Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations.

Supplemental Material

Phillippa C. Taberlay[1,2,#], Joanna Achinger-Kawecka[1,2,#], Aaron T.L. Lun[4,5], Fabian A. Buske[1], Kenneth Sabir[1], Cathryn M. Gould[1], Elena Zotenko[1,2], Saul A. Bert[1], Katherine A. Giles[1], Denis C. Bauer[3], Gordon K. Smyth[4,6], Clare Stirzaker[1,2], Sean I. O'Donoghue[1,3], Susan J. Clark[1,2,*]

* Corresponding author

# Equal contributors

Correspondence should be addressed to: Susan J. Clark (s.clark@garvan.org.au)

1. Epigenetics Research Laboratory, Genomics and Epigenetics Division, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010, Australia
2. St. Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Darlinghurst, NSW 2010, Australia
3. CSIRO, 11 Julius Avenue, North Ryde, New South Wales 2113, Australia
4. Bioinformatics Division, Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria 3052, Australia
5. Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia
6. Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia

March 30, 2016

TABLE OF CONTENTS

SUPPLEMENTAL METHODS

Prostate Cell Lines

LNCaP cell line was obtained from ATCC (LNCaP clone FGC, ATCC® CRL-1740™). LNCaP cell line has been fully characterized by (Horoszewicz et al. 1983). The PC3 cell line was obtained from ATCC (PC-3, ATCC® CRL-1435™). Full characteristics of the PC3 cell line are included in (Kaighn et al. 1979).

Hi-C chromosome conformation capture

Hi-C experiments were performed based on the original protocol by Lieberman-Aiden et al. (Lieberman-Aiden et al. 2009) with minor modifications. Single cells (10-50x10^6 total) were collected and fixed with a final concentration of 1% formaldehyde for 10 mins at room temperature. Reactions were quenched with glycine and incubated on ice for 15 mins. Cells were centrifuged for 3 mins at 500g then washed in ice-cold PBS followed by an additional centrifugation. Nuclei were extracted by incubation in 1mL ice-cold Nuclei Buffer (10mM Tris, pH 7.4, 10mM NaCl, 3mM MgCl2, 0.1mM EDTA and 0.5% NP-40, plus protease inhibitors) per 5x10^6 cells for at least 60 mins on ice. Nuclei were collected by centrifugation at 4°C for 5 mins and 500g then washed twice in 1× NEBuffer3 (New England Biolabs). Nuclei were resuspended in 1× NEBuffer3 supplemented with 10% SDS then incubated at 65°C for exactly 10 mins and transferred immediately to ice before addition of Triton X-100. Chromatin was digested overnight with 400U BglII at 37°C. Ends were repaired and marked with biotin-14-dCTP using Klenow DNA polymerase at room temperature for 20 mins. Enzymes were inactivated by addition of 10% SDS and incubation at 65°C for 30 mins. Dilute ligations were performed in a final volume of 8mL using 250µL Blunt/TA Ligase Mastermix (New England Biolabs) supplemented with 10× Ligation Buffer (New England Biolabs), 10% Triton X-100, 10mg/mL BSA and 100mM ATP. Ligations were performed at room temperature for 4 hours prior to Proteinase K treatment overnight at 65°C. DNA was extracted twice by Tris-EDTA saturated phenol (pH 8.0) then precipitated with 3M sodium acetate and ethanol overnight at -20°C. The DNA pellet was dissolved in Tris-EDTA Buffer and purified twice with Phenol:Chloroform:Isoamyl Alcohol 25:24:1 saturated with 10mM Tris, pH 8.0, 1 mM EDTA. After the second extraction, DNA was precipitated with 3M sodium acetate and 100% ethanol overnight at -20°C. DNA was collected by centrifugation at 18,000g for 30 mins at 4°C, dissolved in Tris-EDTA and RNaseA treated for 15 mins at 37°C. Hi-C material (10µg) was treated with T4 DNA polymerase in Buffer 2 (New England

Biolabs) supplemented with dATP, dGTP and BSA for 2 hours at 12°C to remove biotin-14-dCTP from non-ligated ends. Reactions were quenched by addition of 0.5M EDTA, then DNA was purified once using Phenol:Chloroform:Isoamyl Alcohol 25:24:1 saturated with 10mM Tris, pH 8.0, 1mM EDTA followed by ethanol precipitation at -20°C. DNA was resuspended in a final volume of 100µL nuclease-free water.

Preparation of Hi-C libraries

Hi-C libraries were prepared using a customized protocol. Details are provided in the Supplmental Methods. Hi-C material was sonicated using a Covaris instrument to an average molecular weight of 300-500bp. Achievement of the desired size range was verified by agarose gel electrophoresis. Fragmented DNA was repaired and blunt ends were dA-tailed using the NEBNext DNA Library Prep Master Mix Set for Illumina (NEB# E6040L) according to the manufacturers' instructions. A size selection was performed using AMPureXP Beads (Beckman Coulter Inc.). Biotin-tagged DNA was bound to MyOne Streptavidin C1 beads (Invitrogen, #65601) using 2× Binding Buffer (10mM Tris-HCl pH 8.0, 1mM EDTA, 2M NaCl) for 20min at room temperature with rotation. Biotin-tagged DNA coupled with MyOne Streptavidin C1 beads was isolated using a magnetic particle concentrator. Beads were washed twice with 200uL 1X Binding Buffer and once with 200uL 1× Tween Wash Buffer (5mM Tris-HCl pH 8.0, 0.5mM EDTA, 1M NaCL, 0.05% Tween). Beads were resuspended in a final volume of 65uL of water and adapters were ligated to DNA ends using the NEBNext Ultra DNA Library Prep kit (NEB# E7370L). PCR enrichment was performed using DNA bound to the MyOne Streptavidin C1 beads and NEBNext Multiplex Oligos for Illumina (Set 1, NEB#E7335) using the NEBNext Ultra DNA Library Prep kit with 8-14 cycles for library amplification. PCR products were purified using 1× volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted in 50µL Tris-EDTA Buffer. The Hi-C libraries were quantified using the KAPA Library Quantification Kit for Illumina platforms (KAPA Biosystems) and qualified using the Bioanalyzer 2100 (Agilent Technologies). Optimal concentrations to get the right cluster density were determined empirically. Resulting libraries were run on the HiSeq 2500s (Illumina) platform configured for 100bp paired-end reads according to manufacturer's instructions.

Reference genome and datasets used in this publication

The human reference genome used throughout was hg19. Realigning the reads to GRCh38 would not significantly alter our conclusions and would constrain the

functional interpretation of our results, as large-scale public data sets are not available in the new GRCh38 genome assembly. Information on datasets used in this publication are included in Supplemental Table S13.

Normalization of Hi-C data

All HiC libraries were processed through the NGSane framework v0.5.2 (Buske et al. 2014) available from Github using the "fastqc", "hicup" and "fithicaggregate" modules as follows: First, quality check of sequence libraries was performed with FastQC v0.11.2 . Raw fastq files were then pre-processed, mapped with bowtie v1.1.0 (Langmead et al. 2009) and assessed for artifact levels through HiCuP v0.5.2 supplying genome assembly (hg19) and the BglII restriction enzyme cut site. Aligned read files in BAM format were sorted with Samtools v1.2 (Li et al. 2009) and duplicates were tagged using MarkDuplicates from Picard tools v1.121. Replicates were pooled using bespoke Python scripts (provided within NGSane) leveraging the sparse matrices formats in the SciPy libraries (Jones et al.). Significant connections were assessed from contact count matrices for multiple resolution (100kb and 1Mb) using a custom adaptation of fit-hi-c (Ay et al. 2014; Libbrecht et al. 2015) (provided within NGSane) supplying iteratively corrected bias offsets calculated through HiCorrector v1.1 (Li et al. 2015) as well as genome mappability tracks from ENCODE . Significant contacts with false discovery rate (FDR) less than 0.01 were imported into Rondo or the WashuU Epigenome Browser (Zhou et al. 2013) for visualization and further analysis.

Data visualisation

To visualise the segmentation of the interaction data into domains, we generated 2D heat maps at 100kb resolution and overlaid them with previously generated ChIP-seq tracks and Topologically Associated Domain (TAD) tracks generated as BED files. Interaction frequencies were calculated as previously described and visualized in The WashU Epigenome Browser (Zhou et al. 2013). Positive score thresholds were adjusted manually to normalise for sequencing depth difference between cell lines.

Interaction data imported into Rondo was processed as described previously and visualised at 100kb resolution, and transcription factor binding sites, histone modification peaks and RNA-seq signal were visualized as either bed files (ChIP-seq) or bam/bigwig files (RNA-seq). Rondo can be accessed at http://odonoghuelab.org:8020 and - upon publication – open source at http://github.org/ODonoghueLab/Rondo.

Identification of Topologically Associated Domains and domain boundaries

The topologically associated domains were identified using a pipeline called "domain-caller" developed by (Dixon et al. 2012) (reviewed in (Ay and Noble 2015)). Briefly, the domain-caller algorithm is based on the imbalance between the upstream and downstream contacts of a region that is created by TAD. This imbalance is an indicator of whether a region is in the topological domain, at the boundary, or far away from a TAD and it can be quantified in a statistic called directionality index (DI). Domain-caller algorithm uses Hidden Markov model (HMM) to determine the underlying bias state for each locus (upstream, downstream, none) and then these HMM calls are used to infer TADs as continuous stretches of downstream bias states followed by upstream bias states. A region in between two TADs is either called a "domain boundary" or "unorganized chromatin" depending on the region's length. We defined unorganized chromatin as regions that are at least 100kb from the topological domain and domain boundaries as regions that are less than 100kb from the topological domain as previously described (Dixon et al. 2012). For each cell type, a combined list of boundaries was generated and overlapping boundaries were trimmed and removed (Dixon et al. 2012). Similarly oriented boundaries within 100kb from each other that were present in three cell lines were considered to be constitutive domain boundaries and remaining boundaries were considered to be cell-type specific domain boundaries. The 100kb window was chosen to mimic the uncertainty of the domain boundary position due to the 40kb resolution of the domain calling.

Transcription factor, histone modification and transcription start site enrichment analysis

We used previously generated ChIP-seq data sets that were mapped using Bowtie (Langmead et al. 2009) (v0.12.8) to hg19 and peaks were called with Peak Ranger (v1.16) (Feng et al. 2011). To determine if domain boundaries were associated with a given factor (histone marks, CTCF and RAD21), we used ngsplot (v2.47.1) (Shen et al. 2014) and plotted the averaged data around the +/- 500kb region of the boundary (Figure 2). We used GAT (v1.0) (Heger et al. 2013) to determine the observed enrichment of CTCF and H3K4me3, as well as Transcription Start Sites (TSS). The observed over expected fold change and statistical significance was calculated (Figure 2d and 2e). Additionally, we defined the percentage of overlap between CTCF and H3K4me3 binding sites and domain boundaries by intersecting peaks identified from ChIP-seq data with the domain boundaries.

GO Term enrichment analysis

GO term enrichment analysis were performed by identifying differentially expressed genes at cancer-specific and normal-specific from RNA-seq data and looking for enriched GO terms (BP – Biological Processes, MF – Molecular Function and CC – Cellular Component) using DAVID database (Huang da et al. 2009; Wishart et al. 2009) (v 6.7). Supplemental Fig. S2 displays all non-redundant GO terms with a Benjamini corrected p-values of less than 10^{-4}.

Copy Number Variation analysis

Copy number estimates for LNCaP and PrEC cell lines were taken from published data (Robinson et al. 2010). Copy number data from Mapping 250k Sty arrays for PC3 and LNCaP cell lines was obtained from GEO (GSM827569) and processed as described previously (Bengtsson et al. 2008; Bengtsson et al. 2009). For all CNV datasets we used the liftOver tool from UCSC to lift the reads to the human genome hg19 version. To test for the association between CNVs and TAD boundaries, we overlapped the combined domain boundaries for each cell type with the locations of identified regions of copy number variation in each cell line. We only analysed CNVs up to 10Mb that were present in both LNCaP and PC3 cells. We considered CNVs within 40kb of the boundary to be associated.

