Chromatin architecture aberrations in prostate cancer and leukemia

by

James Hawley

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Medical Biophysics
University of Toronto

Chromatin architecture aberrations in prostate cancer and leukemia

James Hawley
Doctor of Philosophy

Graduate Department of Medical Biophysics
University of Toronto
2021

Abstract

Abstract text goes here. Maximum 350 words for doctoral or 150 words for master's thesis excluding title. Do not include graphs, charts, tables, or illustrations in the abstract. Uses style "Abstract text" (double spaced).

Acknowledgments

Use Body Text or Normal style for text in this section.

Contents

List of Tables

List of Figures

Chapter 1

Epigenetic dynamics underlying B cell acute lymphoblastic leukemia relapse

J.R.H, L.G.-P., A.M., J.E.D., and M.L. conceptualized the study. S.M.D., L.G.-P., R.J.V., E.W, J.M., O.I.G., I.G., S.Z.X., M.H., S.R.O., G.N., S.M.C., J.E., C.J.G., J.S.D., M.D.M., C.G.M., and J.E.D. were involved with primary data acquisition. J.R.H., L.G.-P., A.M., and M.C.-S.-Y., J.E.D., and M.L. were involved with the statistical and computational data analysis and biological interpretation. J.R.H. performed all analyses with the DNA methylation (DNAme) data, M.C.-S.-Y. with the RNA sequencing (RNA-seq) data, and A.M. with the assay for transposase-accessible chromatin sequencing (ATAC-seq) data and integration. J.R.H, L.G.-P., and A.M. designed the figures. J.E.D. and M.L. oversaw the study.

1.1 Abstract

1. Relapse of B-cell acute lymphoblastic leukemia (B-ALL) is common in both pediatric and adult patients

2. Previous investigation of the origins of B-ALL relapse identifies growth of subclones present at diagnosis as critical factor

3. Here, we use patient matched diagnosis and relapse samples and multiple PDXs to investigate

epigenetic dynamics that occur over B-ALL relapse

4. Find most DMRs are unique to each patient and are not recurrent
5. The few recurrent DMRs occur in the promoter regions of genes associated with differentiation and stem cell characteristics
6. Suggests that relapse selects for clones with stem cell characteristics, both genetically and epigenetically

1.2 Introduction

1. Relapse of B-ALL is common in both pediatric and adult patients
2. Previous investigation of the origins of B-ALL relapse identifies growth of subclones present at diagnosis as critical factor
3. Importance of DNAme, CTCF binding, and chromatin organization in hematopoietic stem and progenitor cells (HSPCs) function raises the possibility of these roles being important in B-ALL as well
4. Previous work has identified stem-associated phenotypes and chromatin remodelling pathways as important in these subclones
5. Here, we use patient matched diagnosis and relapse samples and multiple PDXs to investigate epigenetic dynamics that occur over B-ALL relapse

1.3 Results

1.3.1 Multiomic integration of B-ALL relapse patients links DNA methylation to relapse status

- We profiled primary and patient-derived xenograft (PDX) B-ALL samples using RNA-seq, ATAC-seq, and DNA methylation capture sequencing (MeCapSeq) Figure 1.1a
- To contextualize any changes we may find in B-ALL relapse, we first looked to the hematopoietic hierarchy to characterize DNAme changes that occur over the course of B-cell differentiation

- To identify which molecular marker was most indicative of disease stage, we integrated these three data types with disease stage classifications from [REF 1]
- We then performed similarity network fusion (SNF) to construct clusters of samples from all three molecular datasets, both combined and individually
- We found for the three patients whose multiomic data passed quality control checks for all datasets, that DNAme was the individual dataset that most strongly correlated with disease state, behind the combined graphs Figure 1.1b

1.3.2 Widespread loss of DNA methylation over normal B-cell differentiation

- Given the strong correlation between DNAme signal and disease state, we decided to investigate DNAme changes in B-ALL relapse
- We performed MeCapSeq on 8 normal cord blood pools, separating into various cell types based on cell surface markers Table 1.1
- This identified 540 DMRs, 500 (92.6 %) of which become hypomethylated over the course of differentiation
- No DMRs change methylation levels then revert back; once a region changes its methylation, that change persists throughout differentiation
- These findings are corroborated by previous studies of B-cell differentiation using the Illumina 450K array [3–5]
- In summary, normal hematopoietic stem cells permanently change DNAme over the course of differentiation, predominantly by losing DNAme

1.3.3 Recurrent DNA methylation changes identify stem cell pathways in relapse

- We searched for DMRs that are present at B-ALL relapse using the using primary and PDX Dx and Rel B-ALL samples

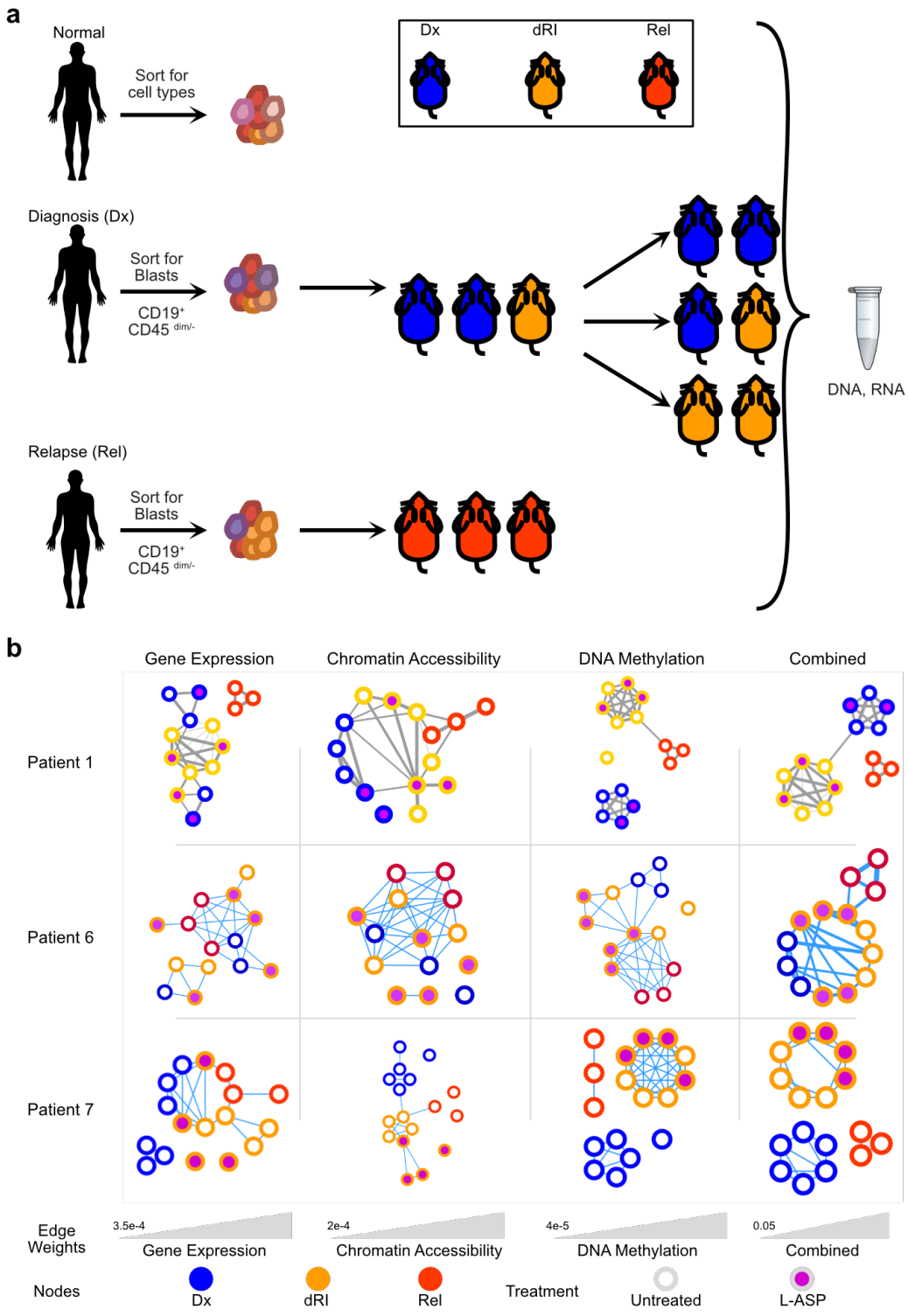

Figure 1.1: **Experimental design and data integration. a.** Experimental design of samples used in this study. Normal samples were obtained from cord blood pools and sorted into various hematopoietic cell types. B-ALL patients who experienced relapse has sorted leukemic blasts collected at Dx and Rel. Based on the mutation profiles from [REF 2] some Dx samples are labelled as dRI. **b.** Individual and fused networks of samples from three patients with complete multiomic profiling. Nodes represent individual samples (either primary or PDX), edges represent similarities between the connected samples.

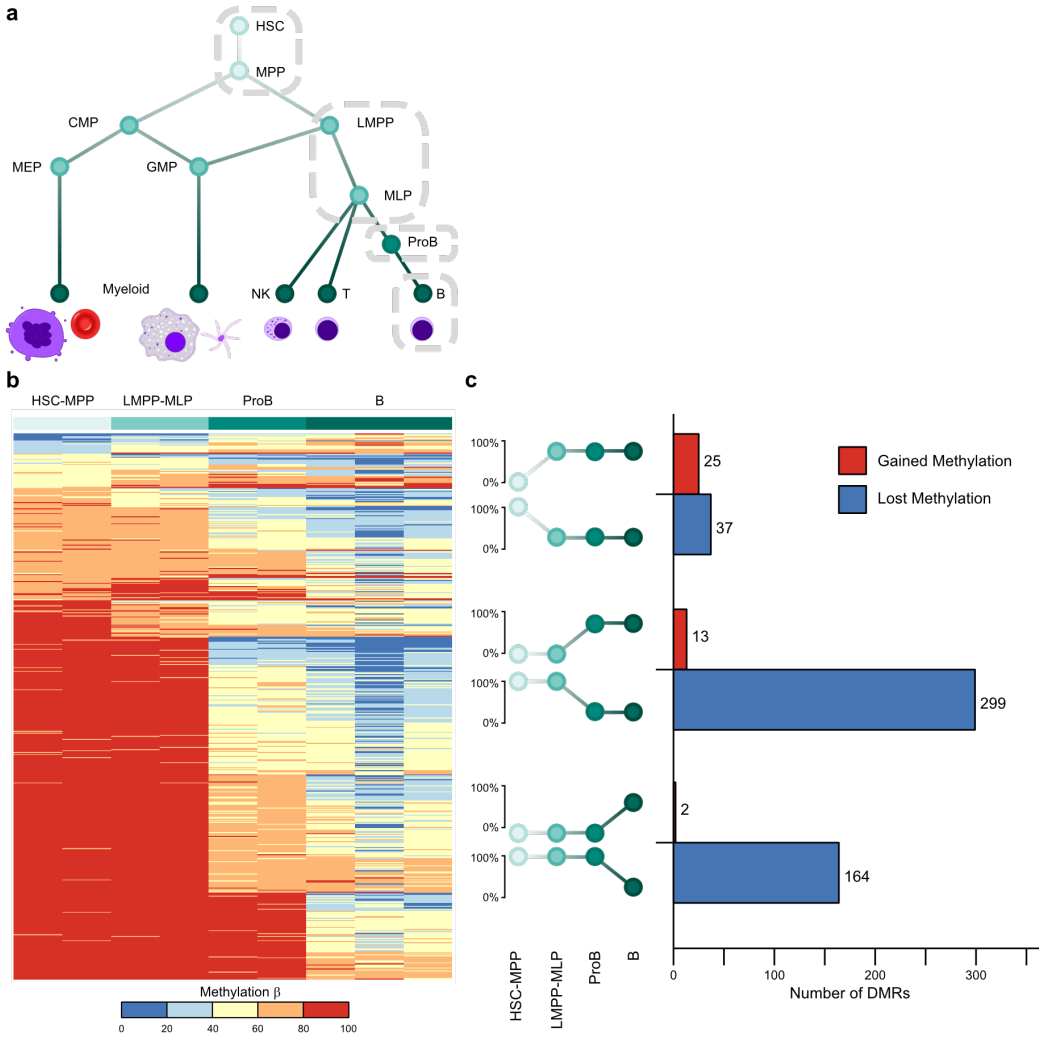

Figure 1.2: **Loss of DNA methylation over B-cell differentiation.** **a.** Schematic of the hematopoietic hierarchy and the grouping of B-cell progenitors into the groups isolated in this study. **b.** Heatmap of DMRs identified between B lineage cell types. Columns are samples ordered by cell type and rows are DMRs identified in at least one pairwise comparison between cell types (dmrseq, FDR < 0.1). **c.** Bar plot of DMRs classified by which step in differentiation they were identified as significantly changed.

