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

Contents

1 Introduction **1**

2 Noncoding mutations target *cis*-regulatory elements of the *FOXA1* plexus in prostate cancer **2**

2.1 Abstract . . . 3

2.2 Introduction . . . 3

2.3 Results . . . 5

2.3.1 *FOXA1* is essential for prostate cancer proliferation . . . 5

2.3.2 Identifying putative *FOXA1* CREs . . . 5

2.3.3 Putative *FOXA1* CREs harbour TF binding sites and SNVs . . . 7

2.3.4 Disruption of CREs reduces *FOXA1* mRNA expression . . . 10

2.3.5 *FOXA1* CREs collaborate to regulate its expression . . . 13

2.3.6 Disruption of *FOXA1* CREs reduces prostate cancer cell growth . . . 13

2.3.7 SNVs mapping to *FOXA1* CREs can alter their activity . . . 15

2.3.8 SNVs mapping to *FOXA1* CREs can modulate the binding of TFs . . . 15

2.4 Discussion . . . 17

2.5 Methods . . . 18

2.5.1 Cell Culture . . . 18

2.5.2 Prostate tumours and cancer cell lines expression . . . 19

2.5.3 Prostate cancer cell line gene essentiality . . . 19

2.5.4 siRNA knockdown and cell proliferation assay . . . 19

2.5.5 Identifying putative FOXA1 CREs . . . 20

2.5.6 Hi-C and TADs in LNCaP cells . . . 20

2.5.7 Clonal wild-type Cas9 and dCas9-KRAB mediated validation . . . 20

2.5.8 Transient Cas9-mediated disruption of CREs . . . 21

2.5.9 RT-PCR assessment of gene expression upon deletion of CREs . . . 22

2.5.10 Confirmation of Cas9-mediated deletion of CREs . . . 22
2.5.11 Cell proliferation upon deletion of FOXA1 CREs . . . 22
2.5.12 Luciferase reporter assays . . . 23
2.5.13 Allele-specific ChIP-qPCR . . . 23
2.6 Data availability . . . 24

3 Recurrent reorganization of the three-dimensional genome pinpoint non-coding drivers of primary prostate tumours 25

3.1 Abstract . . . 25
3.2 Introduction . . . 26
3.3 Results . . . 27
3.3.1 Three-dimensional chromatin organization is stable over oncogenesis . . . 27
3.3.2 Focal chromatin interactions shift over oncogenesis . . . 30
3.3.3 Cataloguing structural variants from Hi-C data . . . 30
3.3.4 SVs alter gene expression independently of intra-TAD contacts . . . 34
3.3.5 SVs alter focal chromatin interactions to hijack CREs and alter antipode gene expression . . . 36
3.3.6 Discussion . . . 38
3.4 Methods . . . 40
3.4.1 Patient selection criteria . . . 40
3.4.2 Patient Tumour in situ low-input Hi-C Sequencing . . . 40
3.4.3 Hi-C Sequencing and Data Pre-processing . . . 43
3.4.4 Hi-C Data Analysis . . . 45
3.4.5 Patient Tumour Tissue H3K27ac ChIP-seqs . . . 49
3.4.6 Primary Tissue RNA Data Analysis . . . 51

4 Hedging uncertainty in differential gene expression analyses with James-Stein estimators 52

4.1 Abstract . . . 52
4.2 Introduction . . . 52
4.3 Results . . . 53
4.3.1 Derivation of the James-Stein fold change estimator . . . 53
4.3.2 Comparison between the OLS and James-Stein estimators . . . 55
4.3.3 Empirical analysis of the James-Stein estimator . . . 57

4.4 Discussion 60
4.5 Methods 60
4.5.1 RNA sequencing data collection and pre-processing 60
4.5.2 Differential expression analysis 61
4.5.3 Random sampling and simulations 61
4.5.4 Random sampling of smaller numbers of transcripts 61

5 Epigenetic dynamics underlying B cell acute lymphoblastic leukemia relapse 63

6 Discussion & Future Directions 64

A Supplementary Material for Chapter 2 65

B Supplementary Material for Chapter 3 77

C Supplementary Material for Chapter 4 86
C.1 Differential expression analysis with Sleuth 86
C.2 Statistical moments of the ordinary least squares estimator 88
C.3 Statistical moments of the James-Stein estimator 89
C.3.1 Expected value of the James-Stein estimator 89
C.3.2 Variance of the James-Stein estimator 90

D Supplementary Material for Chapter 5 92

Glossary 93

References 95

Chapter 1

Introduction

Chapter 2

Noncoding mutations target *cis*-regulatory elements of the *FOXA1* plexus in prostate cancer

This chapter is a version of the paper published in *Nature Communications* as follows:

Zhou, S., **Hawley, J. R.**, Soares, F., Grillo, G., Teng, M., Tonekaboni, S. A. M., Hua, J. T., Kron, K. J., Mazrooei, P., Ahmed, M., Arlidge, C., Yun, H. Y., Livingstone, J., Huang, V., Yamaguchi, T. N., Espiritu, S. M. G., Zhu, Y., Severson, T. M., Murison, A., Cameron, S., Zwart, W., van der Kwast, T., Pugh, T. J., Fraser, M., Boutros, P. C., Bristow, R. G., He, H. H., and Lupien, M. Noncoding mutations target *cis*-regulatory elements of the *FOXA1* plexus in prostate cancer. **Nature Communications**, 2020;11:1–13.

Contributions per the manuscript: S.Z. and M.L. conceptualized the study. S.Z. designed and conducted most of the experiments with help from F.S., G.G., M.T., K.J.K., J.T.H., C.A., H.Y.Y., Y.Z. and S.C. J.R.H. implemented most of the computational analyses and statistical approaches with help from S.A.M., P.M., M.A., A.M., V.H., T.N.Y., S.M.G.E., T.M.S. and J.L. under the supervision of W.Z., T.v.d.K., T.J.P., M.F., P.C.B., R.G.B., H.H.H., or M.L. Figures were designed by S.Z. with assistance from J.R.H. and S.A.M. The manuscript was written by S.Z., J.R.H. and M.L. with assistance from all authors. M.L. oversaw the study.

2.1 Abstract

Prostate cancer (PCa) is the second most commonly diagnosed malignancy among men worldwide. Recurrently mutated in primary and metastatic prostate tumours, *FOXA1* encodes a pioneer transcription factor (TF) involved in disease onset and progression through both androgen receptor (*AR*)-dependent and *AR*-independent mechanisms. Despite its oncogenic properties however, the regulation of *FOXA1* expression remains unknown. Here, we identify a set of six *cis*-regulatory elements (CREs) in the *FOXA1* regulatory plexus harboring somatic single nucleotide variantss (SNVs) in primary prostate tumours. We find that deletion and repression of these CREs significantly decreases *FOXA1* expression and PCa cell growth. Six of the ten SNVs mapping to *FOXA1* regulatory plexus significantly alter the transactivation potential of CREs by modulating the binding of TFs. Collectively, our results identify CREs within the *FOXA1* plexus mutated in primary prostate tumours as potential targets for therapeutic intervention.

2.2 Introduction

PCa is the second most commonly diagnosed cancer among men with an estimated 1.3 million new cases worldwide in 2018 [1]. Although most men diagnosed with primary PCa are treated with curative intent through surgery or radiation therapy, treatments fail in 30% of patients within 10 years [2] resulting in a metastatic disease [3]. Patients with metastatic disease are typically treated with anti-androgen therapies, the staple of aggressive PCa treatment [4]. Despite the efficacy of these therapies, recurrence ultimately develops into lethal metastatic castration-resistant prostate cancer (mCRPC) [4]. As such, there remains a need to improve our biological understanding of PCa development and find novel strategies to treat patients. Sequencing efforts identified coding somatic SNVs mapping to FOXA1 in up to 9% [5–10] and 13% [9–11] of primary and metastatic PCa patients, respectively. These coding somatic SNVs target the Forkhead and transactivation domains of FOXA1 [12], altering its pioneering functions to promote PCa development [10, 13]. Outside of coding SNVs, whole genome sequencing (WGS) also identified somatic SNVs and indels in the 3' untranslated region (UTR) and C-terminus of FOXA1 in ~12% of mCRPC patients [14]. In addition to SNVs, the FOXA1 locus is a target of structural rearrangements in both primary and metastatic PCa tumours, inclusive of duplications, amplifications, and translocations [9, 10]. Taken together, FOXA1 is recurrently mutated taking into account both its coding and flanking noncoding sequences across various stages of PCa development.

FOXA1 serves as a pioneer TF that can bind to heterochromatin, promoting its remodelling to increase accessibility for the recruitment of other TFs [15]. FOXA1 binds to chromatin at cell-type specific genomic coordinates facilitated by the presence of mono- and dimethylated lysine 4 of histone H3 (H3K4me1 and H3K4me2) histone modifications [16, 17]. In PCa, FOXA1 is known to pioneer and reprogram the binding of *AR* alongside HOXB13 [18]. Independent from its role in *AR* signalling, FOXA1 also regulates the expression of genes involved in cell cycle regulation in PCa [19, 20]. For instance, FOXA1 co-localizes with CREB1 to regulate the transcription of genes involved in cell cycle processes, nuclear division and mitosis in mCRPC [19–25]. FOXA1 has also been shown to promote feed-forward mechanisms to drive disease progression [26, 27]. Hence, FOXA1 contributes to *AR*-dependent and *AR*-independent processes favouring PCa development.

Despite the oncogenic roles of FOXA1, therapeutic avenues to inhibit its activity in PCa are lacking. In the breast cancer setting for instance, the use of cyclin-dependent kinases inhibitors have been suggested based on their ability to block FOXA1 activity on chromatin [28]. As such, understanding the governance of FOXA1 messenger RNA (mRNA) expression offers an alternative strategy to find modulators of its activity. Gene expression relies on the interplay between distal CREs, such as enhancers and anchors of chromatin interaction, and their target gene promoter(s) [29]. These elements can lie tens to hundreds of kilobases (kbps) away from each other on the linear genome but physically engage in close proximity with each other in the three-dimensional space [30]. By measuring contact frequencies between loci through the use of chromatin conformation capture (3C)-based technologies, it enables the identification of regulatory plexuses corresponding to sets of CREs in contact with each other [31, 32]. By leveraging these technologies, we can begin to understand the three-dimensional organization of the PCa genome and delineate the FOXA1 regulatory plexus.

Here, we integrate epigenetics and genetics from PCa patients and model systems to delineate CREs establishing the regulatory plexus of FOXA1. We functionally validate a set of six mutated CREs that regulate FOXA1 mRNA expression. We further show that SNVs mapping to these CREs are capable of altering their transactivation potential, likely through modulating the binding of key PCa TFs.

2.3 Results

2.3.1 *FOXA1* is essential for prostate cancer proliferation

We interrogated *FOXA1* expression levels across cancer types. We find that *FOXA1* mRNA is consistently the most abundant in prostate tumours compared to 25 other cancer types across patients (Figure 2.1a), ranking in the 95th percentile for 492 of 497 prostate tumours profiled in The Cancer Genome Atlas (TCGA) (Figure A.1a). Using the same dataset we also find that *FOXA1* is the most highly expressed out of 41 other forkhead box (FOX) factors in prostate tumours (Figure A.1b). We next analyzed expression data from Cancer Dependency Map (DEPMAP) and observed *FOXA1* to be most highly expressed in PCa cell lines compared to cell lines of other cancer types (Figure A.2a). Amongst the eight PCa cell lines in the dataset (22Rv1, DU145, LNCaP, MDA-PCa-2B, NCI-H660, PrECLH, PC3, and VCaP), *FOXA1* mRNA abundance is above the 90th percentile in all but one cell line (PrECLH) compared to the $>$ 56,000 protein coding and non-protein coding genes profiled (Figure A.2b). These new results gained from the TCGA and DEPMAP validate previous understanding that *FOXA1* is one of the highest expressed genes in PCa [33].

Following up on *FOXA1* mRNA expression levels, we interrogated the essentiality of *FOXA1* for PCa cell growth. RNAi-mediated essentiality screens compiled in DEPMAP show that *FOXA1* lies in the 94th percentile across 6 of the 8 available PCa cell lines: 22Rv1, LNCaP, MDA PCa 2B, NCI-H660, PC3, and VCaP cells (Figure 2.1b-c). The median RNAi-mediated essentiality score for all prostate cell lines is significantly lower than all other cell lines, suggesting that *FOXA1* is especially essential for PCa cell proliferation (permutation test, $p = 1 \times 10^{-6}$, see Methods) (Figure A.3a). Growth assays in LNCaP and VCaP cells following *FOXA1* knockdown using two independent siRNAs (Figure 2.1d, Figure A.3b) show significant growth inhibition in LNCaP (siRNA #1: 4-fold, siRNA #2: 3.35-fold) and VCaP (siRNA #1: 8.7-fold, siRNA #2: 2-fold) cells five days post-transfection (Mann-Whitney U Test, $p < 0.05$; Figure 2.1e-f). In accordance with previous reports, our results using essentiality datasets followed by knockdown validation reveals that *FOXA1* is oncogenic and essential for PCa cell proliferation.

2.3.2 Identifying putative *FOXA1* CREs

The interweaving of distal CREs with target gene promoters establishes regulatory plexuses with some to be ascribed to specific genes [31, 32]. Regulatory plexuses stem from chromatin inter-

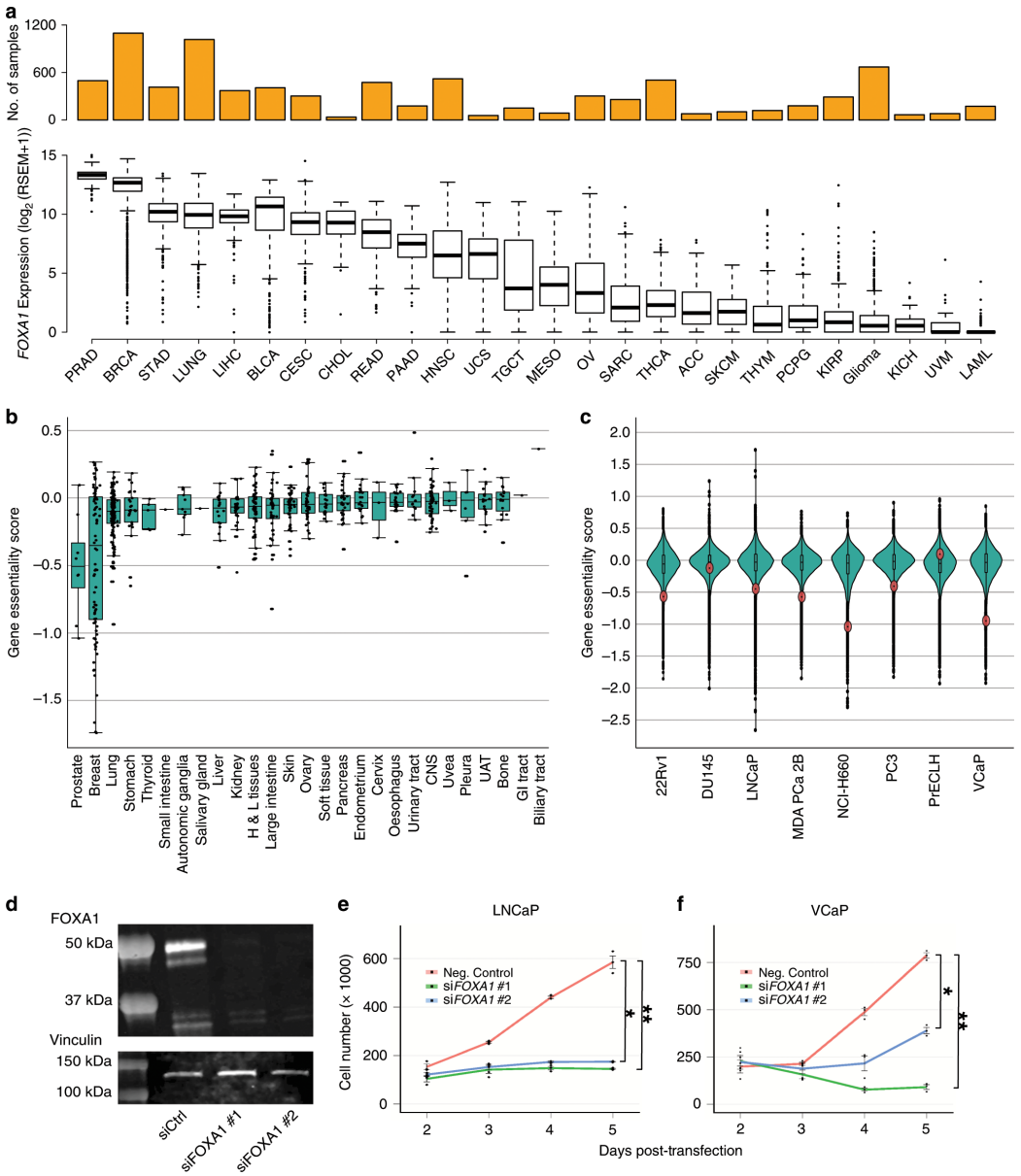

Figure 2.1: ***FOXA1* is highly expressed in PCa and essential for PCa cell proliferation..** **a.** The mRNA expression of *FOXA1* across tumour types ($n = 26$) from RNA-seq data of TCGA. **b.** *FOXA1* essentiality mediated through RNAi across various cell lines ($n = 707$) from DEPMAP. Gene essentiality scores are normalized z-scores. Higher scores indicate less essential, and lower scores indicate more essential for cell proliferation. x-axis indicate tissue of origin for each cell line tested. Each dot indicates one cell line. **c.** Gene essentiality mediated through RNAi across PCa cell lines ($n = 8$) from DEPMAP. Each dot indicates one gene, red indicates *FOXA1*. **d.** Representative Western blot against *FOXA1* in LNCaP cells 5 days post-transfection of non-targeting siRNA and two independent siRNA targeting *FOXA1*. **e.** Cell proliferation assay conducted in LNCaP cells upon siRNA-mediated knockdown of *FOXA1* across 5 days. **f.** Cell proliferation assay conducted in VCaP cells upon siRNA-mediated knockdown of *FOXA1* across 5 days. Error bars indicate ± s.d. $n = 3$ independent experiments. Mann-Whitney U test, * $p < 0.05$, ** $p < 0.01$.

actions orchestrated by various factors including ZNF143, YY1, CTCF and the cohesin complex [34–36]. Motivated by the oncogenic role of *FOXA1* in PCa, we investigated its regulatory plexus controlling its expression. According to chromatin contact frequency maps generated from Hi-C assays performed in LNCaP PCa cells, *FOXA1* lies in a 440 kbp topologically associated domain (TAD) (chr14: 37720002-38160000 ± 40 kbp adjusting for resolution) (Figure 2.2a). By overlaying DNase-seq data from LNCaP PCa cells, there are a total of 123 putative CREs reported as DNase I hypersensitive sitess (DHSs) that populate this TAD (Figure 2.2a). We next inferred the regulatory plexus of *FOXA1* using the C3D method [37]. C3D aggregates and draws correlation of DHS signal intensities between the cell line of choice and the DHS signal across all systems in the database [37]. Anchoring our analysis to the *FOXA1* promoter and using accessible chromatin regions defined in LNCaP PCa cells identified 55 putative CREs to the *FOXA1* regulatory plexus ($r > 0.7$) (Figure 2.2b).

2.3.3 Putative *FOXA1* CREs harbour TF binding sites and SNVs

To delineate the CREs that could be actively involved in the transcriptional regulation of *FOXA1*, we annotated the DHS with available ChIP-seq data for histone modifications and TFs conducted in LNCaP, 22Rv1, VCaP PCa cell lines and primary prostate tumours (Figure 2.2b) [18, 38]. Close to 60% (33/55) of the putative *FOXA1* plexus CREs are positively marked by H3K27ac profiled in primary prostate tumours [38], indicative of active CREs in tumours (Figure 2.2b) [39]. Next, considering that noncoding SNVs can target a set of CREs that converge on the same target gene in cancer [32], we overlapped the somatic SNVs called from WGS across 200 primary prostate tumours to the 33 H3K27ac-marked DHS predicted to regulate *FOXA1* [6, 40]. This analysis identified 6 out of the 33 DHS marked with H3K27ac (18.2%) harboring one or more SNVs (10 total SNVs called from 9 tumours) (Figure 2.2b). We observe that these 6 CREs can be bound by multiple TFs in PCa cells, including *FOXA1*, AR and *HOXB13* (Figure 2.2b, Figure A.4). The Hi-C data from the LNCaP PCa cells corroborates the C3D predictions as demonstrated by the elevated contact frequency between the region harboring the *FOXA1* promoter and where the 6 CREs are located, compared to other loci in the same TAD (Figure 2.3a). The 6 CREs lie in intergenic or intronic regions (Figure 2.3b-h). Together, histone modifications, TF binding sites and noncoding SNVs support that these 6 putative CREs are active in primary PCa. The Hi-C and C3D predictions suggest that they regulate *FOXA1* expression.

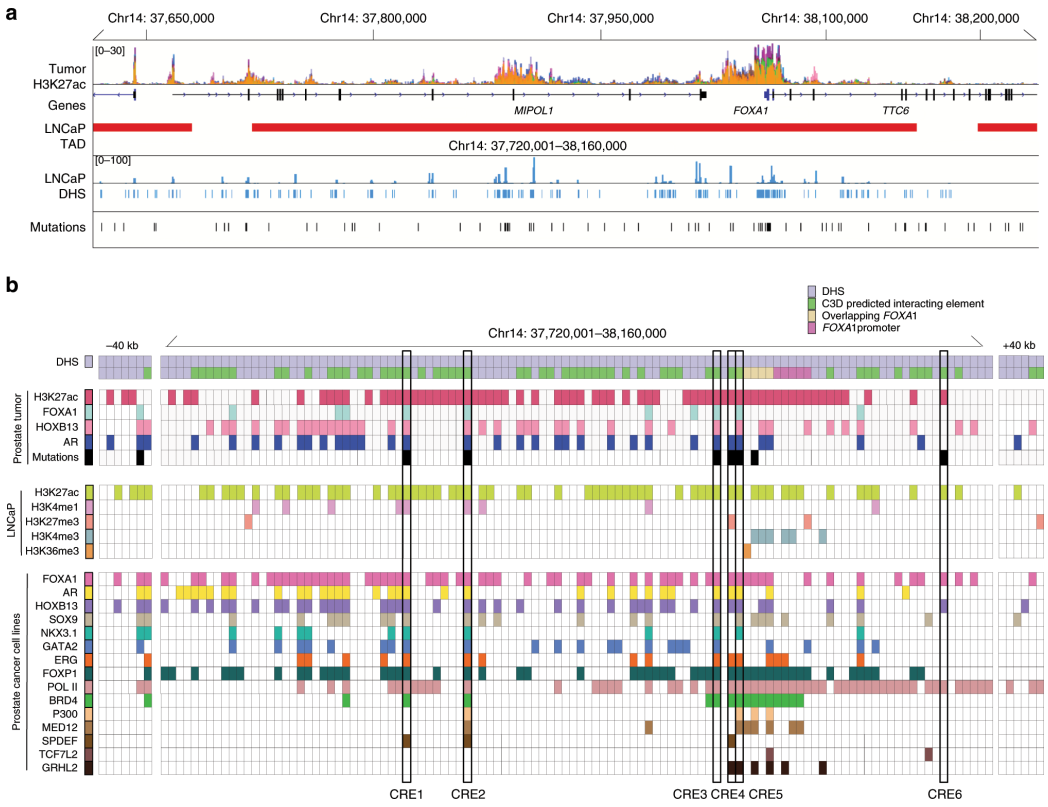

Figure 2.2: **Epigenetic annotation of 14q21.1 locus and identification of *FOXA1* CREs.. a.** Overview of cis-regulatory landscape surrounding *FOXA1* on the 14q21.1 locus. H3K27ac signal track is the ChIP-seq signal overlay of 19 primary prostate tumours. LNCaP Hi-C depicts the TAD structure around *FOXA1*. Mutations indicate SNVs identified in 200 primary prostate tumours. **b.** Functional annotation of putative *FOXA1* CREs using TF and histone modification ChIP-seq conducted in primary tumours and PCa cell lines. Annotated in the matrix are all DHS within the TAD and ± 40 kbp resolution left and right of the TAD. Putative *FOXA1* CREs targeted by noncoding SNVs for downstream validation are boxed.

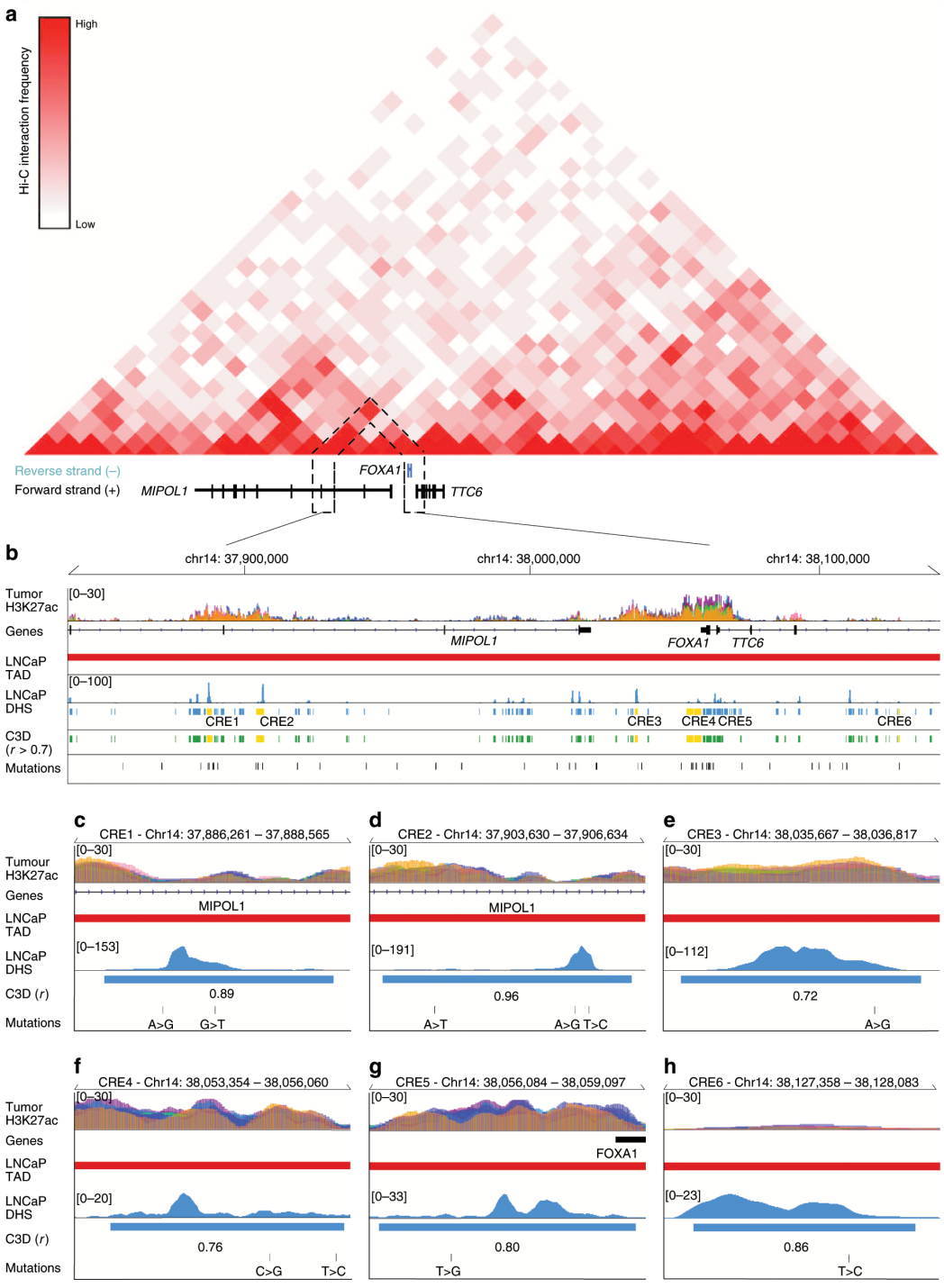

Figure 2.3: **Putative CREs predicted to interact with *FOXA1* promoter. a.** Hi-C conducted in LNCaP cells indicating physical interactions between putative *FOXA1* CREs and the *FOXA1* promoter. Hi-C resolution is 40 kbp. **b.** The six putative *FOXA1* CREs are coloured in yellow. **c-h.** Zoom-in of each individual putative *FOXA1* CRE. C3D value is the Pearson correlation of DHS signal between LNCaP and the DHS reference matrix.

2.3.4 Disruption of CREs reduces *FOXA1* mRNA expression

We next assessed the role of CREs toward *FOXA1* expression using LNCaP and 22Rv1 clones stably expressing the wild-type (WT) Cas9 protein (Figure 2.4a-b). Guide RNAs (gRNAs) designed against the *FOXA1* gene (exon 1 and intron 1) served as positive controls while an outside-TAD region (i.e termed Chr14 (-)), a region on a different chromosome (the human *AAVS1* safe-harbor site at the *PPP1R12C* locus [38, 41]), and three regions within the TAD predicted to be excluded from the *FOXA1* plexus served as negative controls. Individual deletion of the *FOXA1* plexus CREs through transient transfection of gRNAs into the LNCaP cells (See Methods) led to significantly decreased *FOXA1* mRNA expression (Δ CRE1 ~29.3 ± 8.3%, Δ CRE2 ~40.1 ± 11.8%, Δ CRE 3 ~30.6 ± 9.1%, Δ CRE4 ~23.6 ± 8.2%, Δ CRE5 ~25.3 ± 6.6%, Δ CRE6 ~24.5 ± 10.2% and Δ *FOXA1* (exon 1 and intron 1) ~87.4 ± 8.8% reduction relative to basal levels) (Figure 2.4c, Figure A.5a-f). In contrast, deletion of several negative control regions within the same TAD did not significantly reduce *FOXA1* mRNA level (Figure 2.4c, Figure A.5g-i). Similar results were observed in 22Rv1 PCa cells (Figure 2.4d). As each clone expressed Cas9 protein at different levels, there may be a difference between genome editing efficiencies between the clones. We compared the CRISPR/Cas9 on-target genome editing efficiency across the five LNCaP cell line-derived clones with the relative *FOXA1* mRNA levels, and indeed observe a significant inverse correlation across all CREs (Pearson's correlation $r = 0.49, p < 0.005$) (Figure A.6a) and agreeing trends for each individual CRE (Figure A.6b).

Complementary to our findings using the WT CRISPR/Cas9 system, we next generated four LNCaP and four 22Rv1 cell line-derived dCas9-KRAB fusion protein expressing clones (Figure 2.4e-f). Transient transfection of the same gRNAs used in the WT Cas9 experiments, targeting the six *FOXA1* plexus CREs into our dCas9-KRAB LNCaP clones significantly decreased *FOXA1* expression relative to basal levels (iCRE1 ~24.6 ± 6.2%, iCRE2 ~42.2 ± 10.8%, iCRE3 ~25.3 ± 9.2%, iCRE4 ~23.3 ± 4.3%, iCRE5 ~30.2 ± 3.4% and iCRE6 ~23.1 ± 8.1% reduction). Similarly, gRNAs targeting the dCas9-KRAB fusion protein to *FOXA1* decreased its expression (i*FOXA1* ~81.6 ± 11.8% reduction; Student's *t*-test, $p < 0.05$, Figure 2.4g). Analogous results were also observed in our four clonal 22Rv1 dCas9-KRAB cell lines (Student's *t*-test, $p < 0.05$, Figure 2.4h). Collectively, our results suggest that the six CREs control *FOXA1* expression.

