SUPPLEMENTAL DATA

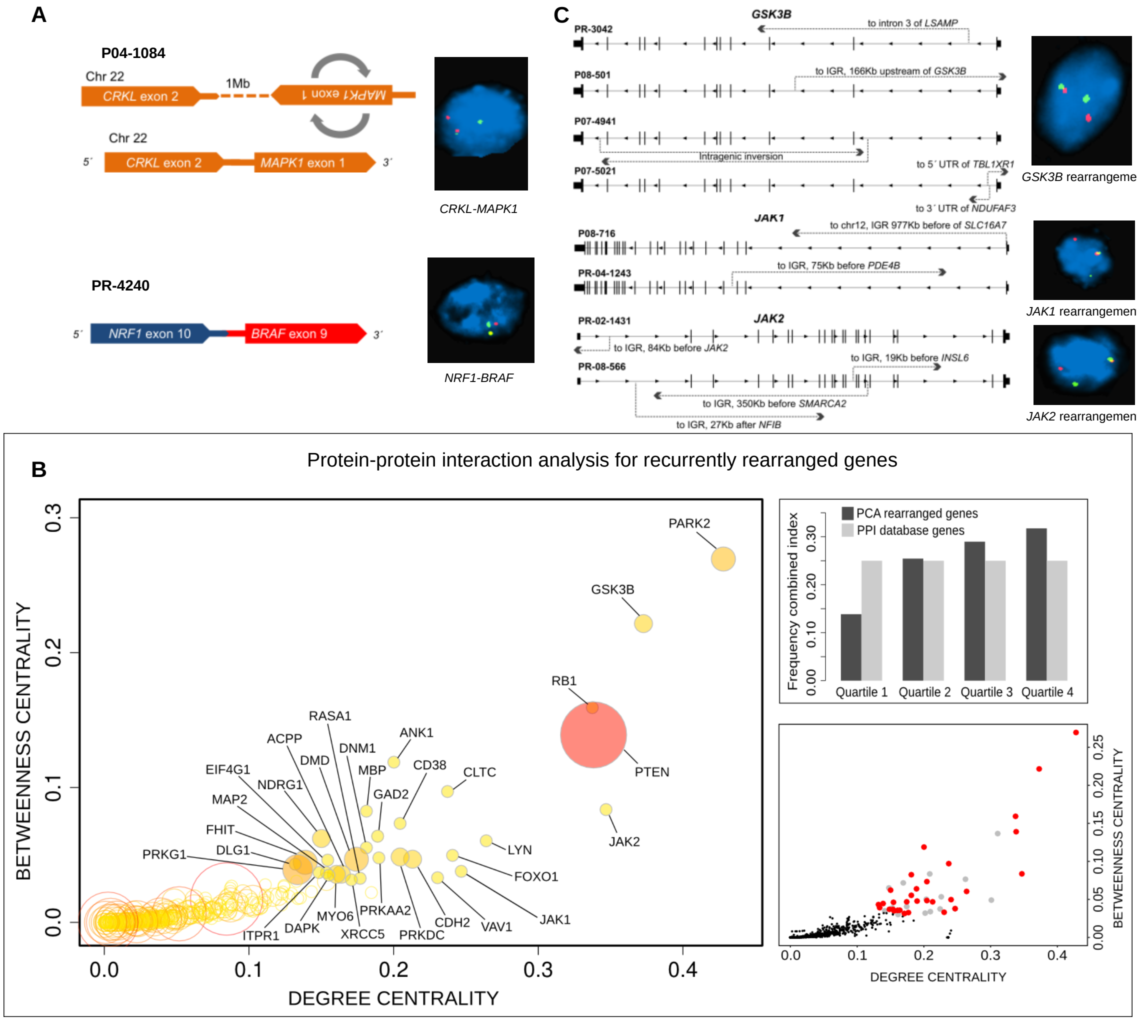

Figure S1. Recurrent rearrangements alter known and putative prostate cancer genes, Related to Figure 1.

(A) Schematic of *CRKL-MAPK1* and *NRF1-BRAF* fusions detected by WGS, along with validation by FISH assay.

(B) Protein-Protein Interaction (PPI) data were analyzed to nominate rearrangements of potential biological consequence. The centrality in a PPI network (Szklarczyk et al., 2011) was assessed for protein products of genes that were rearranged in more than one sample (total 397). X- and Y-axes measure two indexes of centrality, where larger values indicate more central network positions. Circle color and size are proportional to the frequency of gene rearrangement across the tumor cohort. Genes scoring in the 95th percentile are depicted as filled circles. The two panels on the right show the centrality of recurrently rearranged genes (depicted as red circles in the bottom plot) compared to the entire PPI dataset.

(C) Disruptive genomic rearrangement of *JAK1*, *JAK2* and *GSK3B*. Dotted lines show intragenic breakpoints and corresponding text indicates the locus to which the breakpoint is fused (IGR; inter-genic region). Rearrangements depicted above the gene diagrams occurred in a sense-preserving orientation; rearrangements below gene diagrams occurred in an anti-sense orientation. Right, genomic rearrangements were validated by FISH.

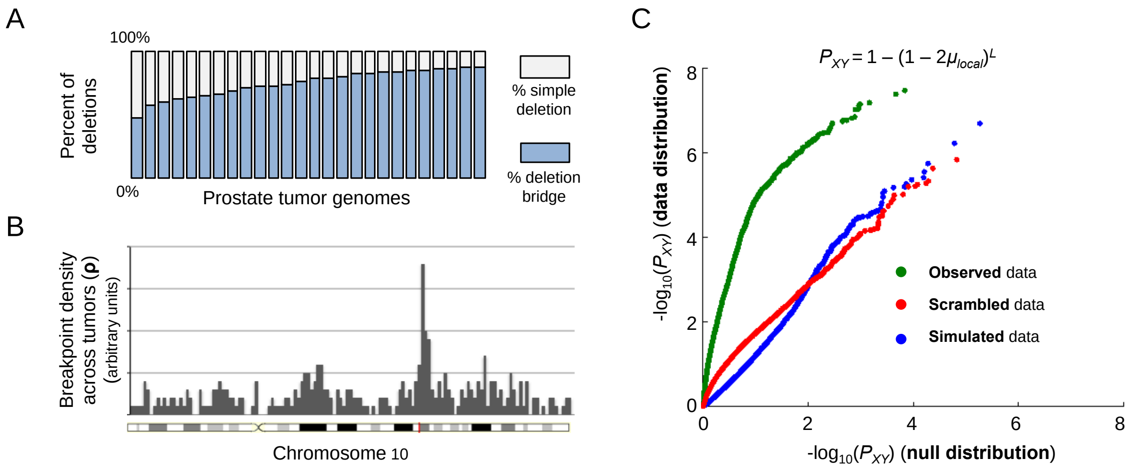

Figure S2. Signatures of coordinately generated rearrangement chains, Related to Figure 2.
(A) Percentage of DNA deletions bounded by fusion breakpoints that were uniquely identified as deletion bridges (blue) or simple deletions (white) in prostate tumors with ten or more deletions in either category.
(B) Probabilistic model of independent rearrangements across the genome. The expected distribution of independent DNA breaks in a given tumor (ρ) is estimated by counting the number of tumors with one or more rearrangements within 1Mb tiling windows across the genome. ρ is used to calculate the value of μ_{local} used by ChainFinder in the null model of independent breakpoints.
(C) Quantile-quantile (Q-Q) plot comparing the distribution of P_{XY} values (the adjacency probabilities for independent breakpoints) for observed, simulated and scrambled rearrangements.