- When grouping all patients together by disease stage, we found no DMRs survived multiple testing corrections
- We reanalyzed the same data in a patient-oriented approach, identifying DMRs that come from each patient's relapse trajectory
- We were able to identify 26 000 DMRs across the cohort of 5 patients Figure 1.3a
- Unlike normal differentiation, most DMRs became hypermethylated at relapse
- In nearly all cases, the Dx PDX samples strongly resembled the Dx primary samples, but one Dx PDX sample for Patient 9 more strongly resembles the methylation profile at Rel Figure 1.3a
- This suggests that, like subpopulations identified via mutation profiles, DNAme of a subpopulation of cells present at diagnosis may give rise to the relapse population
- Combining the patient-oriented DMRs demonstrates that most DMRs are patient-specific (Figure 1.3b, left side)
- Only a small number of DMRs are shared across patients, which are all within promoter regions of the genes highlighted (Figure 1.3b)
- Taken together, we find that the changes to DNAme over the course of B-ALL relapse is antithetical to the changes seen over normal B-cell differentiation
- We investigated the potential effects of these recurrent changes to DNAme using gene ontology (GO) analysis
- We found numerous biological processes positively associated with differentiation, the most significant of which is cell fate determination (Figure 1.3c)
- For the recurrent DMRs, all patients had > 20 % gain in methylation in the promoter regions of these genes ()Figure 1.3d)
- There are some regions where both hyper- and hypomethylation was observed at Rel, but all promoters had increased at least some DMR with increased methylation (Figure 1.3d)
- Taken together, this suggests that the DNAme changes observed at Rel revert to a more de-differentiated, stem-like state

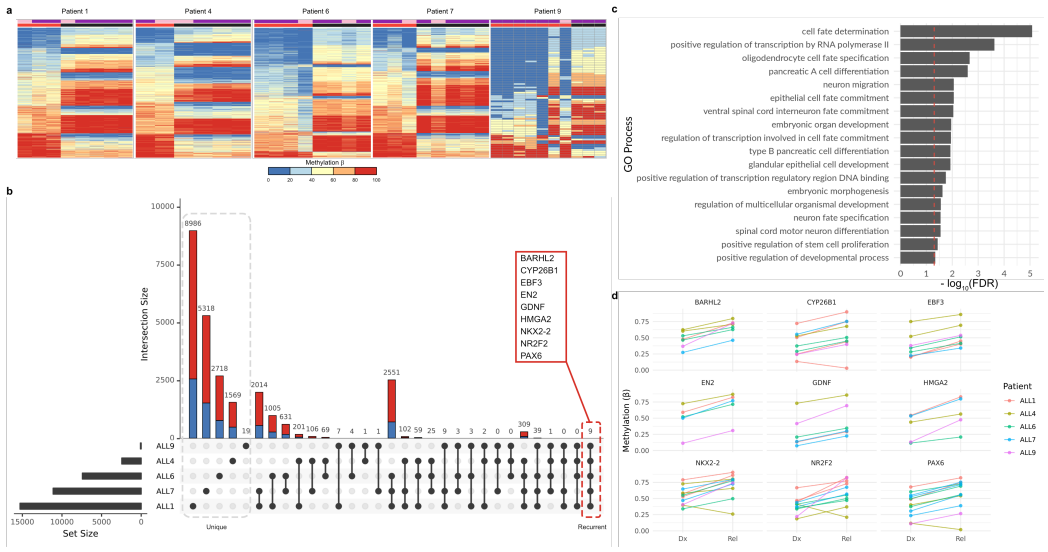

Figure 1.3: **Recurrent relapse DMRs are associated with cell fate decision processes. a.** Heatmaps of DMRs identified between Dx and Rel samples within each patient. **b.** Upset plot showing the shared DMRs between patients. DMRs in the left highlighted block are unique to a single patient, whereas DMRs in the right highlighted block are recurrent changes across all 5 relapse patients. These DMRs are in the promoter regions of the callout genes listed. **c.** GO analysis of genes with recurrently hypermethylated promoters in Rel B-ALL samples. The red dashed line indicates the FDR threshold of 0.05. **d.** Pairwise DNAme changes in each patient at the recurrently hypermethylated loci show increased methylation in all patients.

1.4 Discussion

1. summary of our work
   (a) Relapse is a major barrier to treating B-ALL
   (b) Requires rigorous molecular and multiomic investigation to understand the origins of relapse and whether it can be predicted at diagnosis
   (c) Multiomic studies here identify DNAme as an important biomarker of B-ALL relapse
   (d) DNAme and RNA-seq changes indicate presence of stem-like cells at diagnosis that become dominant at relapse
2. context of current work in regard to relapse treatment and other studies
   (a) The combined networks discriminated samples by disease stage better than any individual dataset, strengthening the claim that multiomic studies provide stronger evidence than individual molecular studies alone
3. future directions for investigation

(a) identifying important molecular changes over relapse may be easier by starting with patient-oriented analyses

(b) effect of targeted DNAme of promoter regions for important stem genes on engraftment

(c) effect of demethylating agents on relapsed B-ALL patients

(d) combination therapy of demethylating agents with chemotherapy to reduce the potential outgrowth of relapse-fated subclones

1.5 Methods

1.5.1 Patient selection and sample collection

Patient samples were obtained at diagnosis and relapse from patients with B-ALL as previously described [2]. All samples were frozen viably and stored long term at -150 °C. Samples were selected retrospectively based on paired-sample availability.

Human cord blood samples were obtained with informed consent from Trillium and Credit Valley Hospital according to procedures approved by the University Health Network Research Ethics Board, as previously described [2]. Cells were stained with the following antibodies (all from BD Biosciences, unless otherwise stated):

- FITC anti-CD45RA (1:50, 555488)
- PE anti-CD90 (1:50, 555596)
- PE-Cy5 anti-CD49f (1:50, 551129)
- V450 anti-CD7 (1:33.3, 642916)
- PE-Cy7 anti-CD38 (1:100, 335790)
- APC anti-CD10 (1:50, 340923)
- APC-Cy7 anti-CD34 (1:200, custom made by BD Biosciences)

Cells were sorted from cord blood cells on the basis of markers listed in Table 1.1, as previously described [6], on a FACSAria III (Becton Dickinson), consistently yielding > 95 % purity.

Table 1.1: **Cell surface markers used to isolate cell populations from cord blood pools.**

Cell type(s)	Surface markers
HSCs & MPPs	CD34+ CD38- CD45RA-
CMPs, GMPs, & MEPs	CD34+ CD38+ CD10- CD19+
LMPPs & MLPs	CD34+ CD38- CD45RA+
EarlyProBs, PreProBs, & ProBs	CD34+ CD38+ CD10+ CD19+
B	CD34- CD38+ CD19+ CD33- CD3- CD56-

1.5.2 Patient-derived xenograft generation and limiting dilution assays

PDXs were generated as previously described [2]. Clinical samples were stained with the following antibodies:

- anti-CD19 PE (BD Biosciences, clone 4G7)
- anti-CD3 FITC (BS Biosciences, clone SK7) or anti-CD3 APC (Beckman Coulter, clone UCHT11)
- anti-CD45 APC (BD Biosciences, clone 2D1) or anti-CD45 FITC (BD Biosciences, clone 2D1)
- anti-CD34 APC-Cy7 (BD Biosciences, clone 581)

Each sample was sorted on a FACSAria III (BD Biosciences) for leukemic blasts ($CD19^{+}CD45^{dim/-}$) and T cells ($CD3^{+}CD45^{hi}$). NOD scid gamma (NSG) mice were bred according to protocols established and approved by the Animal Care Committee at the University Health Network. 8-to-12-week-old mice were sublethally irradiated at 225 cGy 24 hours prior to transplants. Only female mice were used. Intra-femoral injections of 10 to 250 000 sorted leukemic blasts were performed as previously described [7]. Mice were sacrificed 20-to-30 weeks post-transplant or at the onset of disease symptoms. Human cell engraftment in the injected femur, bone marrow (non-injected bones, left tibia, right tibia, left femur), spleen, and central nervous system were assessed using human-specific antibodies for CD45 (PE-Cy7, BD Biosciences, clone HI30; v500 BD Biosciences, clone HI30), CD44 (PE, BD Biosciences, clone 515; FITC, BD Biosciences, clone L178), CD3 (APC, BD Biosciences, clone UCHT1), CD19 (PE-Cy5, Beckman Coulter, clone J3-119), CD33 (PE-Cy7, BD Biosciences, clone P67-6; APC, BD Biosciences, clone P67-6), and CD34 (APC-Cy7, BD Biosciences, clone 581)

analyzed on an LSRII (BD Biosciences). Mice were considered to be engrafted when > 0.1 % of cells in the injected femur were positive for one or more human B-ALL-specific cell surface marker (CD45, CD44, CD19, and CD34). Confidence intervals for the frequency of leukemia initiating cells was calculated using ELDA [8].

1.5.3 Human cell isolation from patient-derived xenografts

Cells from the injected femur, bone marrow, and spleen, were frozen viably after sacrifice. Injected femur and bone marrow of mice engrafted with > 10 % human cells were combined. These cells were depleted of mouse cells using the Miltenyi Mouse Cell Depletion Kit (Miltenyi Biotec; samples with > 20 % engraftment) or by cell sorting with human CD45 and human CD19 and/or CD34 cell surface antibodies to a purity of > 90 %, as determined by post-processing flow cytometry. Central nervous system cells from mice with > 60 % engraftment were used directly for DNA isolation. DNA was isolated using the QIAamp DNA Blood Mini or Micro Kit (Qiagen).

1.5.4 Primary and patient-derived xenograft sample sequencing

RNA sequencing

RNA-seq was performed as previously described [2]. Briefly, amplified complementary DNA (cDNA) was sequenced as paired-end libraries on an Illumina HiSeq2000. The libraries were sequenced as 2×75 bp for the adult and 2×100 bp for the pediatric samples.

DNA methylation capture sequencing

MeCapSeq was performed using the SeqCapEpi CpGiant kit (Roche NimbleGen). Briefly, the DNA library is prepared and bisulfite converted, amplified, and enriched using capture probes for targeted bisulfite-converted DNA fragments, then sequenced on a short-read sequencing machine. More specifically, library preparation for MeCapSeq was performed with the KAPA Library Preparation Kits, bisulfite conversion of genomic DNA was performed with the Zymo EZ DNA Methylation Lightning kit, bisulfite-converted DNA libraries were amplified using the KAPA HiFi HotStart Uracil+ ReadyMix kit, and finally hybridized to probes from the SeqCap Epi Enrichment Kit. Captured DNA fragments were sequenced on an Illumina HiSeq 2500 as 2×125 bp to a target depth of 70×10^6 read pairs per sample.

Assay for transposase-accessible chromatin sequencing

Library preparation for ATAC-seq was performed with the Nextera DNA Sample Preparation Kit (FC-121-1030, Illumina), according to a previously reported protocol [9]. ATAC-seq libraries were sequenced with an Illumina HiSeq 2500 sequencer to generate single-end 50 bp reads.

1.5.5 Sequencing data analysis

Differential gene expression analysis

The methods are described in [REF 2]. Briefly, RNA-seq reads were aligned against the GRCh38 reference human genome with STAR (v2.5.2b) [1] and annotated with the Ensembl reference (v90). Default parameter were used with the following exceptions: chimeric segments were screened with a minimum size of 12 bp, junction overlap of 12 bp, and maximum segment reads gap of 3 bp; splice junction overlap of 10 bp; maximum gap between aligned mates of 100 000 bp; maximum aligned intron of 100 000; and alignSJstitchMismatchNmax of 5 1 5 5. Transcript counts were obtained with HTSeq (v0.7.2) [10]. Data was library size normalized using the RLE method, followed by a variance stabilizing transformation using DESeq2 (v1.22.1) [11]. Principal component analysis plots were generated on a per sample basis using the top 1 000 variable genes. For downstream analysis, the mean expression of each sample clone condition was used. For per-patient analyses, differentially expressed genes were identified between disease stage and clone status using DESeq2. Genes with an FDR < 0.05 and absolute $\log_2$(fold change) > 1 were considered significant.

Identification of accessible chromatin peaks

ATAC-seq reads were aligned against the GRCh38 reference human genome with Bowtie2 (v2.0.5) [12] with default parameters. Accessible peaks were identified with MACS2 (v2.0.10) [13] with the following command:

```
macs2 callpeak -f BED -g hs --keep-dup all -B --SPMR --nomodel --
    shift -75 --extsize 150 -p 0.01 --call-summits -n {sample_name}
     -t {input_bam}
```

A catalogue of peaks from all samples was collected with a custom R script. ATAC-seq signal was mapped from each sample to this catalogue using Bedtools [14] for downstream analysis.

Bisulfite sequencing pre-processing

Sequencing read qualities were assessed with FastQC (v0.11.8) [15]. Low quality bases were trimmed with Trim Galore! (v0.6.3) [16] with the following command:

```
trim_galore --gzip -q 30 --fastqc_args '-o TrimGalore' {
    sample_mate1} {sample_mate2}
```

Trimmed reads were aligned to the GRCh38 reference human genome with Bismark (v0.22.1) [17] with default parameters. Duplicates were removed from the resulting alignment file with the following command:

```
deduplicate_bismark -p --bam {input_bam}
```

The deduplicated BAM file was sorted by position with sambamba (v0.7.0) [18]. M-biases were calculated with MethylDackel (v0.4.0) [19], and methylation β values were extracted from the BAM files with the following command:

```
MethylDackel extract --mergeContext --OT 3,124,3,124 --OB
    3,124,3,124 {ref_genome} {dedup_sorted_bam}
```

Both M and β values were for each CG dinucleotide (CpG) were used in downstream analyses.