We further assessed the regulatory activity of the six *FOXA1* plexus CREs by testing the consequent mRNA expression on other genes within the same TAD, namely *MIPOL1* and *MIPOL1.* Δ CRE1 and Δ CRE2 significantly reduced *MIPOL1* mRNA expression by ~38.4 ± 6.4% and ~48.4

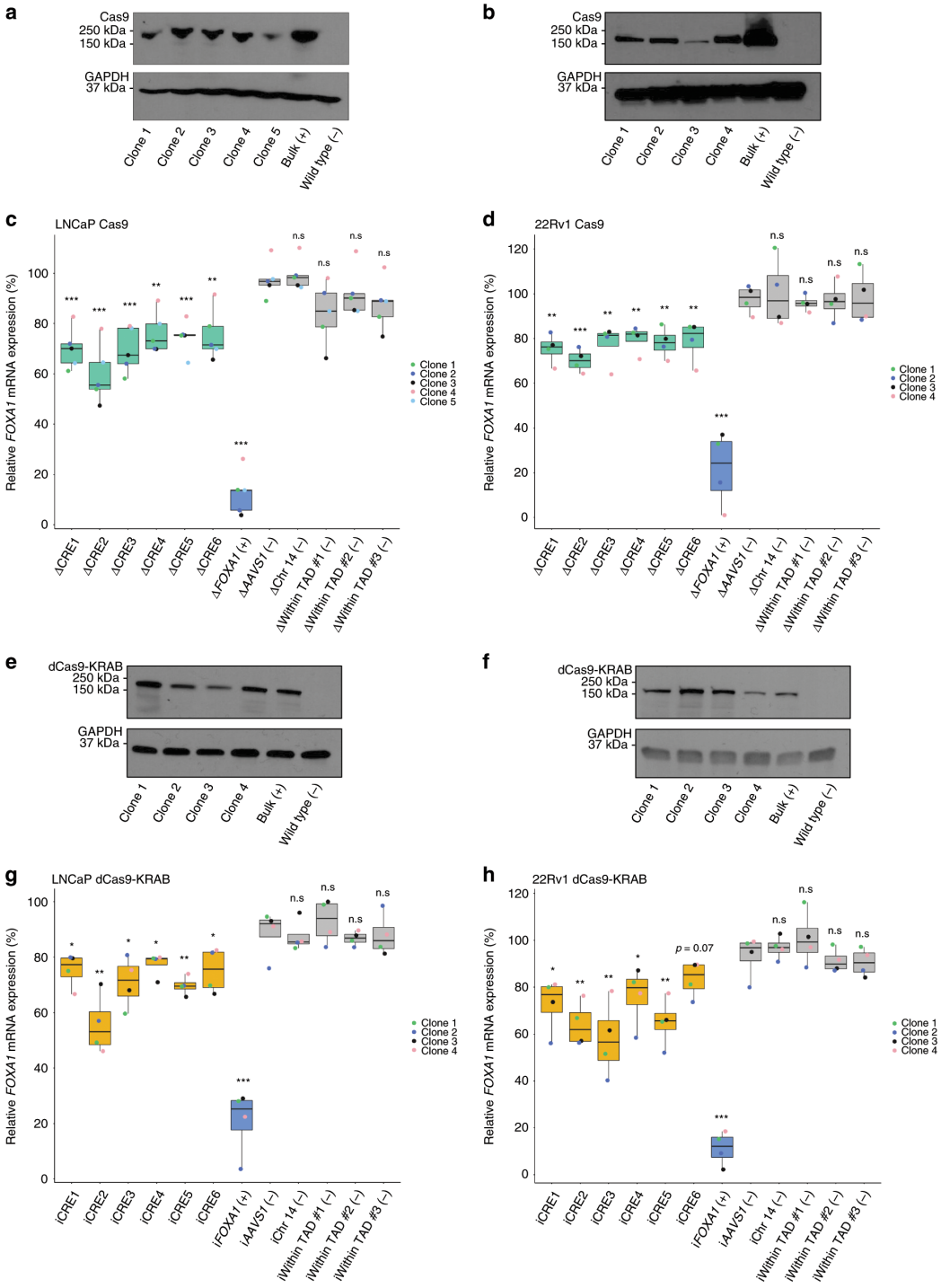

Figure 2.4: **Functional dissection of putative *FOXA1* CREs.** (Continued on the following page)

Figure 2.4: **a.** Representative western blot probed against Cas9 in LNCaP clones ($n = 5$ clones) derived to stably express Cas9 protein upon blasticidin selection. **b.** Representative western blot probed against Cas9 in 22Rv1 clones ($n = 4$ clones) derived to stably express Cas9 protein upon blasticidin selection. **c.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon CRISPR/Cas9-mediated deletion of each CRE using LNCaP clones ($n = 5$ independent experiments, each dot represents an independent clone). **d.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon CRISPR/Cas9-mediated deletion of each CRE using 22Rv1 clones ($n = 4$ independent experiments, each dot represents an independent clone). **e.** Representative western blot probed against Cas9 in LNCaP clones ($n = 4$ clones) derived to stably express the dCas9-KRAB fusion protein upon blasticidin selection. **f.** Representative western blot probed against Cas9 in 22Rv1 clones ($n = 4$ clones) derived to stably express dCas9-KRAB fusion protein upon blasticidin selection. **g.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon dCas9-KRAB-mediated repression of each CRE using LNCaP clones ($n = 4$ independent experiments, each dot represents an independent clone). **h.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon dCas9-KRAB-mediated repression of each CRE using 22Rv1 clones ($n = 4$ independent experiments, each dot represents an independent clone). *FOXA1* mRNA expression was normalized to basal *FOXA1* expression prior to statistical testing. Δ indicates CRISPR/Cas9-mediated deletion, i indicates dCas9-KRAB-mediated repression. Error bars indicate $\pm$ s.d. Student's t-test, n.s. not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

$\pm$ 9%, respectively relative to basal levels, whereas deletion of the other four CREs did not result in any significant *MIPOL1* expression changes (Student's t-test, $p < 0.05$, Figure A.7a). On the other hand, deletion of CREs each significantly reduced *MIPOL1* mRNA expression relative to its basal levels (Δ CRE1 ~52.9% $\pm$ 6.4%, Δ CRE2 ~66 $\pm$ 11.3%, Δ CRE3 ~55.5 $\pm$ 12.8%, Δ CRE4 44.9 $\pm$ 10.6%, Δ CRE5 43.1 $\pm$ 11.9% and Δ CRE6 52.2 $\pm$ 7.3% reduction (Student's t-test, p¡0.05, Figure A.7b), in agreement with the fact that *MIPOL1* shares its promoter with *FOXA1* as both genes are transcribed on opposing strands (Figure A.7c).

Reduction in *FOXA1* mRNA expression resulting from the deletion of *FOXA1* plexus CREs may also impact gene expression downstream of *FOXA1*, we assessed the mRNA expression of several *FOXA1* target genes, namely *SNAI2*, *ACPP*, and *GRIN3A*. Deletion of CREs resulted in significant change in *SNAI2* (up-regulation; Δ CRE1 ~190%, Δ CRE2 ~162.8%, Δ CRE3 ~147.5%, Δ CRE4 ~133.3%, Δ CRE5 ~137.3%, Δ CRE6 ~120.8%, Δ *FOXA1* ~266.7%), *ACPP* (down-regulation; Δ CRE1 ~73.5%, Δ CRE2 ~62.5%, Δ CRE3 ~69.6%, Δ CRE4 ~75.6%, Δ CRE5 ~70.9%, Δ CRE6 ~74.6%, Δ *FOXA1* ~52.2%) and *GRIN3A* expression (up-regulation; Δ CRE1 ~138.2%, Δ CRE2 ~168.8%, Δ CRE3 ~144.6%, Δ CRE4 ~132.1%, Δ CRE5 ~131.4%, Δ CRE6 ~127%, Δ *FOXA1* ~228%) (Student's t-test, $p < 0.05$, Figure A.7d-f). Collectively, our results support the restriction of most *FOXA1* plexus CREs towards *FOXA1* and its target genes.

2.3.5 *FOXA1* CREs collaborate to regulate its expression

Expanding on the idea that multiple CREs can converge to regulate the expression of a single target gene [31, 32, 42], we asked whether the CREs we identified collaboratively regulate *FOXA1* mRNA expression. Here, we applied a transient approach that delivers Cas9 protein:gRNA as a ribonucleoprotein (RNP) complex formed prior to transfection that would avoid the heterogeneity of Cas9 protein expression across the PCa cell clones (See Methods). We first validated this system through single CRE deletions, where we transiently transfected a set of gRNA targeting the CRE of interest. In accordance with data from our PCa cell clones stably expressing WT Cas9 and dCas9-KRAB, individual CRE deletion resulted in a significant reduction in *FOXA1* mRNA expression: (Δ CRE1 ~29.3 $\pm$ 7.3%, Δ CRE2 ~36 $\pm$ 11.8%, Δ CRE3 ~30.6 $\pm$ 12.7%, Δ CRE4 ~24.5 $\pm$ 6.1%, Δ CRE5 ~23.7 $\pm$ 13.2%, Δ CRE6 ~26.8 $\pm$ 14.2% and Δ *FOXA1* ~96.2 $\pm$ 1.4% reduction (Student's t-test, $p < 0.05$, Figure 2.5a, Figure A.8a-f). Next for combinatorial deletions, we prioritized the CREs that harbor more than 1 SNV (i.e CRE1, CRE2, CRE4), and transiently transfected RNP complexes that target both CREs in various combinations (i.e CRE1 + CRE2, CRE1 + CRE4, CRE2 + CRE4), and assessed *FOXA1* mRNA expression. Compared to negative control regions, the combinatorial deletion of Δ CRE1 + Δ CRE2, Δ CRE1 + Δ CRE4, and Δ CRE2 + Δ CRE4 resulted in a significant ~48.5 $\pm$ 4.5%, ~50.4 $\pm$ 2.9% and ~45.2 $\pm$ 5.5% reduction in *FOXA1* mRNA expression, respectively (Student's t-test, $p < 0.05$, Figure 2.5b, Figure A.9a-f) a fold reduction greater than single CRE deletions (Student's t-test, Figure A.10, $p < 0.05$). These results together demonstrate that these CREs collaboratively contribute to the establishment and regulation of *FOXA1* expression in PCa.

2.3.6 Disruption of *FOXA1* CREs reduces prostate cancer cell growth

As *FOXA1* is essential for PCa growth (Figure 2.1b-e), we next sought to assess the importance of the six *FOXA1* plexus CREs towards PCa cell growth. We adapted a lentiviral-based approach that expressed both the Cas9 protein and two gRNA that target each CRE for deletion (See Methods). Upon lentiviral transduction with subsequent selection, we separated LNCaP PCa cells for RNA, DNA and for cell proliferation. We first tested the system by measuring *FOXA1* mRNA expression, and independently observed significant reductions of *FOXA1* mRNA expression (Δ CRE1 ~18%, Δ CRE2 ~30%, Δ CRE3 ~15%, Δ CRE4 ~12%, Δ CRE5 ~35%, Δ CRE6 ~46% and Δ *FOXA1* (exon 1 and intron 1) ~48% reduction (Student's t-test, $p < 0.05$, Figure 2.5c, Figure A.11a-f). We then seeded these cells at equal density. Six days post-seeding, we harvested the cells and observed

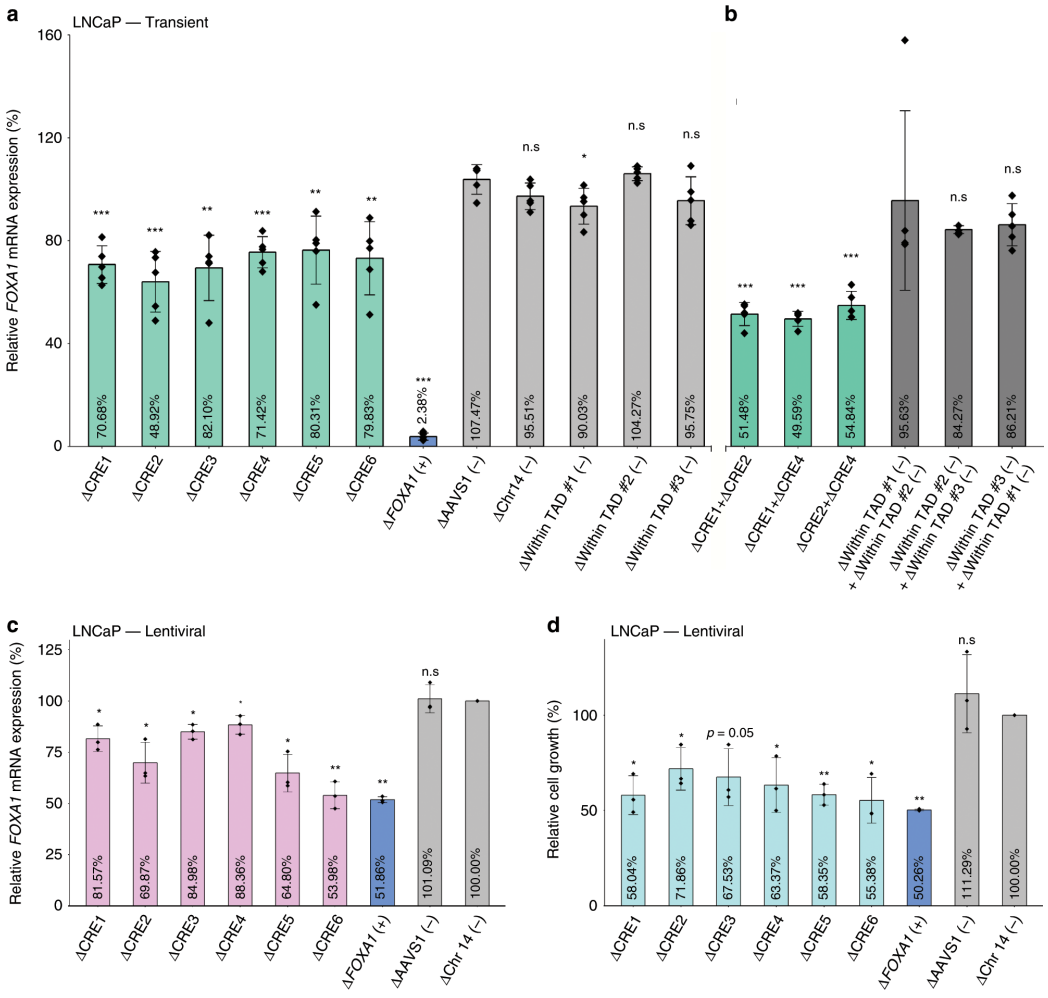

Figure 2.5: ***FOXA1* CREs collaborate to regulate its expression and are critical for PCa cell proliferation. a.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon transient transfection-based CRISPR/Cas9-mediated deletion of CRE1, CRE2, CRE4, and sequential deletion combinations ($n = 5$ independent experiments). **b.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon bulk lentiviral-based CRISPR/Cas9-mediated deletion of each CRE in LNCaP cells ($n = 3$ independent experiments). **c.** Cell proliferation assay conducted after puromycin and blasticidin selection for LNCaP cells carrying deleted regions of interest. Data was based on cell counting 6 days after seeding post-selection ($n = 3$, representative of three independent experiments). *FOXA1* mRNA expression upon deletion was normalized to basal *FOXA1* expression prior to statistical testing. *FOXA1* mRNA expression was normalized to the basal LNCaP *FOXA1* expression prior to statistical testing. Δ indicates CRISPR/Cas9-mediated deletion. Error bars indicate ± s.d. Student's *t*-test, n.s. not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

a significant reduction in cell growth upon deleting any of the six *FOXA1* plexus CREs (Δ CRE1 ~42%, Δ CRE2 ~28%, Δ CRE3 ~33%, Δ CRE4 ~27%, Δ CRE5 ~42%, Δ CRE6 ~44% and Δ *FOXA1* (exon 1 and intron 1) ~50% reduction (Student's *t*-test, $p < 0.05$, Fig 5d). These results suggest that the six *FOXA1* plexus contribute to PCa etiology, in agreement with their ability to regulate *FOXA1* expression and the essentiality of this gene in PCa cell growth.

2.3.7 SNVs mapping to *FOXA1* CREs can alter their activity

SNVs can alter the transactivation potential of CREs [32, 43–51]. In total, we found 10 SNVs called from 9 out of the 200 tumours that map to the six *FOXA1* plexus CREs (Figure 2.6a). To assess the impact of these noncoding SNVs, we conducted luciferase assays comparing differential reporter activity between the variant and the WT allele of each CRE (Figure 2.6b-k). We found that the variant alleles of 6 of the 10 SNVs displayed significantly greater luciferase reporter activity when compared to the WT alleles (Mann-Whitney U test, $p < 0.05$). Specifically, we observed the following fold-changes: chr14:37,887,005 A > G (1.65-fold), chr14:37,904,343 A > T (1.35-fold), chr14:37,905,854 A > G (1.28-fold), chr14:37,906,009 T > C (1.71-fold), chr14:38,036,543 A > G (1.44-fold), chr14:38,055,269 C > G (1.39-fold) (Figure 2.6b, d-h). These results indicate that these SNVs can alter the transactivation potential of *FOXA1* plexus CREs in PCa cells.

2.3.8 SNVs mapping to *FOXA1* CREs can modulate the binding of TFs

We next assessed if the changes in transactivation potential induced by noncoding SNVs related to changes in TF binding to CREs by allele-specific ChIP-qPCR [32, 44, 51] in LNCaP PCa cells. We observed differential binding of *FOXA1*, *AR*, *HOXB13*, *GATA2* and *FOXP1* for the chr14:37887005 (A > G) SNV found in CRE1; the chr14:37904343 (A > T), chr14:37905854 (A > G) and chr14:37906009 (T > C) SNVs found in CRE2; and the chr14:38055269 (C > G) SNV found in CRE4 (Student's *t*-test, $p < 0.05$, Figure 2.6l-p). In contrast, *SOX9* and *NKX3.1* binding was unaffected by these SNVs (Figure 2.6l-q). Compared to the WT sequence, chr14:37,887,005 A > G significantly increased *AR* binding (1.31-fold increase), *GATA2* binding (1.25-fold increase) and *FOXP1* binding (1.23-fold increase); chr14:37,904,343 A > T significant increased *AR* binding (1.30-fold increase), *GATA2* (1.25-fold increase) and *FOXP1* (1.33-fold increase); chr14:37,905,854 A > G significantly increased *FOXA1* binding (1.41-fold increase) and *AR* binding (1.33-fold increase); chr14:37,906,009 T > C significantly increased the binding of *FOXA1* (1.29-fold increase), *AR* (1.31-fold increase), *HOXB13* (1.13-fold increase) and *FOXP1* (1.25-fold increase); and chr14:38,055,269

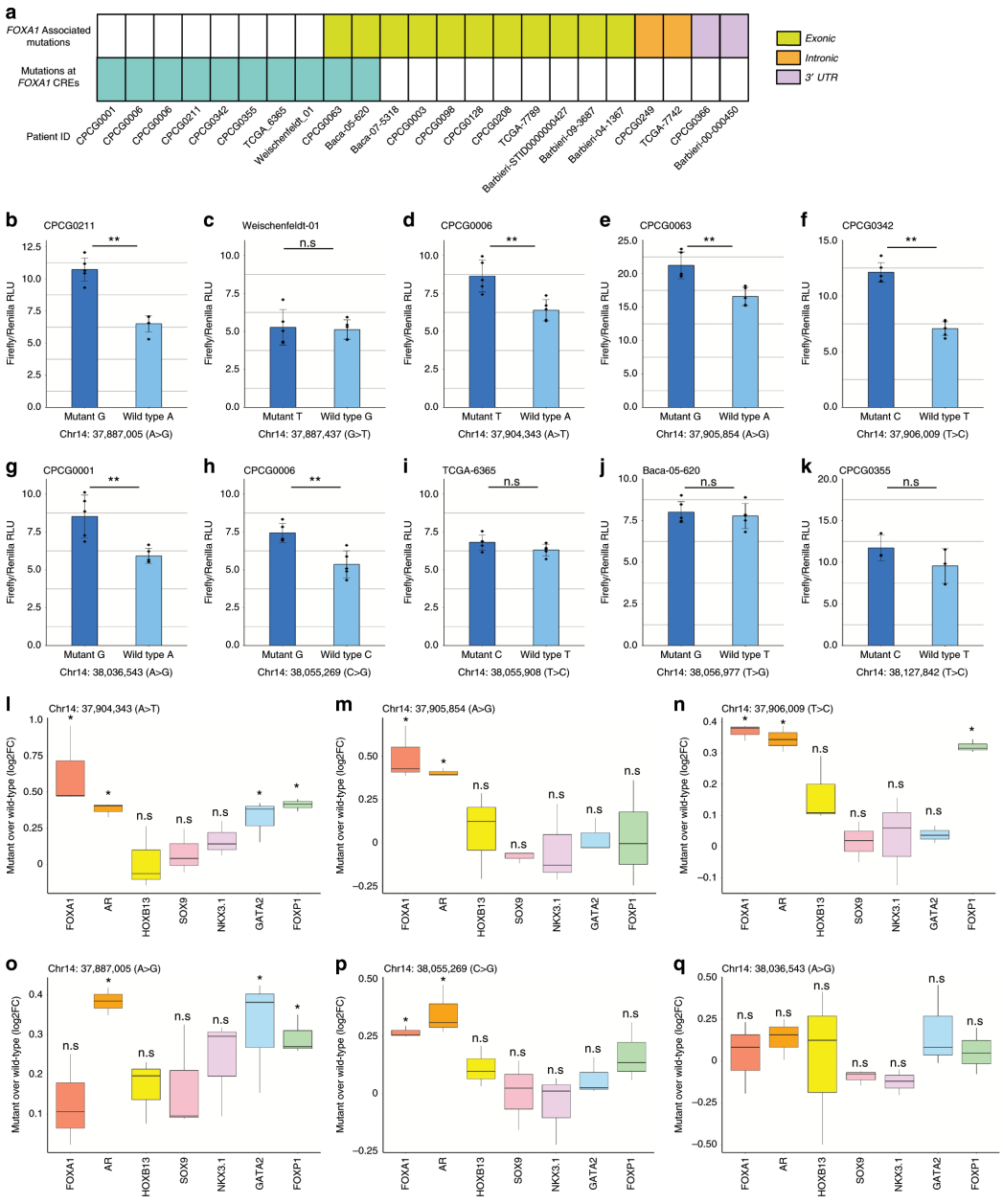

Figure 2.6: **A subset of noncoding SNVs mapping to the *FOXA1* CREs are gain-of-function. a.** Matrix showcasing the patients from the CPC-GENE dataset that harbour SNVs at the *FOXA1* CREs, exons, introns, and the 3' UTR of *FOXA1*. **b-k.** Luciferase assays are conducted in LNCaP cells. Bar plot showcases the mean firefly luciferase activity normalized by *renilla* luciferase activity. Error bars indicate ± s.d. $n = 5$ independent experiments for all CREs except for chr14:38,127,842 T > C where $n = 3$. Each diamond represents an independent experiment. Hypothesis testing done with Mann-Whitney U test. **l-q.** Allele-specific ChIP-qPCR conducted on plasmids carrying the WT or variant sequence upon transient transfection in PCa cells. Data is presented as $\log_2$ fold-change of variant sequence upon comparison to WT sequence ($n = 3$ independent experiments per ChIP). Hypothesis testing don with Student's t-test, n.s. not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

C > G significantly increased *FOXA1* binding (1.20-fold increase). Notably all six SNVs increased the binding of the TFs known to bind at these CREs. In contrast, none of the SNVs significantly decreased the binding of these TFs. Our observations suggest that gain-of-function SNVs populate the *FOXA1* plexus CREs.

2.4 Discussion

Modern technologies and understanding of the epigenome allow the possibility of probing CREs involved in regulating genes implicated in disease. Despite FOXA1 being recurrently mutated [5–8, 11] and playing potent oncogenic roles in PCa etiology [9, 10, 13], the CREs involved in its transcriptional regulation are poorly understood. Understanding how FOXA1 is expressed can provide a complementary strategy to antagonize FOXA1 in PCa.

We used the DHS profiled in PCa cells to identify putative FOXA1 CREs through annotating these regions with five different histone modifications, TF binding sites and noncoding SNVs profiled in PCa cells and primary prostate tumours. Our efforts identified and validated a set of six active CREs involved in FOXA1 regulation, agreeing with a recent report where a subset of our CREs map to loci suggested to be in contact with the FOXA1 promoter [52]. The disruption of these six distal CREs each significantly reduced FOXA1 mRNA levels, similar to what has been demonstrated for ESR1 in luminal breast cancer [32], MLH1 in Lynch syndrome [53], MYC in lung adenocarcinoma and endometrial cancer [54], and *AR* in mCRPC [55, 56]. Through combinatorial deletion of two CREs, FOXA1 mRNA levels were further reduced in comparison with single CRE deletions, raising the possibility of CRE additivity [57]. The deletion of the FOXA1 plexus CREs also significantly reduced PCa cell proliferation at levels comparable to what has been reported upon deletion of the amplified CRE upstream of the *AR* gene in mCRPC [55], suggestive of onco-CREs as reported in lung [54] and prostate [55] cancer.

More than 90% of SNVs found in cancer map to the noncoding genome [58, 59] with a portion of these SNVs mapping to CREs altering their transactivation potential [32, 44–46] and/or downstream target gene expression [48, 58, 60]. We extended this concept with SNVs identified from primary prostate tumours mapping to FOXA1 plexus CREs. We observed that a subset of these SNVs can alter transactivation potential by modulating the binding of specific TFs whose cistromes are preferentially burdened by SNVs in primary PCa [59]. Our findings complement recent reports of SNVs found in the noncoding space of FOXA1 that could affect its expression [14, 61]. The FOXA1 plexus CREs we identified here are also reported to be target of structural variants (SVs) in both the

primary and metastatic settings [9, 62], including tandem duplication in ~14% (14/101) mCRPC tumours over CRE2 [62], amplification, duplication and translocation over CRE3, CRE4, and CRE5 [9]. Notably, the translocation and duplication defining the FOXMIND enhancer driving FOXA1 expression reported in primary and metastatic settings harbors the CRE3 element we characterized [9]. Collectively, these studies combined with our discoveries reveal the fundamental contribution of the FOXA1 plexus in PCa etiology. As a whole, our findings in conjunction with recent reports suggest that CREs involved in the transcriptional regulation of FOXA1 may be hijacked in prostate tumours through various types of genetic alterations.

Despite initial treatment responses from treating aggressive primary and metastatic PCa through castration to suppress *AR* signalling [4], resistance ensues as 80% of mCRPC tumours harbor either *AR* gene amplification, amplification of a CRE upstream of *AR*, or activating *AR* coding mutations [11, 55, 62]. Given the *AR*-dependent [15, 18] and *AR*-independent [25] oncogenic activity of FOXA1 in PCa, its inhibition is an appealing alternative therapeutic strategy. Our dissection of the FOXA1 cis-regulatory landscape complement recent findings through revealing loci that are important for the regulation of FOXA1. Theoretically, direct targeting of the CREs regulating FOXA1 would down-regulate FOXA1 levels and could therefore serve as a valid alternative to antagonize its function.

Taken together, we identified FOXA1 CREs targeted by SNVs that are capable of altering trans-activation potential through the modulation of key PCa TFs. The study supports the importance of considering CREs not only as lone occurrences but as a team that works together to regulate their target genes, particularly when considering the impact of genetic alterations. As such, our work builds a bridge between the understanding of FOXA1 transcriptional regulation and new routes to FOXA1 inhibition. Aligning with recent reports [9, 10, 13], our findings support the oncogenic nature of FOXA1 in PCa. Gaining insight on the cis-regulatory plexuses of important genes such as FOXA1 in PCa may provide new avenues to inhibit other drivers across various cancer types to halt disease progression.

2.5 Methods

2.5.1 Cell Culture

LNCaP and 22Rv1 cells were cultured in RPMI medium, and VCaP cells were cultured in DMEM medium, both supplemented with 10% FBS, and 1% penicillin-streptomycin at 37 °C in a humidified incubator with 5% CO2. These PCa cells originated from ATCC. 293FT cells were purchased from

ThermoFisherScientific (Cat No. R70007) maintained in complete DMEM medium (DMEM with 10% FBS (080150, Wisent), L-glutamine (25030-081, ThermoFisher) and non-essential amino acids (11140-050, ThermoFisher) supplemented with 50mg/mL Geneticin (4727894001, Sigma-Aldrich). The cells are regularly tested for Mycoplasma contamination. The authenticity of these cells was confirmed through Short Tandem Repeat profiling.

2.5.2 Prostate tumours and cancer cell lines expression

Cancer cell line mRNA abundance data were collected from DEPMAP; https://depmap.org/portal/; RNA-seq TPM values from 2018q4 version with all 5 non-cancer cell lines were removed) [63] projects. Prostate tumour mRNA abundance data was collected from TCGA prostate cancer (TCGA-PRAD) project via the Xena Browser (https://xenabrowser.net/; dataset description: TCGA prostate adenocarcinoma gene expression by RNA-seq (polyA+ Illumina HiSeq; RSEM)).

2.5.3 Prostate cancer cell line gene essentiality

Essentiality scores were collected from the DEPMAP Project [64]. To compare gene essentiality between PCa cell lines and others, essentiality scores for FOXA1 were collected from all available cell lines ($n = 707$). To perform a permutation test, the median of 8 randomly selected cell lines was calculated one million times to generate a background distribution of essentiality scores across all cell types available. The median essentiality score from the 8 PCa cell lines was calculated and its percentile within the background distribution is reported.

2.5.4 siRNA knockdown and cell proliferation assay

300,000 LNCaP cells (Day 0) were reverse transfected with siRNA (siFOXA1 using Lipofectamine®mRNAmax reagent (ThermoFisher Scientific, Cat No. 13778150)). Cells were counted using Countess™automated cell counter (Invitrogen). Whole cell lysates LNCaP cells after siRNA-mediated FOXA1 knockdown was collected at 96-hours post-transfection in RIPA buffer. Protein concentrations were determined through the bicinchoninic acid method (ThermoFisher Scientific, Cat No. 23225). Then 25 μg of lysate was subjected to SDS-PAGE. Upon completion of SDS-PAGE, protein was transferred onto PVDF membrane (Bio-Rad, Cat No. 1704156). The membrane was blocked with 5% non-fat milk for one hour at room temperature with shaking. After blocking, anti-FOXA1 (Abcam Cat No. 23737) in 2.5% non-fat milk was added, and was incubated at 4 °C overnight. Next day, the blot was washed and incubated with IRDye®800CW Goat Anti-Rabbit

IgG secondary antibody (LI-COR, Cat No. 925-32211) at room temperature for 1 hour. The blot was then washed and assessed with the Odyssey®CLX imaging system (LI-COR).

2.5.5 Identifying putative FOXA1 CREs

Putative FOXA1 CREs were identified through the use of C3D method based on DNase I Hypersensitivity [37]. Predicted interacting DHS with a Pearson's correlation above 0.7 [65] were kept for downstream analysis.

