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

© Copyright 2021 by James Hawley

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 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
6. Something about DNAme shaping hematopoietic differentiation and the epigenetic factors related to cell type [1, 2]

1.3 Results

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

To investigate the molecular landscape of B-ALL relapse, we profiled gene expression, chromatin accessibility, and DNAme of 3 adult and 2 pediatric B-ALL patients at both diagnosis (Dx) and relapse (Rel) with bulk RNA-seq, ATAC-seq, and DNA methylation capture sequencing (MeCapSeq),

respectively (Figure 1.1a). These patients' tumours contained ≥ 90 % leukemic blasts at diagnosis and were previously profiled using whole exome sequencing (WES) to identify the mutation burden of leukemic driver mutations [3] (see Table D.1; patient numbers used in this study match those from [REF 3]). Matching mutation profiles between Dx and Rel samples allowed for the identification of disease relapse-initiating (dRI) samples, which are cells present at diagnosis that harbour mutations found at relapse, indicating that these cells are relapse-fated. Comprehensive datasets containing RNA-seq, ATAC-seq, and MeCapSeq were produced for 3 patients, with 2 patients lacking RNA-seq data due to source constraints. While expression, chromatin accessibility, and DNAme are each critical for determining cell phenotype and its role in relapse, we sought to investigate the importance of each dataset in an agnostic manner. To achieve this, similarity scores were calculated between all samples using similarity network fusion (SNF) [4]. For each patient, similarity scores between all samples derived from that patient (both primary and patient-derived xenograft (PDX)) were calculated, and weighted graphs to cluster samples together were constructed (see Section 1.5.5). This was done for each individual data type, as well as for a fused network comprised of information by considering all data types simultaneously. To determine the importance of each data type, samples were labelled by their disease stage (Dx, dRI, or Rel; Figure 1.1b). For all 3 patients with complete molecular datasets, the combined networks clustered samples based on disease stage more clearly than each individual dataset (Figure 1.1b). This suggests that disease stages can be more clearly identified from multiple molecular components together than a single component alone [4]. The graphs produced from DNAme data more clearly cluster samples by disease stage than gene expression or chromatin accessibility across all patients, suggesting that DNAme may be a clearer marker of relapse. Taken together, we find that B-ALL disease stage can be identified through non-genetic molecular measurements and that DNAme is mostly closely linked to relapse than gene expression and chromatin accessibility.

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

Given the strong correlation between DNAme signal and disease state, we focused on DNAme changes over B-ALL relapse. To understand the dynamic changes to DNAme that happen over the course of B-ALL relapse, we first looked to the hematopoietic hierarchy and DNAme changes over normal B-cell differentiation. Using normal cord blood pools, sorted into B-cells and multiple B-progenitor cell types, we performed MeCapSeq on 8 pools separated into 4 cell types: hematopoietic

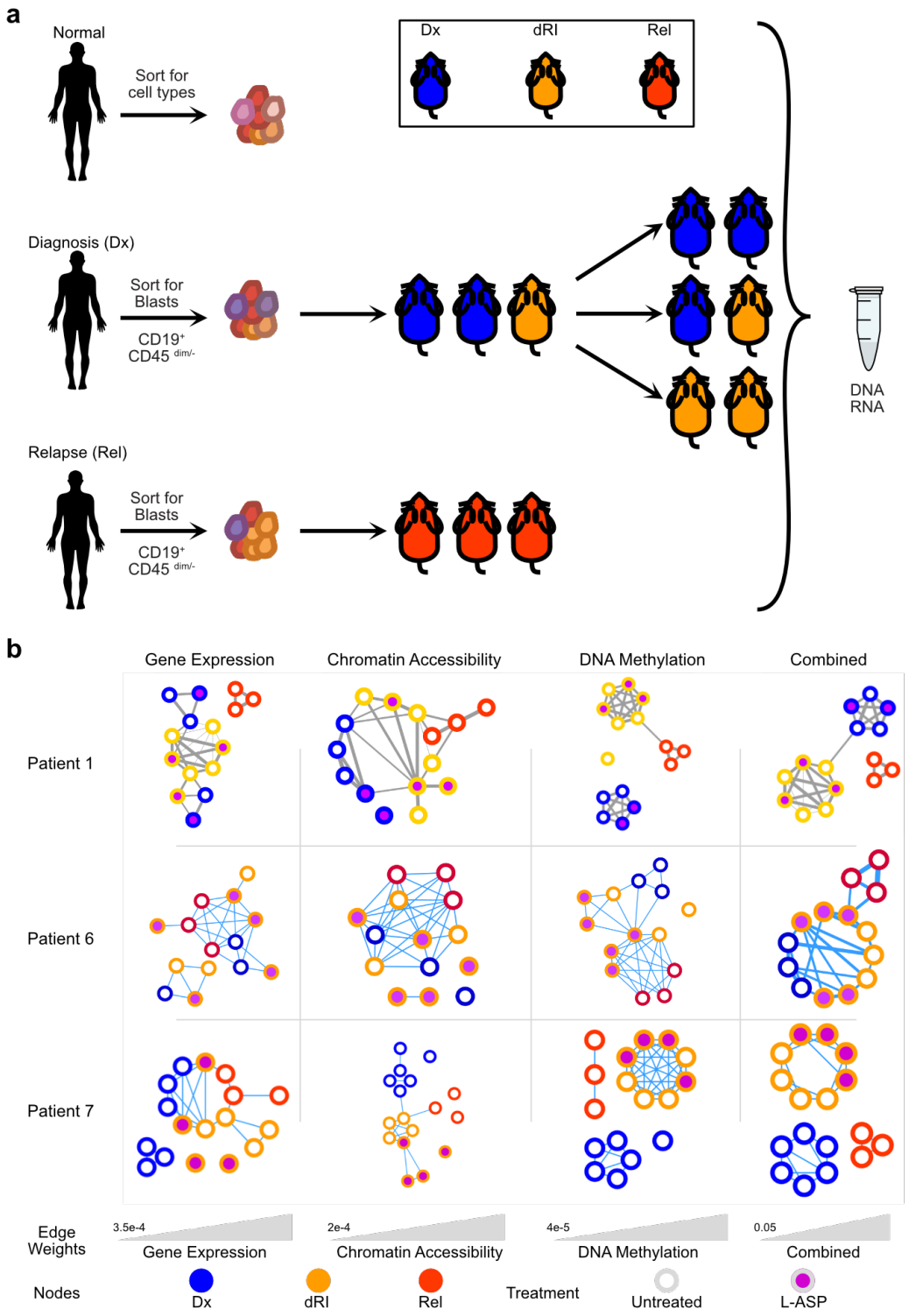

Figure 1.1: **Experimental design and data integration. a.** Experimental design of samples used in this study. Normal samples were obtained from cord blood pools and sorted into various hematopoietic cell types. B-ALL patients who experienced relapse has sorted leukemic blasts collected at Dx and Rel. Based on the mutation profiles from [REF 3] 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.

stem cells (HSCs) and multi-potent progenitors (MPPs); lymphoid-primed multi-potent progenitors (LMPPs) and monocyte-lymphoid progenitors (MLPs); early progenitor B cells (EarlyProBs), pre-progenitor B cells (PreProBs), and progenitor B cells (ProBs) (collectively labelled as ProB); and B-cells (Figure 1.2a; see Table 1.1). Using pairwise comparisons between these cell types, we identified 540 DMRs over the course of B-cell differentiation from HSCs (Figure 1.2b). Significant changes to DNAme occurred in 62 regions from HSC-MPP to LMPP-MLP, 312 regions from LMPP-MLP to ProB, and 166 regions from ProB to fully differentiated B-cells (Figure 1.2c). While roughly equal numbers of loci gained and lost DNAme in the transition from HSCs-MPPs to LMPPs-MLPs, after lymphoid commitment, nearly all regions lost DNAme in later differentiation transitions (Figure B.2c). Overall, 500 (92.6 %) of DMRs identified were loci that became hypomethylated over differentiation. These changes are in agreement with earlier studies profiling DNAme changes over B-cell differentiation using the Illumina 450K arrays [5–7], and provide an expanded set of DMRs with which to track differentiation. Notably, no DMR identified in an earlier transition was found as differentially methylated in a later transition. Regions with altered DNAme in one cell type persisted for all downstream cell type transitions. This suggests that DNAme at these loci can be used as a marker of differentiation. In summary, we find that normal HSCs 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

With the predominant loss of DNAme established in the normal setting, we identified DMRs between Dx and Rel primary and PDX B-ALL samples. When considering all patients together and grouping by disease stage, we found no DMRs remained statistically significant after multiple testing corrections. This result conflicted with previous observations about DNAme changes in B-ALL relapse [6, 7] as well as the earlier SNF analysis. We hypothesized that DNAme changes in across patients was heterogeneous, which limited the ability to detect significant changes. Using a patient-oriented approach, we identified DMRs between Dx and Rel for each patient, separately, to track changes over each patient's relapse trajectory. This identified 25 761 DMRs across the cohort of (range 98 - 15 296, median 7 426, $\delta\beta \geq 20$ %, FDR < 0.1, Figure 1.3a). Unlike the process of normal differentiation, most DMRs were hypermethylated at relapse (Figure 1.2b). 18 610 (72.2 %) DMRs were specific to a single patient and did not overlap DMRs from others (Figure 1.3b, left), as expected from the lack of significant DMRs from the cohort-oriented analysis. Notably,