Similarity network fusion

Preprocessed data from each sample was collected with the following features: normalized gene expression abundance for all genes, chromatin accessibility signal within previously identified accessible peaks, and mean β value for all CpGs listed in the manifest for targeted bisulfite sequencing kit. These features and sample labels were processed with the SNFtool R package [20] to perform the similarity network fusion analysis. Graphs were constructed for all samples deriving from a single patient where each node is a sample and each edge is weighted according to the determined similarity between the samples. Edges whose weights were below specific thresholds were removed from the graph. The threshold weight for the fused graph was 0.05. Similar graphs were constructed using the individual components for each sample (e.g. using just the similarity in RNA-seq data), and the component graphs were compared to the fused graph, to compare the importance of each feature. Threshold weights for these individual graphs were determined to be 6×10^{-5} for DNAme, 4×10^{-4} for gene expression, and 2×10^{-4} for chromatin accessibility.

Differentially methylated region identification

DMRs were identified using the dmrseq R package (v1.3.8) [21] with an absolute filtering cutoff value of 0.05 and using the sequencing batch as an adjustment covariate. Normal samples from all donors were compared pairwise based on their sorted cell type. B-ALL samples were compared by their designated disease stage (Dx, DRI, or Rel), and were compared both across all patients (e.g. all Dx samples against all Rel samples), or within a single patient (e.g. all Dx samples from Patient 1 against all Rel samples from Patient 1). A multiple testing correction with the FDR method was performed [22]. Regions with an FDR < 0.1 were determined to be significant.

Gene ontology enrichment analysis

Gene ontology enrichment analysis was performed using the PANTHER classification system (database version 2019-10-08) [23]. Gene symbols for the genes whose promoter regions contained the recurrently hyper-methylated regions in all B-ALL patient samples were supplied, with the entire human genome as the background. An over-representation Fisher test for biological processes was performed with an FDR correction. Biological processes at the top of the hierarchy with an FDR < 0.05 were determined to be significant.

Appendix A

Supplementary Material for Chapter 2

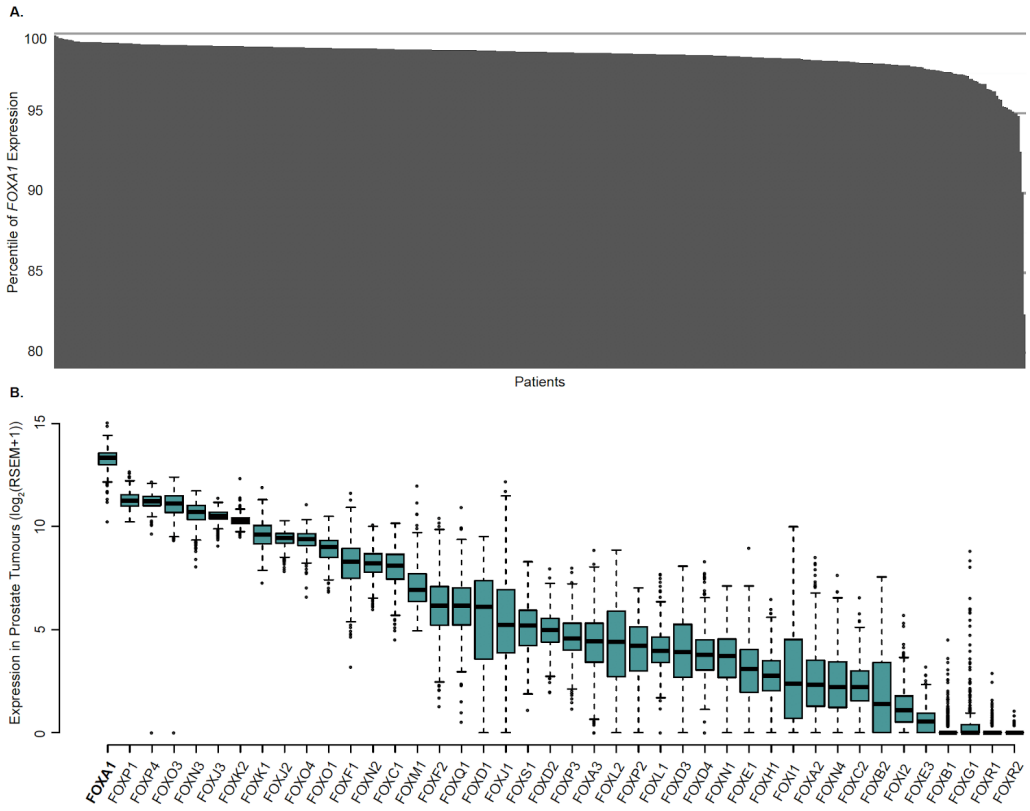

Figure A.1: ***FOXA1* mRNA expression in prostate tumours. a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

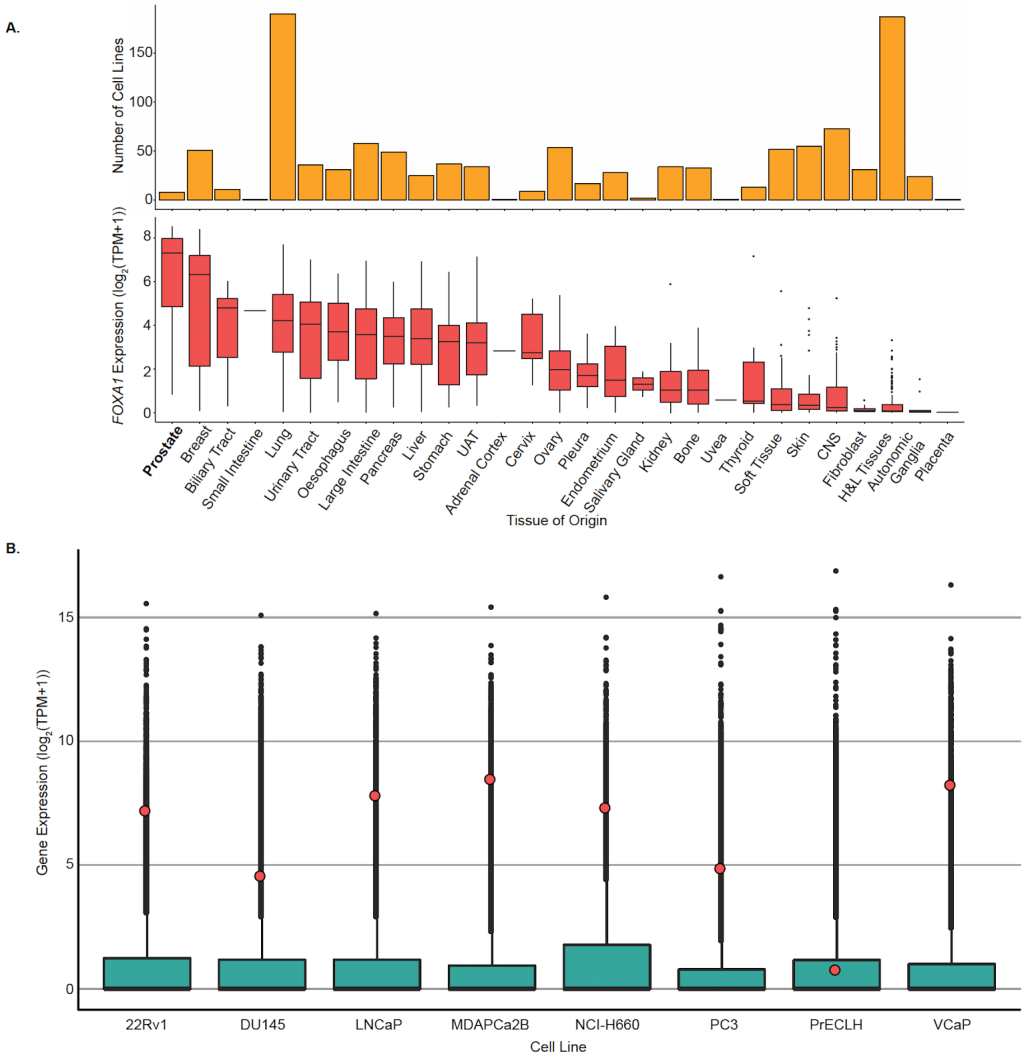

Figure A.2: ***FOXA1* mRNA expression across PCa cell lines. a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Methods). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA expression across eight PCa cell lines from DEPMAP, profiled by RNA-seq (see Methods). Red dots indicate *FOXA1*.

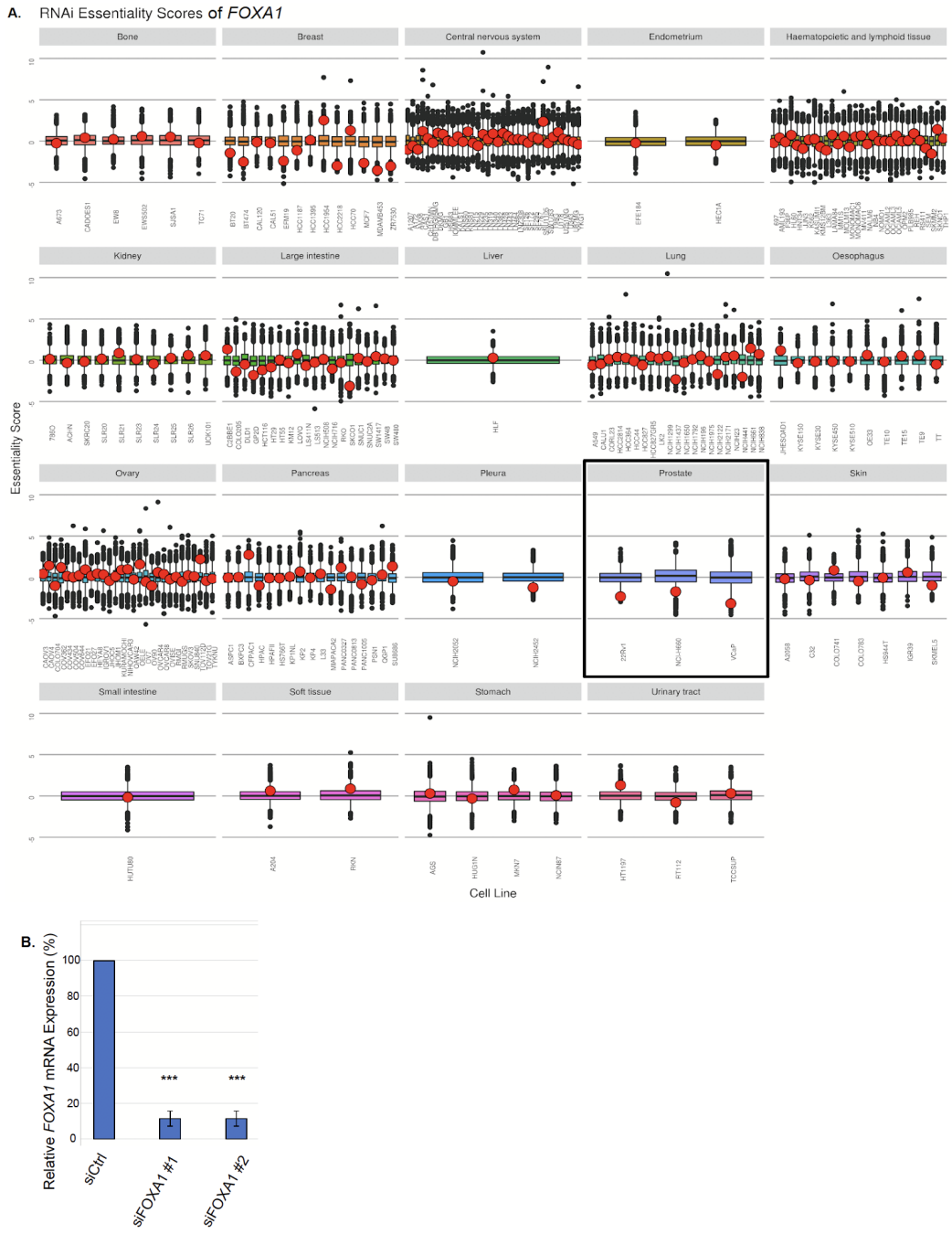

Figure A.3: **Essentiality of *FOXA1* across cancer cell lines of various cancer types. a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot indicates *FOXA1*. **b.** *FOXA1* mRNA expression normalized to housekeeping TBP mRNA expression upon siRNA-mediated knockdown, five days post-transfection ($n = 3$ independent experiments). Error bars indicate ± s.d, Student's t-test, *** $p < 0.001$.