2.5.6 Hi-C and TADs in LNCaP cells

Hi-C and TADs conducted and called, respectively, in LNCaP cells are publicly available off ENCODE portal (ENCSR346DCU). Visualization of the Hi-C dataset is available on the Hi-C Browser [66].

2.5.7 Clonal wild-type Cas9 and dCas9-KRAB mediated validation

Lentiviral particles were generated in 293FT cells (ThermoFisher) using the pMDG.2 and psPAX2 packaging plasmids (Addgene; #12259 and #12260, a gift from Didier Trono) alongside the Lenti-Cas9-2A-Blast plasmid (Addgene #73310, a gift from Jason Moffat) and collected 72 hrs post transfection. LNCaP and 22Rv1 cells were then transduced for 24-48 hours with equal amounts of virus followed by selection with media containing blasticidin (7.5 μg/mL for LNCaP cells, 6 μg/mL for 22Rv1 cells). Upon selection, clones were derived by serial dilution with subsequent single cell seeding into 96-well plates containing selection media. Cas9 protein expression for each clone was then assessed through Western blotting (1 °Ms-Cas9 (Cell Signalling Technology, Cat No. #14697) 1:1000, Ms-GAPDH 1:5000 (Santa Cruz Biotechnology, Cat No. #sc47724) in 5% non-fat milk; 2 °HRP-linked Anti-Mouse IgG (Cell Signalling Technology, Cat No. #7076S) 1:10 000 in 2.5% non-fat milk. The full unprocessed blot is in the Source Data File.

Lentiviral particles were generated in 293FT cells (ThermoFisher) using the pMDG.2 and psPAX2 packaging plasmids (Addgene; #12259 and #12260, a gift from Didier Trono) alongside the Lenti-dCas9-KRAB-blast plasmid (Addgene #89567, a gift from Gary Hon) and collected 72 hrs post transfection. LNCaP and 22Rv1 cells were then transduced for 24-48 hours with equal amounts of virus followed by selection with media containing blasticidin (7.5 μg/mL for LNCaP cells, 6 μg/mL for 22Rv1 cells). Upon selection, clones were derived by serial dilution with subsequent single cell seeding into 96-well plates containing selection media. dCas9-KRAB protein expression for each

clone was then assessed through Western blotting (1 °Ms-Cas9 (Cell Signalling Technology, Cat No. #14697) 1:1000, Ms-GAPDH 1:5000 (Santa Cruz Biotechnology, Cat No. #sc47724) in 5% non-fat milk; 2 °HRP-linked Anti-Mouse IgG (Cell Signalling Technology, Cat No. #7076S) 1:10 000 in 2.5% non-fat milk. The full unprocessed blot is in the Source Data File.

For gRNA design, five to six unique CRISPR RNA (crRNA) molecules (Integrated DNA Technologies) were designed to tile across the region of interest using the CRISPR (http://crispor.tefor.net/) [67] and the Zhang lab CRISPR Design tools (http://crispr.mit.edu/) [68]. See published manuscript for gRNA. Each crRNA and trans-activating CPRISR RNA (tracrRNA) (Integrated DNA Technologies) were duplexed according to company supplier protocol to a concentration of 50 μM. Upon generation of the clones, six guides (crRNA-tracrRNA duplexes) for each region of interest were pooled into a single tube (1μL each guide, 6 μL per reaction) (Integrated DNA Technologies). Lastly, 1 μL (100 μM) of electroporation enhancer (Integrated DNA Technologies) was added to the mix (7 μL total) prior to transfection. The entire transfection reaction was transfected into 350 000 cells through Nucleofection (SF Solution EN120 - 4D Nucleofector, Lonza). Cells were then harvested 24 hours post-transfection for RNA and DNA for RT-PCR and confirmation of deletion, respectively.

2.5.8 Transient Cas9-mediated disruption of CREs

Deletion of elements through this method were achieved through the transfection of Cas9 nuclease protein complexed with the crRNA (Integrated DNA Technologies). Briefly, five to six unique crRNA molecules (Integrated DNA Technologies) were designed to tile across the region of interest using the CRISPR (http://crispor.tefor.net/) [67] and the Zhang lab CRISPR Design tools (http://crispr.mit.edu/) [68]. Each crRNA and tracrRNA (Integrated DNA Technologies) were duplexed according to company supplier protocol to a concentration of 50 μM. The six crRNA-tracrRNA duplexes were pooled into a single tube (6 μL per reaction), prior to adding 1 μL (5 μg) of Alt-R ®S.p HiFi Cas9 Nuclease 3NLS (Integrated DNA Technologies). The reaction was incubated at room temperature for 10 minutes for ribonucleoprotein (RNP) complex formation. Lastly, 1 μL (100 μM) of electroporation enhancer (Integrated DNA Technologies) was added to the mix prior to transfection. The entire transfection reaction was transfected into 350 000 cells through Nucleofection (SF Solution EN120 - 4D Nucleofector, Lonza). Cells were then harvested 24 hours post-transfection for RNA and DNA for RT-PCR and confirmation of deletion, respectively. For double deletions, two sets of gRNA-RNP complex (10 μg of Alt-R ®S.p HiFi Cas9 Nuclease 3NLS) were transfected and harvested 24 hours post-transfection for RNA and DNA for RT-PCR and

confirmation of deletion, respectively. To control for double deletions, two negative control regions within the TAD were also compounded. Due to size, see published manuscript for primers.

2.5.9 RT-PCR assessment of gene expression upon deletion of CREs

DNA and RNA were harvested with Qiagen AllPrep RNA/DNA Kit (Qiagen, Cat No. 80204). Next, cDNA was synthesized from 300 ng of RNA using SensiFast cDNA Synthesis kit (Bioline, Cat No. BIO-65054), and mRNA expression levels for various genes of interest were assessed. Due to size, see published manuscript for the primer sequences used for expression evaluation. Differential gene expression was calculated upon normalization with TBP (housekeeping gene). Statistical significance was calculated using Student's t-test in R.

2.5.10 Confirmation of Cas9-mediated deletion of CREs

Deletion of CREs were confirmed through PCR amplification of the intended region for deletion, followed by the T7 Endonuclease Assay (Integrated DNA Technology). Due to size, see published manuscript for primer sequences used for PCR amplification. PCR products were then loaded onto a 1% agarose gel for visualization. The agarose gel to assess the on-target genome editing efficiency was done through densitometry using ImageJ. The correlation between on-target genome editing efficiency and FOXA1 mRNA expression reduction was drawn through Pearson's correlation in R.

2.5.11 Cell proliferation upon deletion of FOXA1 CREs

Pairs of gRNAs flanking the CREs of interest, FOXA1 promoter and control regions were designed using CRISPOR (http://crispor.tefor.net/) and Zhang lab CRISPR Design tool (http://crispr.mit.edu/) (due to size, see published manuscript). Each pair of gRNAs were cloned into the lentiCRISPRv2 (Addgene; a gift from Feng Zhang #52961) and the lentiCRISPRv2-Blast (Addgene; a gift from Feng Zhang #83480) plasmid as previously described [69]. Lentiviral particles were generated in 293FT cells (ThermoFisher) using the pMDG.2 and psPAX2 packaging plasmids (Addgene; #12259 and #12260, a gift from Didier Trono), and collected 72 hrs post transfection. LNCaP cells were transduced for 24-48hrs with equal amounts of virus, followed by selection with media containing puromycin (3.5 μg/mL, ThermoFisher) and blasticidin (7 μg/mL, Wisent). Cells were harvested upon selection for RNA and DNA for RT-PCR and confirmation of DNA cleavage, respectively. For cell proliferation, cells were seeded at equal density per well (on a 96-well plate; Day 1) upon puromycin and blasticidin selection. Growth of the cells were monitored through cell counting us-

ing Countess ™automated cell counter (Invitrogen). Cell numbers were calculated as a percentage compared to negative control. Statistical significance was calculated using Student's t-test.

2.5.12 Luciferase reporter assays

Each region of interest was ordered as gBlocks from Integrated DNA Technologies. The regions were cloned into the BamHI restriction enzyme digest site of the pGL3 promoter plasmid (Promega). On Day 0, 90 000 LNCaP cells were seeded in 24-well plates. Next day (Day 1), pGL3 plasmids harboring the WT and variant sequences were co-transfected with the pRL Renilla plasmid (Promega) using Lipofectamine 2000. 48-hours later, the cells were harvested, and dual luciferase reporter assays were conducted (Promega). Notably, inserts of both forward and reverse directions were tested using this assay as enhancer elements are known to be direction-independent. Final luminescence readings are reported as firefly luciferase normalized to renilla luciferase activity. The assessment of each mutation was conducted in five biological replicates. Statistical significance was assessed by Mann-Whitney U test in R. See published manuscript for gBlock sequences.

2.5.13 Allele-specific ChIP-qPCR

Briefly, pGL3 plasmids containing the WT sequence and the mutant sequence used in the luciferase reporter assay were transfected into 7 million cells (2 μg per allele, per 1 million cells) using Lipofectamine 2000 (ThermoFisher Scientific), per manufacturer's instructions. Next day, each antibody (FOXA1 5 μg, Abcam, ab23738; *AR* 5 μg, Abcam, ab1083241; HOXB13 5 μg, Abcam, ab201682; SOX9 5 μg, Abcam, ab3697; GATA2 5 μg, Abcam, ab22849; FOXP1 5 μg, Abcam, ab16645; NKX3.1 10 μl, Cell Signalling Technology, #83700) was conjugated with 10 uL of each Dynabeads A and G (Thermo Fisher Scientific) for each ChIP for 6 hours with rotation at 4 °C. When antibody-beads conjugates were ready for use, cells were lifted using trypsin and fixed by re-suspending with 300 uL of 1% formaldehyde in PBS for 10 minutes at room temperature. 2.5M Glycine was added to quench excess formaldehyde (final concentration 0.125 M). Cells were then washed with cold PBS and lysed using 300 uL of Modified RIPA buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA; 140 mM NaCl; 1% Triton X-100; 0.1% SDS; 0.1% sodium deoxycholate) supplemented with protease inhibitor. The lysate was subject to 25 cycles of sonication (30s ON 30s OFF) using Diagenode Bioruptor Pico (Diagenode). 15 uL of sonicated lysate was set aside as input with the rest used for chromatin pulldown through addition of antibody-beads conjugates and overnight incubation at 4 °C with rotation. Next day, the beads were washed once with Modified RIPA

buffer, washed once with Modified RIPA buffer + 500 mM NaCl, once with LiCl buffer (10 mM TrisHCl, pH 8.0; 1 mM EDTA; 250 mM LiCl; 0.5% NP-40; 0.5% sodium deoxycholate) and twice with Tris-ETDA buffer (pH 8). After washes, beads and input were de-crosslinked by addition of 100 μL De-crosslinking buffer and incubation at 65 °C for 6 hours. Samples were then purified and eluted. ChIP and input DNA were then used for allele-specific ChIP-qPCR using MAMA primers as described previously. Fold-change significance was calculated using Student's t-test in R.

All analyses were done using hg19 reference genome coordinates.

2.6 Data availability

Genomic and Epigenomic data sets used to support this study can be found from the following accession codes: primary tumors—H3K27ac ChIP-seq (GSE96652), SNVs called from primary tumors (https://dcc.icgc.org/projects/PRAD-CA), FOXA1, *AR*, and HOXB13 ChIP-seq in primary prostate tumors is available under the following accession code: GSE137527 and EGAS00001003928, TF ChIP-seq data were from public databases of ReMap and ChIP-Atlas. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon reasonable request.

Chapter 3

Recurrent reorganization of the three-dimensional genome pinpoint non-coding drivers of primary prostate tumours

S.Z., J.R.H., and M.L. conceptualized the study. J.R.H. and S.Z. co-led the study with equal contributions and can be interchangeably listed as first author. S.Z. designed and conducted all the experiments with help from C.A. G.G., and K.K. J.R.H. implemented all the computational and statistical approaches and analyses. R.H.-W. pre-processed the RNA-seq data from the primary tumours. Figures were designed by S.Z. and J.R.H. The manuscript was written by S.Z., J.H., and M.L with assistance from all authors. T.v.d.K., M.F., P.C.B., R.G.B., and M.L. supervised the study. M.L. oversaw the study.

3.1 Abstract

Prostate cancer is a heterogeneous disease whose progression is linked to genome instability. However the impact of this instability on the three-dimensional chromatin organization and how this drives progression is unclear. Using primary benign and tumour tissue, we find a high concordance in the higher-order three-dimensional genome organization across normal and prostate cancer cells.

This concordance argues for constraints to the topology of prostate tumour genomes. Nonetheless, we identify changes to focal chromatin interactions and show how SVs can induce these changes to guide *cis*-regulatory element hijacking. Such events result in opposing differential expression on genes found at antipodes of rearrangements. Collectively, our results argue that *cis*-regulatory element hijacking from SV-induced altered focal chromatin interactions overshadows higher-order topological changes in the development of primary prostate cancer.

3.2 Introduction

The human genome is organized into hubs of chromatin interactions within the nucleus, setting its three-dimensional topology [70]. Two classes of higher-order topology, TADs and compartments, define clusters of contacts between DNA elements that are linearly distant from each other, such as CREs and their target gene promoters [71, 72]. Insulating these hubs to prevent ectopic interactions are TAD boundaries, maintained by CCCTC-binding Factor (CTCF) and the cohesin complex [73]. Disruption of TAD boundaries through genetic or epigenetic variants can activate oncogenes, as observed in medulloblastoma [74], acute myeloid leukemia [75], gliomas [76], and salivary gland acinic cell carcinoma [77]. However, recent studies depleting CTCF or the cohesin complex produced little effect on gene expression despite global changes to the three-dimensional chromatin organization [78–80]. In contrast, CRE hijacking caused by genetic alterations can result in large changes to gene expression, despite having little impact on the higher-order chromatin organization [48, 74]. These contrasting observations raise questions about the interplay between components of the genetic architecture, namely, how genetic alterations, chromatin states, and the three-dimensional genome cooperate to misregulate genes in disease. Understanding the roles that chromatin organization and cis-regulatory interactions play in gene regulation is crucial for understanding how their disruption can promote oncogenesis.

The roles of noncoding mutations targeting CREs in cancer are becoming increasingly clear [48, 81, 82]. Mutations to the TERT promoter, for example, lead to its over-expression and telomere elongation in multiple cancer types [45, 83, 84]. Similarly, mutations targeting CREs of the ESR1 and FOXA1 oncogenes in breast and PCas, respectively, lead to their sustained over-expression [9, 32, 85], which is associated with resistance to hormonal therapies [86–89]. Point mutations have the potential to alter three-dimensional chromatin organization, albeit indirectly, by modifying TF or CTCF binding sites [90, 91]. SVs, on the other hand, are large rearrangements of chromatin that can directly impact its structure [92, 93]. This can establish novel CRE interactions from

separate TADs or chromosomes, as has been observed in leukemia [94] and multiple developmental diseases [95, 96]. But how prevalent and to what extent these rearrangements affect the surrounding chromatin remains largely unstudied in primary tumours [82, 93, 97]. Hence, to understand gene misregulation in cancer, it is critical to understand how SVs impact three-dimensional chromatin organization and CRE interactions in primary tumours.

SVs play an important role in PCa, both for oncogenesis and progression. An estimated 97% of primary tumours contain SVs [6, 82], and translocations and duplications of CREs for oncogenes such as *AR* [55], ERG [98], FOXA1 [9, 62] and MYC [9] are highly recurrent. While coding mutations of FOXA1 are found in ~10% of mCRPC patients, SVs that target FOXA1 CREs are found in over 25% of metastatic prostate tumours [9]. In addition to oncogenic activation, SVs in prostate tumours disrupt and inactivate key tumour suppressor genes including PTEN, BRCA2, CDK12, and TP53 [5, 62]. Furthermore, over 90% of prostate tumours contain complex SVs, including chromothripsis and chromoplexy events [99], making it a prime model to study the effects of SVs. However, despite large-scale tumour sequencing efforts, investigating the impact of SVs on three-dimensional prostate genome remains difficult, owing to constraints from chromatin conformation capture (i.e. Hi-C) assays. In this work, we build on recent technological advances in Hi-C protocols to investigate the three-dimensional chromatin organization of the prostate from primary benign and tumour tissues. Using patient-matched WGS, RNA-seq, and ChIP-seq data, we show that SVs in PCa repeatedly hijacking CREs to disrupt the expression of multiple genes with minimal impact to higher-order three-dimensional chromatin organization.

3.3 Results

3.3.1 Three-dimensional chromatin organization is stable over oncogenesis

3C technologies enable the measurement of three-dimensional chromatin organization. These assays, however, are often limited to cell lines, animal models and liquid tumours due to the amount of input required [100]. Here, we optimized and conducted low-input Hi-C [101] on 10 μm thick cryosections from 12 primary prostate tumours and 5 primary benign prostate sections (see Methods, Figure 3.1a, Figure B.1a). The 12 tumours were selected from the CPC-GENE cohort previously assessed for WGS [6], RNA-seq [102], and H3K27ac ChIP-seq [38, 59] (Supplementary Table 1). All 12 of these PCa patients previously underwent radical prostatectomies and 6 of our 12 samples (50%)

harbour the TMPRSS2-ERG genetic fusion (T2E) found in approximately half of the primary PCa patients [6]. The total percent of genome altered ranges from 0.99%-18.78% (Supplementary Table 1) [6]. The 12 tumour samples were histopathologically assessed to have $\geq$ 70% cellularity while the cellularity was $\geq$ 60% for our group of 5 normal prostate samples. Upon Hi-C sequencing, we reached an average of 9.90×10^8 read pairs per sample (range $5.84 \times 10^8 - 1.49 \times 10^9$ read pairs) with minimal duplication rates (range 10.6% - 20.8%) (Supplementary Table 2). Pre-processing resulted in an average of 6.23×10^8 (96.13%) valid read pairs per sample (range $3.95 \times 10^8 - 9.01 \times 10^8$, or 82.42% - 99.22%; Supplementary Table 2). Hence, we produced a high depth, high quality Hi-C library on 17 primary prostate tissue slices.

To characterize the higher-order organization of the primary prostate genome, we first identified TADs. Across the 17 primary tissue samples, we observed an average of 2,305 TADs with a median size of 560 kbp (Supplementary Tables 3-4). However, when considering all hierarchical levels of TAD organization, we did not observe significant differences in the number of TADs identified across length scales (Figure 3.1b), nor in the persistence of their boundaries (Figure 3.1c). This suggests few, if any, differences in three-dimensional chromatin organization at the TAD level between benign and tumour tissue. Notably, we observed differences in organization around essential genes for PCa between primary tissue and previously profiled cell lines. For example, chromatin around the *AR* gene that was previously found enriched in the 22Rv1 compared to RWPE1 prostate cell lines [52] were not recapitulated in either benign or tumour primary samples (Figure 3.1d). Moreover, when compared to other Hi-C datasets, the primary prostate samples clustered separately from cell lines (Figure B.1b), despite similar enrichment of CTCF binding sites near TAD boundaries (Figure B.1c). These results suggest that TADs are constrained over oncogenesis and that cell line models may not harbour disease-relevant three-dimensional chromatin organization.

We next investigated compartmentalization changes, the second class of higher-order three-dimensional chromatin organization. Recurrent changes to segments nearly the size of chromosome arms showed differential compartmentalization in multiple tumour samples compared to benign samples, such as compartment B-to-A transitions on 19q and A-to-B transitions on chromosome Y (Figure B.2a-c). Only two genes on chromosome 19 were differentially expressed between the 8 tumours with benign-like compartmentalization and the other 4 (Figure B.2d). Similarly, no genes on chromosome Y were differentially expressed between the 4 tumours with benign-like compartmentalization and the remaining samples (Figure B.2e). Both arms on chromosome 3 show differential mean compartmentalization, but this appears to be driven by one tumour sample and one benign sample for each arm and is not recurrent (Figure B.2f). Collectively, these results suggest that phe-

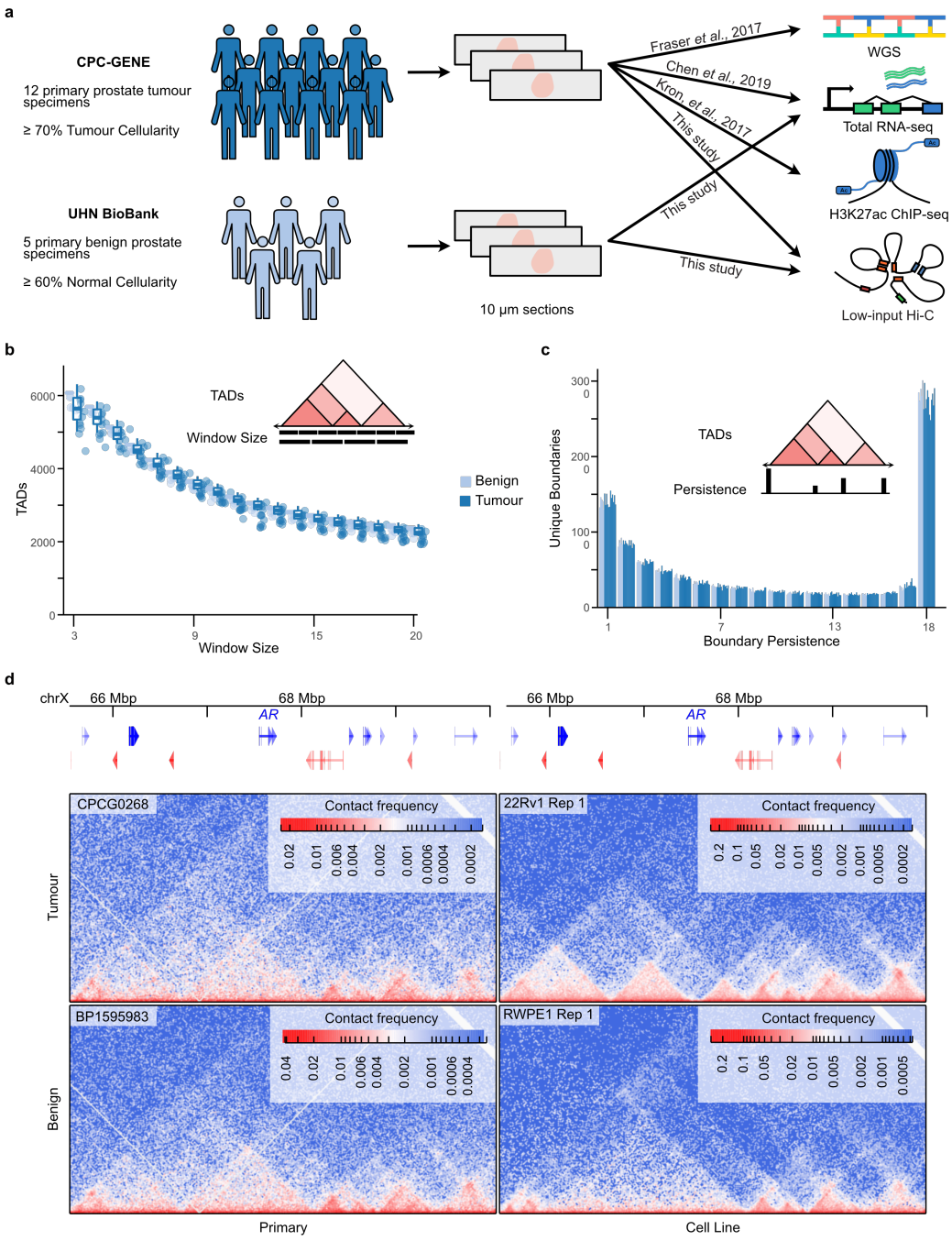

Figure 3.1: **Topologically associated domains are stable over prostate oncogenesis. a.** The sample collection and data usage of primary prostate samples in this study. 10 μm sections from 6 tumours previously identified as T2E+ and 6 T2E- were used for Hi-C sequencing. 5 additional 10 μm sections were collected from benign prostate specimens in the UHN BioBank. **b-c.** A comparison of the number of TADs detected at multiple window sizes (**b**) and boundary persistence (**c**) in each patient sample, with inset schematics. **d.** Contact matrices around the AR gene in primary samples and cell lines. Hi-C data for 22Rv1 and RWPE1 cell lines obtained from Rhie *et al.*, 2019.

notypic differences between benign and tumour tissues do not stem from differences in higher-order three-dimensional chromatin organization alone.

3.3.2 Focal chromatin interactions shift over oncogenesis

Changes to focal chromatin interactions have been observed in the absence of higher-order chromatin changes [103, 104], and we hypothesized that this may be the case in PCa. We detected chromatin interactions, identifying a median of 4,395 interactions per sample (range 1,286 - 6,993; Figure B.3a, Supplementary Table 5). Among these detected interactions, we identified known contacts in PCa such as those between two distal CREs on chromosome 14 and the FOXA1 promoter [85] (Figure B.3b), and CREs upstream of MYC on chromosome 8 that are frequently duplicated in metastatic disease [62] (Supplementary Table 5). 16,474 unique chromatin interactions were identified in at least one sample (Figure 3.2a), reaching an estimated ~80% saturation of detection (Figure B.3c). Restricting our analysis to the 8,486 interactions present in at least two samples (51.5% of all interactions) yielded 1,405 tumour- and 273 benign-specific interactions, suggesting focal changes in three-dimensional chromatin organization occur over oncogenesis. Aggregate peak analysis revealed Hi-C contact enrichment at all detected interactions in all samples (Figure 3.2b-c), demonstrating that tumors- and benign-specific interactions are not binary. Rather, the contacts at "tumour-specific" loci are more enriched than those at "benign-specific" loci in tumour samples (Figure 3.2b). Similarly, the contacts at "benign-specific" loci are more enriched than those at "tumour-specific" loci in benign samples (Figure 3.2c). Together, these results suggest that more focal changes to chromatin interactions are present in prostate oncogenesis despite the stable higher-order organization.

3.3.3 Cataloguing structural variants from Hi-C data

In prostate tumours, SVs populate the genome to aid disease onset and progression [6, 62]. Advances in computational methods now enable the identification of SVs from Hi-C datasets [92, 105]. Applying an SV caller to our primary prostate tumour Hi-C dataset [92], we detected a total of 317 unique breakpoints with a median of 15 unique breakpoints per tumour (range 3-95; Figure 3.3a; Supplementary Table 6). As an example, we found evidence of the TMPRSS2-ERG (T2E) genetic fusion spanning the 21q22.2-3 locus in 6/12 (50%) patients (CPCG0258, CPCG0324, CPCG0331, CPCG0336, CPCG0342, and CPCG0366) (Figure 3.3b), in accordance with previous WGS findings [6]. Combining unique breakpoint pairs into rearrangement events yielded 7.5 total events on average

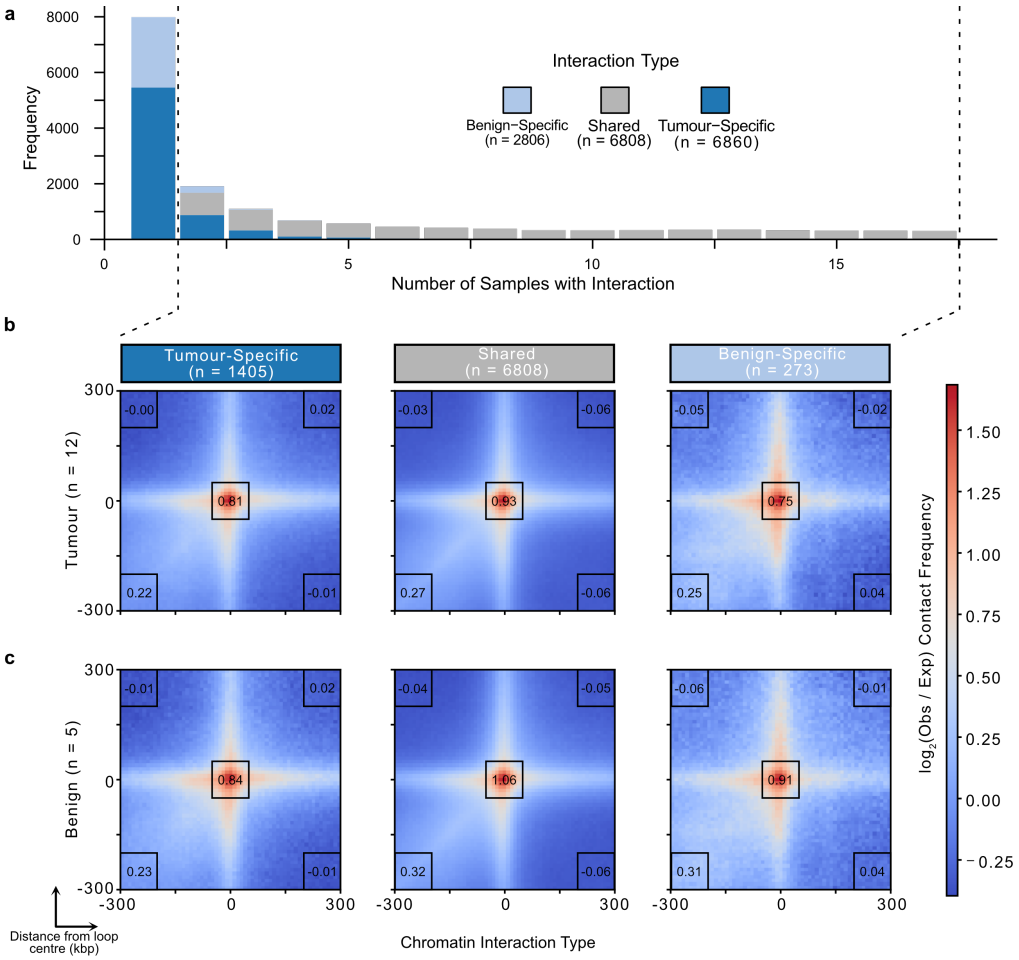

Figure 3.2: **Focal chromatin interactions display subtle differences between benign and tumour tissue. a.** Stacked bar plots of the number of samples that chromatin interactions were identified in. **b-c.** Aggregate peak analysis of tumour (**b**) or benign (**c**) contacts in tumour-specific, benign-specific, and shared interactions identified in two or more samples. Regions plotted are ± 300 kbp around the centre of each identified interaction. Inset numbers are the mean $\log_2$(obs/exp) contact frequencies within the 100 kbp × 100 kbp black boxes.

per patient (range 1 - 36, Figure B.4a-b). We also identified more inter-chromosomal breakpoint pairs with the Hi-C data in 11 of 12 tumours (Figure 3.3b), including a novel translocation event that encompasses the deleted region between TMPRSS2 and ERG into chromosome 14. Few loci contained SV breakpoints recurrent between patients (Figure B.4c). These numbers are smaller than previously reported from matched WGS data [6]; however, the median distance between breakpoints on the same chromosome was much larger at 31.6 Mbp for Hi-C-identified breakpoints, compared to 1.47 Mbp from WGS-identified breakpoints (Figure 3.3c). This is consistent with the inherent nature and resolution of the Hi-C method to detect larger, inter-chromosomal events [92]. No SVs were detected in the 5 primary benign prostate tissue samples from Hi-C data. While this does not rule out the presence of small rearrangements undetectable by Hi-C limited by its resolution, the absence of large and inter-chromosomal SVs further supports a difference in genome stability between benign and tumour tissues [6, 59, 99, 106]. Collectively, Hi-C defines a valid method to interrogate for the presence of SV in tumour samples, compatible with the detection of intra- and inter-chromosomal interactions otherwise missed in WGS analyses.