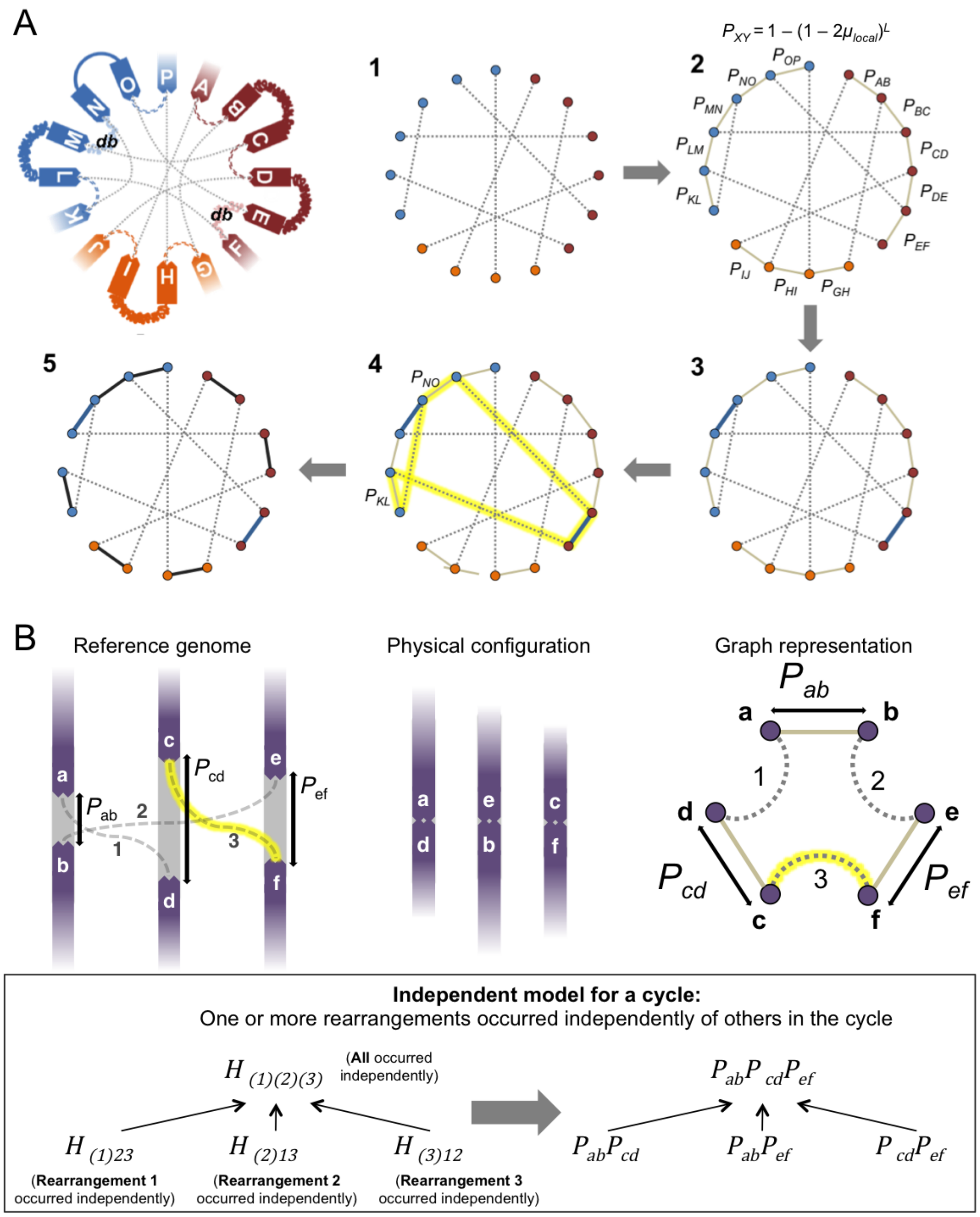

Figure S3. ChainFinder analysis of cancer genomes, Related to Figure 3.
(A) ChainFinder creates a graph representation of somatic rearrangement breakpoints and corresponding deletions (see Supplemental Experimental Procedures for an extended explanation). (1) Breakpoints of somatic fusions are represented as nodes connected by edges. (2) The adjacency probability (P_{XY}) is calculated for pairs of neighboring breakpoints based on their reference genome distance (L) and the local rate of rearrangements (μ_{local}). (3) Breakpoints at either boundary of a deletion bridge are joined by edges. (4) The graph is searched for cycles connecting breakpoints that are unlikely to have arisen independently, based on P_{XY} values of corresponding intervals. (5) The final graph contains sets of rearrangements and associated deletions that are unlikely to have occurred independently.
(B) For a hypothetical cycle involving three rearrangements, the independent breakpoint model constitutes all scenarios by which any rearrangement could have arisen independently of others in the cycle (see Supplemental Experimental Procedures for an extended explanation).

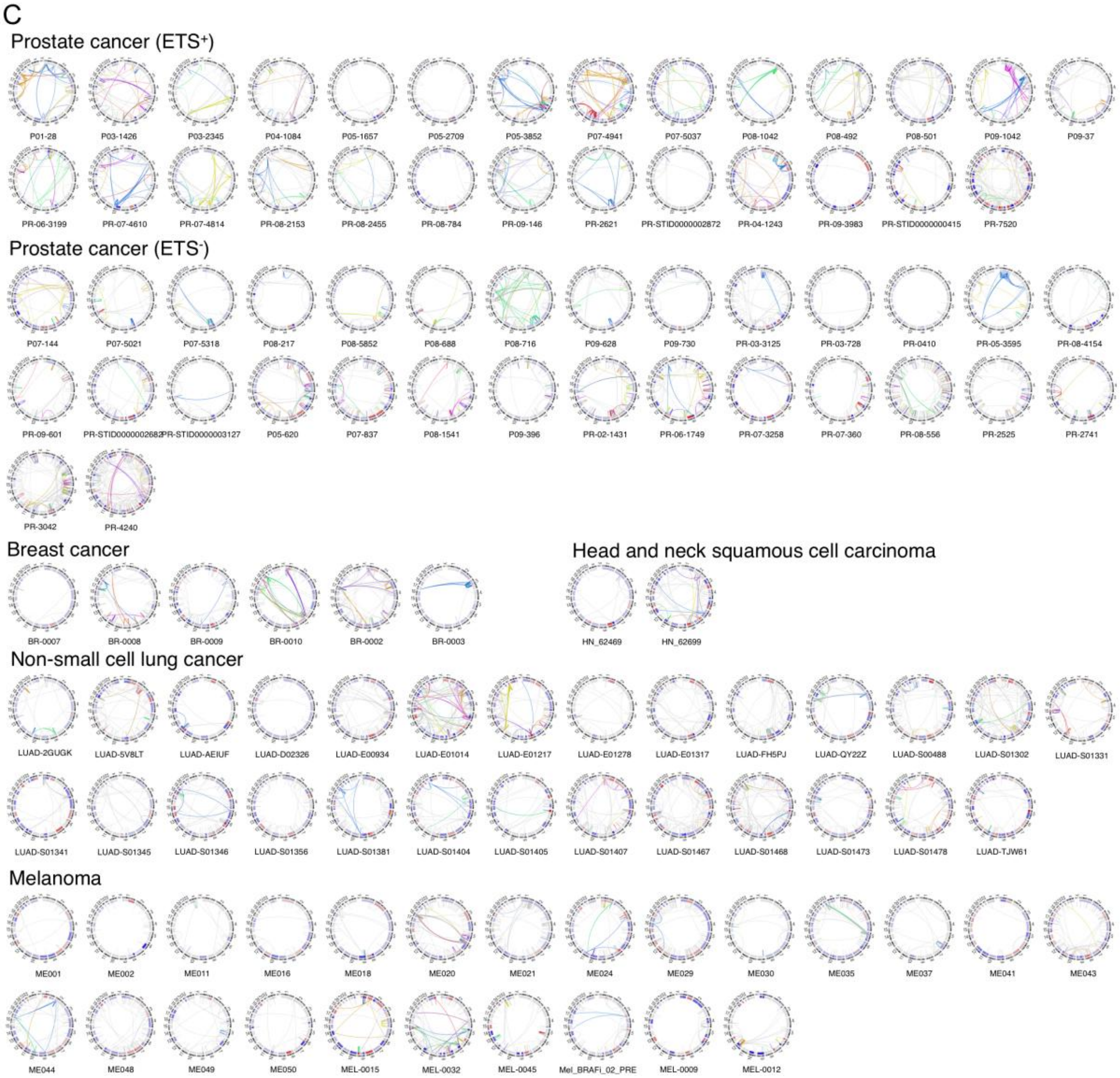

Figure S3, continued. (C) Circos plots of rearrangements color-coded by chain for 57 prostate tumors and 59 previously sequenced cancer genomes (please see Table S5B for references). Rearrangements in gray were not assigned to a chain. Copy number alteration is shown in blue (deletion) and red (amplification) in the inner ring of each plot.