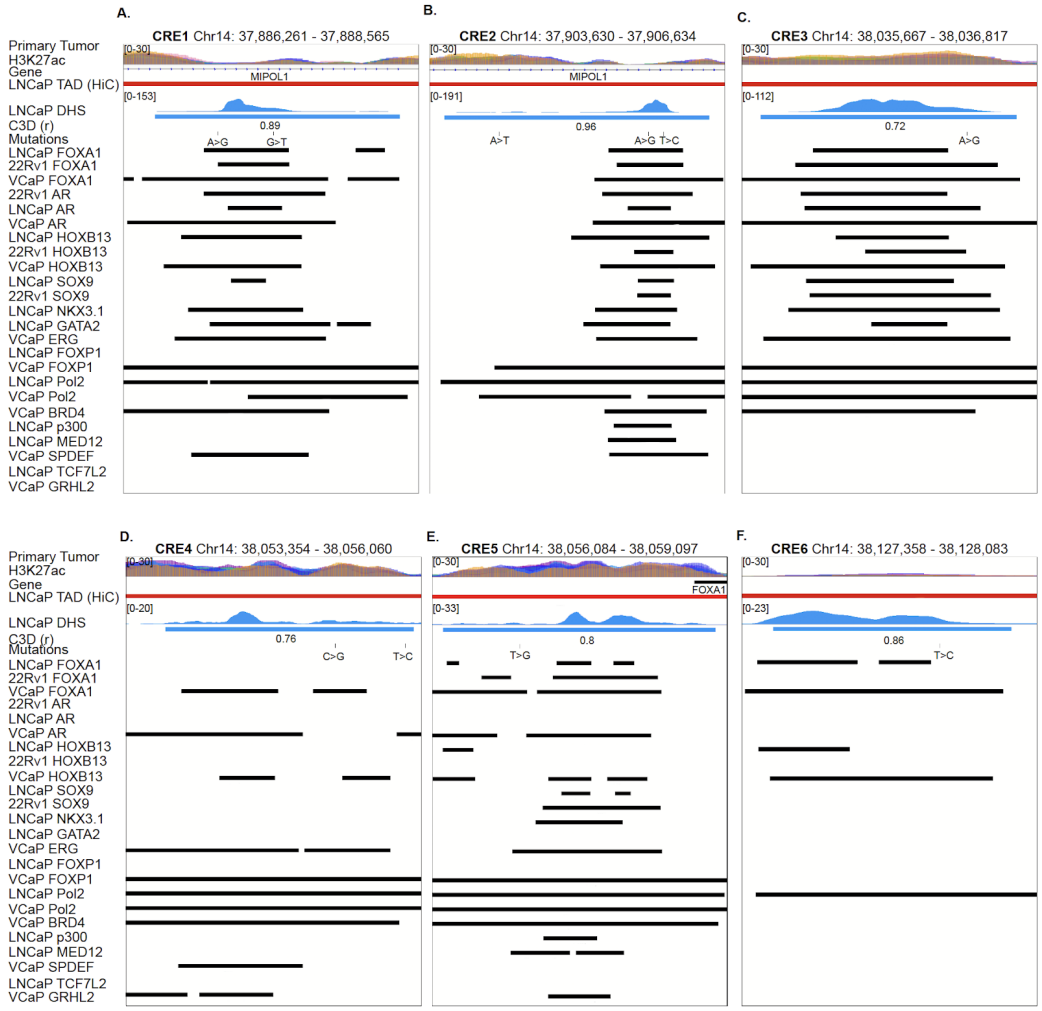

Figure A.4: **Visualization of the functional annotation of the six *FOXA1* CREs. a-f.** Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

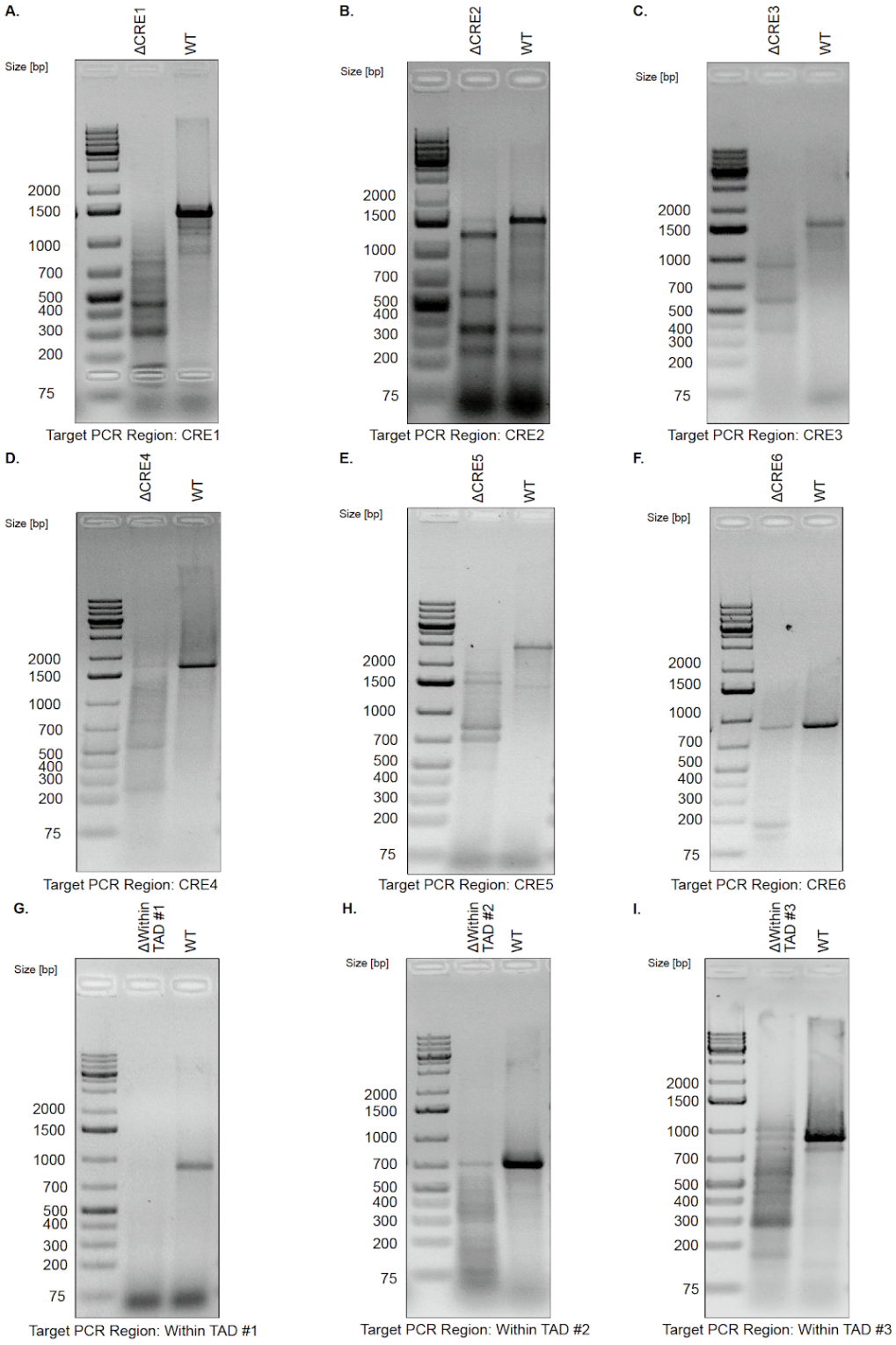

Figure A.5: **Validation of clonal Cas-mediated deletions of CREs. a-f**. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

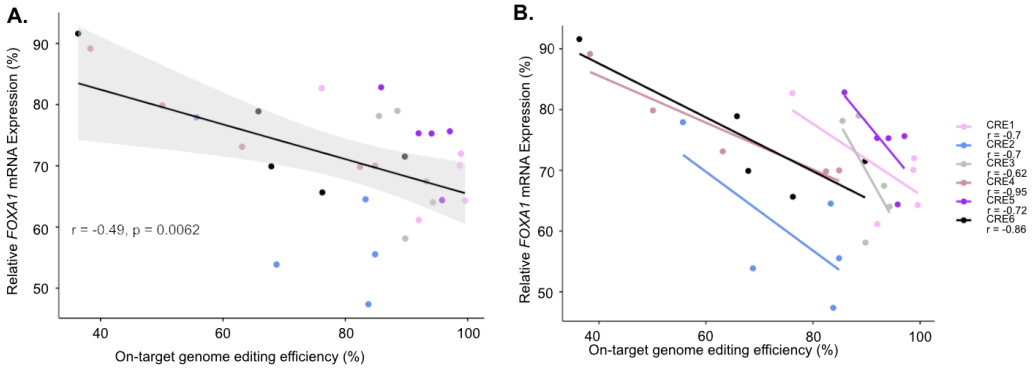

Figure A.6: **Genome editing efficiency (%) is inversely correlated with *FOXA1* mRNA expression. a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all of the CREs. **b.** Pearson's correlation based on each individual CRE, correlation between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells.

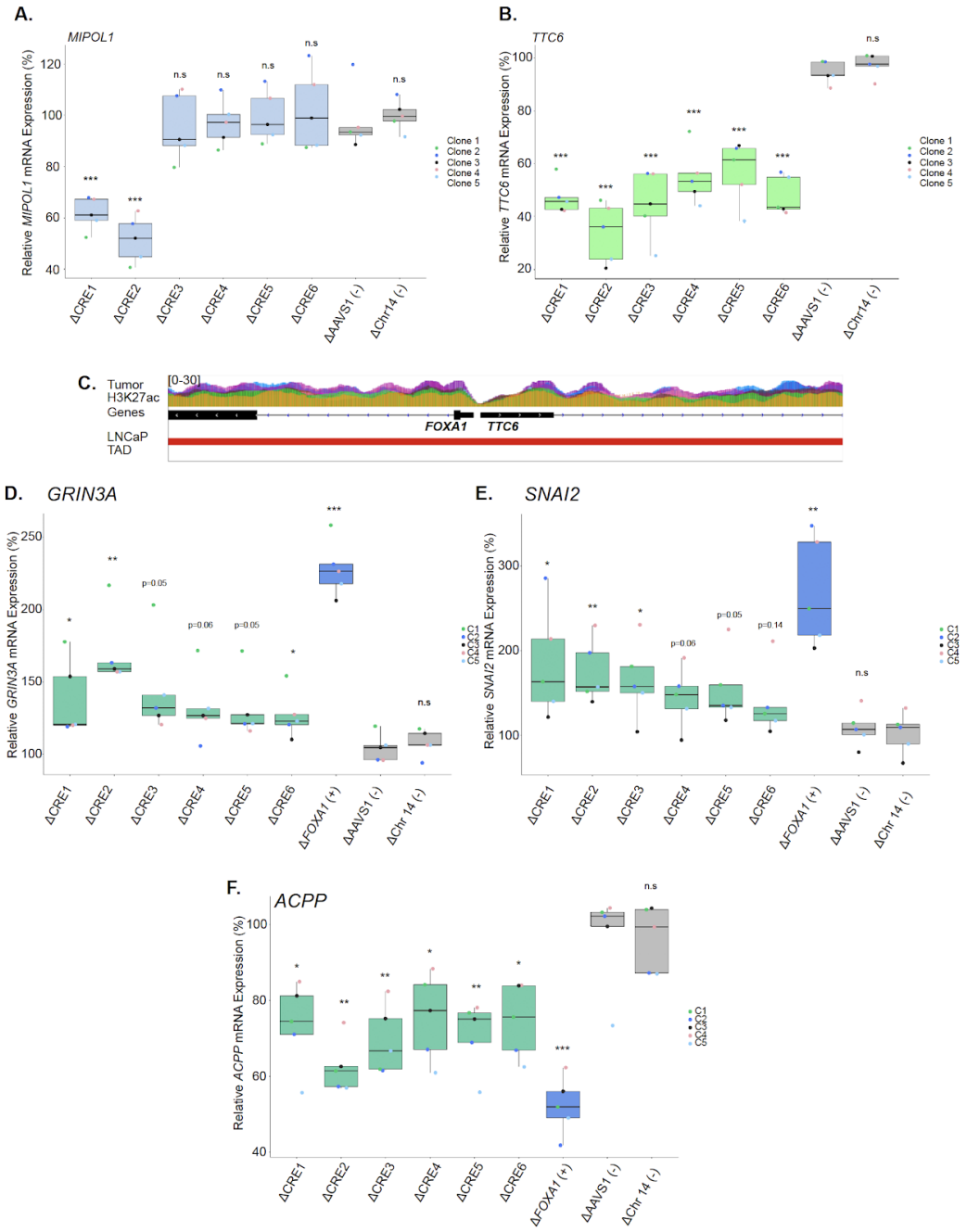

Figure A.7: **Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a.** *MIPOL1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. **b.** *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each CRE. **c.** Zoom-in view of the *FOXA1* and *TTC6* locus. **d-f.** mRNA expression of *GRIN3A*, *SNAI2* and *ACPP* normalized to housekeeping gene *TBP* upon deletion of each region of interest. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments, each dot represents an independent clone). Error bars indicate ± s.d. Student's *t*-test, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