Among SVs detected in primary prostate tumours, we identified both simple and complex chains of breakpoints. While simple SVs correspond to fusion between two distal DNA sequences, complex chains are evidence of chromothripsis and chromoplexy [99]. These genomic aberrations affecting multiple regions of the genome are known to occur in both primary and metastatic PCa [6, 82, 99]. The chains can be pictured as paths connecting breakpoints in the contact matrix (Figure B.4d). 8 of the 12 (66.7%) tumour samples contained these chains, including one patient (CPCG0331) harbouring 11 complex events and three patients (CPCG0246, CPCG0345, and CPCG0365) each harbouring a single complex event. We observed a median of 1 complex event per patient (range 0-11) consisting of a median of 3 breakpoints (range 3-7) spanning a median of 2 chromosomes per event (range 1-4, Supplementary Table 7, Figure B.4b). Patient CPCG0331 had 11 complex events, including a 6-breakpoint event spanning 3 chromosomes (Figure B.4b). A highly rearranged chromosome 3 was also found in the same patient (Figure 3.3d). The most common type of complex event involved 3 breakpoints and spanned 2 chromosomes, occurring 9 times across 5 of the 8 patients with complex events. In summary, using Hi-C, we detected both simple and complex SVs in primary prostate tumours not previously identified using WGS-based methods. We were able to identify known observations, such as a highly mutated region on chromosome 3 and subtype-specific differences in abundance, as well as find novel inter-chromosomal events not previously reported.

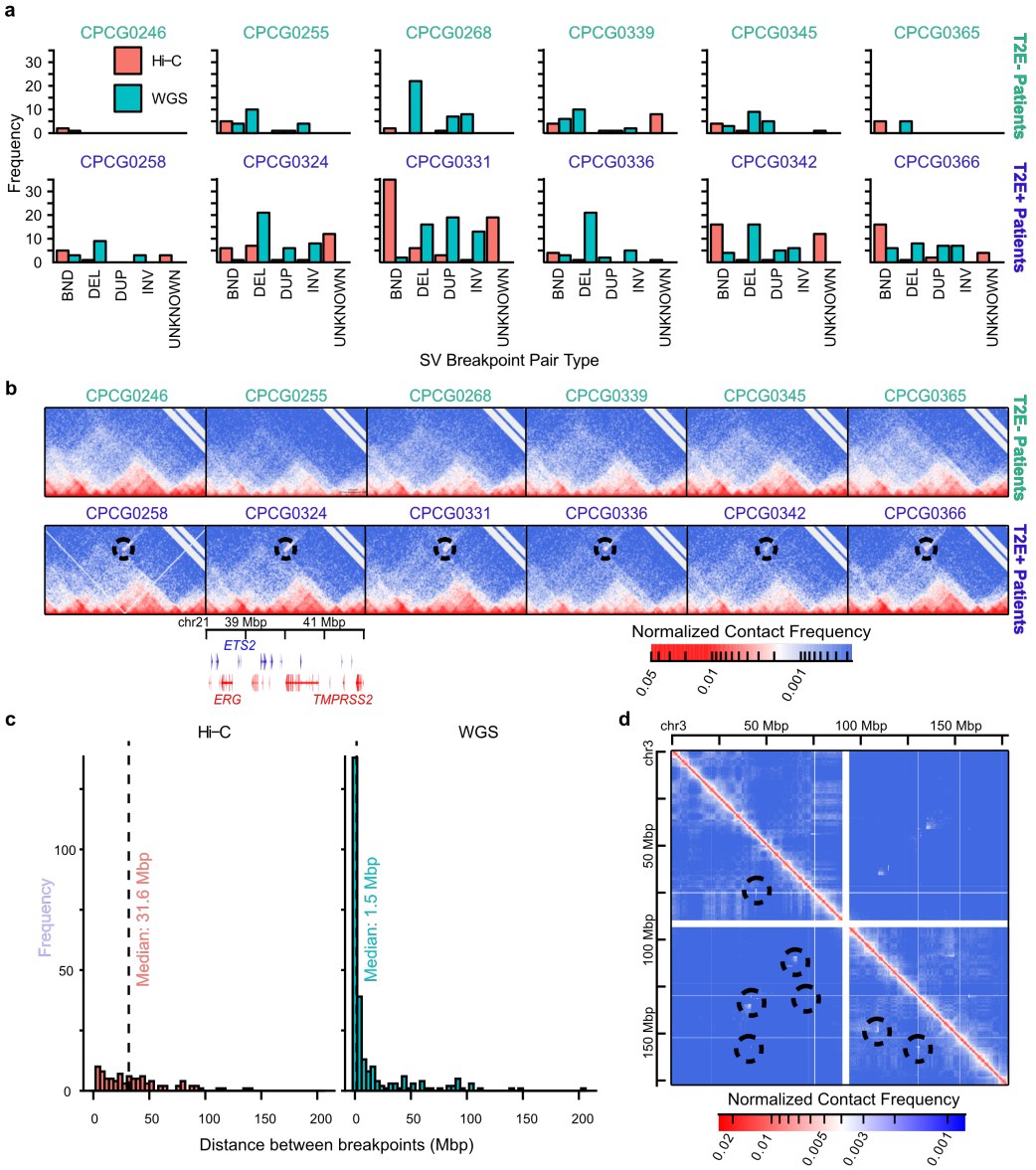

Figure 3.3: **SVs are identified in primary tissue through chromatin conformation capture. a.** Bar plot of SV breakpoint pairs identified by Hi-C and WGS on matched samples. BND = inter-chromosomal translocation, DEL = deletion, DUP = duplication, INV = inversion, UNKNOWN = breakpoint pair of unknown type. **b.** Hi-C contact matrices of the chr21:37-42 Mbp locus harbouring the *TMPRSS2* and *ERG* genes. Circles indicate increased contact between *TMPRSS2* and *ERG* in the T2E+ tumours. **c.** Histogram showing the distance between breakpoints on the same chromosome detected by Hi-C (left) versus WGS (right). **d.** An example of a complex set of rearrangements spanning both arms of chromosome 3 in a single patient.

3.3.4 SVs alter gene expression independently of intra-TAD contacts

Using combined WGS called SVs with those from Hi-C data, we next systematically examined the impact of SVs on TAD structure. This led us to look at the intra-TAD and inter-TAD interactions around each breakpoint. We observed that only 18 of the 260 (6.9%) TADs containing SV breakpoints were associated with decreased intra-TAD or increased inter-TAD interactions (Figure 3.4a). 12 of 18 (66.7%) occurrences were within T2E+ tumours. We found no evidence that simple versus complex SVs were a factor in determining whether a TAD was altered (Pearson's χ^2 test, $\chi^2 = 0.0166$, $p = 0.8974$, $df = 1$). Similarly, the type of SV (deletion, inversion, duplication, or translocation) was not predictive of whether the TAD would be altered (Pearson's χ^2 test, $\chi^2 = 4.7756$, $p = 0.3111$, $df = 4$). Overall, we find that SVs are associated with higher-order topological changes in a small percentage of cases and that the presence of an SV breakpoint is not predictive alone of an altered TAD.

Despite the evidence that SVs rarely impact higher-order chromatin topology, we evaluated whether SVs affected the expression of genes within the TADs surrounding the breakpoint using patient-matched RNA-seq data [102]. We found that 23 of 260 breakpoints (8.8%) are associated with significant changes to local gene expression (Figure 3.4b). Complex events can have opposite effects at each breakpoint. For example, while the T2E fusion across all tumours leads to over-expression of ERG and under-expression of TMPRSS2 [6, 38], the deleted locus between these two genes was inserted into chromosome 14 as part of a complex translocation event in one patient (Figure 3.4c-f). This inserted fragment positions ERG towards the 5' end of the RALGAPA1 gene and TMPRSS2 towards the 3' end (Figure 3.4c) resulting in a significant drop in intra-TAD contacts at the RALGAPA1 locus on chromosome 14 (two-sample unpaired t-test, $t = 6.38$, $p = 1.04 \times 10^{-9}$; Figure 3.4d). Despite the significant topological change on chromosome 14, no significant changes to expression was detectable across genes within the same TAD on chromosome 14 (Figure 3.4e). Conversely, TAD alterations are not required changes to gene expression. As part of a complex SV involving the RIMBP2 gene (Figure 3.4g-j), both ends of the gene contain breakpoints (Figure 3.4g). This rearrangement is not associated with changes to intra-TAD contacts (two-sample unpaired t-test, $t = 0.8101$, $p = 0.4183$; Figure 3.4h). However, RIMBP2 is over-expressed in this patient (Figure 3.4i). More generally, only a single breakpoint was observed with both TAD contact and gene expression changes, although we did not find evidence to suggest these are dependent events (Pearson's χ^2 test, $\chi^2 = 6.31 \times 10^3$, $p = 0.9367$, $df = 1$). For TADs where at least one gene was differentially expressed, 19 (83%) of them had at least one gene with doubled or halved expression.

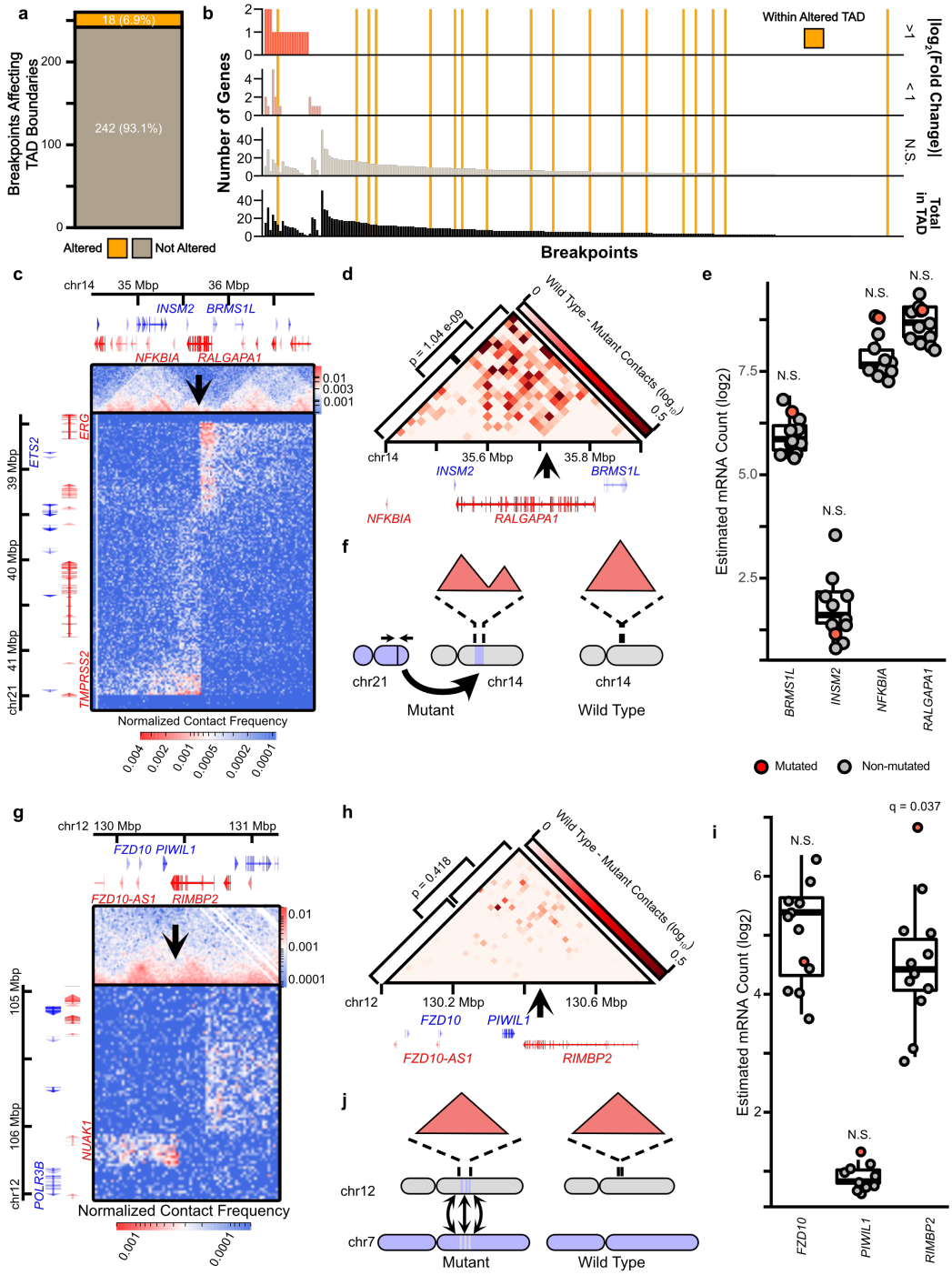

Figure 3.4: **SVs can alter TADs or gene expression around breakpoints, but rarely alters both.** (Continued on the following page)

Figure 3.4: **a.** A count of the number of SV breakpoints associated with altered TAD boundaries. **b.** Bar plot showing the number of genes differentially expressed around SV breakpoints. **c-f.** An example of an SV that alters intra-TAD contacts without significantly affecting gene expression of the nearby genes. **c.** The contact matrix showing a translocation of the *TMPRSS2-ERG* locus into chromosome 14 in the *RALGAPA1* gene. **d.** The differential contact matrix between the tumour containing this translocation and another tumour without it. **e.** Gene expression scatterplot and boxplot of genes within the affected TAD for each sample. **f.** A schematic representation of the translocation. **g-j.** An example of an SV that does not alter intra-TAD contacts but does alter the expression of the nearby genes. **g.** The contact matrix showing a complex rearrangement around the *RIMBP2* gene. **h.** The differential contact matrix between the tumour containing this rearrangement and another tumour without it. **e.** Gene expression scatterplot and boxplot of genes within the affected TAD for each sample. **j.** A schematic representation of the rearrangement. Boxplots highlight the first, second, and third quartiles of expression in the tumours without the example SV. Red dots represent the tumour with the example SV, grey dots represent the tumours without.

Notably, we found that inter-chromosomal translocations are associated with altering the expression of genes nearby their breakpoints compared to intra-chromosomal breakpoints (Pearson's χ^2 test, $\chi^2 = 7.0088$, $p = 0.00811$, $df = 1$; Figure B.5). Taken together, these results suggest that while SVs can alter contacts within TADs, this is neither necessary nor sufficient to alter gene expression.

3.3.5 SVs alter focal chromatin interactions to hijack CREs and alter antipode gene expression

Mutations in PCa have previously been found to converge on active CREs [59]. To assess if SVs function in a similar fashion, we investigated the convergence of SV breakpoints in active CREs. We find that SV breakpoints are enriched in the catalogue of CREs captured by H3K27ac ChIP-seq from our 12 primary prostate tumours compared to the rest of the genome (one-sided permutation z-test, $z = 25.591$, $p = 0.0099$, $n = 100$; Figure 3.5a-b). This is similar to the enrichment of point mutations in CREs active in PCa [59], suggesting that SVs which alter gene expression may do so by recurrently targeting CREs. Since individual CREs can regulate multiple genes [107], we suspected that SVs that do alter gene expression may predominantly affect multiple genes at the same time, instead of single genes. In agreement, when considering all SVs associated with altered gene expression near a breakpoint we find 16 of the 22 (72.7%) SVs are associated with altered expression of multiple genes (Figure 3.5c-d). Notably, 15 of these 16 SVs (93.8%) are associated with both over- and under-expression of genes, instead of genes all being either over-expressed or under-expressed (Figure 3.5d). 12 of these 15 (80%) SVs are associated with expression changes at SV antipodes, opposite ends of a breakpoint pair (Figure B.6). The recurrent targeting of active CREs, combined with the opposite gene expression changes at SV antipodes, suggests that SVs may

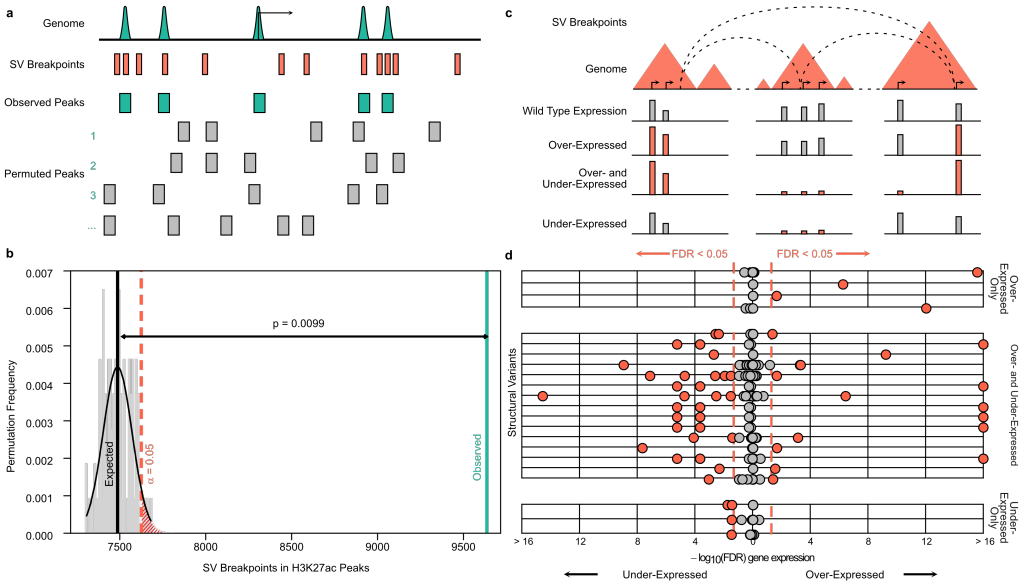

Figure 3.5: **SV breakpoints are enriched in active CREs and repeatedly alter the expression of multiple genes. a.** Schematic of permutation testing for the overlap between SV breakpoints in all CPC-GENE prostate tumours and the catalogue of active CREs in the 12 tumour samples in this study. **b.** Histogram of permutation test results in grey. The vertical black and green bars refer to the expected and observed overlap of SV breakpoints and CREs, respectively. P-value is obtained from the permutation test, $n = 100$. **c.** Schematic of how the expression of genes within TADs containing SV breakpoints are compared between mutant and WT tumours are compared. **d.** Scatterplot of FDR values obtained from differential gene expression analysis as outlined in **c.** Red dots are differentially expressed genes (FDR ¡ 0.05), grey dots are genes not differentially expressed between the mutant and WT tumours.

repeatedly alter expression by CRE hijacking.

The fusion of *PMEPA1* and *ZNF516* is an example of CRE hijacking resulting in opposite differential gene expression. Specifically, the fusion results in the *PMEPA1* promoter being hijacked to the 5' end of the *ZNF516* gene. This is concomitant with the over-expression of *ZNF516* and under-expression of *PMEPA1* (Figure 3.6a-c). In addition to hijacking the *PMEPA1* promoter to the *ZNF516* gene, this fusion also coincides with gains in H3K27ac over the *ZNF516* gene body and of H3K27ac histone hypo-acetylation over the 3' end of *PMEPA1*'s gene body. This mirrors the creation of a Cluster Of Regulatory Elements (COREs) reported for the T2E fusion, reflective of new CREs enabling ERG over-expression and the concomitant under-expression of TMPRSS2 (Figure B.7) [38, 108, 109]. CRE hijacking is also observed with inter-chromosomal rearrangements such as seen at the SV connecting chromosomes 7 and 19, creating 2 fusion products (termed C2B and B2C; Figure 3.6d). This SV separates the 3' end of BRAF from its promoter and upstream enhancers on chromosome 7 (C2B; Figure 3.6d), fusing it to the 3' end of CYP4F11 (Figure 3.6e). Focal chromatin interactions between BRAF and multiple active CREs are only observed in the

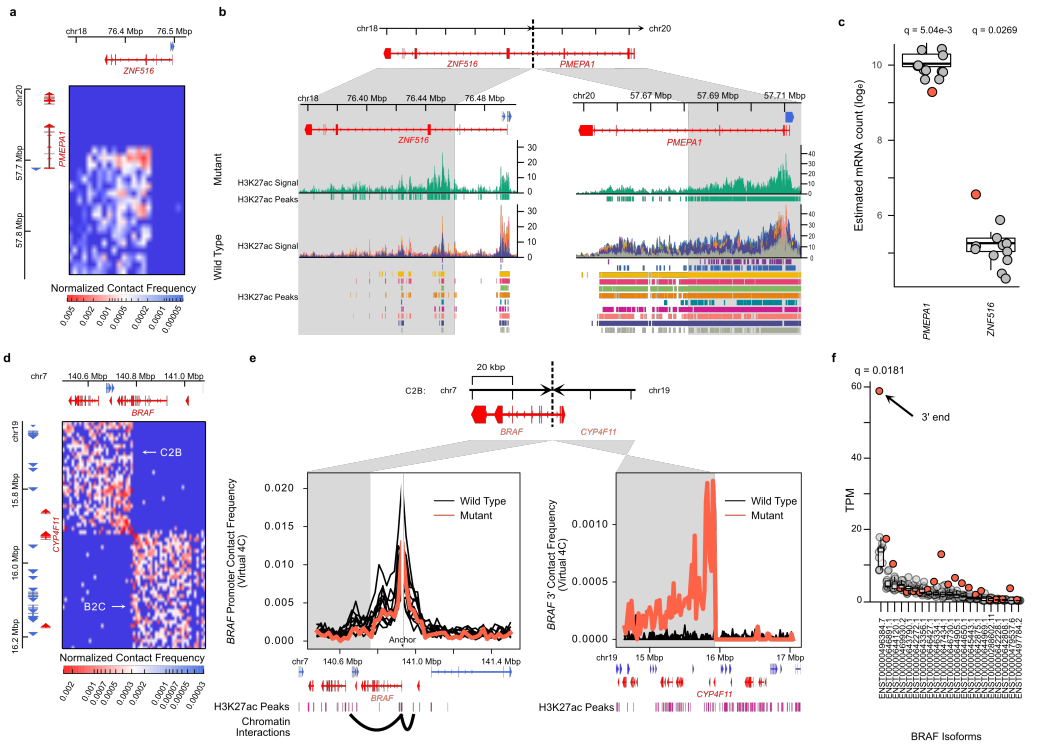

Figure 3.6: **SVs altering gene expression by rewiring focal chromatin interactions. a.** The contact matrix of the deletion between *PMEPA1* and *ZNF516*. **b.** Genome tracks of H3K27ac ChIP-seq signal around the *ZNF516* and *PMEPA1* genes with the rearrangement. Grey regions are loci brought into contact by the SV. **c.** Gene expression of *PMEPA1* and *ZNF516* in all tumour samples. Boxplots represent the first, second, and third quartiles of WT patients (grey dots). Red dots are the gene expression for the mutated patient. **d.** The contact matrix of an inter-chromosomal break between chromosome 7 and chromosome 19. **e.** Contact frequencies of the *BRAF* promoter on chromosome 7 (left) and the 3' end of *BRAF* on chromosome 19 (right). SV-associated contacts between the 3' end of *BRAF* on chromosome 19 (right) are focally enriched at H3K27ac peaks downstream of *CYF4P11*. Bar plot of SVs categorized by how differentially expressed genes altered.

fusion on chromosome 19 (Figure 3.6e). Using matched RNA-seq data, we observe an estimated 5 fold increase in expression for the 3' exons of BRAF in the mutated tumour compared to others (fold-change = 4.976, FDR = 0.0181; Figure 3.6f). Collectively, over-expression of the oncogenes, such as ERG and BRAF, and suppression of the tumour suppressor *PMEPA1* demonstrates the disease-relevant effects of CRE hijacking mediated by SVs in primary PCa resulting in changes to focal chromatin interactions, and that these effects overshadow the effect on higher-order topology in primary PCa.

3.3.6 Discussion

Genetic alterations that subvert the higher-order chromatin organization to allow for aberrant focal interactions may be more common in cancer than previously recognized. In this work we

demonstrated that CRE hijacking by SVs is often associated with opposing gene expression changes at SV antipodes, whereby genes on one flank of the breakpoint are up-regulated while genes on the other flank are repressed. Complex SVs, such as chromoplexy and chromothripsis, are found in numerous cancer types [82, 99], providing many opportunities for widespread effects on gene expression and CRE hijacking. This is in addition to many known cancer drivers that alter CRE interactions, including the *AR* and FOXA1 enhancer amplifications in primary and metastatic prostate tumours [9, 38, 55, 56, 62, 85]. More recent findings also fit this model, such as accumulation of extrachromosomal circular DNA activating oncogenes that would otherwise be constrained by chromatin topology [110–113]. These insights stress the importance of investigating all ends of an SV to assess the biological impact of these mutations on the cis-regulatory landscape as a whole, as opposed to focusing on CREs or SV breakpoints as single entities.

Changes to the three-dimensional genome reported in disease onset or development are often inferred from alterations in TAD boundaries [78, 93]. For instance, CTCF activity is targeted by somatic mutations that enrich at its binding sites in colorectal, esophageal, and liver cancers [91, 114]. Furthermore, gains in DNA methylation at CTCF binding sites are linked to altered TAD structures in gliomas [76]. In primary PCa 97% of differentially methylated regions genome-wide in primary PCa are losses of DNA methylation [115, 116], an epigenetic process previously shown to have limited impact on CTCF chromatin binding [117]. This suggests that aberrant CTCF binding at TAD boundaries is not a hallmark of prostate oncogenesis. Our observation of stable chromatin organization supports this model. Notably, stable TAD structures observed in these primary tissues contrast previous reports of chromatin organization in cell lines derived from prostate cells [52, 118], highlighting the necessity of low-input protocols and primary tissues [101]. Our findings further support recent reports of shared higher-order chromatin organization among phenotypically distinct cell types in model organisms [71, 93, 97, 119–121]. Taken together, this body of evidence suggests that large disruptions to TADs and compartments may constrain the transformation of normal to cancer cells or the divergent subtyping within prostate tumours. Instead, changes to focal chromatin interactions seem to reflect alterations in the genetic architecture leading to cancer development. Investigating these focal chromatin interactions may provide insights on the relationship between CREs, such as between enhancers and their target gene promoter [122, 123] to better understand the etiology of disease.

In conclusion, by bypassing technical limitations to characterize the three-dimensional genome organization across benign and tumour prostate tissue, our work reveals the predominant stable nature of genome topology across prostate oncogenesis. Instead, alterations to discrete chromatin

interactions populate the PCa genome. These impact the function of CREs, such as we report for SV-mediated CRE hijacking events. Considering the contribution of SVs across human cancers [124], our collective work presents a framework inclusive of genetics, chromatin state, and three-dimensional genome organization to understand the genetic architecture across individual primary tumours.

3.4 Methods

3.4.1 Patient selection criteria

Patients were selected from the CPC-GENE cohort of Canadian men with indolent PCa, Gleason scores of 3+3, 3+4, and 4+3. All primary human material was obtained with informed consent with approval of our institutional research ethics board (UHN 11-0024). The intersection of previously published data for WGS [6], RNA abundance [102], and H3K27ac ChIP-seq [38] led to 25 samples having data for all assays. 11 of these tested positive for ETS gene family fusions (T2E status), and 14 without. To accurately represent the presence of this subtype of PCa in the disease generally, and to ensure minimum read depths required to perform accurate analysis on chromatin conformation data, we selected approximately half of these remaining samples (6 T2E+ and 6 T2E-).

3.4.2 Patient Tumour in situ low-input Hi-C Sequencing

We followed the general in situ low input Hi-C (Low-C) protocol from Diaz *et al.*, 2018 [101] with our own re-optimization for solid tumour tissue sections. It is worth noting that throughout the protocol, the pellet would be hardly visible and would require careful pipetting. The specific modifications of the protocol are described below.

Tumour Tissue Preparation

Twelve cryopreserved-frozen PCa tumour tissue specimens were obtained from primary PCa patients as part of the CPC-GENE effort [6]. Informed consent was obtained from all patients with REB approval (UHN 11-0024). These tumour specimens were sectioned into 10 μm sections. Sections before and after the sections used for Hi-C were stained with hematoxylin and eosin (H&E) and assessed pathologically for $\geq$ 70% PCa cellularity. The percentage of infiltrating lymphocytes was also estimated by pathological assessment to be $\leq$ 3%. Stratification into T2E+ or T2E- was determined through either WGS detection of the rearrangement, immunohistochemistry, or mRNA

expression microarray data [6].

Normal Tissue Preparation

Five snap-frozen prostate tumour-adjacent normal tissue specimens were obtained. Informed consent was obtained from all patients with REB approval (UHN 11-0024). Tissue specimens were sectioned into 5, 10, and 20 μm sections. Sections used for Hi-C and RNA-seq were stained with H&E and assessed pathologically for $\geq$ 60% prostate glandular cellularity.

Fixation and Lysis

One or two sections (consecutive; depending on surface area) for each patient were thawed and fixed by adding 300 μL of 1% formaldehyde in PBS directly onto the tissue sample, followed by a 10-minute incubation at room temperature (RT) (Supplementary Figure 1a). The formaldehyde was quenched by adding 20 μL of 2.5M glycine to the sample reaching a final concentration of 0.2M followed by 5 minutes of incubation at RT. The samples were then washed three times with 500 μL cold PBS and scraped off the microscope slide with a scalpel into 1.5 mL centrifuge tube containing 250 μL of ice-cold Low-C lysis buffer (10 mM Tris-Cl pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630 (Sigma-Aldrich)) supplemented with protease inhibitor. The samples were then mixed thoroughly by gentle pipetting and left on ice for 20 minutes with intermittent mixing. Upon lysis, the samples were snap-frozen with liquid nitrogen and stored at -80 ˚C until processing the next day. As a note, stagger fixation times when processing multiple samples to prevent needless rush and chance of under/over-fixation.

Enzyme Digestion and Overhang Fill-In

The samples stored at -80 ˚C were thawed on ice and spun down at 300 $\times$ g for 5 minutes at 4 ˚C. The samples were then re-suspended in 125 μL of ice-cold 10X NEB2 Buffer (New England Biolabs), and again spun down at 13,000 $\times$ g for 5 minutes at 4 ˚C. The pellet was then re-suspended in 25 μL of 0.4% SDS and incubated at 65 ˚C for 10 minutes without agitation for permeabilization. To quench the SDS, 10% Triton X-100 in water (12.5 μL + 75 μL water) was then added to the samples and incubated at 37 ˚C for 45 minutes at 650 rpm. For enzymatic digestion, 35 μL of 10X NEB2.1 buffer (New England Biolabs) was added to each sample, follow by the addition of 50 U of MboI and 90 minutes incubation at 37 ˚C with gentle agitation (add 30 U first, incubate 45 minutes, followed by the addition of another 20 U and another 45 minutes of incubation). Upon digestion, the MboI enzyme was inactivated by incubating at 62 ˚C for 20 minutes. The overhangs generated

by the MboI enzyme was then filled-in by adding a mix of dNTPs and DNA Polymerase I Klenow Fragment directly to each sample (10 μL of 0.4 mM biotin-14-dCTP, 0.5 μL of 10 mM dATP, 0.5 μL of 10 mM dGTP, 0.5 μL of 10 mM dTTP, 4 μL of 5U/μL DNA Polymerase I Klenow Fragment). The samples were then mixed by gentle pipetting followed by incubation at 37 ˚C for 90 minutes with gentle agitation.