Identification of differential interactions

Differential interactions between normal PrEC and cancer (LNCaP and PC3) cells were identified with Bioconductor diffHiC (v.1.0.1) R package (Lun and Smyth 2015). Paired-end reads were aligned to hg19 with Bowtie2 (Langmead and Salzberg 2012), low-abundance reads were filtered out and they resulting data was normalized for trended or CNV-driven biases. The statistical framework of the edgeR package to model the biological variability and to test for significance of identified differential genomic interactions.

Annotation of differential interactions to distinct chromatin states

Chromatin states were defined as previously described (Taberlay et al. 2014). To test for the association between differential chromatin interactions and chromatin states, we annotated the regulatory elements (CTCF, enhancer, enhancer + CTCF, promoter, promoter + CTCF, transcribed and repressed) to anchor points of differential

interactions at 100kb resolution and calculated the observed over expected enrichment using the GAT software (v1.0) (Heger et al. 2013).

Association of differential interactions with gene expression

To test for the association between differential chromatin interactions and gene expression, we annotated genomic locations of transcription starts sites of genes that are expressed in either of the two cell lines (PrEC or PC3; TPM $\geq$ 1) to anchor points of differential interactions at 100kb resolution (minimum required overlap $\geq$ 1bp). Gene expression levels between PrEC and PC3 were compared using unpaired t-test. Heat map was plotted using R package ggplot2.

TCGA data analysis

Gene expression analysis utilized clinical data available through the TCGA Prostate Adenocarcinoma (PRAD) cohort (The Cancer Genome Atlas Research Network 2015). Processed RNA-seq V2 data (level 3) were obtained from the TCGA data portal (normal samples=50, tumors=278). Patient outcome data was obtained from Taylor el al., 2010 through The Project Betastasis (www.betastasis.com) (Taylor et al. 2010).

RNA-seq

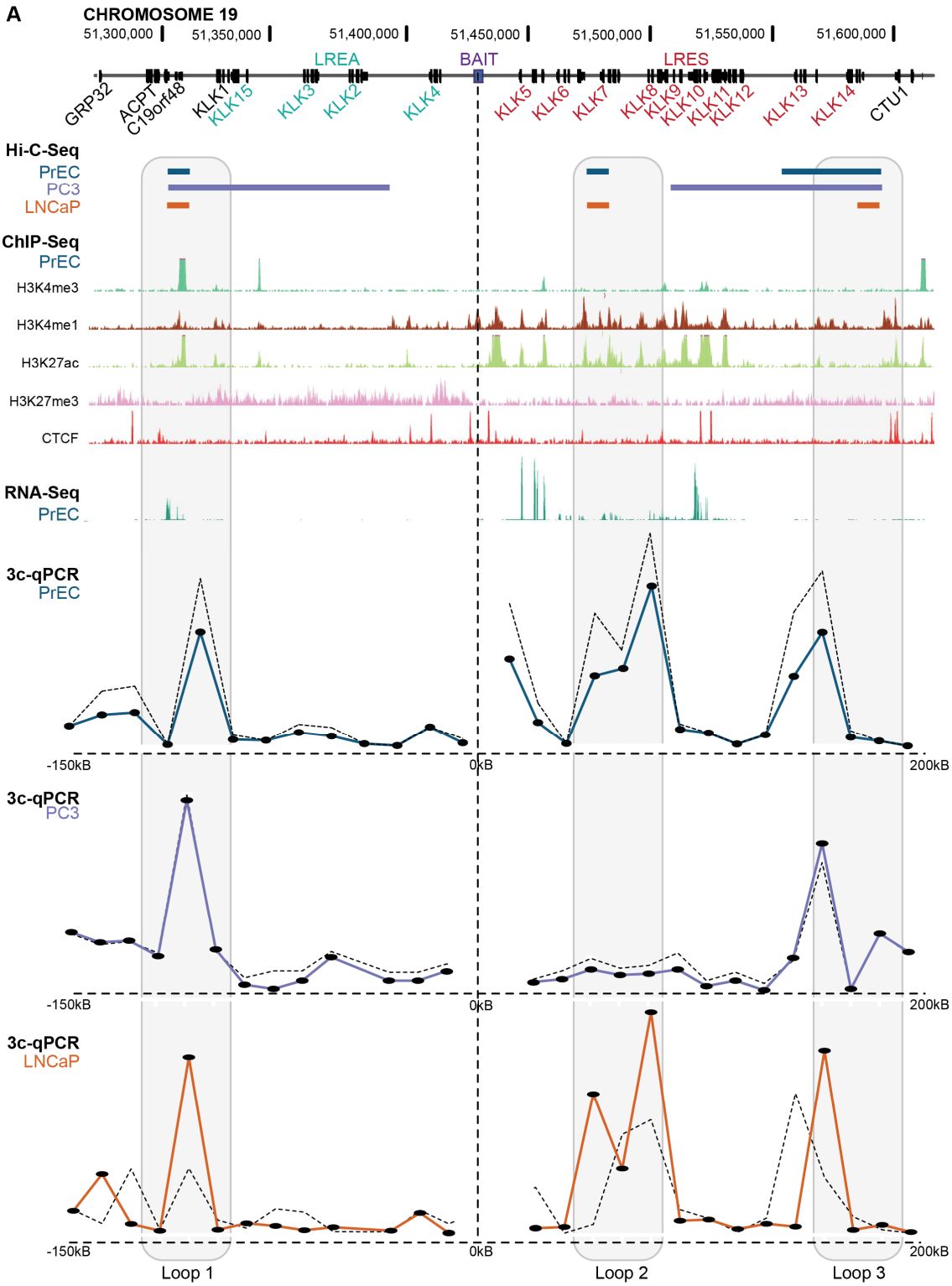
Total RNA was isolated with TRIzol reagent from exponentially growing LNCaP and PrEC cells that were 80% confluent and treated with DNaseI. Samples were quantified by Nanodrop and quality checked on the Bioanalyzer with the Agilent RNA 6000 Nano Kit. Total RNA (500ng) was spiked with external controls (ERCC RNA spike-in Mix, Thermo Fischer #4456740) and libraries were constructed with the Illumina TruSeq Stranded mRNA sample preparation kit. Paired-end reads (100bp) in biological triplicate were processed using Trim Galore (version 0.11.2) for adapter trimming (parameter settings: --fastqc --paired --retain_unpaired --length 16) and STAR (version 2.4.0j) (Dobin et al. 2013) for mapping reads to the hg19 human genome build with GENCODE 19 (Harrow et al. 2012) used as a reference transcriptome (parameter settings: --quantMode TranscriptomeSAM --outFilterMatchNmin 101). Paired-end reads for PC3 and PrEC RNA-seq data were downloaded from GEO (GSE25183) and mapped to the reference genome (hg19) using STAR (version 2.4.0j) (Dobin et al. 2013) as described previously. Mapped reads where counted into genes using featureCounts (Liao et al. 2014) program belonging to the Subread suite (version 1.4.6-p4) (Liao et al. 2013). Fold changes (FC) were computed as the log2 ratio of

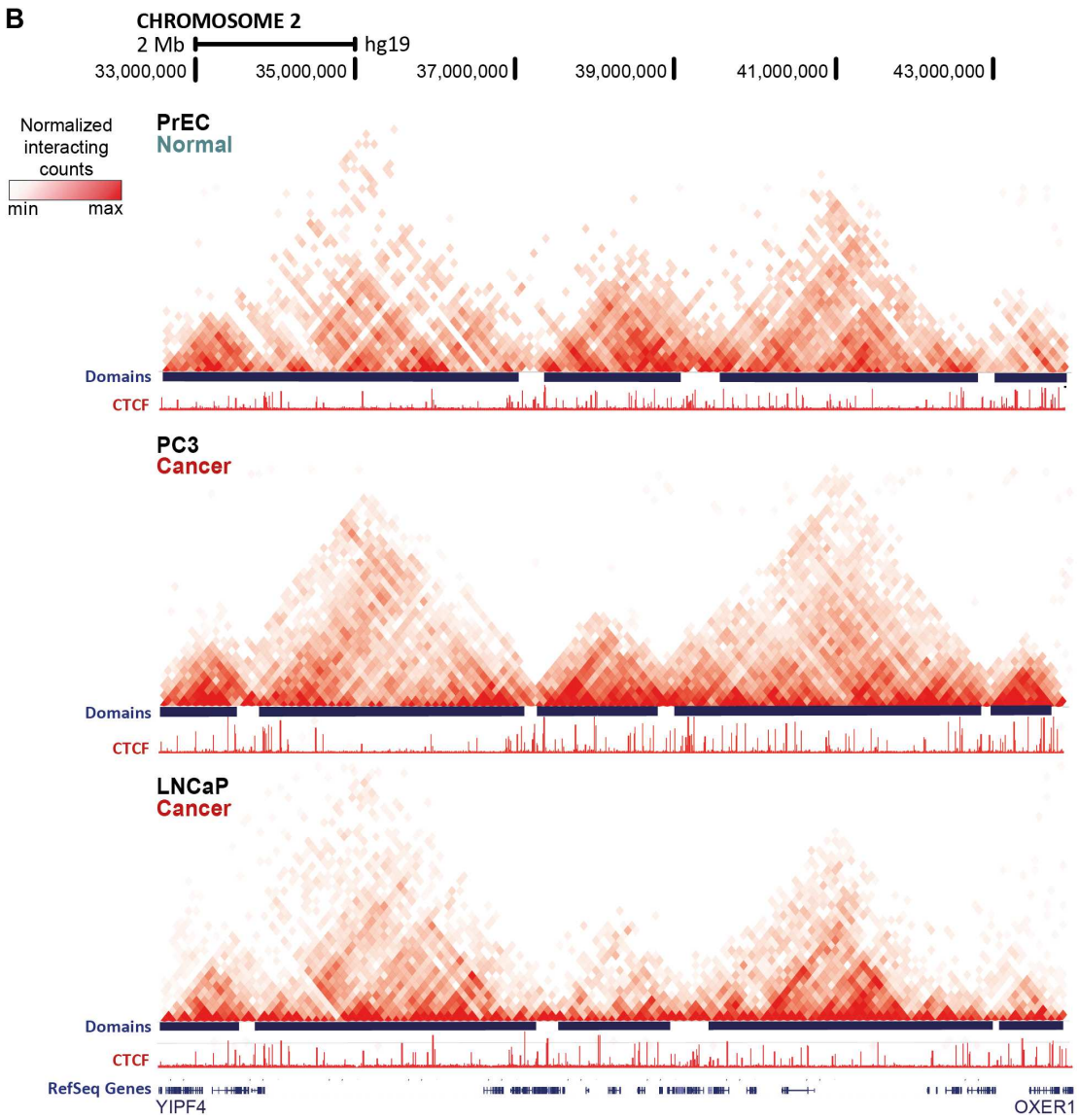
normalized reads per gene using the edgeR R package. Genes with fold change ± 1.5 ($P < 0.05$; FDR < 0.01) were considered as significantly altered.

