Supplementary Information for Berger *et al.*, The genomic complexity of primary human prostate cancer

I. List of Supplementary Figures and Tables

II. Supplementary Methods

A. Sample Attributes, DNA Preparation, and Quality Control

Description of the clinical cohort

The prostate cancer samples used for this study came from a cohort of men undergoing surgery by one surgeon (A.T.) for clinically localized prostate cancer at the Institute of Prostate Cancer and Lefrak Center of Robotic Surgery, Weill Cornell Medical College and New York Presbyterian Hospitals (New York, NY).

Prostate cancer selection and DNA extraction

All of the prostate cancer samples were collected under an IRB approved protocol. Hematoxylin and eosin (H&E) slides were prepared from frozen tissue blocks and evaluated for cancer extent and tumor grade by the study pathologist (M.A.R./R.E.). To ensure for high purity of cancer cells and minimize benign tissue, tumor isolation was performed by first selecting for high-density cancer foci (<10% stromal or other non-tumor tissue contamination) and then taking 1.5 mm biopsy cores from the frozen tissue block for DNA extraction. DNA was extracted using phenol-chloroform and purified by ethanol precipitation method. Frozen tissue cores were homogenized and incubated in lysis solution made up of TE, NaCl, SDS, Proteinase K and nuclease-free water for 16 hours at 55°C. Next, phenol/chloroform/isoamyl alcohol (24:25:1, pH 8) mixture was added and DNA isolated from the aqueous phase. Into the removed supernatant, chloroform/isoamyl alcohol (49:1) mixture was added, centrifuged, and the aqueous phase re-extracted. For purification, DNA was precipitated in isopropanol solution containing glycogen and 2M sodium perchlorate, and the pellet washed twice with 70% ethanol at 4°C. The purified DNA was suspended in nuclease-free water. DNA from whole blood was extracted using Gentra® Puregene® Kit (Qiagen, Valencia, CA). DNA from tissue and blood were treated with RNase A (Qiagen) according to the manufacturer's instructions and then run on a 2% agarose gel to assess for structural integrity.

Determining ETS rearrangement status by interphase FISH and RT-PCR

The ETS rearrangement status was assessed on tissue slides or RNA taken from the same tumor nodule used for DNA sequencing. The methods for fluorescence in situ hybridization (FISH) and RT-PCR for *TMPRSS2-ETS* gene fusion have been previously described[1]. We used an *ERG* break-apart FISH assay followed by *TMPRSS2* break-apart assay to confirm the genes' rearrangement on the DNA level. Confirmation of the *TMPRSS2-ERG* fusion on the transcript level was performed by RT-PCR. In brief, RNA was reverse transcribed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster City, CA). The *TMPRSS2-ERG* PCR was performed using Platinum *Taq* DNA Polymerase (Invitrogen, Carlsbad, CA) with 1.5 mM $MgCl_2$, 0.1 μM of each primer (forward: *TMPRSS2* exon 1 –TAGGCGCGAGCTAAGCAGGAG – and reverse: *ERG* exon 5 – GTAGGCACACTCAAACAACGACTGG (ref [1]); and 50 ng cDNA at an annealing temperature (Ta) of 63°C for 35 cycles and the PCR products were separated on a 2.5% agarose gel. For the detection of *TMPRSS2-ERG* isoform IV, the PCR was performed, using a reverse primer in ERG exon 7 (CCATATTCTTTCACCGCCCACTCC), under the same conditions but with 1 mM $MgCl_2$ and 40 cycles. The obtained products were isolated from the gel using the MinElute™ Gel Extraction Kit (Qiagen, Valencia, CA) and subsequently sent for Sanger sequencing at the Life Sciences Core Laboratories Center of Cornell University (Ithaca, NY).

Quality assessment of DNA and tumor purity

Concentrations of tumor and normal DNA were measured using PicoGreen® dsDNA Quantitation Reagent (Invitrogen, Carlsbad, CA). We required a minimum DNA concentration of 60 ng/μl for sequencing. In one case (PR-1701), the initial concentration was <60 ng/μl and was increased following ethanol precipitation and re-suspension. To confirm that the DNA samples were of sufficiently high quality and not degraded, we performed gel electrophoresis and observed that in each case the large majority of DNA was high molecular weight. We prepared reserve stocks of each sample using whole genome amplification (WGA) for use in subsequent validation efforts, though the Illumina sequencing libraries were created with the native DNA. The identities of all tumor and normal DNA samples (native and WGA product) were confirmed by mass spectrometric fingerprint genotyping of 24 common SNPs (Sequenom, San Diego, CA). Finally, tumor DNA was hybridized to genome-wide human SNP microarrays (Affymetrix SNP Array 6.0) and analyzed as described previously[2]. We used a novel algorithm, termed ABSOLUTE, to infer the tumor purity and average ploidy from the allele-specific copy number levels (Carter S.L. *et al.*, manuscript in preparation). We then calculated the "allelic fraction" for each tumor, indicative of the fraction of sequence reads expected to harbor the non-reference allele at a locus with a somatic mutation existing at a single copy per nucleus. We selected samples for sequencing with an allelic fraction > 0.25.

B. Sequence Data Generation and Processing

We sequenced the complete genomes of tumor and normal samples according to the manufacturer's protocols (Illumina, San Diego, CA) and as described elsewhere (Chapman *et al.*, in press). A brief summary is provided below.

Whole genome shotgun (WGS) library construction

3 μg of native DNA from each tumor and normal sample was sheared to a range of 100-700 basepairs using the Covaris E210 instrument (Covaris, Woburn, MA). DNA fragments were end-repaired, phosphorylated, and modified by adenylation of 3' ends. Following the ligation of standard paired end adaptors, fragments were purified by gel electrophoresis (4% agarose, 85 volts, 3 hours) and gel excision of two bands (500–520bp and 520–540bp). This resulted in two libraries for each sample, with inserts

averaging 380bp and 400bp, respectively. Qiagen min-elute columns were used for DNA purification after each step. Final purified fragments were enriched by PCR amplification (10 cycles).

Illumina sequencing

The quantity of fragments with properly ligated adapter was measured for each library by qPCR. Each library was then normalized to 2 nM and denatured using 0.1 N NaOH. Cluster amplification was performed according to the manufacturer's protocols using v2 Chemistry and v2 Flowcells. Cluster densities were measured using SYBR Green dye. We performed paired-end sequencing (2 × 101bp) on the Illumina Genome Analyzer II platform, using v3 Sequencing-by-Synthesis kits and the Illumina v1.3.4 analysis pipeline. Both libraries from each individual sample were sequenced to approximately equal depth. Each library was sequenced on an average of 15 flow cell lanes, resulting in a haploid genomic coverage of approximately 30x for each tumor and normal. Standard quality control metrics for each lane are listed in **Supplementary Table S9**.

Data processing pipeline (Picard)

The data-processing pipeline "Picard" was developed by the Sequencing Platform at the Broad Institute for the pre-processing, alignment, and post-filtering of massively parallel sequencing data (Fennell T. *et al.*, unpublished). The output of Picard is a single BAM file[3] (http://samtools.sourceforge.net/SAM1.pdf) storing all reads with well-calibrated quality scores together with their alignments to the reference genome. The Picard pipeline consists of four steps (outlined below): (1) alignment to the genome; (2) recalibration of base qualities; (3) aggregation of lane-level data; and (4) flagging artifactual duplicate read pairs. Many individual tools in the Picard pipeline are available for download at http://picard.sourceforge.net/.

Each base is initially assigned a Phred-like quality Q score[4] by the Illumina pipeline, representing the probability that the base call is erroneous. In the first step of Picard, read pairs are aligned to human genome (NCBI build 36.3) using MAQ and sorted according to their chromosomal position[5]. (The fractions of mapping reads in each lane are listed in **Supplementary Table S9**.) In the second step, these original Q scores are empirically recalibrated based on the read-cycle, the lane, the flow cell tile, the base in question, and the preceding base. (The original quality scores are kept in the BAM file in the OQ tag for each read.) In the third step, lane-level BAM files are aggregated to library-level BAM files, which are then combined to sample-level BAM files. This information is captured in the read group tag and the BAM header. In the final step, molecular duplicate reads are flagged to indicate artifacts from PCR amplification of library fragments.

Several of the tools in the Picard pipeline, as well as in the Firehose pipeline (discussed below), were developed in collaboration with the Broad Institute's Medical and Population Genetics Program as part of the Genome Analysis Toolkit (GATK). An introduction to the GATK may be found at http://www.broadinstitute.org/gatk.

C. Identification of Somatic Mutations

In order to characterize the full spectrum of somatic mutations from the BAM files produced for each Tumor/Normal sample pair, the Cancer Genome Analysis group at the Broad Institute developed a suite of tools that together comprise the "Firehose" pipeline. Firehose manages the input files, analysis tools, and output files as well as the analysis workflow: *i.e.*, where the data reside, what needs to be executed on each file and in what order, and what is currently running (Voet D. *et al*., unpublished). Firehose uses GenePattern[6] as its execution engine, which runs the pipelines and modules based on specified parameters and ensures that the analysis results are reproducible.