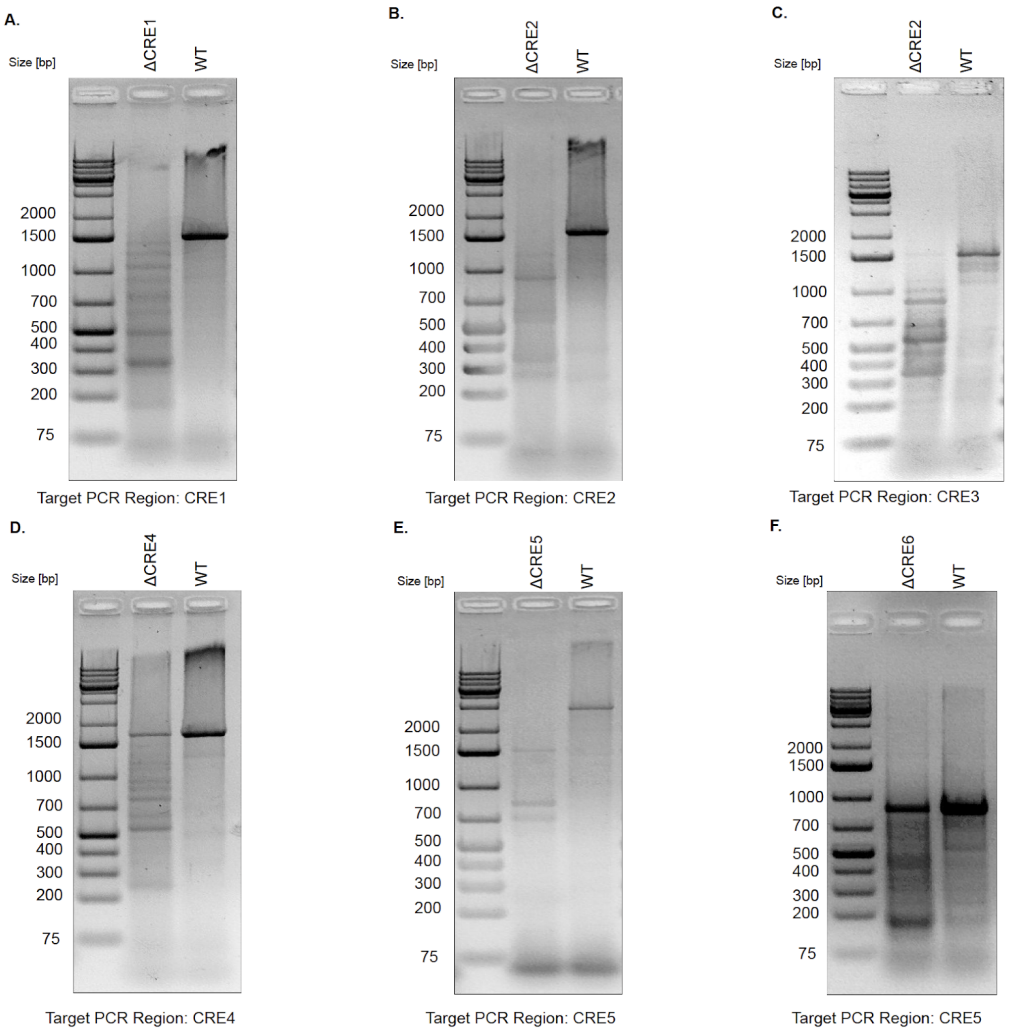

Figure A.8: **Validation of transient Cas9-mediated single deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

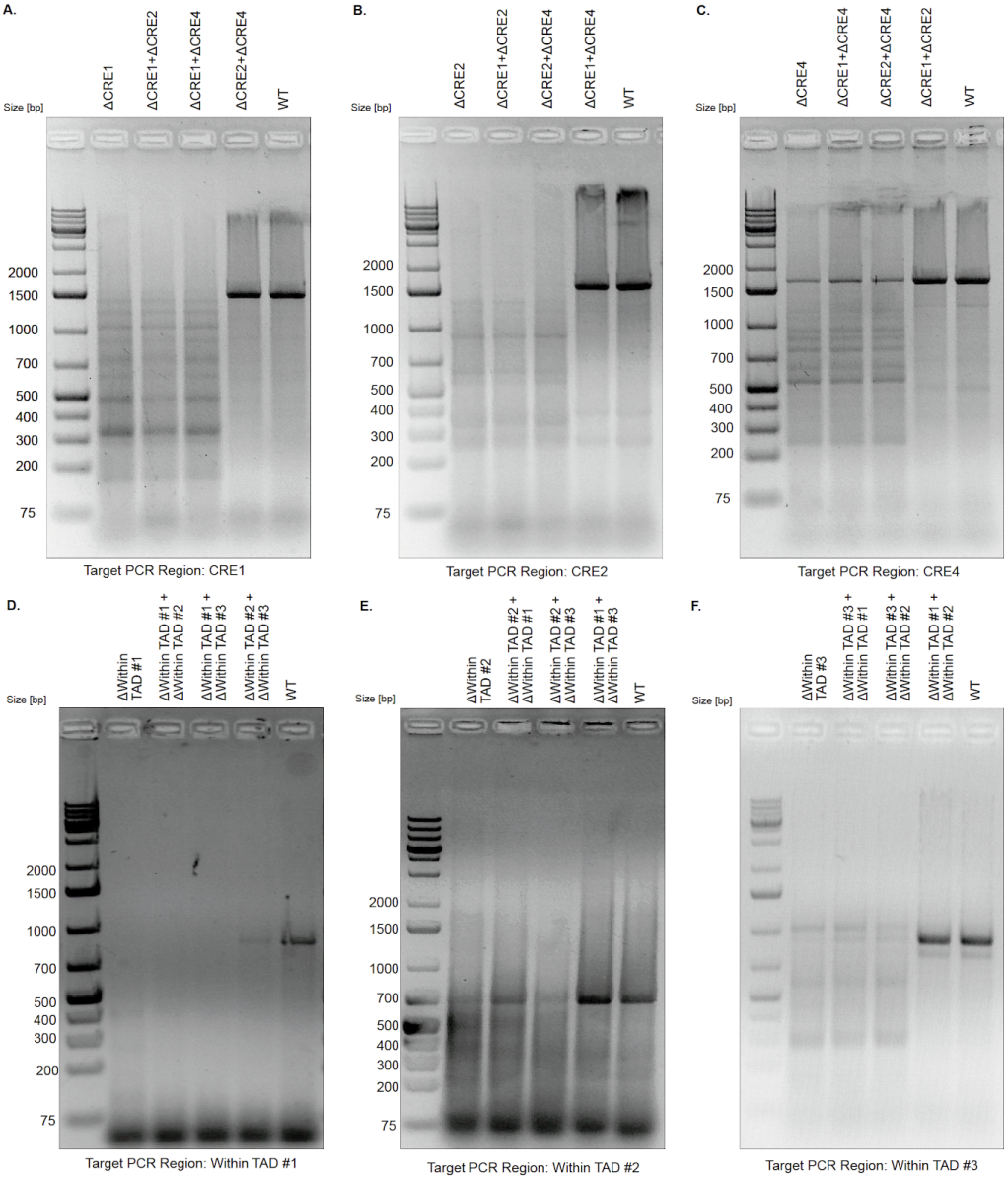

Figure A.9: **Validation of transient Cas9-mediated double deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

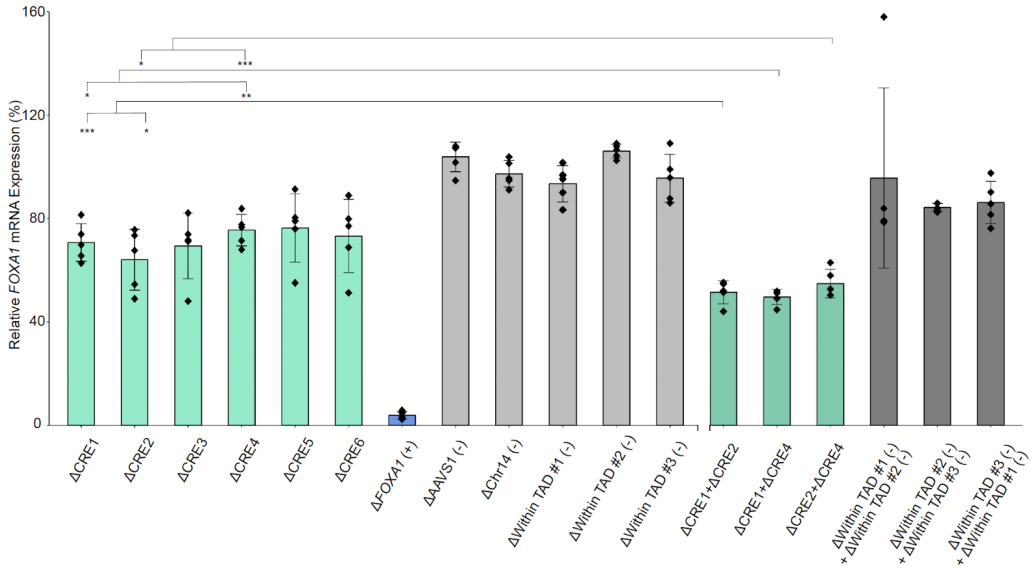

Figure A.10: **Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s).** *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate ± s.d., Student's t-test, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

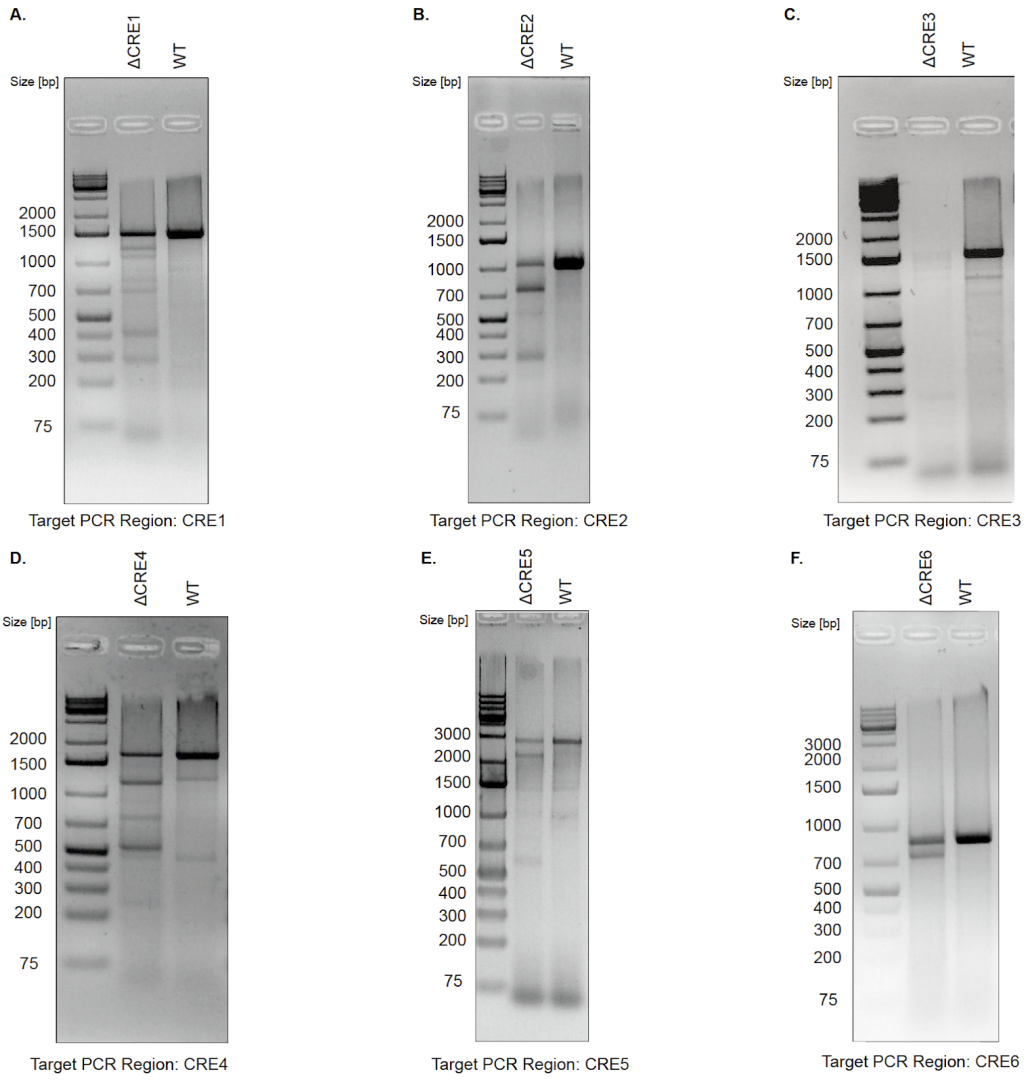

Figure A.11: **Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f.** Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

Appendix B

Supplementary Material for Chapter 3

Table B.1 Clinical information of samples involved in this study.

Table B.2 Sequencing metrics as calculated by HiCUP for all Hi-C libraries generated in this study.

Table B.3 Summary statistics for TAD counts in all 12 tumour and 5 benign samples, across multiple window sizes.

Table B.4 Individual TAD calls in all 12 tumour and 5 benign samples.

Table B.5 Detected chromatin interactions in all 12 tumour and 5 benign samples.

Table B.6 SV breakpoints detected by Hi-C in each tumour sample.

Table B.7 Simple and complex SVs reconstructed from SV breakpoints.

Table B.8 H3K27ac peaks identified in each of the 12 primary PCa patients.

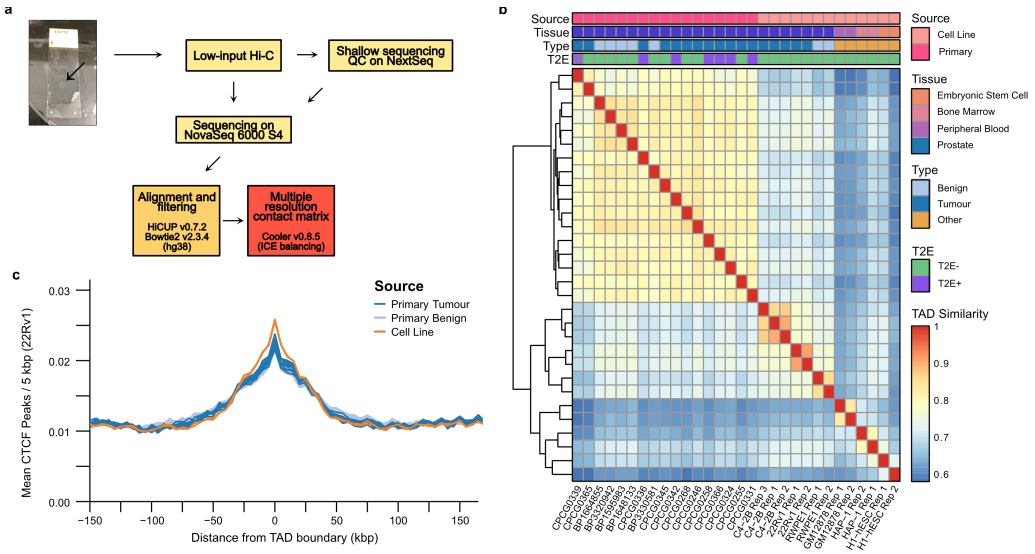

Figure B.1: **Sample processing and TAD similarity between samples. a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate cell lines. Median similarity scores between TADs in primary prostate tissues and cell lines is 72.1%, 66.9% between prostate and non-prostate cell lines, and 63.5% between primary prostate and non-prostate lines. **c.** Local enrichment of CTCF binding sites from the 22Rv1 PCa cell line around TAD boundaries identified in the primary samples.