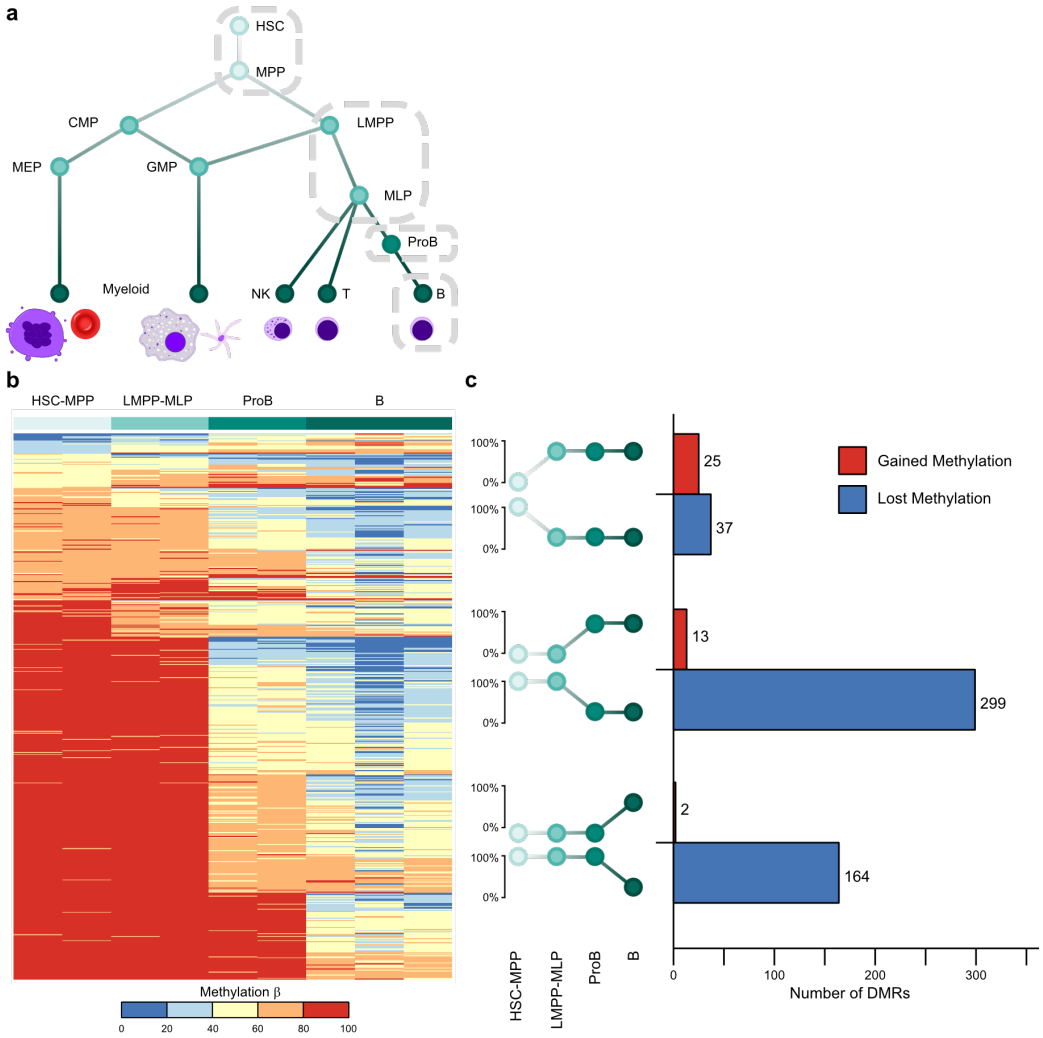

Figure 1.2: **Widespread loss of DNA methylation over B-cell differentiation. a.** Schematic of the hematopoietic hierarchy and the grouping of B-cell progenitors into the groups isolated in this study. **b.** Heatmap of DMRs identified between B lineage cell types. Columns are samples ordered by cell 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.

the 9 recurrently DMRs in all 5 patients are all in the promoter regions of the following genes: *BARHL2*, *CYP26B1*, *EBF3*, *EN2*, *GDNF*, *HMGA2*, *NKX2-2*, *NR2F2*, and *PAX6*. Using gene ontology (GO) analysis, we find that these genes with nearby recurrent differential methylation are positively associated with differentiation, with the most statistically significant pathway being cell fate determination (Figure 1.3c). For these genes, we find that the promoter regions become hypermethylated at B-ALL relapse (Figure 1.3d). Some genes, like *CYP26B1*, have multiple short DMRs in the promoter and one gains DNAme while the other loses DNAme at relapse, but all these promoters gain DNAme overall. Given the association between hypermethylation in promoter regions and decreased expression [8], these results suggest that these genes are under-expressed at relapse. 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, and that recurrent DNAme changes suggest that B-ALL relapse reverts to a more de-differentiated, stem-like DNAme state.

1.3.4 Relapse DNA methylation profiles are present at diagnosis in some patients

Relapse-fated subpopulations of cells present at diagnosis were detected in these patients by their mutations [3]. Yet some Dx samples harboured similar DNAme profiles to the Rel samples (e.g. column 2 for Patient 7 and column 6 for Patient 9 in Figure 1.3a). We hypothesized whether these same populations could be detected by their DNAme profile at diagnosis. By identifying DMRs across all three disease stages (Dx, dRI, and Rel), we identified a median of 2 784 DMRs between disease stages across each patient (range 376 - 4 098; Figure 1.4). There is heterogeneity in DNAme profiles across samples derived from the same patient, and even within the same disease stage (Figure 1.4a). The heterogeneity within disease stages resulted identifying DMRs specific to the dRI samples that were shared between Dx and Rel samples, even when some dRI samples showed similar methylation rates (e.g. leftmost dRI sample for Patient 4). Based on which disease stage the DMRs were identified in, each glsdmr was classified as Dx-specific (shared between dRI and Rel), dRI-specific (shared between Dx and Rel), Rel-specific (shared between Dx and dRI), or unique (significantly differentially methylated in all three stages). All patients harboured Rel-specific DMRs, and a majority of DMRs in total were detected in Patients 1, 4, and 6 (range 71.6 %, 100 %, and 100 %, respectively; Figure 1.4b). For these three patients, a majority of the relapse-fated cells shared the DNAme profile of the neighbouring cells, suggesting that these DNAme changes occurred after mutation. Patients 1, 7, and 9, dRI-specific DMRs were found, suggesting that the DNAme

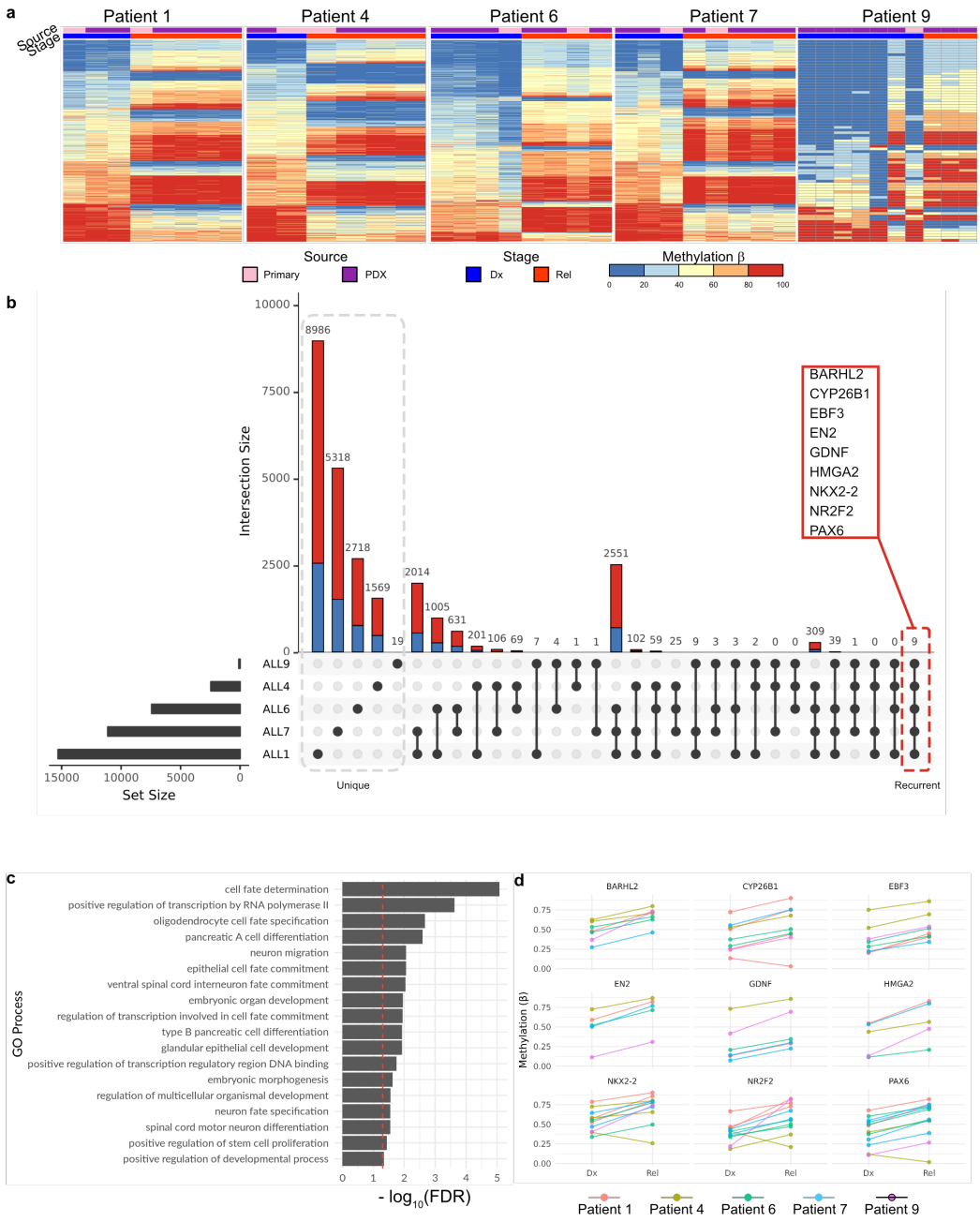

Figure 1.3: **Recurrent relapse DMRs are associated with cell fate decision processes. a.** Heatmaps of DMRs identified between Dx and Rel samples within each patient. **b.** Upset plot showing the shared DMRs between patients. DMRs in the left highlighted block are unique to a single patient, whereas 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.

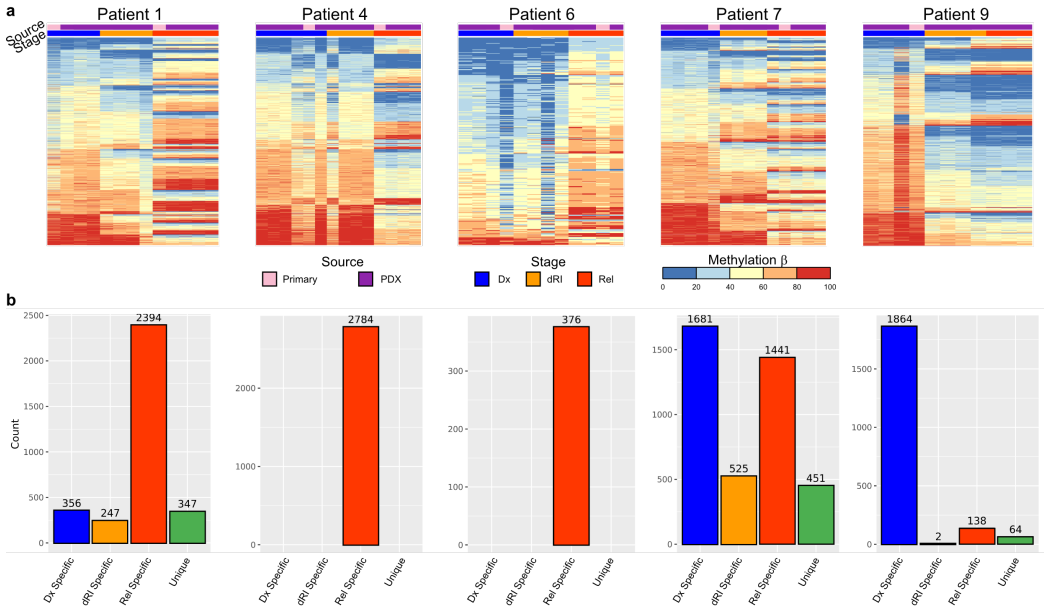

Figure 1.4: **Subpopulations present at diagnosis can harbour relapse-like DNAme profiles. a.** Heatmaps of expanded DMRs identified in dRI PDX samples. **b.** Bar plot showing the number of DMRs classified by which disease stage it is specific to. "Dx Specific" DMRs have shared DNAme between dRI and Rel samples. "DRI Specific" DMRs have shared DNAme between Dx and Rel samples. "Rel Specific" DMRs have shared DNAme between Dx and dRI samples. "Unique" DMRs are regions that have significantly different DNAme at each stage.

profile of cells at diagnosis is not necessarily linked to their mutation status. Further, 41.1 % and 90.1 % of DMRs were found to be Dx-specific for Patients 7 and 9, respectively (Figure 1.4b). Patient 1 also harboured 356 (10.6 %) Dx-specific DMRs. This suggests that some DNAme changes are linked to the mutation status of relapse-fated cells. Taken together, these results suggest that relapse-fated DNAme profiles can be detected at diagnosis, but that the trajectory of DNAme changes over the course of relapse is heterogeneous across patients.