The tools contained within Firehose execute the following analyses, as also described elsewhere (Chapman M. *et al*., in press; Berger M.F. *et al*., submitted, Bass A. *et al*., submitted):

Quality control

The first step in Firehose is to ensure that all sequence data match their corresponding patient and that there are no swaps between the tumor and normal samples for the same individual. To test whether sequence data match their corresponding patient, base calls are compared to genotypes determined from Affymetrix SNP 6.0 microarrays[7]. Homozygous non-reference genotypes are compared to the observed bases at the corresponding genomic positions for each separate Illumina lane. Lanes with <95% concordance are excluded from the analysis.

To test whether there are swaps between tumor and normal samples, we use two pieces of information: (1) insert size distribution and (2) copy number profile. Each sequencing library has a characteristic insert size distribution whose mean and standard deviation can be precisely defined from the tens of millions of read pairs in a given Illumina lane. Lanes with an insert size distribution that do not match the distribution of the other lanes for the same library are excluded. We also determine the copy number profile of each lane (using the depth of coverage in windows of 100kb along the genome) and compare to the profile determined by SNP microarrays. Tumor lanes that do not match the expected profile, or normal lanes that deviate from the expected flat profile, are excluded.

Local Realignment

The presence of insertions or deletions (indels) with respect to the reference genome can lead to multiple unwanted scenarios. The indel may not be properly recognized by the sequence aligner, leading to the accumulation of erroneous mutations in the flanking sequence. Alternatively, the indel may be placed in the wrong position within the read and/or in inconsistent positions within the collection of reads that map to the locus. In order to take into account all the evidence for an indel available from multiple reads mapping to the locus, we perform a multiple sequence alignment of reads in the vicinity of all putative indel sites. (Putative indel sites are denoted based on the presence of indels and/or consecutive mismatches within individual reads.) This is accomplished using the IndelRealigner module of the Genome Analysis Toolkit (http://www.broadinstitute.org/gatk).

Identification of base pair substitutions

Somatic base pair substitutions are identified using a highly sensitive and specific method developed by the Broad's Cancer Genome Analysis group, called *muTector*. The basic steps are outlined below, though more details will be presented elsewhere (Cibulskis K. *et al*., manuscript in preparation).

First, aligned reads in the tumor and normal BAM files are filtered out if they harbor too many mismatches or very low quality scores, as these introduce unnecessary noise. Second, candidate somatic mutations are identified according to the observed allele counts, base quality scores, and sequence coverage at a given genomic position in the tumor and normal BAM files. Third, candidate mutations are subject to a series of empirical filters designed to eliminate false positives calls. Finally, mutations are annotated according to their genomic region (*e.g.*, exon, intron, promoter, intergenic region), amino acid change, protein domain, etc.

The calling algorithm utilizes a Bayesian statistical framework to compare the probabilities of generating the observed sequence data given underlying reference or non-reference genotypes. For each sample pair, we calculate two LOD scores (log odds) to express our confidence that the tumor is non-reference and that the normal is reference at a given position. The tumor LOD score and normal LOD score are compared to separate cutoffs reflecting the prior probabilities of mistakenly calling a non-reference base in the tumor that is really reference and of mistakenly calling a reference base in the normal that is really a germline single nucleotide variant (SNV).

Once the candidate mutation calls are made, filters are applied to account for commonly observed error modes. For instance, local sequence context can occasionally lead to incorrect Illumina base calls, but often only in reads sequenced in a single direction. Therefore, we require that the observed orientations of reads carrying the variant allele not significantly differ from the observed orientations of all reads mapping to the locus. Occasionally there is not enough evidence to apply this strand filter, either because there are not enough total reads mapping to the locus or because there are not enough variant alleles to achieve a statistically significant result. We categorize each mutation according to the power of the filters given the observed data (*i.e.*, could the filters achieve a statistically significant result given the observed sequence coverage and variant allele count?). Based on independent validation of 562 predicted mutations (discussed below), we consider those calls where the filters are sufficiently powered as "high confidence" and those calls where the filters are underpowered as "moderate confidence".

Of all candidate somatic mutations identified, 46% were categorized as "high confidence" and 54% as "moderate confidence" mutations (Supplementary Table 2). The validation rates of high confidence and moderate confidence mutations in coding regions were 96% and 47%, respectively, (88% and 37% in non-coding regions) based on validation of 562 mutations by mass spectrometric genotyping. All predicted somatic base pair mutations (and accompanying validation data) are listed in **Supplementary Tables S2** and **S3**.

Identification of short insertions and deletions

Putative indel events are called from locally realigned data (see above) based on the fraction of supporting reads at a given locus in the tumor BAM file. These high sensitivity calls are then subject to a series of filters including the average number of mismatches and the distribution of base qualities in the reads containing indels. Events are categorized as germline or somatic according to whether there is evidence for the same event at the same locus in the normal BAM file. Independent validation experiments (Sequenom) have shown a high false positive rate (~60%), consistent with other groups, but that manual inspection of putative indels using the Integrative Genomics Viewer (Robinson J.T. *et al.*, submitted; http://www.broadinstitute.org/igv) enables the identification of the vast majority of false positive calls. Therefore, all indels predicted within protein coding exons are subject to manual review. Further details will be presented elsewhere (Sivachenko A. *et al.*, manuscript in preparation).

All predicted indels in coding regions are listed in **Supplementary Table S4**.

Identification of chromosomal rearrangements

Rearrangements were identified from discordant paired sequence reads mapping to different chromosomes (translocations), different positions on the same chromosome (large deletions, inversions, and duplications), or in unexpected orientations (small inversions and tandem duplications).

Chromosomal rearrangements are identified by a novel method developed by the Broad's Cancer Genome Analysis group, called *dRanger* (Lawrence M.S. *et al.*, manuscript in preparation). First, discordant read pairs are identified in the tumor. These are read pairs that map to different chromosomes or in unexpected positions (>600bp apart) or unexpected orientations (incorrect order on opposite strands or any order on the same strand) on the same chromosome. Second, clusters of discordant pairs are used to nominate potential rearrangement events. Candidate rearrangements are removed if there are any supporting discordant pairs for the same event in its corresponding matched normal or in a panel of additional normal genomes sequenced at the Broad Institute. Third, a series of additional filtering metrics is computed for each candidate rearrangement: (1) the fraction of nearby reads with a mapping quality of zero; (2) the number and diversity of other discordant pairs in the vicinity of the breakpoints; and (3) the standard deviation of the starting positions of the supporting read pairs. These filtering metrics are combined into an overall quality measure (0 to 1), which serves as a multiplicative scaling factor to convert the number of supporting read pairs to a score for the rearrangement. Based on independent validation experiments (discussed below), we consider rearrangements with a score of 3.0 or higher.

Approximate locations of rearrangement breakpoints are assigned based on the boundary of all reads in supporting read pairs. Breakpoints are then annotated as intronic, exonic, or intergenic according to the RefSeq database. Rearrangements with both breakpoints located in genes are further annotated as to whether they are consistent with a gene fusion, and whether it would be in-frame or out-of-frame.

When possible, breakpoints are mapped to basepair resolution using *BreakPointer* (Drier Y. *et al*., manuscript in preparation). *BreakPointer* searches for read pairs where one read mapped on either side of the breakpoint and the pair mate is partly mapped on the breakpoint, or failed to align anywhere. It is expected that many of these reads span the actual fusion point. These unmapped reads are subjected to a modified Smith-Waterman alignment procedure with the ability to jump between the two reference sequences at the most fitting point. Further details will be presented elsewhere (Drier Y. *et al*., manuscript in preparation). Using *BreakPointer*, we were able to map the breakpoints to base pair resolution in 88% of cases (663/755).

Rearrangements are illustrated using the "CIRCOS" program (http://mkweb.bcgsc.ca/circos) and are shown in **Figure 1**. All predicted rearrangements and breakpionts (and accompanying validation results) are listed in **Supplementary Table S5**. Across the seven prostate tumors, 56% of rearrangements involved at least one intragenic breakpoint.

D. Experimental Validation of Somatic Mutations

Mass spectrometric genotyping of point mutations

In order to estimate the specificity of our method for calling somatic mutations, we obtained independent validation data for 562 candidate mutations using mass spectrometric genotyping (Sequenom) of the whole genome amplified tumor and normal DNA. This collection included 283 "high confidence" and 279 "moderate confidence" predictions. We tested 157 candidate protein-coding mutations, including the vast majority of all non-silent mutations in all 7 samples, and 405 non-coding mutations (202 intronic and 203 intergenic). The genotyping data confirmed that 96% and 88% of high confidence calls, and 47% and 37% of moderate confidence calls, were *bona fide* somatic mutations in coding and non-coding regions, respectively. Sequenom failures may account for some of the false positive calls. One mechanism by which this may arise is through loss of mutant alleles during whole genome amplification. Additionally, we have observed that Sequenom exhibits a high failure rate for mutations with an allelic fraction <20% (data not shown).