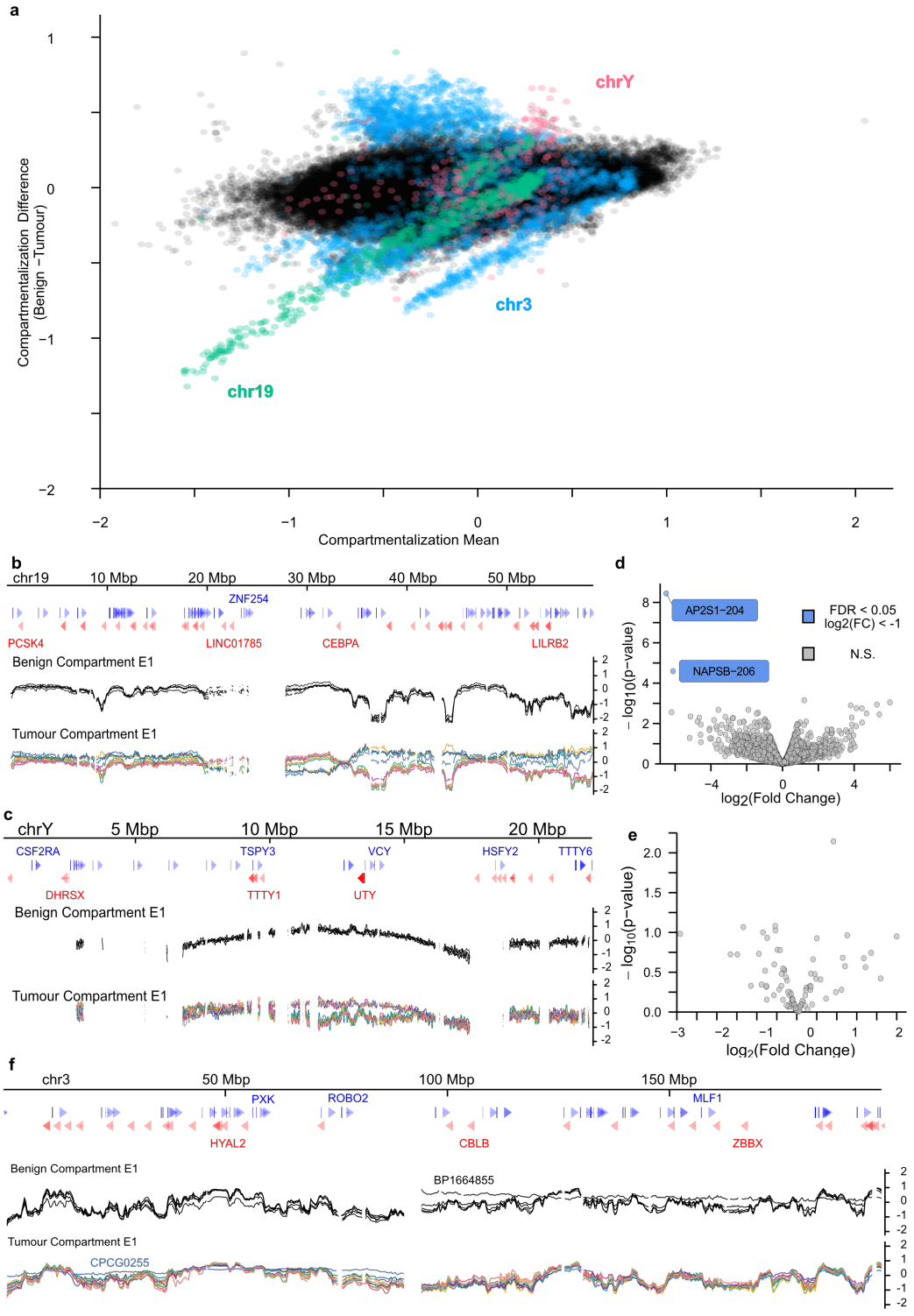

Figure B.2: **Compartmentalization changes in tumours is not associated with widespread differential gene expression.** (Continued on the following page)

Figure B.2: **a.** Bland-Altman plot of the mean compartmentalization score between tumour and benign samples. Chromosomes 3, 19, and Y are highlighted for their consistent deviation between the tissue types. **b-c.** Compartmentalization genome tracks across chromosomes 19 (**b**) and Y (**c**) in all primary samples. **d-e.** Volcano plot of differential transcript expression between the tumour samples with benign-like compartmentalization and altered compartmentalization in chromosomes 19 (**d**) and Y (**e**). Grey dots are transcripts without significant differential expression, blue dots are differentially expressed transcripts ($FDR < 0.05$) that are under-expressed in the altered compartment samples. **f.** Compartmentalization genome tracks across chromosome 3.

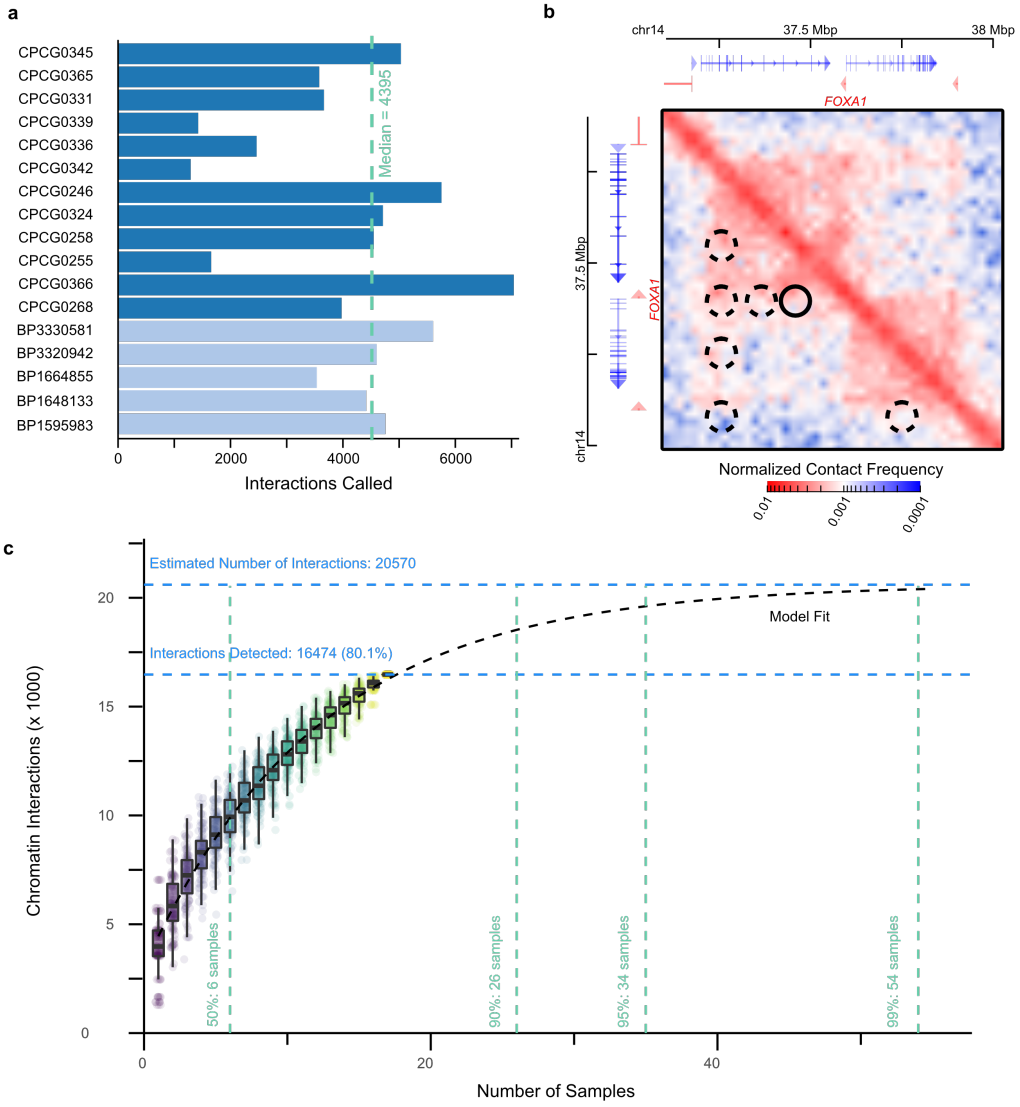

Figure B.3: **Characterization of chromatin interactions in benign and tumour tissue. a.** Bar plot of the number of significant chromatin interactions identified in each of the primary prostate samples. **b.** A snapshot of significant chromatin interactions called around the *FOXA1* gene. Identified interactions are highlighted as circles. The interaction marked by the solid border contains two CREs of *FOXA1* identified in Zhou *et al.*, 2020 (listed in that publication as CRE1 and CRE2). The interactions marked by the dashed border indicate regions of increased contact that may contain more distal CREs of *FOXA1*. **c.** Saturation analysis of chromatin interactions detected in our cohort of prostate samples versus the theoretical estimation obtained through asymptotic estimation from bootstraps. Boxplots show the first, second, and third quartiles of the identified interactions across the bootstrap iterations. The dashed black line corresponds to the asymptotic model of estimated mean unique interactions obtained from an increasing number of samples. Horizontal blue dashed lines indicate the number of observed unique interactions and theoretical maximum. Vertical green dashed lines indicate the number of samples required to reach as estimated 50%, 90%, 95%, and 99% of the theoretical maximum.

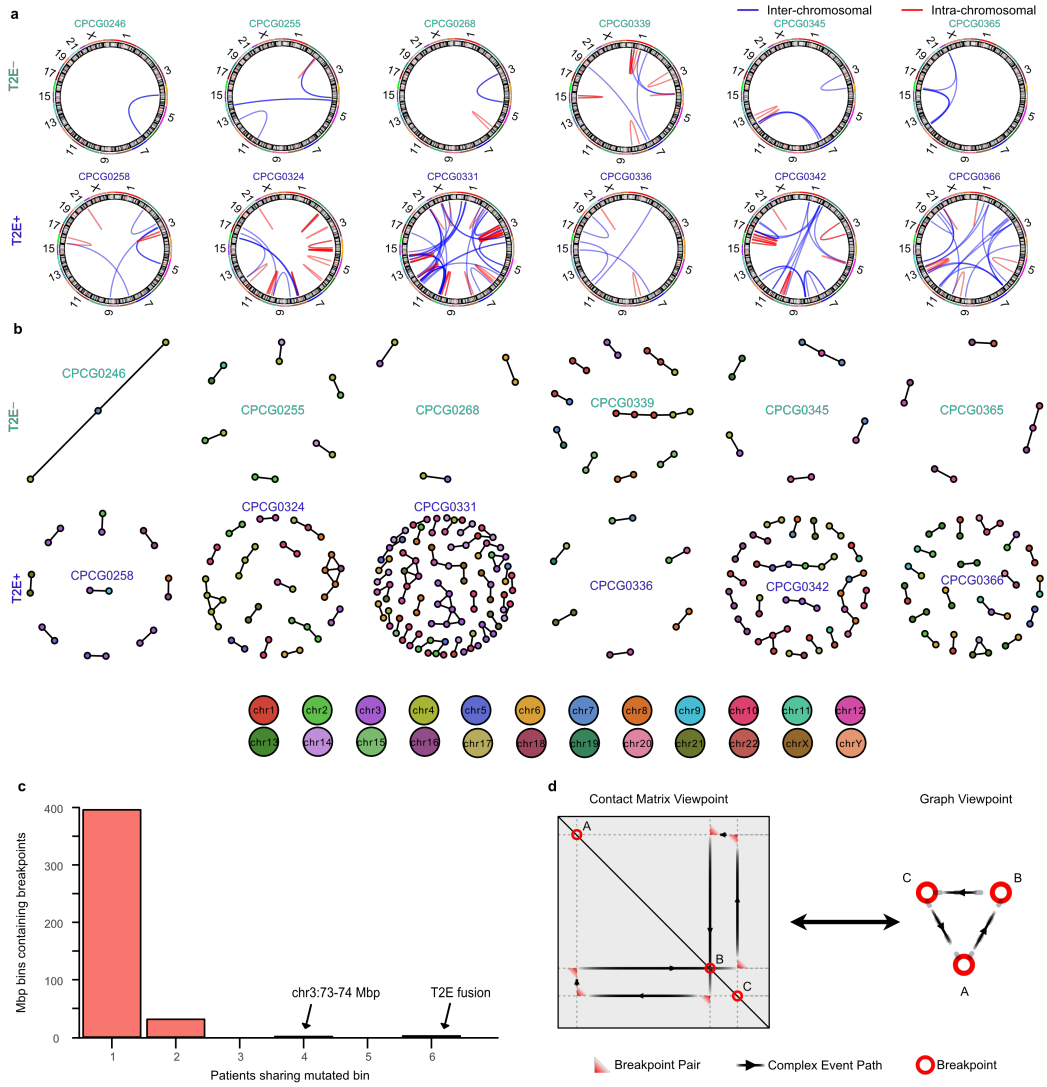

Figure B.4: **Structural variant detection from Hi-C data. a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1 Mbp bins with SV breakpoints from multiple patients. The previously-reported highly-mutated regions on chr3 and T2E fusion are highlighted. **d.** Correspondence between the breakpoint representation in the contact matrices and a graph representation. Each node represents a breakpoint and each edge determines whether the breakpoints were directly in contact, as identified by the Hi-C contact matrix.