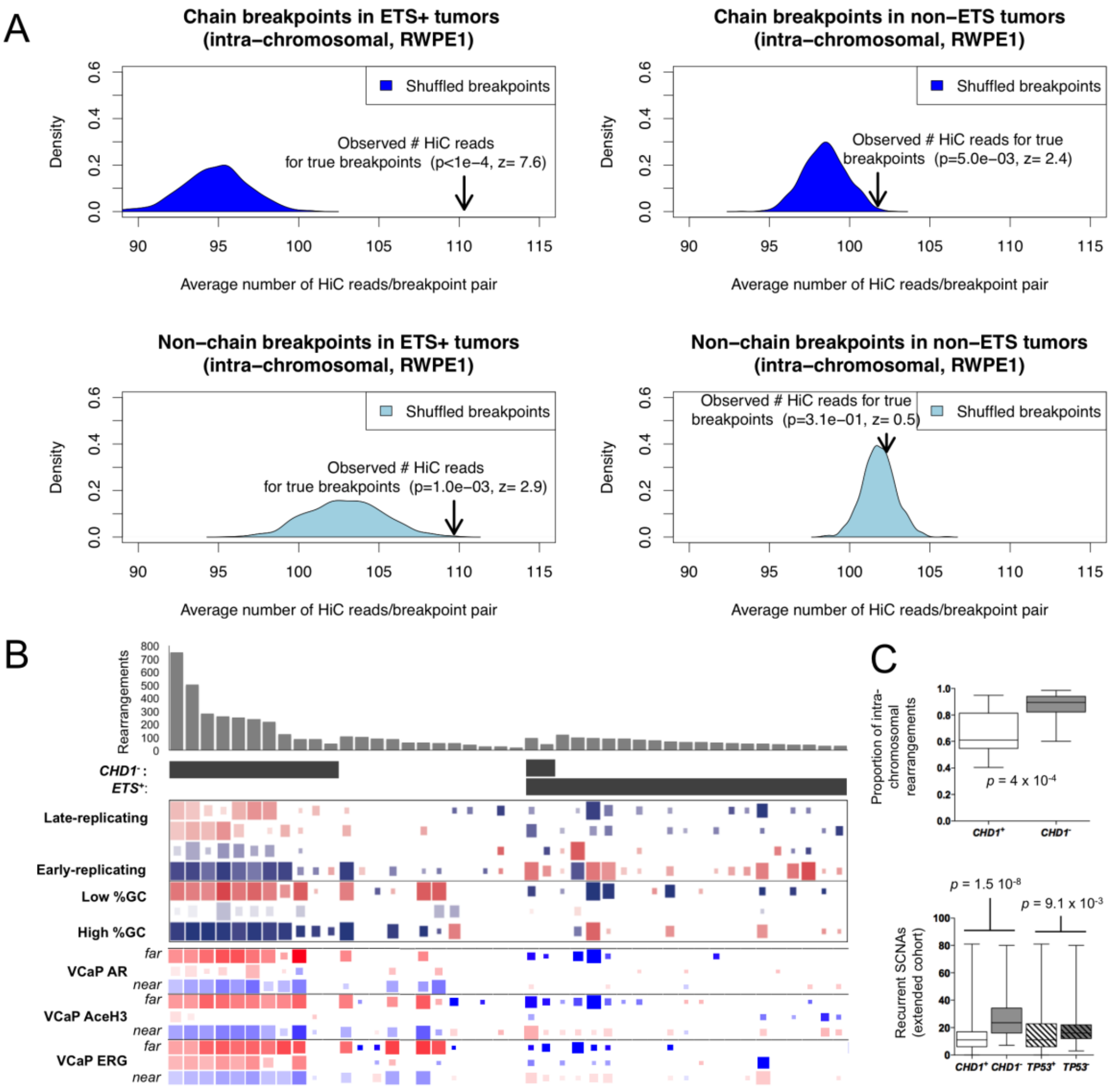

Figure S4. Rearrangement profiles of prostate tumor genomes, Related to Figure 4.

(A) Chromoplexy arises in physically interacting chromatin. Chains are enriched for rearrangements that fuse portions of the genome in close physical proximity as determined by Hi-C analysis of the RWPE-1 prostate epithelial cell line (Rickman et al., 2012). Further details are provided in the Supplemental Experimental Procedures.

(B) Enrichment of rearrangement breakpoints near to and distant from various genomic features, including ChIP-seq peaks from ERG$^+$ VCaP prostate cancer cells (Yu et al., 2010). Color hue reflects the degree of enrichment (red) or depletion (blue) and box area reflects statistical significance. “Near” and “Far” correspond to within 100kb and further than 500kb, respectively. The number of rearrangements for each tumor is depicted in the gray columns (see Supplemental Experimental Procedures).

(C) Recurrent somatic copy number alterations (SCNAs) across an extended panel of 199 prostate tumors grouped by *CHD1* deletion status. For comparison, the same samples are also grouped by *TP53* deletion status. Median, middle quartiles, and range are indicated.

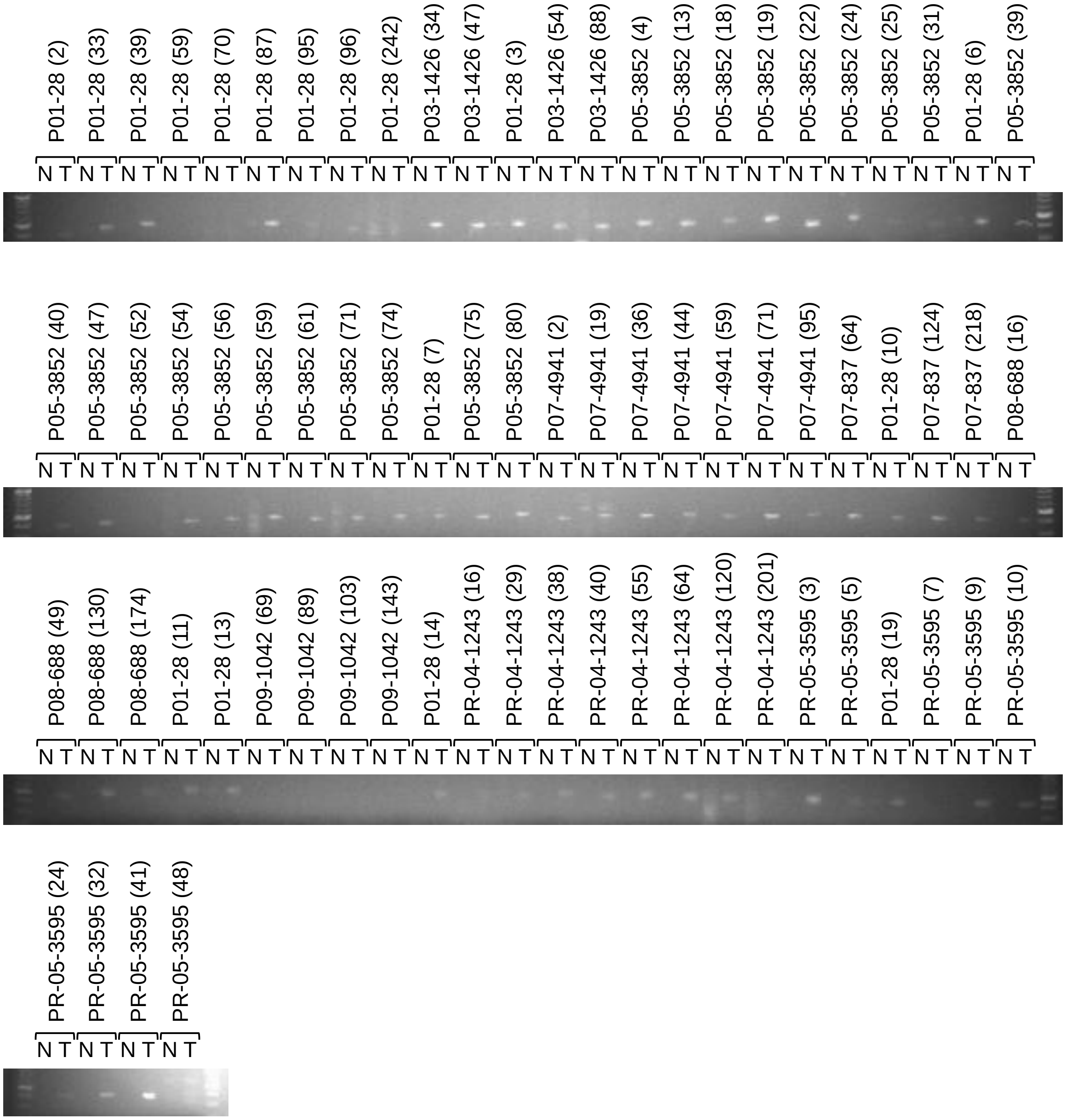

Figure S5. PCR validation of chained DNA rearrangements, Related to Figure 5.
PCR reactions were run on tumor and normal DNA to amplify across the junctions of 76 somatic fusions. Rearrangements are numbered as in Table S3C. Please see Table S3C for a list of additional rearrangements that were validated by PCR and deep sequencing on the MiSeq platform.

