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.