Using the Clopper-Pearson method to calculate 95% confidence intervals, we infer that our accuracy rates for high confidence mutations are 96% (CI: 89–99%) in protein coding regions and 88% (CI: 83–92%) in non-coding regions, and our accuracy rates for moderate confidence mutations are 47% (CI: 36–58%) in protein coding regions and 37%

(CI: 31–45%) in non-coding regions. However, as suggested above, the true accuracy of our method may be greater, on account of false negative calls arising from mass spectrometric genotyping. Further, this value is, in theory, dependent upon the overall mutation burden per patient, as a tumor harboring more true somatic mutations would be expected to exhibit a higher validation rate. As a mitigating factor, we note that the estimated mutation rate per patient varies by only two-fold. We confirmed that our validation rate in the 3 samples with the highest mutation rates was indistinguishable from our validation rate in the 4 samples with the lowest mutation rates:

High confidence mutations (coding and non-coding)

All samples (283 mutations)	90% (CI: 86–93%)
Highest 3 mutation rates (144 mutations)	91% (CI: 84–95%)
Lowest 4 mutation rates (139 mutations)	90% (CI: 83–94%)

Moderate confidence mutations (coding and non-coding)

All samples (279 mutations)	40% (CI: 34–46%)
Highest 3 mutation rates (116 mutations)	42% (CI: 35–50%)
Lowest 4 mutation rates (163 mutations)	37% (CI: 28–48%)

Each fusion-positive sample harbored an overall excess of mutations at CpG dinucleotides that was out of proportion with the pattern observed in the four ETS-negative prostate tumors (p=0.0031; **Supplementary Fig. S1**).

PCR and massively parallel sequencing of structural rearrangements

Rearrangements predicted by *dRanger* were validated by PCR followed by pooled 454 sequencing. PCR primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3) such that they spanned the predicted chimeric junction and would produce an amplicon approximately 300–350bp long. PCRs were performed on whole genome amplified product for both tumor and normal DNA. (For somatic breakpoints, only the tumor DNA would be expected to yield a product.) Each PCR product was quantified using a NanoDrop Spectrophotometer (Thermo Scientific, Wilmington, DE). PCR products were pooled such that: (1) equal amounts of tumor products were combined, (2) the same volumes were taken from the corresponding normal products, and (3) matching tumor and normal products were placed in separate pools. Libraries for 454 sequencing were prepared from each pool and sequenced separate regions of a 454 Genome Sequencer FLX System (454 Life Sciences, Branford, CT). Primer sequences served as unique barcodes for identifying the source PCR product for each 454 read. A rearrangement was judged to be somatic if the predicted chimeric product was detectable in tumor DNA and not normal DNA. Out of 594 predicted rearrangements tested, 464 were confirmed as somatic, yielding a validation rate of 78%. However, based on discordant results for independent primer pairs designed to target the same loci, we estimate the overall sensitivity of the PCR validation assay to be 80–90%. (In 24/209 cases where we designed two different primer pairs, only one successfully amplified the chimeric product.) This suggests that the true specificity of *dRanger* is likely >90%.

Fluorescence in situ hybridization (FISH) for MAGI2, PTEN, CADM2, and CSMD3 rearrangements

To assess the status of *PTEN*, we used a locus specific probe and a reference probe (see details below). To assess for inversion of the *MAGI2* gene, a unique FISH assay was designed. Probes spanning the ends of the gene were labeled red (3' end) or green (5' end). A third probe, also labeled green, acted as a reference for the arrangement of the gene. A chromosome with no gene inversion showed a red signal (3' end of *MAGI2*) followed by two green signals (5' end of *MAGI2* then the reference probe at 7q36). A chromosome with the gene inversion showed the red signal between the two green ones, indicating that the 3' end and the 5' end have been inverted. To identify rearrangements disrupting *CADM2* and *CSMD3*, we utilized break apart FISH assays with probes positioned on both sides of the gene.

PTEN (10q23)	BAC#
PTEN gene (red)	CTD-2047N14
reference (green)	RP11-431P18

MAGI2 (7q11)	BAC#
5' end (green)	CTD-2014F20
3' end (red)	CTD-2517A17
reference (green)	RP11-28C14 (7q36.1)

CADM2 (3p12)	BAC#
5' end	RP11-164K14
3' end	RP11-781J22

CSMD3 (8q23)	BAC#
5' end	RP11-644M14
3' end	RP11-88L22

To determine the prevalence of each class of rearrangement, we surveyed an independent cohort of 90 patients (mean age 63.2 years) who underwent radical prostatectomy at Weill Cornell Medical College (New York, NY) as a monotherapy. The pathological stages ranged from organ confined to cases with extra-prostatic tumor extension.

E. Calculation of Somatic Mutation Rates

For the purposes of calculating the genome-wide mutation rate, we defined a base as "covered" if there were at least 14 and 8 reads that overlapped the position in the tumor and normal, respectively. We considered only mutations called at covered positions; the total number of covered positions ranged from 2.51–2.67 Gigabases per sample. To account for the variable specificity of high confidence and moderate confidence calls, we prorated each class according to their empirical validation rates in *non-coding* regions

(88% for high confidence, 37% for moderate confidence). As a result, we estimated the average genome-wide mutation rate to be 0.9 per megabase.

There are at least two reasons why this might be an underestimate of the true mutation rate. First, the empirical validation rates, which we used as scaling factors, may themselves be underestimates due to the imperfect sensitivity of our mass spectrometric (Sequenom) validation assay. Second, this calculation does not take into account the sensitivity of *muTector*. Modeling suggests that the sensitivity of *muTector* is approximately 90% (not shown), though it may be somewhat lower for samples with low purity and allelic fraction, discussed above.

F. Determination of Significantly Mutated Genes and Pathways

We identified significantly mutated genes based on the observed number of mutations for each gene in each mutation class (defined below), the sample-specific and class-specific background mutation rates, and the number of covered bases per gene. This analysis was performed using a novel algorithm, MutSig (Lawrence M.S. *et al.*, manuscript in preparation), based partly on methods published elsewhere[8,9]. Mutations were divided into classes according to sequence context: CpG, other C:G, and A:T. For each gene, we calculated the probability of obtaining the observed set of mutations (or a more extreme one) given the observed background mutation rates. P-values are converted to Q-values using the Benjamini-Hochberg procedure for controlling False Discovery Rate (FDR). We repeated this analysis at the pathway-level, considering a list of 616 gene sets corresponding to known pathways or gene families. For this analysis, we tabulated the number of mutations and the number of covered bases in all component genes of each gene set.

G. High-Density SNP Array Analysis of 66 Prostate Tumors

Genomic DNA from tumor and paired blood samples was processed using Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc.) according to manufacturer's protocols. The DNA was digested with NspI and StyI enzymes (New England Biolabs), ligated to the respective Affymetrix adapters using T4 DNA ligase (New England Biolabs), amplified (Clontech), purified using magnetic beads (Agencourt), labeled, fragmented, and hybridized to the arrays. Following hybridization, the arrays were washed and stained with streptavidin-phycoerythrin (Invitrogen Corporation). Following array scanning, data preprocessing was performed using Affymetrix Power Tools. Copy number data was evaluated after segmenting the log 2 ratios between tumor and paired normal levels on a sample basis. Quality control, data integrity, segmentation and copy number analysis were performed as previously described by Demichelis *et al.*[10]