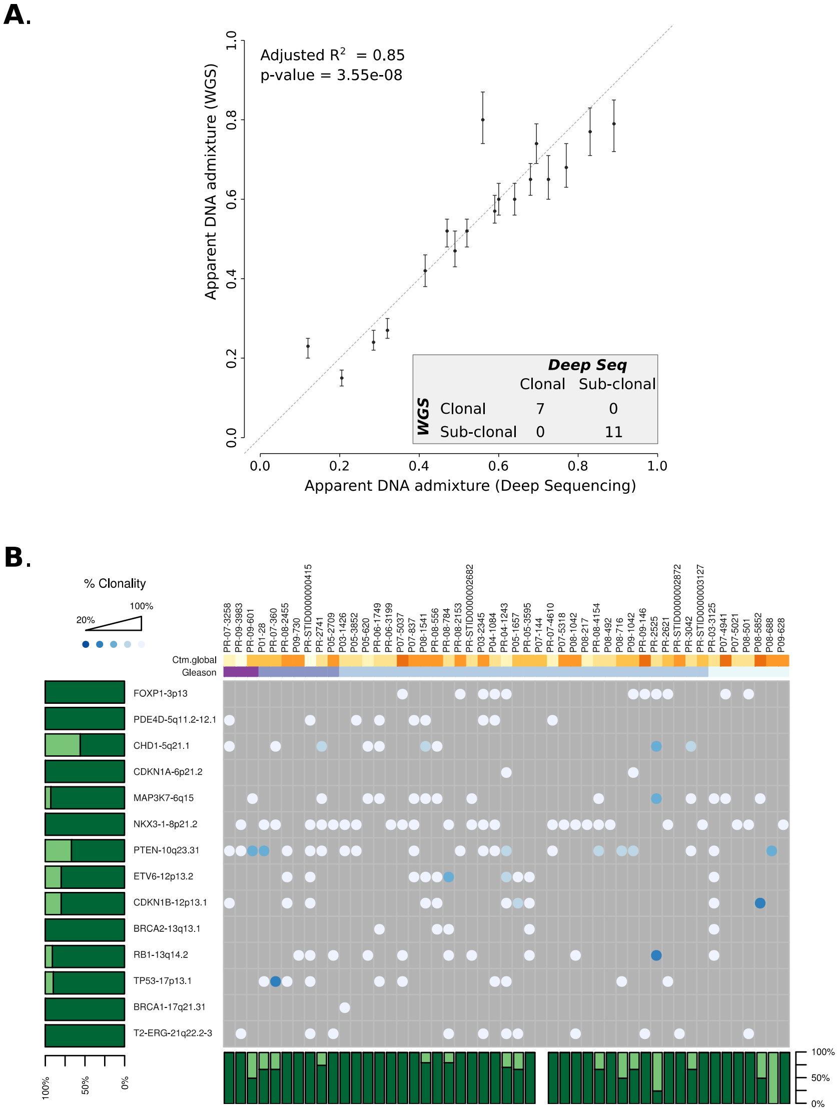

Figure S6. Estimation of clonality and stromal DNA admixture, Related to Figure 6.
(A) Apparent DNA admixture computed from WGS and MiSeq deep-sequencing data for 18 somatically deleted genes in 7 samples. Error bars for WGS estimations are computed according to Table S6A. Clonality calls on WGS data were made with a minimum of 20 informative hemizygous SNPs covered to an average depth of 20x or greater. MiSeq calls are based on 4 SNPs with average local coverage of >65,000x. The contingency table (bottom-right) shows the agreement for clonality and sub-clonality calls between MiSeq and WGS based data (Cochran test, p-value=1).
(B) Clonal status of deletions at 14 loci inferred across 49 prostate cancers. The central panel denotes the clonal status of a gene lesion in a sample. Empty dark gray rectangles indicate either that the gene was not deleted or that there were insufficient informative SNPs to determine clonality status. White circles indicate a 100% clonal deletion. Colored circles indicate sub-clonal deletions, where darker color indicates a more subclonal deletion. Top rows report Gleason scores, ranging from 6 (light blue) to 9 (violet), and global stromal DNA admixture, where darker color signifies more admixture. Green bars summarize lesion clonality on a per-sample and per-gene basis. Dark and light green denote the proportion of clonal and sub-clonal deletions, respectively.

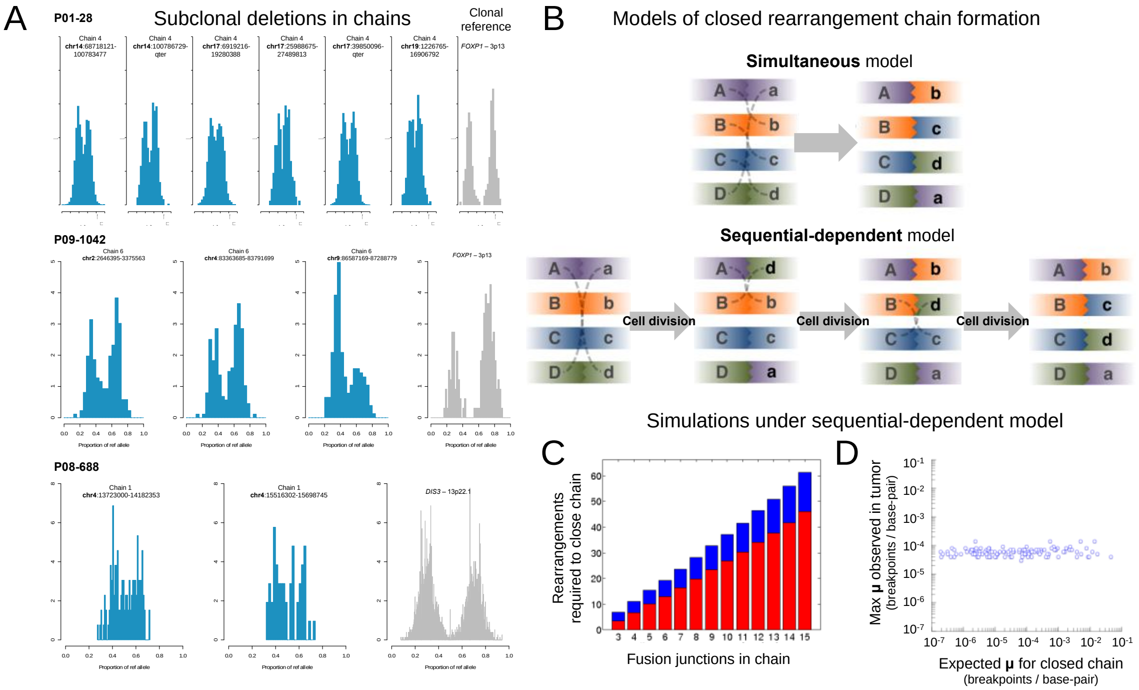

Figure S7. Chromoplexy continues during outgrowth of tumor sub-clones and generates multiple rearrangements in closed chains at once, Related to Figure 7.

(A) Three examples of subclonal chains identified by clonality analysis of deletion bridges. Allelic fraction distributions of heterozygous SNPs within the deleted segments are indicated. On the right, a clonal deletion bridge from the same sample is shown for comparison.

(B) Closed chains of non-independent rearrangements could arise from (1) a series of balanced translocations over multiple cell generations (the "sequential-dependent model") or (2) concerted rearrangements within one cell cycle (the "simultaneous model").

(C) For closed chains of rearrangements, bars indicate the median number of sequential balanced translocations required to close a chain under the sequential-dependent model (assuming translocations occur randomly between breakpoints within the chain). Average values from 10,000 simulations per chain size are shown in blue. Red bars indicate the number of rearrangements that disrupt a previously formed fusion junction.

(D) For 121 observed closed chains, the values from (C) and genomic distances between chain breakpoints were used to calculate the local rate of rearrangements required to close the chain under the sequential-dependent model. This density is compared to the maximum density of rearrangements observed in the tumor containing the chain (assessed in 10kb windows genome-wide).

Table S1. Clinical characteristics of genome-sequenced prostate tumors, Related to Figure 1.

Table S2. Sequencing metrics of 57 prostate cancer whole genomes, Related to Figure 1.

Table S3. Somatic DNA alterations in 57 prostate cancers, Related to Figure 1.
(A) Somatic point mutations and small insertions/deletions in protein-coding genes identified by WGS.
(B) Segmented copy number data for the tumor cohort from Affymetrix SNP 6.0 arrays (seg format).
(C) Somatic genomic rearrangements identified by WGS.

Table S4. Outlier expression of rearranged genes, Related to Figure 1.
For genes scoring above the 85th percentile in the PPI centrality analysis shown in Figure S1, expression was assessed from transcriptome sequencing data. Median expression of the gene in all samples is listed alongside the expression percentile in sample(s) with rearrangement of the gene. Fields are highlighted where samples with disruption of the gene show outlier expression in the top or bottom tenth percentile of all other tumors. Empty fields denote genes where transcriptome data were not obtained from samples with rearrangement of the gene.

Table S5. Summary of rearrangement chains, Related to Figure 3.
(A) Chains of co-generated somatic rearrangements and deletions detected by ChainFinder.
(B) Summary of chromoplexy in prostate cancer, other tumor types and simulated or scrambled tumor genomes.
(C) Known and putative prostate cancer genes dysregulated by chromoplexy. Genes were chosen from a list of potential tumor suppressor genes from the KEGG database (Kanehisa et al., 2012) ("KEGG") or were curated from published literature ("Additional").

Table S6. Clonality analysis of prostate tumor genome alterations, Related to Figure 6.
(A) Uncertainty in estimates of apparent DNA admixture. For a somatic deletion with a given number of informative heterozygous SNPs with a given mean sequence coverage, the table reports the uncertainty in the estimate of the apparent DNA admixture (*adm*). Estimations are created by randomly sampling 1,800 simulations created using various combinations of mean coverage, SNP number and true DNA admixture. The uncertainty is the mean difference between the computed versus the true *adm* values.
(B) Estimation of deletion clonality. For each sample and each deleted gene included in the study, the estimated level of deletion clonality is listed with the range of possible clonality values computed according to the uncertainty table in (A).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Description of the Tumor Cohort

Prostate cancers analyzed in this study originated from two cohorts: Weill Cornell Medical College (WCMC; New York, NY) and Uropath Pty Ltd. (Perth, Australia), a provider of banked urological tissues. All prostate cancer samples were collected under an Institutional Review Board-approved protocol with the informed consent of patient donors. Sixteen tumors were characterized by exome-sequencing in a previous study (Table S1) (Barbieri et al., 2012). Previous analyses of SNP data from these cohorts indicated that patients were primarily of Caucasian ancestry (Barbieri et al., 2012). Primary adenocarcinomas were removed prior to any additional treatment for prostate cancer, including radiation therapy, brachytherapy or hormone ablation therapy.