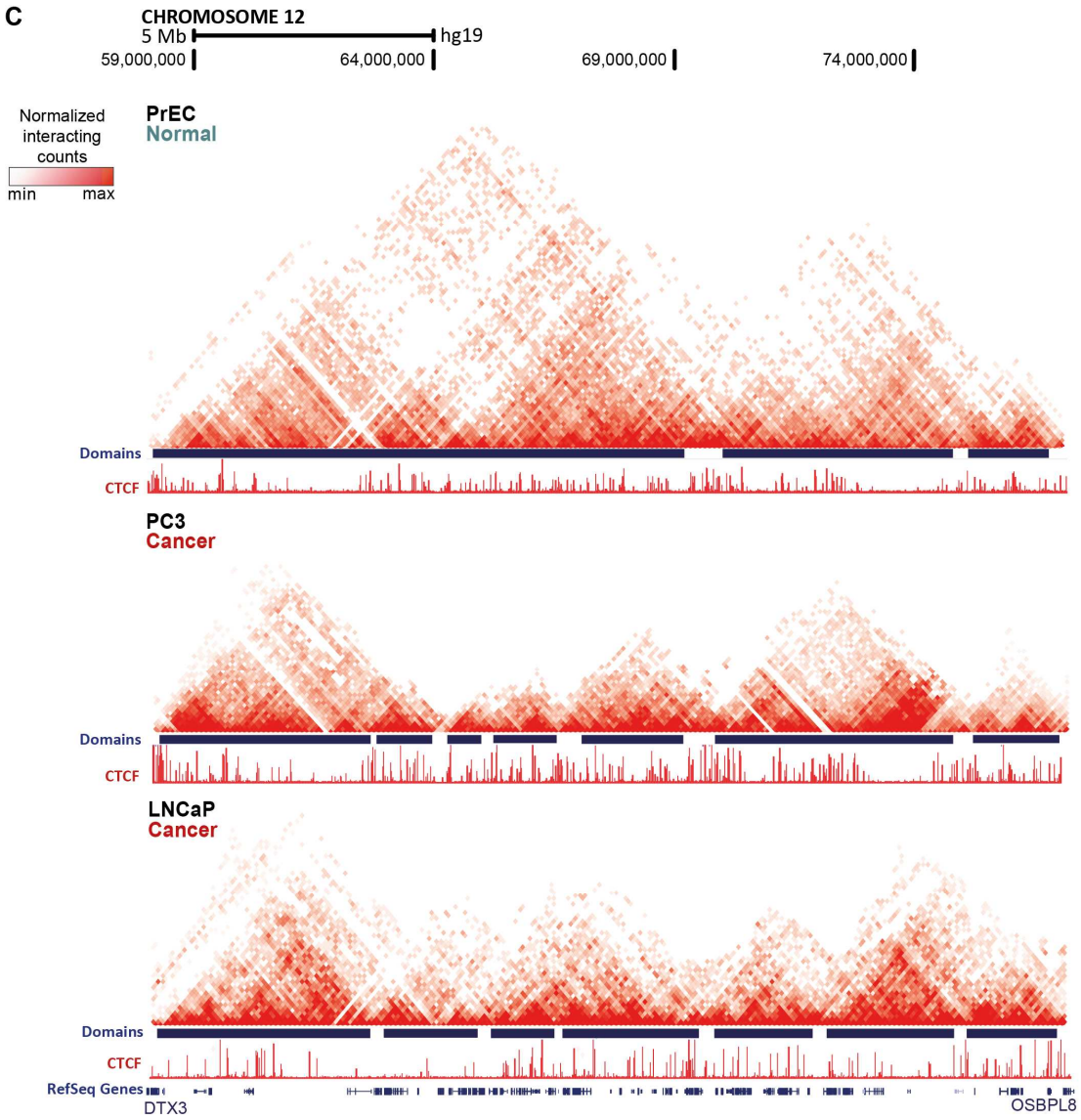
ChIP-seq

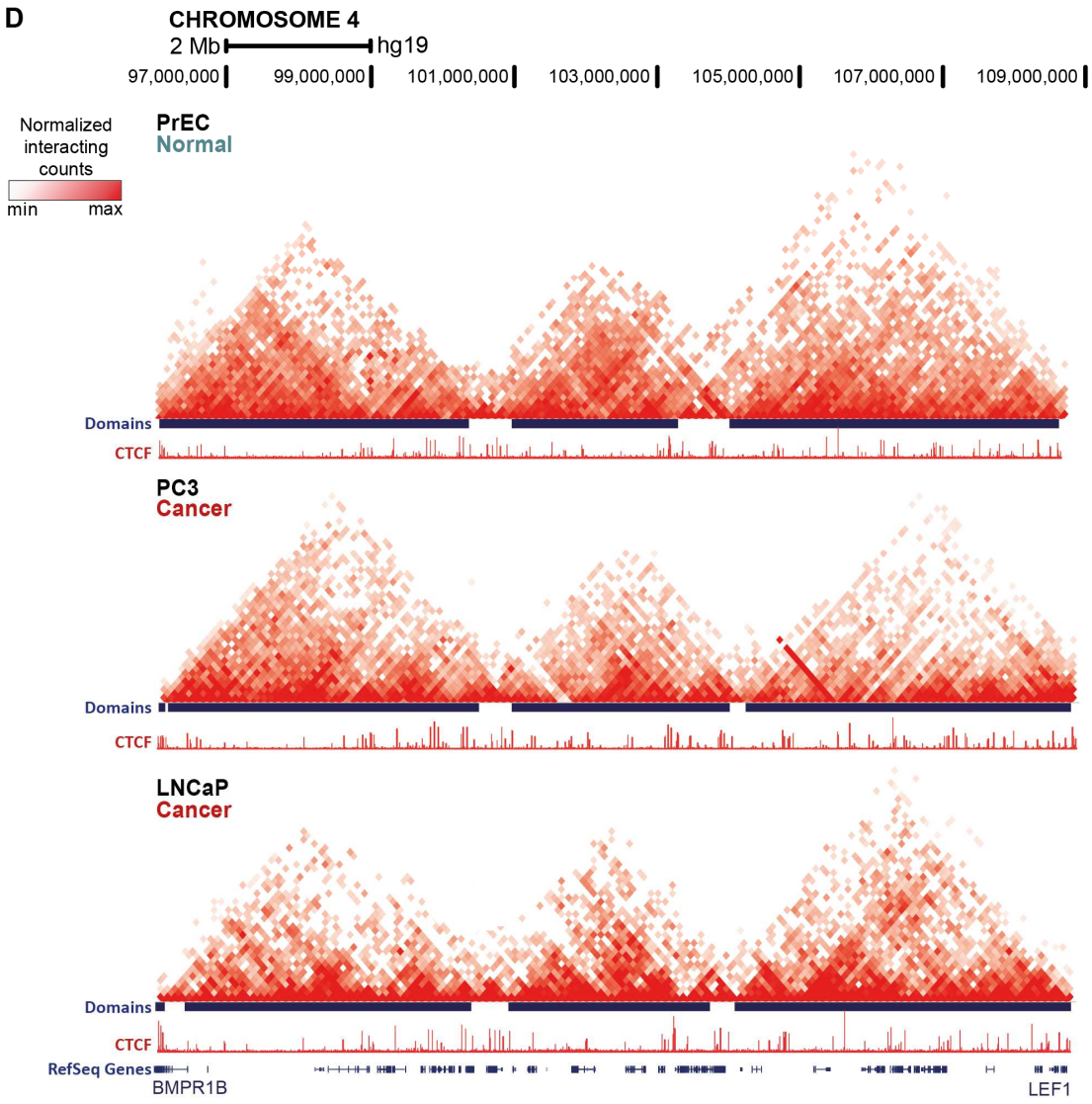
ChIP assays were performed as previously (Oakford et al. 2010; Taberlay et al. 2011; Taberlay et al. 2014). Briefly, nuclei were purified (described above for Hi-C) after formaldehyde crosslinking, collected, and resuspended in SDS Lysis buffer before sonication. Antibodies (10 μg) used for ChIP experiments included H3K4me1 (#39297, Active Motif), H3K27ac (#39133, Active Motif) and RAD21 (#ab992l, Abcam). Libraries for ChIP-seq were prepared following Illumina protocols. The resulting libraries were sequenced on the Illumina HiSeq 2000 platform configured for 50-bp single-end reads. Bowtie (Langmead et al. 2009) was used to align ChIP-seq reads to hg19 allowing up to three mismatches, discarding reads mapping to multiple positions in the genome and removing clonal reads. ChIP-seq broad peaks were called using ccat algorithm from Peak Ranger (v1.16) with following parameters: format: bam, read extension length: 200, FDR cut off: 0.11, sliding window size: 500, window moving step: 50, min window reads: 4 and min window fold change: 5 (Feng et al. 2011).

SUPPLEMENTAL FIGURES

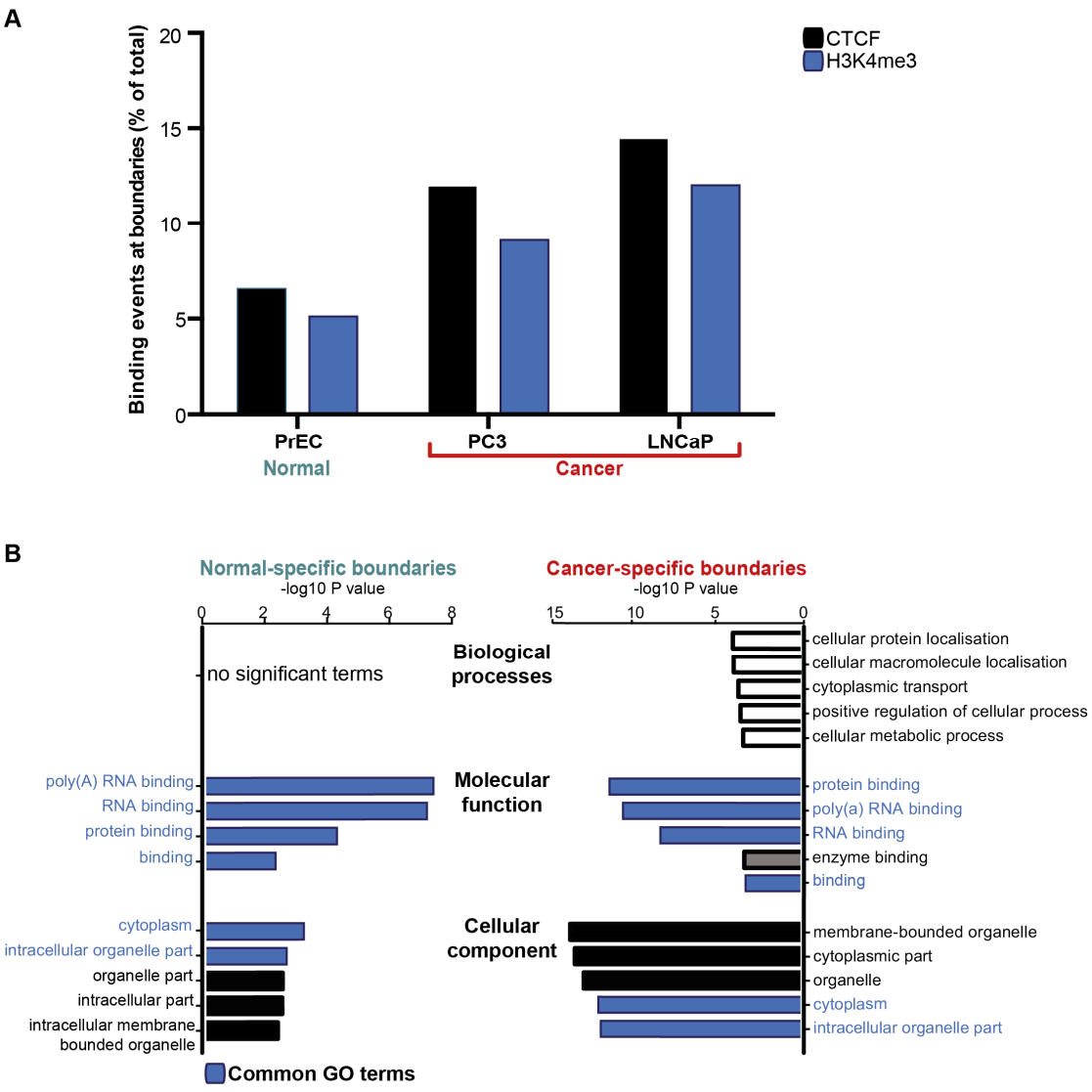



B
CHROMOSOME 2
2 Mb
hg19
33,000,000
35,000,000
37,000,000
39,000,000
41,000,000
43,000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
CTCF
PC3
Cancer
Domains
CTCF
LNCaP
Cancer
Domains
CTCF
RefSeq Genes
YIPF4
OXER1

