1 consisted primarily of ETS-negative tumors and mRNA clusters 2 and 3 were split among ETS fusion-positive tumors (Figures 1 and S4). miRNA clustering showed a similar pattern, revealing a general difference in miRNA expression between ETS-positive and -negative tumors (Figures 1 and S6). Clustering of RPPA data identified three distinct subgroups, with cluster 3 exhibiting elevated PI3K/AKT, MAP-kinase, and receptor tyrosine kinase activity (Figure S7A). The cluster was not enriched, however, in genomic alterations in these pathways, and in general, there was little correlation of increased pathway activity (as measured by phospho-AKT and other downstream phospho-proteins) with the frequent genomic alterations in the pathways (see the example of *PTEN* deletions in Figure S7B).

Recurrently Altered Genes and Their Patterns across Subtypes

The overall mutational burden of the cohort, inferred from whole-exome sequencing, was 0.94 mutations per megabase (median, range 0.04–28 per megabase), which corresponds to 19 non-synonymous mutations per tumor genome (median; 13–25, 25th and 75th, percentiles respectively). This is consistent with prior exome and genome-scale sequencing results for localized prostate cancers (Barbieri et al., 2012; Baca et al., 2013) and is lower than the mutational burden of metastatic prostate cancers (Gundem et al., 2015; Grasso et al., 2012; Robinson et al., 2015). These results reaffirm that prostate cancer possesses a lower mutational burden than many other epithelial tumor types that are not associated with a strong exogenous mutagen (Alexandrov et al., 2013; Lawrence et al., 2013). Prior exome sequencing of 112 prostate cancers identified 12 recurrently mutated genes through focused assessment of point mutations and short insertions and deletions (Barbieri et al., 2012). By comparison, mutational significance analysis of these 333 tumor-normal pairs by MutSigCV (Lawrence et al., 2013, 2014) identified 13 significantly mutated genes (q value < 0.1), seven of which had not been previously identified (Figure 2 and Tables S1B and S1C). Among the significantly mutated genes, *SPOP*, *TP53*, *FOXA1*, *PTEN*, *MED12*, and *CDKN1B* were previously identified as recurrently mutated. Additional clinically relevant genes were identified with lower mutation frequencies; these included genes within canonical kinase signaling pathways (*BRAF*, *HRAS*, *AKT1*), the beta-catenin pathway (*CTNNB1*), and the DNA repair pathway (*ATM*). The rate of *BRAF* mutations (2.4%) seen in this study is higher than previously reported; these include several known activating mutations but, curiously, not the canonical V600E hotspot. We identified no *BRAF* fusions, which had previously been reported in a subset of clinically advanced prostate cancer (Palanisamy et al., 2010). *NKX3-1*, previously implicated in familial prostate cancer syndromes and often found to be deleted, was also somatically mutated in this cohort (1% of tumors). While its functional significance is unknown, *ZMYM3*, an epigenetic regulatory protein not previously

implicated in prostate cancer but infrequently mutated in Ewing sarcomas (Tirode et al., 2014) and various pediatric cancers (Huether et al., 2014), was also recurrently mutated (2% of tumors). Genes with known biological relevance that were mutated at frequencies just below the threshold of significance (q value < 0.01) included *KMT2C* (*MLL3*), *KMT2D* (*MLL4*), *APC*, *IDH1*, and *PIK3CA* (Figure 2 and Tables S1B and S1C). Mutations in the tumor suppressor genes *KMT2C*, *KMT2D*, and *APC* were mostly truncating; the *IDH1* and *PIK3CA* mutations occurred in previously characterized hotspots and thus may have therapeutic relevance for those occasional tumors with these mutations.

Notwithstanding these key somatic mutations, the most frequent molecular abnormalities involved chromosomal arm-level copy-number alterations (Taylor et al., 2010). These alterations included recurrent genomic gains of chromosome 7 and 8q and heterozygous losses of 8p, 13q, 16q, and 18 (Figure S3A). Significance analysis of recurrent focal DNA copy-number alterations revealed 20 amplifications and 35 deletions (q value < 0.25, GISTIC 2.0; Figure S3A and Table S1D). Recurrent focal amplifications included those spanning known oncogenes such as *CCND1* (11q13.2, 2%), *MYC* (8q24.21, 8%), and *FGFR1* and *WHSC1L1* (8p11.23, 8%). Recurrent focal deletions were much more common. Homozygous deletions spanning the *PTEN* locus occurred at one of the highest rates of any tumor type studied thus far (15%). Focal deletions of the region between the *TMPRSS2* and *ERG* genes on 21q22.3, which result in TMPRSS2-ERG fusions, were unique to prostate cancers, as expected. Other focal deletions include those spanning tumor suppressors *TP53* (17p13.1), *CDKN1B* (12p13.1), and *MAP3K1* (5q11.2), *FANCD2* (3p26), as well as *SPOPL* (2q22.1) and the complex locus spanning *FOXP1/RYBP/SHQ1* (3p13). *MAP3K7* (6q.12–22) was also frequently deleted, along with deletion of *CHD1* (5q15–q21); co-deletion of these loci has been associated with aggressive ETS-negative prostate cancer (Kluth et al., 2013; Rodrigues et al., 2015).

As the pattern and extent of SCNAs in prostate cancer genomes have been associated with probability of disease recurrence and metastasis in primary prostate cancers (Taylor et al., 2010; Hieronymus et al., 2014; van Dekken et al., 2004; Paris et al., 2004), we sought to identify similar structure in the burden of SCNAs by performing hierarchical clustering of arm-level alterations. We identified three major groups of prostate cancers, one with mostly unaltered genomes (hereafter referred to as *quiet*), a second group encompassing 50% of all tumors with an intermediate level of SCNAs, and a third group with a high burden of arm level genomic gains and losses (Figures S3B and S3C). While a formal outcome analysis was not possible due to the limited clinical follow-up available for this cohort, the subset of tumors with the greatest burden of SCNAs had significantly higher Gleason scores and PSA levels than the other two groups (Figures S3B–S3D). The tumors in this group also had significantly higher tumor cellularity (Figure S3C).

Epigenetic Changes Define Molecularly Distinct Subtypes of Prostate Cancer

Integrative analysis of genetic and epigenetic changes revealed a diversity of DNA methylation changes that defined molecularly distinct subsets of primary prostate cancer (Figure 3). Unsupervised hierarchical clustering of the most variably hypermethylated CpGs identified four epigenetically distinct groups of prostate cancers (Figures S5A and S5B). When integrated with the molecular taxonomy defined above, we found a number of striking associations. Among these was a notable pattern within *ERG* fusion-positive tumors. Specifically, while nearly two-thirds of all *ERG* fusion-positive tumors belonged to an unsupervised cluster with only moderately elevated DNA methylation (DNA methylation cluster 3), the remaining *ERG* fusion-positive tumors comprised a distinct hypermethylated cluster (cluster 1) that was almost exclusively associated with *ERG* fusions. On average, this cluster contained twice the number of hypermethylated loci as DNA methylation cluster 3 (Figure S5A), and the epigenetic patterns were largely distinct from those of *ETV1* and *ETV4* fusion-positive tumors, which showed more heterogeneous methylation. What drives these epigenetically distinct groups of ETS fusion-positive tumors is unknown, but there is considerable diversity in their DNA methylation profiles that may reflect altered epigenetic silencing (Figures S5A and S5B). Together, these results support further ETS fusion-based subtyping of disease but also reveal a greater molecular and likely biological diversity among *ERG* fusion-positive tumors than previously appreciated. Likewise, these results are consistent with in vivo mouse modeling and expression profiling studies that suggest important molecular and clinicopathological differences between *ERG* and non-*ERG* ETS fusion-positive tumors (Baena et al., 2013; Tomlins et al., 2015).

SPOP and *FOXA1* mutant tumors exhibited homogeneous epigenetic profiles. These tumors belonged almost exclusively to DNA methylation cluster 2, a group that also contained a majority of the *ETV1* and *ETV4* but not *ERG*-positive tumors. Lastly, the *IDH1* mutant tumors were notable given their strongly elevated levels of genome-wide DNA hypermethylation (Figure S5B). While of low incidence, these *IDH1* R132 mutant tumors defined a distinct subgroup of what appears to be early-onset prostate cancer (Figure 3B) that possesses fewer DNA copy-number alterations (see Figure 1) or other canonical genomic lesions that are common to most other prostate cancers. *IDH1* and *IDH2* mutations have been associated with a DNA methylation phenotype in other tumor types, most notably in gliomas (Noushmehr et al., 2010) and acute myeloid leukemias (AML, Figueroa et al., 2010). Curiously, *IDH1* mutant prostate cancers possessed even greater levels of genome-wide hypermethylation than either glioma or AML *IDH1* mutant tumors (Figure 3B). After further investigating DNA methylation differences between IDH mutant and wild-type tumors among prostate cancers, gliomas, and AMLs, we found that hypermethylated loci were specific to the cancer type rather than IDH mutants (Figure S5F).

Integrating these epigenetic data with mRNA expression levels, we identified 164 genes that were epigenetically silenced in subsets of the cohort (Figure S5C and Table S1F). These silenced genes were significantly enriched for genes previously found to be differentially expressed in prostate cancer—specifically, genes that are downregulated in metastatic prostate cancer (Chandran et al., 2007) and genes involved in prostate organ development (Schaeffer et al., 2008) (q value $< 2.0 \times 10^{-5}$). These 164 silenced genes displayed heterogeneous frequencies of

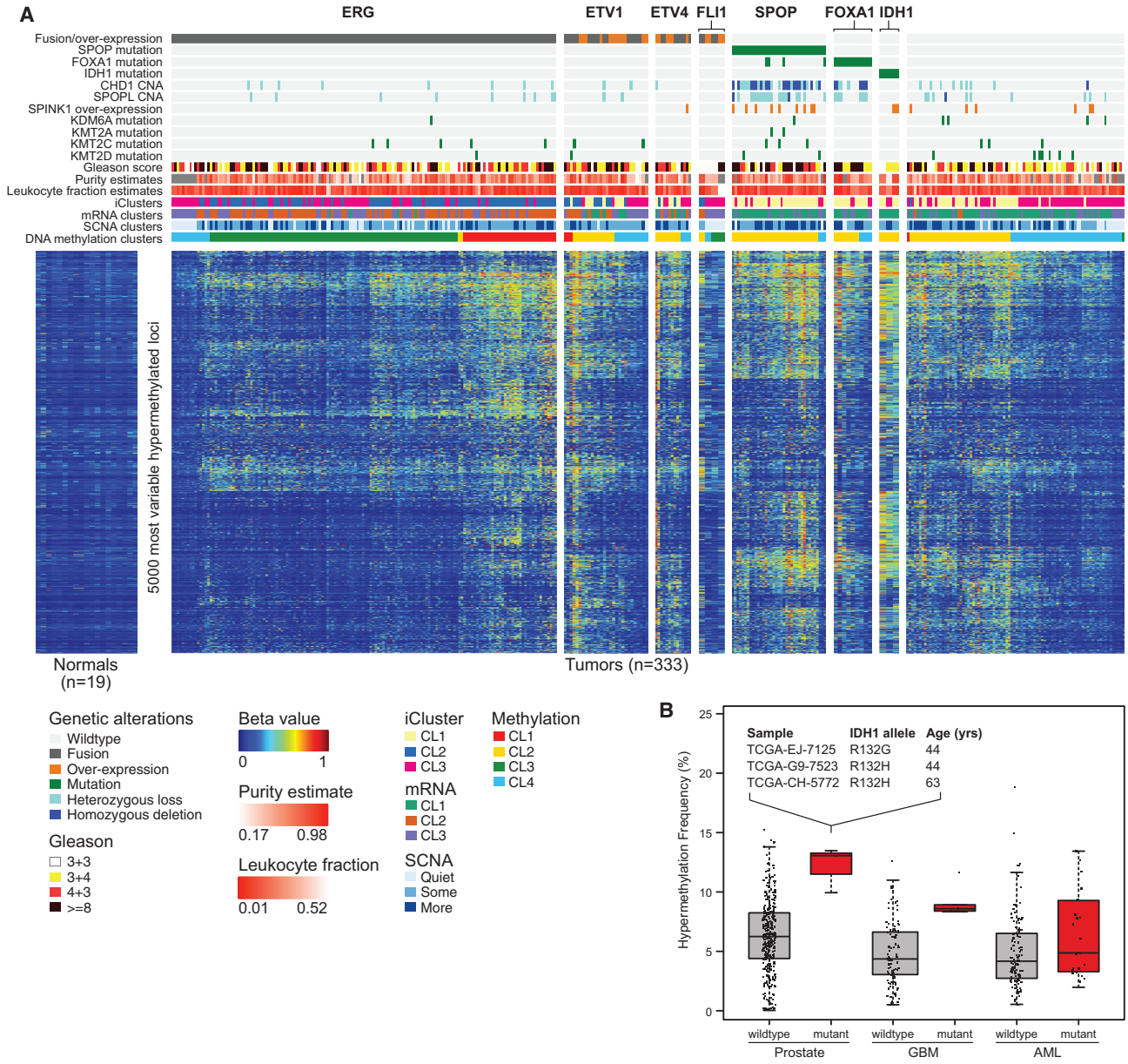


Figure 3. Hypermethylation Is Common across Primary Prostate Cancer

(A) Primary prostate cancers show diverse methylation changes compared to normal prostate samples (left). Unsupervised clustering was performed on the beta-values of the 5,000 most hypermethylated loci, and the results mapped to the genomic subtypes. *ERG*-positive tumors had a high diversity of methylation changes, with a distinct subgroup (cluster 1) nearly unique to this group. *SPOP* and *FOXA1* mutant tumors also exhibited global hypermethylation.

(B) *IDH1* mutant prostate cancers, which are associated with younger age, are among the most hypermethylated tumors, as in glioblastoma (GBM) and AML. See also Figure S4 and Table S1F.

epigenetic silencing across the cohort. For example, *SHF*, *FAXDC2*, *GSTP1*, *ZNF154*, and *KLF8* were epigenetically silenced in almost all tumors (>85%) whereas *STAT6* was silenced predominantly in ETS fusion-positive tumors and not in *SPOP* and *IDH1* mutant tumors. Conversely, *HEXA* was silenced preferentially in *SPOP* mutant tumors compared to *ERG* fusion-positive tumors (86.5 versus 14.5%, respectively, $p < 5.4 \times 10^{-15}$). Consistent with their increased DNA hypermethylation, the *IDH1* mutant prostate tumors also possessed the greatest number of epigenetically silenced genes among all prostate tumors (Table S1F).

AR Activity Is Variable in Primary Prostate Cancers

The androgen receptor (AR) regulates normal prostate development, as well as critical growth and survival programs in prostate carcinoma. Primary prostate cancer is androgen dependent, and

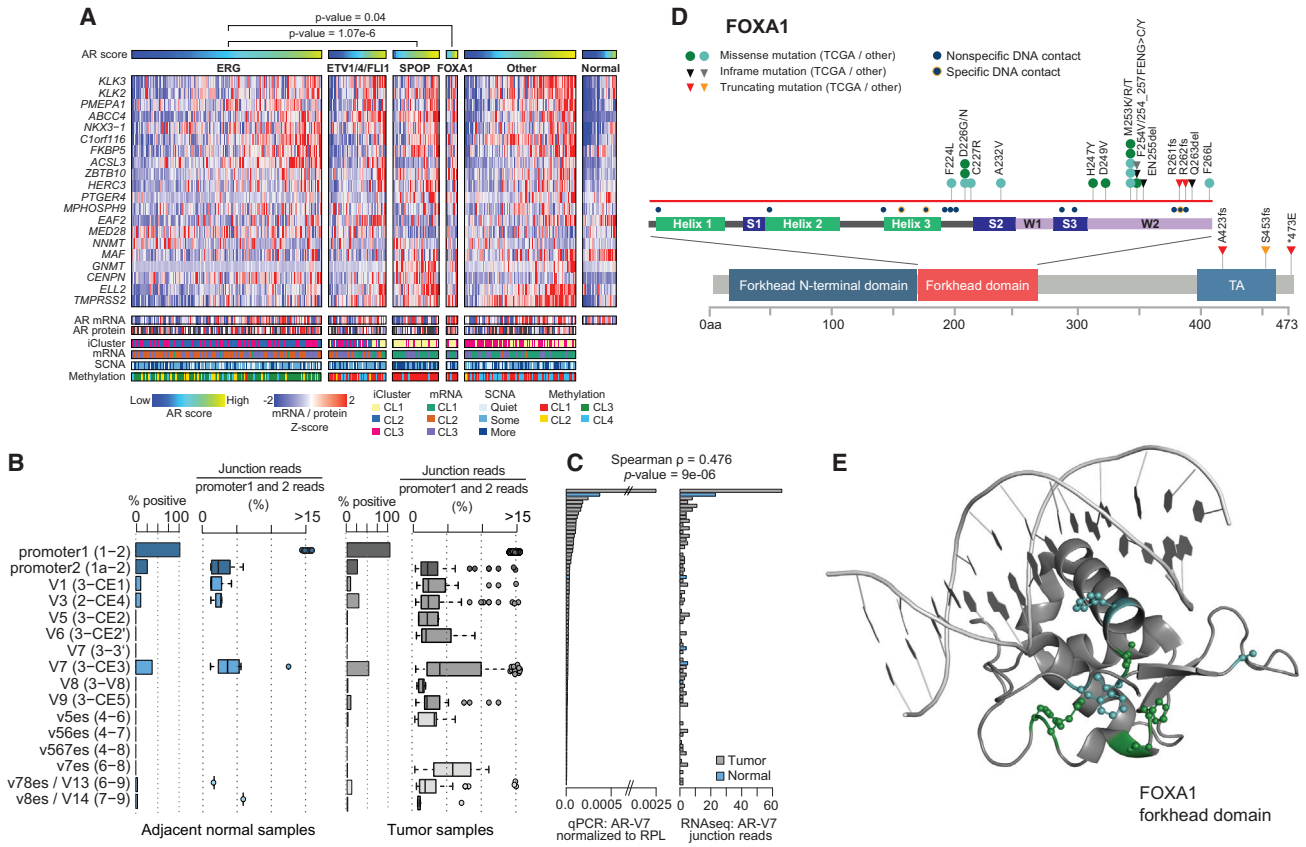


Figure 4. The Diversity of Androgen Receptor Activity in Primary Prostate Cancer

(A) Androgen receptor activity, as inferred by the induction of *AR* target genes, was significantly increased in *SPOP* and *FOXA1* mutant tumors when compared to normal prostate or *ERG*-positive tumors. This increase in activity cannot be fully explained by AR mRNA or protein levels.

(B) Multiple known *AR* splice variants were detected in benign prostate (left) and primary prostate cancer (right), with the AR-V7 variant detected in 50% of tumors.

(C) Real-time qPCR comparison of AR-V7 in 74 tumor samples (gray) and 5 adjacent-normal samples (blue).