Primary adenocarcinomas from the WCMC cohort were collected from patients undergoing radical prostatectomy by one surgeon (A. Tewari; under IRB number: 0407007351) 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 Hospital (New York, NY). This cohort included neuroendocrine prostate cancer (NEPC) cases, which were obtained under an IRB approved protocol from lung (PR-4240) and abdominal wall soft tissue (PR-7520) metastases by H. Beltran. Patient-matched normal DNA was obtained from whole blood samples for this cohort.

Tumors from the Uropath cohort were obtained from men undergoing radical prostatectomy for clinically localized prostate cancer across multiple medical centers in Western Australia. Radical prostatectomies were performed by one of 30 clinicians between 2000 and 2010. Samples were stored at -84°C. Paired normal DNA was obtained from benign prostate tissue. To extract normal DNA, we identified a frozen tissue block with no histological evidence of neoplasia in order to minimize the possibility of contamination from tumor DNA.

In both cohorts, Hematoxylin and Eosin (H&E)-stained tissue sections were reviewed by M. Rubin and K. Park to verify Gleason score and to determine the percentage of Gleason pattern 4 and 5 histology at the site selected for DNA extraction. NEPC samples were reviewed by the study pathologists and confirmed as neuroendocrine carcinomas of prostatic origin based on clinical history and/or presence of ERG fusion (PR-7520). Immuno-histochemistry was negative for PSA and positive for the neuroendocrine marker synaptophysin in both cases.

High-density SNP Array Analysis

Genomic DNA from tumor samples was profiled with Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc.) according to the 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). After array scanning, data preprocessing was performed using Affymetrix Power Tools. Quality control, data integrity, segmentation and copy number analysis were performed as previously described (Demichelis et al., 2009), except that tumor copy number data were compared to a reference model built using a set of peripheral blood cell DNA samples. Regions of copy deletion and amplification were defined with a $\log_2$ ratio cutoff of <-0.1 and > 0.1, respectively.

Sequencing Data Generation

WGS Library Construction

Libraries were prepared as described previously (Fisher et al., 2011) with slight modifications. First, the genomic DNA input into shearing was reduced from 3µg to 100ng in 50µL of solution. In addition, for adapter ligation, Illumina paired-end adapters were replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter. Size selection was then performed using Sage Bioscience’s Pippin Prep, with a target insert size of either 340bp or 370bp +/- 10%.

Following sample preparation, libraries were quantified using quantitative PCR (KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2nM and then denatured with 0.1 N NaOH using Perkin-Elmer's MiniJanus liquid handling platform.

RNA-Seq Library Construction

RNA was isolated using a Dynabeads® mRNA Purifcation Kit (Life Technologies). Two rounds of poly-A selection (with bead regeneration) were performed to achieve rRNA contamination of less than 10%, as assessed by the Bioanalyzer mRNA Pico program (Agilent). Eluate was treated with DNase (TURBO DNA-free™ kit, Ambion) at 37°C for 30 minutes then immediately cleaned using RNAClean XP beads (Agencourt). RNA was fragmented in Fragmentation Buffer (Affymetrix, 900371) at 80˚C for 4 minutes. First- and second-strand cDNA synthesis were performed with SuperScript Double-Stranded cDNA Synthesis Kit (Life Technology). Library construction proceeded as described previously (Fisher et al., 2011), except that SPRI beads were used in the end-repair cleanup and standard paired-end adapters were replaced with barcoded adapters each containing a unique 8-base index sequence. After adapter ligation, two sequential cleanups were performed to remove adapter dimers, followed by 8 cycles of cDNA PCR amplification and SPRI cleanup. Before sequencing, samples were pooled and normalized according to qPCR results.

Cluster Amplification and Sequencing

Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using HiSeq v3 cluster chemistry and flowcells. Flowcells were sequenced with 101-bp paired end reads on a HiSeq 2000 using HiSeq v3 Sequencing-by-Synthesis Kits and analyzed using RTA v.1.12.4.2.

Genome Sequence Analysis

Sequencing Data Management and Processing

A BAM file was generated for each sample from Illumina sequence reads using the Picard pipeline (http://picard.sourceforge.net/). Reads were mapped to the NCBI Human Reference Genome GRCh37 (hg19) with the Burrows-Wheeler Aligner (BWA) (http://bio-bwa.sourceforge.net). The cancer genome analysis pipeline known as Firehose (Principal author D. Voet) was used to manage and coordinate analysis of WGS data. Firehose submits input files and parameters to GenePattern (DePristo et al., 2011), which executes a series of analyses to verify data quality and detect somatic alterations by comparing tumor and normal sequences.

Quality Control

We employed several quality control modules to monitor for contamination or potential sample mix-ups. To ensure that tumor- and normal-DNA were properly matched for a given individual and free of contaminating human sequences, we generated SNP fingerprints from 24 highly polymorphic sites for each sequencing lane. Lanes with outlier fingerprint genotypes for a given individual were discarded. In addition, we used the ContEst algorithm (Cibulskis et al., 2011) to analyze homozygous non-reference SNPs to estimate levels of contamination with foreign human DNA, and required that samples demonstrate >95% concordance.

Normal DNA sequences were assessed for admixture with cancer DNA by examining copy number profile estimates based on sequence coverage in 100kb bins across the genome. Three samples (PR-07-3258, PR-09-3983 and P05-2709) demonstrated low-level contamination with tumor DNA, based on similar patterns of DNA gains and losses between tumor and normal in a pair. In these cases, the histologically benign prostate tissue used as a source of normal DNA likely contained neoplastic or pre-neoplastic cells. The detection of somatic alterations in these samples was therefore limited. We specifically analyzed discarded rearrangement calls from these samples for prostate cancer-associated fusions, and identified cases in which the *TMPRSS2-ERG* fusion was filtered out due to presence in normal (see below).

Detection of Chromosomal Rearrangements

Detection of somatic rearrangements was performed using the dRanger algorithm (Berger, 2011) to identify sequence reads from paired ends that map to the reference genome with unexpected orientations or intervals between read pairs. Candidate rearrangements were identified from clusters of such reads. They were then assigned a score by multiplying the number of chimeric reads supporting the fusion by a quality multiplier between 0 and 1. The quality multiplier takes into account the following factors: (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. Rearrangements with score of 4 or greater were classified as high-confidence. Rearrangements were categorized as deletions, inversions, interchromosomal translocations or tandem duplications based on the locations and strand directions of reads at fusion breakpoints.

In three cases, the *TMPRSS2-ERG* fusion was detected but filtered out, either due to low levels of tumor contamination in the adjacent prostate tissue used as a normal comparator (PR-09-3983, P05-2709), or to an abundance of breakpoints at the locus that resulted in a low quality multiplier (PR-STID0000000415). The *TMRPSS2-ERG* fusion was confirmed by fluorescence in situ hybridization (FISH) in these and all other fusion-positive cases.

Some loci that were rearranged in the context of a chain harbored many breakpoints that decreased the rearrangement quality score and caused the rearrangement to be rejected, despite support from multiple tumor reads and the absence of the rearrangement in normal DNA. To improve our ability to detect chains in these situations, we adjusted the parameters of dRanger so that rearrangements were considered if they demonstrate five supporting reads in tumor DNA, no reads in the panel of normal genomes and a score of 1 or greater. Rearrangements falling into this category were retained in the final dataset only if they were assigned to a chain.

Breakpoint fusion junctions were mapped to base-pair resolution where possible using the BreakPointer algorithm (Drier et al., 2012). BreakPointer searches for read pairs where one read maps near a breakpoint and the pair mate partially overlaps with the fusion junction, or fails to align anywhere. 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. BreakPointer mapped the breakpoints to base pair resolution in 94% of the 5596 high-confidence rearrangements. In these cases, sequence homology at fusion junctions and any foreign sequence insertions were annotated.

Rearrangements were annotated with transcript information from the UCSC Genome Browser’ s UCSC Genes track (Table S3C) (Fujita et al., 2011) and illustrated using Circos (http://mkweb.bcgsc.ca/circos)

Identification of Somatic Single Nucleotide Variants (SSNVs)

We used the MuTect algorithm from the Broad Institute Genome Analysis Toolkit (GATK) to identify SSNVs (~~Cibulsksis K. et al., in preparation; https://confluence.broadinstitute.org/display/CGATools/MuTect~~Cibulskis et al., 2013; www.broadinstitute.org/cancer/cga). As previously described (Berger et al., 2011; Stransky et al., 2011), MuTect identifies candidate SSNVs by performing a statistical analysis of the bases and their qualities in the tumor and normal BAMs at the genomic locus under examination. Base-pairs were required to be covered by at least 14 reads in the tumor and 8 in the normal for mutation detection.