H. Co-Occurrence of Breakpoint Locations, Mutations, and Published ChIP-Seq Binding Data

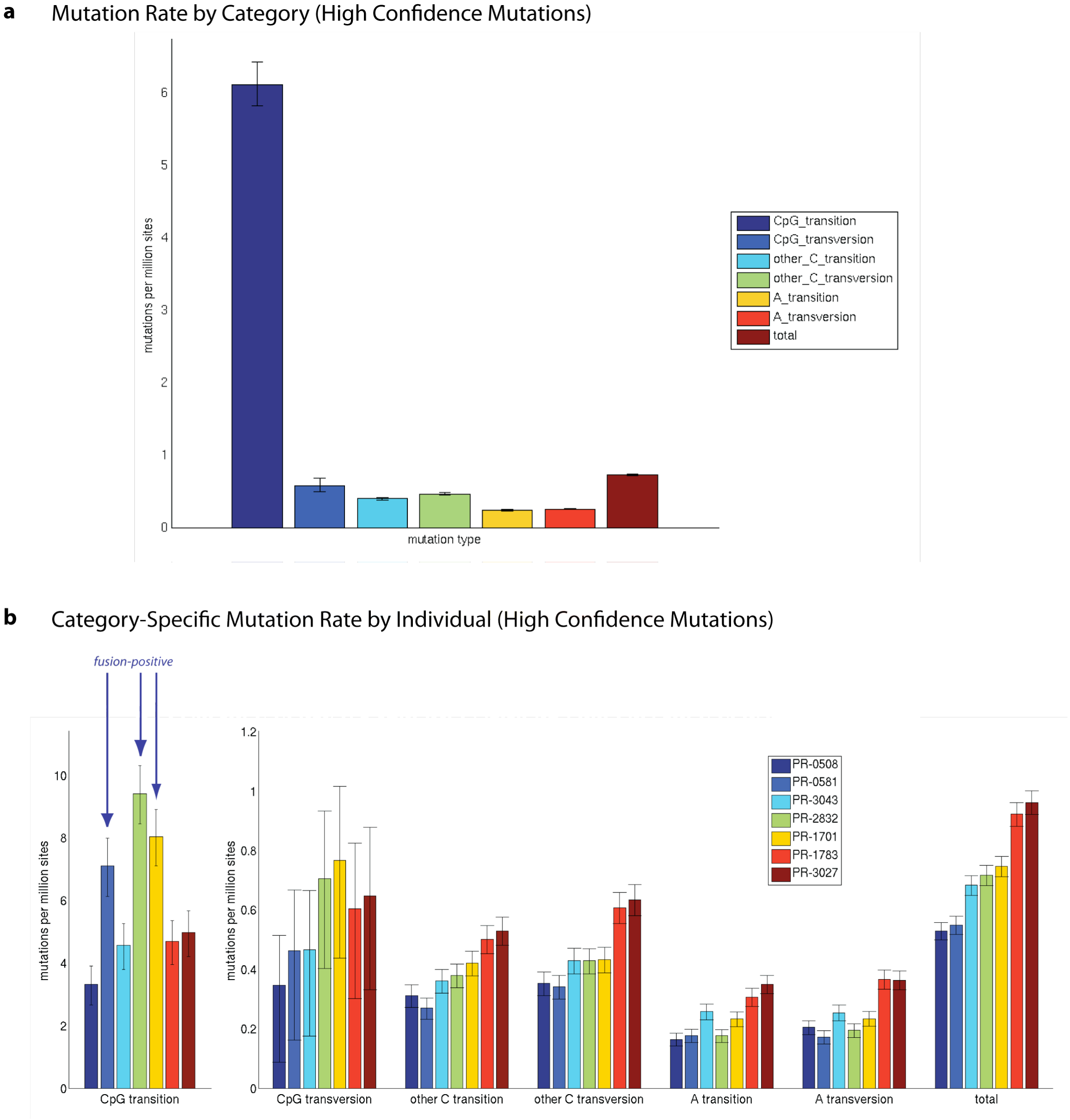
For each prostate cancer genome (as well as published melanoma, lung, and breast cancer genomes), we tested whether the associated rearrangement breakpoints occurred closer to or farther from a given set of ChIP binding sites than expected by chance. We downloaded pre-computed ChIP-Seq binding peaks for the following transcription factors and chromatin marks in the androgen-sensitive, *TMPRSS2-ERG* fusion positive prostate cancer cell line VCaP: AR (followed by treatment of R1881, a synthetic agonist of the androgen receptor), ERG, RNA polymerase II, acetylated histone H3, trimethylated histone H3K4, trimethylated histone H3K36, trimethylated histone H3K9, and trimethylated histone H3K27 (ref [11]). The number of peaks in each experiment ranged from 1,725 to 42,568. We also downloaded pre-computed genome-wide ChIP-Seq binding peaks for AR, H3K4me3, H3K36me3, H3K9me3, and acetylated histone H3 in the ETV1+ prostate cancer cell line LNCaP (ref [11]); for AR in the ETS- prostate cancer cell line PC3 (ref [12]); for H3K4me3, H3K36me3, and H3K27me3 in 3 cell lines from the ENCODE project[13] (GM12878, K-562, and H1ES); and ChIP-chip binding peaks for estrogen receptor (ER) in the breast cancer cell line MCF7 (ref [14]). In addition to the prostate cancer rearrangements presented here, we considered pre-computed rearrangements for published genomes in a melanoma cell line[15], a small cell lung cancer cell line[16], a primary lung cancer[17], and 24 breast cancer cell lines and primary tumors[18]. (We later discarded 6/24 breast cancers with fewer than 20 rearrangements.)

To test for enrichment or depletion of a prostate tumor's rearrangements near a given set of ChIP-Seq peaks, we calculated the rate of breakpoints within the aggregate of all sequence intervals +/- 50 kb surrounding each peak. This was compared to the background rate of breakpoints, which we estimated by taking the average of 1,000 simulations in which we controlled for *coverage* and *structure*. Simulated breakpoints were randomly generated at positions matched in sequence *coverage* to the observed breakpoints, to control for hidden correlations between breakpoints and ChIP-Seq peaks due to sequencing bias. (Background sampling considered the mean sequence coverage across all 7 prostate genomes in bins of size 5 (top bin ≥ 50-fold depth).) To control for *structure*, simulated breakpoint pairs corresponding to intrachromosomal rearrangements were preserved at fixed distances such that one end was perfectly matched in sequence coverage and the other end occurred at a site with no less than (but possibly greater than) the observed sequence coverage. (Controlling for structure was necessary to account for non-independent events from small intrachromosomal inversions and deletions.) Significance of enrichment or depletion of observed breakpoints compared to background was calculated according to the binomial distribution (**Supplementary Table 6**). In addition to a binomial p-value, we also computed the *ratio* of the observed rate to the background rate to determine the effect size independent of the total number of rearrangements detected in a given sample.

We repeated this calculation using different window sizes and found that the effects were consistent for intervals ranging from +/- 1 kb to +/- 1 Mb surrounding each ChIP-Seq peak (**Supplementary Fig S12a**). To be sure that significant associations were not the result of a small number of driver genes, we repeated calculations upon removing all rearrangements involving any of 16 genes we found to be recurrently disrupted in the 7 prostate tumors presented here (57 rearrangements total; **Supplementary Fig. S12b**).

To test for enrichment or depletion of point mutations near a given set of ChIP-Seq peaks, we repeated the calculations exactly as described above using a coverage-matched simulated background (**Supplementary Fig. S7a**). (We did not attempt to preserve the distance between mutations on the same chromosome.) To test for enrichment or depletion of point mutations near rearrangements in the corresponding prostate genome, we repeated the calculations using a coverage-matched simulated background in different window sizes surrounding each breakpoint (**Supplementary Fig. S7b**).

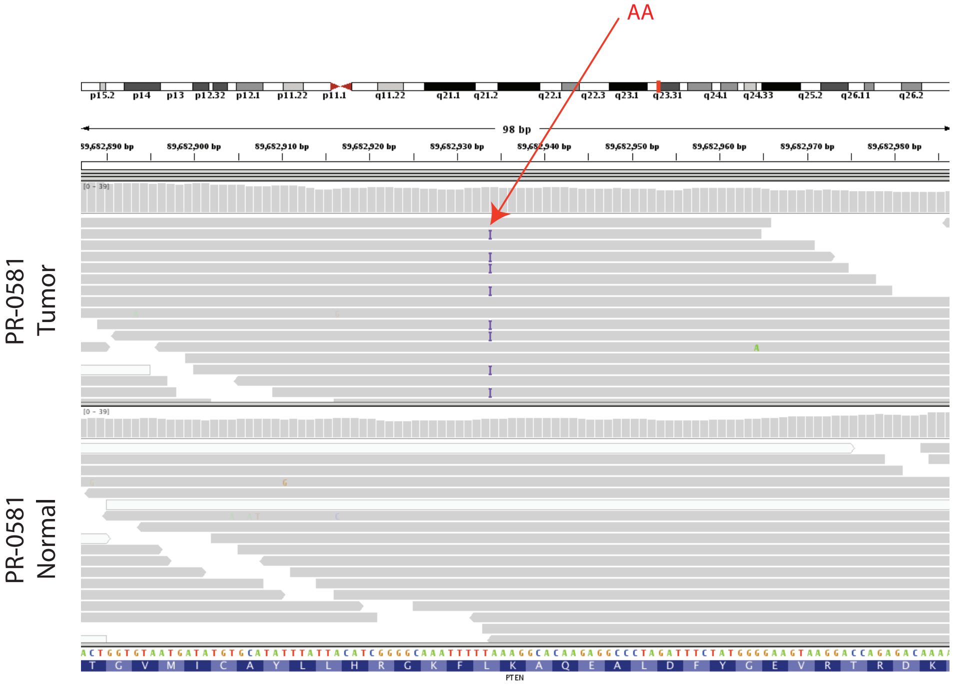
III. Supplementary Figures



a Mutation Rate by Category (High Confidence Mutations)

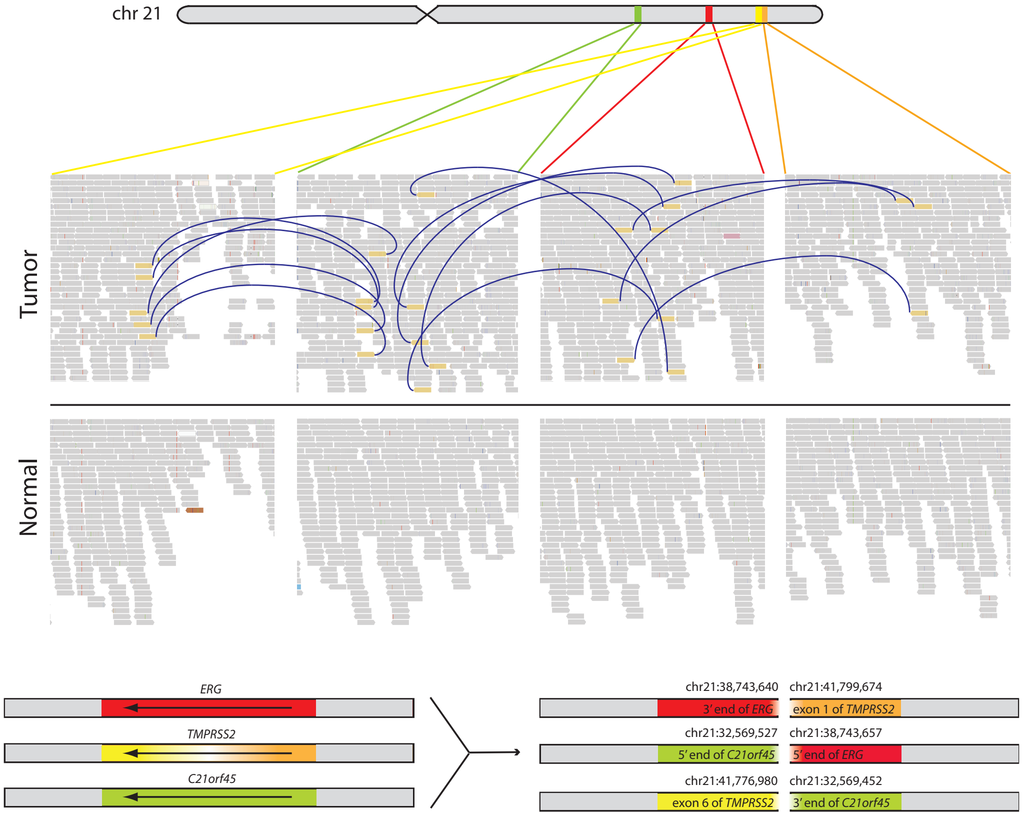
b Category-Specific Mutation Rate by Individual (High Confidence Mutations)

Supplementary Figure S1: Category-specific point mutation rates. Rates represent high confidence mutations at all "covered" positions (as defined in Methods); error bars represent 95% confidence intervals. (a) Average mutation rate for all 7 prostate tumors, broken down by category: CpG transition, CpG transversion, other C transition, other C transversion, A transition, A transversion. (b) Category-specific mutation rates, broken down by individual. The three *TMRPSS2-ERG* fusion-positive samples exhibit an excess of CpG transitions that is out of proportion with the other samples (p=0.0031; two-sided t-test).