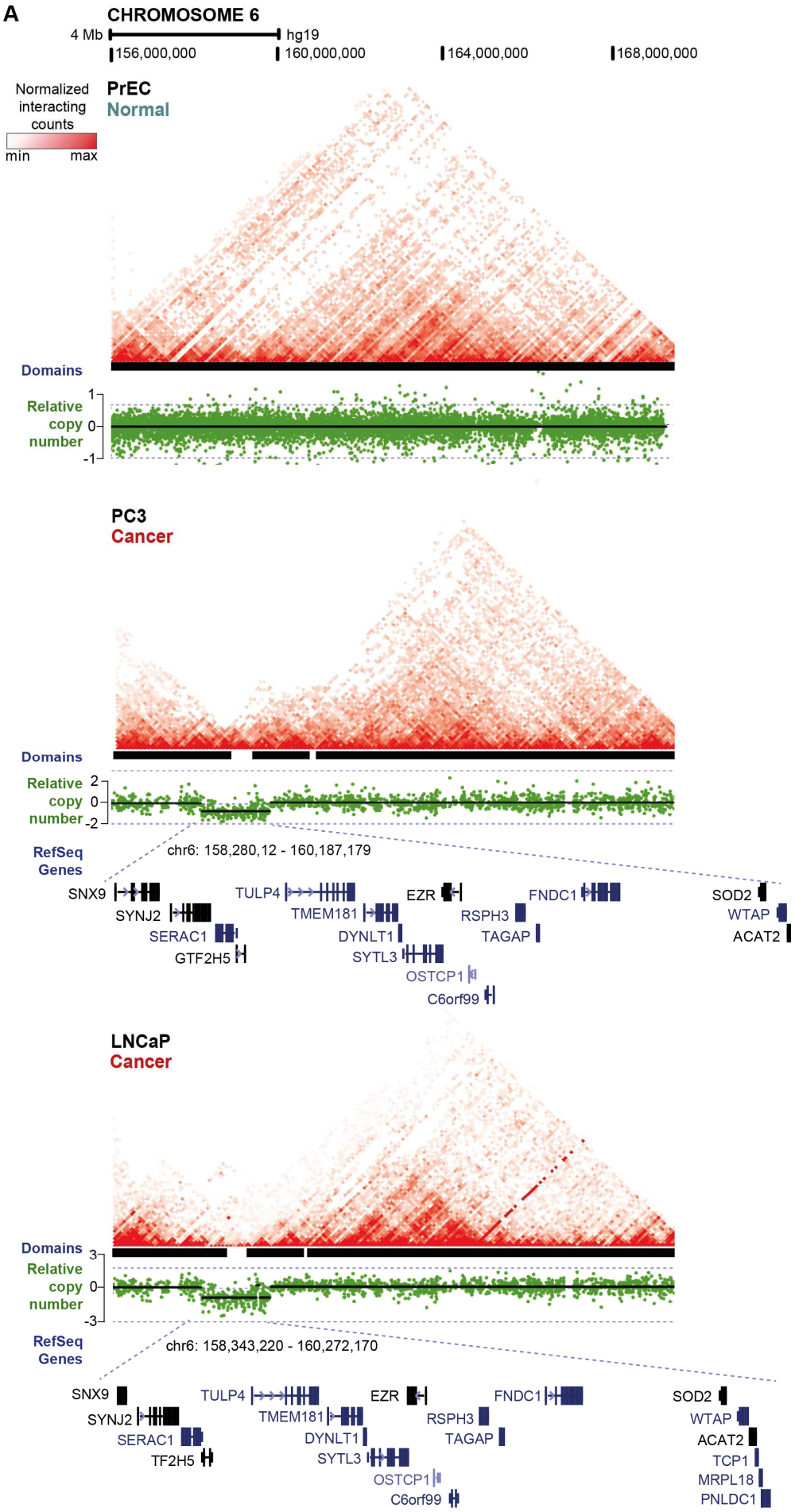

C
CHROMOSOME 12
5 Mb
hg19
59,000,000
64,000,000
69,000,000
74,000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
CTCF
PC3
Cancer
Domains
CTCF
LNCaP
Cancer
Domains
CTCF
RefSeq Genes
DTX3
OSBPL8

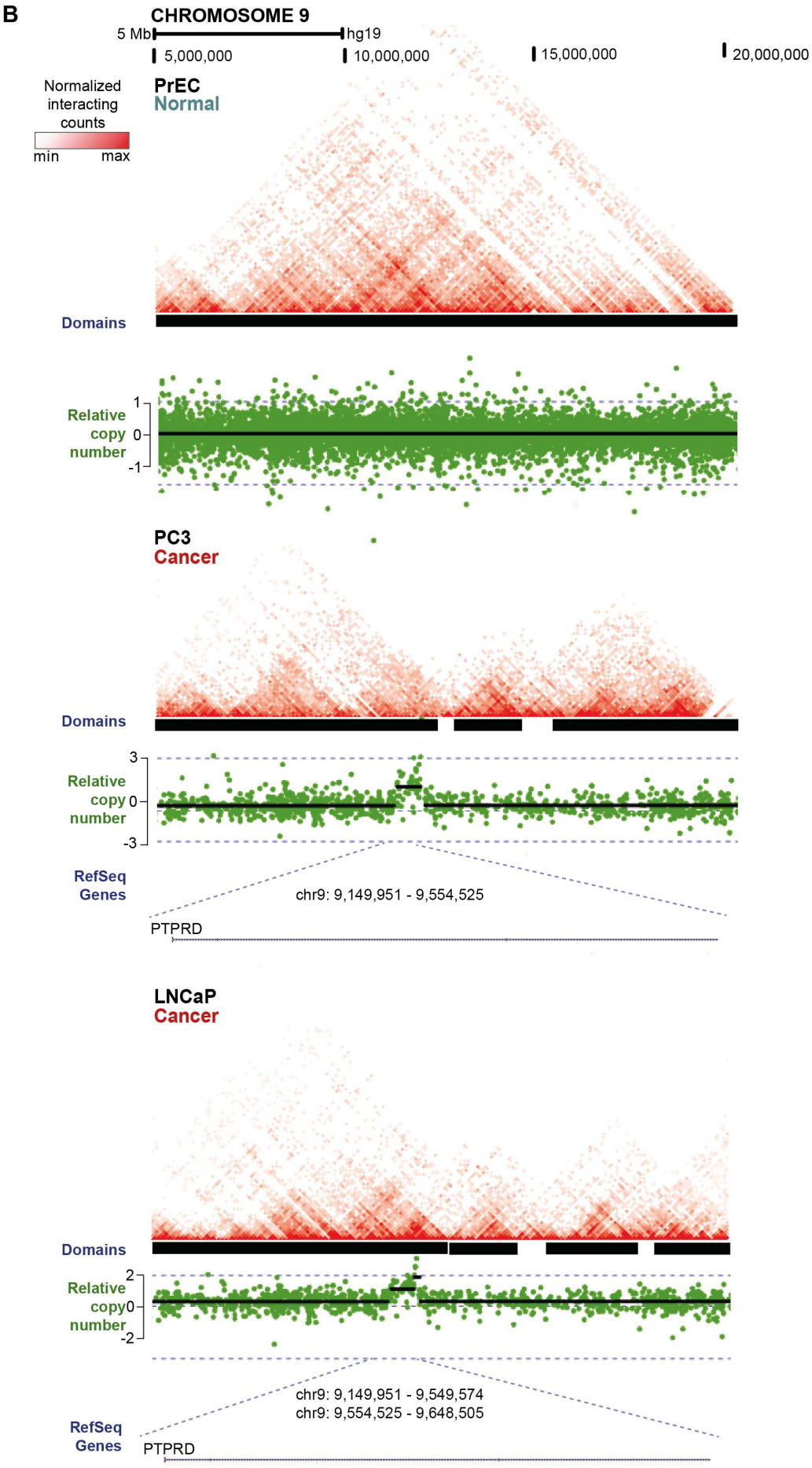


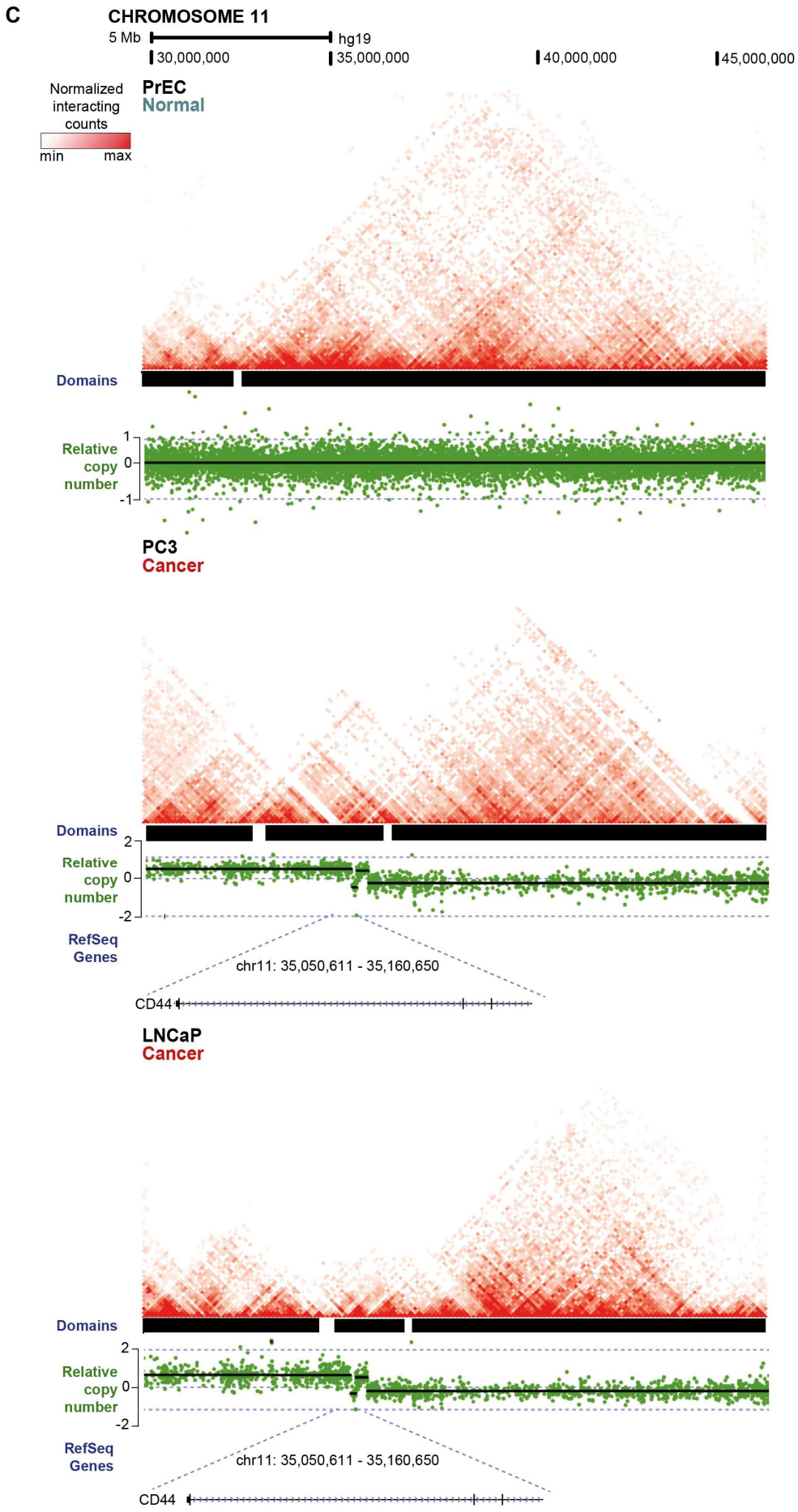
Supplemental Figure S1: 3C validation and comparison of Topological-Associated Domains (TADs) across cell lines. (A) 3C validation of the Hi-C data across the Kallikrein gene locus. Top panel: Chromatin loops identified by 3C in PrEC cells are plotted below the location of RefSeq genes across the KLK region. Enrichment of active histone marks (H3K4me3, H3K4me1, H3K27ac) and repressive histone marks (H3K27me3), as well as CTCF binding and RNA-seq track are shown. Bottom panel: Hi-C interaction data is visualised as arcs connecting each end of a chromatin loop. Normalised 3C-qPCR values are plotted against the genomic distance from the "bait" locus. **(B-D)** Chromatin interaction heat maps in PrEC, PC3 and LNCaP cells visualised as two-dimensional interaction matrices in WashU Epigenome Browser. The interaction data is aligned with RefSeq genes and CTCF binding sites.