MuTect first filters out reads with low quality scores or excessive mismatches. A statistical analysis is then performed to identify somatic mutations using Bayesian classifiers for the tumor and normal sequences at a given locus:

$$LOD_T = log_{10}\left(\frac{P(\text{observed data in tumor} \mid \text{site is mutated}}{P(\text{observed data in tumor} \mid \text{site is reference}}\right)$$

$$LOD_N = log_{10}\left(\frac{P(\text{observed data in normal} \mid \text{site is reference}}{P(\text{observed data in normal} \mid \text{site is mutated}}\right)$$

Thresholds were chosen for each statistic to achieve a sufficiently low false positive rate. Several post-processing filters are applied to remove artifactual calls. For example, mutations are excluded that appear solely at the 5' or 3' end of reads or that are identified in panels of genomes from non-cancerous tissue. The accuracy of MuTect has been estimated to be greater than 95% by orthogonal validation of mutations in previous sequencing studies (Barbieri et al., 2012; Stransky et al., 2011). Mutations in known cancer-associated genes were reviewed manually using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

Local Realignment and Detection of Indels

To improve detection of small insertions and deletions (indels), reads in tumor and paired normal were jointly realigned at genomic locations harboring putative indels by the local realignment module in the GATK (DePristo et al., 2011). Putative indels were then considered at sites that were well covered in tumor and normal where the indel-containing allele was supported by 8 or more reads or 30% of all reads from the locus. Next, these indel calls were filtered based on local alignment statistics around the putative event, including the average number of additional mismatches per indel-supporting read, average mismatch rate and base quality in a small window around the indel (The Cancer Genome Atlas Research Network, 2011).

Mutation Annotation

Point mutations and indels were annotated with information about relevant genes, transcripts, proteins and other features using publicly available databases. A set of reference transcripts was compiled for annotation from the UCSC Genome Browser’ s UCSC Genes track as provided in the TCGA General Annotation Files (GAF) hg19 June 2011 bundle (https://tcga-data.nci.nih.gov/docs/GAF/). Variants were also annotated using the following resources: dbSNP build 134 (Sherry et al., 2001), UCSC Genome Browser’ s ORegAnno track (Griffith et al., 2008), UniProt release 2011_09 (Consortium, 2011) and COSMIC v55 (Forbes et al., 2011).

Validation of Somatic Mutations and Rearrangements

Mutation Validation from Transcriptome Sequencing

We assessed 818 somatic point mutations covering annotated transcripts in RNA-Seq data from 20 tumors profiled by transcriptome sequencing (Table S3A). Of the mutated sites, 92 were covered by 40 or more RNA-Seq reads and present in WGS reads at an allele fraction of 0.2 or greater. Of these mutations, 84 (91%) showed at least two reads supporting the alternate allele.

Validation of Somatic Rearrangements

Rearrangements were validated by two approaches. We assessed a set of 73 rearrangements, enriched for events affecting cancer genes, by PCR and deep sequencing on a MiSeq instrument (Table S3C). Reads from tumor and normal DNA were aligned to a custom genome that contained the hg19 reference genome along with sequences of all predicted fusion junctions across samples. Rearrangements were classified as somatic if tumor, but not normal alignments, showed multiple high-quaity reads spanning the predicted fusion junction.

In addition, we selected 76 chromoplexy-associated rearrangements for validation by PCR alone (Figure S5, Table S3C). Primers were designed to amplify approximately 200bp containing the predicted fusion junction. Rearrangements were annotated as somatic if a band of the predicted size was amplified from tumor DNA but not from normal DNA.

Fluorescence in Situ Hybridization Validation of Rearrangements

ETS rearrangement was assessed using break-apart assays for *ERG* and *ETV1* as described previously (Berger et al., 2011). To assess genomic deletion, gene fusion and disruptive translocations, we used locus-specific dual-color FISH assays following a previously described approach (Berger et al., 2011; Perner et al., 2006). At least 50 nuclei were evaluated per tissue section using a fluorescence microscope (Olympus BX51; Olympus Optical, Tokyo, Japan). The following probes were used for FISH assays:

Locus	**BAC #**
CHD1	RP11-58M12
CHD1 Reference (5p13.1)	RP11-429D13
GSK3B 3'	RP11-59M4
GSK3B 5'	RP11-113H22
JAK1 3'	RP11-1061K17
JAK1 5'	RP11-76O023
JAK2 3'	RP11-274A3
JAK2 5'	RP11-259N10
CRKL 3'	RP11-76I4
CRKL 5'	RP11-1152E2
MAPK1 3'	RP11-317J15
MAPK1 5'	RP11-179H3
PTEN	CTD-2047N14
PTEN Reference (10q25)	RP11-431P18
FOXP1	RP11-410B2
FOXP1 Reference (3p11)	RP11-91M15
BRAF 3'	RP11-248P7
BRAF 5'	RP11-248O23

Protein-Protein Interaction (PPI) Analysis of Somatically Rearranged Genes

To identify gene rearrangements of potential biological consequence in Figure S1, we searched for recurrently rearranged genes whose protein products occupy central positions in interaction networks. To assess protein-protein interaction (PPI) network centrality, we considered the product of two measures of degree centrality and betweenness centrality:

1. **Degree centrality:** Given a protein p and a PPI network N, the index Degree(p,N) measures the number of interactions incident upon p. The index is normalized by dividing D(p,N) with the maximum index in the network.

2. **Betweenness centrality:** Given a protein p and a PPI network N, the index Betweenness(p,N) measures the number of shortest paths from all proteins to all others that pass through protein p. The index is normalized by dividing Betweenness(p,N) with the maximum index in the network.

We assessed centrality with the STRING database (Szklarczyk et al., 2011), and considered two other databases for independent support (Human Protein Reference Database (HPRD) (Prasad et al., 2009) and I2D (Brown and Jurisica, 2007)). The top quartile of centrality indexes in the entire network of 18,583 proteins was significantly enriched with protein products of the 397 genes with rearrangements in more than one sample (p = 0.002).

For rearranged genes that scored highly in the centrality analysis, we assessed gene expression levels in the subset of transcriptome-sequenced samples using RSEQtools (Habegger et al., 2011) (Table S4). To evaluate the effects of the rearrangements on gene transcription, we noted genes that were expressed in the bottom or top tenth percentiles in samples harboring rearrangement of the locus compared to all other tumors.

Detection of Chained Rearrangements and Deletions

Overview of the ChainFinder Algorithm

ChainFinder analyzes somatic DNA rearrangements from WGS data (e.g., deletions, inversions or translocations) and infers whether the rearrangement likely occurred in the context of a “chain” with two or more other rearrangements. Chained rearrangements are identified by searching for sets of breakpoints that are distributed about the genome in a configuration that would be improbable if the rearrangements had occurred independently of one another. The ability to detect chains is enhanced by also considering copy number profiles for signatures of chained rearrangements.

ChainFinder is implemented in MATLAB, and formulates the detection of rearrangement chains as a graph theory problem, in which breakpoints are treated as nodes that may be inter-connected by graph edges (Figures 3A and S3A). Edges connect pairs of breakpoints that are either (1) somatically fused to each other (2) involved in two distinct rearrangements that are unlikely to have arisen independently or (3) at either end of a deletion bridge. An initial graph is constructed by searching for sets of breakpoints and associated deletions for which the independent model can be rejected after correction for multiple hypothesis testing (see below). The initial graph is then refined by considering any alternative valid assignments of breakpoints and deletion segments to deletion bridges. In the final graph, breakpoints connected by edges correspond to collections of rearrangements that may have arisen concertedly in the context of a chain. These steps are described in detail in the following sections and diagramed in Figure S3A.

Assessment of Adjacent Breakpoints

Each pair of breakpoints joined by a somatic DNA fusion is first connected by an edge on the graph (Figure S3A). For each pair of neighboring breakpoints on the reference genome within 1Mb of each other, the probability of two breakpoints arising independently within the observed distance of one another (P_{XY}) is calculated as follows. We assume that the probability of a DNA breakage event per nucleotide is uniform near the breakpoint and equal to μ_{local}. The probability of a second event *not* occurring within a distance *L* from the reference event (either upstream or downstream) is $(1\text{-}2\mu_{local})^{L}$. Therefore, the probability P_{XY} of observing a second breakpoint *Y* within distance L of an index breakpoint *X* is:

$$P_{XY} = 1 - (1 - 2\mu_{local})^{L}$$

The rate μ_{local} is calculated based on (1) the number of breakpoints per base-pair observed in a given tumor (μ_{global}) and (2) the density of breaks near the rearranged locus across the panel of 57 prostate tumors (ρ):

$$\mu_{local} = \mu_{global}\,\rho$$

We estimate the breakpoint density ρ as a function of genomic location by dividing the genome into 1Mb windows and counting the number of tumors with one or more breaks within a given window (Figure S2B). Values of ρ are scaled uniformly such that the sum of μ_{local} across all windows is equal to μ_{global}.