1.4 Discussion

Disease relapse remains a major barrier in treating B-ALL [9–11]. While the genetic origins of relapse have been characterized, epigenetic aberrations underlying relapse have been less well-studied. In this work, we investigated the epigenetic and transcriptomic changes of 5 B-ALL patients over the course of relapse to identify non-genetic changes in tumours that may lead to relapse. DNAme is more highly correlated with disease stage than RNA or chromatin accessibility and changes to DNAme are antithetical to DNAme changes seen in normal B-cell differentiation. While most DNAme changes are patient-specific, a small number of recurrent changes indicate a more

stem-like state at relapse. In some cases, these stem-like DNAme profiles are present at diagnosis, indicating that subclones defined by DNAme may also contribute to B-ALL relapse.

Both genetic and epigenetic aberrations in tumours play important roles in determining disease relapse [6, 7]. Previous reports highlight the frequency that epigenetic regulators are mutated in B-ALL [12] and leukemias more generally, such as *DNMT3A*, *TET2*, *IDH1*, and *IDH2* in acute myeloid leukemia (AML) [13–15] and *CHD2*, *HIST1H1E*, and *ZMYM3* in chronic lymphocytic leukemia (CLL) [16–18]. These findings demonstrate that epigenetic modifications, in conjunction with genetic aberrations, discriminate disease outcomes and can share an evolutionary trajectory in cancer. However, it is not the case that genetic and epigenetic states always behave similarly. In this study we found both Dx and dRI PDXs samples that share DNAme profiles with the Rel tumours, suggesting that DNAme states can vary independently of mutations. Moreover, the differences between PDX methylomes derived from the same primary sample demonstrates that subpopulations of cells can have differing DNAme states while sharing mutations. This decoupling between genome and epigenome has been observed in other tumours, such as pediatric ependymomas, where recurrent DNAme profiles were found in the absence of recurrent mutations and was associated with outcome [19, 20], and glioblastoma, where stem cells are characterized by widespread changes in chromatin accessibility [21] and histone modifications [22]. These studies highlight the role of epigenetic plasticity and intra-tumour heterogeneity in cancers [23]. With similar results found in leukemias that are linked to disease outcome [24–28], it is likely that epigenetic plasticity and heterogeneity are also key factors in therapeutic response and relapse. Taken together, these results suggest that the epigenome can provide mechanisms independent of genetic aberrations, to respond and adapt to therapies, but are often guided by genetic aberrations. This complexity of disease response will need to be addressed to design treatment regimens for patients with an increased propensity towards relapse.

Previous investigations of DNAme aberrations in B-ALL have primarily focused on a select few genes, or single CG dinucleotides (CpGs) in promoter regions [6, 7, 29, 30]. While the recurrent DMRs in this study were found in these same regions, most DMRs were identified in intergenic regions. This suggests that important changes in the epigenetic landscape is currently unidentified, and future studies investigating DNAme aberrations in B-ALL should prioritize genome-wide approaches. The phenotypic impact of focal hypermethylation on engraftment and self-renewal capacity has not been assessed here, so experiments should be conducted to validate these findings (this is a bad sentence but this idea is important). For patients undergoing B-ALL treatment, DNAme has the potential to be used as early indicators of relapse. Moreover, treatment with DNA demethylating

agents, such as 5-aza-citidine and 5-aza-2'-deoxycytadine, may be effective at preventing relapse. These treatments have been approved for use in patients with myelodisplastic syndrome (MDS) and AML in adult populations and early clinical trials have demonstrated their safety [31, 32], although some toxic effects have been identified in drug combination trials [33]. Taken together, therapeutic targeting of DNAme may be an effective method to prevent B-ALL relapse by preventing the outgrowth of stem-like subpopulations that survive chemotherapy.

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 [3]. 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 [3]. 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 [34], 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 [3]. 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 [35]. 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 [36].

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 [3]. 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 [37]. 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 3]. Briefly, RNA-seq reads were aligned against the GRCh38 reference human genome with STAR (v2.5.2b) [38] 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) [39]. Data was library size normalized using the RLE method, followed by a variance stabilizing transformation using DESeq2 (v1.22.1) [40]. 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) [41] with default parameters. Accessible peaks were identified with MACS2 (v2.0.10) [42] 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 [43] for downstream analysis.

Bisulfite sequencing pre-processing

Sequencing read qualities were assessed with FastQC (v0.11.8) [44]. Low quality bases were trimmed with Trim Galore! (v0.6.3) [45] 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) [46] 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) [47]. M-biases were calculated with MethylDackel (v0.4.0) [48], 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 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 [4] 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) [49] 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 [50]. 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) [51]. 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

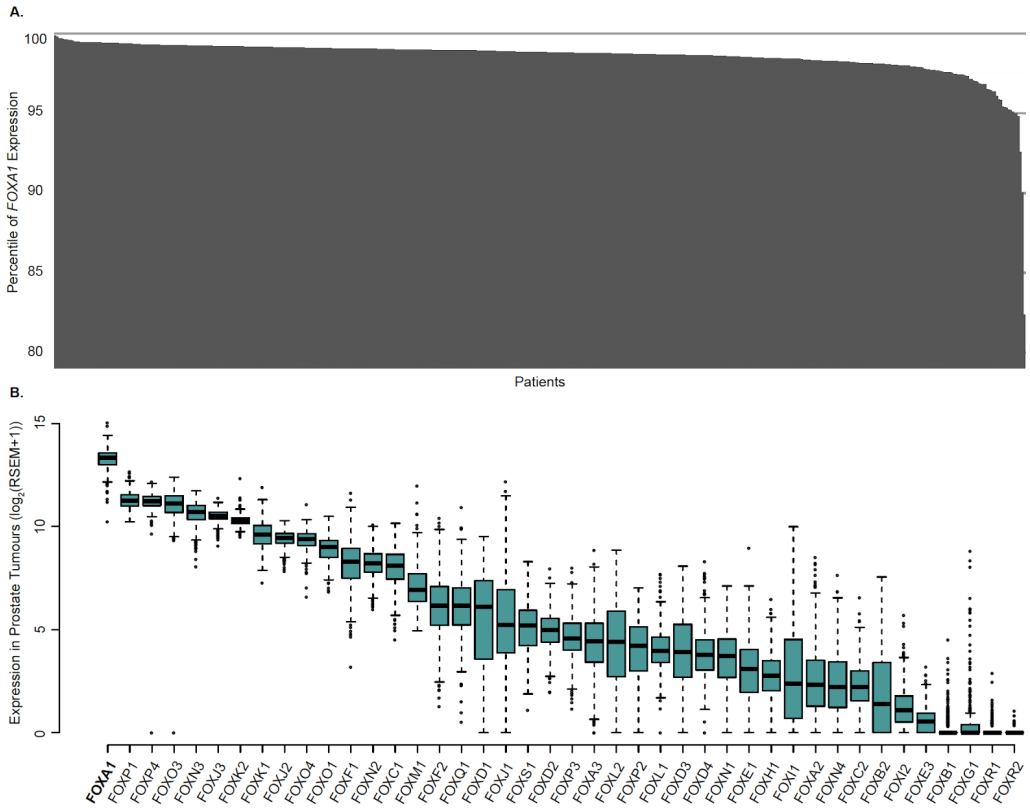

Figure A.1: ***FOXA1* mRNA expression in prostate tumours. a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

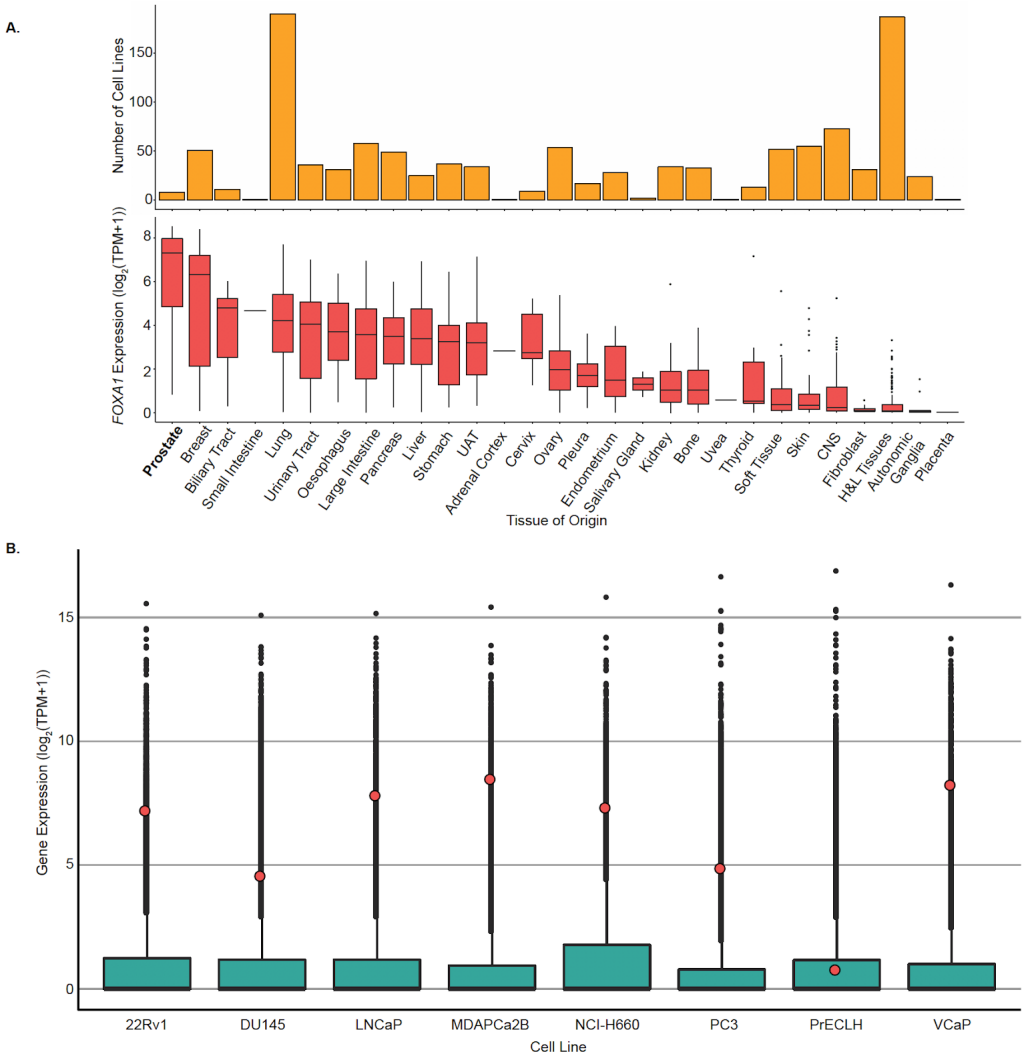

Figure A.2: ***FOXA1* mRNA expression across PCa cell lines. a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Methods). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA expression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