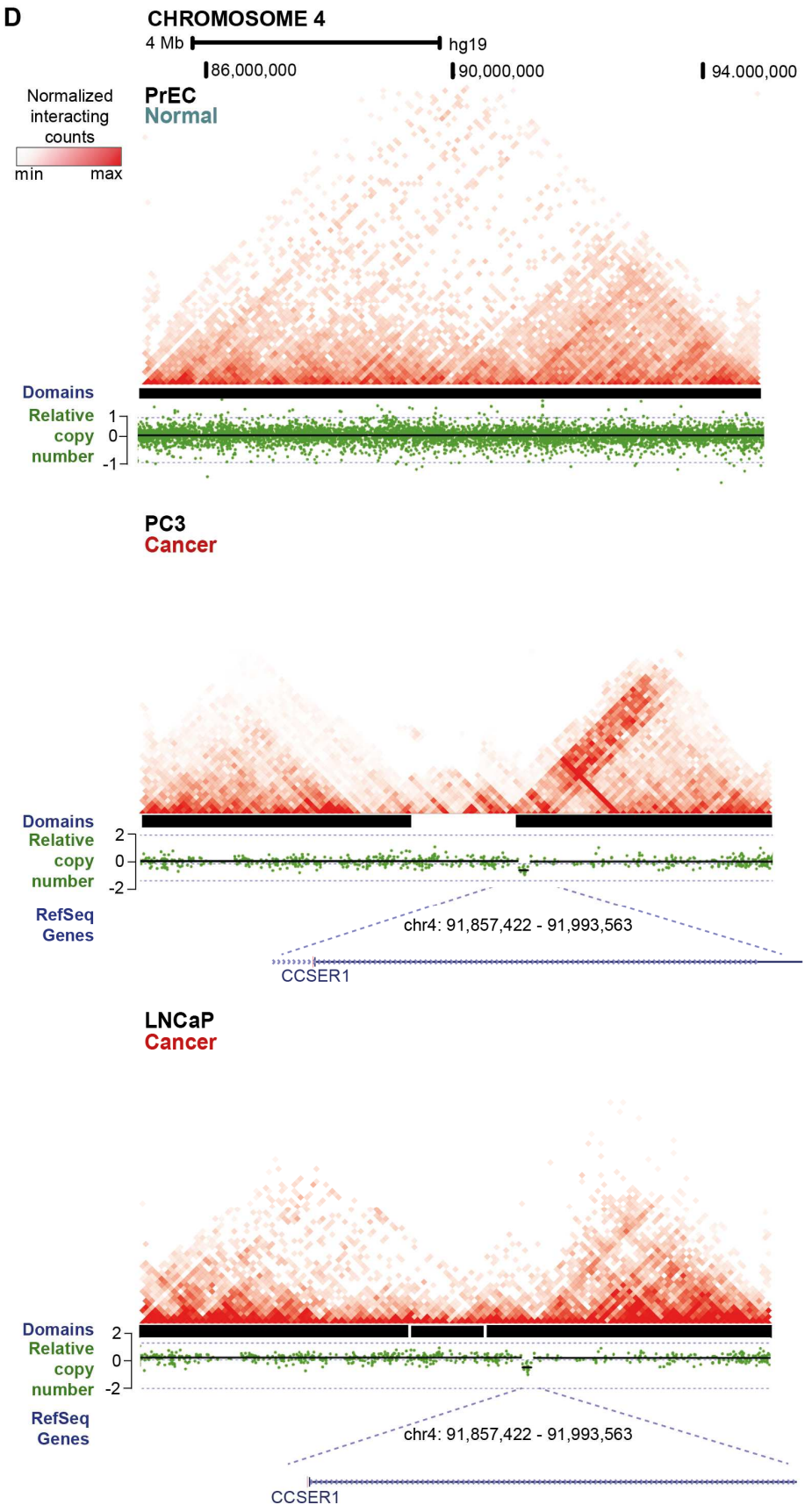


Supplemental Figure S2: Enrichment of CTCF binding and H3K4me3 at domain boundaries and gene ontology analysis of genes located at domain boundaries. (A) The proportion of CTCF binding sites that are considered to be 'associated' with a domain boundary (±20kb window was used due to 40Kb binning of topological domains). **(B)** The proportion of H3K4me3 binding sites that are considered to be 'associated' with a domain boundary. **(C)** Gene Ontology *P* value chart for genes considered to be 'associated' with a domain boundary present only in normal cells (normal-specific boundary) or present only in cancer cells (cancer-specific boundary).

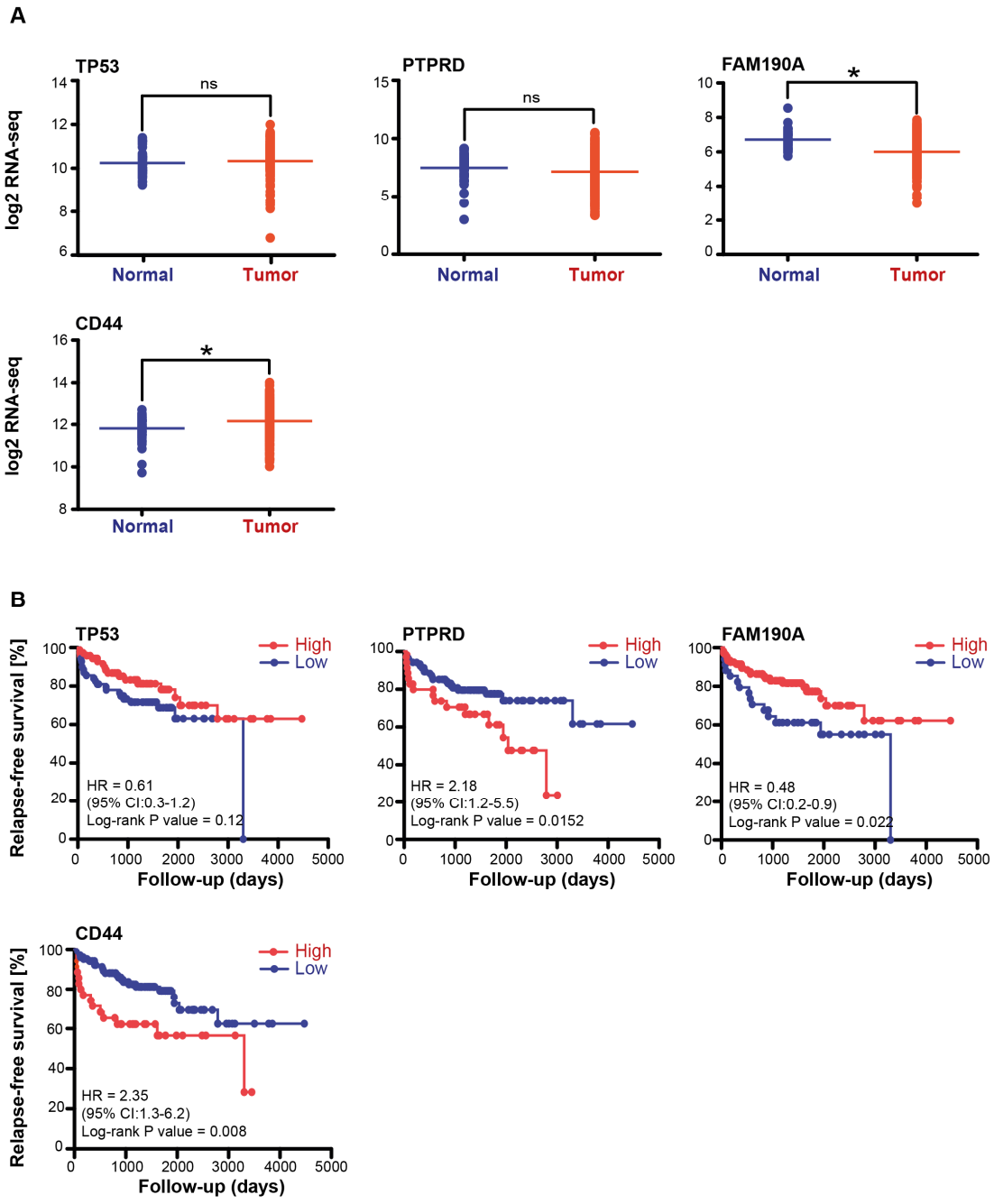

A
CHROMOSOME 6
4 Mb
hg19
156,000,000
160,000,000
164,000,000
168,000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
Relative copy number
1
0
-1
PC3
Cancer
Domains
Relative copy number
2
0
-2
RefSeq Genes
chr6: 158,280,12 - 160,187,179
SNX9
SYNJ2
SERAC1
GTF2H5
TULP4
TMEM181
DYNLT1
SYTL3
EZR
OSTCP1
C6orf99
FNDC1
RSPH3
TAGAP
SOD2
WTAP
ACAT2
LNCaP
Cancer
Domains
Relative copy number
3
0
-3
RefSeq Genes
chr6: 158,343,220 - 160,272,170
SNX9
SYNJ2
SERAC1
TF2H5
TULP4
TMEM181
DYNLT1
SYTL3
EZR
OSTCP1
C6orf99
FNDC1
RSPH3
TAGAP
SOD2
WTAP
ACAT2
TCP1
MRPL18
PNLDC1


B
CHROMOSOME 9
5 Mb
hg19
5,000,000
10,000,000
15,000,000
20,000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
Relative copy number
1
0
-1
PC3
Cancer
Domains
Relative copy number
3
0
-3
RefSeq Genes
chr9: 9,149,951 - 9,554,525
PTPRD
LNCaP
Cancer
Domains
Relative copy number
2
0
-2
chr9: 9,149,951 - 9,549,574
chr9: 9,554,525 - 9,648,505
RefSeq Genes
PTPRD

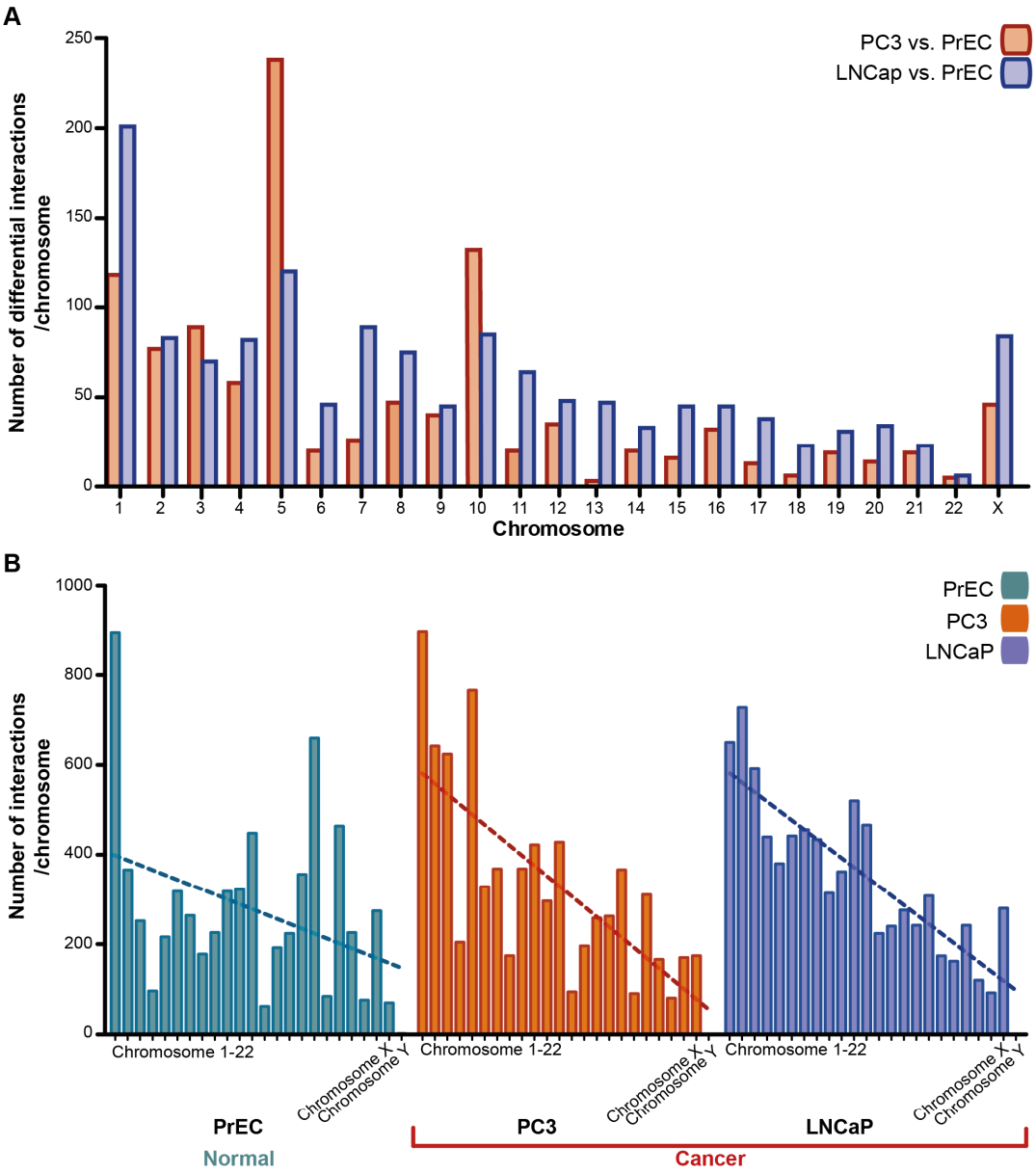

C
CHROMOSOME 11
5 Mb
hg19
30,000,000
35,000,000
40,000,000
45,000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
Relative copy number
1
0
-1
PC3
Cancer
Domains
Relative copy number
2
0
-2
RefSeq Genes
chr11: 35,050,611 - 35,160,650
CD44
LNCaP
Cancer
Domains
Relative copy number
2
0
-2
RefSeq Genes
chr11: 35,050,611 - 35,160,650
CD44