Proximity Ligation and De-crosslinking

Upon overhang fill-in, each sample was subject to proximity ligation through the addition of 328.5 μL water, 60 μL of 10X T4 DNA Ligase Buffer (ThermoFisher Scientific), 50 μL of 10% Triton X-100, 6 μL of 20 mg/mL BSA (New England Biolabs) and 3.5 μL of 5 Weiss U/μL T4 DNA Ligase (ThermoFisher). The samples were mixed through gentle pipetting and incubated at RT (20-22 ˚C) with rotation for 4 hours. The samples were then spun down at 13,000 × g for 5 minutes at RT and re-suspended in 250 μL of Extraction Buffer (50 mM Tris-Cl pH 8.0, 50 mM NaCl, 1 mM EDTA, 1% SDS) upon removal of supernatant. Next, 10 μL of 20 mg/mL Proteinase K (New England Biolabs) was added to each sample and incubated at 55 ˚C for 30 minutes at 1,000 rpm. Then 65 μL of 5 M NaCl was added to each sample and incubated at 65 ˚C at 1,000 rpm overnight.

DNA Extraction

Phenol-chloroform extraction columns were spun down at 17,000 × g for 1 minute at 4 ˚C to get gel down to the bottom of the tube. The samples incubated overnight were then added to the column. Next, an equal volume (~325 μL) of phenol-chloroform-isoamyl alcohol mixture (25:24:1) (Sigma) was also added to the column. The column was then inverted for thorough mixing and spun down at 17,000 × g for 5 minutes at 4 ˚C. The surface layer on top of the gel upon spinning contains the sample and is transferred to a clean 1.5 mL tube (~325 μL). Each sample was mixed with 31.5 μL of 3M sodium acetate, 2 μL of GlycoBlue (ThermoFisher Scientific), and 504 μL of 100% ethanol for DNA precipitation. The samples were inverted several times for mixing and incubated at -80 ˚C for 20 minutes, followed by a centrifuge spin at 17,000 × g for 45 minutes at 4 ˚C. The supernatant was carefully discarded and the pellet was washed with 800 μL of ice-cold 70% ethanol followed by a centrifuge spin at 17,000 × g for 5 minutes at 4 ˚C. The supernatant was then discarded and the tube was air-dried until no traces of ethanol was left prior to dissolving the DNA pellet with 30 μL of Elution Buffer (Qiagen PCR Clean-Up Kit). 1 μL of RNase A (ThermoFisher Scientific) was added to each sample followed by incubation at 37 ˚C for 15 minutes. A mix of 5 μL of 10X NEB2.1 buffer (New England Biolabs), 1.25 μL of 1 mM dATP, 1.25 μL of 1 mM dCTP, 1.25 μL of 1 mM dGTP,

1 mM of dTTP, 0.5 μL of 10 mg/mL BSA, 5 μL of water, 3.5 μL of 3 U/μL T4 DNA Polymerase (New England Biolabs) was added to each sample. The samples were mixed thoroughly by gentle pipetting, and then incubated at 20 °C for 4 hours.

Fragmentation and Biotin Pull-down

70 μL of water was added to each sample bringing total volume up to 120 μL, and the samples were transferred into Covaris sonication tubes. The samples were then sonicated using Covaris M220 sonicator to attain 300-700 bp fragments. For biotin pull-down using a magnetic rack, 30 μL of Dynabeads MyOne Streptavidin C1 beads (Life Technologies) for each sample was washed once with 400 μL of 1X B&W buffer + 0.1% Triton X-100. The beads were then re-suspended in 120 μL of 2X B&W buffer and transferred to the 120 μL of sample (1:1 ratio). The sample was then incubated with gentle rotation at RT for 20 minutes. The supernatant was discarded and the beads were re-suspended with 400 μL of 1X B&W buffer + 0.1% Triton X-100 followed by a 2-minute incubation at 55 °C with mixing. The wash was repeated once more, then re-suspended in 400 μL of 1X NEB2 buffer (New England Biolabs).

Library Preparation and Size Selection

The beads containing the Hi-C samples were separated on a magnetic rack to remove the supernatant. The beads were then re-suspended in a total volume of 10 μL for library preparation using the SMARTer ThruPLEX DNA-seq library preparation kit (Takara Biosciences) per manufacturer's protocol with an adjustment on the last step, a PCR reaction for library amplification. Upon reaching that step, the reaction was carried out on a regular PCR for two cycles to amplify the Hi-C samples off the streptavidin beads. Next, the samples were transferred onto a new tube where 20X SYBR was added. The samples were then subject to real-time qPCR and pulled out from the qPCR machine mid-exponential phase. Ultimately, this is done to reduce PCR duplication rates, a huge limitation for low-input Hi-C protocols. The Hi-C libraries were then double size-selected for 300-700 bp using Ampure XP beads and sent for BioAnalyzer analysis prior to sequencing.

3.4.3 Hi-C Sequencing and Data Pre-processing

Sequencing

The Hi-C libraries for each tumour sample were sent for shallow paired-end 150 bp sequencing (~10-15 million reads per sample) on a NextSeq 500. Upon confirming library quality and low

duplication rates ($< 2\%$), samples were sent for deep paired-end 150 bp sequencing with the aim of 800 million raw read pairs per sample on NovaSeq 6000.

Sequence alignment and Hi-C artefact removal

Paired-end FASTQ files were pre-processed with HiCUP (v0.7.2) [125]. Reads were truncated at MboI ligation junction sites prior to alignment with `hicup_digester`. Each mate was independently aligned to the hg38 genome and were then paired and assigned to MboI restriction sites by `hicup_map`. `hicup_map` uses Bowtie2 (v2.3.4) [126] as the underlying aligner which has the following parameters: `--very-sensitive --no-unal --reorder`. Reads that reflect technical artefacts were filtered out with `hicup_filter`. Duplicate reads were removed with `hicup_deduplicator`.

Reads that came from different sequencing batches were then aggregated for each tumour sample at this stage using `sambamba merge` (v0.6.9) [127]. This resulted in an average of 1.12×10^9 read per tumour sample (Supplementary Table 2).

Contact matrix generation and balancing

Aggregated binary alignment map (BAM) files were converted to the pairs format using pairtools (v0.2.2) [128] and then the cooler format using the cooler package (v0.8.5) [129]. The pairs files were generated with the following command:

```
pairtools parse -c {genome} --assembly hg38 -o {output_pairs} {
    input_bam}
```

The cooler files were generated at an initial matrix resolution of 1000 bp with the following command:

```
cooler cload pairs --assembly hg38 -c1 2 -p1 3 -c2 4 -p2 5 {genome
    }:1000  {input_pairs} {output_cooler}
```

The raw contact matrices stored in the cooler file format were balanced using cooler's implementation of the ICE algorithm [130] using the `cooler balance` command. Contact matrices at different resolutions were created with the `cooler zoomify` command.

3.4.4 Hi-C Data Analysis

TAD identification

Contact matrices were binned at a resolution of 40 kbp. To remove sequencing depth as a confounding factor, contact matrices for all samples were first downsampled to match the sequencing depth of the shallowest sample. For comparisons including cell lines, this was 120×10^6 contacts. For comparisons only involving primary samples, this was 300×10^6 contacts. This was achieved with Cooltools (v0.3.2) [131] with the following command:

```
cooltools random-sample -c 120000000 {input}::/resolutions/40000 {
    output}
```

TADs were identified using TopDom [132] on the downsampled, ICE-normalized contact matrices. To identify domains at multiple length scales, similar in concept to Artamus' gamma parameter [133], TopDom was run multiple times per sample, with the window size parameter set at values between 3 and 40, inclusive (corresponding to 120 kbp and 1.6 Mbp). The lower bound for the window size parameter allowed for the identification of domains multiple megabases in size at the upper end and domains < 100 kbp at the lower end without being dominated by false calls due to sparsity of the data. Despite TopDom being more resistant to confounding by sequencing depth than other TAD calling tools [134], biases in boundary persistence were evident between samples of different sequencing depth. Downsampling contact matrices to similar depths resolved these biases.

Given the stochasticity of Hi-C sequencing, boundaries called at one window size may not correspond to the exact same location at a different window size. To attempt to resolve these different boundary calls and leverage power from multiple window sizes, boundaries for a given patient were considered at all window sizes. Boundaries within one bin (40 kbp) of each other and called at different window sizes were marked as conflicting calls. If only two boundaries were in conflict and all the window sizes where the first boundary was called are smaller than the window sizes where the second boundary was called, the second boundary was selected since larger smoothing windows are less sensitive to small differences in contact counts. If only two boundaries were in conflict but there is no proper ordering of the window sizes, the boundary that was identified most often between the two was selected. If three boundaries are in conflict, the middle boundary was selected. If four or more boundaries were in conflict, the boundary that was identified most often was selected.

To determine the maximum window size for TAD calls, TAD calls were compared across window sizes for the same patient using the BPscore metric [135]. TAD calls are identical when the BPscore

is 0, and divergent when 1. The cut-off window size for a single patient was determined when the difference between TAD calls at consecutive window sizes was < 0.005, twice in a row. The maximum window size was determined by the maximum window size cut-off across all samples in a comparison. For comparisons involving only primary samples, the maximum window size was determined to be $w = 20 \times 40$ kbp. For comparisons involving cell lines, this was $w = 32 \times 40$ kbp.

The persistence of a TAD boundary was calculated as the number of window sizes where this region was identified as a boundary.

Sample clustering by TADs

Using the TAD calls at the window size $w = 32 \times 40$ kbp, the similarity between samples was calculated with BPscore. The resulting matrix, containing the similarity between any two samples, was used as the distance matrix for unsupervised hierarchical clustering with `ward.D2` linkage.

Compartment identification

Contact matrices were binned at a resolution of 40 kbp, similarly to TAD identification. To remove sequencing depth as a confounding factor, contact matrices for all samples were first down-sampled to match the sequencing depth of the shallowest sample. Contact matrix eigenvectors were calculated with Cooltools. To standardize the sign of each eigenvector, the GC content of the reference genome, binned at 40 kbp, was used as a phasing reference track. This reference track was calculated with the `frac_gc` function from the Bioframe Python package (v0.0.12) [136]. The first eigenvector was used to identify compartments with the following command:

```
cooltools call-compartments --bigwig --reference-track gc-content-
    phase.bedGraph -o {output} {input}
```

Identification of significant chromatin interactions

Chromatin interactions were identified in all 17 primary samples with Mustache (v1.0.2) [137]. Using the Cooler files from above, Mustache was run on the ICE-normalized 10 kbp contact matrix for each chromosome with the following command:

```
mustache -f {input} -r 10000 -ch {chromosome} -p 8 -o {output}
```

Interaction calls on each chromosome were merged for each sample to create a single table of interaction calls across the entire genome.

To account for variances in detection across samples and to identify similarly called interactions across samples, interaction anchors were aggregated across all samples to form a consensus set. Interaction anchors were merged if they overlapped by at least 1 bp. Interaction anchors for each sample were then mapped to the consensus set of anchors, and these new anchors were used in all subsequent analyses.

Chromatin interaction saturation analysis

To estimate the detection of all chromatin interactions across all samples, a nonlinear regression on an asymptotic model was performed. This is similar in method to peak saturation analysis used to assess peaks detected in ChIP-seq experiments from a collection of samples [38]. Bootstrapping the number of unique interactions detected in a random selection of n samples was calculated for n ranging from 1 to 17. 100 iterations of the bootstrapping process were performed. An exponential model was fit against the mean number of unique interactions detected in n samples using the `nls` and `SSaymp` functions from the stats R package (v3.6.3). The model was fit to the following equation:

$$\mu = \alpha + (R_0 - \alpha) \exp(kn)$$

where μ is the mean number of chromatin interactions for a given number of samples, n, α is the asymptotic limit of the total number of mean detected interactions, R_0 is the response for $n = 0$, and k is the rate constant. The estimated fit was used to predict the number of samples required to reach 50%, 90%, 95%, and 99% saturation of the asymptote (Supplementary Figure 3c).

Structural variant breakpoint pair detection

Breakpoint pairs for each patient were called on the merged BAM files using `hic_breakfinder` (commit 30a0dcc6d01859797d7c263df7335fd2f52df7b8) [92]. Pre-calculated expected observation files for the hg38 genome were downloaded from the git repository on July 24, 2019, as per the instructions. Breakpoints were explicitly called with the following command:

```
hic_breakfinder --bam-file {BAM} --exp-file-inter inter_expect_1Mb.
    hg38.txt --exp-file-intra intra_expect_100kb.hg38.txt --name {
    Sample_ID} --min-1kb.
```

For the T2E fusion, only one patient had the deletion identified by `hic_breakfinder` with default parameters (CPCG0336). Difficulties identifying SVs with `hic_breakfinder` have been previously noted [105]. After adjusting the detection threshold, we were able to identify the fusion in other

samples. To ensure the T2E+ tumours were effectively stratified for future analyses, the fusion was annotated using the same coordinates for the other T2E+ samples. No other additions to breakpoint calls were made. Certain breakpoints that appeared to be artefacts were removed, as described below.

Structural variant annotation and graph construction

The contact matrix spanning 5 Mbp upstream and downstream around the breakpoint pairs were plotted and annotated according to previously published heuristics (Supplementary Figure 4 for [92]). Breakpoint pairs that were nearby other breakpoints or did not match the heuristics in this figure were labelled as `UNKNOWN`. These annotations were matched against the annotations identified from the previously published WGS SVs [6]. Breakpoint pairs matching the following criteria were considered as detection artefacts and were ignored.

1. At least one breakpoint was > 1 Mbp
2. At least one breakpoint was surrounded by empty regions of the contact matrix
3. At least one breakpoint corresponded to a TAD or compartment boundary shared across all samples that lacked a distinct sharp edge that is indicative of a chromosomal rearrangement

To identify unique breakpoints that were identified in multiple breakpoint pairs, breakpoints that were within 50 kbp of each other were considered as possibly redundant calls. This distance was considered as the resolution of the non-artefactual calls is 100 kbp. Plotting the contact matrix 5 Mbp around the breakpoint, breakpoints calls were considered the same breakpoint if the sharp edge of each breakpoint was equal to within 5 kbp. Similar in concept to the ChainFinder algorithm [99], we consider each breakpoint as a node in a graph. Nodes are connected if they are detected as a pair of breakpoints by `hic_breakfinder`. Simple SVs are connected components in the breakpoint graph containing only two nodes, and complex variants those with greater than two nodes. A visual representation of these graphs can be found in Figure B.4b. Graphs are displayed with a spring-force layout, adjusted using the Kamada Kawai optimization [138] from the NetworkX Python package (v2.4) [139].

Determination of SV breakpoints altering intra-TADs contacts

Patients are assigned into one of two groups using hierarchical clustering (complete linkage) with the matrix of pairwise BPscore [135] values as a distance matrix. If the clustering equals the

mutated samples from the non-mutated samples (i.e. the clustering matches the mutation status in this locus), then the local topology was classified as `altered` because of the SV.

Virtual 4C

Two parts of the BRAF gene were used as anchors for virtual 4C data: the promoter region (1500 bp upstream, 500 bp downstream of the TSS) and the entire gene downstream of the breakpoint. Contact frequencies from the ICE-normalized, 20 kbp contact matrices were extracted, with the rows as the bins containing the anchor and the columns as the target regions (the x-axes in Figure 3.6e). The row means were calculated to produce a single vector where each element is the average normalized contact frequency between the anchor of interest and the distal 20 kbp bin. These vectors were plotted as lines in Figure 3.6e.

3.4.5 Patient Tumour Tissue H3K27ac ChIP-seqs

ChIP-seq against H3K27ac was previously published for these matching samples in [38]. Sequencing data was processed similarly to the previous publication of this data [38]; however, the hg38 reference genome was used instead of hg19.

Sequence alignment

FASTQ files from single-end sequencing were aligned to the hg38 genome using Bowtie2 (v2.3.4) with the following command:

```
bowtie2 -x {genome} -U {input} 2> {output_report} | samtools view -u
    > {output_bam}
```

For FASTQ files from paired-end sequencing, only the first mate was considered and reads were aligned with the following command:

```
bowtie2 -x {genome} -U {input} -3 50 2> {output_report} | samtools
   view -u > {output_bam}
```

This ensured that all H3K27ac ChIP-seq data had the same format (single-end) and length (52 bp) before alignment to mitigate possible differences in downstream analyses due to different sequencing methods. Duplicate reads were removed with sambamba (v0.6.9) via `sambamba markdup -r` and were then sorted by position using `sambamba sort`.

Peak calling

Peak calling was performed using MACS2 (v2.1.2) [140] with the following command:

```
macs2 callpeak -g hs -f BAM -q 0.005 -B -n {output_prefix} -t {
    seq_chip} -c {seq_input}
```

ENCODE hg38 blacklist regions were then removed from the narrow peaks [141]. Peaks calls are in Supplementary Table 8.

Differential acetylation analysis

Unique peak calls and de-duplicated pull-down and control BAM files from tumour samples were loaded into R with the DiffBind package (v2.14.0) [142] using DESeq2 (v1.26.0) 84 as the differential analysis model. 3 of the 12 samples had low quality peak calls compared to the other 9 and were not considered when calculating differential acetylation (CPCG0268, CPCG0255, and CPCG0336). We considered each unique breakpoint one at a time in the remaining 9 samples. Samples were grouped by their mutation status (i.e. a design matrix where the mutation status is the only covariate) and DiffBind's differential binding analysis method was performed to identify all differentially acetylated regions between the two groups. Acetylation peaks outside of the TADs overlapping the breakpoint were filtered out. Multiple test correction with the Benjamini-Hochberg FDR method [143] was performed on all peaks after all breakpoints were considered, due to similar group stratifications depending on the breakpoint under consideration.

Structural variant breakpoint enrichment

Structural variant breakpoint coordinates from WGS data from the CPC-GENE cohort were obtained from the International Cancer Genome Consortium (structural somatic mutations from the PRAD-CA dataset, release 28). Breakpoint coordinates were lifted over to hg38 coordinates using the liftOver function from the rtracklayer R package (v1.46.0) [144]. Permutation tests were performed with the regioneR R package (v1.18.0) [145], selecting randomized regions from the hg38 genome, excluding the ENCODE blacklist regions [141] and masked loci. 100 permutations were calculated and a one-sided permutation z-test was used to calculate statistical significance.

3.4.6 Primary Tissue RNA Data Analysis

Tumour sample RNA sequencing

Total RNA was extracted for the CPC-GENE tumour samples as previously described [102]. Briefly, total RNA was extracted with mirVana miRNA Isolation Kit (Life Technologies) according to the manufacturer's instructions. RNA samples were sent to BGI Americas where it underwent QC and DNase treatment. For each sample, 200 ng of total RNA was used to construct a TruSeq strand-specific library with the Ribo Zero protocol (Illumina, Cat. #RS-122-2203). The libraries were sequenced on a HiSeq 2000 to a minimal target of 180 million, 2 $\times$ 100 bp paired-end reads.

RNA sequencing data pre-processing

RNA-seq FASTQ files were pseudo-aligned to the hg38 genome using Kallisto (v0.46.1) [146] with the following command:

```
kallisto quant --bootstrap-samples 100 --pseudobam --threads 8 --
    index /path/to/GRCh38.idx --output-dir {output_dir} {input_R1.
    fastq.gz} {input_R2.fastq.gz}
```

Differential gene expression analysis

To assess whether SVs were associated with local gene expression changes, we considered each unique breakpoint one at a time. For each breakpoint, we compared the gene expression between the mutated and non-mutated tumour samples using Sleuth (v0.30.0) [147, 148] with a linear model where the mutation status was the only covariate. To reduce the chance of falsely identifying genes as differentially expressed, only genes located within the TADs (window size $w = 20$) containing breakpoints were considered. Fold-change estimates of each transcript were assessed for significance using a Wald test. Transcript-level p-values are combined to create gene-level p-values using the Lancaster aggregation method provided by the Sleuth package [148]. Correcting for multiple tests was then performed with the Benjamini-Hochberg FDR correction for all genes that were potentially altered in the mutated sample(s).

Chapter 4

Hedging uncertainty in differential gene expression analyses with James-Stein estimators

J.R.H., and M.L. conceptualized the study. J.R.H. derived the statistical estimates and designed and conducted all the experiments. Figures were designed by J.R.H. The manuscript was written by J.H., and M.L. M.L. oversaw the study.

4.1 Abstract

4.2 Introduction

- the two main approaches for reducing error in a model are to reduce the model variance or model bias Figure 4.1
- here we attempt to decrease mean square error (MSE) by simultaneously increasing the bias and decreasing the variance in fold change coefficient estimators
- derivation for the James-Stein (JS) estimator can be found in Section 4.3.1
- Equation for the JS estimator can be seen in Figure 4.1b.
- in theory, this may increase the error of some transcripts, but will decrease MSE for a set of transcripts in aggregate Appendix C.3

First, we derive the JS estimator fold gene expression fold change and relate it to the ordinary least squares (OLS) estimator. Then, using simulations from a highly replicated RNA-seq experiment [149], we compare the differences in statistical inferences between the JS and OLS estimators. Finally, we investigate how the number of transcripts under consideration affects the reduction in MSE, suggesting how this method can be best used in practice.

4.3 Results

4.3.1 Derivation of the James-Stein fold change estimator

For a p-variate normal distribution $Z \sim \mathcal{N}_p(\mu, \Sigma)$ where μ is unknown and Σ is known, the following theorem holds [150]:

Theorem 1 *The estimator $\hat{\mu}^{(0)} = Z$, for any mean μ, does not minimize the MSE $\mathbb{E}\left[(\mu - \hat{\mu})^2\right]$ for a single observation of Z when $p \geq 3$ and $\Sigma = I_p$. Namely, the estimator $\hat{\mu}^{(JS)} = \left(1 - \frac{b}{a+\|Z\|^2}\right) Z$ has a smaller MSE than $\hat{\mu}^{(0)}$ for a single observation for sufficiently small b and large a.*

This result was generalized to non-singular covariance matrices that were not necessarily the identity matrix (Theorem 2 of [REF 151]):

Theorem 2 *Let $Z \sim \mathcal{N}_p(\mu, \Sigma)$. Let $\hat{\mu}^{(JS)} = \left(1 - \frac{c}{Z^T \Sigma^{-1} Z}\right) Z$. If $p \geq 3$, $Tr(\Sigma) \geq 2\lambda_L$ (where λ_L is the largest eigenvalue of the covariance matrix, Σ), and $0 \leq c \leq 2\left(\frac{Tr(\Sigma)}{\lambda_L} - 2\right)$, then $\hat{\mu}^{(JS)}$ is the estimator for the mean, μ, that minimizes the MSE for a single observation of Z.*

Consider the differential analysis model used in the Sleuth R package [147, 148] for a single transcript, s, with the simple experimental design of n_{wt} WT samples and 1 mutant sample:

$$D_s | Y_s \sim \mathcal{N}_N\left(\beta_{s,0} + \mathbb{I}_{mut}\beta_{s,1}, (\sigma_s^2 + \tau_s^2) I_N\right)$$

For the n_{wt} WT samples, this is equivalent to

$$D_s | Y_s \sim \mathcal{N}_{n_{wt}}\left(\beta_{s,0}, (\sigma_s^2 + \tau_s^2) I_{n_{wt}}\right)$$

which can be fit with the same model process that Sleuth employs. For the single mutated sample, this model is

$$D_s | Y_s \sim \mathcal{N}\left(\beta_{s,0} + \beta_{s,1}, \max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2\right) \tag{4.1}$$

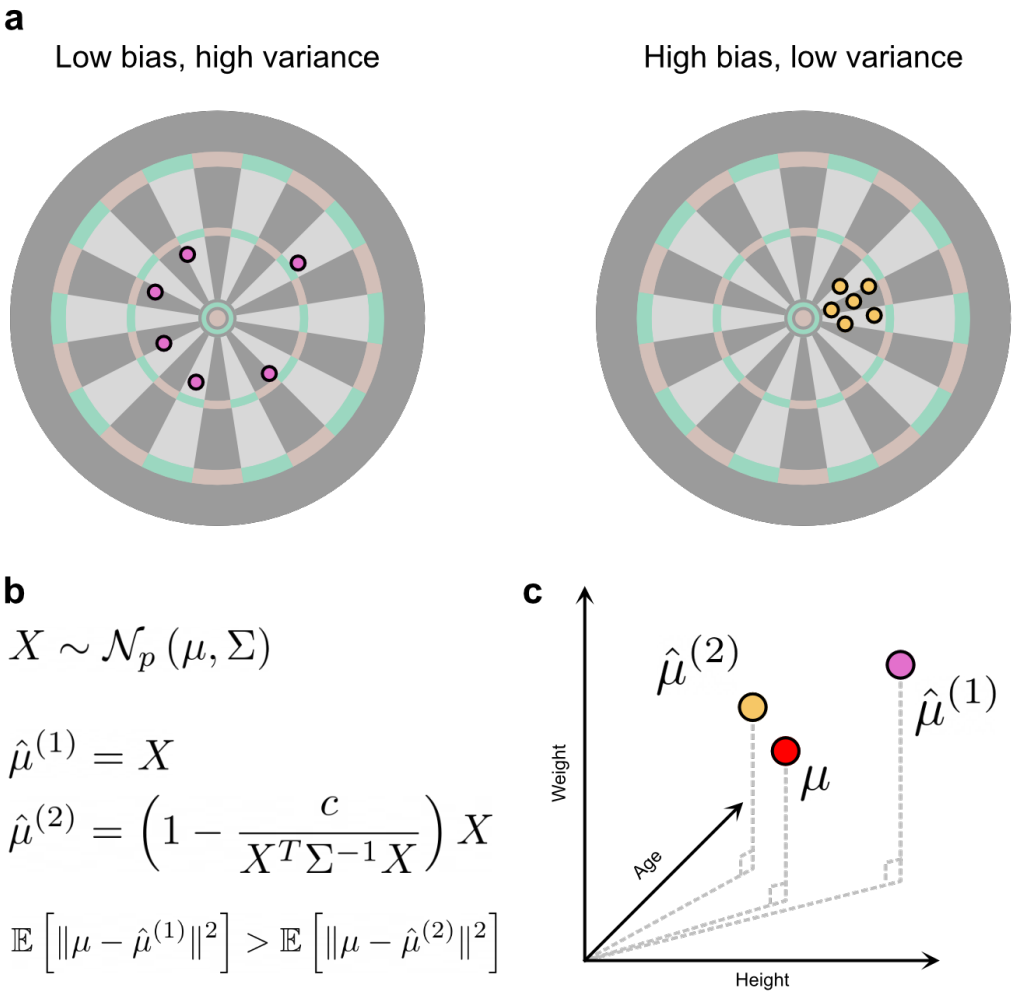

Figure 4.1: **Reducing the bias-variance tradeoff by combining information across multiple features. a.** Schematic of the bias-variance tradeoff for assessing model performance. Dartboard on the left shows low bias of the darts (mean is close to the bullseye) but a large variance. Dartboard on the right shows a high bias of the darts (mean is off-centre), but a small variance. **b.** For a p-variate normal distribution from which a single observation is made, the naive estimator has a higher MSE than the JS estimator, defined as $\hat{\mu}^{(2)}$. **c.** An analogy showing how the JS estimators work in theory. Trying to estimate the mean height, weight, and age for the entire population (μ) from a single person will give an estimate that is likely far from the truth ($\hat{\mu}^{(1)}$). Combining information from the three variables together can produce an estimate that is closer to the truth ($\hat{\mu}^{(2)}$).

The covariance matrix is the same as the mutated samples, but the mean $\beta_{s,0} + \beta_{s,1}$ is unknown. By reparameterizing this model to consider every transcript in the single mutated sample, Equation (4.1) can be re-written as follows:

$$\Delta \sim \mathcal{N}_{|S|}(\mathrm{B}_0 + \mathrm{B}_1, \Sigma) \tag{4.2}$$

where

$$\mathrm{B}_{i,s} = \beta_{s,i} \forall s \in S$$

$$\Sigma = \begin{bmatrix} \max\{\hat{\sigma}_1^2, \tilde{\sigma}_1^2\} + \hat{\tau}_1^2 & & 0 \\ & \ddots & \\ 0 & & \max\{\hat{\sigma}_{|S|}^2, \tilde{\sigma}_{|S|}^2\} + \hat{\tau}_{|S|}^2 \end{bmatrix}$$

We switch from using coefficients $\beta_{t,i}$ to $\mathrm{B}_{i,s}$ to avoid confusion, since $\beta_{t,i} \in \mathbb{R}^p$ (a p-dimensional vector for each covariate in the design) whereas $\mathrm{B}_{i,s} \in \mathbb{R}^{|S|}$ (an $|S|$-dimensional vector for only a single coefficient over all transcripts in S).

Observations of a single mutated sample from this model meet the criteria for the JS estimators. A JS estimator for the unknown fold change, B_1, can be constructed.

$$\hat{\mathrm{B}}_1^{(JS)} = \left(1 - \frac{c}{(\Delta - \hat{\mathrm{B}}_0)^T \Sigma^{-1} (\Delta - \hat{\mathrm{B}}_0)}\right)(\Delta - \hat{\mathrm{B}}_0) \tag{4.3}$$

where $\hat{\mathrm{B}}_0$ is the estimate obtained from the non-mutated samples for all transcripts $s \in S$.

It is straightforward to see that $\mathrm{Tr}(\Sigma) = \sum_{s \in S} \max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2$ and that $\lambda_L = \max_{s \in S} \left\{\max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2\right\}$.

4.3.2 Comparison between the OLS and James-Stein estimators

For a simple experimental design where the mutation status is the only coefficient the OLS estimator is given by:

$$\begin{bmatrix} \hat{\beta}_{s,0}^{(OLS)} \\ \hat{\beta}_{s,1}^{(OLS)} \end{bmatrix} = \hat{\beta}_s^{(OLS)} = (X^T X)^{-1} X^T d_s = \begin{bmatrix} \bar{d}_s^{(wt)} \\ d_s^{(mut)} - \bar{d}_s^{(wt)} \end{bmatrix}$$

where

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

is the design matrix. Looking closely at the OLS estimator for the mutation coefficient, $\beta_{s,1}$, it is clear that it is given by:

$$\hat{\beta}_{s,1}^{(OLS)} = d_s^{(mut)} - \hat{\beta}_{s,0}^{(OLS)} = \delta_s - \hat{\beta}_{0,s} \tag{4.4}$$

which is used directly in the definition of the JS estimator in Equation (4.3). The JS estimator for B_1 can then be expressed simply as:

$$\hat{\mathrm{B}}_1^{(JS)} = \left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T \Sigma^{-1} \hat{\mathrm{B}}_1^{(OLS)}}\right) \hat{\mathrm{B}}_1^{(OLS)} \tag{4.5}$$

From this definition, it is easy to see that the JS estimate is colinear with the OLS estimate but uniformly shrunk towards 0. For a more general experimental design, the above can be extended.

Theorem 3 *Given an experimental design matrix $X \in \mathbb{R}^{n\times p}$ where $n > p$, rank$(X) = p$ and rank$(X^*) = p - 1$ where $X^* \in \mathbb{R}^{(n-1)\times p}$ is the same design matrix but with one sample removed, a JS estimator for the linear coefficient uniquely specified by the one sample is given by*

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

It can be shown that the JS estimator is biased towards 0 with a smaller variance than the OLS estimator (see Appendix C.3).