For neighboring breakpoint pairs, P_{XY} is considered as a p-value for the hypothesis that the two breakpoints arose independently. Pairs of breakpoints are connected by an edge (assigned to the same chain) if the corresponding P_{XY} can be rejected with control of the false discovery rate at 10^{-2} (Benjamini, 1995).

Assignment of Deletion Bridges

Next, segmented copy number data are overlaid with breakpoint locations to identify rearrangement breakpoints that correspond to deletion events. This step connects breakpoints on the graph with edges corresponding to deletion bridges in cases where the breakpoints may have originated from the same DNA deletion event.

Each breakpoint is provisionally paired to a boundary of a deletion segment if the breakpoint lies within 8 SNP probes of the boundary (typically a span of several thousand base-pairs). Breakpoints at either boundary of a deletion segment are potentially joined by a deletion bridge if:

A. The breakpoints on either end of the deletion are not fused to each other; i.e., the deletion must correspond to a deletion bridge (involving two rearrangements) rather than a “simple deletion” (involving one rearrangement) (Figure 2A).
B. The sequencing reads supporting the breakpoints at either end of the deleted segment must “point towards” the deletion, such that the deleted sequence would lie directly downstream of the reads.

Edges are added to the graph to denote potential deletion bridges. In cases where pairs of breakpoints cannot be uniquely assigned to a single bridge, multiple interpretations are tested in a subsequent step (see Finalization of the Graph, below)

Evaluation of Graph Cycles

In some cases, P_{XY} is extremely small – for instance, when breakpoints from separate fusions map within several hundred base pairs of one another – and the breakpoints clearly did not originate independently. However, borderline cases often arise where P_{XY} is not sufficiently small to reject the independent model for two breakpoints unequivocally. In such cases, additional evidence that rearrangements were generated coordinately can be obtained by considering sets of breakpoints whose nodes on the graph are contained within cycles (paths along edges that begin and end at the same node).

Each cycle is evaluated under the independent breakpoint model based on P_{XY} values for adjacent breakpoints within the cycle (Figure S3A). Specifically, all possible scenarios are considered by which one or more rearrangements within the cycle could have arisen independently. For example, three rearrangements involving six breakpoints in a hypothetical cycle (Figure S3B) could have arisen by the following (non-mutually exclusive) scenarios, where subscripted numbers in parentheses denoted rearrangements that occurred independently:

$$\{H_{(1)23}, H_{(2)13}, H_{(3)12}, H_{(1)(2)(3)}\}$$

This set of scenarios represents the independent model for the cycle, which encompasses all alternative possibilities to the breakpoints in the cycle arising coordinately (H_{123}).

ChainFinder considers the probability of detecting the independently generated breakpoints under each scenario within the observed distance of each other. Each scenario in the independent model requires that *two or more pairs* of adjacent breakpoints from separate rearrangements arise independently, in order to “split” the cycle into two or more separate events. Each such scenario can be expressed in terms of combinations of P_{XY} values from edges within the cycle corresponding to adjacent breakpoints (i.e., P_{ab}, P_{cd} and P_{ef}; Figure S3B)

$$\{(P_{ab}\,P_{cd}), (P_{ab}\,P_{ef}), (P_{cd}\,P_{ef}), (P_{ab}\,P_{cd}\,P_{ef})\}$$

As shown in Figure S3B, all scenarios involving three or more independent events require the co-occurrence of two or more scenarios involving only two events. Therefore, rejecting all scenarios involving only two events is sufficient to reject the independent model overall for the cycle. To assess all scenarios involving two independent events, ChainFinder tests the pairwise products of all P_{XY} values within a cycle (corresponding to all two-event scenarios) with control of the family-wise error rate (FWER; (Holm, 1979)) at 10^{-2} across all scenarios. Control of the FWER ensures that, if the independent model is rejected for a cycle, there is a 1% chance that one or more of the independent rearrangement scenarios for the corresponding cycle were mistakenly rejected. All cycles for which the independent model is rejected are linked within a chain.

Finalization of the Graph

Finally, the graph is refined by considering deletion bridges that could not be uniquely assigned. Although a single deletion bridge may exist that connects two breakpoints, frequently multiple interpretations are possible due to overlapping regions of deletion from separate alleles or distinct tumor subclones. In these cases, a single choice must be made from a set of mutually exclusive possible bridges. Bridges are mutually permissible only if the following conditions are met:

1. The bridges do not share the same breakpoints at either deletion segment boundary
2. If the bridges overlap, the deletion segment in the region of overlap must demonstrate a consistently lower copy number than segments outside the region of overlap.

ChainFinder tests permutations of mutually permissible bridges to find the combination that incorporates the most breakpoints into deletion bridges, because this solution best reconciles the copy number and rearrangement data. If a unique valid combination of bridges exists that maximizes the number of breakpoints in deletion bridges, the bridges are accepted and any distinct chains that they link are combined. If multiple optimal interpretations exist, only bridges that are included in all of these interpretations are kept.

The graph is finalized by removing any edge between neighboring breakpoints for which the independent-generation model could not be rejected. In addition, deletion bridge edges are retained in the graph only if the breakpoints on either end of the deletion arose non-independently (e.g., within a cycle).

Evaluation of Genes Disrupted in Chains

Once chains have been assigned, a list of genes disrupted in each chain is compiled. Genes are included if they fall at least partially within a deletion bridge in the chain or within 10kb of a copy-neutral rearrangement in the chain. Circos plots are generated in which all rearrangements in a given chain are depicted in the same color (e.g., Figure 2B).

Assessment of False-Positive Rate with Simulated Tumor Genomes

In order to test the false-positive rate of ChainFinder, we created “scrambled” tumor genomes by simulating the independent accumulation of rearrangements based on observed data. For each tumor, ten “scrambles” were created that combined rearrangements from other tumors. Each scramble contained the same number of rearrangements as the corresponding sequenced tumor. Any two rearrangements were combined in a scramble only if they were not part of the same chain from the same sequenced tumor. The scrambles served as “true negative” cases in which all rearrangements were generated independently, while preserving genome-position specific influences on breakage and fusion, since the data are drawn from observed rearrangements. Copy number profiles were simulated based on observed data as well. Segments of copy number alteration were generated that maintained (1) the number of breakpoints at the boundaries of potential deletion bridges and (2) the overall ratio of copy number gains to losses. The simulated rearrangement and copy number data were profiled with ChainFinder, and the proportion of breakpoints assigned to a chain was compared between observed and simulated data.

For each sequenced tumor, we also created ten simulations matched for rearrangement number and chromosomal connectivity. The rearrangement breakpoints were further matched to observed data with respect to (1) sequence coverage, (2) guanine and cytosine content of local sequence, (3) expression levels of nearby genes, (4) replication timing of DNA and (5) reference genome distance between breakpoints for intrachromosomal rearrangements (within 5%). Coverage was matched within 5x to the coverage near the observed breakpoint. Suitable locations for simulated breakpoints were identified by creating bins for the values of parameters (2) through (5) for each chromosome, and randomly choosing a location that falls within the same bin as the corresponding observed breakpoint. For each feature (e.g., GC content), we created bins containing the bottom and top fifth percentiles across the chromosome. We then split the middle 90% evenly into three additional evenly spaced bins. Copy number profiles were simulated such that breakpoints at edges of deletion segments were preserved. In most cases where ChainFinder identified chains within simulated tumors, the simulations were too restrictive, so that the only matched location for a set of rearrangements in a chain was near to the location where they were observed.

Quantification of Gene Expression near Rearrangement Breakpoints

Expression was quantified in terms of gene-level FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values from 16 prostate tumor transcriptomes using CuffLinks (Trapnell et al., 2012). The transcription levels near rearrangements were estimated from median values of $\log_{10}$ (1+FPKM) across the tumor transcriptomes in 10kb windows on either side of the breakpoint. Where this window overlapped multiple genes, the largest FPKM value was used. For the analysis shown in Figure 4D, the statistical enrichment of chained breakpoints near highly expressed DNA in ETS-positive tumors was robust to exclusion of the *TMPRSS2* and *ERG* loci from expression level estimates.

Assessment of Nuclear Proximity of Fused Loci from Hi-C Data

We sought to determine whether breakpoints involved in structural rearrangements are in close physical proximity in nuclei in which these breakpoints have not yet occurred. For this, we used filtered chromatin interaction data (Hi-C) from experiments performed in prostate epithelial cells (RWPE1) stably expressing a GFP reporter (RWPE1-GFP) (Rickman et al., 2012). To determine whether a set of breakpoint pairs are in close proximity, we defined a 1Mb window centered on each breakpoint and counted the Hi-C reads connecting the two windows for all breakpoint pairs. The average Hi-C read counts were determined separately for chained rearrangement breakpoints and for breakpoints that were not assigned to a chain for comparison. Rearrangements were further subdivided by ETS-status of the tumor in which they were observed.