D
CHROMOSOME 4
4 Mb
hg19
86,000,000
90,000,000
94.000,000
Normalized interacting counts
min
max
PrEC
Normal
Domains
Relative copy number
1
0
-1
PC3
Cancer
Domains
Relative copy number
2
0
-2
RefSeq Genes
chr4: 91,857,422 - 91,993,563
CCSER1
LNCaP
Cancer
Domains
Relative copy number
2
0
-2
RefSeq Genes
chr4: 91,857,422 - 91,993,563
CCSER1

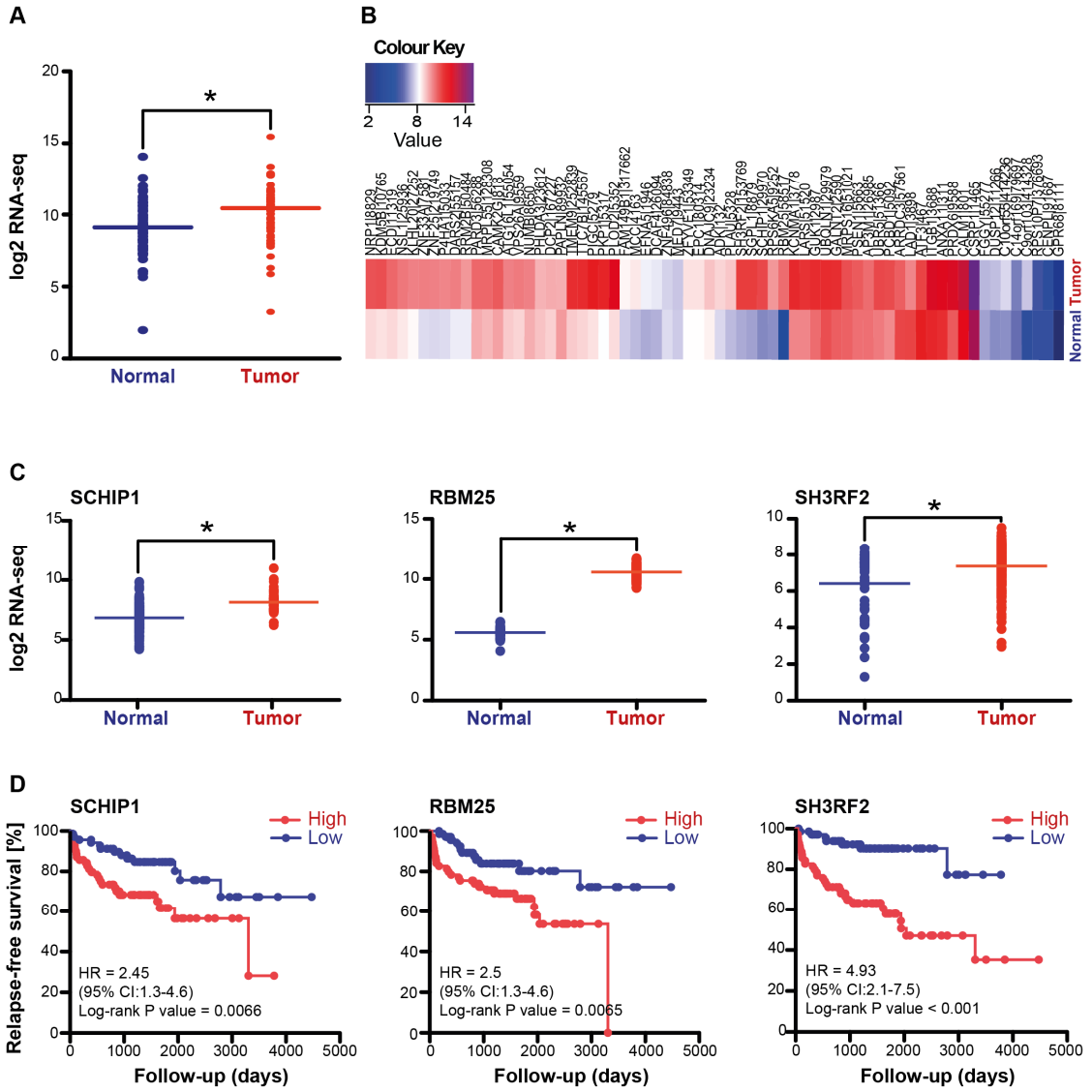
Supplemental Figure S3: Copy number variants (CNVs) are associated with the formation of new domain boundaries in cancer cells. CNV regions visualized as relative copy number estimates for each cell type (presented in green) are aligned with chromatin interaction heatmaps (presented as normalized interaction counts visualised in WashU Epigenome Browser) and TADs, demonstrating that cancer-specific domain boundaries are located at regions of CNVs in cancer cell lines. The location of RefSeq genes in the region of copy number variation is indicated below. **(A)** An example from chromosome 6 is shown, where a ~ 2Mb deletion that is present in both cancer cell lines is associated with establishment of a new domain boundary **(B)** An example from chromosome 9 is shown, where a ~400kb amplification at the PTPRD gene in both cancer cell lines is associated with establishment of a new domain boundary **(C)** An example from chromosome 11 is shown, where a ~10kb amplification at the CD44 gene in both cancer cell lines is associated with establishment of a new domain boundary **(D)** An example from chromosome 4 is shown, where a ~150kb deletion at the *CCSER1* gene in both cancer cell lines is associated with establishment of a new domain boundary.



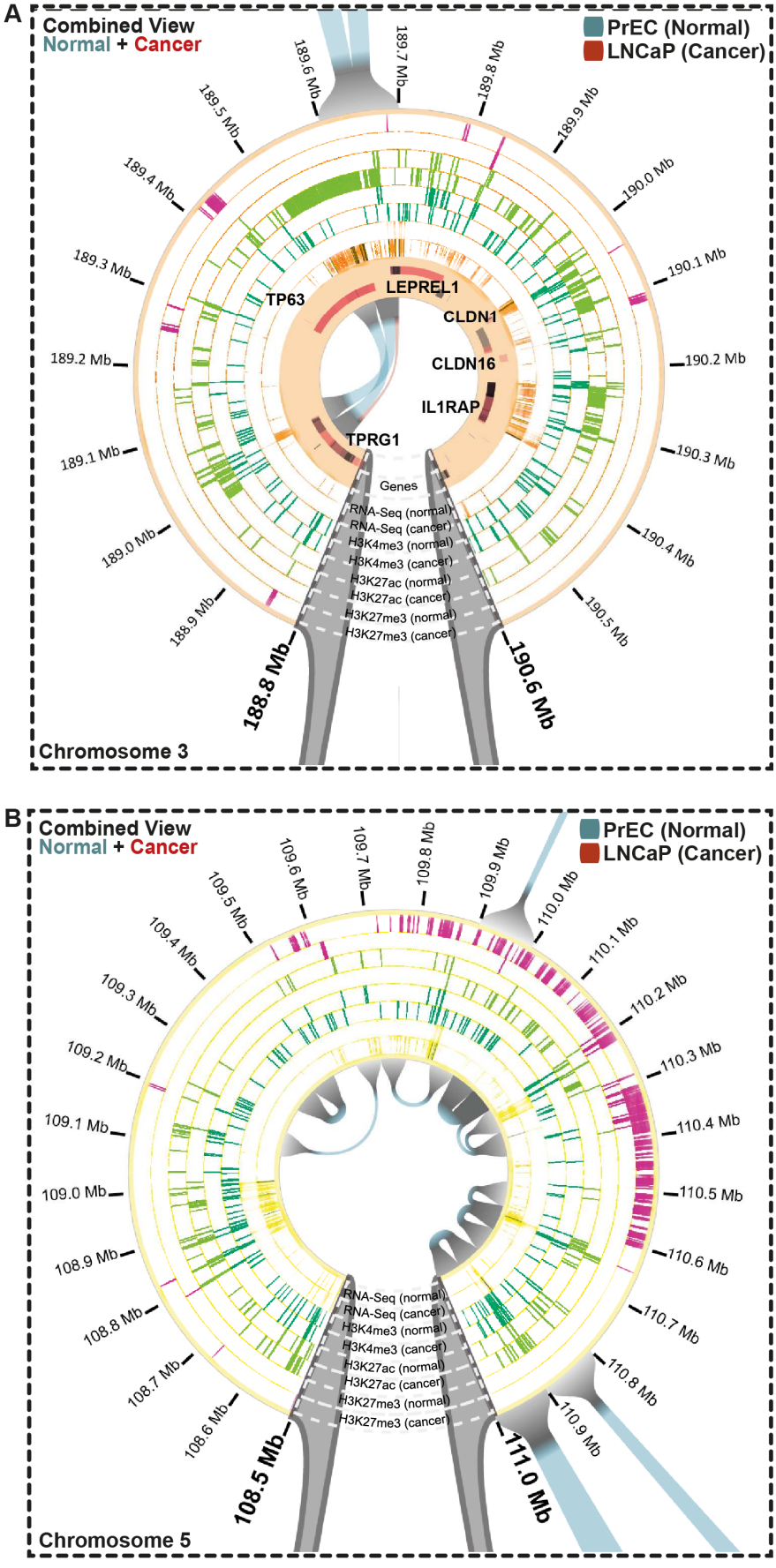
Supplemental Figure S4: Genes located at TAD-altering CNVs are associated with altered expression in primary tumours and survival. (A) Expression of TP53 (deletion on chr17) PTPRD (amplification on chr6), FAM190A (CCSER1) (deletion on chr4) and CD44 (amplification on chr11) in normal and tumour prostate samples from PRAD TCGA dataset. **(B)** Kaplan-Meier survival curves for RFS for *TP53*, *PTPRD*, *FAM190A* (*CCSER1*) and *CD44* genes.