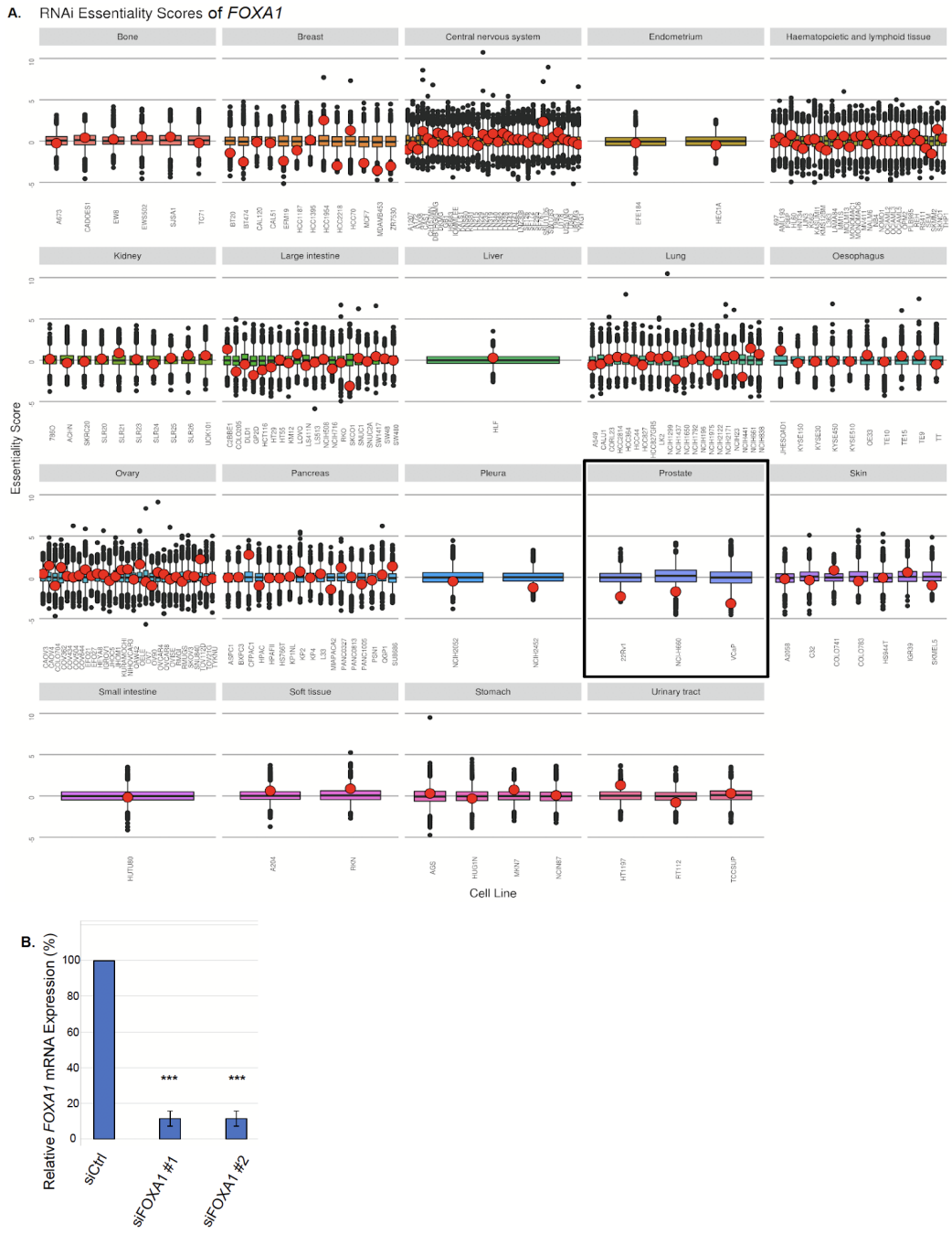

Figure A.3: **Essentiality of *FOXA1* across cancer cell lines of various cancer types. a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot 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, *** $p < 0.001$.

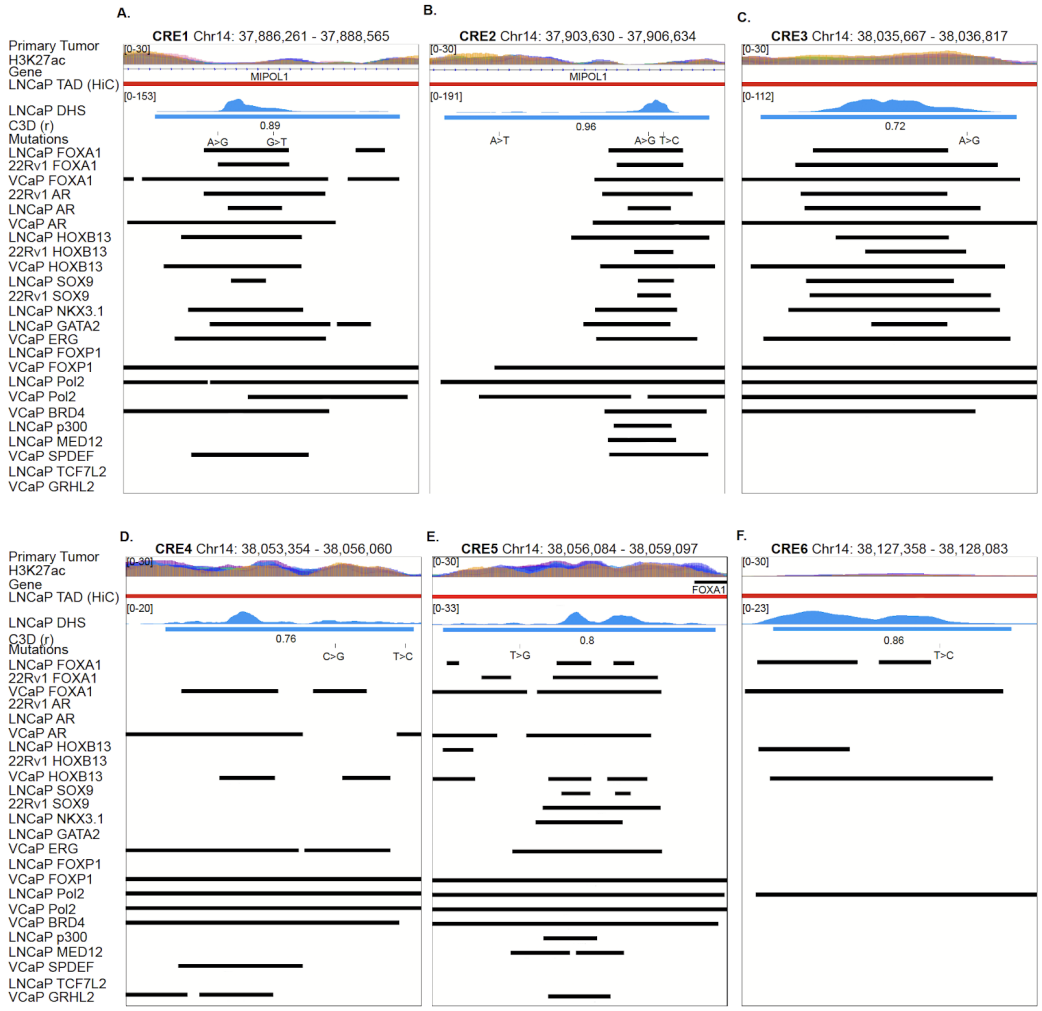

Figure A.4: **Visualization of the functional annotation of the six *FOXA1* CREs. a-f.** Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

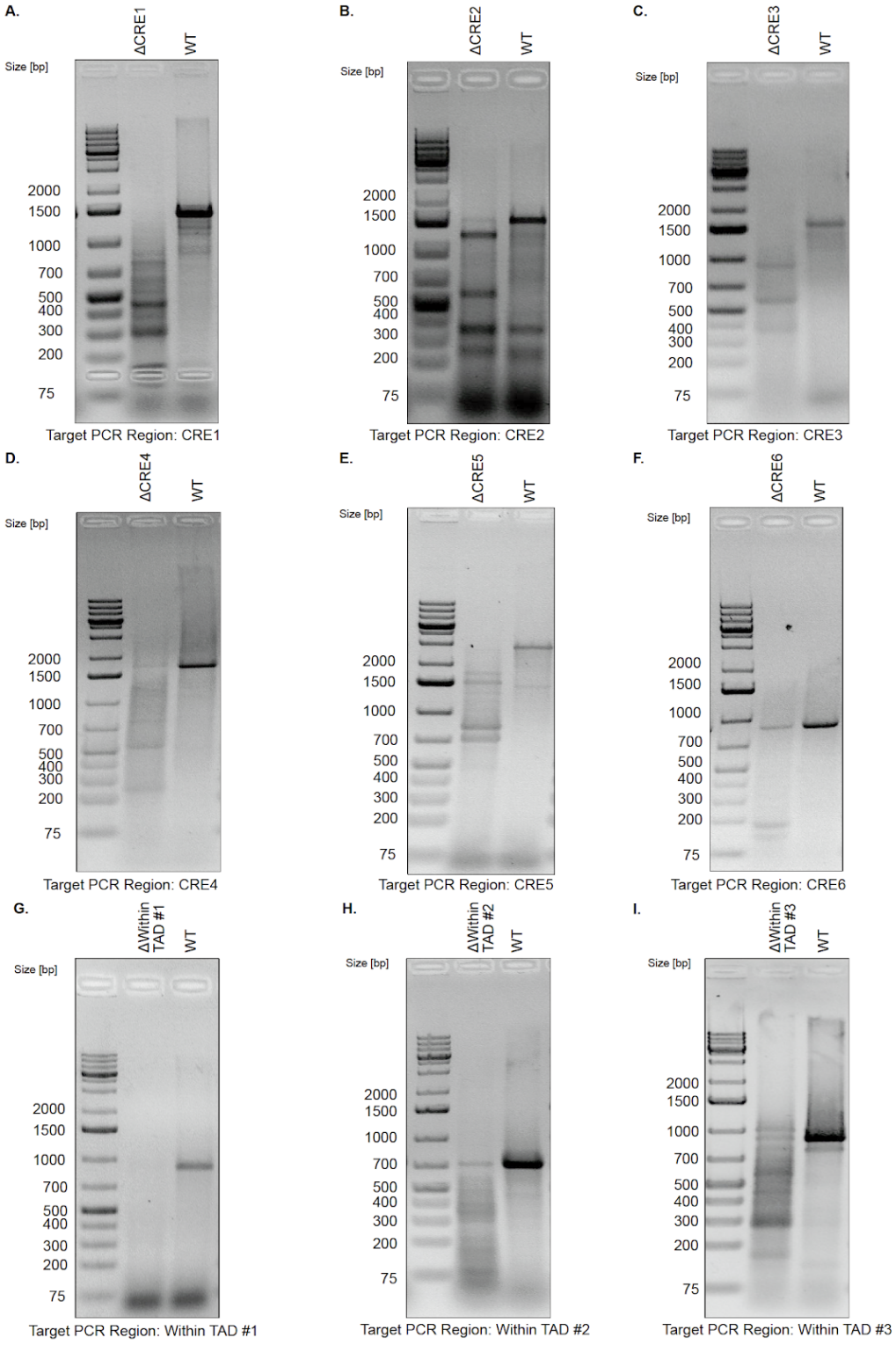

Figure A.5: **Validation of clonal Cas-mediated deletions of CREs. a-f**. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

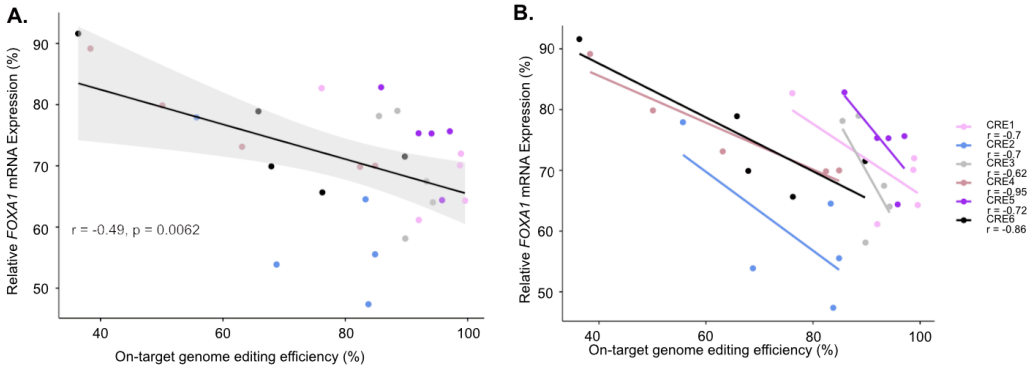

Figure A.6: **Genome editing efficiency (%) is inversely correlated with *FOXA1* mRNA expression. a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all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