(D and E) (D) *FOXA1* missense mutations were clustered in the forkhead domain, mostly in residues that do not form contacts with DNA (see also the 3D structure in panel E).

androgen activity is a central axis in prostate cancer pathogenesis, driving the creation and overexpression of most ETS fusion genes (Lin et al., 2009; Mani et al., 2009; Tomlins et al., 2005). However, the extent to which individual primary prostate cancers differ in androgen sensitivity or dependence is unknown, and the issue has translational implications because AR targeting is therapeutically important. To address these questions, we sought to infer the AR output of tumors by calculating an AR activity score from the expression pattern of 20 genes that are experimentally validated AR transcriptional targets (Hieronymus et al., 2006). This score suggested that a broad spectrum of AR activity exists across all prostate tumors, as well as between genomic subtypes (Figure 4A). Although ETS fusion genes are under AR control, the ETS fusion-positive groups had variable AR transcriptional activity. In contrast, we found that tumors with *SPOP* or *FOXA1* mutations had the highest AR transcriptional activity of all genotypically distinct subsets of prostate cancer ($p = 1.1 \times 10^{-6}$ and 0.04, respectively, t test). Consistent with this, *SPOP* mutations have been previously implicated in androgen signaling in model systems, since both AR and AR coactivators are substrates deregulated by *SPOP* mutation (Geng et al., 2013; An et al., 2014; Geng et al., 2014), providing a possible explanation for the associated increase in AR activity seen in this subtype of prostate cancers.

While AR transcriptional output is a proxy for ligand-driven AR activity in many tumors, *AR* transcript variants have been described that encode truncated AR proteins that lack the ligand-binding domain and hence are capable of activating AR target genes in the absence of androgens (Dehm et al., 2008; Watson et al., 2010). Using RNA sequencing reads that spanned the splice junctions unique to each *AR* variant, we quantified the expression of these *AR* transcript variants. This analysis revealed that several *AR* splice variants, most notably AR-V7, can be detected at low levels in primary tumors and, in a few cases, in adjacent benign prostate tissue (Figure 4B), and we validated these expression levels with qPCR (Figure 4C). However, their expression was not associated with differential expression of known AR target genes or with the seven previously defined genomic subtypes. Most detected splice forms were truncated after the DNA-binding domain by the presence

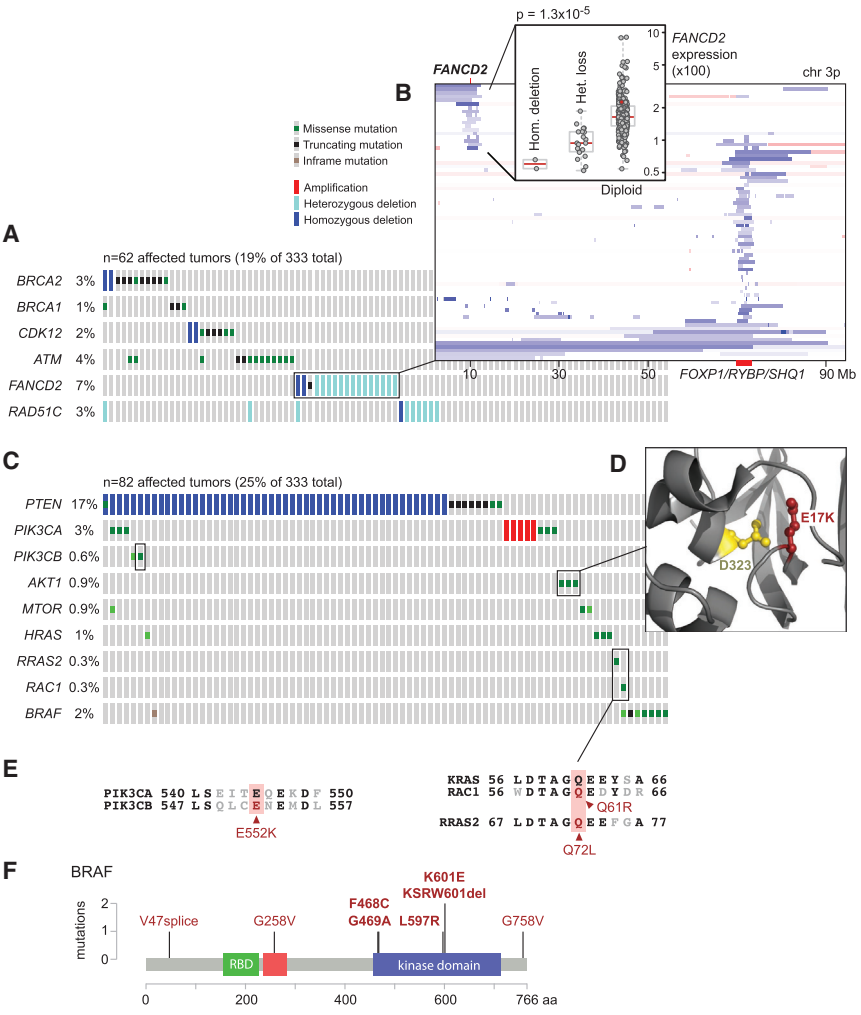


Figure 5. Alterations in Clinically Relevant Pathways

(A) Alterations in DNA repair genes were common in primary prostate cancer, affecting almost 20% of samples through mutations or deletions in *BRCA2*, *BRCA1*, *CDK12*, *ATM*, *FANCD2*, or RAD51C.

(B) Focal deletions of *FANCD2* were found in 7% of samples and were associated with reduced mRNA expression of *FANCD2*.

(C) The RAS or PI-3-Kinase pathways were altered in about a quarter of tumors, mostly through deletion or mutation of *PTEN*, but also through rare mutations in other pathway members.

(D) *AKT1* mutations were found in three samples. Two of them were the known activating E17K, and the third one affected the D323 residue, which is adjacent to E17 in the protein structure.

(E) One of the observed *PIK3CB* mutations, E552K, is paralogous to the known activating E545K mutation in *PIK3CA*, and the *RAC1* Q61 and *RRAS2* Q72 mutations are paralogous to the Q61 mutations in KRAS.

(F) BRAF mutations were found in 2% of samples, mostly in known non-V600E hotspots in the kinase domain.

See also Figure S3.

of a cryptic exon rather than by skipping those exons encoding the ligand-binding domain. Truncated *AR* splice variants were previously assumed to be expressed primarily in metastatic castration-resistant prostate cancers, where, at least for AR-V7, their presence was associated with resistance to hormone therapy (Antonarakis et al., 2014). Hence, our finding that they are expressed in hormone-naive primary prostate cancers is notable.

In prostate cancers, the degree of AR pathway output is controlled not only by AR mRNA and protein expression levels, but also by expression of and mutations in AR cofactors (Heemers and Tindall, 2007). It is therefore notable that *FOXA1* was recurrently mutated in our cohort, as it is a pioneering transcription factor that targets AR and has a demonstrated role in prostate cancer oncogenesis (Jin et al., 2013). We identified *FOXA1* mutations in 4% of the primary prostate cancers studied here, which is similar to the mutation frequency observed previously (Barbieri et al., 2012; Grasso et al., 2012) (Figure 4A). While a subset of these mutations was present in tumors that also possessed *SPOP* mutations and had elevated levels of AR output, *FOXA1* mutations were mutually exclusive with all other alterations that define the genomic subclasses described here. While there were some truncating mutations near the C terminus and the C-terminal part of the forkhead domain, the majority of the mutations found here and in other prostate cancer cohorts were missense mutations that primarily affect the winged-helix DNA binding domain of FOXA1. Curiously, these mutations do not directly alter FOXA1 DNA-binding residues (Figures 4D and 4E), a pattern similar to the *FOXA1* mutations recently found in lobular breast cancers (TCGA, unpublished data), which suggests that the impact of *FOXA1* mutations has less to do with altering DNA binding than with disrupting or altering interactions with other chromatin-bound cofactors.

Clinically Actionable DNA Repair Defects in Primary Prostate Cancers

Prior data indicate that several DNA repair pathways are disrupted in a subset of prostate cancers (Karanika et al., 2014; Pritchard et al., 2014). Moreover, the PARP inhibitor olaparib is effective in some patients with prostate cancer (Mateo et al., 2014). Here, we found inactivation of several DNA repair genes that collectively affected 19% of affected individuals (Figure 5A). While we found only one inactivating *BRCA1* germline mutation, a frameshift at V923 caused by a 4 bp deletion (Clinvar RCV000083190.3), *BRCA2* inactivation affected 3% of tumors, including both germline and somatic truncating mutations. All six *BRCA2* germline mutations were K3326*, a C-terminal truncating mutation with debated functional impact but increased prevalence in several tumor types (Farrugia et al., 2008; Martin

et al., 2005; Delahaye-Sourdeix et al., 2015). Two additional tumors possessed focal *BRCA2* homozygous deletions that were accompanied by very low *BRCA2* transcript expression. Four tumors (1%) possessed either loss-of-function mutations or homozygous deletion of *CDK12*, a gene that has been implicated in DNA repair by regulating expression levels of several DNA damage response genes (Blazek et al., 2011) and is recurrently mutated in metastatic prostate cancer (Grasso et al., 2012). *ATM*, an apical kinase of the DNA damage response, which is activated by the Mre11 complex and mediates downstream checkpoint signaling, was affected by a nonsense mutation in one case and by a likely kinase-dead hotspot N2875 mutation in two cases. *FANCD2* was similarly affected by diverse uncommon lesions, including a truncating mutation in one tumor, homozygous deletion in two tumors, and focal heterozygous losses in 6% of the cohort (Figure 5B). *RAD51C* (3%) was affected by focal DNA losses, most of which were heterozygous. Finally, it was notable that heterozygous losses of *BRCA2* (13q13.1) almost always coincided with concurrent loss of the distant *RB1* tumor suppressor gene (13q14.2) (Figure S3D). The observation that nearly 20% of primary prostate cancers bear genomic defects involving DNA repair pathways is remarkably consistent with the recently announced TOPARP-A Phase II trial results in patients with metastatic castration-resistant prostate cancer, indicating that clinical responses to the PARP inhibitor olaparib likely occurred in the subgroup of tumors bearing defects in DNA repair genes (Mateo et al., 2014; Robinson et al., 2015).

Clinically Actionable Lesions in PI3K and Ras Signaling

The long tail of the frequency distribution of molecular abnormalities is particularly notable among primary prostate cancers. Beyond *PTEN*, which was deleted or mutated in 17% of the cohort, various driver mutations in effectors of PI3K signaling were present at low incidence (Figure 5C). *PIK3CA*, which encodes the 110 kDa catalytic subunit of phosphatidylinositol 3-kinase, was mutated in six tumors, including one case possessing coincident activating mutations (E542A and N345I), both of which appeared to be subclonal. The other four *PIK3CA* mutations were all known activating mutational hotspots (E545K, Q546K, N345I, and C420R), while one had a mutation of unknown function (E474A). Focal *PIK3CA* amplification with associated mRNA overexpression occurred in ~1% of cases. Interestingly, *PIK3CB* was mutated in two tumors that also possessed coincident homozygous deletions of *PTEN*, both of which were clonal. *PIK3CB* E552K was found in one tumor at a paralogous residue to the canonical *PIK3CA* helical domain E545K mutant and is presumably activating (Figure 5E). As *PTEN*-deleted tumors are likely *PIK3CB*-dependent due to the feedback inhibition of *PIK3CA*, co-existent loss and mutation of *PTEN* and *PIK3CB* may be elevating PI3K pathway output and perhaps indicating a set of tumors in which combined PI3K and androgen signaling inhibition may be effective (Schwartz et al., 2015). Among other lesions that drive PI3K signaling, *AKT1* was mutated in three tumors. Two tumors had the known E17K hotspot mutation, while another encoded a D323Y mutation. Whereas E17K is the most common hotspot in *AKT1* across human cancer, the D323Y variant is uncommon, having been identified previously in one lung adenocarcinoma (Cancer Genome Atlas Research Network, 2014) and one urothelial bladder cancer (Guo et al., 2013). Nevertheless, while distant linearly from the activating E17K hotspot, in three dimensions, this D323Y kinase domain mutant directly abuts the PH-domain containing E17K (Figure 5D) and has been described as potentially activating (Parikh et al., 2012).

We also identified known or presumed driver mutations in several other genes of the MAPK pathway, affecting 25% of the tumors (Figure 5A). *HRAS* was mutated in four tumors, of which three were Q61R hotspot mutations. Two mutations arose in other Ras family small GTPases. While both *RAC1* Q61R and *RRAS2* Q72L occurred only once each, they affected residues paralogous to the RAS Q61 hotspot (Figure 5E) (Chang et al., 2015). We also identified eight *BRAF* mutations, though, curiously, none were the common V600E mutation that is prevalent in cutaneous melanomas, thyroid cancers, and many other tumor types. Five *BRAF* mutations are likely activating, including known hotspots (K601E, G469A, L597R), two of which confer sensitivity to MEK inhibitors (Dahlman et al., 2012; Bowyer et al., 2014). Another mutation was a likely activating in-frame 3 amino acid deletion at K601 (Figure 5F), while the final mutation (F468C) affected the adjacent residue to the known G469 hotspot. Together, these findings reveal a long tail of low-incidence potentially actionable predicted driver mutations present across the molecular taxonomy of prostate cancer.

Comparison with Metastatic Prostate Cancer

To put these results in context, we compared our findings with those from a recently published cohort of 150 castration-resistant metastatic prostate cancer samples (Robinson et al., 2015). The analysis revealed some similarities and many differences between primary and treated metastatic disease. Although the overall burden of copy-number alterations and mutations was significantly higher in the metastatic samples (Figure 6A), consistent with previous findings (Taylor et al., 2010; Grasso et al., 2012), the primary and metastatic samples were remarkably similar in their subtype distribution, with the exception that the metastatic dataset contained no *IDH1* mutant tumors (Figure 6B). We compared the frequencies of all recurrently altered genes described in both studies and found that, similar to the overall burden of genomic alterations (Figure 6A), many genes and pathways have increased alteration rates in the metastatic samples (Figure 6C and Table S3). Androgen receptor signaling was more frequently altered in the metastatic samples, most often by amplification or mutation of *AR*, events that were essentially absent in primary samples. Interestingly, *SPOP* mutations were somewhat less frequent in the metastatic samples (8% versus 11% in the primary samples). DNA repair and PI3K pathway alterations were more frequent in the metastatic samples, as were mutations or deletions of *TP53*, *RB1*, *KMT2C*, and *KMT2D*. Interestingly, we found no focal, clonal *MYCL* amplifications, which were recently described in primary prostate cancer (Boutros et al., 2015), in either dataset nor in a separate set of 63 untreated prostate cancer samples (Hovelson et al., 2015).

DISCUSSION

The comprehensive molecular analyses of primary prostate cancers presented here reveal highly diverse genomic, epigenomic,

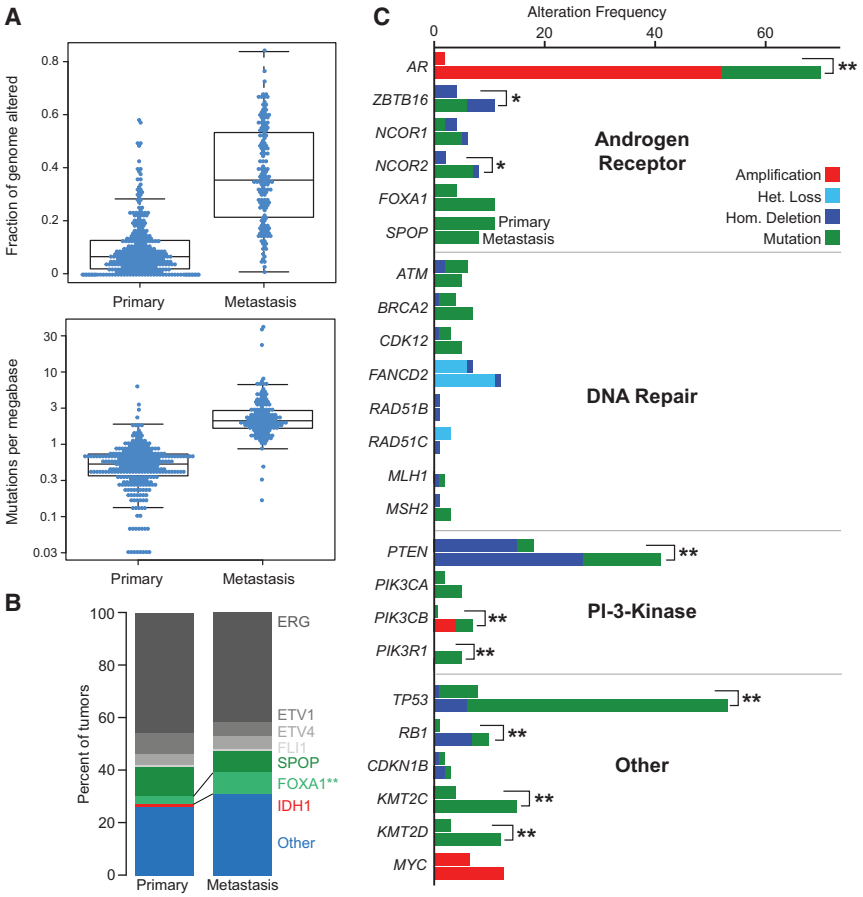


Figure 6. Comparison of Primary with Metastatic Prostate Cancer

(A) Metastatic prostate cancer samples have more copy-number alterations (top, measured as fraction of genome altered) and mutations (bottom).

(B) The relative distribution of main subtypes (*ERG*, *ETV1/4*, *FLI1*, *SPOP*, *FOXA1*, *IDH1*, other) is similar in primary and metastatic samples.

(C) The alteration frequencies of several genes and pathways are higher in metastatic samples. The upper bar for each gene indicates the alteration frequency in primary samples, the lower bar for metastatic samples. The most notable differences in alteration frequencies involve the Androgen Receptor pathway, the PI3K pathway, and *TP53*. See also Table S3.

and transcriptomic patterns. Major subtypes could be defined by fusions of the ETS family genes *ERG*, *ETV*, *ETV4*, or *FLI1* and by mutations in *SPOP*, *FOXA1*, or *IDH1*. However, even within the groups, there was significant diversity in DNA copy-number alterations, gene expression, and DNA methylation. The mutational heterogeneity mirrors the heterogeneous natural history of primary prostate cancers.

Although the broad spectrum of copy-number alterations in tumors with ETS fusions has been previously characterized (Demichelis et al., 2009; Taylor et al., 2010), here we uncovered additional differences between the epigenetic profiles of those tumors. We found that *ERG* fusion-positive tumors can be subdivided into two methylation subtypes: one with lower levels of methylation, and one with a distinct spectrum of hypermethylation. Many genes were epigenetically silenced as a result of the hypermethylation in the latter tumors. While further studies will be required to determine which silencing events are linked to prostate cancer pathogenesis, the findings presented here reveal variability among what was previously considered to be genetically homogeneous prostate cancer subtypes.

We have also identified a distinct subgroup of tumors with *IDH1* R132 mutations that is associated with younger age at diagnosis. Although *IDH1* mutations have previously been identified in prostate cancer with a similar incidence (2.7%) (Ghiam et al., 2012; Kang et al., 2009), we show here that those tumors are all ETS fusion negative and *SPOP* wild-type, have little SCNA burden, and possess elevated levels of genome-wide methylation. The levels of methylation observed in this methylator phenotype are higher than those observed in *IDH1* mutant GBMs and AMLs. Consistent with our observations, a recently published clinical study of 117 prostate cancers identified a single *IDH1* mutant prostate cancer from 56-year-old affected individual that also lacked significant copy number alterations, ETS gene fusions, or driver mutations (Hovelson et al., 2015). Future studies in cohorts with sufficient clinical follow-up will be able to ask whether the *IDH1* mutant prostate cancers are prognostically distinct, as noted for gliomas (Noushmehr et al., 2010) and AMLs (Mardis et al., 2009), and if they are sensitive to newly developed IDH1-targeted therapeutics (Rohle et al., 2013).