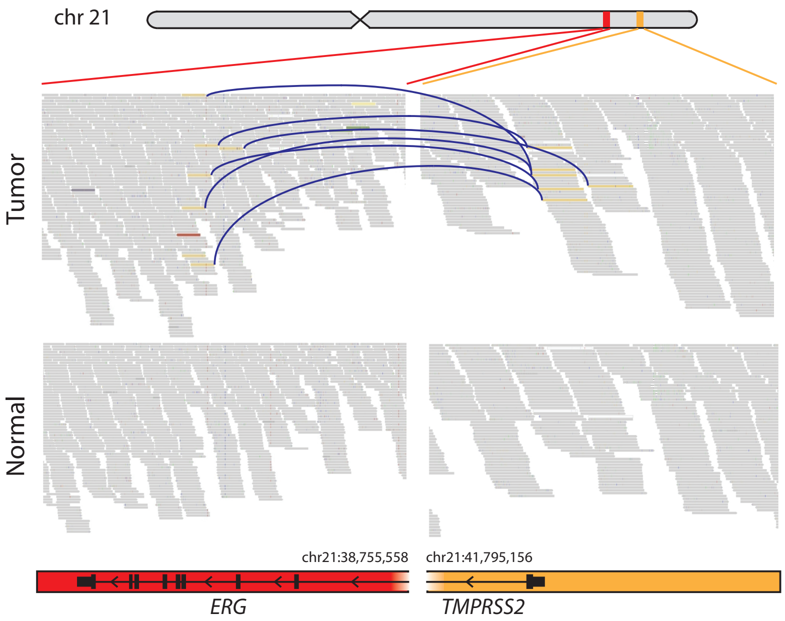

AA
p15.2 p14 p13 p12.32 p12.1 p11.22 p11.1 q11.22 q21.1 q21.2 q22.1 q22.3 q23.1 q23.31 q24.1 q24.33 q25.2 q26.11 q26.2
98 bp
89,682,890 bp 89,682,900 bp 89,682,910 bp 89,682,920 bp 89,682,930 bp 89,682,940 bp 89,682,950 bp 89,682,960 bp 89,682,970 bp 89,682,980 bp
[0 - 39]
PR-0581 Tumor
PR-0581 Normal
ACTGGTGTAATGATATGTGCATATTTATTACATCGGGGCAAATTTTTAAAGGCACAAGAGGCCCTAGATTTCTATGGGGAAGTAAGGACCAGAGACAAAA
T G V M I C A Y L L H R G K F L K A Q E A L D F Y G E V R T R D K
PTEN

PTEN exon 5:
c.437_438insAA
p.146fs

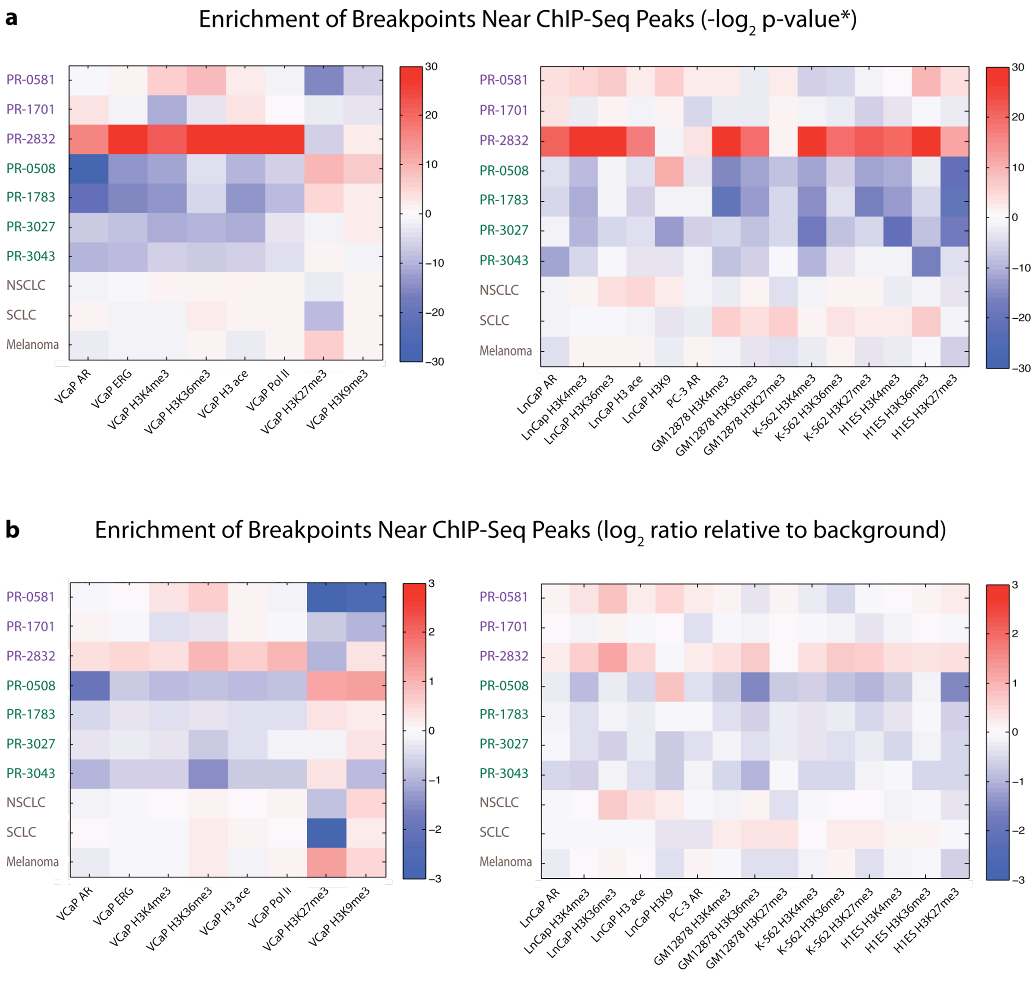
Supplementary Figure S2: Frameshift insertion in *PTEN*. Prostate PR-0581 harbors a 2bp frameshift insertion, which is evident in the sequence reads from the tumor genome but not the matched normal genome. Sequence data are visualized using the Integrative Genomics Viewer (IGV; Robinson J.T. *et al.*, submitted): http://www.broadinstitute.org/igv.



Supplementary Figure S3: Complex rearrangement producing *TMPRSS2-ERG* in prostate PR-0581. A trio of balanced (copy neutral) intrachromosomal rearrangements on chromosome 21 involving *ERG*, *C21orf45*, and two separate introns of *TMPRSS2* leads to the creation of *TMPRSS2-ERG*. Exon 1 of *TMPRSS2* is joined to the 3' end of *ERG*; the 5' end of *ERG* is joined to the 5' end of *C21ORF45*; and the 3' end of *C21ORF45* is joined to exon 6 of *TMPRSS2*. Discordant read pairs in the tumor genome but not the normal genome (yellow bars connected by blue lines) indicate somatic breakpoints.