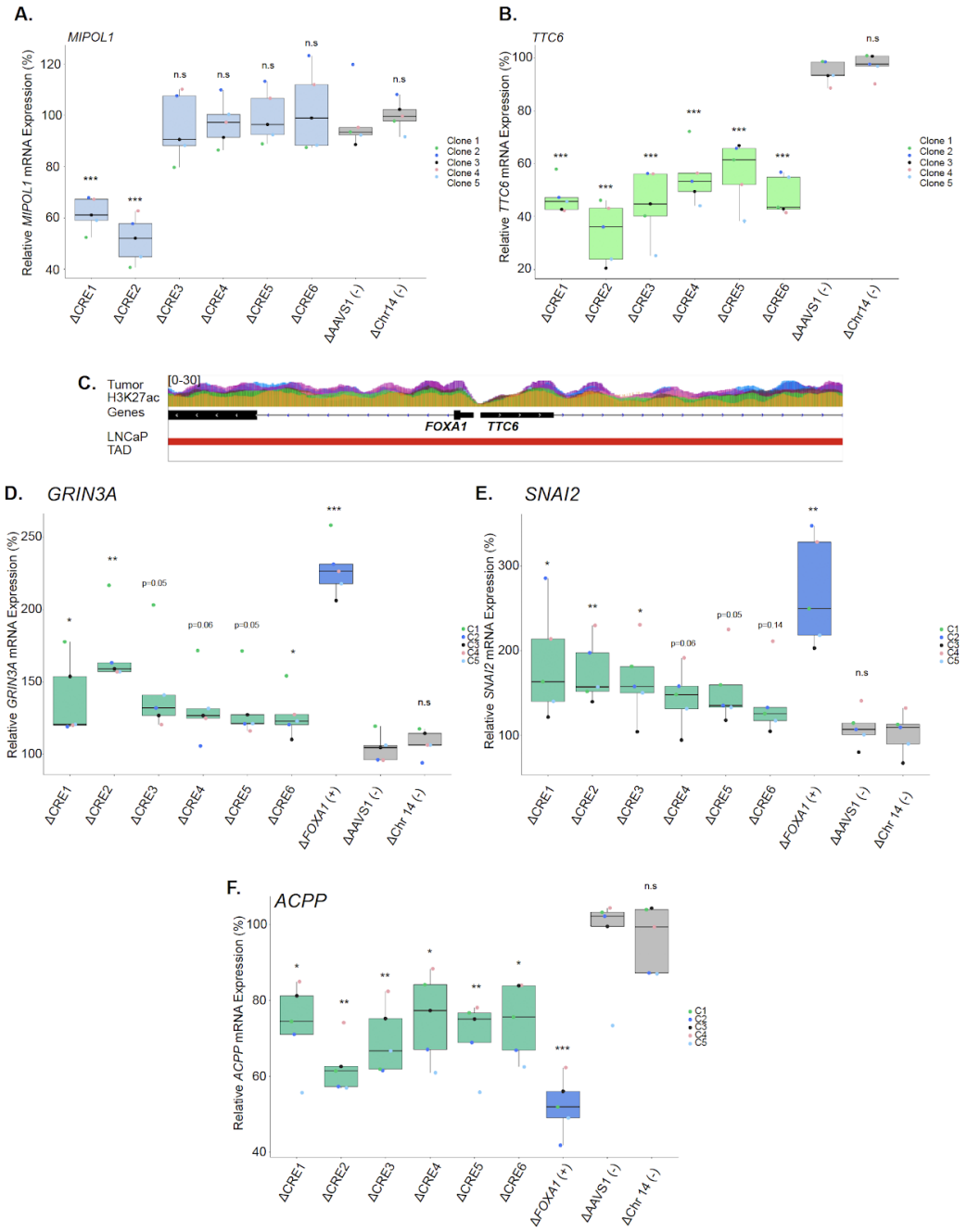

Figure A.7: **Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a.** *MIPOL1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. **b.** *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of 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, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

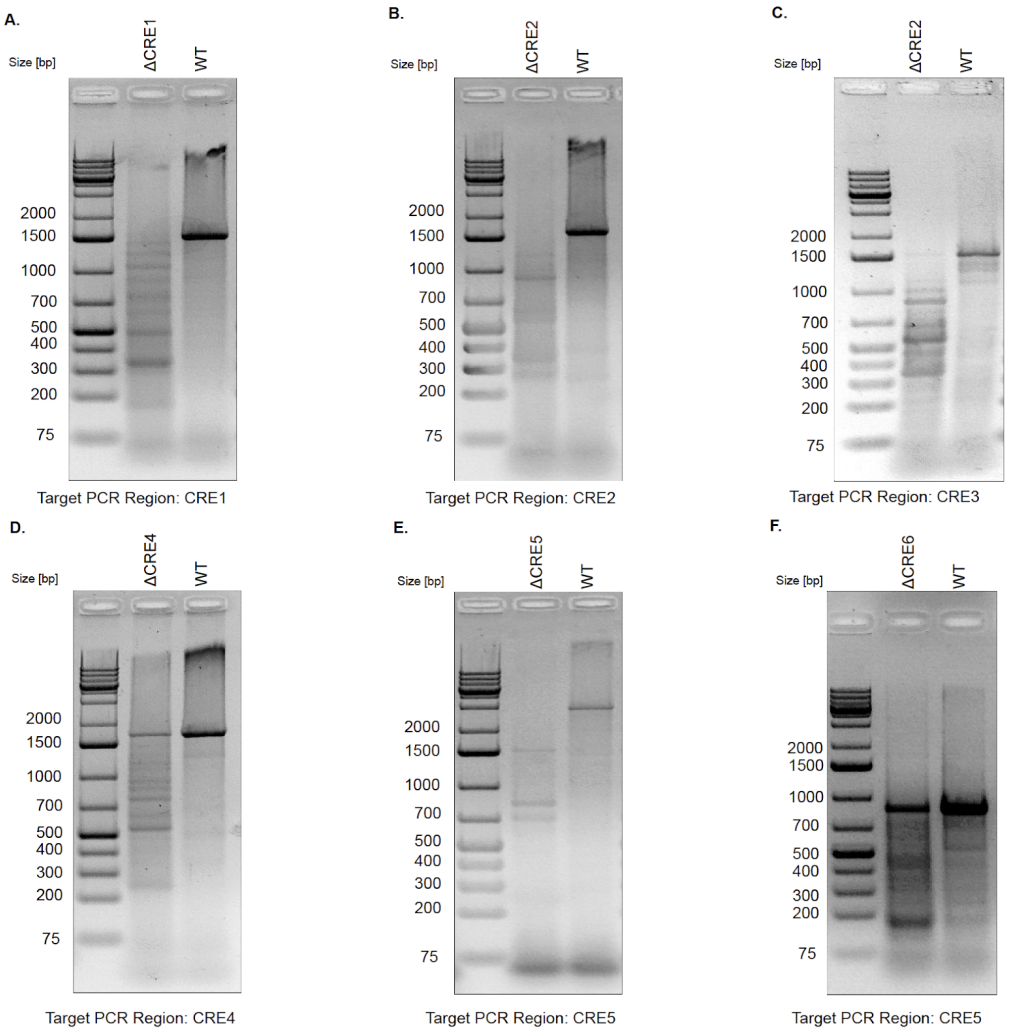

Figure A.8: **Validation of transient Cas9-mediated single deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

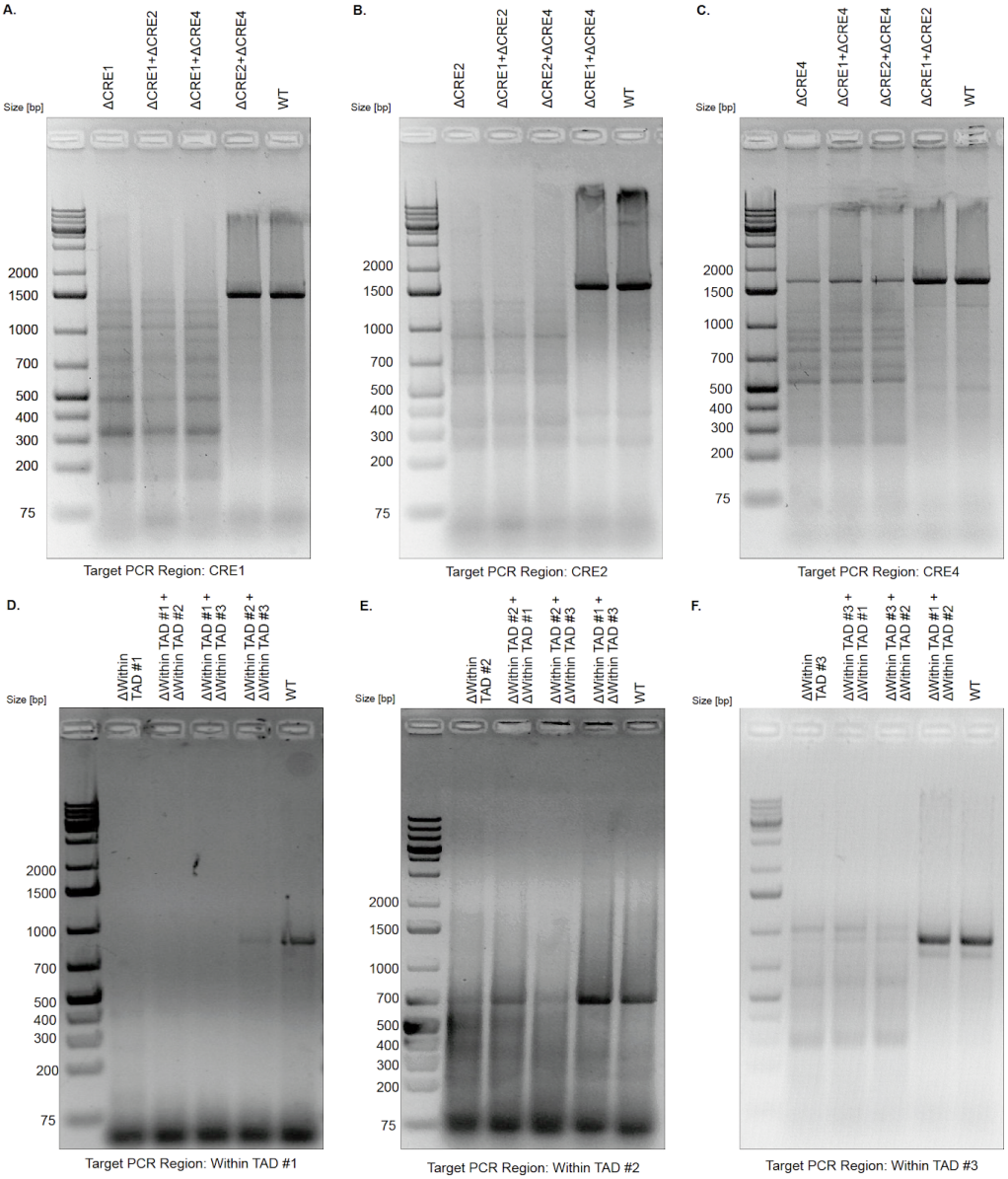

Figure A.9: **Validation of transient Cas9-mediated double deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