From this definition, we can see two parameters of this model that will affect the amount of biasing: the scaling coefficient, c, and the total number of transcripts being considered, $|S|$. Firstly, the scaling coefficient can be manually specified, and the largest biasing occurs when c is its maximum value, $2\left(\frac{\mathrm{Tr}(\Sigma)}{\lambda_L} - 2\right)$. Secondly, the transcripts under consideration can also be manually specified, which will affect the value of the denominator $\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^T \Sigma^{-1} \hat{\mathrm{B}}_1^{(OLS)}$, and thus the amount of biasing. The more transcripts under consideration, the larger the denominator, and thus the smaller the effect, compared to the OLS method. Thus, we have produced a high-bias, low-variance fold change estimator that has a lower MSE than the OLS estimator and two tunable parameters.

4.3.3 Empirical analysis of the James-Stein estimator

Now that the statistical properties of the JS estimator have been shown, we use empirical RNA-seq data to measure the performance of this method in practice. To demonstrate this, we make use of a highly replicated RNA-seq experiment involving Δ *Snf2* knockout (KO) and WT yeast cells [149]. This dataset contains 48 biological replicates of each condition, an infeasible sample size for most RNA-seq experiments. Experiments with small sample sizes can be compared to the full dataset to estimate the number of true and false detections for a given experimental design and a given method. We randomly select N total samples from the full dataset with an optimal even split between the two groups (i.e. $N/2$ Δ *Snf2* and $N/2$ WT) or a suboptimal $N-1$-vs-1 split (i.e. $N-1$ Δ *Snf2* and 1 WT or 1 Δ *Snf2* and $N-1$ WT). To measure the effect of total sample size, these simulations are repeated for multiple values of N.

We find that for all values of N, the JS method produces more true positive (TP) and false positive (FP) calls, as well as fewer true negative (TN) and false negative (FN) calls, than the OLS method with suboptimal designs, on average (two-way Analysis of Variance (ANOVA), $p < 2.2 \times 10^{-16}$; Figure 4.2). For example, for $N = 6$, the JS method identified 642.87 TP, 58 FP, 2098.3 TN, and 2995.7 FN calls on average, compared the to 520.7 TP, 41.37 FP, 2114.97 TN, and 3117.93 FN calls for the OLS method (23.5 % more TP, 40 % more FP, 1.8 % fewer TP and 3.9 % fewer FN calls; Figure 4.2). The strength of this effect appears to decrease as the total number of samples increases. Notably, for the $N = 4$ case where a suboptimal design would be most common in practice, the JS method had more TP and fewer FN than the optimal experimental design. In all other cases, however, the optimal even split between Δ *Snf2* and WT groups results in the most TP and fewest FN calls, as expected. Thus, for differential expression hypothesis testing, the OLS method can identify more TP and fewer FN calls than the OLS method when dealing with suboptimal experimental designs.

To investigate where the changes in statistical inferences come from, we can view a representative simulation (Figure 4.3). In the full dataset with 48 biological replicates, Snf2, *Snf2* is the most under-expressed gene with an estimated 99 % reduction in expression (Figure 4.3a). Three other example genes, *PHO12*, *TIS11*, and *TYE7*, are also under-expressed in the Δ *Snf2* cells. Using 4 samples in total, evenly split between the two groups, all four genes remain detected as differentially expressed (Figure 4.3b). Using a suboptimal design with 1 Δ *Snf2* and 3 WT samples, *TIS11* and *TYE7* are no longer detected as differentially expressed using the OLS method (Figure 4.3c). However, the *TIS11* gene is detected as differentially expressed using the JS method (Figure 4.3d). The fold

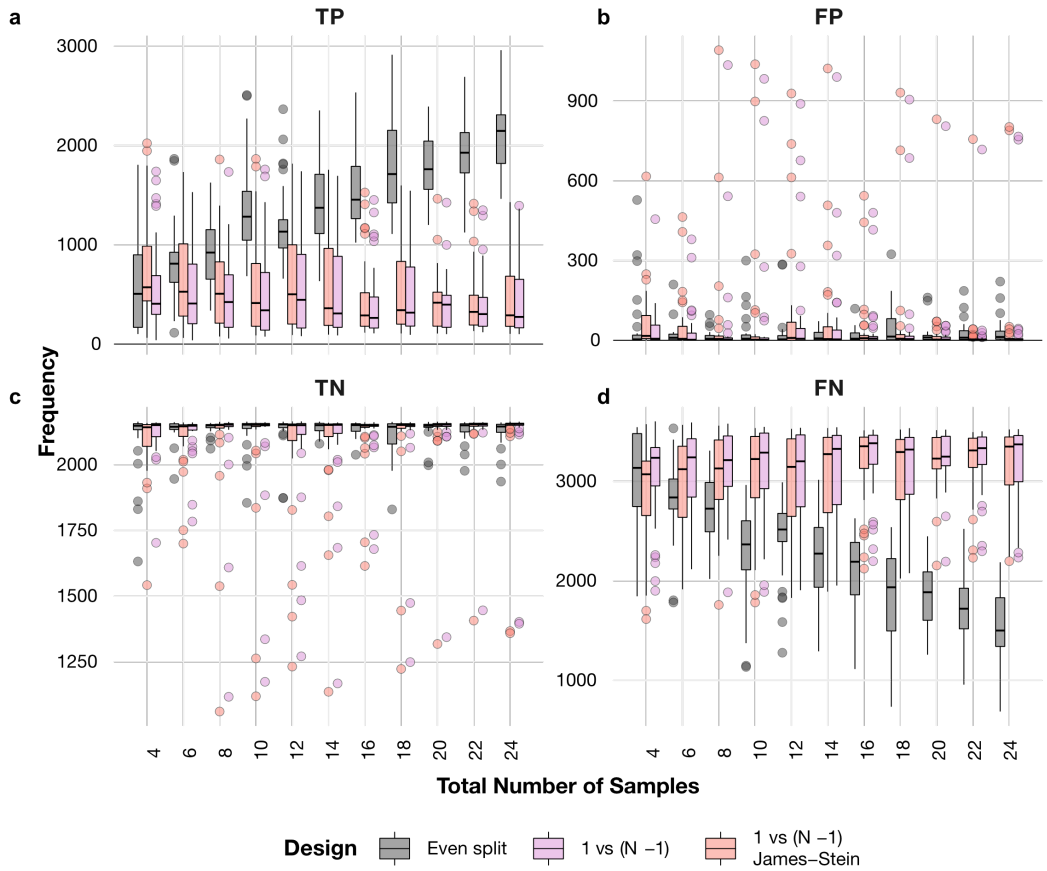

Figure 4.2: **Differential gene expression analysis of the entire yeast transcriptome with differently sized experimental designs.** Simulations ($n = 30$) using randomly selected samples which were then compared to the full dataset of 48 Δ *Snf2* vs 48 WT to calculate TP (**a**), FP (**b**), TN (**c**), and FN (**d**).

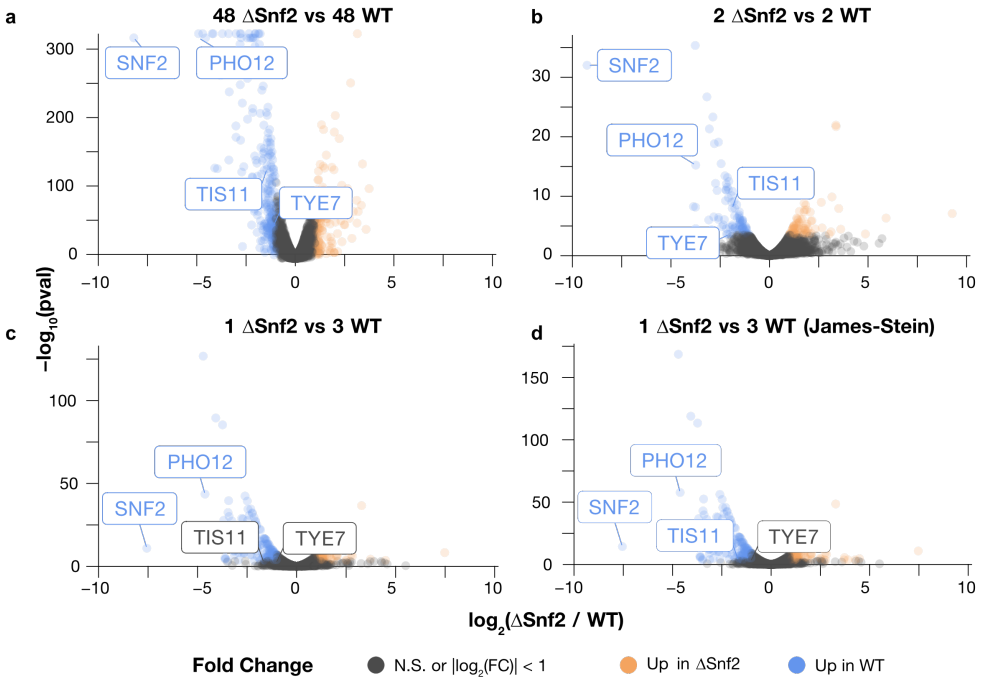

Figure 4.3: **Differential gene expression analysis of Δ *Snf2* vs WT yeast cells using different sample sizes and experimental designs. a.** Volcano plot of differential expression results with OLS estimates in a highly replicated experiment consisting of 48 biological replicates of each condition. **b.** The same analysis as (**a**) using 4 samples in total, 2 Δ *Snf2* and 2 WT samples. **c.** The same analysis as (**c**) using 1 Δ *Snf2* and 3 WT. **d.** The same analysis as (**c**) using the JS method instead of OLS.

change estimates in Figure 4.3c-d are similar, since the biasing is small effect. Thus, the differences in statistical inference result from the smaller variance of the JS estimator, as predicted.

Finally, we investigated the impact of the number of transcripts considered on MSE reduction. Using the same yeast RNA-seq data, we randomly selected $N = 6$ samples and a small number of transcripts from the entire transcriptome, then performed differential expression analysis with the OLS and JS methods. This was repeated 30 times (15 simulations of 5 Δ *Snf2* and 1 WT and 15 simulations of 1 Δ *Snf2* and 5 WT) with a total of $|S| \in \{3, 10, 25, 50, 100, 250\}$ transcripts. Nearly all simulations show a smaller MSE when using the JS method compared to the OLS method (dots below the dotted lines in Figure 4.4a). Moreover, the MSE is reduced by 7.73 - 22.68 % using the JS method, on average, regardless of the number of transcripts considered. The largest percentage of MSE reduction is observed when 3 transcripts are chosen, with an ~86 % error reduction (Figure 4.4b). However, the effect is also more variable when fewer transcripts are considered, as some simulations with 3 transcripts resulted in an ~150 % increase in error (Figure 4.4b). Taken together, we find that the JS method significantly reduces the fold change MSE, with greater reductions found by considering a smaller number of transcripts.

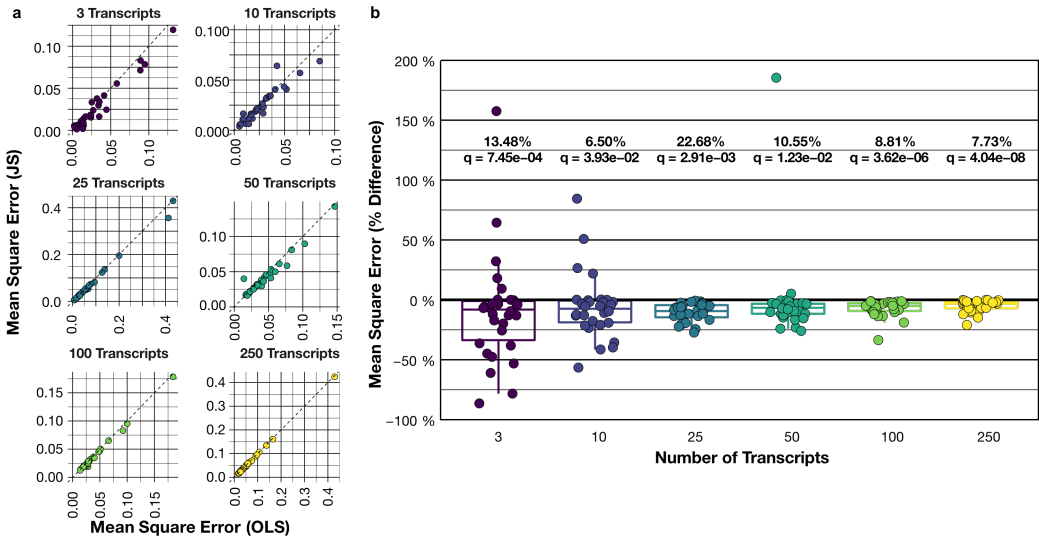

Figure 4.4: **Differential gene expression analysis focusing on a subset of trancsripts, not the entire transcriptome.** All experiments use 1 Δ *Snf2* vs 5 WT samples (or vice versa). **a.** Comparison of the MSE of the JS estimates (y-axis) against the OLS estimates (x-axis). The total number of transcripts in each comparison is specified above each facet. **b.** Percent different in MSE between the JS and OLS estimates. One-sided, two-sample paired Student's t-test, $n = 30$, FDR multiple test corrections.

4.4 Discussion

4.5 Methods

4.5.1 RNA sequencing data collection and pre-processing

RNA-seq reads were obtained from [REF 149] from the Seqeuence Read Archive (SRA) (Project Accession PRJEB5348). Briefly, the Δ *Snf2* KO and WT cells were both sequenced as single-end, 50 bp reads, split across seven lanes of a single Illumina HiSeq 2000 flow cell. The quality of the sequencing reads was assessed with FastQC [152]. The *Saccharomyces cerevisiae* R64-1-1 reference transcriptome index from Ensembl (v96) was downloaded from https://github.com/pachterlab/kallisto-transcriptome-indices/. RNA-seq reads from all technical replicates of a single biological replicate were quantified against the reference transcriptome with Kallisto (v0.46.2) [146] with the following command:

```
kallisto quant --bootstrap-samples 100 --pseudobam --threads 8 --
    index /path/to/R64-1-1.idx --output-dir {output_dir} {input.fastq
    .gz} > {output_report}
```

4.5.2 Differential expression analysis

Differential expression analysis was conducted in R (v4.0.2) [153] with the Sleuth package (v0.30.0) [147, 148]. Statistical significance for the fold change of transcripts from the mutation status was determined with a two-sided Wald test. Multiple testing correction was performed using the Benjamini-Hochberg FDR method [143]. Transcripts were determined to be significantly differentially expressed if $FDR < 0.01$.

This method was used for differential expression analysis for the full experimental design (48 Δ *Snf2* replicates and 48 WT replicates), as well as all simulations with smaller sample sizes.

4.5.3 Random sampling and simulations

From the 48 biological replicates of the Δ *Snf2* and WT samples, N total samples were randomly selected, where $N \in \{4, 6, 8, ..., 22, 24\}$. The sampling was repeated 30 times, independently. The samples were chosen according to of the following experimental designs:

1. 30 iterations of $N/2$ Δ *Snf2* vs $N/2$ WT, compared using the OLS estimator
2. 15 iterations of 1 Δ *Snf2* vs $N-1$ WT + 15 iterations of $N-1$ Δ *Snf2* vs 1 WT, compared using the OLS estimator
3. 15 iterations of 1 Δ *Snf2* vs $N-1$ WT + 15 iterations of $N-1$ Δ *Snf2* vs 1 WT, compared using the JS estimator

Differential expression analysis was calculated with the Sleuth package [147, 148]. For the JS calculations, the scaling coefficient, c, was set to its largest value of $2\left(\frac{\mathrm{Tr}(\Sigma)}{\lambda_L} - 2\right)$ to reduce error as much as possible. For each simulation, the number of TP, FP, TN, and FN calls was calculated by comparing the classification of a given transcript as significantly differentially expressed or not (i.e. if $FDR < 0.01$ for a given transcript) in the simulation to the full experiment of 48 Δ *Snf2* vs 48 WT. With this confusion matrix, derived rates, such as the true positive rate, were then calculated.

4.5.4 Random sampling of smaller numbers of transcripts

In a similar fashion to above, 30 iterations of randomly selected samples were chosen to match the same three scenarios as above, with a fixed $N = 6$ (15 iterations with 1 Δ *Snf2* vs 5 WT and 15 iterations with 5 Δ *Snf2* vs 1 WT). A subset of transcripts, S, were randomly selected from those transcripts with ≥ 10 estimated reads in all 6 of the randomly selected samples. Fold change

estimates for the three scenarios were calculated as above. The number of transcripts selected was chosen from $|S| \in \{3, 10, 25, 50, 100, 250\}$.

MSE was calculated by comparing the fold change estimate from each iteration to the fold change estimate from the full experiment with 48 Δ *Snf2* vs 48 WT. Percent differences in MSE were calculated as:

$$\text{Percent difference } = \frac{MSE_{JS} - MSE_{OLS}}{MSE_{OLS}}$$

Chapter 5

Epigenetic dynamics underlying B cell acute lymphoblastic leukemia relapse

Chapter 6

Discussion & Future Directions

Appendix A

Supplementary Material for Chapter 2

Table A.1 Prostate cancer SNVs within the *FOXA1* TAD

Table A.2 gRNA for clonal and transient CRISPR/Cas9 and dCas9-KRAB experiments

Table A.3 CRISPR/Cas9 Deletion PCR Validation Primers

Table A.4 RT-PCR mRNA Expression Primers

Table A.5 gRNA for lentiviral-based CRISPR/Cas9 deletion proliferation assays

Table A.6 Primers for MAMA ChIP-qPCR

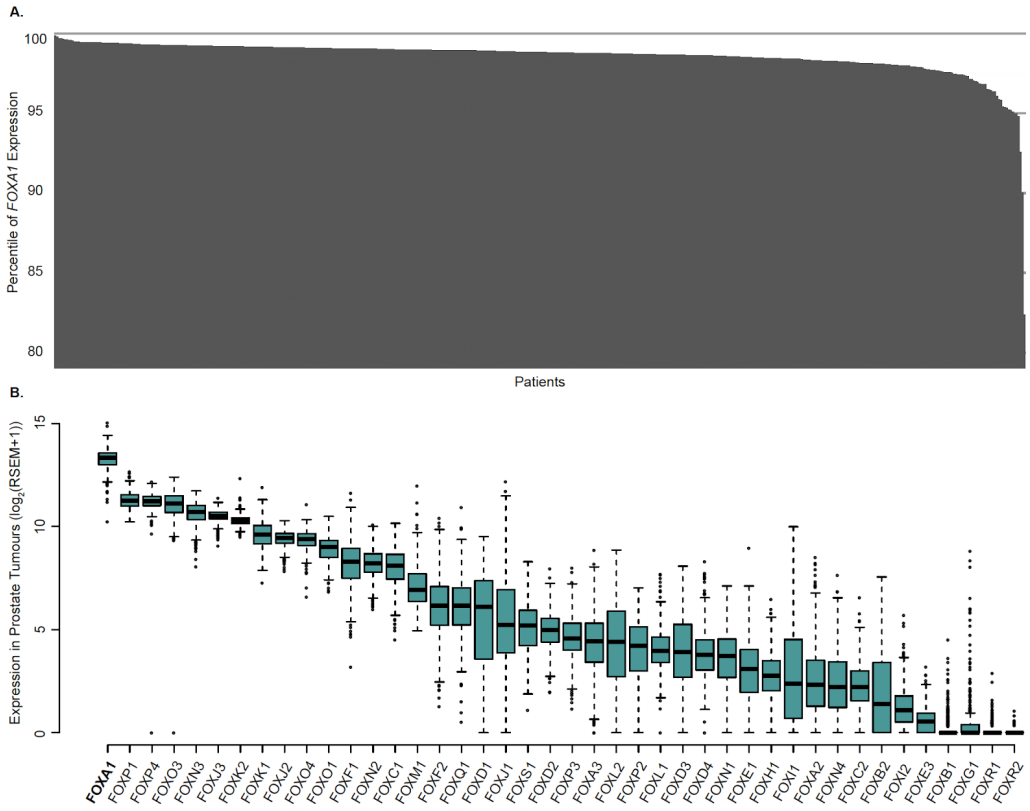

Figure A.1: ***FOXA1* mRNA expression in prostate tumours. a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

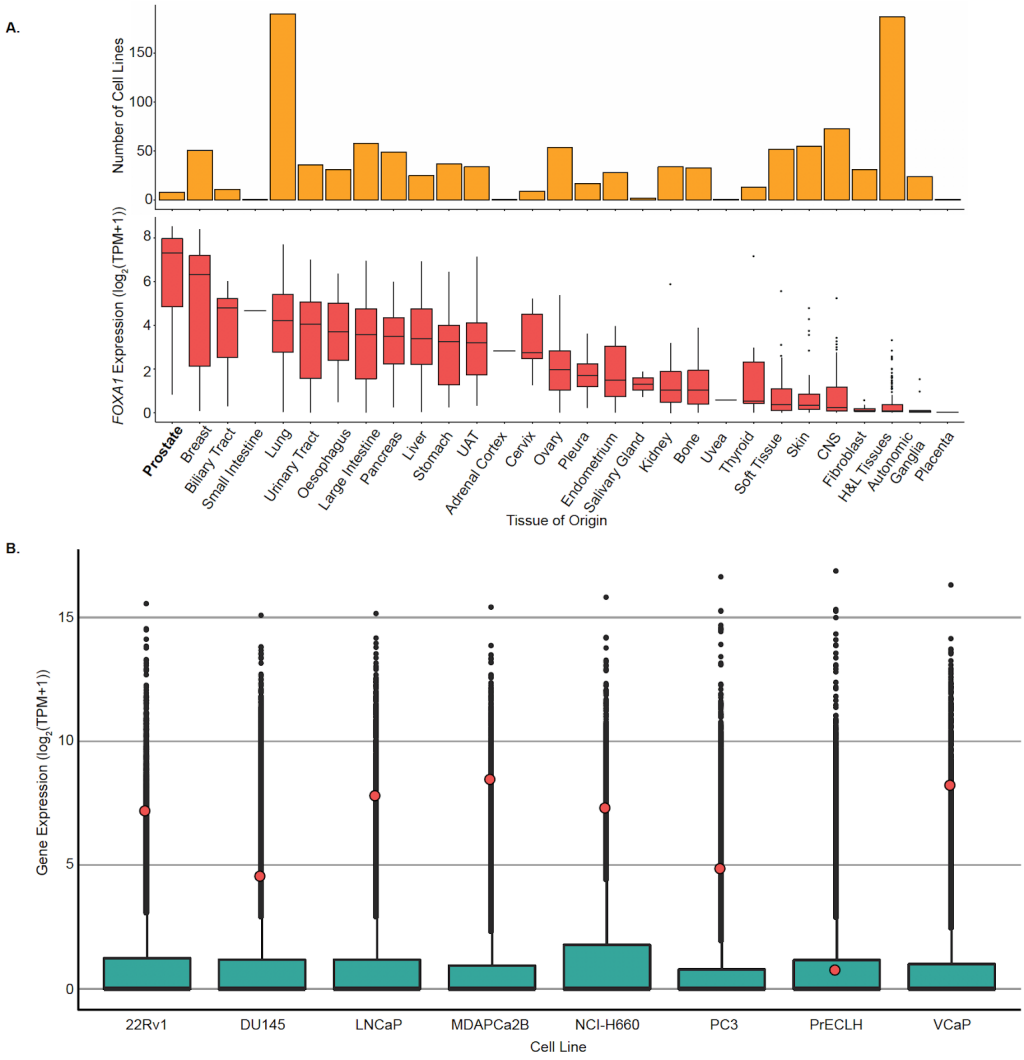

Figure A.2: ***FOXA1* mRNA expression across PCa cell lines. a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Methods). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA expression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

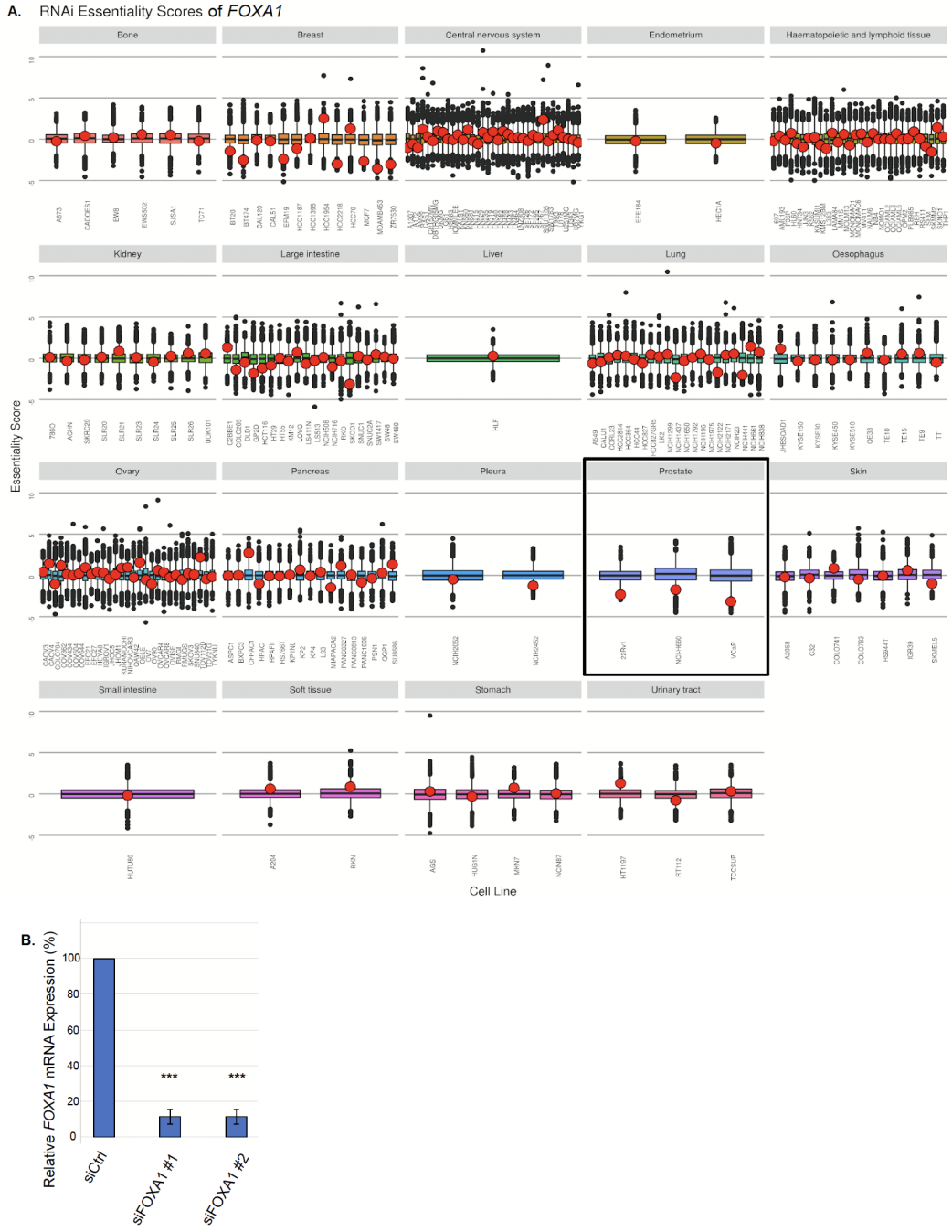

Figure A.3: **Essentiality of *FOXA1* across cancer cell lines of various cancer types. a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot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

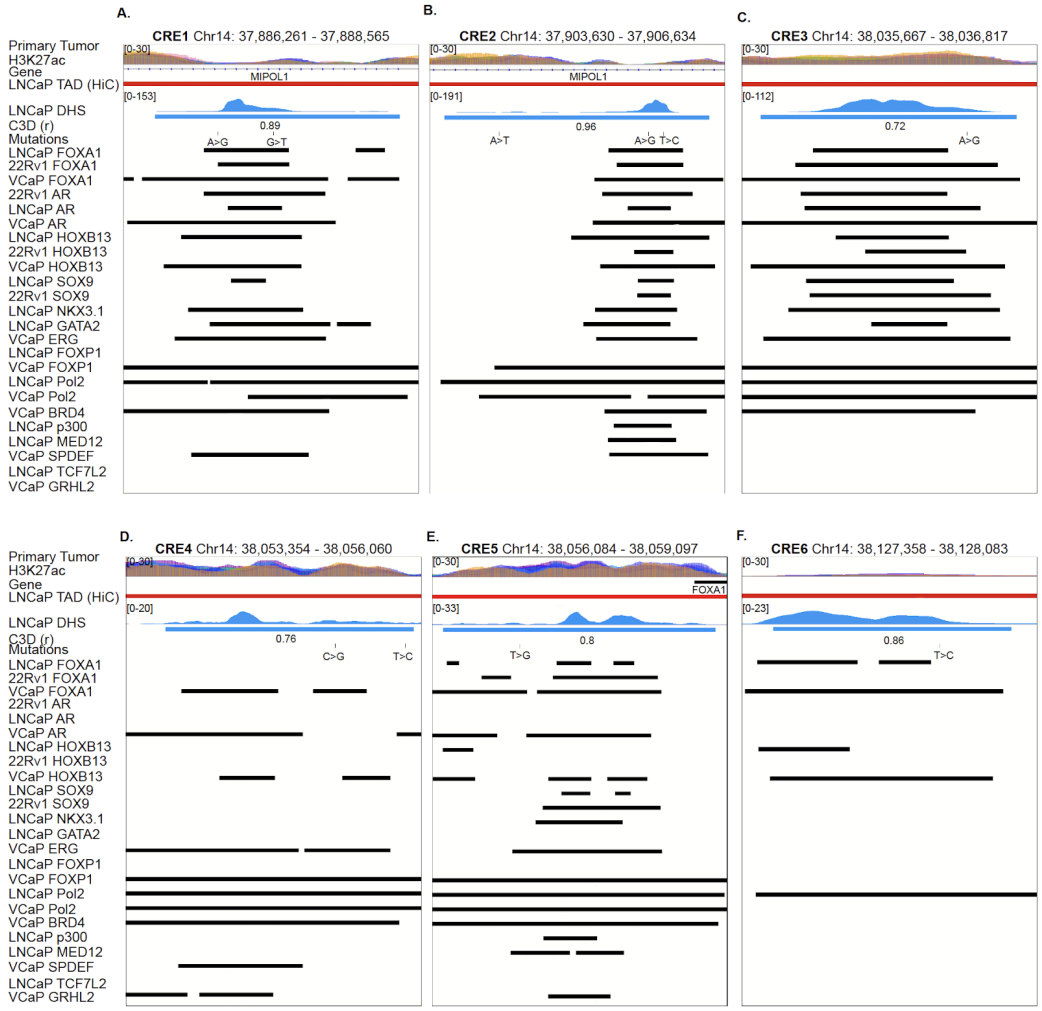

Figure A.4: **Visualization of the functional annotation of the six *FOXA1* CREs. a-f.** Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