Interestingly, 26% of the tumors in this study could not be characterized by one of the taxonomy-defining cardinal genomic alterations. The 26% were clinically and genomically heterogeneous, with some tumors exhibiting extensive DNA copy-number alterations and high Gleason scores indicative of poorer prognosis. About a third of them were genomically similar to *SPOP* and *FOXA1* mutant tumors but lacked any canonical mutation (iCluster 1, methylation cluster 2, mRNA cluster 1); others were enriched for mutations of *TP53*, *KDM6A*, and *KMT2D* or specific SCNAs spanning *MYC* and *CCND1*. Many of the tumors had low Gleason score with few if any DNA copy-number alterations and a normal-like DNA methylation pattern. As previously reported, tumors with fewer genomic alterations were also more commonly Gleason score 6 tumors (38% in the "quiet" class versus 8% in the class with the greatest burden of alterations).

Tumor cellularity, as assessed by pathology review, was lower among tumors with fewer SCNAs (one-sided Mann-Whitney test, $p = 0.0002$), indicating that the apparent lower burden of alterations in tumors with smaller volumes may be due in part to their tumor purities being lower. However, the lower cellularity of these tumors did not limit the detection of clonal molecular

lesions since tumor cellularity between ETS fusion-positive and these fusion-negative tumors was not significantly different (two-sided Mann-Whitney test, p = 0.32). One must also keep in mind that this study was limited to a single tumor focus for each affected individual, even though the vast majority of primary prostate tumors are multifocal and molecular heterogeneity between different foci has been demonstrated (Cooper et al., 2015; Boutros et al., 2015; Lindberg et al., 2015). Such issues must be considered when designing new therapeutic approaches and biomarker panels for clinical use, as affected individuals likely have more than one of these molecular subtypes present due to this commonly occurring tumor multifocality and molecular heterogeneity.

Primary prostate cancers exhibit a wide range of androgen receptor activity. This study demonstrates for the first time a direct association between mutations in *SPOP* or *FOXA1* and increased AR-driven transcription in human prostate cancers. Further studies in preclinical models, as well as in clinical trial settings, will be required to understand the implications of variable AR activity in the contexts of chemoprevention and prostate cancer-directed treatment strategies (Mostaghel et al., 2010). Other, more immediately actionable opportunities for targeted therapy exist for the 19% of primary prostate cancers that have defects in DNA repair and for the nearly equal number of cancers with altered key effectors of both PI3K and MAPK pathways. While the numbers of DNA repair defects found in organ-confined prostate tumors may be lower than those found in metastatic prostate cancer (Robinson et al., 2015), an increase in the number of such defects with disease progression suggests a possible advantage to targeting DNA repair-deficient tumors at an earlier stage of disease, perhaps at initial diagnosis. Such strategies may include preventing DNA damage, as well as targeting deficient DNA repair (Ferguson et al., 2015). Alterations in the PI3K/MTOR pathway also play an important role: beyond the frequent inactivation of *PTEN*, we document rare activation of *PIK3CA*, *PIK3CB*, *AKT1*, and *MTOR*, and of several small GTPases, including *HRAS*, as well as *BRAF*. As DNA sequencing of tumor samples becomes more widely adopted earlier in the clinical care of cancer patients, such alterations may emerge as candidates for inclusion in clinical trials after front-line therapy.

In summary, our integrative assessment of 333 primary prostate cancers has confirmed previously defined molecular subtypes across multiple genomic platforms and identified novel alterations and subtype diversity. It provides a resource for continued investigation into the molecular and biological heterogeneity of the most common cancer in American men.

EXPERIMENTAL PROCEDURES

Tumor and matched normal specimens were obtained from prostate cancer patients who provided informed consent and were approved for collection and distribution by local Institutional Review boards. Blocks frozen in OCT were made of all tumors and of paired benign tissue when present. A 5 micron section was cut from both the top and bottom of the OCT block of 111 tumor cases and from the top or bottom only of the OCT block of 222 tumor cases. Out of 39 normal samples included in the freeze, 23 underwent pathology review, and prostate origin (i.e., no seminal vesicles) and absence of tumor and high grade prostate intraepithelial neoplasia (HGPIN) were confirmed. Tissue images were reviewed by eight genitourinary pathologists, who reported the primary and secondary Gleason patterns of cancer for each slide and estimates of tumor cellularity in 10% increments (from 0%–100%). In case of discrepancies of Gleason scores between the top and bottom sections, the Gleason scores of cancer in the section with the largest area of tumor were used. A subset of 54 cases was reviewed by two pathologists. Discrepancies that occurred between the two pathologists were reconciled by blind review by a third pathologist.

DNA, RNA, and protein were purified and distributed throughout the TCGA network. Samples with evidence for RNA degradation were excluded from the study (Supplemental Experimental Procedures). In total, 333 primary tumors with associated clinicopathologic data were assayed on at least four molecular profiling platforms. Platforms included exome and whole genome DNA sequencing, RNA sequencing, miRNA sequencing, SNP arrays, DNA methylation arrays, and reverse phase protein arrays. Integrated multiplatform analyses were performed.

The data and analysis results can be explored through the Broad Institute FireBrowse portal (http://firebrowse.org/?cohort=PRAD), the cBioPortal for Cancer Genomics (http://www.cbioportal.org/study.do?cancer_study_id=prad_tcga_pub), TCGA Batch Effects (http://bioinformatics.mdanderson.org/tcgambatch/), Regulome Explorer (http://explorer.cancerregulome.org/), and Next-Generation Clustered Heat Maps (http://bioinformatics.mdanderson.org/TCGA/NGCHMPortal/). See also Supplemental Information and the TCGA publication page (https://tcga-data.nci.nih.gov/docs/publications/prad_2015/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.10.025.

CONSORTIA

The members of The Cancer Genome Atlas Research Network are Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D. Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, J. Todd Auman, Miruna Balasundaram, Saianand Balu, Christopher E. Barbieri, Thomas Bauer, Christopher C. Benz, Alain Bergeron, Rameen Beroukhim, Mario Berrios, Adrian Bivol, Tom Bodenheimer, Lori Boice, Moiz S. Bootwalla, Rodolfo Borges dos Reis, Paul C. Boutros, Jay Bowen, Reanne Bowlby, Jeffrey Boyd, Robert K. Bradley, Anne Breggia, Fadi Brimo, Christopher A. Bristow, Denise Brooks, Bradley M. Broom, Alan H. Bryce, Glenn Bubley, Eric Burks, Yaron S.N. Butterfield, Michael Button, David Canes, Carlos G. Carlotti, Rebecca Carlsen, Michel Carmel, Peter R. Carroll, Scott L. Carter, Richard Cartun, Brett S. Carver, June M. Chan, Matthew T. Chang, Yu Chen, Andrew D. Cherniack, Simone Chevalier, Lynda Chin, Juok Cho, Andy Chu, Eric Chuah, Sudha Chudamani, Kristian Cibulskis, Giovanni Ciriello, Amanda Clarke, Matthew R. Cooperberg, Niall M. Corcoran, Anthony J. Costello, Janet Cowan, Daniel Crain, Erin Curley, Kerstin David, John A. Demchok, Francesca Demichelis, Noreen Dhalla, Rajiv Dhir, Alexandre Doueik, Bettina Drake, Heidi Dvinge, Natalya Dyakova, Ina Felau, Martin L. Ferguson, Scott Frazer, Stephen Freedland, Yao Fu, Stacey B. Gabriel, Jianjiong Gao, Johanna Gardner, Julie M. Gastier-Foster, Nils Gehlenborg, Mark Gerken, Mark B. Gerstein, Gad Getz, Andrew K. Godwin, Anuradha Gopalan, Markus Graefen, Kiley Graim, Thomas Gribbin, Ranabir Guin, Manaswi Gupta, Angela Hadjipanayis, Syed Haider, Lucie Hamel, D. Neil Hayes, David I. Heiman, Julian Hess, Katherine A. Hoadley, Andrea H. Holbrook, Robert A. Holt, Antonia Holway, Christopher M. Hovens, Alan P. Hoyle, Mei Huang, Carolyn M. Hutter, Michael Ittmann, Lisa Iype, Stuart R. Jefferys, Corbin D. Jones, Steven J.M. Jones, Hartmut Juhl, Andre Kahles, Christopher J. Kane, Katayoon Kasaian, Michael Kerger, Ekta Khurana, Jaegil Kim, Robert J. Klein, Raju Kucherlapati, Louis Lacombe, Marc Ladanyi, Phillip H. Lai, Peter W. Laird, Eric S. Lander, Mathieu Latour, Michael S. Lawrence, Kevin Lau, Tucker LeBien, Darlene Lee, Semin Lee, Kjong-Van Lehmann, Kristen M. Leraas, Ignaty Leshchiner, Robert Leung, John A. Libertino, Tara M. Lichtenberg, Pei Lin, W. Marston Linehan, Shiyun Ling, Scott M. Lippman, Jia Liu, Wenbin Liu, Lucas Lochovsky, Massimo

Loda, Christopher Logothetis, Laxmi Lolla, Teri Longacre, Yiling Lu, Jianhua Luo, Yussanne Ma, Harshad S. Mahadeshwar, David Mallery, Armaz Mariamidze, Marco A. Marra, Michael Mayo, Shannon McCall, Ginette McKercher, Shaowu Meng, Anne-Marie Mes-Masson, Maria J. Merino, Matthew Meyerson, Piotr A. Mieczkowski, Gordon B. Mills, Kenna R. Mills Shaw, Sarah Minner, Alireza Moinzadeh, Richard A. Moore, Scott Morris, Carl Morrison, Lisle E. Mose, Andrew J. Mungall, Bradley A. Murray, Jerome B. Myers, Rashi Naresh, Joel Nelson, Mark A. Nelson, Peter S. Nelson, Yulia Newton, Michael S. Noble, Houtan Noushmehr, Matti Nykter, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Joel S. Parker, Joseph Paulauskis, Robert Penny, Charles M. Perou, Alain Piché, Todd Pihl, Peter A. Pinto, Davide Prandi, Alexei Protopopov, Nilsa C. Ramirez, Arvind Rao, W. Kimryn Rathmell, Gunnar Rätsch, Xiaojia Ren, Victor E. Reuter, Sheila M. Reynolds, Suhn K. Rhie, Kimberly Rieger-Christ, Jeffrey Roach, A. Gordon Robertson, Brian Robinson, Mark A. Rubin, Fred Saad, Sara Sadeghi, Gordon Saksena, Charles Saller, Andrew Salner, Francisco Sanchez-Vega, Chris Sander, George Sandusky, Guido Sauter, Andrea Sboner, Peter T. Scardino, Eleonora Scarlata, Jacqueline E. Schein, Thorsten Schlomm, Laura S. Schmidt, Nikolaus Schultz, Steven E. Schumacher, Jonathan Seidman, Luciano Neder, Sahil Seth, Alexis Sharp, Candace Shelton, Troy Shelton, Hui Shen, Ronglai Shen, Mark Sherman, Margi Sheth, Yan Shi, Juliann Shih, Ilya Shmulevich, Jeffry Simko, Ronald Simon, Janae V. Simons, Payal Sipahimalani, Tara Skelly, Heidi J. Sofia, Matthew G. Soloway, Xingzhi Song, Andrea Sorcini, Carrie Sougnez, Serghei Stepa, Chip Stewart, John Stewart, Joshua M. Stuart, Travis B. Sullivan, Charlie Sun, Huandong Sun, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Katherine Tarvin, Barry S. Taylor, Patrick Teebagy, Imelda Tenggara, Bernard Têtu, Ashutosh Tewari, Nina Thiessen, Timothy Thompson, Leigh B. Thorne, Daniela P. Tirapelli, Scott A. Tomlins, Felipe Amstalden Trevisan, Patricia Troncoso, Lawrence D. True, Maria Christina Tsourlakis, Svitlana Tyekucheva, Eliezer Van Allen, David J. Van Den Berg, Umadevi Veluvolu, Roel Verhaak, Cathy D. Vocke, Doug Voet, Yunhu Wan, Qingguo Wang, Wenyi Wang, Zhining Wang, Nils Weinhold, John N. Weinstein, Daniel J. Weisenberger, Matthew D. Wilkerson, Lisa Wise, John Witte, Chia-Chin Wu, Junyuan Wu, Ye Wu, Andrew W. Xu, Shalini S. Yadav, Liming Yang, Lixing Yang, Christina Yau, Huihui Ye, Peggy Yena, Thomas Zeng, Jean C. Zenklusen, Hailei Zhang, Jianhua Zhang, Jiashan Zhang, Wei Zhang, Yi Zhong, Kelsey Zhu, and Erik Zmuda.

AUTHOR CONTRIBUTIONS

Project leaders: M. Loda and C. Sander; analysis leaders: N.S. and B.S.T.; project coordinators: I.F. and M. Sheth; data coordinators: J. Armenia and N.S.; supplement coordinator: J. Armenia; pathology review: M.I., M. Loda, V.E.R., B.R., M.A.R., P. Troncoso, L.D.T., and H.Y.; clinical data: J. Bowen, N.M.C., L.I., K.M.L., and T.M.L.; tumor cellularity analysis: J. Ahn, P.C.B., A.D.C., F.D., S.H., D.P., M.A.R., J. Shih, S.T., and W.W.; exome sequencing analysis: M. Gupta, C. Sougnez, and E.V.A.; whole-genome sequencing analysis: J. Armenia, Y.F., M.B.G., M. Gupta, E.K., L. Lochovsky, A. Pantazi, C. Sougnez, and E.V.A.; copy-number analysis: A.D.C., and J. Shih; mRNA expression and fusion analysis: A. Sboner, N.S., M.D.W., and C.-C.W.; androgen receptor analysis: J. Armenia, R.K.B., H.D., R.A.M., A.J.M., P.S.N., and N.S.; DNA methylation analysis: P.W.L., S.K.R., and H.S.; miRNA analysis: A.G.R.; RPPA analysis: R.A., W.L., Y.L., and G.B.M.; integrative analysis: A. Arora, A.D.C., D.I.H., L.I., N.S., R. Shen, B.S.T., and M.D.W.; batch effects analysis: R.A., A.K., K.-V.L., S. Ling, A.R., and J.N.W.; mRNA degradation analysis: A.K., K.-V.L., G.R., M.A.R., N.S., and M.D.W.; manuscript writing: C.E.B., C.C.B., P.R.C., Y.C., A.D.C., F.D., S.M.L., M. Loda, J. Simko, P.S.N., S.K.R., A.G.R., M.A.R., C. Sander, A. Sboner, N.S., B.S.T., S.A.T., E.V.A., and J.N.W.

ACKNOWLEDGMENTS

We are grateful to all the affected individuals and families who contributed to this study. We thank Margi Sheth and Ina Felau for project management. This work was supported by the following grants from the NIH: 5U24CA143799, 5U24CA143835, 5U24CA143840, 5U24CA143843, 5U24CA143845, 5U24CA143848, 5U24CA143858, 5U24CA143866, 5U24CA143867, 5U24CA143882, 5U24CA143883, 5U24CA144025, U54HG003067, U54HG003079, U54HG003273, and P30CA16672. D.J.W. is a consultant for Zymo Research Corporation. S.A.T., M.A.R., and F.D. are co-authors on a patent issued to the University of Michigan and the Brigham and Women's Hospital regarding ETS gene fusions in prostate cancer. The diagnostic field of use has been licensed to Hologic/Gen-Probe, Inc., who has sub-licensed some rights to Ventana Medical Systems. S.A.T. is a coauthor on a patent filed by the University of Michigan regarding SPINK1 in prostate cancer. The diagnostic field of use has been licensed to Hologic/Gen-Probe, Inc., who has sub-licensed some rights to Ventana Medical Systems. S.A.T. has served as a consultant and received honoraria from Ventana Medical Systems. M.A.R. is a co-inventor of the patent for the detection and therapeutic field of SPOP mutations in prostate cancer filed by Cornell University. E.V.A. is a consultant for Syapse and Ventana Medical Systems and owns equity in Syapse and Microsoft. A.D.C. has received research support from Bayer, AG. M.B.G. serves on the SAB of BINA. R. Beroukhim is a consultant for and received grant funding from Novartis. M. Meyerson received research support from Bayer and has equity in and is a consultant for Foundation Medicine.

Received: June 9, 2015
Revised: August 14, 2015
Accepted: October 6, 2015
Published: November 5, 2015

REFERENCES

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M.B., Diao, L., Wistuba, I.I., and Wang, W. (2013). DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinformatics *29*, 1865–1871.

Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., Conti, D.V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D.J., Wang, Z., Saunders, E., et al.; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; COGS (Collaborative Oncological Gene-environment Study) Consortium; GAME-ON/ELLIPSE Consortium (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet. *46*, 1103–1109.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

An, J., Wang, C., Deng, Y., Yu, L., and Huang, H. (2014). Destruction of full-length androgen receptor by wild-type SPOP, but not prostate-cancer-associated mutants. Cell Rep. *6*, 657–669.

Antonarakis, E.S., Lu, C., Wang, H., Luber, B., Nakazawa, M., Roeser, J.C., Chen, Y., Mohammad, T.A., Chen, Y., Fedor, H.L., et al. (2014). AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. N. Engl. J. Med. *371*, 1028–1038.

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

Baena, E., Shao, Z., Linn, D.E., Glass, K., Hamblen, M.J., Fujiwara, Y., Kim, J., Nguyen, M., Zhang, X., Godinho, F.J., et al. (2013). ETV1 directs androgen metabolism and confers aggressive prostate cancer in targeted mice and patients. Genes Dev. *27*, 683–698.

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

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Blattner, M., Lee, D.J., O'Reilly, C., Park, K., MacDonald, T.Y., Khani, F., Turner, K.R., Chiu, Y.L., Wild, P.J., Dolgalev, I., et al. (2014). SPOP mutations in prostate cancer across demographically diverse patient cohorts. Neoplasia *16*, 14–20.

Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. Genes Dev. *25*, 2158–2172.

Börno, S.T., Fischer, A., Kerick, M., Fälth, M., Laible, M., Brase, J.C., Kuner, R., Dahl, A., Grimm, C., Sayanjali, B., et al. (2012). Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. Cancer Discov. *2*, 1024–1035.

Boutros, P.C., Fraser, M., Harding, N.J., de Borja, R., Trudel, D., Lalonde, E., Meng, A., Hennings-Yeomans, P.H., McPherson, A., Sabelnykova, V.Y., et al. (2015). Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat. Genet. *47*, 736–745.

Bowyer, S.E., Rao, A.D., Lyle, M., Sandhu, S., Long, G.V., McArthur, G.A., Raleigh, J.M., Hicks, R.J., and Millward, M. (2014). Activity of trametinib in K601E and L597Q BRAF mutation-positive metastatic melanoma. Melanoma Res. *24*, 504–508.

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.

Chandran, U.R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M., and Monzon, F.A. (2007). Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer *7*, 64.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2015). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. Published online November 9, 2015. http://dx.doi.org/10.1038/nbt.3391.

Cooper, C.S., Eeles, R., Wedge, D.C., Van Loo, P., Gundem, G., Alexandrov, L.B., Kremeyer, B., Butler, A., Lynch, A.G., Camacho, N., et al.; ICGC Prostate Group (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat. Genet. *47*, 367–372.

Cooperberg, M.R., Broering, J.M., and Carroll, P.R. (2009). Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. J. Natl. Cancer Inst. *101*, 878–887.

D'Amico, A.V., Whittington, R., Malkowicz, S.B., Schultz, D., Blank, K., Broderick, G.A., Tomaszewski, J.E., Renshaw, A.A., Kaplan, I., Beard, C.J., and Wein, A. (1998). Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. JAMA *280*, 969–974.