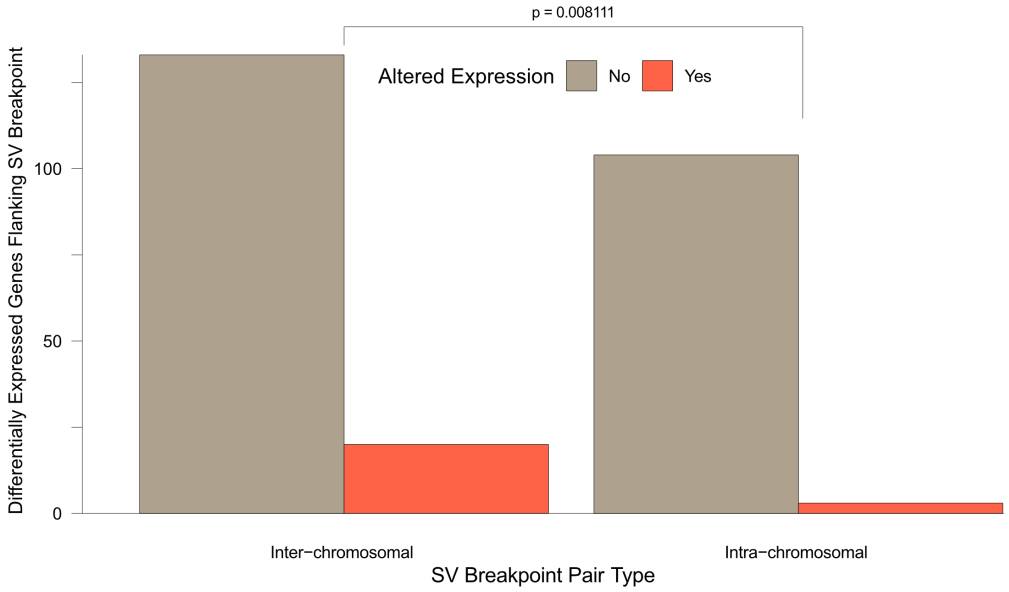

Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes.

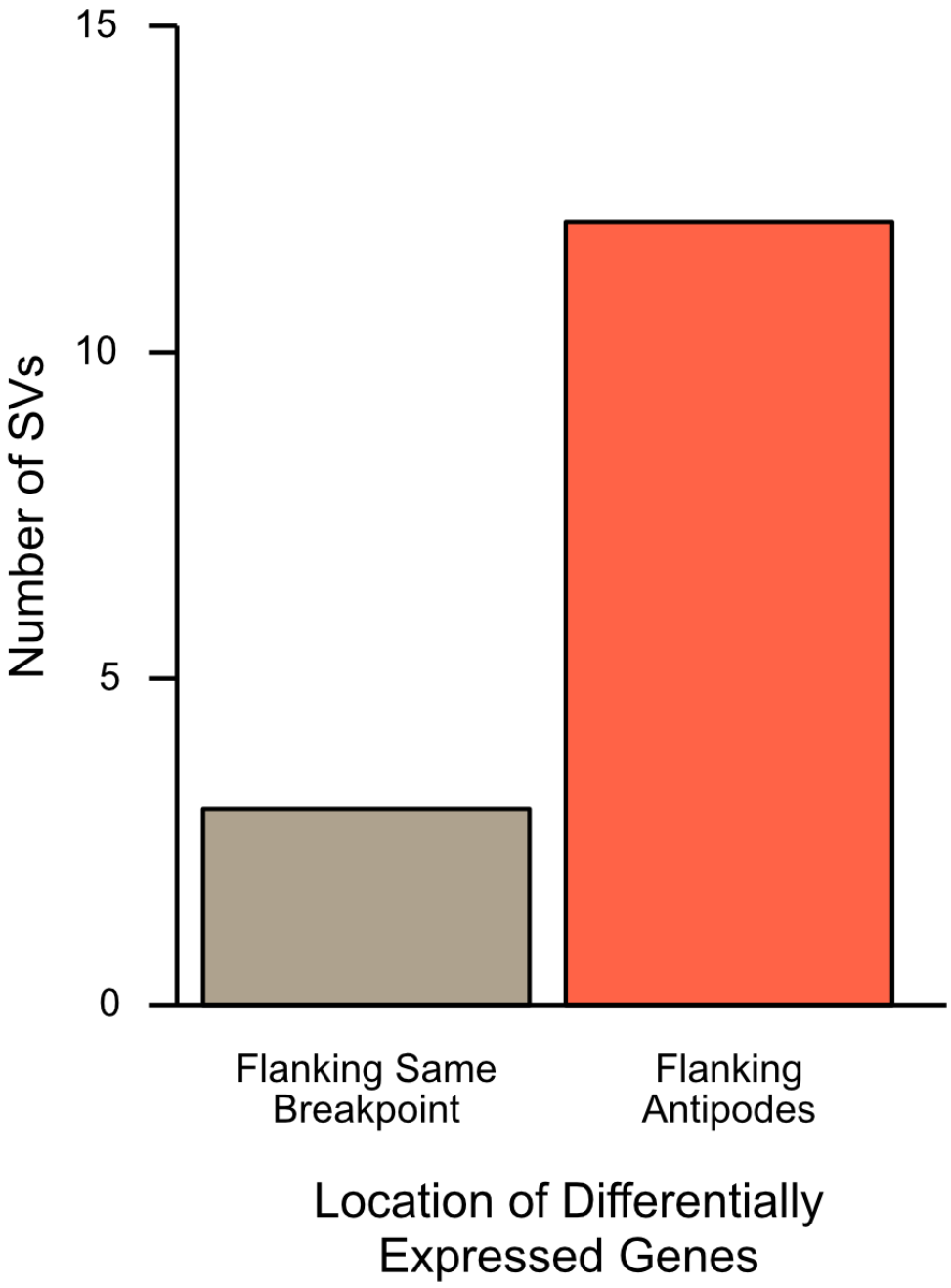

Figure B.6: **Location of differentially expressed genes around SV breakpoints.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

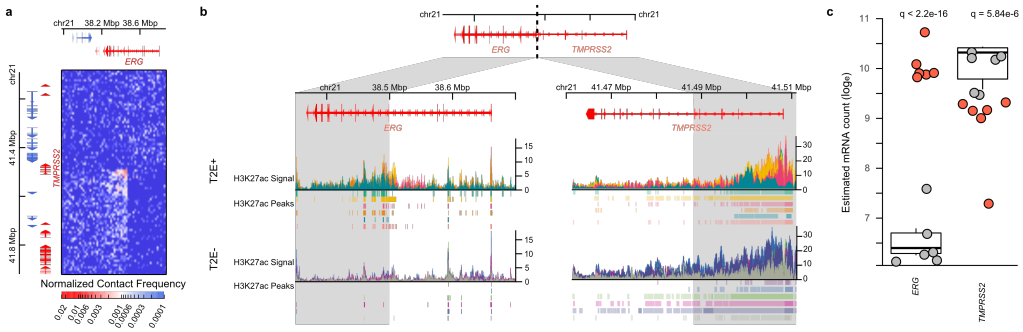

Figure B.7: **Chromatin organization of the *TMPRSS2-ERG* fusion. a.** Contact matrix of the deletion between *TMPRSS2* and *ERG*. **b.** Genome tracks of H3K27ac ChIP-seq signal in T2E+ and T2E- patients. The grey region highlights the loci that come into contact as a result of the deletion. **c.** Expression of *TMPRSS2* and *ERG* genes. Boxplots represent first, second, and third expression quartiles of T2E- patients (grey dots). T2E+ patients are represented by red dots.

Appendix C

Supplementary Material for Chapter 4

C.1 Differential expression analysis with Sleuth

The differential expression model employed in the Sleuth (v0.30.0) [24, 25] can be described as follows. Consider a set of transcripts, S, measured in N samples with an experimental design matrix, $X \in \mathbb{R}^{N \times p}$, where p is the number of covariates considered. Let Y_{si} be the natural log of the abundance of transcript s in sample i. Given the design matrix

$$X = [x_1^T; x_2^T; ...x_n^T], x_i \in \mathbb{R}^p$$

the abundance of transcripts can be modelled as a generalized linear model (GLM)

$$Y_{si} = x_i^T \beta_s + \epsilon_{si} \tag{C.1}$$

where $\epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2)$ is the biological noise of transcript s in sample i and $B_s \in \mathbb{R}^p$ is the fixed effect of the covariates on the expression of transcript s.

Due to inferential noise from sequencing, each Y_{si} are not observed directly, but indirectly through the observed perturbations, D_{si}. This can be modelled as

$$D_{si}|Y_{si} = Y_{si} + \zeta_{si} \tag{C.2}$$

where $\zeta_{si} \sim \mathcal{N}(0, \tau_s^2)$ is the inferential noise of transcript s in sample i. Both biological and inferential noise for each transcript are independent and identically distributed (IID) and independent of each other. Namely:

$$\begin{aligned}
\mathbb{C}ov[\epsilon_{si}, \epsilon_{rj}] &= \sigma_s^2 \delta_{i,j} \delta_{s,r} \\
\mathbb{C}ov[\zeta_{si}, \zeta_{rj}] &= \tau_s^2 \delta_{i,j} \delta_{s,r} \\
\mathbb{C}ov[\epsilon_{si}, \zeta_{rj}] &= 0 \\
&\forall s, r \forall i, j
\end{aligned}$$

The abundances for transcript s in all N samples can then modelled as a multivariate normal distribution

$$D_s | Y_s \sim \mathcal{N}_N(X\beta_s, (\sigma_s^2 + \tau_s^2) I_N) \tag{C.3}$$

where $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix.

The goal of the differential analysis is to estimate the $|S| \times p$ coefficients in $B_s \forall s \in S$, and to determine which coefficients differ significantly from 0. This is achieved through a Wald test or likelihood ratio test after estimating the inferential variance, τ_s^2, through bootstrapping and the biological variance, σ_s^2, through dispersion estimation and shrinkage.

The estimator for the differential effect is the ordinary least squares (OLS) estimate:

$$\hat{\beta}_s = (X^T X)^{-1} X^T d_s$$

where d_s is the observed abundances given by

$$\begin{aligned}
d_{si} &= \ln\left(\frac{k_{si}}{\hat{f}_i} + 0.5\right) \\
\hat{f}_i &= \underset{s \in S^*}{\text{median}} \frac{k_{si}}{\sqrt[N]{\prod_{j=1}^{N} k_{sj}}}
\end{aligned}$$

where k_{si} is the estimated read count from the Kallisto package (v0.46.1) [26] for transcript s in

sample i and $\hat{f}_i$ is the scaling factor for sample i, calculated from the set of all transcripts that pass initial filtering, S^*.

C.2 Statistical moments of the ordinary least squares estimator

As shown in Supplementary Note 2 of [REF 24], the estimator is unbiased, Namely

$$\mathbb{E}\left[\hat{\beta}_s^{(OLS)}\right] = B_s \tag{C.4}$$

It can also be shown that, for a covariance matrix Σ,

$$\mathbb{V}\left[\hat{\beta}_s^{(OLS)}\right] = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

In the case where $\Sigma = (\sigma_s^2 + \tau_s^2) I_N$, this reduces to

$$\mathbb{V}\left[\hat{\beta}_s{}^{(OLS)}\right] = (\sigma_s^2 + \tau_s^2)(X^T X)^{-1}$$

Consider a simple experimental design where the only covariate of interest is the presence of a mutation. Then the design matrix, with the first column being the intercept and the second being the mutation status, looks like so:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times 2}$$

The variance of the OLS estimator is then

$$\mathbb{V}\left[\hat{\beta}_s{}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)}{n_{mut} n_{wt}} \begin{bmatrix} n_{mut} & -n_{mut} \\ -n_{mut} & n_{mut} + n_{wt} \end{bmatrix}$$

Importantly, the estimate for the coefficient measuring the effect that the presence of the mutation has variance

$$\mathbb{V}\left[\beta_{s,mut}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)(n_{mut} + n_{wt})}{n_{mut} n_{wt}}$$

When there is only 1 mutated sample, as per the motivation of this work, this reduces to

$$\mathbb{V}\left[\beta_{s,mut}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)(1 + n_{wt})}{n_{wt}} \tag{C.5}$$

C.3 Statistical moments of the James-Stein estimator

C.3.1 Expected value of the James-Stein estimator

We can use a Taylor expansion around B_1 to approximate the expectated value of $\hat{\mathrm{B}}_1^{(JS)}$. Consider:

$$\hat{\mathrm{B}}_1^{(JS)} = \left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T \Sigma^{-1} \hat{\mathrm{B}}_1^{(OLS)}}\right) \hat{\mathrm{B}}_1^{(OLS)}$$

where

$$\hat{\mathrm{B}}_1^{(OLS)} \sim N_{|S|}(\mathrm{B}_1, \Sigma)$$
$$\Sigma_{s,t} = \begin{cases} \left(\frac{n_{wt}+1}{n_{wt}}\right)(\sigma_s^2 + \tau_s^2) & s = t \\ 0 & s \neq t \end{cases}$$

Let $u = \Sigma^{-1/2}\hat{\mathrm{B}}_1^{(OLS)}$. Then

$$\begin{aligned} \mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right] &= \mathbb{E}\left[\hat{\mathrm{B}}_1^{(OLS)}\right] - c\Sigma^{1/2}\mathbb{E}\left[\frac{u}{\|u\|^2}\right] \\ &= \mathrm{B}_1 - c\Sigma^{1/2}\mathbb{E}\left[\frac{u}{\|u\|^2}\right]\Sigma^{1/2} \end{aligned}$$

Expanding $\frac{u}{\|u\|^2}$ around $a = \Sigma^{-1/2}\mathrm{B}_1$ gives:

$$\begin{aligned} \mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right] &= \mathrm{B}_1 - c\Sigma^{1/2}\mathbb{E}\left[\frac{a}{\|a\|^2} + \left(\frac{1}{\|a\|^2} - \frac{2}{\|a\|^4}aa^T\right)(u-a) + \mathcal{O}(\|u-a\|^2)\right] \\ &= \left(1 - \frac{c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right)\mathrm{B}_1 + \mathcal{O}(\|u-a\|^2) \end{aligned}$$

As long as the number of transcripts being considered, $|S|$, is not large, and that the true coefficient of variation is not large (i.e. that $\|u-a\|^2 \ll \|\mathrm{B}_1\|^2$), the Taylor approximation is close to

$$\mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right] \approx \left(1 - \frac{c}{\mathrm{B}_1^T \Sigma^{-1} \mathrm{B}_1}\right) \mathrm{B}_1 \tag{C.6}$$

Thus the James-Stein (JS) estimator is an estimate of B_1 that is biased towards 0.