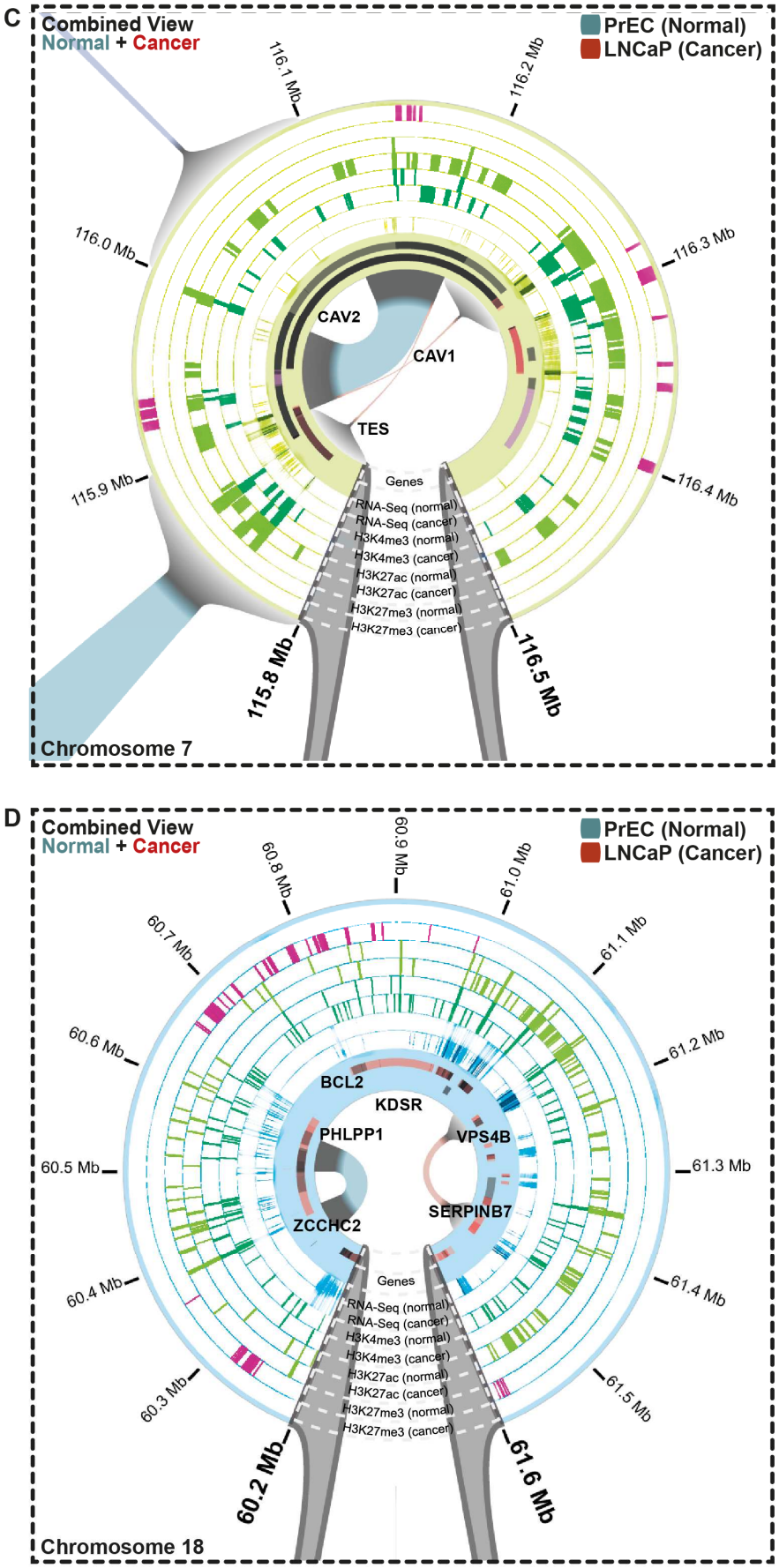


Supplemental Figure S5: Distribution of intra-chromosomal interactions and identified differential interactions by chromosome. (A) Numbers and distribution by chromosome of identified differential intra-chromosomal interactions. **(B)** Numbers and distribution by chromosome of intra-chromosomal interactions in normal (PrEC), and cancer (PC3 and LNCaP) cells. **(C)** Pie chart of inter- and intra-domain differential interactions (at 100kb resolution).

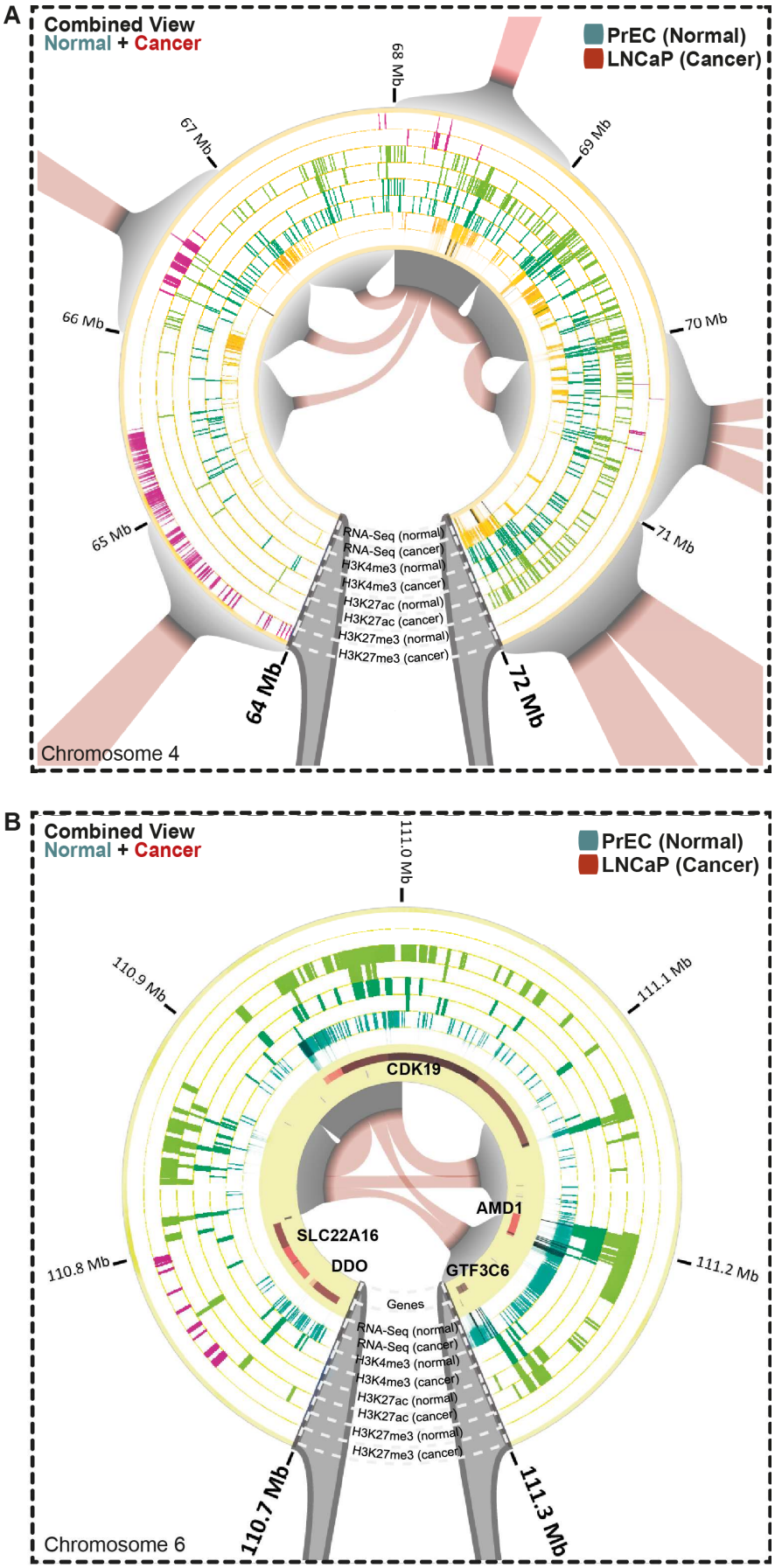


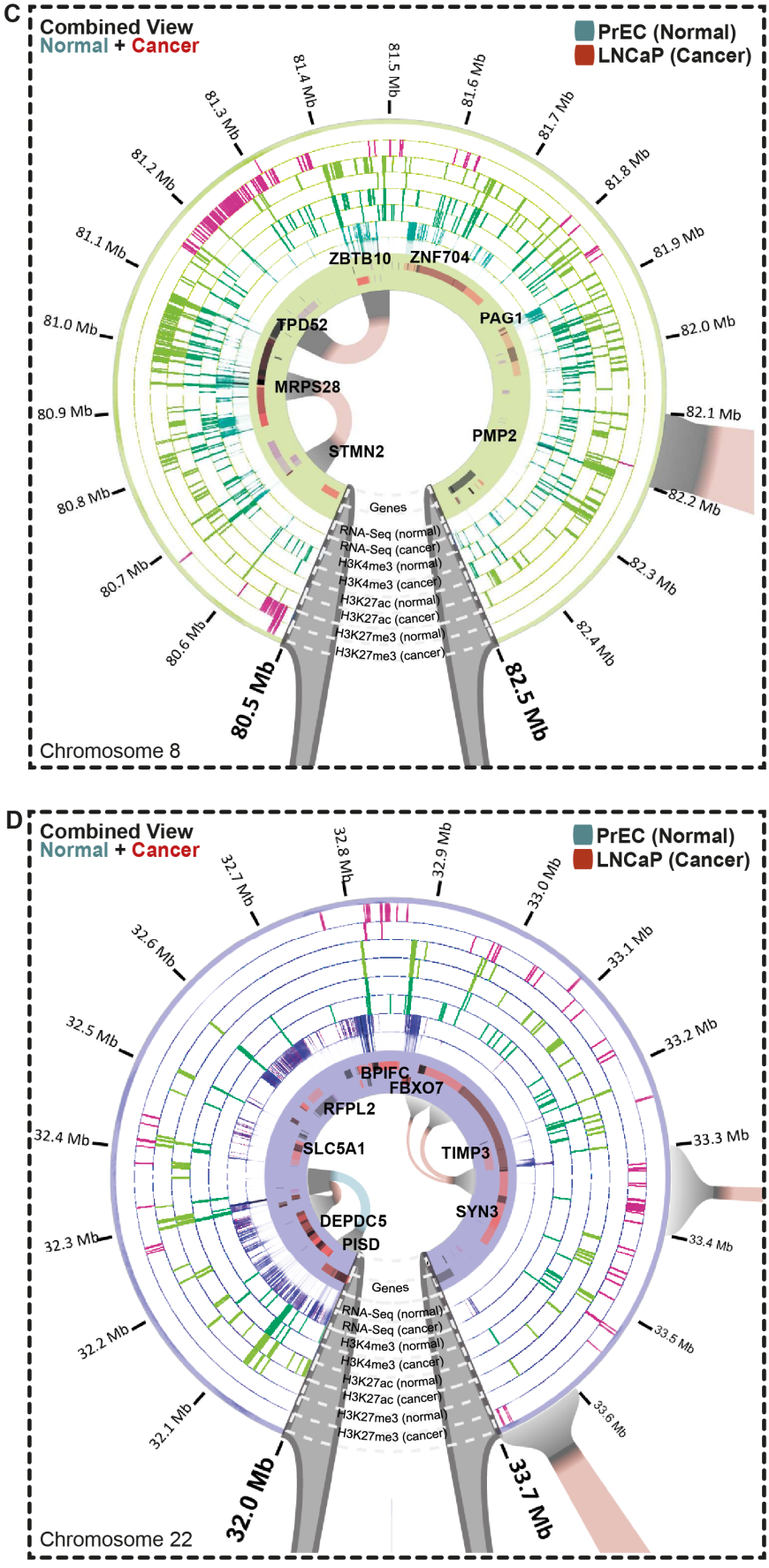
Supplemental Figure S6: Genes located at differential interactions are associated with altered expression in primary prostate cancer. (A-B) Aligned dot plot and a heatmap showing significantly altered expression (log2 RNA-seq, $P < 0.001$) of genes at differential interactions in primary prostate cohort (TCGA PRAD). **(C)** SCHIP1, RBM25 and SH3RF2 gene expression in tumour and normal samples in TCGA PRAD dataset. **(D)** Kaplan-Meier survival curves for RFS for *SCHIP1*, *RBM25* and *SH3RF2* genes.