Dahlman, K.B., Xia, J., Hutchinson, K., Ng, C., Hucks, D., Jia, P., Atefi, M., Su, Z., Branch, S., Lyle, P.L., et al. (2012). BRAF(L597) mutations in melanoma are associated with sensitivity to MEK inhibitors. Cancer Discov. *2*, 791–797.

Dehm, S.M., Schmidt, L.J., Heemers, H.V., Vessella, R.L., and Tindall, D.J. (2008). Splicing of a novel androgen receptor exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. Cancer Res. *68*, 5469–5477.

Delahaye-Sourdeix, M., Anantharaman, D., Timofeeva, M.N., Gaborieau, V., Chabrier, A., Vallée, M.P., Lagiou, P., Holcátová, I., Richiardi, L., Kjaerheim, K., et al. (2015). A rare truncating BRCA2 variant and genetic susceptibility to upper aerodigestive tract cancer. J. Natl. Cancer Inst. *107*, djv037.

Demichelis, F., Setlur, S.R., Beroukhim, R., Perner, S., Korbel, J.O., Lafargue, C.J., Pflueger, D., Pina, C., Hofer, M.D., Sboner, A., et al. (2009). Distinct genomic aberrations associated with ERG rearranged prostate cancer. Genes Chromosomes Cancer *48*, 366–380.

Farrugia, D.J., Agarwal, M.K., Pankratz, V.S., Deffenbaugh, A.M., Pruss, D., Frye, C., Wadum, L., Johnson, K., Mentlick, J., Tavtigian, S.V., et al. (2008). Functional assays for classification of BRCA2 variants of uncertain significance. Cancer Res. *68*, 3523–3531.

Ferguson, L.R., Chen, H., Collins, A.R., Connell, M., Damia, G., Dasgupta, S., Malhotra, M., Meeker, A.K., Amedei, A., Amin, A., et al. (2015). Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. Semin. Cancer Biol. Published online April 11, 2015. http://dx.doi.org/10.1016/j.semcancer.2015.03.005.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 (Lyon, France: International Agency for Research on Cancer).

Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F., et al. (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell *18*, 553–567.

Friedlander, T.W., Roy, R., Tomlins, S.A., Ngo, V.T., Kobayashi, Y., Azameera, A., Rubin, M.A., Pienta, K.J., Chinnaiyan, A., Ittmann, M.M., et al. (2012). Common structural and epigenetic changes in the genome of castration-resistant prostate cancer. Cancer Res. *72*, 616–625.

Gasi, D., van der Korput, H.A., Douben, H.C., de Klein, A., de Ridder, C.M., van Weerden, W.M., and Trapman, J. (2011). Overexpression of full-length ETV1 transcripts in clinical prostate cancer due to gene translocation. PLoS ONE *6*, e16332.

Geng, C., He, B., Xu, L., Barbieri, C.E., Eedunuri, V.K., Chew, S.A., Zimmermann, M., Bond, R., Shou, J., Li, C., et al. (2013). Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover. Proc. Natl. Acad. Sci. USA *110*, 6997–7002.

Geng, C., Rajapakshe, K., Shah, S.S., Shou, J., Eedunuri, V.K., Foley, C., Fiskus, W., Rajendran, M., Chew, S.A., Zimmermann, M., et al. (2014). Androgen receptor is the key transcriptional mediator of the tumor suppressor SPOP in prostate cancer. Cancer Res. *74*, 5631–5643.

Ghiam, A.F., Cairns, R.A., Thoms, J., Dal Pra, A., Ahmed, O., Meng, A., Mak, T.W., and Bristow, R.G. (2012). IDH mutation status in prostate cancer. Oncogene *31*, 3826.

Gopalan, A., Leversha, M.A., Satagopan, J.M., Zhou, Q., Al-Ahmadie, H.A., Fine, S.W., Eastham, J.A., Scardino, P.T., Scher, H.I., Tickoo, S.K., et al. (2009). TMPRSS2-ERG gene fusion is not associated with outcome in patients treated by prostatectomy. Cancer Res. *69*, 1400–1406.

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

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M., Papaemmanuil, E., Brewer, D.S., Kallio, H.M., Högnäs, G., Annala, M., et al.; ICGC Prostate UK Group (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353–357.

Guo, G., Sun, X., Chen, C., Wu, S., Huang, P., Li, Z., Dean, M., Huang, Y., Jia, W., Zhou, Q., et al. (2013). Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat. Genet. *45*, 1459–1463.

Heemers, H.V., and Tindall, D.J. (2007). Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex. Endocr. Rev. *28*, 778–808.

Hieronymus, H., Lamb, J., Ross, K.N., Peng, X.P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S.M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer Cell *10*, 321–330.

Hieronymus, H., Schultz, N., Gopalan, A., Carver, B.S., Chang, M.T., Xiao, Y., Heguy, A., Huberman, K., Bernstein, M., Assel, M., et al. (2014). Copy number alteration burden predicts prostate cancer relapse. Proc. Natl. Acad. Sci. USA *111*, 11139–11144.

Hovelson, D.H., McDaniel, A.S., Cani, A.K., Johnson, B., Rhodes, K., Williams, P.D., Bandla, S., Bien, G., Choppa, P., Hyland, F., et al. (2015). Development and validation of a scalable next-generation sequencing system for assessing relevant somatic variants in solid tumors. Neoplasia *17*, 385–399.

Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C., et al. (2014). The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. Nat. Commun. *5*, 3630.

Jin, H.J., Zhao, J.C., Ogden, I., Bergan, R.C., and Yu, J. (2013). Androgen receptor-independent function of FoxA1 in prostate cancer metastasis. Cancer Res. *73*, 3725–3736.

Kang, M.R., Kim, M.S., Oh, J.E., Kim, Y.R., Song, S.Y., Seo, S.I., Lee, J.Y., Yoo, N.J., and Lee, S.H. (2009). Mutational analysis of IDH1 codon 132 in glioblastomas and other common cancers. Int. J. Cancer *125*, 353–355.

Karanika, S., Karantanos, T., Li, L., Corn, P.G., and Thompson, T.C. (2014). DNA damage response and prostate cancer: defects, regulation and therapeutic implications. Oncogene *34*, 2815–2822.

Kattan, M.W., Eastham, J.A., Stapleton, A.M., Wheeler, T.M., and Scardino, P.T. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J. Natl. Cancer Inst. *90*, 766–771.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al.; 1000 Genomes Project Consortium (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.

Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M., et al. (2011). Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res. *21*, 1028–1041.

Kluth, M., Hesse, J., Heinl, A., Krohn, A., Steurer, S., Sirma, H., Simon, R., Mayer, P.S., Schumacher, U., Grupp, K., et al. (2013). Genomic deletion of MAP3K7 at 6q12-22 is associated with early PSA recurrence in prostate cancer and absence of TMPRSS2:ERG fusions. Mod. Pathol. *26*, 975–983.

Kobayashi, Y., Absher, D.M., Gulzar, Z.G., Young, S.R., McKenney, J.K., Peehl, D.M., Brooks, J.D., Myers, R.M., and Sherlock, G. (2011). DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome Res. *21*, 1017–1027.

Kunju, L.P., Carskadon, S., Siddiqui, J., Tomlins, S.A., Chinnaiyan, A.M., and Palanisamy, N. (2014). Novel RNA hybridization method for the in situ detection of ETV1, ETV4, and ETV5 gene fusions in prostate cancer. Appl. Immunohistochem. Mol. Morphol. *22*, e32–e40.

Lalonde, E., Ishkanian, A.S., Sykes, J., Fraser, M., Ross-Adams, H., Erho, N., Dunning, M.J., Halim, S., Lamb, A.D., Moon, N.C., et al. (2014). Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. Lancet Oncol. *15*, 1521–1532.

Lapointe, J., Li, C., Giacomini, C.P., Salari, K., Huang, S., Wang, P., Ferrari, M., Hernandez-Boussard, T., Brooks, J.D., and Pollack, J.R. (2007). Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res. *67*, 8504–8510.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.G., Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K., and Rosenfeld, M.G. (2009). Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell *139*, 1069–1083.

Lindberg, J., Kristiansen, A., Wiklund, P., Grönberg, H., and Egevad, L. (2015). Tracking the origin of metastatic prostate cancer. Eur. Urol. *67*, 819–822.

Mahapatra, S., Klee, E.W., Young, C.Y., Sun, Z., Jimenez, R.E., Klee, G.G., Tindall, D.J., and Donkena, K.V. (2012). Global methylation profiling for risk prediction of prostate cancer. Clin. Cancer Res. *18*, 2882–2895.

Mani, R.S., Tomlins, S.A., Callahan, K., Ghosh, A., Nyati, M.K., Varambally, S., Palanisamy, N., and Chinnaiyan, A.M. (2009). Induced chromosomal proximity and gene fusions in prostate cancer. Science *326*, 1230.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. *361*, 1058–1066.

Martin, S.T., Matsubayashi, H., Rogers, C.D., Philips, J., Couch, F.J., Brune, K., Yeo, C.J., Kern, S.E., Hruban, R.H., and Goggins, M. (2005). Increased prevalence of the BRCA2 polymorphic stop codon K3326X among individuals with familial pancreatic cancer. Oncogene *24*, 3652–3656.

Mateo, J., Hall, E., Sandhu, S., Omlin, A.G., Miranda, S., Carreira, S., Goodall, J., Gillman, A., Mossop, H., Ralph, C., et al. (2014). LBA20 - Antitumour activity of the PARP inhibitor olaparib in unselected sporadic castration-resistant prostate cancer (CRPC) in the TOPARP trial. Ann. Oncol. *25*, 1–41.

Mostaghel, E.A., Geng, L., Holcomb, I., Coleman, I.M., Lucas, J., True, L.D., and Nelson, P.S. (2010). Variability in the androgen response of prostate epithelium to 5alpha-reductase inhibition: implications for prostate cancer chemoprevention. Cancer Res. *70*, 1286–1295.

Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P., et al.; Cancer Genome Atlas Research Network (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell *17*, 510–522.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., et al. (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat. Med. *16*, 793–798.

Parikh, C., Janakiraman, V., Wu, W.I., Foo, C.K., Kljavin, N.M., Chaudhuri, S., Stawiski, E., Lee, B., Lin, J., Li, H., et al. (2012). Disruption of PH-kinase domain interactions leads to oncogenic activation of AKT in human cancers. Proc. Natl. Acad. Sci. USA *109*, 19368–19373.

Paris, P.L., Andaya, A., Fridlyand, J., Jain, A.N., Weinberg, V., Kowbel, D., Brebner, J.H., Simko, J., Watson, J.E., Volik, S., et al. (2004). Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. Hum. Mol. Genet. *13*, 1303–1313.

Penney, K.L., Stampfer, M.J., Jahn, J.L., Sinnott, J.A., Flavin, R., Rider, J.R., Finn, S., Giovannucci, E., Sesso, H.D., Loda, M., et al. (2013). Gleason grade progression is uncommon. Cancer Res. *73*, 5163–5168.

Pettersson, A., Graff, R.E., Bauer, S.R., Pitt, M.J., Lis, R.T., Stack, E.C., Martin, N.E., Kunz, L., Penney, K.L., Ligon, A.H., et al. (2012). The TMPRSS2:ERG rearrangement, ERG expression, and prostate cancer outcomes: a cohort study and meta-analysis. Cancer Epidemiol. Biomarkers Prev. *21*, 1497–1509.

Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P.C., Svensson, M.A., Kitabayashi, N., Moss, B.J., MacDonald, T.Y., et al. (2011). Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res. *21*, 56–67.

Prandi, D., Baca, S.C., Romanel, A., Barbieri, C.E., Mosquera, J.M., Fontugne, J., Beltran, H., Sboner, A., Garraway, L.A., Rubin, M.A., and Demichelis, F. (2014). Unraveling the clonal hierarchy of somatic genomic aberrations. Genome Biol. *15*, 439.

Pritchard, C.C., Morrissey, C., Kumar, A., Zhang, X., Smith, C., Coleman, I., Salipante, S.J., Milbank, J., Yu, M., Grady, W.M., et al. (2014). Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. Nat. Commun. *5*, 4988.

Quon, G., Haider, S., Deshwar, A.G., Cui, A., Boutros, P.C., and Morris, Q. (2013). Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. Genome Med. *5*, 29.

Robinson, D., Van Allen, E.M., Wu, Y.M., Schultz, N., Lonigro, R.J., Mosquera, J.M., Montgomery, B., Taplin, M.E., Pritchard, C.C., Attard, G., et al. (2015). Integrative clinical genomics of advanced prostate cancer. Cell *161*, 1215–1228.

Rodrigues, L.U., Rider, L., Nieto, C., Romero, L., Karimpour-Fard, A., Loda, M., Lucia, M.S., Wu, M., Shi, L., Cimic, A., et al. (2015). Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. Cancer Res. *75*, 1021–1034.

Rohle, D., Popovici-Muller, J., Palaskas, N., Turcan, S., Grommes, C., Campos, C., Tsoi, J., Clark, O., Oldrini, B., Komisopoulou, E., et al. (2013). An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. Science *340*, 626–630.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D.Z., Rozowsky, J.S., Tewari, A.K., Kitabayashi, N., Moss, B.J., Chee, M.S., et al. (2010). FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. *11*, R104.

Schaeffer, E.M., Marchionni, L., Huang, Z., Simons, B., Blackman, A., Yu, W., Parmigiani, G., and Berman, D.M. (2008). Androgen-induced programs for prostate epithelial growth and invasion arise in embryogenesis and are reactivated in cancer. Oncogene *27*, 7180–7191.

Schwartz, S., Wongvipat, J., Trigwell, C.B., Hancox, U., Carver, B.S., Rodrik-Outmezguine, V., Will, M., Yellen, P., de Stanchina, E., Baselga, J., et al. (2015). Feedback suppression of PI3Kα signaling in PTEN-mutated tumors is relieved by selective inhibition of PI3Kβ. Cancer Cell *27*, 109–122.

Shen, R., Olshen, A.B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics *25*, 2906–2912.

Siegel, R.L., Miller, K.D., and Jemal, A. (2015). Cancer statistics, 2015. CA Cancer J. Clin. *65*, 5–29.

Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., Arora, V.K., Kaushik, P., Cerami, E., Reva, B., et al. (2010). Integrative genomic profiling of human prostate cancer. Cancer Cell *18*, 11–22.

Tirode, F., Surdez, D., Ma, X., Parker, M., Le Deley, M.C., Bahrami, A., Zhang, Z., Lapouble, E., Grossetête-Lalami, S., Rusch, M., et al.; St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project and the International Cancer Genome Consortium (2014). Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov. *4*, 1342–1353.

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

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

Tomlins, S.A., Bjartell, A., Chinnaiyan, A.M., Jenster, G., Nam, R.K., Rubin, M.A., and Schalken, J.A. (2009). ETS gene fusions in prostate cancer: from discovery to daily clinical practice. Eur. Urol. *56*, 275–286.

Tomlins, S.A., Alshalalfa, M., Davicioni, E., Erho, N., Yousefi, K., Zhao, S., Haddad, Z., Den, R.B., Dicker, A.P., Trock, B.J., et al. (2015). Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. Eur. Urol. *68*, 555–567.

van Dekken, H., Paris, P.L., Albertson, D.G., Alers, J.C., Andaya, A., Kowbel, D., van der Kwast, T.H., Pinkel, D., Schröder, F.H., Vissers, K.J., et al. (2004). Evaluation of genetic patterns in different tumor areas of intermediate-grade prostatic adenocarcinomas by high-resolution genomic array analysis. Genes Chromosomes Cancer *39*, 249–256.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Wang, X.S., Shankar, S., Dhanasekaran, S.M., Ateeq, B., Sasaki, A.T., Jing, X., Robinson, D., Cao, Q., Prensner, J.R., Yocum, A.K., et al. (2011). Characterization of KRAS rearrangements in metastatic prostate cancer. Cancer Discov. *1*, 35–43.

Watson, P.A., Chen, Y.F., Balbas, M.D., Wongvipat, J., Socci, N.D., Viale, A., Kim, K., and Sawyers, C.L. (2010). Constitutively active androgen receptor splice variants expressed in castration-resistant prostate cancer require full-length androgen receptor. Proc. Natl. Acad. Sci. USA *107*, 16759–16765.

Cell

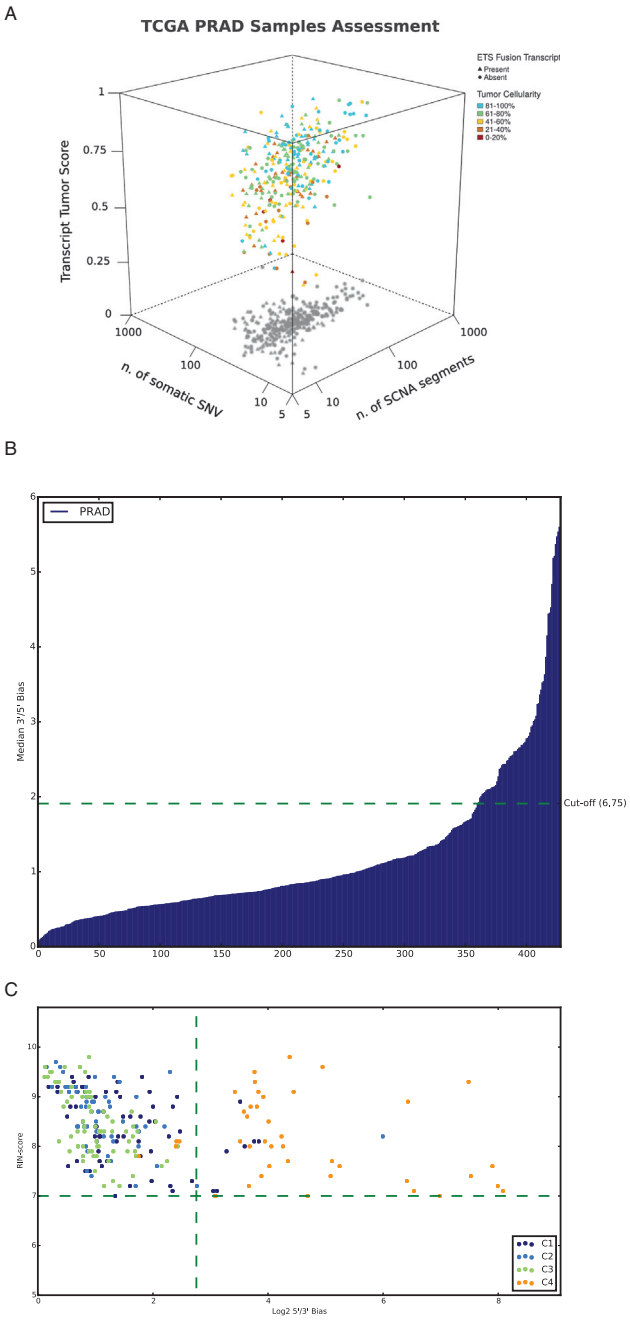


Figure S1. Assessment of Clonality and RNA Degradation, Related to Figure 1, Table 1, and Experimental Procedures

(A) The plot summarizes the characteristics of the 333 PRAD study samples. Each element represents a tumor sample that is annotated with five features: (i) the number of somatic copy number changes (SCNA) on the x axis, defined as the number of genomic segments with log2 ratio outside of the interval [-0.15, 0.15]; (ii) the number of somatic single nucleotide variant (SNV) on the y axis, including silent and non-silent mutations; (iii) the Transcript Tumor Score on the z-axis (the higher the score the more representation of prostate cancer tumor transcripts); (iv) the presence or absence of tumor specific fusion transcripts (ETS fusions), presence or absence indicated with triangle or circle, respectively; and (v) the consensus pathology estimate of cellularity represented by the element color.
(B) 3′/5′ bias across 539 RNA-seq samples, including tumor and normal samples, prior to sample removal.
(C) Comparison of RIN-scores against 3′/5′ bias. Color shows RNA-based subtype clusters.