We then compared the observed average Hi-C count to Hi-C counts that would be observed by chance if the breakpoint pairs were randomly distributed on the genome. We generated random sets of breakpoints matched to the observed breakpoints for intra-chromosomal distances, chromosomal distribution and short read mappability. We again defined 1Mb windows centered in the random breakpoints and counted Hi-C reads connecting each pair of simulated breakpoints. We repeated this analysis 1,000 times to generate a null distribution of average Hi-C read counts for random breakpoint pairs. To generate a p-value, we counted how many of the 1,000 sets of random breakpoints had an average Hi-C read count greater than or equal to the average read count for the observed breakpoints. Of note, only intra-chromosomal rearrangements were considered for this analysis, as inter-chromosomal breakpoints were supported by very few Hi-C reads even when considering large windows centered on the breakpoints.

Breakpoint Enrichment Analysis

Enrichment and depletion of breakpoints was assessed across the genome with respect to replication time, guanine/cytosine (GC) content and distance to transcribed genes. Observed distributions were compared to randomly generated distributions controlled for chromosome and coverage. First, nearby breakpoints (up to 2,500bp away) were consolidated into a single “event.” For each event, 100,000 locations (one per iteration) were generated uniformly from all locations on the same chromosome having the same coverage. The genome was considered in the following bins: low GC (0-36%), medium GC (36%-45%) and high GC (45%-100%). Replication time was binned according to late/early ratio (Ryba et al., 2010) at $(-\infty,-0.8),[-0.8,0), [0,0.8),[0.8,\infty)$. Changing the thresholds did not affect the essence of the results, other than losing sensitivity for very large or small bins (data not shown). For every bin we counted the number of breakpoints for both the observed breakpoints and the random breakpoints. All of these counts were used to compute nonparametric p-values (observed rates). Enrichment or depletion was determined by picking the lower of the one-sided p-values, and p-values were then corrected for multiple hypotheses by the Benjamini-Hochberg FDR procedure (Benjamini, 1995).

An analogous procedure was used to detect enrichment near ChIP-Seq peaks (Yu et al., 2010). Here, bins were defined as within 50kb of a peak (“near”) or >100kb from a peak (“far”; see Figure S4).

Quantification of Tumor Purity and Subclonality

Prior to sequencing, estimates of tumor purity and ploidy were derived from Affymetrix SNP6.0 data using ABSOLUTE (Carter et al.). These estimates were used to select high purity samples for whole genome sequencing (median purity 70%; ploidy range 1.84 – 2.21; Table S1A).

Analyses of tumor purity and subclonality from WGS data were performed by exploiting individuals' genotypes at polymorphic loci within somatically altered regions of the genome, using considerations related to previously described methods (Carter et al., 2012; Nik-Zainal et al., 2012; Landau et al., 2013). For each tumor sample included in the study, we estimated stromal DNA admixture and lesion clonality using CLONET (CLONality Estimate in Tumors; Prandi D. et al., manuscript in preparation). The approach behind CLONET and the MiSeq-based validation we performed are outlined hereafter.

For a tumor sample *TS* containing a mixture of N_{TS} normal (diploid) cells and T_{TS} tumor cells, the percentage of admixed normal cells is:

$$Adm(TS) = \frac{N_{TS}}{N_{TS} + T_{TS}}$$

Based on the above equation, the admixture can be estimated from sequencing reads covering a site of hemizygous deletion *s* as:

$$Adm_s(TS) = \frac{\beta_s(TS)}{2 - \beta_s(TS)}$$

where $\beta_s(TS)$ is the proportion of reads at locus *s* that originated from normal cells in TS.

In order to calculate *Adm*(*TS*), we first selected informative heterozygous SNPs within regions of somatic deletion that were identified from copy number array data. For each hemizygous deletion *H*, we considered the distribution of the allelic fractions (i.e., the fraction of reference sequence reads) from selected SNPs within *H*. Using "particle swarm optimization" (Kennedy, 1995) we calculated a composite value $\beta_H(TS)$ for the deleted region that best accounted for the observed distribution of allelic fractions at each heterozygous SNP *s* across the region. For every deletion, a value $Adm.apparent_H(TS)$ was computed that describes the apparent admixture at that locus. $Adm.apparent_H(TS)$ reflects both stromal admixture and potentially subclonal lesions that increase the apparent level of normal DNA at that locus. For each hemizygous somatic deletion *H,* the values of $Adm.apparent_H(TS)$ were grouped if the difference between the values could be explained by the estimation error determined by simulation-based error estimations. The smallest mean value of $Adm.apparent_H(TS)$ across a set of grouped deletions was taken as the candidate value of *Adm*(*TS*).

Estimates of cancer DNA purity by this procedure are listed in Table S1 and compared to estimates from the same tumors by ABSOLUTE run on SNP array data. The estimates were highly consistent across the samples ($R^2 = 0.99$; $p < 10^{-4}$) with the exception of two samples (PR-STID0000002682 and PR-07-360), where stromal admixture was detected in WGS data but not SNP array data.

We analyzed the clonality of gene deletions based on normalized $\log_2$ ratios of tumor and normal WGS sequence coverage, after correction for the estimated normal DNA admixture in tumor samples. Deletions where *Adm* and $Adm.apparent_H$ differ beyond the error estimation are potential sub-clonal lesions. We estimated the percentage of tumor cells that harbor a somatic hemizygous deletion *H*, i.e., the clonality of *H*, as:

$$Clonality_H(TS) = \frac{1 - Adm.apparent_H(TS)}{1 - Adm(TS)}$$

In the case of a 100% clonal hemizygous somatic deletion *H* the value of $Clonality_H(TS)$ is 1, as $Adm.apparent_H(TS)$ equals *Adm*(TS); otherwise $Clonality_H(TS)$ is less than 1. In the presence of high coverage, small variations in $\text{Clonality}_H(\text{TS})$ can demonstrate differences in sub-clonality along a continuous scale. Here, in order to avoid false positive calls for borderline subclonality, we adopted a conservative approach and only considered two classes of deletions: clonal ($Clonality_H(TS) \geq 0.8$) and subclonal ($Clonality_H(TS) < 0.8$). After *Adm* was calculated,we executed a similar procedure to estimate the clonal status of somatic homozygous deletions and point mutations.

The sensitivity of clonality detection depends upon the number of heterozygous SNPs within a deletion of interest and the depth of sequence coverage at these SNPs. We evaluated the uncertainty in clonality estimates as a function of these parameters by randomly sampling 1,800 simulations and averaging the difference between the true clonality and computed clonality for a given coverage and number of SNPs (Table S6). To ensure robust clonality calls, we considered only deletions with 20 or more informative SNPs with average sequence coverage of 20x (corresponding to a 5.4% estimation error). Table S7 lists the percentage of tumor cells found to harbor a specific lesion together with the associated uncertainty range.

Validation of Clonality Estimates

To assess our ability to estimate apparent DNA admixture from our WGS data, we generated independent validation data for a set of 18 aberrant genes with four heterozygous SNPs each from seven tumor samples by PCR and deep sequencing (>65,000x coverage). The deep coverage provided a precise estimate of the ratio of alleles at SNP sites. Figure S6A compares the local apparent DNA admixture for the 18 genes computed using WGS data to the estimates computed using deep sequencing data ($R^2 = 0.85$, p-value = $3.55x10^{-8}$). The contingency table inset in Figure S6A demonstrates agreement between WGS- and deep sequencing-based calls of clonality status (Cochran test, p-value = 1). Importantly, these results show that the conservative approach taken in this study to call clonality in two states (clonal or sub-clonal) is robust in the presence of poorly defined lesion breakpoints, as in the case of a hemizygous deletion of *PTEN* in sample P08-688 (outlier, apparent DNA admixture ~0.6, correctly classified as sub-clonal).

SUPPLEMENTAL REFERENCES

Brown, K.R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. *8*, R95.

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

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.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491-498.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L.*, et al.* (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. *12*, R1.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. *39*, D945-950.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res. *39*, D876-882.

Griffith, O.L., Montgomery, S.B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M.*, et al.* (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. *36*, D107-113.

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.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat 6, 65-70.

Kennedy, J.E., R. (1995). Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks *IV*, 1942-1948.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L.*, et al.* (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714-726.

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol. Biol. *577*, 67-79.

Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V.*, et al.* (2012). Oncogene-mediated alterations in chromatin conformation. Proc. Natl. Acad. Sci. USA *109*, 9083-9088.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-26.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761-770.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308-311.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A.*, et al.* (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157-1160.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P.*, et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. *39*, D561-568.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562-578.

The Uniprot Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. *39*, D214-219.

Yu, J., Mani, R.S., Cao, Q., Brenner, C.J., Cao, X., Wang, X., Wu, L., Li, J., Hu, M., Gong, Y.*, et al.* (2010). An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell *17*, 443-454.