A
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
189.6 Mb
189.7 Mb
189.8 Mb
189.9 Mb
190.0 Mb
190.1 Mb
190.2 Mb
190.3 Mb
190.4 Mb
190.5 Mb
190.6 Mb
189.5 Mb
189.4 Mb
189.3 Mb
189.2 Mb
189.1 Mb
189.0 Mb
188.9 Mb
188.8 Mb
TP63
LEPREL1
CLDN1
CLDN16
IL1RAP
TPRG1
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
Chromosome 3
B
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
109.6 Mb
109.7 Mb
109.8 Mb
109.9 Mb
110.0 Mb
110.1 Mb
110.2 Mb
110.3 Mb
110.4 Mb
110.5 Mb
110.6 Mb
110.7 Mb
110.8 Mb
110.9 Mb
111.0 Mb
109.5 Mb
109.4 Mb
109.3 Mb
109.2 Mb
109.1 Mb
109.0 Mb
108.9 Mb
108.8 Mb
108.7 Mb
108.6 Mb
108.5 Mb
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
Chromosome 5


C
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
116.1 Mb
116.2 Mb
116.0 Mb
116.3 Mb
CAV2
CAV1
TES
115.9 Mb
116.4 Mb
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
115.8 Mb
116.5 Mb
Chromosome 7
D
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
60.8 Mb
60.9 Mb
61.0 Mb
60.7 Mb
61.1 Mb
60.6 Mb
61.2 Mb
BCL2
KDSR
PHLPP1
VPS4B
60.5 Mb
61.3 Mb
SERPINB7
ZCCHC2
60.4 Mb
61.4 Mb
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
60.3 Mb
61.5 Mb
60.2 Mb
61.6 Mb
Chromosome 18

Supplemental Figure S7: Long-range epigenetically silenced (LRES) domains occur at differential interactions in cancer cells. (A-D) Four different examples of LRES regions showing anchor points of differential interactions between normal PrEC and LNCaP cancer cells are visualised in *Rondo* simultaneously with ChIP-Seq (H3K27ac, H3K4me3 and H3K27me3), RefSeq genes and RNA-seq data (circular tracks). Those interactions unique to PrEC (normal; teal) and LNCaP (cancer; orange) are evident, while the shared interaction is shown in yellow. The circular tracks depict gene expression (RNA-seq) and histone marks (H3K4me3, dark green; H3K27ac, light green; H3K27me3, pink). Teal lines in the circle depict a loss of chromatin interactions in cancer and the orange lines depict a gain of interaction in the cancer cells.


A
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
68 Mb
67 Mb
69 Mb
66 Mb
70 Mb
65 Mb
71 Mb
64 Mb
72 Mb
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
Chromosome 4
B
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
111.0 Mb
110.9 Mb
111.1 Mb
CDK19
AMD1
SLC22A16
DDO
GTF3C6
110.8 Mb
111.2 Mb
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
110.7 Mb
111.3 Mb
Chromosome 6


C
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
81.4 Mb
81.5 Mb
81.6 Mb
81.7 Mb
81.8 Mb
81.9 Mb
82.0 Mb
82.1 Mb
82.2 Mb
82.3 Mb
82.4 Mb
82.5 Mb
81.3 Mb
81.2 Mb
81.1 Mb
81.0 Mb
80.9 Mb
80.8 Mb
80.7 Mb
80.6 Mb
80.5 Mb
ZBTB10
ZNF704
TPD52
PAG1
MRPS28
PMP2
STMN2
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
Chromosome 8
D
Combined View
Normal + Cancer
PrEC (Normal)
LNCaP (Cancer)
32.8 Mb
32.9 Mb
33.0 Mb
33.1 Mb
33.2 Mb
33.3 Mb
33.4 Mb
33.5 Mb
33.6 Mb
33.7 Mb
32.7 Mb
32.6 Mb
32.5 Mb
32.4 Mb
32.3 Mb
32.2 Mb
32.1 Mb
32.0 Mb
BPIFC
FBXO7
RFPL2
SLC5A1
TIMP3
SYN3
DEPDC5
PISD
Genes
RNA-Seq (normal)
RNA-Seq (cancer)
H3K4me3 (normal)
H3K4me3 (cancer)
H3K27ac (normal)
H3K27ac (cancer)
H3K27me3 (normal)
H3K27me3 (cancer)
Chromosome 22

Supplemental Figure S8: Long-range epigenetically activated (LREA) domains occur at differential interactions in cancer cells. (A-D) Four different examples of LREA regions showing anchor points of differential interactions between normal PrEC and LNCaP cancer cells are visualised in *Rondo* simultaneously with ChIP-Seq (H3K27ac, H3K4me3 and H3K27me3), RefSeq genes and RNA-seq data (circular tracks). The circular tracks depict gene expression (RNA-seq) and histone marks (H3K4me3, dark green; H3K27ac, light green; H3K27me3, pink). Teal lines in the circle depict a loss of chromatin interactions in cancer and the orange lines depict a gain of interaction in the cancer cells. Four additional examples are shown.

SUPPLEMENTAL TABLES

Supplemental Table S1

Cell Line	Raw Read Pairs	Times Coverage
PrEC Replicate #1	344689196	11.49
PrEC Replicate #2	357794528	11.93
PrEC Replicate #3	368347682	12.28
PrEC pooled	1070831406	35.7
PC3 Replicate #1	184867580	6.16
PC3 Replicate #2	158413376	5.28
PC3 pooled	343280956	11.44
LNCaP Replicate #1	225174794	7.51
LNCaP Replicate #2	231360118	7.71
LNCaP Replicate #3	334794636	11.16
LNCaP Replicate #4	198184060	6.61
LNCaP Replicate #5	182724438	6.09
LNCaP Replicate #6	257772746	8.6
LNCaP Replicate #7	136613906	4.55
LNCaP Replicate #8	212618640	7.09
LNCaP pooled	1779243338	59.31
Total	4607468062	153.58

Supplemental Table S5

Groups	Count	Sum	Mean	SD	Variance	SS	Std Err	Lower	Upper
LNCaP [Mb]	1111	2479.8	2.23	2.23	4.89	5429.29	0.11	2.02	2.44
PC3 [Mb]	622	2438.16	3.92	3.54	12.52	7772.61	0.14	3.64	4.2
PrEC [Mb]	317	2497.07	7.88	6.5	42.19	13332.21	0.2	7.48	8.27
ANOVA									
Sources	**SS**	**df**	**MS**		**F**	***P* value**	**F crit**	**RMSSE**	**Omega Sq**
Between Groups	7941.41	2	3970.71		306.32	4.18^{E-117}	3	0.805	0.23
Within Groups	26534.1	2047	12.96						
Total	34475.52	2049	16.82						

Count – Number of TADs
Sum – Sum of the TAD sizes
Mean – Mean size of TADs
SD – Standard deviation
SS – Sum of squares
Std Err – Standard error
Lower– Lower 95% confidence interval
Upper – Upper 95% confidence interval
df – Degrees of freedom
MS – Mean Square
F – F-statistic value
F crit – F-critical value
RMSSE – Root Square Standardised Effect
Omega Sq – Omega squared

Supplemental Table S6

	PrEC	Random (n=100)	*P* value
ESC	132 (21.71%)	101 (16.6%)	$P < 0.01$
IMR90	145 (23.85%)	92.85 (15.27%)	$P < 0.01$
LNCaP	471 (77.45%)	197 (32.4%)	$P < 0.01$
PC3	472 (77.63%)	92.8 (15.26%)	$P < 0.01$

Supplemental Table S13

Cell Line	Data Type	Source	Accession
PrEC	HiC-seq	This publication	GSE73785
PrEC	CTCF ChIP-seq	Bert et al., 2013	GSE38685
PrEC	RAD21 ChIP-seq	This publication	GSE73785
PrEC	H3K4me3 ChIP-seq	Bert et al., 2013	GSE38685
PrEC	H3K4me1 ChIP-seq	Taberlay et al., 2014	GSE57498
PrEC	H3K27ac ChIP-seq	Taberlay et al., 2014	GSE57498
PrEC	H3K27me3 ChIP-seq	Taberlay et al., 2014	GSE57498
PrEC	RNA-seq	This publication	GSE73785
PrEC	ChromHMM	Taberlay et al., 2014	GSE57498
PrEC	SNP array	Robinson et al., 2010	GSE24546
PC3	HiC-seq	This publication	GSE73785
PC3	CTCF ChIP-seq	Taberlay et al., 2014	GSE57498
PC3	RAD21 ChIP-seq	This publication	GSE73785
PC3	H3K4me3 ChIP-seq	Taberlay et al., 2014	GSE57498
PC3	H3K4me1 ChIP-seq	Taberlay et al., 2014	GSE57498
PC3	H3K27ac ChIP-seq	Taberlay et al., 2014	GSE57498
PC3	RNA-Seq	Prensner et al., 2011	GSE25183
PC3	ChromHMM	Taberlay et al., 2014	GSE57498
PC3	SNP array	Rothenberg et al., 2010	GSE20306
LNCaP	HiC-seq	This publication	GSE73785
LNCaP	CTCF ChIP-seq	Bert et al., 2013	GSE38685
LNCaP	RAD21 ChIP-seq	This publication	GSE73785
LNCaP	H3K4me3 ChIP-seq	Bert et al., 2013	GSE38685
LNCaP	H3K4me1 ChIP-seq	This publication	GSE73785
LNCaP	H3K27ac ChIP-seq	This publication	GSE73785
LNCaP	H3K27me3 ChIP-seq	Bert et al., 2013	GSE38685
LNCaP	RNA-seq	This publication	GSE73785
LNCaP	SNP array	Robinson et al., 2010	GSE24546
LNCaP	SNP array	Rothenberg et al., 2010	GSE20306
ESC	HiC-seq	Dixon et al., 2012	GSE35156
IMR90	HiC-seq	Dixon et al., 2012	GSE35156

SUPPLEMENTAL REFERENCES

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research* **24**: 999-1011.

Ay F, Noble WS. 2015. Analysis methods for studying the 3D architecture of the genome. *Genome biology* **16**: 183.

Bengtsson H, Irizarry R, Carvalho B, Speed TP. 2008. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**: 759-767.

Bengtsson H, Wirapati P, Speed TP. 2009. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* **25**: 2149-2156.

Bert SA, Robinson MD, Strbenac D, Statham AL, Song JZ, Hulf T, Sutherland RL, Coolen MW, Stirzaker C, Clark SJ. 2013. Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer cell* **23**: 9-22.

Buske FA, French HJ, Smith MA, Clark SJ, Bauer DC. 2014. NGSANE: a lightweight production informatics framework for high-throughput data analysis. *Bioinformatics* **30**: 1471-1472.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376-380.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Feng X, Grossman R, Stein L. 2011. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* **12**: 139.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**: 1760-1774.

Heger A, Webber C, Goodson M, Ponting CP, Lunter G. 2013. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**: 2046-2048.

Horoszewicz JS, Leong SS, Kawinski E, Karr JP, Rosenthal H, Chu TM, Mirand EA, Murphy GP. 1983. LNCaP model of human prostatic carcinoma. *Cancer research* **43**: 1809-1818.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**: 44-57.

Jones E, Oliphant E, Peterson P. SciPy: Open Source Scientific Tools for Python, 2001-

Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW. 1979. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Invest Urol* **17**: 16-23.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li W, Gong K, Li Q, Alber F, Zhou XJ. 2015. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* **31**: 960-962.

Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic acids research* **41**: e108.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.

Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS. 2015. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome research* **25**: 544-557.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

Lun AT, Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**: 258.

Oakford PC, James SR, Qadi A, West AC, Ray SN, Bert AG, Cockerill PN, Holloway AF. 2010. Transcriptional and epigenetic regulation of the GM-CSF promoter by RUNX1. *Leuk Res* **34**: 1203-1213.

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD et al. 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29**: 742-749.

Robinson MD, Stirzaker C, Statham AL, Coolen MW, Song JZ, Nair SS, Strbenac D, Speed TP, Clark SJ. 2010. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome research* **20**: 1719-1729.

Shen L, Shao N, Liu X, Nestler E. 2014. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**: 284.

Taberlay PC, Kelly TK, Liu CC, You JS, De Carvalho DD, Miranda TB, Zhou XJ, Liang G, Jones PA. 2011. Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell* **147**: 1283-1294.

Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. 2014. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome research* **24**: 1421-1432.

Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B et al. 2010. Integrative genomic profiling of human prostate cancer. *Cancer cell* **18**: 11-22.

The Cancer Genome Atlas Research Network. 2015. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**: 1011-1025.

Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S et al. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research* **37**: D603-610.

Zhou X, Lowdon RF, Li D, Lawson HA, Madden PA, Costello JF, Wang T. 2013. Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat Methods* **10**: 375-376.