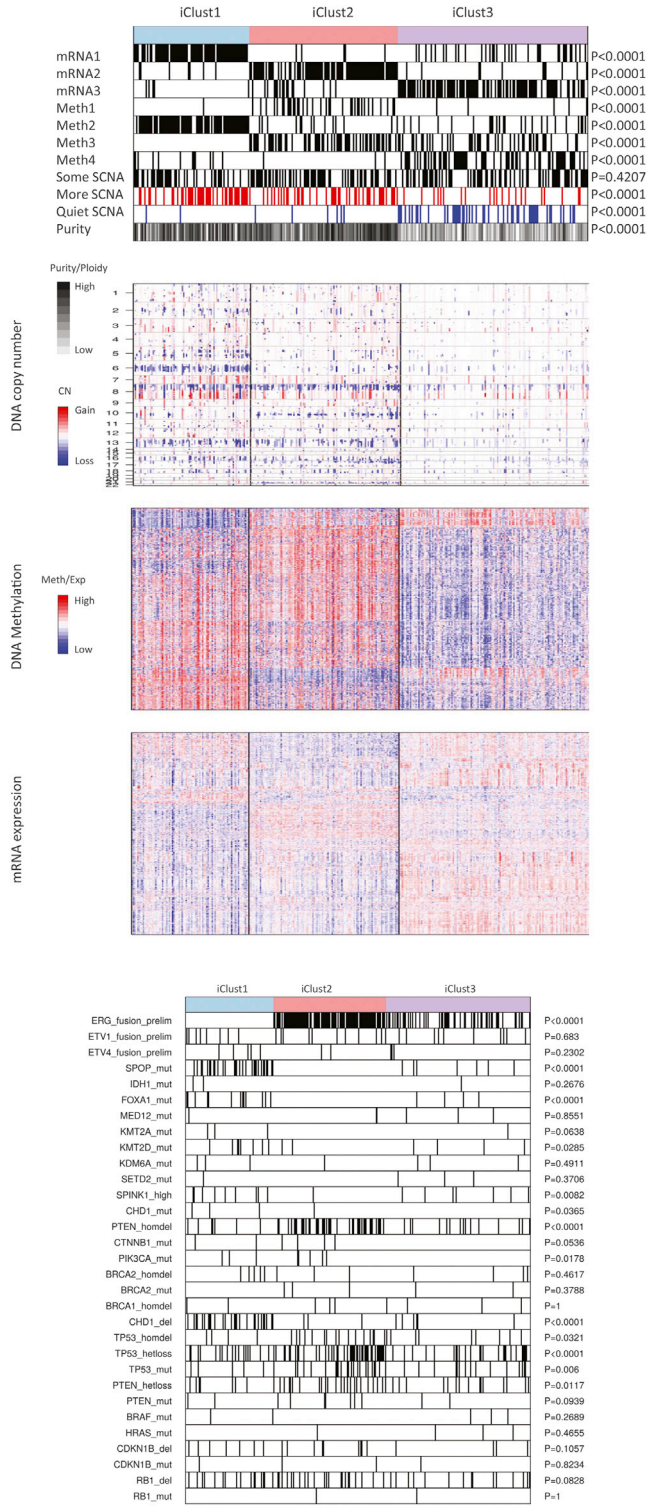


Figure S2. Integrative Clustering Revealed Three Distinct Molecular Subgroups of Prostate Cancer, Related to Figure 1

The top section illustrates the relationship between the iCluster subtypes and the platform-specific (mRNA, methylation, SCNA) subtypes. The middle section presents heatmaps of DNA copy-number, DNA methylation, and mRNA expression by the iCluster subtypes. The association between the iCluster membership and other sample attributes is illustrated in the bottom panel.

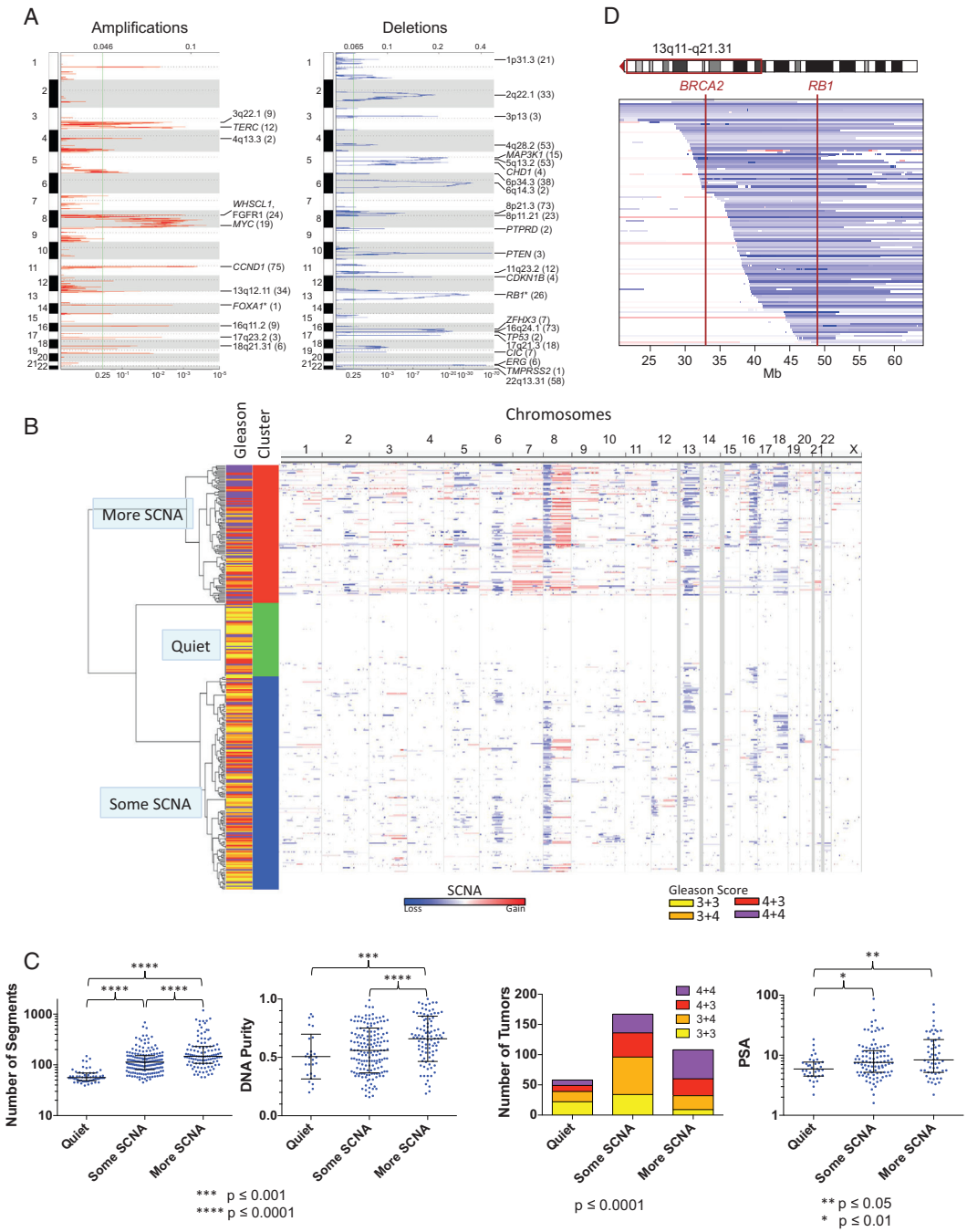


Figure S3. DNA Copy-Number Alterations and Subtypes in Prostate Cancer, Related to Figures 1, 2, and 5

(A) GISTIC 2.0 analysis of focal amplification and deletions in prostate tumors. Significantly recurring focal amplifications (left) and deletions (right) are plotted by false discovery rates along chromosomal locations. Annotated peaks are those with q values $\leq$ 0.1 and have fewer the 75 genes. Possible driver oncogenes, tumor suppressor and fragile sites are indicated along with the number genes within a peak. Oncogenes or tumor suppressor marked with an * are adjacent to minimally defined GISTIC peaks.

(B) Prostate tumors clustering using chromosomal arm level copy number changes. In the heatmap, SCNA in tumors (vertical axis) are plotted by chromosomal location (horizontal axis). Vertical side bars show the division of the major cluster group, mutation rates, and gleason scores.

(C) Distribution of chromosomal copy number breaks, tumor DNA purity, gleason scores and PSA in each copy number cluster.

(D) Co-deletion of BRCA2 and RB1. Related to Figure 5. BRCA2, which is located on the q arm of chromosome 13, is frequently co-deleted with RB1 in prostate cancer and almost never by itself. A 40 Mb region on chromosome 13 is shown with tumors in rows from top to bottom and shades of blue showing different deletion intensities (white regions are diploid and red indicates copy-number gain).

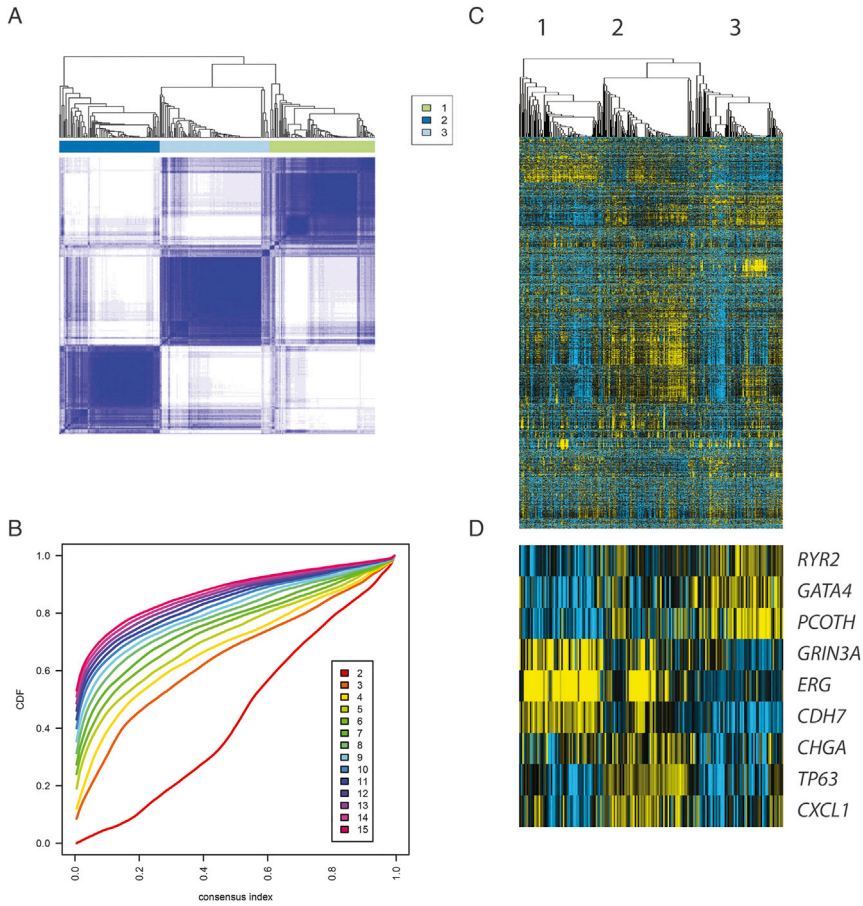


Figure S4. mRNA Expression Subtypes, Related to Figure 1

Unsupervised expression clustering of prostate tumor mRNA-seq. The top 3,000 most variable genes, by median absolute deviation, were selected from prostate tumor mRNA-seq. Expression data were log2 transformed and gene median centered. Consensus average linkage hierarchical clustering (ConsensusClusterPlus; Wilkerson et al., 2010) supported the presence of 3 clusters.

(A–D) The consensus matrix (A. rows and columns are tumors) indicates the frequency that samples occur in the same cluster, over repeated sub-samplings of the cohort, with increasing dark blue shading representing increasing co-clustering. The consensus cumulative distribution function plot indicates an appreciable increase in consensus is achieved at 3 clusters and subsequent divisions yield minor increases (B). Expression heatmap of the genes used in clustering and consensus dendrogram are displayed in (C), with selected exemplar genes displayed in (D).

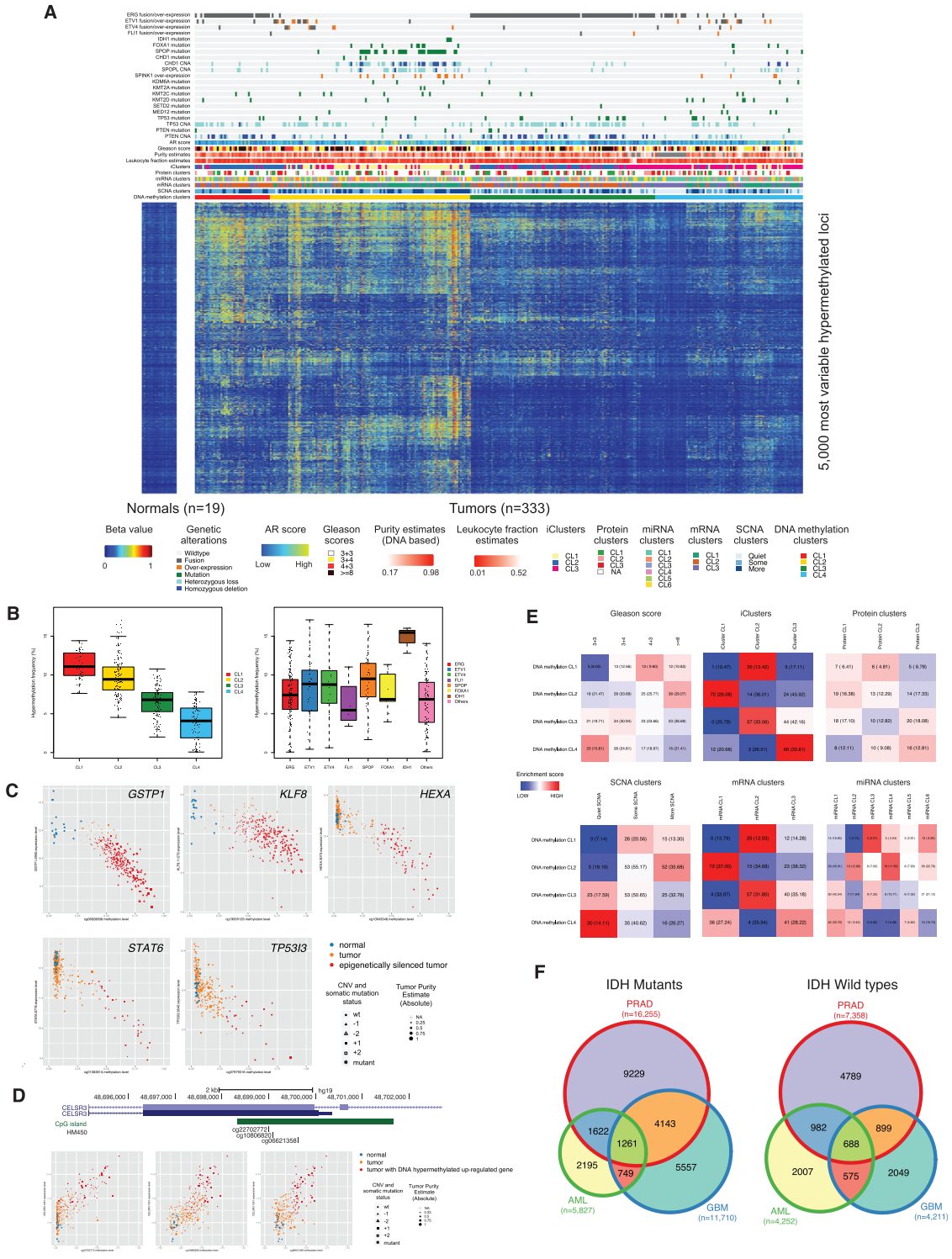


Figure S5. DNA Hypermethylation and Epigenetically Silenced Genes, Related to Figures 1 and 3

(A) DNA hypermethylation heatmap with subtypes and annotations. Unsupervised clustering of DNA hypermethylation data from 333 tumor samples detected four DNA methylation clusters.

(B) DNA hypermethylation frequencies among tumors, grouped by DNA methylation clusters (left) and fusion/mutation subgroups (right).

(C) Example scatter plots of identified epigenetically silenced genes. In each scatterplot, each sample is colored according to its epigenetic silencing status: normals are shown in blue, tumors without epigenetically silenced genes are shown in orange, and tumors with epigenetically silenced genes are shown in red.

(legend continued on next page)

(D) Example scatter plots of a gene, *CELSR3*, displayed increased DNA methylation associated with elevated expression. A UCSC genome browser screen shot shows the location of the CpG island and HM450 probes near transcription start site of *CELSR3* gene. In each scatterplot, each sample is colored according to its DNA methylation and gene expression status: normals are shown in blue, tumors without DNA hypermethylated upregulated genes are shown in orange, and tumors with DNA hypermethylated upregulated genes are shown in red.
The x axis represents DNA methylation level of individual CpG site, and y axis represents the gene expression level. For tumor samples, purity estimates are indicated as a size of each dot (small – lower purity, large – higher purity). Somatic mutation and CNV status of each gene is shown as a shape (filled circle – wild-type, asterisk – mutant, filled triangle – one allele deletion, open triangle – both alleles deleted, filled square – one allele amplified, open square – both alleles amplified).
(E) Comparison of DNA methylation clusters with Gleason scores and other platform clusters. The number of tumors in each group is shown in the heatmap with the expected number of tumors in parentheses. Each cell of the heatmap is colored based on the enrichment scores, ranged from blue to red.
(F) Comparison of hypermethylated probes between IDH mutant (left) and wild-type (right) tumors for PRAD, GBM, and AML.