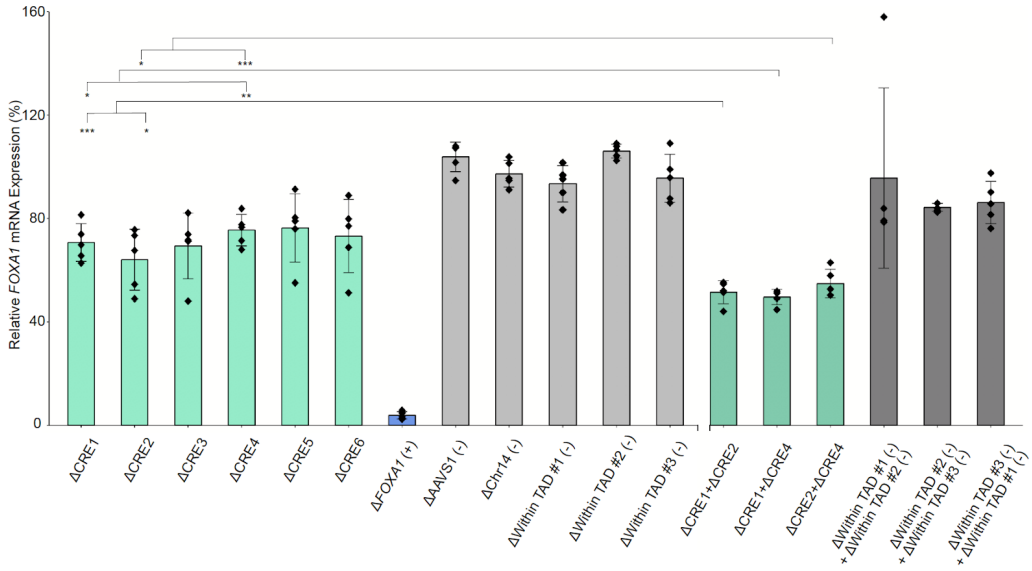

Figure A.10: **Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s).** *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate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

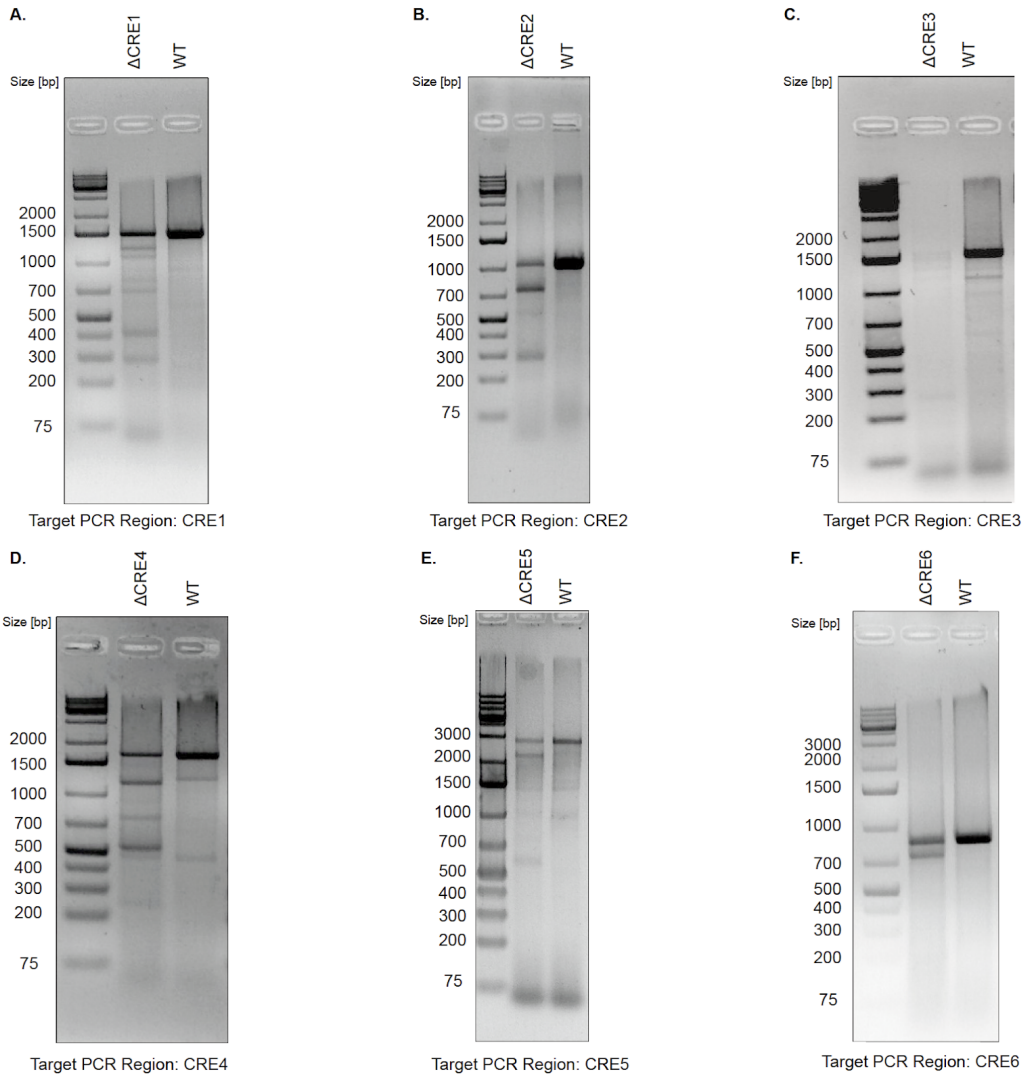

Figure A.11: **Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f.** Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs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

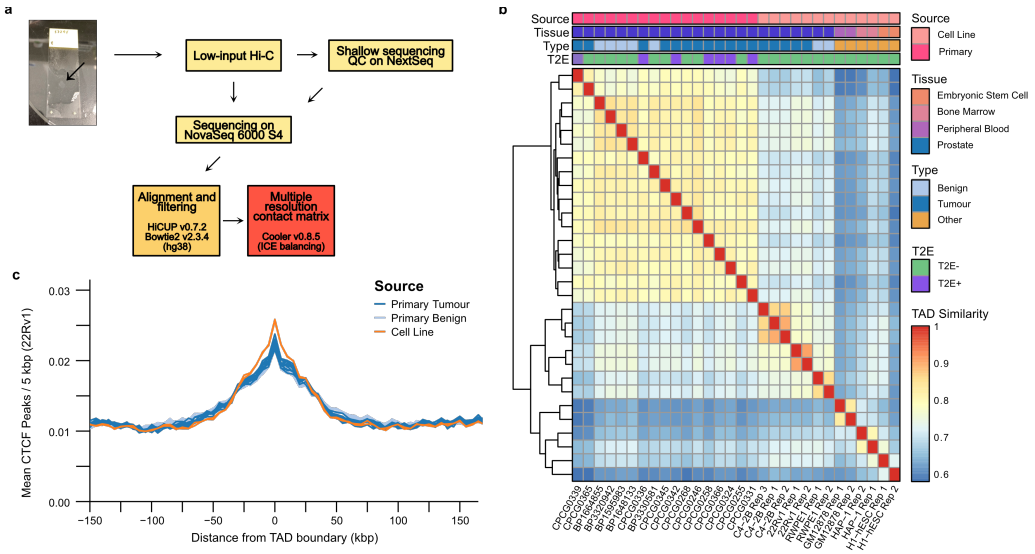

Figure B.1: **Sample processing and TAD similarity between samples. a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate 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.

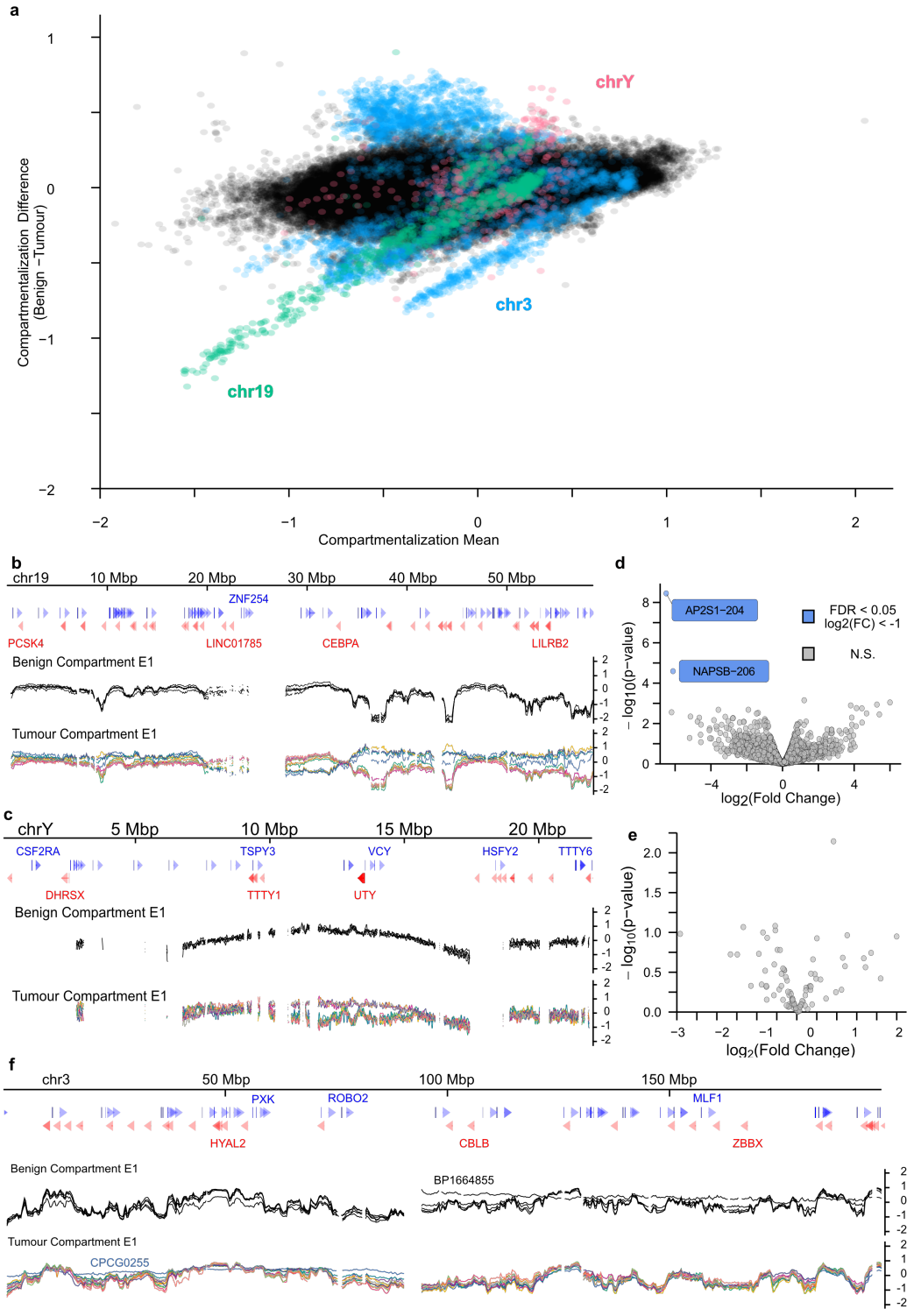

Figure B.2: **Compartmentalization changes in tumours is not associated with widespread differential gene expression.** (Continued on the following page)

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.