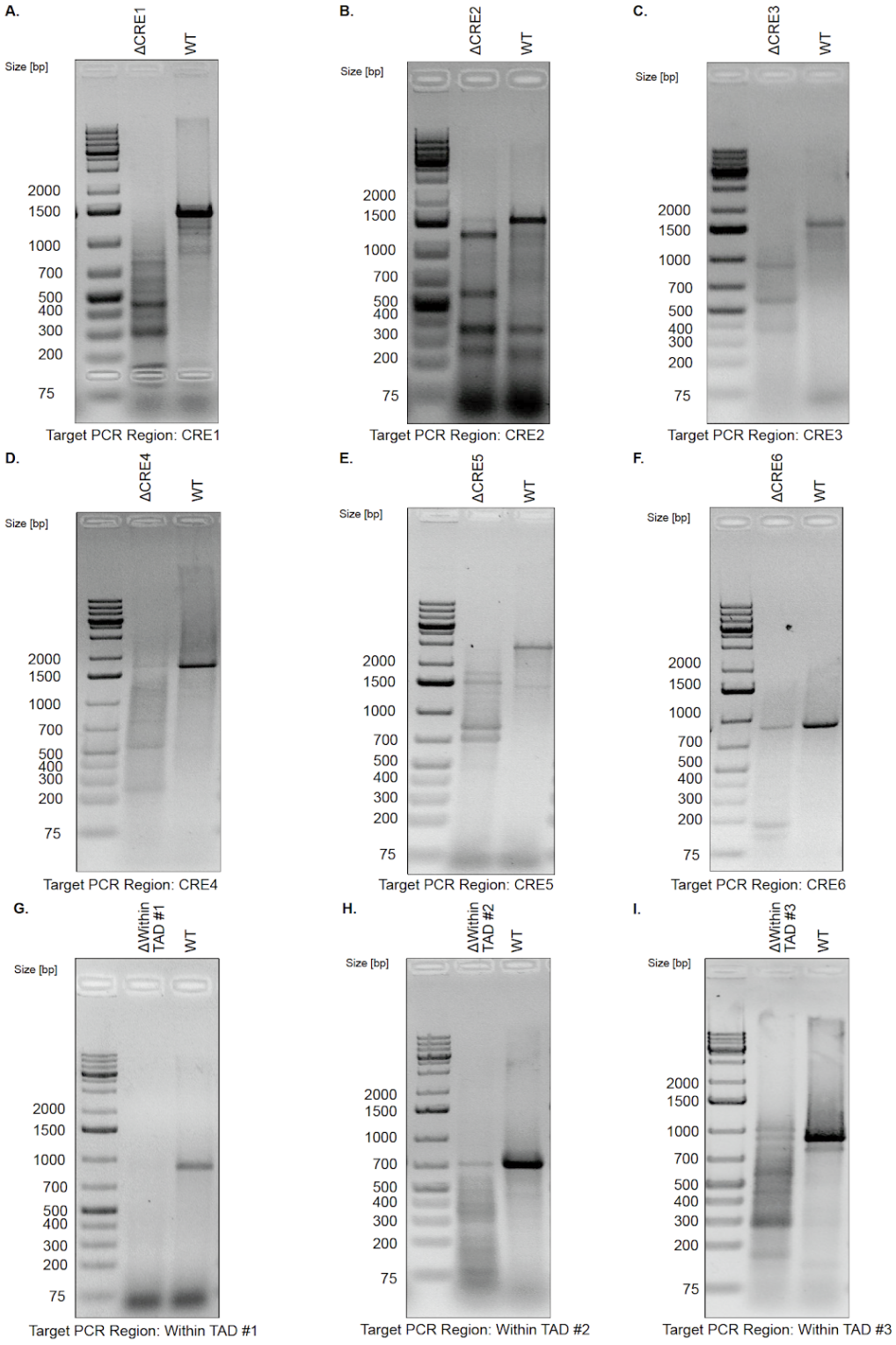

Figure A.5: **Validation of clonal Cas-mediated deletions of CREs. a-f**. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

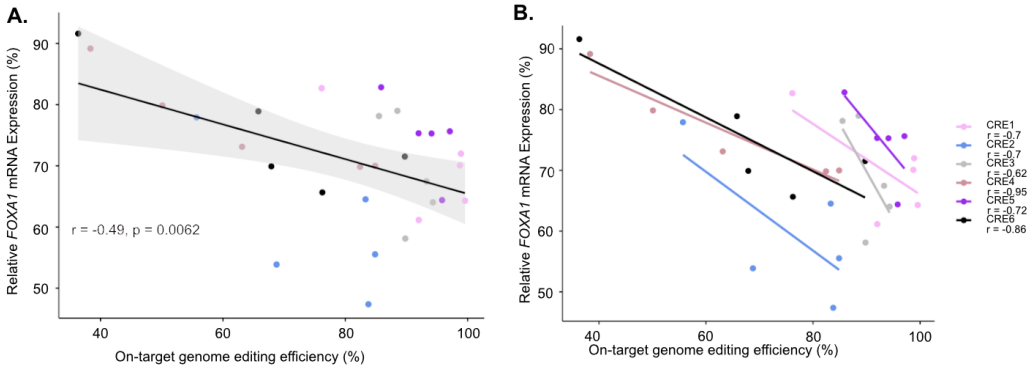

Figure A.6: **Genome editing efficiency (%) is inversely correlated with *FOXA1* mRNA expression. a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all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

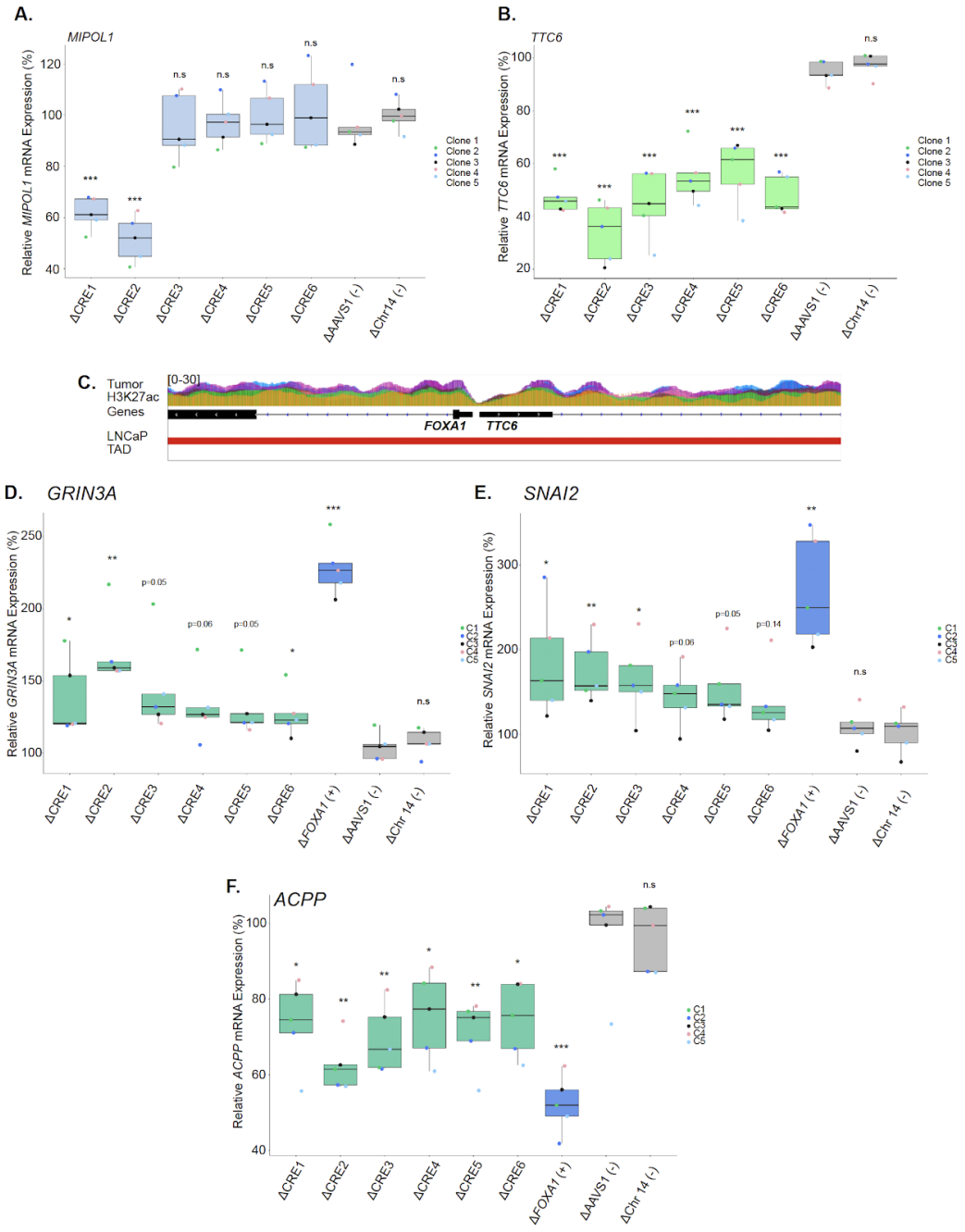

Figure A.7: **Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a.** *MIPOL1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. **b.** *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of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

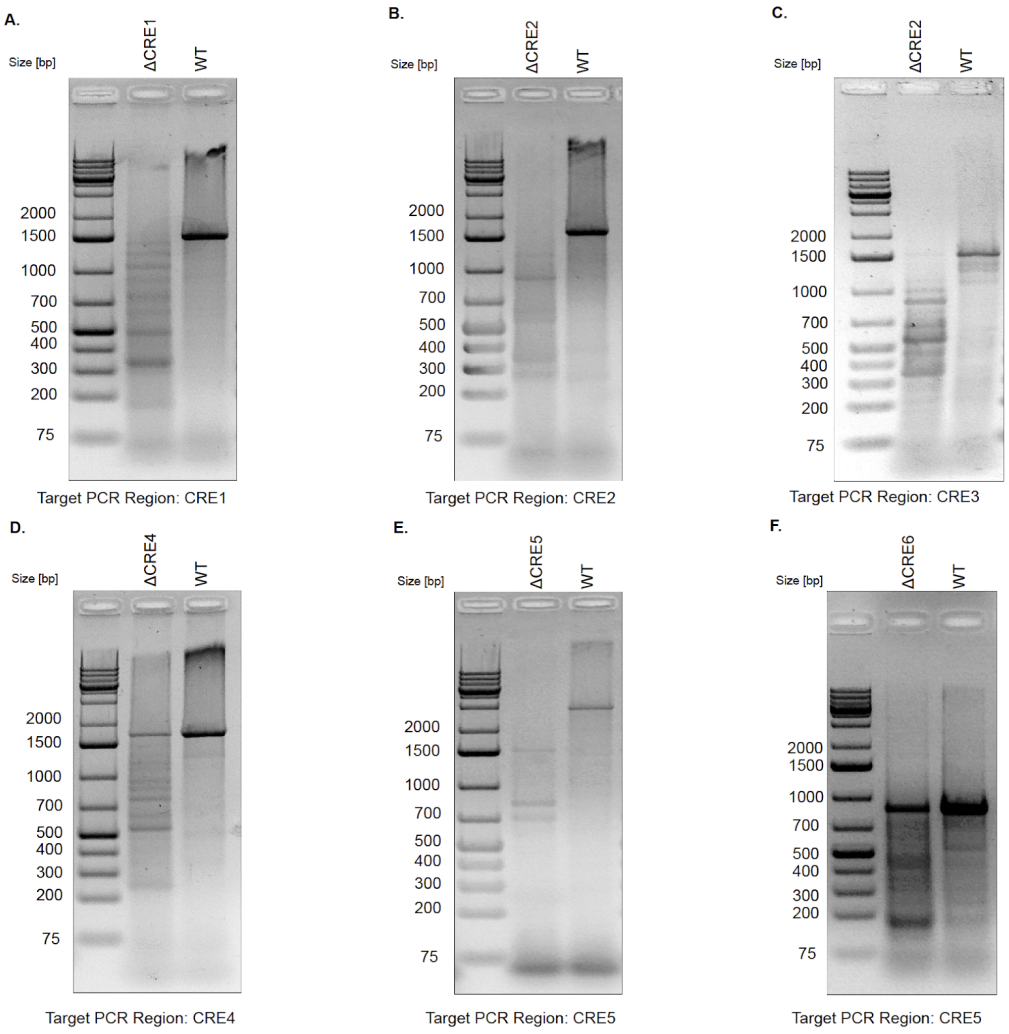

Figure A.8: **Validation of transient Cas9-mediated single deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

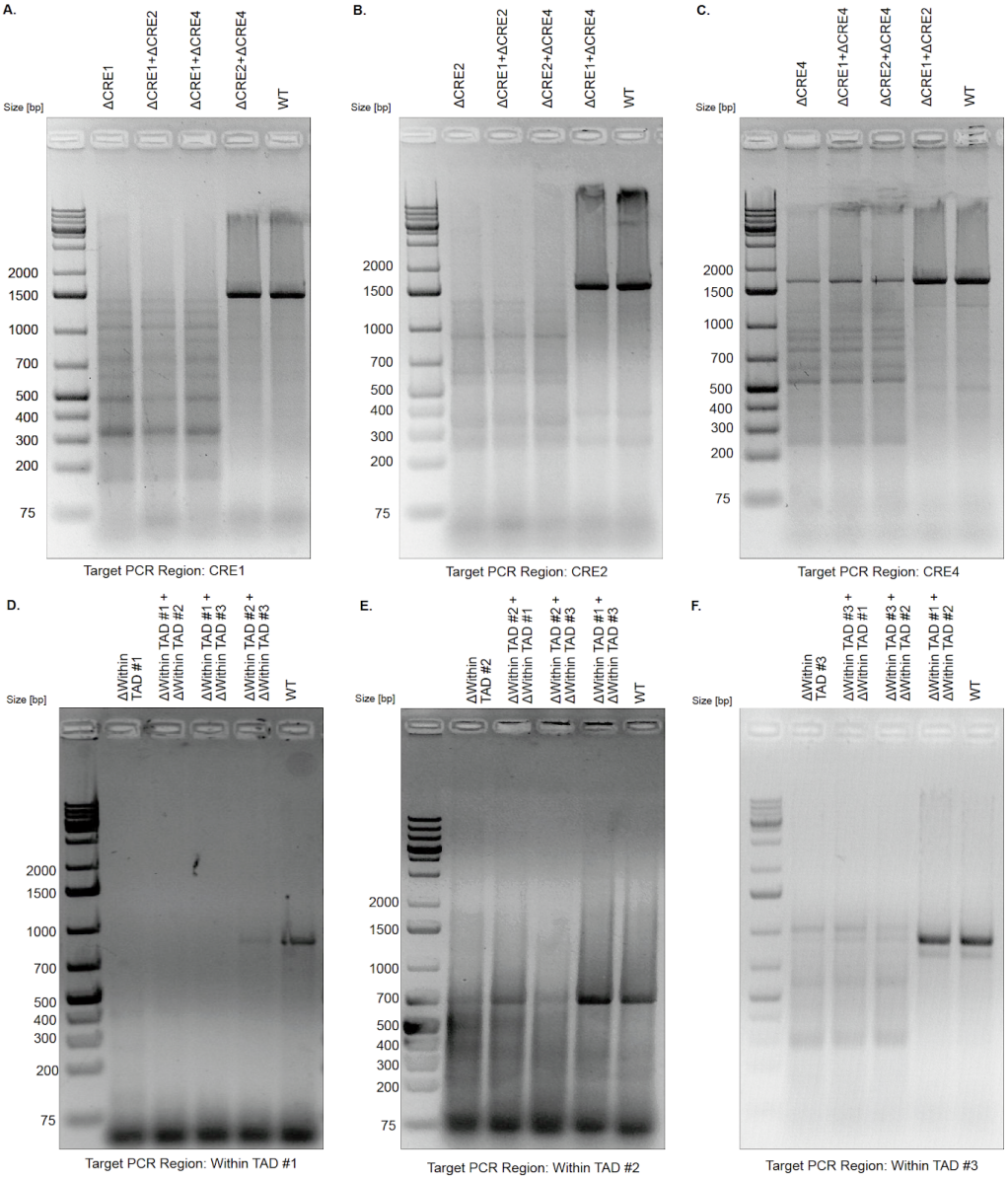

Figure A.9: **Validation of transient Cas9-mediated double deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

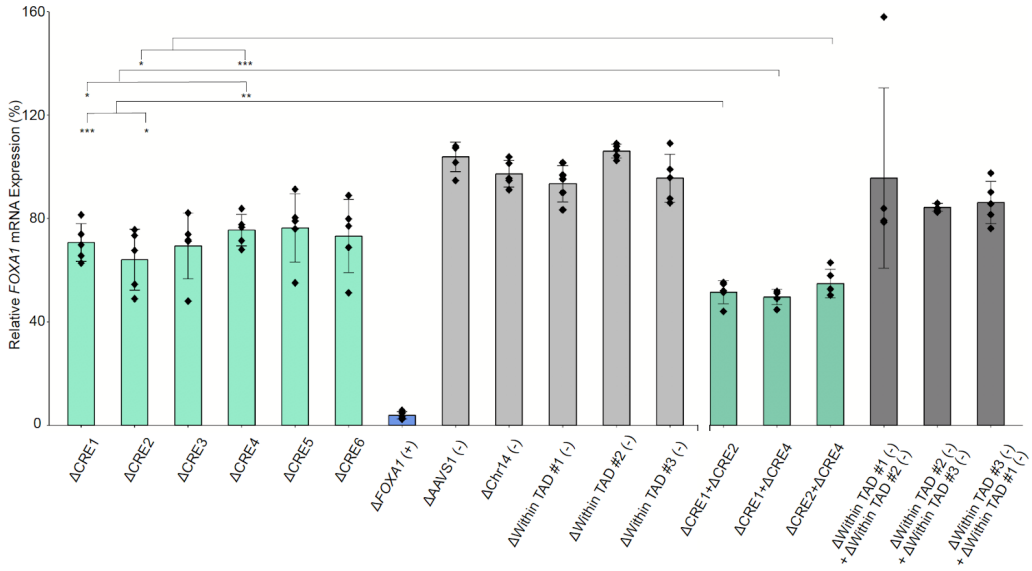

Figure A.10: **Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s).** *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate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

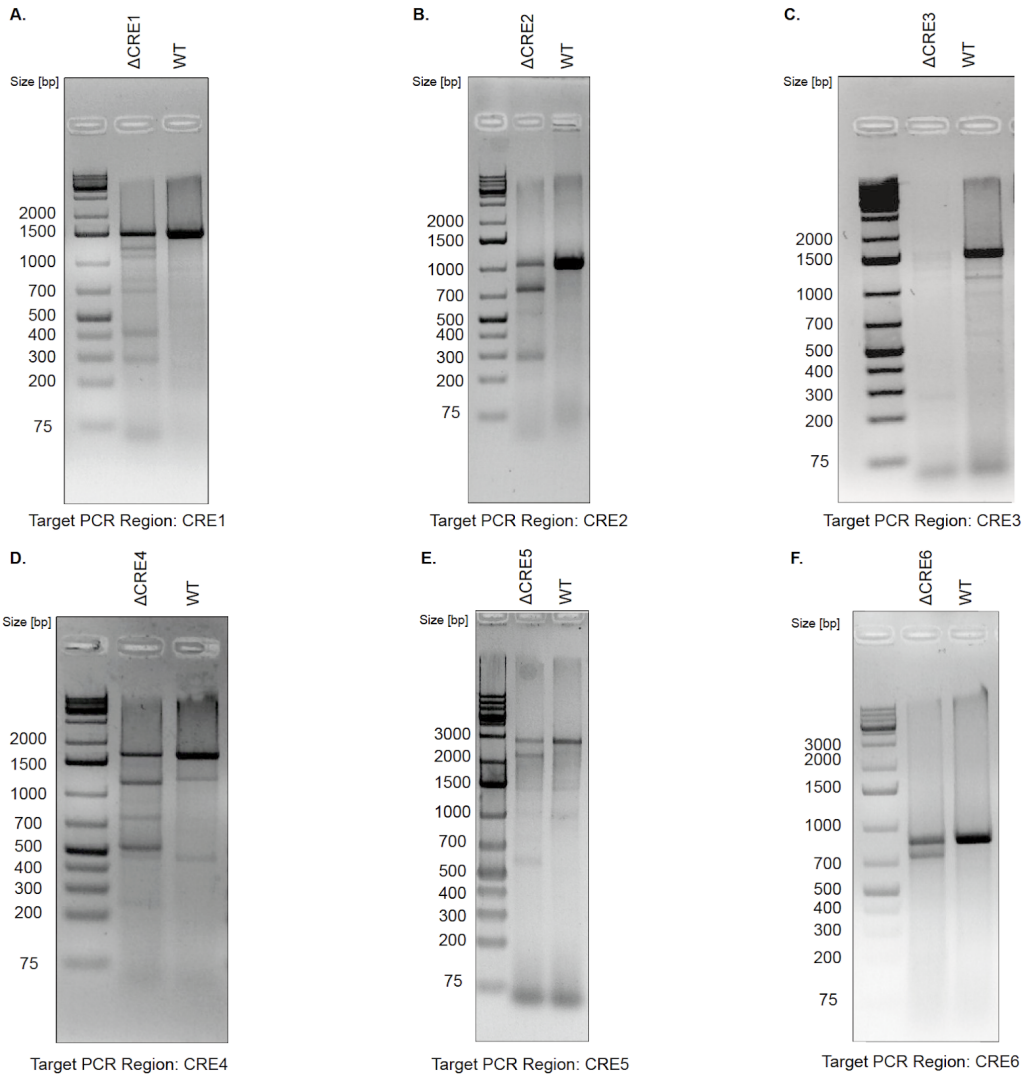

Figure A.11: **Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f.** Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs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

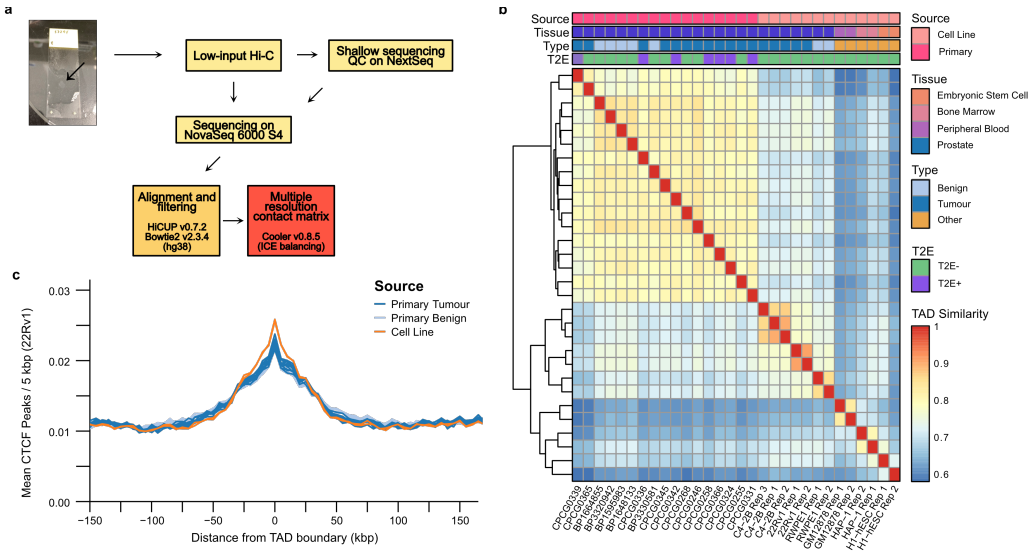

Figure B.1: **Sample processing and TAD similarity between samples. a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate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

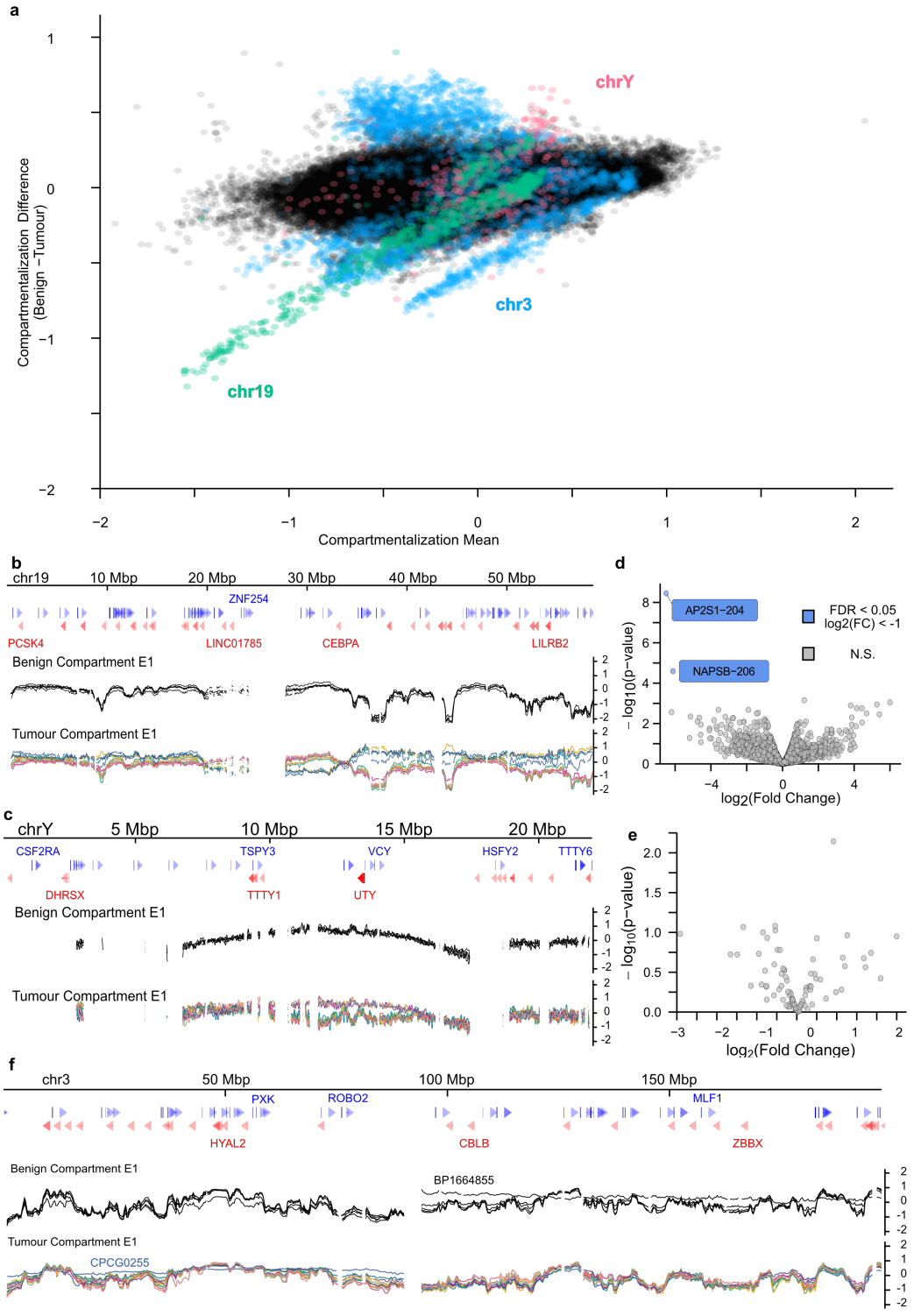

Figure B.2: **Compartmentalization changes in tumours is not associated with widespread differential gene expression.** (Continued on the following page)

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.