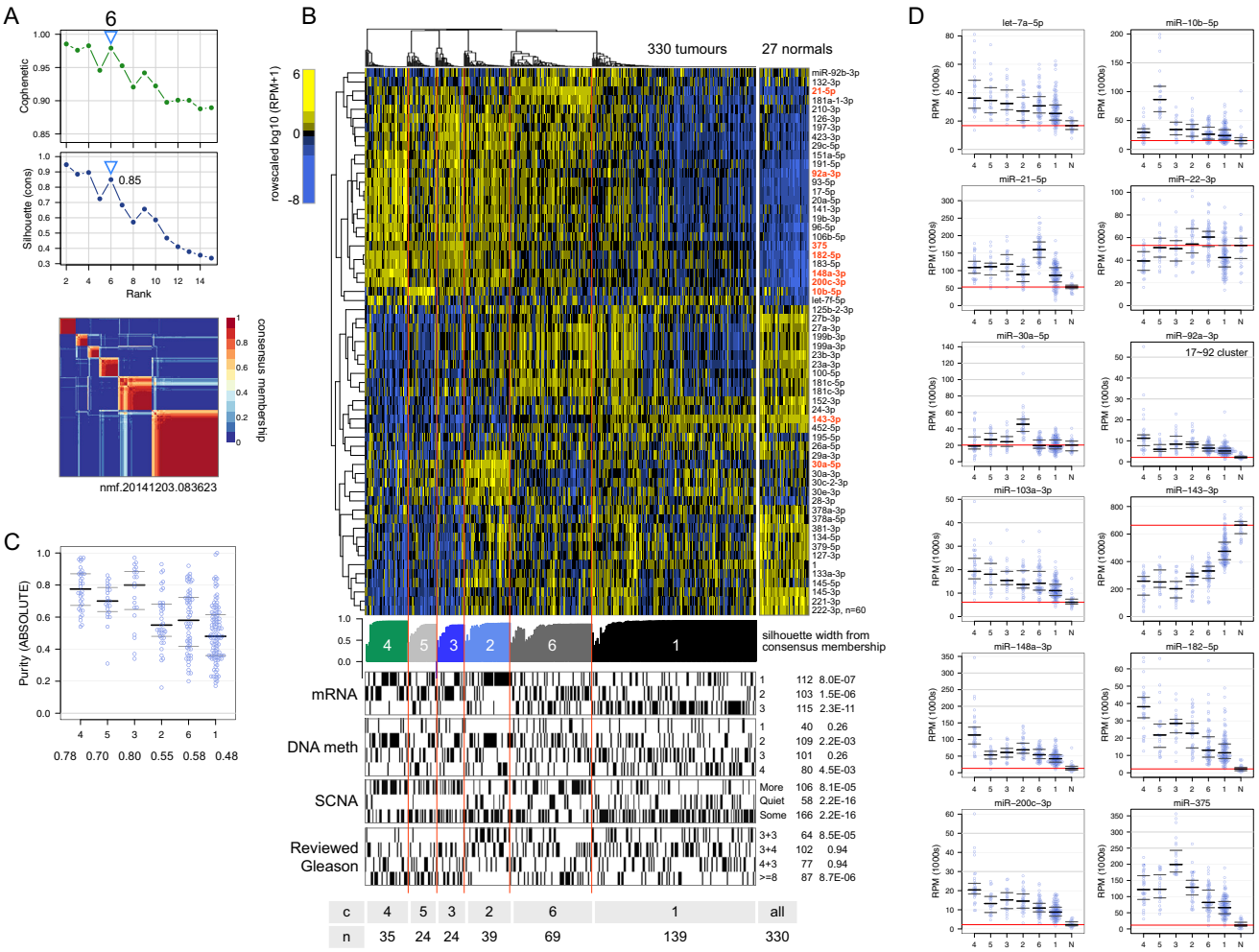


Figure S6. Unsupervised NMF Consensus Clustering of miRNA-Seq 5p and 3p Mature Strand Data for 330 Tumor Samples, Related to Figure 1

(A) Profiles of cophenetic correlation coefficient and consensus average silhouette width from the NMF rank survey, with a red/blue consensus membership heatmap for a six-cluster solution. The clustering input was the miR abundance profiles for the mature strands that were in the upper quartile of variance across the tumor samples.

(B) Normalized abundance heatmap for the six-group solution, with a heatmap for 27 adjacent normal tissues. The heatmaps show the 60 miRs that had the largest scores in a multiclass differential abundance analysis that was done on tumors and tissue normals together. The scale bar shows row-scaled, $\log_{10}$ normalized (reads-per-million, RPM) miR abundance. The miRs whose names are highlighted in red are shown in panel (d). Below the heatmap are (top to bottom) a profile of silhouette width calculated from the consensus membership matrix; tracks for a subset of molecular and clinical covariates, with Fisher exact P-values; and a summary table.

(C) Distributions of tumor purity as reported by ABSOLUTE.

(D) Distributions of RPM abundance for a subset of miRs that were both differentially abundant across the tumor clusters and highly abundant in at least one cluster, or in tissue normal samples. Red horizontal lines mark median RPMs for tissue normals.

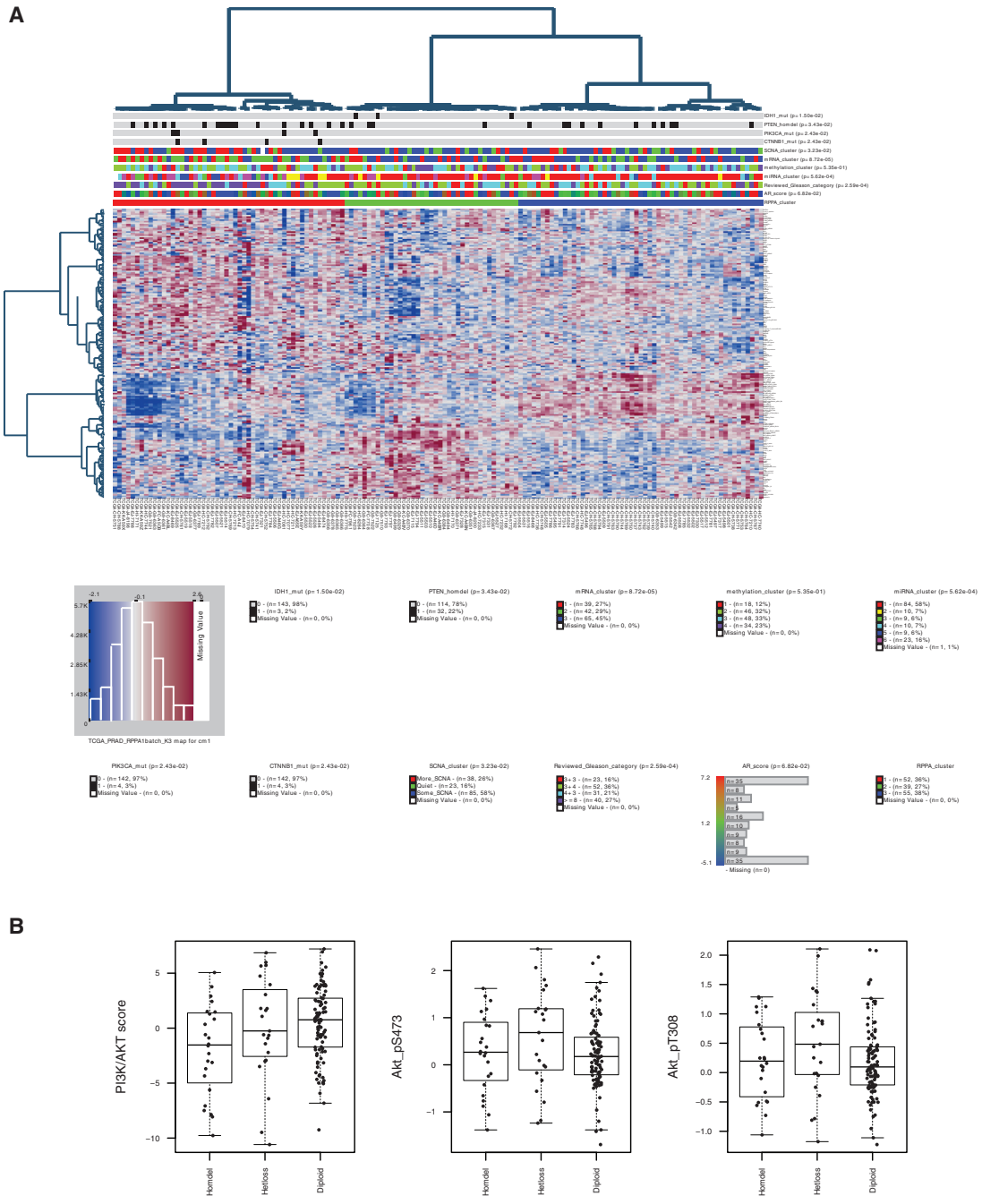


Figure S7. RPPA Clustering and Correlations with PI3K Signaling, Related to Figures 1 and 5

(A) Unsupervised clustering of protein data (RPPA) identified three clusters.

(B) PTEN copy-number status versus PI3K pathway activation. There is no correlation of PTEN deletion status with an overall PI3K/AKT pathway activation score (based on phosphorylation levels of AKT, GSK3, P27, PRAS40, Tuberin, and absolute levels of PTEN and INPP4B) or levels of phospho-Akt alone (S473 and T308).

Cell

Supplemental Information

The Molecular Taxonomy of Primary Prostate Cancer

The Cancer Genome Atlas Research Network

TCGA Prostate Cancer Manuscript - Supplementary Information

Table of Contents

Supplemental Experimental Procedures

Section 1: Biospecimens and Analysis Centers

Sample inclusion criteria

Surgical resection biospecimens were collected from patients diagnosed with prostate adenocarcinoma, and had not received prior treatment for their disease (chemotherapy, radiotherapy, or hormonal ablation therapy). Institutional review boards at each tissue source site reviewed protocols and consent documentation and approved submission of cases to TCGA. Cases were staged according to the American Joint Committee on Cancer (AJCC) and assigned a Gleason score. Shipments from Tissue Source Sites (TSSs) were restricted to increase the number of patients of African descent.

Each frozen primary tumor specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the tissue source site). Seminal vesicle was accepted as a germline control in lieu of blood, and tumor-adjacent prostate was characterized if it was found not to contain tumor by pathology review and was accompanied by DNA from a patient-matched blood specimen. Specimens were shipped overnight from TSSs using a cryoport that maintained an average temperature of less than -180°C.

Pathology quality control was performed on each tumor and normal tissue (if available) specimen from either a frozen section slide prepared by the BCR or from a frozen section slide prepared by the TSS. Hematoxylin and eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumor specimen was histologically consistent with the allowable prostate adenocarcinoma subtypes and the adjacent normal specimen contained no tumor cells. The percent tumor nuclei, percent necrosis, and other pathology annotations were also assessed. Tumor samples with ≥60% tumor nuclei and ≤20% or less necrosis were submitted for nucleic acid extraction.

The TSSs contributing biospecimens used as part of this manuscript include: ABS, Asterand, Inc., Cornell, Fox Chase Cancer Center, Global BioClinical – Georgia, Global Bioclinical – Moldova, Harvard Beth Israel, International Genomics Consortium, Indivumed GmbH, The University of Texas MD Anderson Cancer Center, Melbourne Health, Memorial Sloan-Kettering Cancer Center, National Cancer Institute Urologic Oncology Branch, PROCURE Biobank, Roswell Park Cancer Institute, Stanford University School of Medicine, University Medical Center Hamburg-Eppendorf, University of Arizona, University of California San Francisco, University of Kansas, University of Minnesota, University of North Carolina, University of Pittsburgh, University of Sao Paulo Brazil, Wake Forest University, and Washington University.

Approximately 70% of prostate cancer cases (consisting of a primary tumor and a germline control) submitted to the BCR and processed passed quality control metrics. Tumor tissue from 352 cases was submitted for reverse phase protein array analysis. The data freeze included

333 cases from TCGA batches 91, 108, 161, 184, 221, 244, 268, 285, 308, 315, 320, 331, 348, 357, 370, and 389.

Sample Processing

RNA and DNA were extracted from tumor and adjacent normal tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp blood midi kit (Qiagen).

RNA samples were quantified by measuring Abs_{260} with a UV spectrophotometer and DNA quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify that tumor DNA and germline DNA representing a case were derived from the same patient. Five hundred nanograms of each tumor and normal DNA were sent to Qiagen (Hilden, Germany) for REPLI-g whole genome amplification using a 100 µg reaction scale. RNA was analyzed via the RNA6000 nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only analytes with RIN ≥7.0 were included in this study. Only cases yielding a minimum of 6.9 µg of tumor DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study.

Samples with residual tumor tissue were considered for proteomics analysis. When available, a 10 to 20 mg piece of snap-frozen tumor adjacent to the piece used for molecular sequencing and characterization was submitted to the University of Texas MD Anderson Cancer Center for reverse phase protein array analysis.

Section 2: Clonality Analysis

CLONET DNA based estimates of Tumor Purity and Ploidy

To handle highly heterogeneous and aberrant tumor samples CLONET (CLONality Estimate in Tumors) assesses tumor purity and tumor ploidy by considering the most informative genomic areas (local approach) and then infers clonality of each aberration by taking advantage of the genetic background of each individual. Within its mathematical framework the tool uniformly quantifies clonality of point mutations and copy number changes from DNA sequence based data. Methodological details including the estimates and propagation of uncertainty are described in Prandi et al Genome Biology 2014 (Prandi et al., 2014). Briefly, starting from a set of segmented genomic intervals with uniform tumor over normal signal ratio (referred to as *Log R*) and the read count at germline heterozygous SNP loci for the individual (referred to as *informative SNPs*) compares the empirical distribution of the allelic fraction (AF) of the informative SNPs within a segment *S* with the expected binomial distribution where the distance between the two modes is proportional to the percentage of neutral reads *β*. Neutral reads are those reads that equally represent parental chromosomes (copy number neutral reads), in contrast to reads that originate from only one parent chromosome. For each segment *S*, it then exploits optimization based on swarm intelligence to find a *β* that minimizes the difference between the expected (binomial) and the observed AF distribution. Then, the *Log R* of *S* allows computing a local estimate of the purity. In particular, if the *Log R* value of *S* is compatible with a mono-allelic deletion, a local estimate of the purity is:

$$\text{Purity}_S = 1 - \frac{\beta}{2-\beta}$$

The global estimate of the sample purity is obtained by applying spatial clustering to *β* estimates and selecting the one with the highest median value. The clonality of *S* (the percentage of tumor cells harboring the lesions *S)* is then computed. The wider the difference between local purity and global purity the more *S* is subclonal. Tumor aneuploidy causes a shift in the values of the *Log R* of *S* and may result in the misinterpretation of the copy number of *S* and in turn in a poor estimation of the sample purity. To assess the extent of *Log R* shift to correctly interpret the copy number of *S,* CLONET utilizes genomic segments with neutral reads only (β equal to 1). The ploidy of a sample is then inferred using the shift in the *Log R* values of the neutral segment that best accounts for the observed *Log R* values (**Fig. S1A**).

Tumor purity and ploidy were estimated using default parameters for each study sample and also applied to adjust segmented data. Tumor evolution patterns are built upon clonality estimates and precedence relations among co-occurring aberrations as previously reported in (Baca et al., 2013).

Transcript Tumor Score

Using six large studies represented in the Oncomine software (Grasso et al., 2012; Wallace et al., 2008; Tomlins et al., 2007; Lapointe et al., 2004; Yu et al., 2004; Singh et al., 2002) we

selected genes that were consistently up- or down- regulated in prostate tumors versus normal comparisons. Median rank of the differential expression for the selected genes across six studies was less or equal to 150. Transcript Tumor Score (TTS) was computed as (t.s + (1-n.s))/2, where t.s and n.s were single sample Gene Set Enrichment Analysis (ssGSEA) scores computed separately for genes up-regulated in tumors and up-regulated in normals.

Section 3: Mutational Analysis

DNA sequencing and data processing

Whole exome sequencing (WES, n=333 tumor/normal sample pairs) and high coverage whole genome sequencing (WGS, n=20 tumor/normal sample pairs) were performed as previously described (The Cancer Genome Atlas Network, 2014). In brief, 0.5–3 micrograms of DNA from each sample were used to prepare the sequencing library through shearing of the DNA followed by ligation of sequencing adaptors. Whole exome capture was performed using Agilent SureSelect Human All Exon (http://www.genomics.agilent.com/en/Exome-Sequencing/SureSelect-Human-All-Exon-Kits) protocol according to the manufacturers' instructions. Exome capture regions were based on protein coding regions as defined by the consensus CDS (CCDS; http://www.ncbi.nlm.nih.gov/projects/CCDS/) and RefSeq genes (http://www.ncbi.nlm.nih.gov/RefSeq/). Thus, 188,260 exons from ~18,560 genes (93% of known, non-repetitive protein coding genes) that spans ~1% of the genome (32.7 Mbps) were sequenced. Whole exome and whole genome sequencing was performed on the Illumina HiSeq 2000 platform using the V3 Sequencing Kits (http://support.illumina.com/sequencing/sequencing_instruments) and the Illumina 1.3.4 pipeline to produce paired-end sequenced data (2x101 bp for WGS to roughly 30x read coverage and 2x76 bp for WES to roughly 100x read coverage). The "Picard" and "Firehose" pipelines at the Broad Institute were used for basic alignment and sequence QC.

Sequencing data-processing pipeline ("Picard pipeline")

The "Picard" pipeline (http://picard.sourceforge.net/) generates a BAM file (http://samtools.sourceforge.net/SAM1.pdf) for each sample. Specifically, the pipeline aggregates data from multiple libraries and flow cell runs into a single BAM file for a given sample. The BAM file contains reads aligned to the human genome with quality scores recalibrated using Genome Analysis Toolkit's Table Recalibration tool. The reads in the file were aligned to the Human Genome Reference Consortium build 37 (GRCh37) using BWA v0.5.9 (Li and Durbin, 2010) (http://biobwa.sourceforge.net/). Unaligned reads that passed the Illumina's quality filter (PF reads) were stored in the BAM file as well. All duplicate reads were marked and removed, and thus, unique sequenced DNA fragments were used in subsequent analysis. Sequence reads corresponding to genomic regions that may have small insertions or deletions (indels) were jointly realigned to improve detection of indels and to decrease the number of false positive single nucleotide variations which are a consequence of misaligning reads, particularly at the 3' end. Thus, all sites potentially harboring small insertions or deletions in either the tumor or the matched normal were realigned in all samples. Finally, the Picard pipeline provided summary QC metrics such as the target coverage and an estimated level of "oxo-G" artifacts (Costello et al., 2013) for each BAM that were used in subsequent processing.

Cancer genome analysis pipeline ("Firehose")

The Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose) performed additional QC on the BAM files, mutation calling, small insertion and deletion detection, and annotation of point mutations and indels. These steps are described in further detail below.

1. QC on BAM files: The sample cross-individual contamination levels were estimated using the ContEst program (Cibulskis et al., 2011). Tumor normal pairs of samples with contamination less than 4% were used further downstream for analysis.
2. Somatic mutation calling: The MuTect algorithm (Cibulskis et al., 2013) was used to detect somatic single nucleotide variants (SSNVs).
3. Small insertion and deletion detection: The Indelocator algorithm (https://www.broadinstitute.org/cancer/cga/indelocator) was used to detect small indels.
4. Mutations and indels annotations: Point mutations and indels detected by respective MuTect and Indelocator were annotated using utility named Oncotator (Lee et al., 2015). Oncotator mapped somatic mutations to respective genes, transcripts, and other relevant features. These annotations correspond to the fields in the Mutation Annotation Format (MAF) files version 2.4 (https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification).

Post-processing ("Panel of Normals filtering")

Following Firehose processing, we employed various strategies to identify and filter out false somatic point mutations and indels. We used Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013) for the manual review of sequencing evidence in the tumor and normal samples. In several cases, IGV was even used to identify mutations that were otherwise missed by the detection pipeline (e.g. orientation bias artifacts in MLLT10). In addition, we used a representative panel of 4513 normal WES bams to model a wide range of sequencing or alignment artifacts, or rare germline mutations that might be misidentified as a somatic point mutation or indel. The Panel of Normals (PoN) filter removed any mutation with a corresponding alternate allele appearing in more than 0.2% of reads covering a given site or alternate allele appearing in more than 0.2% of the PoN samples. The PoN filter removed nearly half of SSNV and nearly all of the somatic indels detected by the Firehose pipeline. The large proportion of the calls removed by the PoN filter is a consequence of the low density of true somatic mutations in PRAD compared to the rate of false detection inherent in DNA sequencing technologies and detection methods. In order to ensure that no candidate driving mutations were mistakenly removed by the post-processing filtering, previously implicated cancer genes' candidate mutations were manually reviewed using IGV.

Significantly Mutated genes

After filtering for artifacts and defining a final set of mutations, the MAF was analyzed to determine significantly mutated genes. This was accomplished using the MutSig2CV v1.2 (Lawrence et. al., 2014).