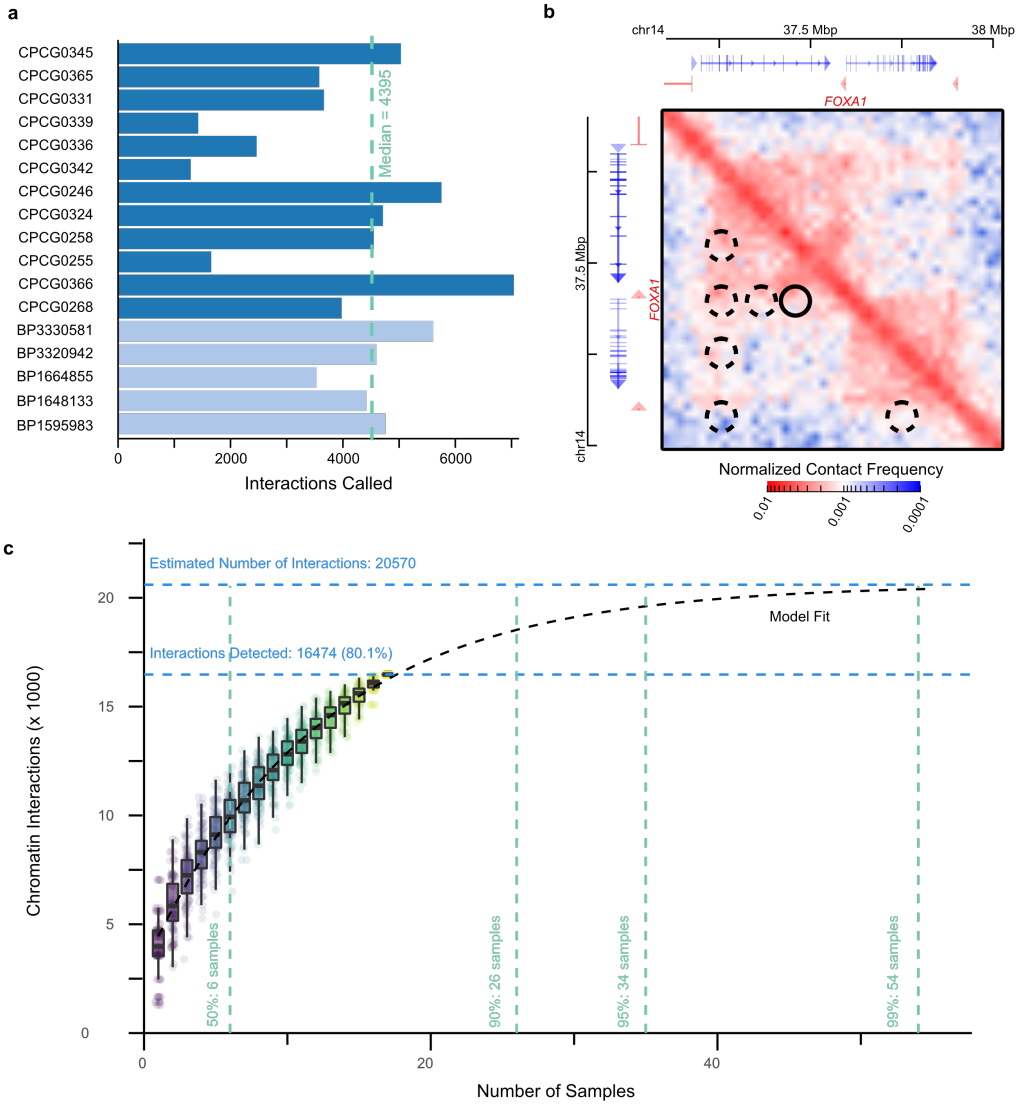

Figure B.3: **Characterization of chromatin interactions in benign and tumour tissue. a.** Bar plot of the number of significant chromatin interactions identified in each of the primary prostate samples. **b.** A snapshot of significant chromatin interactions called around the *FOXA1* gene. Identified 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.

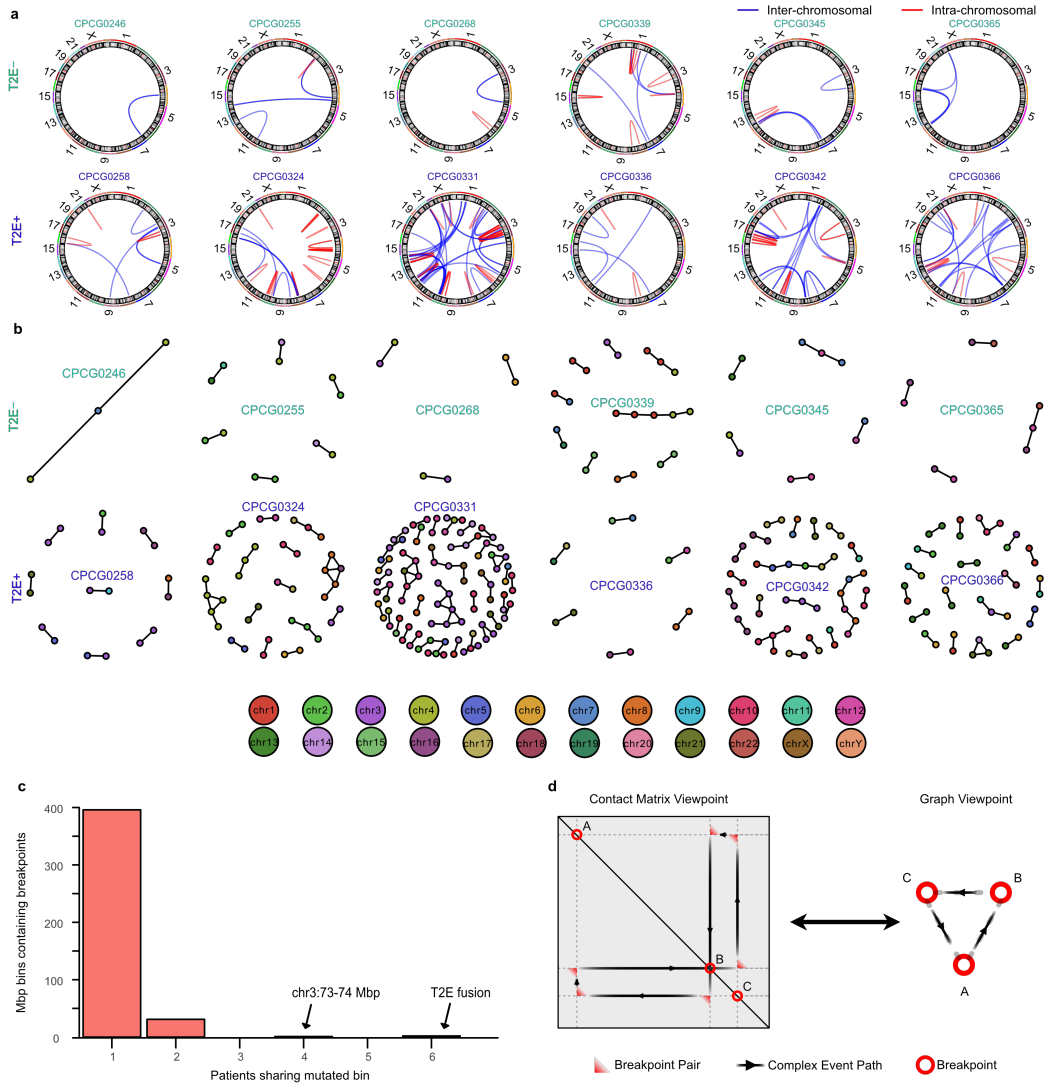

Figure B.4: **Structural variant detection from Hi-C data. a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1 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.

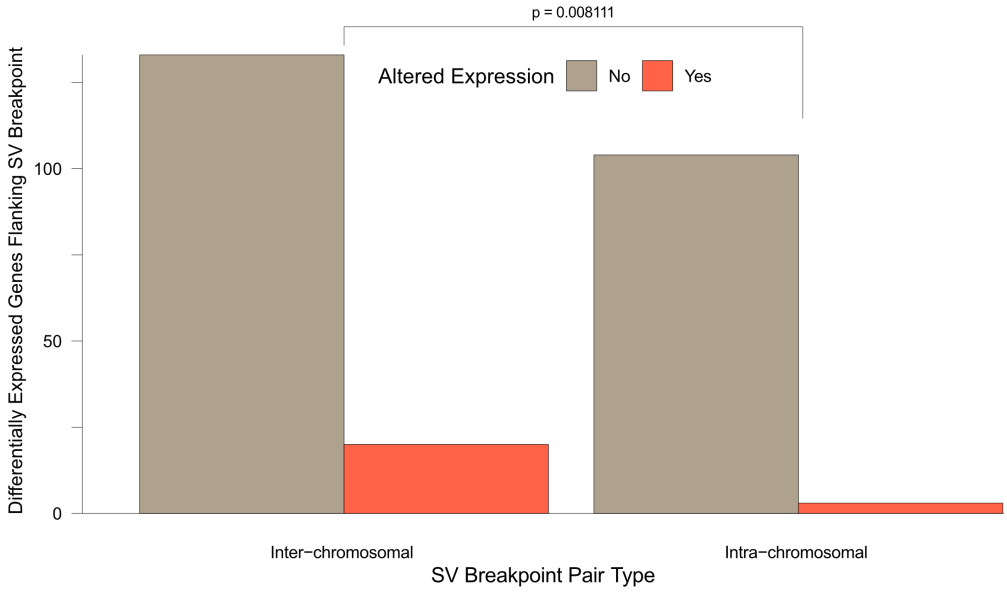

Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes.