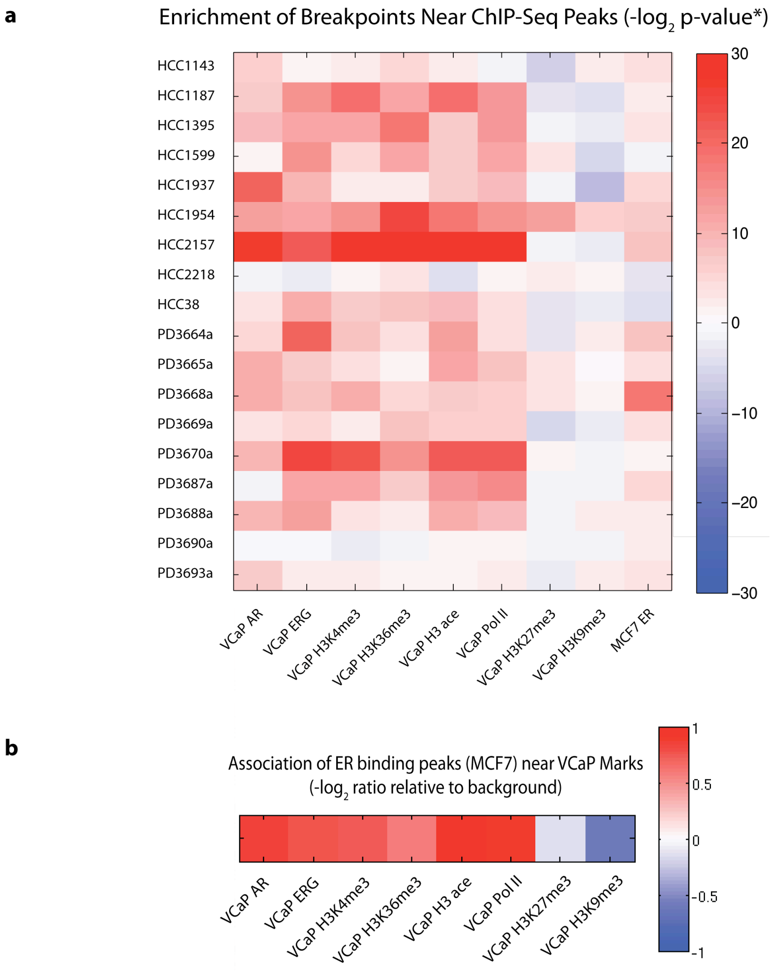


Supplementary Figure S4: Hemizygous deletion producing *TMPRSS2-ERG* in prostate PR-2832. A somatic 3-megabase deletion on chromosome 21 is implicated by the presence of discordant read pairs in the tumor genome but not the normal genome (yellow bars connected by blue lines), indicative of intragenic breakpoints joining the 5' end of *TMPRSS2* and the 3' end of *ERG*.



* For *depletion* of breakpoints (rather than enrichment), -log p-values are multiplied by -1.

Supplementary Figure S5: Association between rearrangement breakpoints and chromatin marks for prostate cancer, lung cancer, and melanoma. (a) For each genome, enrichment of breakpoints within 50 kb of each set of ChIP-Seq binding peaks was determined relative to a distance-matched and coverage-matched simulated background, as described in Supplementary Methods. P-values were calculated according to the binomial distribution. (For display purposes, -log p-values were multiplied by -1 in cases where breakpoints were depleted rather than enriched.) Marks of active transcription include AR, ERG, H3K4me3, H3K36me3, H3ace, and Pol II; marks of closed chromatin include H3K27me3 and H3K9me3. Breakpoints in *TMPRSS2-ERG* fusion-positive prostate PR-2832 are enriched near marks of active transcription in VCaP (left) and other cell lines (right), while breakpoints in ETS-negative prostates (green labels) are depleted. Breakpoints in 2 lung cancer genomes[16,17] and 1 melanoma genome[15] do not exhibit significant associations with VCaP chromatin marks. (b) Enrichment is displayed as the ratio of the observed breakpoint rate to the background rate near each indicated set of ChIP-Seq peaks. Unlike p-values (above), the effect size is independent of the total number of rearrangements per sample.

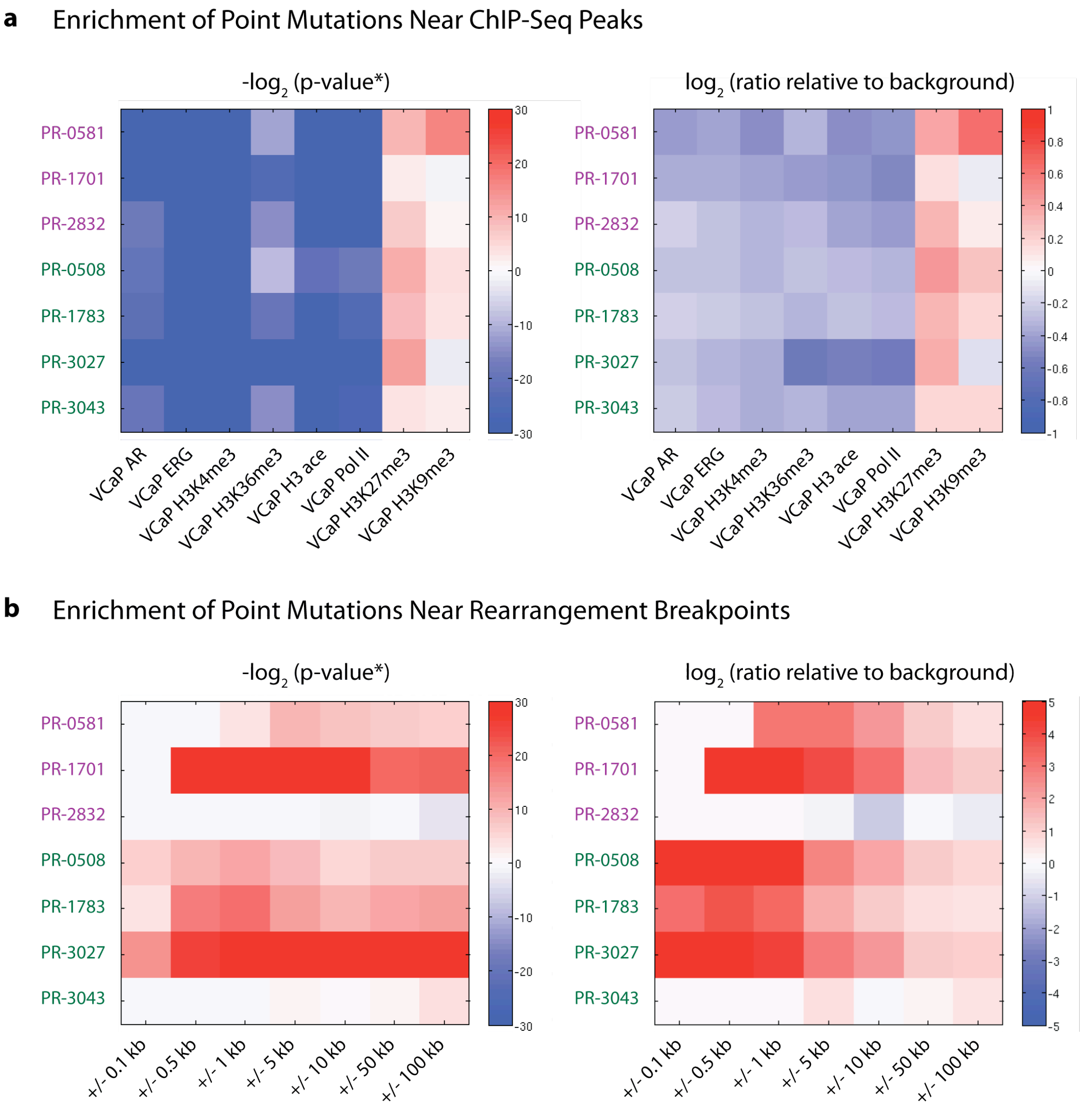


Enrichment of Breakpoints Near ChIP-Seq Peaks (-log$_2$ p-value*)

Association of ER binding peaks (MCF7) near VCaP Marks (-log$_2$ ratio relative to background)

* For *depletion* of breakpoints (rather than enrichment), -log p-values are multiplied by -1.

Supplementary Figure S6: Association between rearrangements in breast cancer and chromatin marks in prostate cancer. Breast cancer rearrangement breakpoints were identified previously from low coverage paired-end genome sequencing[18]. ChIP-Seq binding peaks were defined previously for the ERG+ prostate cancer cell line VCaP[11], and ChIP-chip binding peaks were defined previously for the breast cancer cell line MCF7 (ref [14]). (a) Enrichment of breakpoints near ChIP-Seq peaks, as shown in Supplementary Figure S5. Breast cancers show a positive association between rearrangement breakpoints and marks of active chromatin in prostate cancer, and also with estrogen receptor (ER) binding sites in breast cancer. For each chromatin mark, enrichment of breakpoints was determined for 9 cell lines (top) and 9 primary tumors (bottom) with at least 20 rearrangements. (b) High correlation between ChIP binding peaks in MCF7 for ER and binding peaks in VCaP for markers of open chromatin (AR, ERG, H3K4 me3, H3K36 me3, Pol II) but not closed chromatin (H3K9 me3, H3K27 me3). Enrichment of ER binding peaks is calculated as above and represented as a ratio of the observed peak rate to the background rate.