Section 4: Somatic Copy-Number Alterations

SNP-based copy number analysis

DNA from each tumour or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described (McCarroll et., 2008). Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus (Korn et al., 2008). For each tumour, genome-wide copy number estimates were refined using tangent normalization, in which tumour signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumour (Cancer Genome Atlas 2011; Tabak and Beroukhim Manuscript in preparation). This linear combination of normal samples tends to match the noise profile of the tumour better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation (Olshen et al., 2004). As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumour and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile (Mermel et al., 2011). Significant focal copy number alterations were identified from segmented data using GISTIC 2.05. Tumors were clustered based on chromosomal arm level alterations. In arm level analysis, chromosomal arms were considered altered if at least 60% of the arm was lost or gained with a relative log2 copy number change greater the 0.1. Clustering was done in R based on Manhattan distance using Ward's method. Purity and ploidy estimates, were calculated using the ABSOLUTE algorithm (Carter et al., 2011). Allelic copy number derived from ABSOLUTE was used along with relative copy number to determine regions of loss of heterozygosity and homozygous deletions.

Section 5: Gene Fusions

Detection of chimeric transcripts with FusionSeq

FusionSeq, a modular computational framework for the detection of chimeric transcripts, was applied on the set of 333 tumor specimens (Sboner et al., 2010). FusionSeq is composed of two main modules: 1. Fusion Detection module: it detects all chimeric paired-end (PE) reads, i.e. those reads from a single fragment of mRNA but with their ends mapped to different genes, thus suggesting the mRNA fragment was generated by a gene fusion event; 2. Filtration Cascade module: it removes many artifacts due to several sources of errors and provides a high-confidence list of fusion candidates. For the TCGA-PRAD study, FASTQ files were re-mapped to the human genome reference sequence (hg19) using STAR v2.3.0e (Dobin et al., 2013), and converted into Mapped Read Format (MRF) using RSEQtools, a suite of tools for RNA-seq data processing and analysis (Habegger et al., 2011). MRF files include only the primary alignments as determined by STAR and do not include reads mapped to the mitochondrial chromosome. MRF files were used as input to FusionSeq. The Fusion Detection module was applied (geneFusions) and detected all chimeric PE reads involving genes reported in the UCSC knownGenes annotation dataset (downloaded from UCSC on 10 September 2013). FusionSeq was run in two separate modes: 1. High Sensitivity, where the focus was to identify well-characterized chimeric transcripts; and 2. High Specificity, where the focus was to provide a high-confidence list of fusion candidates involving any gene.

FusionSeq – High Sensitivity mode

In this mode, FusionSeq is presented with a list of well-characterized fusions and it will report any evidence of these chimeric fusions in the data. Fusions involving ERG, ETV1, ETV4, or FLI (ETS canonical fusions) were considered here.

FusionSeq – High Specificity mode

In this mode, FusionSeq is applied using a more conservative approach for evidence of fusion transcripts. After the Fusion Detection module, a series of computational filters are applied to reduce the amount of artifacts (Filtration Cascade module). Artifacts arise for different reasons: mis-mapping because of sequencing errors, of homologous regions, of SNPs or mutations in the tumor, or because of the generation of chimeric fragments during library preparation (Sboner et al., 2010). The remaining candidates were classified as inter-, intra-chromosomal, and read-through events (chimeric transcripts involving nearby genes in the genomic space). Finally, the candidates are then ranked by using DASPER, a statistical measures that takes into account the relative expression of the genes involved in the chimeric transcript and the evidence of the fusion provided by the chimeric reads (Sboner et al., 2010). The higher the DASPER score, the higher the chance of a real fusion transcript.

The results obtained with both high-sensitive and high-specificity modes were then combined with the predictions from Mapsplice and presented in **Table S1E**.

Low-pass WGS library construction

Between 500 and 700 ng of each gDNA sample were sheared using Covaris E220 to about 250 bp fragments, then converted to a pair-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturer's protocols. All libraries were sequenced on HiSeq2000 using one sample per lane, with the pair-end 2 x 51bp setup. Tumor and its matching normal were usually loaded on the same flow cell. Raw data were converted to the FASTQ format and BWA alignment was used to generate bam files.

Detection of structural rearrangements in low-pass WGS data

We used two algorithms, BreakDancer (Chen et al., 2009) and Meerkat (Yang et al., 2013), to detect structural variation. The first step in BreakDancer requires a configuration file of each bam file for each tumor pair with the bam2cfg.pl perl module of the program. The perl module BreakDancerMax.pl is then run on the configuration file to call structural variants in the tumor and control files. The set of structural variant calls from each tumor sample is filtered by the calls from its matched normal to remove germ-line variants. Structural variations were also detected by Meerkat, which requires at least two discordant read pairs supporting each event and at least one read covering the breakpoint junction. Variants detected from tumor genomes were filtered by the variants from all normal genomes to remove germ-line events and were also filtered out if both breakpoints fall into simple repeats or satellite repeats. The final call has to fulfill the following: (i) the read identified to span the breakpoint junction hit predicted breakpoint region uniquely by BLAT, or (ii) the mate of the read spanning the breakpoint junction is mapped near the predicted breakpoint.

Section 6: Androgen Receptor Analysis

AR output score analysis

The AR output score is derived from the mRNA expression of genes that are experimentally validated AR transcriptional targets (Hieronymus et al., 2006). Precisely, a list of 20 genes upregulated in LNCaP cells stimulated with the synthetic androgen R1881 was used as a gene signature of androgen-induced genes. An AR output score was defined by the quantification of the composite expression of this 20-gene signature in each sample. Here, we measured differential AR activity between genomic subtypes (ERG, ETV1/4/FLI1, SPOP, FOXA1, other, normal prostate). To this aim, we computed a Z-score for the expression of each gene in each sample by subtracting the pooled mean from the RNA-seq expression values and dividing by the pooled standard deviation.

The AR output score for each sample is then computed as the sum of the Z-scores of the AR signaling gene signature.

AR-V7 splice variant detection analysis

RNA-seq reads were mapped to all known genes and isoforms as previously described (Dvinge et al., 2014), in addition to a manually curated list of all AR splice variants. Gapped reads spanning uniquely identifying splice junctions were used as a measure of the presence of processed AR variants if there was at least one read spanning the 3' end of the upstream exon and the 5' end of the cryptic or downstream exon, with a minimum of 6nt overhang on either side without mismatches. The number of reads spanning AR variant splice junctions were scaled using the total number of reads spanning the first splice junction (exon 1-2 or exon 1a-2), which is present in all AR transcripts.

AR Splice Variant Verification by qPCR

2-3 µg total RNA samples were arrayed into a 96-well plate and polyadenylated (PolyA+) RNA was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) with on-column DNaseI-treatment as per the manufacturer's instructions. The eluted PolyA+ RNA was ethanol precipitated and resuspended in 10µL of DEPC treated water with 1:20 SuperaseIN (Life Technologies, USA). Double-stranded cDNA was synthesized from the purified polyA+RNA using Maxima H Minus reverse transcriptase (Life Technologies, USA) and random hexamer primers at a concentration of 5µM.

The qPCR primer sequences for the variant AR-V7 (AR-V7-F: 5'-CCATCTTGTCGTCTTCGGAAATGTTATGAAGC, and AR-V7-R: 5-TTTGAATGAGGCAAGTCAGCCTTTCT)
were as reported in a previous publication (Hu et al., 2009). The wild-type AR qPCR primers consisted of AR-F (5'- ATCCTCATATGGCCCAGTGTCAAG) and AR-R (5'-GCTCTCTAAACTTCCCGTGGCATA). The qPCR primers for the control gene consisted of RPL13A-F (5'- CCTGGAGGAGAAGAGGAAAGAGA) and RPL13A-R (5'-TTGAGGACCTCTGTGTATTTGTCAA).

cDNA for 80 samples (75 primary tumors, 5 adjacent normals) was quantified using the Quant-iT dsDNA HS Assay Kit (Life Technologies) on a VICTOR3V plate reader (Perkin Elmer). qPCR was performed in triplicate using the iTaq Universal SYBR Green Supermix kit (Bio-Rad) on a CFX384 Touch Real-Time System (Bio-Rad). All three primers pairs for each template were processed on the same 384-well plate. qPCR reactions were set up according to manufacturer's specifications in 10μL reaction volume with 5μL of 2x Supermix, 0.25μL of forward primer (10μM), 0.25μL of reverse primer (10μM), and 100ng of cDNA template. The cycling conditions consisted of one cycle of 50°C for 30 sec and 95°C for 10 min, followed by 40 cycles of 95°C for 10 sec and 60°C for 30 sec. The conditions for the melting curve analysis were according to the instrument's default setting of 65°C to 95°C with 0.5°C increments.

Section 7: DNA Methylation Analysis

Array-based DNA methylation assay

We used Illumina Infinium HumanMethylation450 (HM450) platform (Bibikova et al., 2011) to obtain DNA methylation profiles of 333 prostate cancer tissue samples and 19 adjacent non-malignant prostate tissue samples. Each batch of samples was assayed with control cell line technical replicates to monitor technical variations. The HM450 array contains 485,777 probes, which include 482,421 CpG sites, 3,091 non-CpG (CpH) sites, and 65 SNPs in the human genome. The array interrogates 96% of CpG islands, 92% of CpG shores, and 99% of RefSeq genes with multiple probes per gene located in promoter, 5'UTR, first exon, gene body, and 3'UTR regions. The detailed information of the HM450 array is available on the Illumina website (www.illumina.com).

Sample and data processing

For each sample, 1ug of genomic DNA was converted with sodium bisulfite using the EZ-96 DNA methylation kit as recommended by the manufacturer (Zymo Research, Irvine, CA). Quality control (QC) assays were performed on each sample to assess the amount of converted DNA and the completeness of bisulfite conversion using MethyLight reactions as previously described (Campan et al., 2009). All samples passed the QC tests, and bisulfite-converted DNAs were then whole-genome amplified (WGA), enzymatically fragmented, and hybridized to HM450 arrays. HM450 arrays were subsequently scanned with Illumina iScan technology. Level 1 IDAT data files were imported for processing using the R/Bioconductor package *methylumi* (Methylumi, 2014, Triche et al., 2013). Level 2 and level 3 DNA methylation data of TCGA prostate samples were generated by using the *EGC.tools* R package (https://github.com/uscepigenomecenter/EGC.tools).

TCGA DNA methylation Data packages

Three levels of TCGA prostate adenocarcinoma (PRAD) DNA methylation data are available from TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga).

Level 1 data contain raw IDAT files, in which the cy3 and cy5 signal intensities are embedded. Level 2 data contain background-corrected intensities of methylated (M) and unmethylated (U), and detection P-values of each probe. Level 3 data contain β values (M/(M+U)) with annotations for the HUGO Gene Nomenclature Committee (HGNC) gene symbol, chromosome, and genomic coordinate of each CpG/CpH site (UCSC hg19, Feb 2009). Probes that meet the following criteria are masked as "NA": Probes that 1) overlap with a common SNP (dbSNP build 135, minor allele frequency >1%) within 10 bp of the interrogated CpG site, 2) are located within 15bp of a repetitive element (RepeatMasker and Tandem Repeats Finder from UCSC hg19, Feb 2009), 3) are aligned to multiple sites on human genome (UCSC hg19, Feb 2009), and 4) have detection P values greater than 0.05 for a specific data point.

The following data archives were used for the analyses described in this manuscript.

jhu-usc.edu_PRAD.HumanMethylation450.Level_3.1.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.2.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.3.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.4.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.5.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.6.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.7.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.8.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.9.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.10.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.11.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.12.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.13.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.14.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.15.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.16.12.0
jhu-usc.edu_PRAD.HumanMethylation450.Level_3.17.12.0

Unsupervised clustering of DNA methylation data

To identify cancer-specific DNA hypermethylation events, we selected 155,708 probes that were unmethylated in adjacent non-malignant prostate tissue samples (mean $\beta<0.2$). Among these probes, we selected 32,936 probes that were methylated ($\beta>0.3$) in at least 5% of tumors (n=17) for a preliminary clustering analysis. However, a clustering analysis using β values for this set of probes was strongly confounded by tumor purity. In order to alleviate the influence of variable tumor purity levels on a clustering result, the following steps were conducted: First, among 32,936 probes, we only included probes that were not associated with tumor purity after performing linear regression between DNA methylation levels and tumor purity using ABSOLUTE (adj. p-value > 0.05) (n=11,927). Next, we selected the most variably hypermethylated probes for clustering in order to assess heterogeneous DNA methylation levels among tumor samples (n=5,000). Finally, to reduce additional influence of tumor purity on DNA methylation levels, we dichotomized the data using a β value of >0.3 as a threshold for positive DNA methylation. Unsupervised hierarchical clustering was performed with the dichotomized data using the binary distance metric for clustering and Ward's method for linkage from the R/Bioconductor software packages (http://www.bioconductor.org). DNA methylation cluster assignments were generated by cutting the resulting dendrogram. **Fig. S5A** displays a heat map of the original β values for the 5,000 most variably hypermethylated probes. The heatmap was organized based on the order of unsupervised hierarchical clustering of the dichotomized data.

Clustering analysis with rules-based classification

Prostate tumors were classified to eight groups based on somatic mutation and fusion states: ERG, ETV1, ETV4, FLI1, SPOP, FOXA1, IDH1, and others. The heatmap of DNA methylation data for the 5,000 most variably hypermethylated probes, organized by these eight subgroups, is shown in **Fig. 3A**. Within each group, tumors were ordered by DNA methylation cluster assignments. Probes were sorted based on the order of unsupervised hierarchical clustering results shown in **Fig. S5A**.

DNA hypermethylation frequency

We identified a set of 155,708 probes that were unmethylated in adjacent non-malignant prostate tissue samples (mean $\beta<0.2$). In order to compare cancer-specific DNA hypermethylation events in DNA methylation clusters and rules-based classification, we dichotomized the PRAD DNA methylation data using a β value of >0.3 as a threshold for positive DNA methylation. The hypermethylation frequencies of tumors, grouped by DNA methylation clusters and fusion/mutation subgroups, are shown in **Fig. S5B**.

Comparison of DNA methylation clusters with Gleason scores and other platform clusters

We calculated enrichment scores for each DNA methylation cluster with other features, including Gleason Score and cluster assignments from iCluster, RPPA, SCNA, mRNA and miRNA data sets (**Fig. S5E**). The enrichment score was calculated by using the ratio between the observed number of tumors and the expected number of tumors.

Comparison of *IDH* mutations among PRAD, GBM, and AML tumors

In order to compare cancer-specific DNA hypermethylation events in tumors harboring *IDH* mutations, we obtained TCGA HM450 data for prostate cancer (PRAD), glioblastoma multiforme (GBM), and acute myeloid leukemia (AML) and normal samples for prostate and brain tissues from the TCGA Data Portal. Specifically, we identified three *IDH1* mutated PRAD tumors, five *IDH1* mutated GBM tumors, as well as 14 *IDH1* and 12 *IDH2* mutated AML tumors. In comparison, we identified 330 PRAD, 104 GBM, 108 AML, 19 normal prostate tissue samples, and two normal brain samples with wild type *IDH*. For normal blood samples, CD34+CD38- hematopoietic stem/progenitor cells, promyelocytes, neutrophils and monocytes, were obtained from the GSE49618 data set in the GEO database (n=15). After identifying unmethylated probes in corresponding normal tissue samples (155,708, 155,806, and 183,030 probes for prostate, brain, and blood, respectively), we selected a set of 139,905 probes that were unmethylated in all normal tissue samples to calculate hypermethylation frequencies. β values of tumors were dichotomized using a $\beta>0.3$ as a threshold for positive DNA methylation. The hypermethylation frequency of each *IDH* genotype across PRAD, GBM and AML tumors is indicated in **Fig. 3B**. In order to compare DNA hypermethylation events between *IDH* mutant and wild type PRAD, GBM and AML tumors, the intersection of hypermethylated probes in each group (mean $\beta>0.3$) was determined by generating Venn diagrams using the R/bioconductor package *Vennerable* (**Fig. S5F**).

Identification of Epigenetically Silenced Genes

Level 3 HM450 PRAD data and $\log_2$-transformed level 3 PRAD RNA-seq RSEM data ($\log_2$(RSEM+1)) were used to identify epigenetically silenced genes, following the criteria listed below:

1. *Identification of hypermethylated probes located in gene promoter regions*: Probes located within a 3kb spanning from 1,500bp upstream to 1,500bp downstream of transcription start sites were selected as promoter loci. Among these probes, only those unmethylated in normal prostate tissues (mean $\beta<0.2$) but methylated ($\beta>0.3$) in at least 5% of tumors (n=17) were chosen (21,256 probes for 6,321 genes).
2. *Selection of epigenetically silenced genes*: In order to find genes whose expression levels were inversely correlated with DNA methylation levels, we selected tumors for which at least one allele of the afore described hypermethylated promoter loci was present using GISTIC2 CNV calls. Tumor samples were then grouped as methylated ($\beta>0.3$) or unmethylated ($\beta<0.3$) for each probe, and the corresponding expression levels were computed. We identified probes for which the mean expression in the methylated group was lower than 1.645 standard deviations (bottom 5%) of the mean expression in the unmethylated group. Among the identified probes/genes, only genes with a maximum $\beta>0.6$ and having epigenetically silencing events in more than 1% of tumors (n=3) were included.
3. *Identification of tumors with epigenetically silenced genes:* Each tumor sample was labeled as epigenetically silenced when it belonged to the methylated group with lower gene expression level than mean expression in the unmethylated group. In the instances in which multiple probes were available for a specific gene, we only included tumor samples that were silenced at more than half of the probes for the promoter region of that gene.

Using the above method, we identified 164 epigenetically silenced genes in prostate tumor samples (**Table S1F**). Gene set enrichment analysis of identified epigenetically silenced genes was performed using the GSEA tool (http://www.broadinstitute.org/gsea, Subramania et al., 2005)(ref: 16199517). Hypergeometric test was used to calculate p-value, and false discovery rate (q-value) < 0.05 was used to select significantly enriched gene sets. Associations of epigenetically silenced genes with DNA methylation clusters and fusion/mutation subgroups were tested by Fisher's exact test. To account for multiple-testing bias, the p-value was adjusted using the Benjamini-Hochberg correction.

Scatterplot visualization of epigenetically silenced genes

In order to visualize epigenetically silenced genes in prostate tumors, we used level 3 HM450 PRAD DNA methylation data and $\log_2$-transformed level 3 PRAD RNA-seq RSEM data ($\log_2$(RSEM+1)) to generate scatterplots. We identified 693 probe/gene pairs for 164

epigenetically silenced genes in TCGA PRAD tumors. Scatterplots of five epigenetically silenced genes were displayed in **Fig. -S5C** as examples.