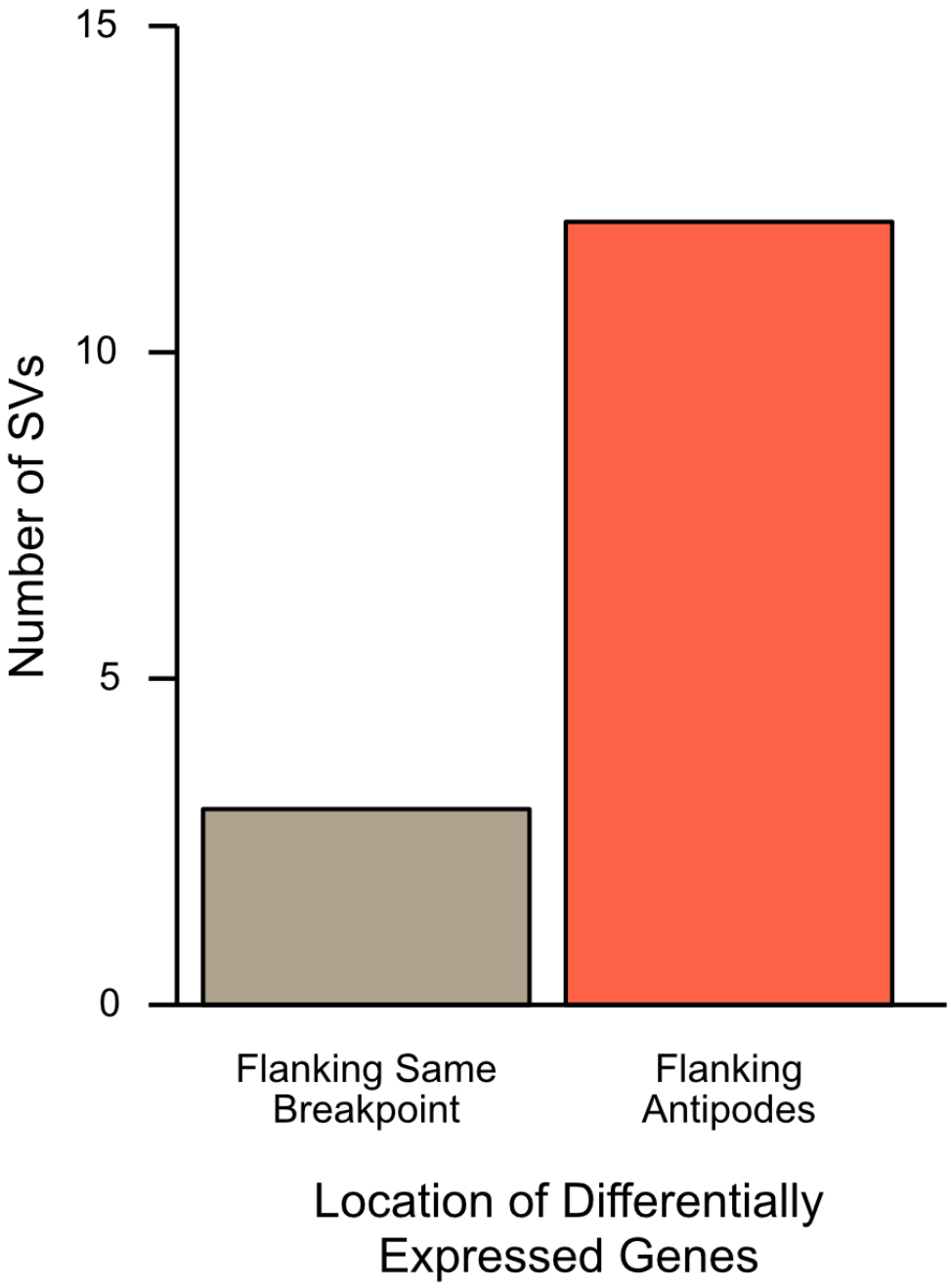

Figure B.6: **Location of differentially expressed genes around SV breakpoints.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

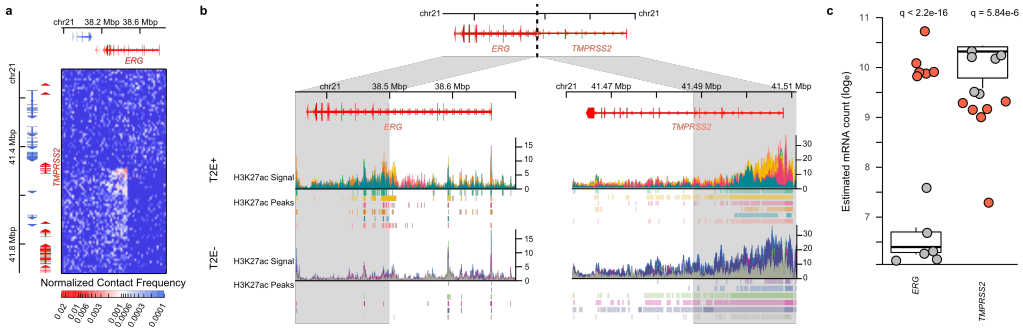

Figure B.7: **Chromatin organization of the *TMPRSS2-ERG* fusion. a.** Contact matrix of the deletion between *TMPRSS2* and *ERG*. **b.** Genome tracks of H3K27ac ChIP-seq signal in T2E+ and T2E- patients. The grey region highlights the loci that come into contact as a result of the deletion. **c.** 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) [52, 53] 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) [54] 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 52], 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 55], $\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] \leq \mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(OLS)} - \mathrm{B}_1\|^2\right]$. 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

Table D.1: **Clinical characteristics of patients participating in this study. .**

Patient	Dx Subtype	Age	Sex	Bone Marrow Blast Count	Cytogenetics
1	DUX4	> 18	M	90%	46,XY,del(6)(q21q23), t(11;13)(p1
4	B-other	> 18	M	90%	NA
6	B-other	> 18	M	90%	46 XY t(2;3) (p23,q21)[20]
7	DUX4	< 18	F	92%	46,XX,del(9)(p22) [20/100%
9	B-other	< 18	M	96%	46,XY [1/5%]/47,XY,del(9)(p13),+18

Glossary

3C chromatin conformation capture

AML acute myeloid leukemia

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

CLL chronic lymphocytic leukemia

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

MDS myelodisplastic syndrome

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

WES whole exome sequencing

WGS whole genome sequencing

WT wild-type

References

1. Izzo, F. *et al.* DNA Methylation Disruption Reshapes the Hematopoietic Differentiation Landscape. en. *Nature Genetics,* 1–10. ISSN: 1546-1718 (Mar. 2020).

2. Takayama, N. *et al.* The Transition from Quiescent to Activated States in Human Hematopoietic Stem Cells Is Governed by Dynamic 3D Genome Reorganization. en. *Cell Stem Cell* **28,** 488–501.e10. ISSN: 19345909 (Mar. 2021).

3. 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).

4. 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).

5. 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).

6. 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).

7. 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).

8. Jones, P. A. Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nature reviews. Genetics* **13,** 484–92. ISSN: 1471-0064 (Electronic)\r1471-0056 (Linking) (2012).

9. Forman, S. J. & Rowe, J. M. The Myth of the Second Remission of Acute Leukemia in the Adult. en. *Blood* **121,** 1077–1082. ISSN: 0006-4971 (Feb. 2013).

10. Liew, E. *et al.* Outcomes of Adult Patients with Relapsed Acute Lymphoblastic Leukemia Following Frontline Treatment with a Pediatric Regimen. *Leukemia Research. Special Section: Symposium on Myeloid Neoplasms - June 9, 2012* **36,** 1517–1520. ISSN: 0145-2126 (Dec. 2012).

11. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. en. *New England Journal of Medicine* **373** (ed Longo, D. L.) 1541–1552. ISSN: 0028-4793, 1533-4406 (Oct. 2015).

12. Ma, X. *et al.* Rise and Fall of Subclones from Diagnosis to Relapse in Pediatric B-Acute Lymphoblastic Leukaemia. en. *Nature Communications* **6,** 1–12. ISSN: 2041-1723 (Mar. 2015).

13. Kishtagari, A., Levine, R. L. & Viny, A. D. Driver Mutations in Acute Myeloid Leukemia. en-US. *Current Opinion in Hematology* **27,** 49–57. ISSN: 1065-6251 (Mar. 2020).

14. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine* **374,** 2209–2221. ISSN: 0028-4793 (June 2016).

15. Ley, T. J. *et al.* DNMT3A Mutations in Acute Myeloid Leukemia. *New England Journal of Medicine* **363,** 2424–2433. ISSN: 1533-4406 (Electronic)\r0028-4793 (Linking) (Dec. 2010).

16. Billot, K. *et al.* Deregulation of Aiolos Expression in Chronic Lymphocytic Leukemia Is Associated with Epigenetic Modifications. *Blood* **117,** 1917–1927. ISSN: 0006-4971 (Feb. 2011).

17. Landau, D. A. & Wu, C. J. Chronic Lymphocytic Leukemia: Molecular Heterogeneity Revealed by High-Throughput Genomics. en. *Genome Medicine* **5,** 47. ISSN: 1756-994X (May 2013).

18. Landau, D. A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152,** 714–726. ISSN: 0092-8674 (Feb. 2013).

19. Mack, S. C. *et al.* Epigenomic Alterations Define Lethal CIMP-Positive Ependymomas of Infancy. en. *Nature* **506,** 445–450. ISSN: 0028-0836, 1476-4687 (Feb. 2014).

20. Pajtler, K. W. *et al.* Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. English. *Cancer Cell* **27,** 728–743. ISSN: 1535-6108, 1878-3686 (May 2015).

21. Guilhamon, P. *et al.* Single-Cell Chromatin Accessibility Profiling of Glioblastoma Identifies an Invasive Cancer Stem Cell Population Associated with Lower Survival. *eLife* **10** (eds Postovit, L.-M., Struhl, K. & Verhaak, R.) e64090. ISSN: 2050-084X (Jan. 2021).

22. Liau, B. B. *et al.* Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. en. *Cell Stem Cell* **20,** 233–246.e7. ISSN: 1934-5909 (Feb. 2017).

23. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic Plasticity and the Hallmarks of Cancer. *Science* **357,** eaal2380–eaal2380 (July 2017).

24. Pastore, A. *et al.* Corrupted Coordination of Epigenetic Modifications Leads to Diverging Chromatin States and Transcriptional Heterogeneity in CLL. En. *Nature Communications* **10,** 1874. ISSN: 2041-1723 (Apr. 2019).

25. Landau, D. A. *et al.* Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell* **26,** 813–825. ISSN: 1878-3686 (Electronic)\r1535-6108 (Linking) (2014).

26. Gaiti, F. *et al.* Epigenetic Evolution and Lineage Histories of Chronic Lymphocytic Leukaemia. en. *Nature* **569,** 576–580. ISSN: 1476-4687 (May 2019).

27. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating Genetic and Non-Genetic Determinants of Cancer Evolution by Single-Cell Multi-Omics. en. *Nature Reviews Genetics* **22,** 3–18. ISSN: 1471-0064 (Jan. 2021).

28. Li, S. *et al.* Distinct Evolution and Dynamics of Epigenetic and Genetic Heterogeneity in Acute Myeloid Leukemia. *Nature Medicine* **22,** 792–799 (June 2016).

29. Garcia-Manero, G. *et al.* DNA Methylation of Multiple Promoter-Associated CpG Islands in Adult Acute Lymphocytic Leukemia. eng. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **8,** 2217–2224. ISSN: 1078-0432 (July 2002).

30. Garcia-Manero, G. *et al.* Aberrant DNA Methylation in Pediatric Patients with Acute Lymphocytic Leukemia. eng. *Cancer* **97,** 695–702. ISSN: 0008-543X (Feb. 2003).

31. Benton, C. B. *et al.* Safety and Clinical Activity of 5-Aza-2′-Deoxycytidine (Decitabine) with or without Hyper-CVAD in Relapsed/Refractory Acute Lymphocytic Leukaemia. en. *British Journal of Haematology* **167,** 356–365. ISSN: 1365-2141 (2014).

32. National Cancer Institute (NCI). *A Groupwide Pilot Study to Test the Tolerability and Biologic Activity of the Addition of Azacitidine (NSC# 102816) to Chemotherapy in Infants With Acute Lymphoblastic Leukemia (ALL) and KMT2A (MLL) Gene Rearrangement* Clinical Trial Registration NCT02828358 (clinicaltrials.gov, May 2021).

33. Therapeutic Advances in Childhood Leukemia Consortium. *A Pilot Study of Decitabine and Vorinostat With Chemotherapy for Relapsed ALL* Clinical Trial Registration NCT01483690 (clinicaltrials.gov, Oct. 2020).

34. 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).

35. 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).

36. 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).

37. 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).

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

39. 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).

40. 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).

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

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

43. 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).

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

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

46. 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).

47. 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).

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

49. 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).

50. 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).

51. 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).

52. 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).

53. 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).

54. 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).

55. 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).