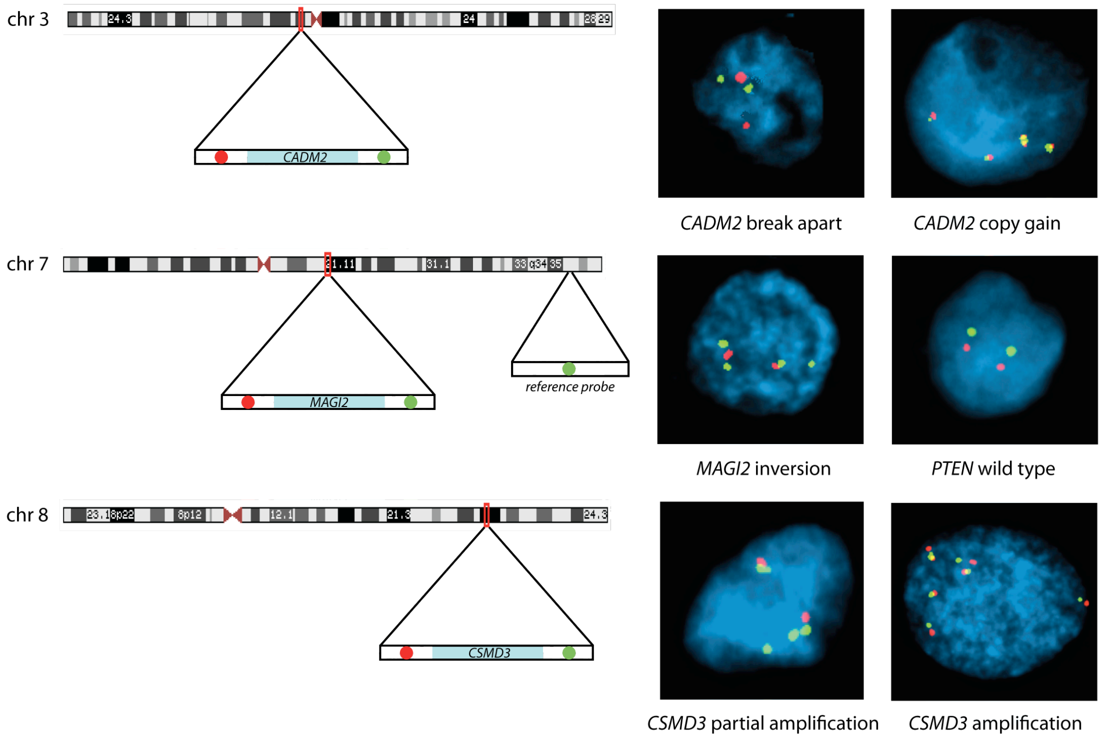


a Enrichment of Point Mutations Near ChIP-Seq Peaks

b Enrichment of Point Mutations Near Rearrangement Breakpoints

* For *depletion* of mutations (rather than enrichment), -log p-values are multiplied by -1.

Supplementary Figure S7: Association of point mutations with chromatin marks and rearrangement breakpoints. (a) Point mutations are depleted near VCaP ChIP-Seq binding peaks indicative of open chromatin (AR, ERG, H3K4me3, H3K36me3, H3 ace, Pol II) and enriched near VCaP ChIP-Seq binding peaks indicative of closed chromatin (H3K27me3, H3K9me3) in all 7 prostate tumors. P-values and enrichment ratios are calculated using a coverage-matched simulated background as above, for intervals of 50 kb surrounding each set of ChIP-Seq binding peaks. These results are consistent with both negative selection and transcription-coupled DNA repair. (b) Point mutations are enriched near rearrangement breakpoints at multiple distances in 5 prostate tumors, including 2/3 *TMPRSS2-ERG* fusion-positive tumors (purple labels) and 3/4 ETS-negative tumors (green labels).



CADM2 break apart

CADM2 copy gain

MAGI2 inversion

PTEN wild type

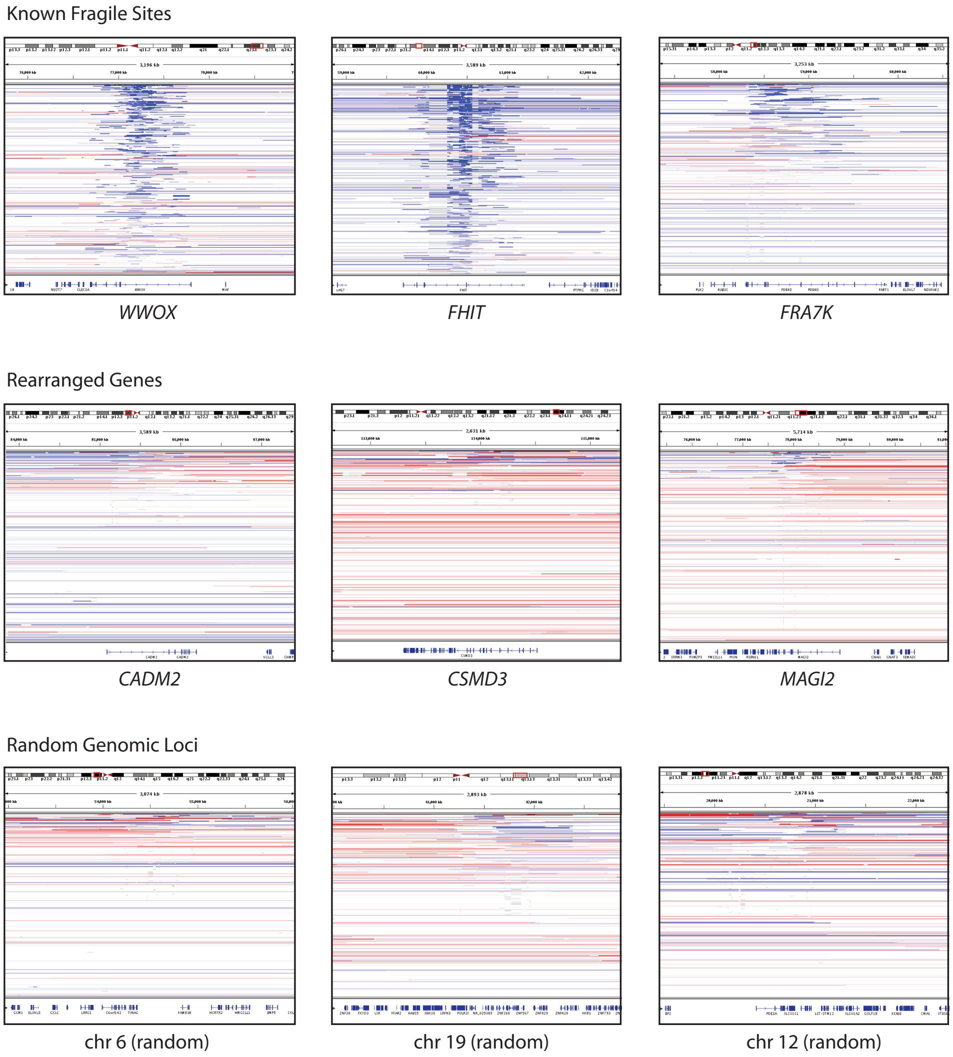
CSMD3 partial amplification

CSMD3 amplification

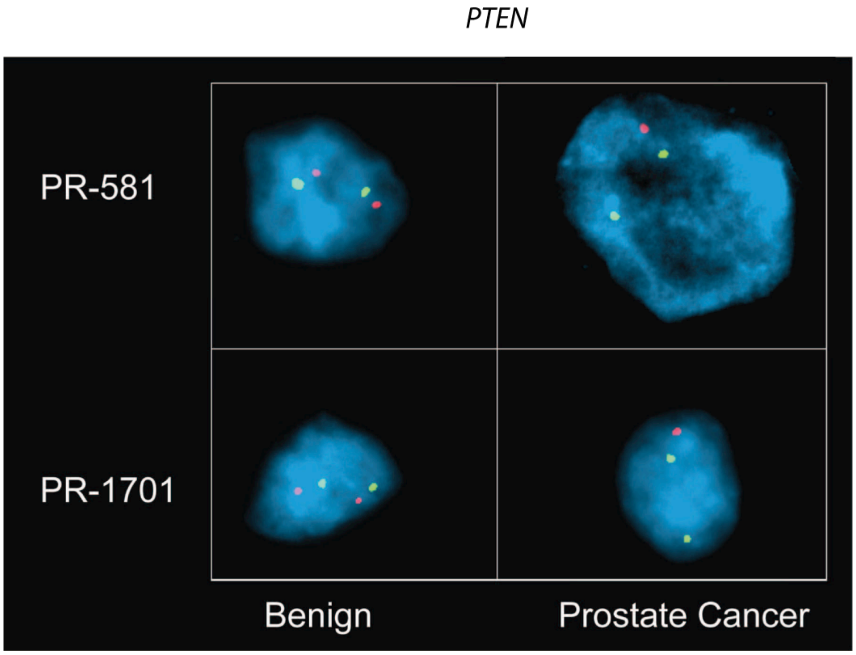
EVENT	CASES	SAMPLES
CADM2 break apart	5/90	STID 0410, 0415, 0432, 0560, 3080
CADM2 copy gain	1/90	STID 3085
MAGI2 inversion	3/88	STID 0427, 1023, 1045
CSMD3 amplification	20/94	
CSMD3 deletion	2/94	