C.3.2 Variance of the James-Stein estimator

The mean square error (MSE) of the JS estimator is related to its variance.

$$\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] = \sum_{s\in S} \mathbb{E}\left[\left(\hat{\mathrm{B}}_{1,s}^{(JS)} - \mathrm{B}_{1,s}\right)^2\right] = \sum_{s\in S} \mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(JS)}\right]$$

By [REF 27], $\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] \leq \mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(OLS)} - \mathrm{B}_1\|^2\right]$. However, this does not imply that $\mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(JS)}\right] \leq \mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(OLS)}\right] \forall s \in S$. Some transcripts may have larger variances than the OLS estimator, but all transcripts in aggregate will have a smaller MSE. This is still desirable if the goal is to find if there is an effect on any transcripts in the set S, instead of a particular one within the set.

To calculate the variance for each individual transcript, a similar approach with Taylor expansions can be used, as above.

$$\begin{aligned}
&\mathbb{V}\left[\hat{\mathrm{B}}_1^{(JS)}\right] \\
&\approx \mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\left(\hat{\mathrm{B}}_1^{(JS)}\right)^T\right] - \left(1 - \frac{c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right)^2 \mathrm{B}_1\mathrm{B}_1^T \\
&= \Sigma^{1/2}\mathbb{E}\left[uu^T - \frac{2c}{u^Tu}uu^T + \left(\frac{c}{u^Tu}\right)^2 uu^T\right]\Sigma^{1/2} - \left(1 - \frac{c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right)^2 \mathrm{B}_1\mathrm{B}_1^T
\end{aligned}$$

where, again, $u = \Sigma^{-1/2}\hat{\mathrm{B}}_1^{(OLS)}$. Expanding about $a = \Sigma^{-1/2}\mathrm{B}_1$ yields:

$$\mathbb{V}\left[\hat{\mathrm{B}}_1^{(JS)}\right] = \left(1 - \frac{2c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right)\Sigma - \frac{2c}{\left(\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^T + \mathcal{O}(\|u-a\|^4)$$

Under similar conditions of the number of transcripts under consideration, $|S|$, and $\|u-a\|^2$, we then have that

$$\mathbb{V}\left[\hat{\mathrm{B}}_1^{(JS)}\right] \approx \left(1 - \frac{2c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right)\Sigma - \frac{2c}{\left(\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^T \tag{C.7}$$

Since the diagonal elements of $\frac{2c}{\left(\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^T$ are all ≥ 0 and $0 \leq \left(1 - \frac{2c}{\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1}\right) \leq 1 \forall c > 0$, the variance than of the JS estimators are smaller than the OLS estimators. The resulting Wald test statistics for the fold change coefficient of transcript s in the OLS and JS cases can be summarized as follows:

$$W_s^{(OLS)} = \frac{\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2}{\Sigma_{s,s}} \tag{C.8}$$

$$W_s^{(JS)} = \frac{\left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T\Sigma^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right)^2\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2}{\left(1 - \frac{2c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T\Sigma^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right)\Sigma_{s,s} - \frac{2c}{\left(\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T\Sigma^{-1}\hat{\mathrm{B}}_1^{(OLS)}\right)^2}\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2} \tag{C.9}$$

The coefficient of $\hat{\mathrm{B}}_{1,s}^{(OLS)}$ in the numerator is larger than the coefficient of Σ in the denominator since $(1-a)^2 = 1 - 2a + a^2 > 1 - 2a \forall a \in \mathbb{R}$. This implies that the Wald test statistics will be larger for the JS estimator than for the OLS estimator. Thus the JS method will produce more positive calls, in general, than the OLS method.

Notably, the variance of the JS estimator is a function of both the mean and variance of the transcripts under consideration. This is in contrast to the OLS estimator, which is solely a function of the variance. Additionally, the off-diagonal elements of the matrix $\mathrm{B}_1\mathrm{B}_1^T$ imply that the JS fold change estimates are not independent of each other. This, again, contrasts with the OLS estimator, where the diagonal covariance matrix, Σ, implies that the fold change estimates are themselves independent of each other. The effect of this dependence on statistical inference is a function of the variance and true fold change, as can be seen from the $\frac{2c}{\left(\mathrm{B}_1^T\Sigma^{-1}\mathrm{B}_1\right)^2}$ coefficient. While rarely true in practice, this statistical dependence can affect the results of statistical inference, in theory. For most purposes, is not expected to have a large effect on the results of statistical inference.

Appendix D

Supplementary Material for Chapter 5

Glossary

3C chromatin conformation capture

ANOVA Analysis of Variance

AR androgen receptor

ATAC-seq assay for transposase-accessible chromatin sequencing

B-ALL B-cell acute lymphoblastic leukemia

cDNA complementary DNA

ChIP-seq chromatin immunoprecipitation sequencing

CMP common myeloid progenitor

CPC-GENE Canadian Prostate Cancer Genome Network

CpG CG dinucleotide

crRNA CRISPR RNA

CRE *cis*-regulatory element

DEPMAP Cancer Dependency Map

DHS DNase I hypersensitive sites

DMR differentially methylated region

DNAme DNA methylation

dRI disease relapse-initiating

Dx diagnosis

EarlyProB early progenitor B cell

FDR false discovery rate

FN false negative

FP false positive

FOX forkhead box

GLM generalized linear model

GMP granulocyte-macrophage progenitor

GO gene ontology

gRNA guide RNA

HSC hematopoietic stem cell

HSPC hematopoietic stem and progenitor cell

IID independent and identically distributed

JS James-Stein

kbp kilobase

KO knockout

LDA limiting dilution assay

LMPP lymphoid-primed multi-potent progenitor

MeCapSeq DNA methylation capture sequencing

MEP megakaryocyte-erythrocyte progenitor

MSE mean square error

mCRPC metastatic castration-resistant prostate cancer

MLP monocyte-lymphoid progenitor

MPP multi-potent progenitor

NSG NOD scid gamma

OLS ordinary least squares

mRNA messenger RNA

PCa prostate cancer

PDX patient-derived xenograft

PreProB pre-progenitor B cell

ProB progenitor B cell

Rel relapse

RNAi RNA interference

RNA-seq RNA sequencing

shRNA small hairpin RNA

siRNA small interfering RNA

SNV single nucleotide variants

SRA Seqeuence Read Archive

SNF similarity network fusion

SV structural variant

TAD topologically associated domain

TCGA The Cancer Genome Atlas

TN true negative

TP true positive

TF transcription factor

tracrRNA trans-activating CPRISR RNA

UTR untranslated region

WGS whole genome sequencing

WT wild-type

References

1. Dobin, A. *et al.* STAR: Ultrafast Universal RNA-Seq Aligner. en. *Bioinformatics* **29,** 15–21. ISSN: 1460-2059, 1367-4803 (Jan. 2013).

2. Dobson, S. M. *et al.* Relapse-Fated Latent Diagnosis Subclones in Acute B Lineage Leukemia Are Drug Tolerant and Possess Distinct Metabolic Programs. en. *Cancer Discovery* **10,** 568–587. ISSN: 2159-8274, 2159-8290 (Apr. 2020).

3. Lee, S.-T. *et al.* A Global DNA Methylation and Gene Expression Analysis of Early Human B-Cell Development Reveals a Demethylation Signature and Transcription Factor Network. en. *Nucleic Acids Research* **40,** 11339–11351. ISSN: 0305-1048 (Dec. 2012).

4. Lee, S.-T. *et al.* Epigenetic Remodeling in B-Cell Acute Lymphoblastic Leukemia Occurs in Two Tracks and Employs Embryonic Stem Cell-like Signatures. en. *Nucleic Acids Research* **43,** 2590–2602. ISSN: 1362-4962, 0305-1048 (Mar. 2015).

5. Nordlund, J. *et al.* Genome-Wide Signatures of Differential DNA Methylation in Pediatric Acute Lymphoblastic Leukemia. *Genome Biology* **14,** r105. ISSN: 1474-760X (Sept. 2013).

6. Notta, F. *et al.* Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment. en. *Science* **333,** 218–221. ISSN: 0036-8075, 1095-9203 (July 2011).

7. Mazurier, F., Doedens, M., Gan, O. I. & Dick, J. E. Rapid Myeloerythroid Repopulation after Intrafemoral Transplantation of NOD-SCID Mice Reveals a New Class of Human Stem Cells. en. *Nature Medicine* **9,** 959–963. ISSN: 1078-8956, 1546-170X (July 2003).

8. Hu, Y. & Smyth, G. K. ELDA: Extreme Limiting Dilution Analysis for Comparing Depleted and Enriched Populations in Stem Cell and Other Assays. en. *Journal of Immunological Methods* **347,** 70–78. ISSN: 00221759 (Aug. 2009).

9. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nature Methods* **10,** 1213–8 (Dec. 2013).

10. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python Framework to Work with High-Throughput Sequencing Data. en. *Bioinformatics* **31,** 166–169. ISSN: 1367-4803, 1460-2059 (Jan. 2015).

11. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474-760X (Dec. 2014).

12. Langmead, B. & Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. en. *Nature Methods* **9,** 357–359. ISSN: 1548-7105 (Apr. 2012).

13. Zhang, Y. *et al.* Model-Based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137. ISSN: 1474-760X (Sept. 2008).

14. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis: BEDTools: The Swiss-Army Tool for Genome Feature Analysis. en. *Current Protocols in Bioinformatics* **47,** 11.12.1–11.12.34. ISSN: 19343396 (Sept. 2014).

15. Simon Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data* 2010.

16. Felix Krueger. *Trim Galore* Mar. 2012.

17. Krueger, F., Kreck, B., Franke, A. & Andrews, S. R. DNA Methylome Analysis Using Short Bisulfite Sequencing Data. *Nature Methods* **9,** 145–151. ISSN: 1548-7105 (Electronic)\r1548-7091 (Linking) (Jan. 2012).

18. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast Processing of NGS Alignment Formats. en. *Bioinformatics* **31,** 2032–2034. ISSN: 1367-4803 (June 2015).

19. Ryan, D. P. *MethylDackel* Apr. 2019.

20. Wang, B. *et al.* Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. en. *Nature Methods* **11,** 333–337. ISSN: 1548-7091, 1548-7105 (Mar. 2014).

21. Korthauer, K., Chakraborty, S., Benjamini, Y. & Irizarry, R. A. Detection and Accurate False Discovery Rate Control of Differentially Methylated Regions from Whole Genome Bisulfite Sequencing. en. *Biostatistics.* ISSN: 1465-4644, 1468-4357 (Feb. 2018).

22. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300. ISSN: 0035-9246 (1995).

23. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-Scale Gene Function Analysis with the PANTHER Classification System. en. *Nature Protocols* **8,** 1551–1566. ISSN: 1754-2189, 1750-2799 (Aug. 2013).

24. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty. en. *Nature Methods* **14,** 687–690. ISSN: 1548-7105 (July 2017).

25. Yi, L., Pimentel, H., Bray, N. L. & Pachter, L. Gene-Level Differential Analysis at Transcript-Level Resolution. *Genome Biology* **19,** 53. ISSN: 1474-760X (Apr. 2018).

26. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-Optimal Probabilistic RNA-Seq Quantification. en. *Nature Biotechnology* **34,** 525–527. ISSN: 1546-1696 (May 2016).

27. Bock, M. E. Minimax Estimators of the Mean of a Multivariate Normal Distribution. en. *The Annals of Statistics* **3,** 209–218. ISSN: 0090-5364 (Jan. 1975).