Identification of Up-regulated genes with DNA hypermethylation

Level 3 HM450 PRAD data and $\log_2$-transformed level 3 PRAD RNA-seq RSEM data ($\log_2$(RSEM+1)) were used to identify up-regulated genes with DNA hypermethylation, following the criteria listed below:

1. *Identification of hypermethylated probes located in gene promoter regions*: Probes located within a 3kb spanning from 1,500bp upstream to 1,500bp downstream of transcription start sites were selected as promoter loci. Among these probes, only those unmethylated in normal prostate tissues (mean $\beta<0.2$) but methylated ($\beta>0.3$) in at least 5% of tumors (n=17) were chosen (21,256 probes for 6,321 genes).
2. *Selection of up-regulated genes with DNA hypermethylation*: In order to find genes whose expression levels were positively correlated with DNA methylation levels, we selected tumors for which at least one allele of the afore described hypermethylated promoter loci was present using GISTIC2 CNV calls. Tumor samples were then grouped as methylated ($\beta>0.3$) or unmethylated ($\beta<0.3$) for each probe, and the corresponding expression levels were computed. We identified probes for which the mean expression in the methylated group was higher than 1.645 standard deviations (bottom 5%) of the mean expression in the unmethylated group. Among the identified probes/genes, only genes with a maximum $\beta>0.6$ and having elevated gene expression with hypermethylation events in more than 1% of tumors (n=3) were included.
3. *Identification of tumors with DNA hypermethylated up-regulated genes*: Each tumor sample was labeled as a tumor with DNA hypermethylated up-regulated genes when it belonged to the methylated group with higher gene expression level than mean expression in the unmethylated group. In the instances in which multiple probes were available for a specific gene, we only included tumor samples that were silenced at more than half of the probes for the promoter region of that gene.

Using the above method, we identified one gene, *CELSR3*, displayed increased DNA methylation associated with elevated expression in some of prostate tumor samples. DNA methylation levels of three HM450 probes near transcription state site of the gene and gene expression levels were visualized in the scatterplots (**Fig. S5D**).

Section 8: MicroRNA Analysis

Libraries and sequencing

We generated microRNA sequence (miRNA-seq) data for 330 tumor and 27 adjacent normal samples using methods described previously (Cancer Genome Atlas Research Network, 2012). We aligned reads to the GRCh37/hg19 reference human genome, and annotated miRNA read count abundance with miRBase v16. While we used only exact-match read alignments for quantifying miRNA abundance, BAM files are available from CGHub (cghub.ucsc.edu, Wilks et

al., 2014) that include all sequence reads. We used miRBase v20 to assign 5p and 3p mature strand (miR) names to MIMAT accession IDs.

Unsupervised clustering

We identified groups of samples that had similar abundance profiles using unsupervised non-negative matrix factorization (NMF) consensus clustering (v0.5.06) in R, with default settings (Gaujoux and Seoighe, 2010). The input was a reads-per-million (RPM) data matrix for the ~300 (25%) most-variant 5p or 3p mature strands, which we parsed (by MIMAT accession ID) from the level 3 isomiR data files that are available from the TCGA data portal. After running a rank survey with 30 iterations per solution, we chose a preferred clustering solution from profiles of the cophenetic correlation coefficient and the average silhouette width calculated from the consensus membership matrix, and performed a 200-iteration run to generate the final clustering solution. To support identifying less-typical cluster members within a cluster, we calculated a profile of silhouette widths from the final NMF consensus membership matrix. To generate a heatmap for the NMF results, we first identified miRs that were differentially abundant between the unsupervised miRNA clusters, using a multiclass analysis with SAMseq (samr 2.0, Li and Tibshirani, 2013) in R, with a read-count input matrix and an FDR threshold of 0.05. For the heatmap, we included miRs that had the largest SAMseq scores and median abundances greater than 25 RPM. The RPM filtering acknowledged potential sponge effects from competitive endogenous RNAs (ceRNAs) that can make weakly abundant miRs less influential (Mullokandov et al., 2012; Tay et al., 2014). We transformed each row of the matrix by $\log_{10}(\text{RPM} + 1)$, then used the pheatmap v0.7.7 R package to scale and cluster only the rows, with a Euclidean distance measure. miR abundance (RPM) distributions across the unsupervised clusters were visualized using the beeswarm (v0.1.6) R package.

Covariates

For clinical and molecular covariates, we calculated contingency table association *P*-values with a Fisher exact test for categorical data.

Purity and ploidy

Tumor sample purity and ploidy were calculated by the Broad Institute using ABSOLUTE (Carter et al., 2012). Purity distributions were visualized using the beeswarm (v0.1.6) R package.

miR targeting

We assessed potential miR-gene targeting in the tumor samples by calculating miR-mRNA and miR-RPPA Spearman correlations with the MatrixEQTL v2.1.1 (Shabalin, 2012) R package, using gene-level normalized abundance RNAseq (RSEM) and RPPA data matrices from Firehose (gdac.broadinstitute.org). We calculated correlations with a P-value threshold of 0.05, and filtered the resulting anticorrelations at FDR<0.05. We then extracted miR-gene pairs that corresponded to functional validation publications reported by MiRTarBase v4.5 (Hsu et al.,

2014), for stronger (luciferase reporter, qPCR, Western blot) and weaker experimental evidence types. We displayed anticorrelation results as networks with Cytoscape 2.8.3.

Differential abundant miRs

We identified miRs that were differentially abundant between pairs of sample groups with unpaired two-class SAMseq analyses, and across sets of more than two groups with multiclass SAMseq analyses, using a read-count input matrix and an FDR threshold of 0.05. We removed miRs that had a Wilcoxon adjusted P-value > 0.05 and with median abundance less than 50 RPM in one of the two groups being compared.

Section 9: RPPA Analysis

RPPA experiments and data processing

Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 mmol/L Hepes (pH 7.4), 150 mmol/L NaCl, 1.5 mmol/L $MgCl_2$, 1 mmol/L EGTA, 100 mmol/L NaF, 10 mmol/L NaPPi, 10% glycerol, 1 mmol/L phenylmethylsulfonyl fluoride, 1 mmol/L Na_3VO_4, and aprotinin 10 µg/mL) from human tumors and RPPA was performed as described previously (Tibes et al., 2006; Liang et al., 2007; Hu et al., 2007; Hennessy et al., 2007; Coombes et al., 2011). Lysis buffer was used to lyse frozen tumors by Precellys homogenization. Tumor lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumor lysates were manually serial diluted in two-fold of 5 dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 190 validated primary antibodies followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation-catalyzed system and DAB colorimetric reaction. Slides were scanned in a CanoScan 9000F. Spot intensities were analyzed and quantified using Array-Pro Analyzer (Media Cybernetics Washington DC) to generate spot signal intensities (Level 1 data). The software SuperCurveGUI (Hu et al., 2007; Coombes et al., 2011), available at http://bioinformatics.mdanderson.org/Software/supercurve/, was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model (Tibes et al., 2006). During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric (Coombes et al., 2011) was returned for each slide to help determine the quality of the slide: if the score is less than 0.8 on a 0-1 scale, the slide was dropped. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained for an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described (Hu et al., 2007;Coombes et al., 2011; Gonzalez-Angulo et al., 2011) using median centering across antibodies (level 3 data). In total, 190 antibodies and 152 samples were used. Final selection of antibodies was also driven by the availability of high quality antibodies that consistently pass a strict validation process as previously described [7]. These antibodies are assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumor tissue. Antibodies are labeled as validated and use with caution based on degree of validation by criteria previously described (Hennessy et al., 2010).

Two RPPA arrays were quantitated and processed (including normalization and load controlling) as described previously, using MicroVigene (VigeneTech, Inc., Carlisle, MA) and the R package SuperCurve (version-1.3), available at http://bioinformatics.mdanderson.org/OOMPA (Tibes et al., 2006; Liang et al., 2007; Hu et al., 2007). Raw data (level 1), SuperCurve non-parametric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the DCC.

Data normalization

We performed median centering across all the antibodies for each sample to correct for sample loading differences. Those differences arise because protein concentrations are not uniformly distributed per unit volume. That may be due to several factors, such as differences in protein concentrations of large and small cells, differences in the amount of proteins per cell, or heterogeneity of the cells comprising the samples. By observing the expression levels across many different proteins in a sample, we can estimate differences in the total amount of protein in that sample vs. other samples. Subtracting the median protein expression level forces the median value to become zero, allowing us to compare protein expressions across samples.

Prostate samples were processed in two RPPA batches. However, batch effects analysis revealed that the samples had large batch effects upstream of the RPPA platform, because controls run on both platforms didn't show any batch effects. Consequently, we decided to use 152 samples from only one RPPA set in our analysis.

Consensus clustering

We used consensus clustering to cluster the prostate samples (**Fig. S7**). Pearson correlation was used as distance metric and Ward was used as a linkage algorithm in the clustering analysis. A total of 152 samples and 190 antibodies were used in the analysis. We identified three robust sample clusters. Cluster 1 had low AKT/PI3K pathway, RTK pathway, RAS/MAPK and TSC/mTOR pathways, but high apoptosis and DNA damage response pathways. It had enrichment for mutations in CTNNB1 gene and a large fraction of high Gleason scores. Cluster 2 had high EMT pathway score and was depleted in CTNNB1 and RYBP mutations. It had a moderate fraction of high Gleason scores. Cluster 3 had low apoptosis and DNA damage response pathways, but high RAS/MAPK, AKT/PI3K, TSC/mTOR and RTK pathways. It had enrichment of RYBP mutations, no TP53 mutations, and had a high proportion of low Gleason scores (<= 7). The analysis revealed strong association of pathways with Gleason scores and RPPA clusters.

Section 10: RNA degradation analysis

Batch effect analysis

To ensure the highest quality data set for analysis, we performed batch effect detection across all cases and data platforms [http://bioinformatics.mdanderson.org/main/TCGABatchEffects:Overview].

This revealed an unusual distribution of shipping batches in individual clusters of hierarchically clustered RNA and protein expression data. Reasoning that identifying the source of this effect and eliminating affected cases across all batches was less susceptible to the introduction of artifacts from batch effect correction, we found that many samples in the affected cluster as well as some samples in other clusters showed increased levels of 5’ RNA degradation. We developed a scoring method to determine the extent of degradation in each samples and removed cases, including all their DNA measurements, from any batch on the basis of increased 3’ / 5’ bias.

3'/5' Bias Calculation

We have retrieved exon quantifications for all PRAD samples from firebrowse.org, which are based on the Firehose pipeline. All exon quantifications are measured in RPKM. We have also retrieved the exon annotations in GAF format used to generate these quantifications. Based on this annotation we have selected all genes which are not alternatively spliced. Further we have also selected genes, which have at least two constitutive exons. Each transcript length is dened as the sum over all coding base pairs plus half the length of the rst and the last exon. The gene length is then defined as the median transcript length. Genes which had less than an average RPKM of one across all samples have been removed. We then calculate the ratio of the median of the RPKM of the two constitutive exons which are furthest away from each other. This quantifications are subsetted to the 25% longest genes only. Subsequently the 3'/5' ratio for each sample is dened as the median ratio across all genes in this set. In order to dene a robust threshold representing excess of 3'/5' bias we have applied Tukey's outlier rule (Tukey, 1977). By this rule we remove all samples for which the bias exceeds the upper percentile plus 1.5 times the interquartile range of the 3'/5' bias across all samples (**Fig. S1B-C**).

Section 11: Integrative Analysis and Exploration

Integrated Analysis and Interactive Exploration with Regulome Explorer

To gain greater insight into the development and progression of prostate adenocarcinoma, we have integrated all of the data types produced by TCGA and described in this paper into a single "feature matrix". From this single heterogeneous dataset, significant pairwise associations have been inferred using statistical analysis and can be visually explored in a genomic context using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). In addition to associations that are inferred directly from the TCGA data, additional sources of information and tools are integrated into the visualization for more extensive exploration (e.g., NCBI Gene, miRBase, the UCSC Genome Browser, etc).

Feature Matrix Construction

A feature matrix was constructed using all available clinical, sample, and molecular data for 333 unique patient/tumor samples. The clinical information includes features such as age and tumor size; while the sample information includes features derived from molecular data such as single- platform cluster assignments. The molecular data includes mRNA and microRNA expression levels (Illumina HiSeq data), protein levels (RPPA data), copy number alterations (derived from segmented Affymetrix SNP data as well as GISTIC regions of interest and arm-level values), DNA methylation levels (Illumina Infinium Methylation 450k array), and somatic mutations. For mRNA expression data, gene level RPKM values from RNA-seq were log2 transformed, and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). For miRNA expression data, the summed and normalized microRNA quantification files were log2 transformed, and filtered to remove low-variability microRNAs (bottom 25% removed, based on interdecile range). For methylation data, probes were filtered to remove the bottom 25% based on interdecile range. For somatic mutations, several binary mutation features indicating the presence or absence of a mutation in each sample were generated. Mutation types considered were synonymous, missense, nonsense and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains.

Pairwise Statistical Significance

Statistical association among the diverse data types in this study was evaluated by comparing pairs of features in the feature matrix. Hypothesis testing was performed by testing against null models for absence of association, yielding a *p*-value. *P*-values for the association between and among clinical and molecular data types were computed according to the nature of the data levels for each pair: categorical vs. categorical (Chi-square test or Fisher's exact test in the case of a 2x2 table); categorical vs. continuous (Kruskal-Wallis test) or continuous vs. continuous (probability of a given Spearman correlation value). Ranked data values were used in each case. To account for multiple-testing bias, the *p*-value was adjusted using the Bonferroni correction.

Exploring significant associations between features

Regulome Explorer allows the user to interactively explore significant associations between various types of features – associations between molecular features, associations between molecular features and derived numeric features (like AR score), and associations between molecular features and categorical features such as clinical features or clusters derived from prior analysis (like iCluster). The examples below are screenshots from Regulome Explorer which illustrate exploration of the TCGA prostate cancer data.

Supplemental References

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell 153, 666-77.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson. K.L., Fan, J.B., Shen R. (2011). High density DNA methylation array with single CpG site resolution. Genomics 98, 288-95.

Campan, M., Weisenberger, D.J., Trinh, B., Laird, P.W. (2009). MethyLight. Methods Mol Biol. 507,325-37.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature. 490,61-70.

Carter, SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 30,413-21.

Chen, K., Wallis, J.W, McLellan, M.D, Larson, D.E., Kalicki, J.M., Pohl, C.S, McGrath, S.D., Wendl, M.C., Zhang. Q., Locke. D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6, 677-81.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 31, 213-219.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27, 2601-2602.

Coombes, K., Neeley, E.S., and Joy, C. (2011). SuperCurve: SuperCurve Package. R package version 1.4.1.

Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research 41, e67.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Dvinge, H., Ries, R.E., Ilagan, J.O., Stirewalt, D.L., Meshinchi, S., Bradley R.K. (2014). Sample processing obscures cancer-specific alterations in leukemic transcriptomes. Proc Natl Acad Sci U S A 111, 16802-7.

Gaujoux, R., Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 11,367.

Grasso, C.S., Wu ,Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature 487, 239-43.

Habegger, L., Sboner, A., Gianoulis, T.A., Rozowsky, J., Agarwal, A., Snyder, M., and Gerstein, M. (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics 27, 281–283.

Hennessy, B.T., Lu, Y., Poradosu, E., Yu, Q., Yu, S., Hall, H., Carey, M.S., Ravoori, M., Gonzalez-Angulo, A.M., Birch, R., et al. (2007). Pharmacodynamic markers of perifosine efficacy. Clinical cancer research : an official journal of the American Association for Cancer Research 13, 7421-7431.

Hieronymus, H., Lamb, J., Ross, K.N., Peng, X.P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S.M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer Cell 10, 321-30.

Hsu, S.D., Tseng, Y.T., Shrestha, S., Lin, Y.L., Khaleel, A., Chou, C.H., Chu, C.F., Huang, H.Y., Lin, C.M., Ho, S.Y., et al. (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Res 42, D78-85.

Hu, J., He, X., Baggerly, K.A., Coombes, K.R., Hennessy, B.T., and Mills, G.B. (2007). Non-parametric quantification of protein lysate arrays. Bioinformatics 23, 1986-1994.

Hu R, Dunn TA, Wei S, Isharwal S, Veltri RW, Humphreys E, Han M, Partin AW, Vessella RL, Isaacs WB, Bova GS, Luo J. Ligand-independent androgen receptor variants derived from splicing of cryptic exons signify hormone-refractory prostate cancer. Cancer Res. 2009 Jan 1;69(1):16-22.

Knijnenburg, T.A., Wessels, L.F., Reinders, M.J., Shmulevich, I. (2009). Fewer permutations, more accurate P-values. Bioinformatics. 25, i161-8.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A

101, 811-6.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. et al., (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Li, H., Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-95.

Li, J., Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res 22,519-36.

Liang, J., Shao, S.H., Xu, Z.X., Hennessy, B., Ding, Z., Larrea, M., Kondo, S., Dumont, D.J., Gutterman, J.U., Walker, C.L., et al. (2007). The energy sensing LKB1-AMPK pathway regulates p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. Nature cell biology 9, 218-224.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A. et al. (2008). Integrated detection and population genetic analysis of SNPs and copy number variation. Nat Genet. 40,1166-1174.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy number alteration in human cancers. Genome Biol. 12, R41.

Mullokandov, G., Baccarini, A., Ruzo, A., Jayaprakash, A.D., Tung, N., Israelow, B., Evans, M.J., Sachidanandam, R., Brown, B.D. (2012). High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nat Methods. 9,840-6.

Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M. (2004). Circular binary segmentation for the analysis of array based DNA copy number data. Biostatistics 5, 557-572.

Prandi, D., Baca, SC., Romanel, A., Barbieri, CE., Mosquera, J.M., Fontugne, J., Beltran, H., Sboner, A., Garraway, L.A., Rubin, M.A., Demichelis, F. (2014). Unraveling the clonal hierarchy of somatic genomic aberrations. Genome Biol. 15(8):439.

Ramos, A. H., Lichtenstein, L., Gupta M., Lawrence M.S., Pugh T.J., Saksena G., Meyerson M. and Getz G. (2015). Oncotator: Cancer Variant Annotation Tool. Human Mutation *36*, E2423-9.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D.Z., Rozowsky, J.S., Tewari,

A.K., Kitabayashi, N., Moss, B.J., Chee, M.S., et al. (2010). FusionSeq: a modular framework for finding gene fusions by analyzing Paired-End RNA-Sequencing data. Genome Biol. 11, R104.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 28,1353-8.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw ,A.A., D'Amico, A.V., Richie, J.P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203-9.

Subramanian, A., Tamayo, P., Mootha ,V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, JP. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 102, 15545-50.

Tay, Y., Rinn, J., Pandolfi, P.P. (2014). The multilayered complexity of ceRNA crosstalk and competition. Nature 505,344-52.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature 474, 609-615.

The Cancer Genome Atlas Network, (2014). Integrated genomic characterization of papillary thyroid carcinoma 159, 676–690.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 14, 178-192.

Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G.B., Kornblau, S.M. (2006). Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther. 5, 2512-21.

Tomlins ,S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M,. Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., Shah, R.B., Chinnaiyan, A.M. (2007). Integrative molecular concept modeling of prostate cancer progression. Nat Genet. 239, 41-51.

Triche, T.J. Jr., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., Siegmund, K.D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 41, e90.

Tukey, J.W. (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

Wallace, T.A., Prueitt, R.L., Yi, M., Howe, T.M., Gillespie, J.W., Yfantis, H.G., Stephens ,R.M., Caporaso, N.E., Loffredo, C.A., Ambs, S. (2008). Tumor

immunobiological differences in prostate cancer between African-American and European-American men. Cancer Res. 1, 927-36.

Wilkerson, M.D., Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics, 2010 Jun 15;26(12):1572-3.

Wilks, C., Cline, M.S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., et al. (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford). pii: bau093.

Yang, L., Luquette, L.J., Gehlenborg ,N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L., Kucherlapati, R, et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153, 919-29.

Yu, Y.P., Landsittel, D., Jing, L., Nelson, J., Ren, B., Liu, L., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., Michalopoulos, G., Becich, M., Luo, JH. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. J Clin Oncol. 15, 2790-9.