ERG rearrangement positive

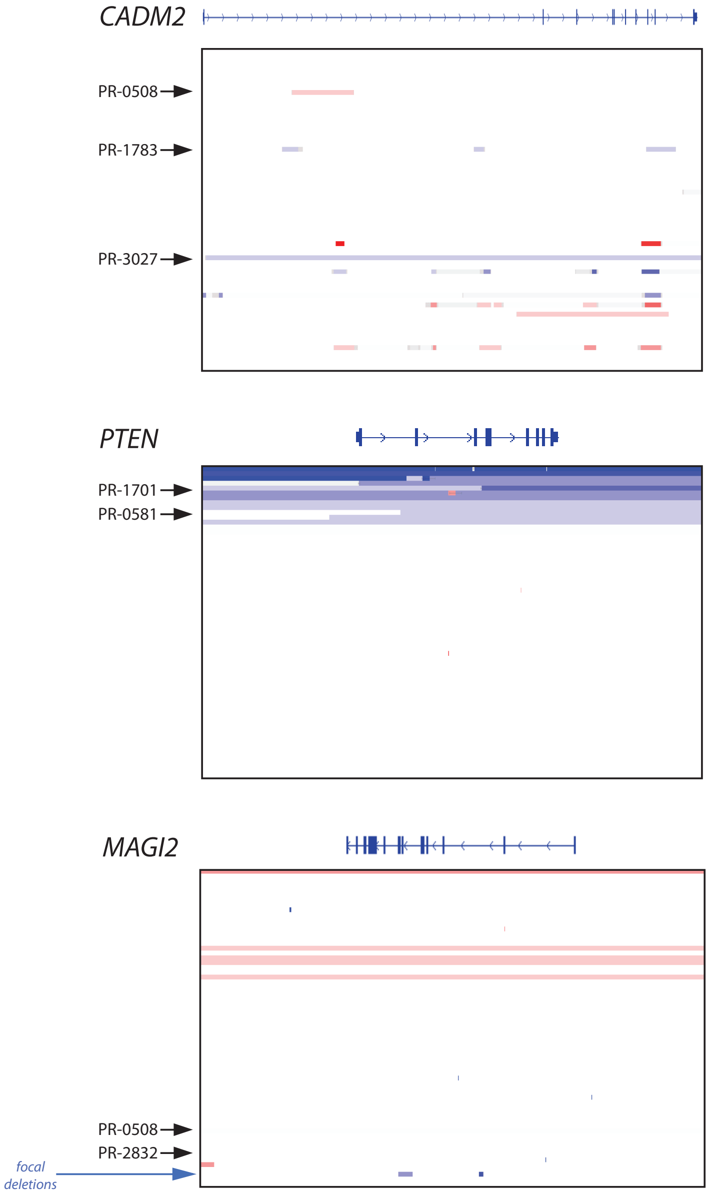
Supplementary Figure S8: FISH validation studies on an independent prostatectomy cohort. Clinically localized prostate cancer samples from Weill Cornell Medical College were screened by FISH to examine the prevalence of particular classes of rearrangements involving *CADM2*, *MAGI2*, and *CSMD3*. The genomic positions of FISH probes are illustrated at left, and representative FISH images from the independent cohort are shown at right. Results are summarized in the table. All samples with detectable *MAGI2* inversions are wild type for *PTEN*. The rates shown here include only those cases detectable by the specific FISH probes used in this experiment and thus represent a lower bound for the true prevalence of rearrangements disrupting these genes.



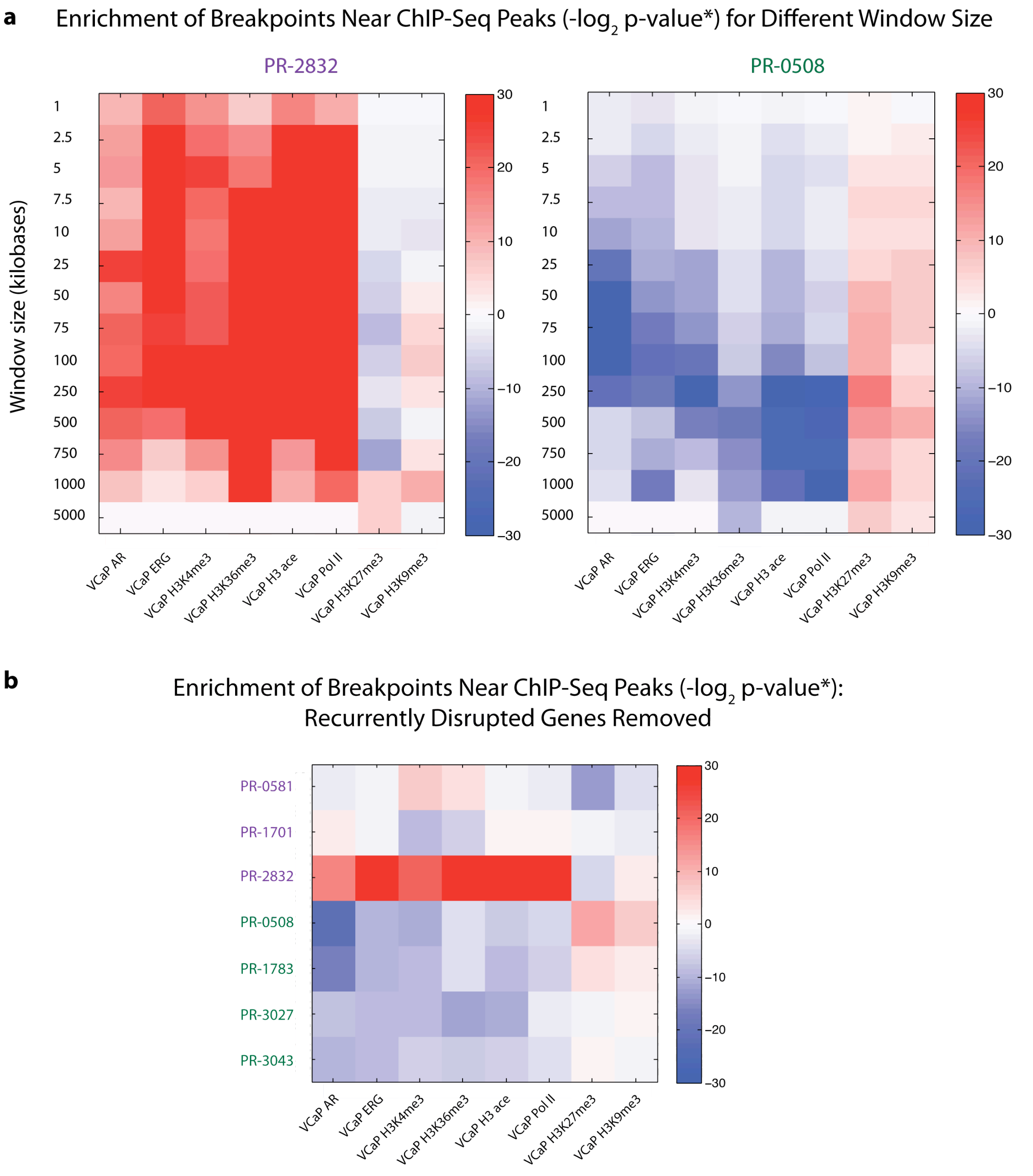
Supplementary Figure S9: Comparison of *CADM2*, *CSMD3*, and *MAGI2* to known fragile sites. Genome-wide copy number profiles from >3,300 human tumors and cell lines, as analyzed by Beroukhim *et al.*[19], were visualized using the Integrative Genomics Viewer (Robinson J.T. *et al.*, submitted; http://www.broadinstitute.org/igv). For each locus, a 1-megabase interval was defined, and all samples were sorted according to the sum of the magnitudes of all copy number alterations within the interval. Illustrated in each panel are the top ~8% of samples sorted in this fashion. The three genes recurrently rearranged in prostate cancer are indistinguishable from the random loci and exhibit far fewer copy number breakpoints than the known fragile sites. FRA7K is a novel fragile site described by Bignell and colleagues[20].



Supplementary Figure S10: FISH confirmation of *PTEN* rearrangements. Intragenic breakpoints in *PTEN* in prostates PR-0581 and PR-1701 generate unbalanced heterozygous deletions that were confirmed by FISH analysis. FISH was performed as described in Supplementary Methods.



Supplementary Figure S11: Copy number analysis of intragenic breakpoints in *CADM2*, *PTEN*, and *MAGI2*. Genome-wide copy number profiles were determined for 66 prostate tumors using Affymetrix SNP 6.0 microarrays (red = copy gain; blue = copy loss). Samples (rows) with intragenic breakpoints identified from whole genome sequencing are labeled.



* For *depletion* of breakpoints (rather than enrichment), -log p-values are multiplied by -1.

Supplementary Figure S12: Robustness of association between rearrangement breakpoints and chromatin marks. (a) Effect of varying window size on enrichment calculation. Enrichment/depletion near ChIP-Seq binding peaks is consistent for PR-2832 (left) and PR-0508 (right) when different sized intervals flanking binding peaks are considered. (b) Effect of removing candidate driver genes. We obtained nearly identical results after removing 57 rearrangements involving 16 genes recurrently disrupted in the 7 prostate tumors (see Supplementary Figure S5), suggesting that the observed associations are not the result of a small number of driver genes.

IV. References

1. Tomlins, S.A*., et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).
2. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
3. Li, H*., et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
4. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185 (1998).
5. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).
6. Reich, M*., et al.* GenePattern 2.0. *Nat Genet* **38**, 500-501 (2006).
7. Korn, J.M*., et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260 (2008).
8. Getz, G*., et al.* Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317**, 1500 (2007).
9. Ding, L*., et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075 (2008).
10. Demichelis, F*., et al.* Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer* **48**, 366-380 (2009).
11. Yu, J*., et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443-454 (2010).
12. Lin, B*., et al.* Integrated expression profiling and ChIP-seq analyses of the growth inhibition response program of the androgen receptor. *PLoS One* **4**, e6589 (2009).
13. Birney, E*., et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
14. Carroll, J.S*., et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297 (2006).
15. Pleasance, E.D*., et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).
16. Pleasance, E.D*., et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-190 (2010).
17. Lee, W*., et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).
18. Stephens, P.J*., et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010 (2009).
19. Beroukhim, R*., et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
20. Bignell, G.R*., et al.* Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893-898 (2010).