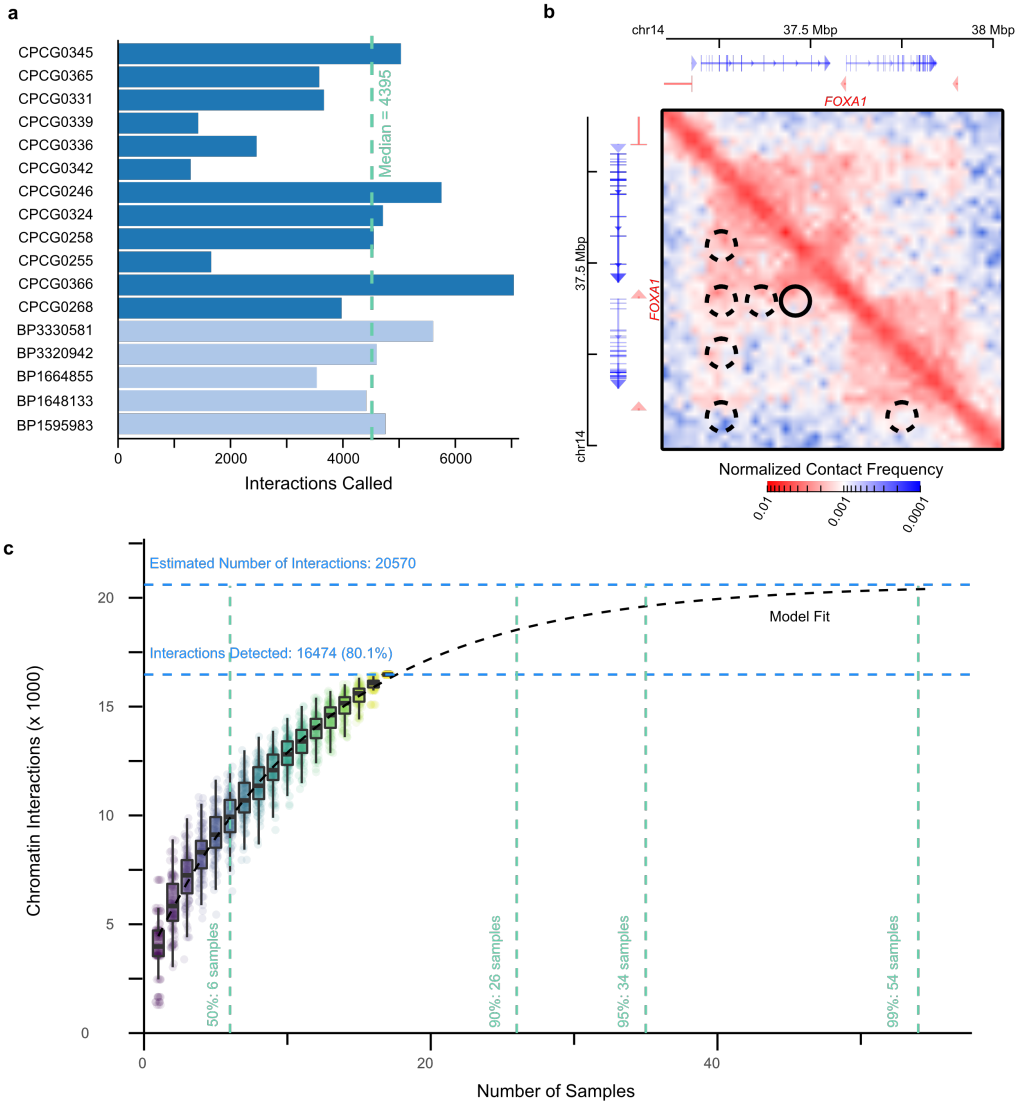

Figure B.3: **Characterization of chromatin interactions in benign and tumour tissue. a.** Bar plot of the number of significant chromatin interactions identified in each of the primary prostate samples. **b.** A snapshot of significant chromatin interactions called around the *FOXA1* gene. Identified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

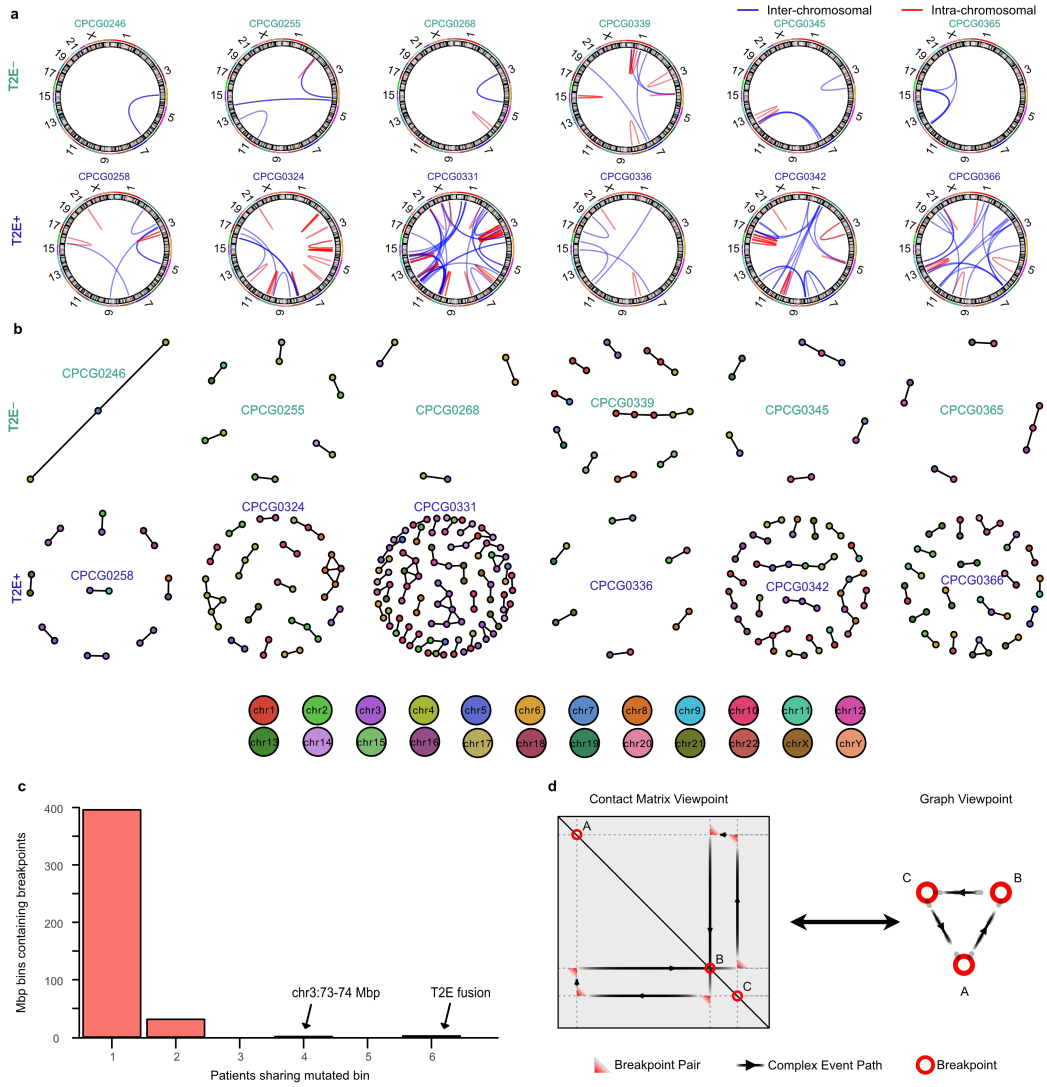

Figure B.4: **Structural variant detection from Hi-C data. a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1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

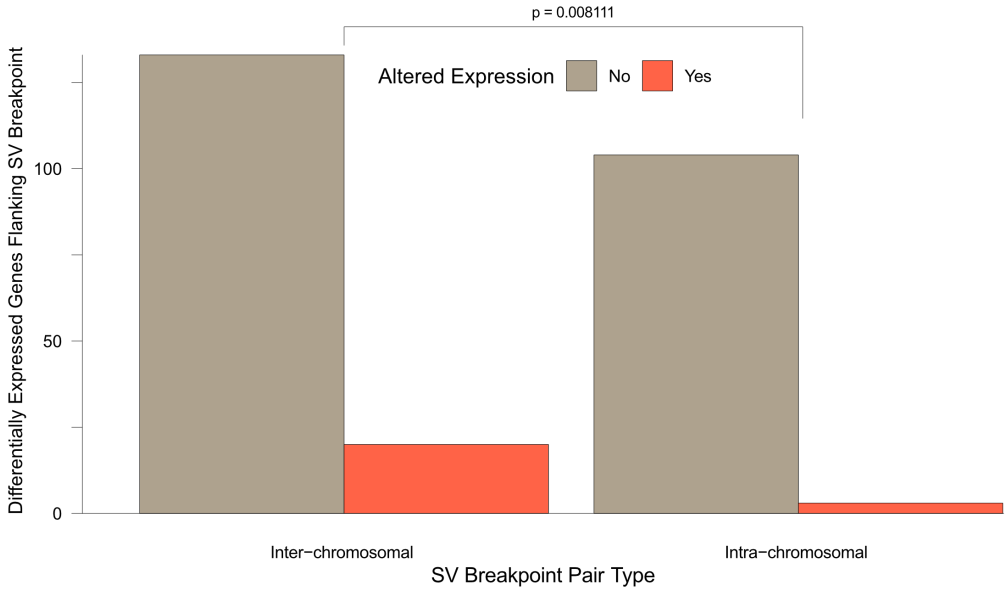

Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes.

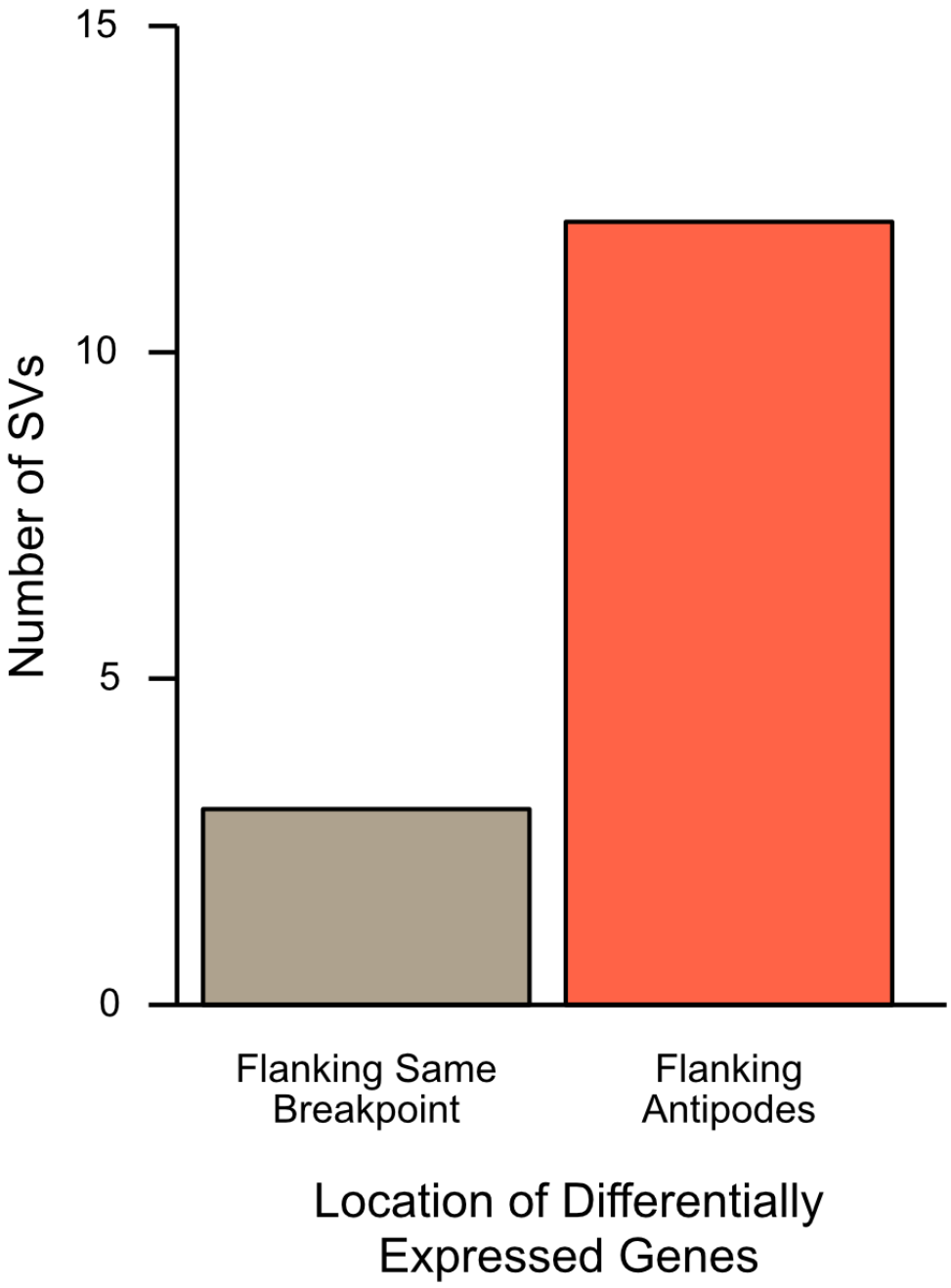

Figure B.6: **Location of differentially expressed genes around SV breakpoints.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

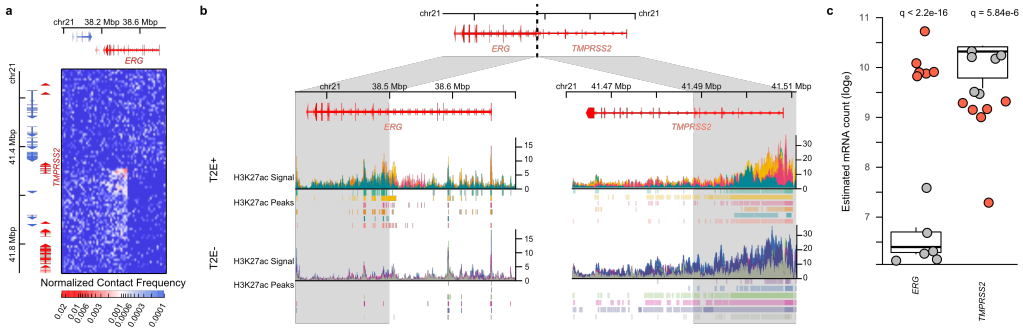

Figure B.7: **Chromatin organization of the *TMPRSS2-ERG* fusion. a.** Contact matrix of the deletion between *TMPRSS2* and *ERG*. **b.** Genome tracks of H3K27ac ChIP-seq signal in T2E+ and T2E- patients. The grey region highlights the loci that come into contact as a result of the deletion. **c.** 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) [147, 148] 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 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) [146] 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 147], 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 JS estimator is an estimate of B_1 that is biased towards 0.

C.3.2 Variance of the James-Stein estimator

The 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 151], $\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] \leq \mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(OLS)} - \mathrm{B}_1\|^2\right]$. 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.

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

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

B-ALL B-cell acute lymphoblastic leukemia

ChIP-seq chromatin immunoprecipitation sequencing

CPC-GENE Canadian Prostate Cancer Genome Network

crRNA CRISPR RNA

CRE *cis*-regulatory element

DEPMAP Cancer Dependency Map

DHS DNase I hypersensitive sites

FN false negative

FP false positive

FOX forkhead box

GLM generalized linear model

gRNA guide RNA

IID independent and identically distributed

JS James-Stein

kbp kilobase

KO knockout

MSE mean square error

mCRPC metastatic castration-resistant prostate cancer

OLS ordinary least squares

mRNA messenger RNA

PCa prostate cancer

RNAi RNA interference

RNA-seq RNA sequencing

shRNA small hairpin RNA

siRNA small interfering RNA

SNV single nucleotide variants

SRA Seqeuence Read Archive

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. Bray, F. *et al.* Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. en. *CA: A Cancer Journal for Clinicians* **68,** 394–424. ISSN: 00079235 (Nov. 2018).

2. Boorjian, S. A. *et al.* Long-Term Outcome After Radical Prostatectomy for Patients With Lymph Node Positive Prostate Cancer in the Prostate Specific Antigen Era. en. *Journal of Urology* **178,** 864–871. ISSN: 0022-5347, 1527-3792 (Sept. 2007).

3. Litwin, M. S. & Tan, H.-J. The Diagnosis and Treatment of Prostate Cancer: A Review. en. *JAMA* **317,** 2532. ISSN: 0098-7484 (June 2017).

4. Attard, G. *et al.* Prostate Cancer. en. *The Lancet* **387,** 70–82. ISSN: 01406736 (Jan. 2016).

5. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. en. *Cell* **163,** 1011–1025. ISSN: 00928674 (Nov. 2015).

6. Fraser, M. *et al.* Genomic Hallmarks of Localized, Non-Indolent Prostate Cancer. en. *Nature* **541,** 359–364. ISSN: 1476-4687 (Jan. 2017).

7. Barbieri, C. E. *et al.* Exome Sequencing Identifies Recurrent SPOP , FOXA1 and MED12 Mutations in Prostate Cancer. en. *Nature Genetics* **44,** 685–689. ISSN: 1546-1718 (June 2012).

8. Grasso, C. S. *et al.* The Mutational Landscape of Lethal Castration-Resistant Prostate Cancer. en. *Nature* **487,** 239–243. ISSN: 0028-0836, 1476-4687 (July 2012).

9. Parolia, A. *et al.* Distinct Structural Classes of Activating FOXA1 Alterations in Advanced Prostate Cancer. en. *Nature* **571,** 413–418. ISSN: 1476-4687 (July 2019).

10. Adams, E. J. *et al.* FOXA1 Mutations Alter Pioneering Activity, Differentiation and Prostate Cancer Phenotypes. en. *Nature* **571,** 408–412. ISSN: 0028-0836, 1476-4687 (July 2019).

11. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. en. *Cell* **161,** 1215–1228. ISSN: 00928674 (May 2015).

12. Robinson, J. L. L., Holmes, K. A. & Carroll, J. S. FOXA1 Mutations in Hormone-Dependent Cancers. *Frontiers in Oncology* **3.** ISSN: 2234-943X (2013).

13. Gao, S. *et al.* Forkhead Domain Mutations in FOXA1 Drive Prostate Cancer Progression. en. *Cell Research* **29,** 770–772. ISSN: 1001-0602, 1748-7838 (Sept. 2019).

14. Annala, M. *et al.* Frequent Mutation of the FOXA1 Untranslated Region in Prostate Cancer. en. *Communications Biology* **1,** 122. ISSN: 2399-3642 (Aug. 2018).

15. Yang, Y. A. & Yu, J. Current Perspectives on FOXA1 Regulation of Androgen Receptor Signaling and Prostate Cancer. en. *Genes & Diseases* **2,** 144–151. ISSN: 23523042 (June 2015).

16. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132,** 958–970. ISSN: 0092-8674 (Mar. 2008).

17. Eeckhoute, J. *et al.* Cell-Type Selective Chromatin Remodeling Defines the Active Subset of FOXA1-Bound Enhancers. en. *Genome Research* **19,** 372–380. ISSN: 1088-9051 (Dec. 2008).

18. Pomerantz, M. M. *et al.* The Androgen Receptor Cistrome Is Extensively Reprogrammed in Human Prostate Tumorigenesis. en. *Nature Genetics* **47,** 1346–1351. ISSN: 1061-4036, 1546-1718 (Nov. 2015).

19. Imamura, Y. *et al.* FOXA1 Promotes Tumor Progression in Prostate Cancer via the Insulin-Like Growth Factor Binding Protein 3 Pathway. en. *PLoS ONE* **7** (ed Agoulnik, I.) e42456. ISSN: 1932-6203 (Aug. 2012).

20. Xu, Y., Chen, S.-Y., Ross, K. N. & Balk, S. P. Androgens Induce Prostate Cancer Cell Proliferation through Mammalian Target of Rapamycin Activation and Post-Transcriptional Increases in Cyclin D Proteins. en. *Cancer Research* **66,** 7783–7792. ISSN: 0008-5472, 1538-7445 (Aug. 2006).

21. Jin, H.-J., Zhao, J. C., Ogden, I., Bergan, R. C. & Yu, J. Androgen Receptor-Independent Function of FoxA1 in Prostate Cancer Metastasis. en. *Cancer Research* **73,** 3725–3736. ISSN: 0008-5472, 1538-7445 (June 2013).

22. Yang, Y. A. *et al.* FOXA1 Potentiates Lineage-Specific Enhancer Activation through Modulating TET1 Expression and Function. en. *Nucleic Acids Research* **44,** 8153–8164. ISSN: 0305-1048, 1362-4962 (Sept. 2016).

23. Zhang, G. *et al.* FOXA1 Defines Cancer Cell Specificity. en. *Science Advances* **2,** e1501473. ISSN: 2375-2548 (Mar. 2016).

24. Augello, M. A., Hickey, T. E. & Knudsen, K. E. FOXA1: Master of Steroid Receptor Function in Cancer: FOXA1: Master of Steroid Receptor Function in Cancer. en. *The EMBO Journal* **30,** 3885–3894. ISSN: 02614189 (Oct. 2011).

25. Sunkel, B. *et al.* Integrative Analysis Identifies Targetable CREB1/FoxA1 Transcriptional Co-Regulation as a Predictor of Prostate Cancer Recurrence. en. *Nucleic Acids Research* **45,** 6993–6993. ISSN: 0305-1048, 1362-4962 (June 2017).

26. Ni, M. *et al.* Amplitude Modulation of Androgen Signaling by C-MYC. en. *Genes & Development* **27,** 734–748. ISSN: 0890-9369 (Apr. 2013).

27. Sasse, S. K. & Gerber, A. N. Feed-Forward Transcriptional Programming by Nuclear Receptors: Regulatory Principles and Therapeutic Implications. en. *Pharmacology & Therapeutics* **145,** 85–91. ISSN: 01637258 (Jan. 2015).

28. Wang, S., Singh, S., Katika, M., Lopez-Aviles, S. & Hurtado, A. High Throughput Chemical Screening Reveals Multiple Regulatory Proteins on FOXA1 in Breast Cancer Cell Lines. en. *International Journal of Molecular Sciences* **19,** 4123. ISSN: 1422-0067 (Dec. 2018).

29. Rowley, M. J. & Corces, V. G. Organizational Principles of 3D Genome Architecture. en. *Nature Reviews Genetics* **19,** 789–800. ISSN: 1471-0056, 1471-0064 (Dec. 2018).

30. Vernimmen, D. & Bickmore, W. A. The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. en. *Trends in Genetics* **31,** 696–708. ISSN: 01689525 (Dec. 2015).

31. Sallari, R. C. *et al.* Convergence of Dispersed Regulatory Mutations Predicts Driver Genes in Prostate Cancer. *bioRxiv,* 38–38 (2016).

32. Bailey, S. D. *et al.* Noncoding Somatic and Inherited Single-Nucleotide Variants Converge to Promote ESR1 Expression in Breast Cancer. *Nature Genetics* **48,** 1260–1269 (2016).

33. Tsourlakis, M. C. *et al.* FOXA1 Expression Is a Strong Independent Predictor of Early PSA Recurrence in ERG Negative Prostate Cancers Treated by Radical Prostatectomy. en. *Carcinogenesis* **38,** 1180–1187. ISSN: 0143-3334, 1460-2180 (Dec. 2017).

34. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. en. *Cell* **137,** 1194–1211. ISSN: 00928674 (June 2009).

35. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. en. *Cell* **171,** 1573–1588.e28. ISSN: 00928674 (Dec. 2017).

36. Bailey, S. D. *et al.* ZNF143 Provides Sequence Specificity to Secure Chromatin Interactions at Gene Promoters. en. *Nature Communications* **6,** 6186. ISSN: 2041-1723 (May 2015).

37. Mehdi, T., Bailey, S. D., Guilhamon, P. & Lupien, M. C3D: A Tool to Predict 3D Genomic Interactions between Cis-Regulatory Elements. en. *Bioinformatics* **35** (ed Kelso, J.) 877–879. ISSN: 1367-4803, 1460-2059 (Mar. 2019).

38. Kron, K. J. *et al.* TMPRSS2–ERG Fusion Co-Opts Master Transcription Factors and Activates NOTCH Signaling in Primary Prostate Cancer. en. *Nature Genetics* **49,** 1336–1345. ISSN: 1546-1718 (Sept. 2017).

39. Creyghton, M. P. *et al.* Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State. en. *Proceedings of the National Academy of Sciences* **107,** 21931–21936. ISSN: 0027-8424, 1091-6490 (Dec. 2010).

40. Espiritu, S. M. G. *et al.* The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. English. *Cell* **173,** 1003–1013.e15. ISSN: 0092-8674, 1097-4172 (May 2018).

41. DeKelver, R. C. *et al.* Functional Genomics, Proteomics, and Regulatory DNA Analysis in Isogenic Settings Using Zinc Finger Nuclease-Driven Transgenesis into a Safe Harbor Locus in the Human Genome. en. *Genome Research* **20,** 1133–1142. ISSN: 1088-9051 (Aug. 2010).

42. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: Five Essential Questions. en. *Nature Reviews Genetics* **14,** 288–295. ISSN: 1471-0056, 1471-0064 (Apr. 2013).

43. Rheinbay, E. *et al.* Recurrent and Functional Regulatory Mutations in Breast Cancer. *Nature* **547,** 55–60. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (June 2017).

44. Zhang, X., Cowper-Sal{middle dot}lari, R., Bailey, S. D., Moore, J. H. & Lupien, M. Integrative Functional Genomics Identifies an Enhancer Looping to the SOX9 Gene Disrupted by the 17q24.3 Prostate Cancer Risk Locus. en. *Genome Research* **22,** 1437–1446. ISSN: 1088-9051 (Aug. 2012).

45. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. en. *Science* **339,** 957–959. ISSN: 0036-8075, 1095-9203 (Feb. 2013).

46. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. en. *Science* **339,** 959–961. ISSN: 0036-8075, 1095-9203 (Feb. 2013).

47. Fuxman Bass, J. I. *et al.* Human Gene-Centered Transcription Factor Networks for Enhancers and Disease Variants. en. *Cell* **161,** 661–673. ISSN: 00928674 (Apr. 2015).

48. Zhou, S., Treloar, A. E. & Lupien, M. Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. *Cancer Discovery* **6,** 1215–1229 (Nov. 2016).

49. Feigin, M. E. *et al.* Recurrent Noncoding Regulatory Mutations in Pancreatic Ductal Adenocarcinoma. en. *Nature Genetics* **49,** 825–833. ISSN: 1061-4036, 1546-1718 (June 2017).

50. Khurana, E. *et al.* Role of Non-Coding Sequence Variants in Cancer. en. *Nature Reviews Genetics* **17,** 93–108. ISSN: 1471-0056, 1471-0064 (Feb. 2016).

51. Cowper-Sal·lari, R. *et al.* Breast Cancer Risk–Associated SNPs Modulate the Affinity of Chromatin for FOXA1 and Alter Gene Expression. en. *Nature Genetics* **44,** 1191–1198. ISSN: 1061-4036, 1546-1718 (Nov. 2012).

52. Rhie, S. K. *et al.* A High-Resolution 3D Epigenomic Map Reveals Insights into the Creation of the Prostate Cancer Transcriptome. en. *Nature Communications* **10,** 1–12. ISSN: 2041-1723 (Sept. 2019).

53. Liu, Q. *et al.* Disruption of a -35 Kb Enhancer Impairs CTCF Binding and *MLH1* Expression in Colorectal Cells. en. *Clinical Cancer Research* **24,** 4602–4611. ISSN: 1078-0432, 1557-3265 (Sept. 2018).

54. Zhang, X. *et al.* Identification of Focally Amplified Lineage-Specific Super-Enhancers in Human Epithelial Cancers. en. *Nature Genetics* **48,** 176–182. ISSN: 1546-1718 (Feb. 2016).

55. Takeda, D. Y. *et al.* A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. English. *Cell* **174,** 422–432.e13. ISSN: 0092-8674, 1097-4172 (July 2018).

56. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. en. *Cell* **174,** 433–447.e19. ISSN: 00928674 (July 2018).

57. Osterwalder, M. *et al.* Enhancer Redundancy Provides Phenotypic Robustness in Mammalian Development. en. *Nature* **554,** 239–243. ISSN: 0028-0836, 1476-4687 (Feb. 2018).

58. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes. en. *Nature Genetics* **47,** 710–716. ISSN: 1061-4036, 1546-1718 (July 2015).

59. Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. English. *Cancer Cell* **36,** 674–689.e6. ISSN: 1535-6108, 1878-3686 (Dec. 2019).

60. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-Wide Analysis of Noncoding Regulatory Mutations in Cancer. en. *Nature Genetics* **46,** 1160–1165. ISSN: 1061-4036, 1546-1718 (Nov. 2014).

61. CAMCAP Study Group *et al.* Sequencing of Prostate Cancers Identifies New Cancer Genes, Routes of Progression and Drug Targets. en. *Nature Genetics* **50,** 682–692. ISSN: 1061-4036, 1546-1718 (May 2018).

62. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. English. *Cell* **174,** 758–769.e9. ISSN: 0092-8674, 1097-4172 (July 2018).

63. The Cancer Cell Line Encyclopedia Consortium & The Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic Agreement between Two Cancer Cell Line Data Sets. en. *Nature* **528,** 84–87. ISSN: 0028-0836, 1476-4687 (Dec. 2015).

64. McFarland, J. M. *et al.* Improved Estimation of Cancer Dependencies from Large-Scale RNAi Screens Using Model-Based Normalization and Data Integration. en. *Nature Communications* **9,** 4610. ISSN: 2041-1723 (Dec. 2018).

65. Thurman, R. E. *et al.* The Accessible Chromatin Landscape of the Human Genome. en. *Nature* **489,** 75–82. ISSN: 0028-0836, 1476-4687 (Sept. 2012).

66. Wang, Y. *et al.* The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions. en. *Genome Biology* **19,** 151. ISSN: 1474-760X (Dec. 2018).

67. Haeussler, M. *et al.* Evaluation of Off-Target and on-Target Scoring Algorithms and Integration into the Guide RNA Selection Tool CRISPOR. en. *Genome Biology* **17,** 148. ISSN: 1474-760X (Dec. 2016).

68. Hsu, P. D. *et al.* DNA Targeting Specificity of RNA-Guided Cas9 Nucleases. en. *Nature Biotechnology* **31,** 827–832. ISSN: 1087-0156, 1546-1696 (Sept. 2013).

69. Sanjana, N. E., Shalem, O. & Zhang, F. Improved Vectors and Genome-Wide Libraries for CRISPR Screening. en. *Nature Methods* **11,** 783–784. ISSN: 1548-7091, 1548-7105 (Aug. 2014).

70. Finn, E. H. & Misteli, T. Molecular Basis and Biological Function of Variability in Spatial Genome Organization. en. *Science* **365,** eaaw9498. ISSN: 0036-8075, 1095-9203 (Sept. 2019).

71. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. en. *Nature* **485,** 376–380. ISSN: 1476-4687 (May 2012).

72. Nora, E. P. *et al.* Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre. en. *Nature* **485,** 381–385. ISSN: 0028-0836, 1476-4687 (May 2012).

73. Pombo, A. & Dillon, N. Three-Dimensional Genome Architecture: Players and Mechanisms. en. *Nature Reviews Molecular Cell Biology* **16,** 245–257. ISSN: 1471-0080 (Apr. 2015).

74. Northcott, P. A. *et al.* Enhancer Hijacking Activates GFI1 Family Oncogenes in Medulloblastoma. en. *Nature* **511,** 428–434. ISSN: 1476-4687 (July 2014).

75. Gröschel, S. *et al.* A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia. en. *Cell* **157,** 369–381. ISSN: 00928674 (Apr. 2014).

76. Flavahan, W. A. *et al.* Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas. en. *Nature* **529,** 110–114. ISSN: 0028-0836, 1476-4687 (Jan. 2016).

77. Haller, F. *et al.* Enhancer Hijacking Activates Oncogenic Transcription Factor NR4A3 in Acinic Cell Carcinomas of the Salivary Glands. en. *Nature Communications* **10,** 368. ISSN: 2041-1723 (Dec. 2019).

78. Oudelaar, A. M. & Higgs, D. R. The Relationship between Genome Structure and Function. en. *Nature Reviews Genetics.* ISSN: 1471-0056, 1471-0064 (Nov. 2020).

79. Despang, A. *et al.* Functional Dissection of the Sox9–Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture. en. *Nature Genetics* **51,** 1263–1271. ISSN: 1061-4036, 1546-1718 (Aug. 2019).

80. Williamson, I. *et al.* Developmentally Regulated *Shh* Expression Is Robust to TAD Perturbations. en. *Development* **146,** dev179523. ISSN: 0950-1991, 1477-9129 (Oct. 2019).

81. Rheinbay, E. *et al.* Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes. en. *Nature* **578,** 102–111. ISSN: 1476-4687 (Feb. 2020).

82. Li, Y. *et al.* Patterns of Somatic Structural Variation in Human Cancer Genomes. en. *Nature* **578,** 112–121. ISSN: 1476-4687 (Feb. 2020).

83. Vinagre, J. *et al.* Frequency of TERT Promoter Mutations in Human Cancers. en. *Nature Communications* **4,** 2185. ISSN: 2041-1723 (Oct. 2013).

84. Stern, J. L., Theodorescu, D., Vogelstein, B., Papadopoulos, N. & Cech, T. R. Mutation of the *TERT* Promoter, Switch to Active Chromatin, and Monoallelic *TERT* Expression in Multiple Cancers. en. *Genes & Development* **29,** 2219–2224. ISSN: 0890-9369, 1549-5477 (Nov. 2015).

85. Zhou, S. *et al.* Noncoding Mutations Target Cis -Regulatory Elements of the FOXA1 Plexus in Prostate Cancer. en. *Nature Communications* **11,** 1–13. ISSN: 2041-1723 (Jan. 2020).

86. Jeselsohn, R., Buchwalter, G., De Angelis, C., Brown, M. & Schiff, R. ESR1 Mutations—a Mechanism for Acquired Endocrine Resistance in Breast Cancer. en. *Nature Reviews Clinical Oncology* **12,** 573–583. ISSN: 1759-4774, 1759-4782 (Oct. 2015).

87. Robinson, J. L. L. & Carroll, J. S. FoxA1 Is a Key Mediator of Hormonal Response in Breast and Prostate Cancer. *Frontiers in Endocrinology* **3.** ISSN: 1664-2392 (2012).

88. Fu, X. *et al.* FOXA1 Overexpression Mediates Endocrine Resistance by Altering the ER Transcriptome and IL-8 Expression in ER-Positive Breast Cancer. en. *Proceedings of the National Academy of Sciences* **113,** E6600–E6609. ISSN: 0027-8424, 1091-6490 (Oct. 2016).

89. Fu, X. *et al.* FOXA1 Upregulation Promotes Enhancer and Transcriptional Reprogramming in Endocrine-Resistant Breast Cancer. en. *Proceedings of the National Academy of Sciences* **116,** 26823–26834. ISSN: 0027-8424, 1091-6490 (Dec. 2019).

90. Maurano, M. T. *et al.* Large-Scale Identification of Sequence Variants Influencing Human Transcription Factor Occupancy in Vivo. en. *Nature Genetics* **47,** 1393–1401. ISSN: 1061-4036, 1546-1718 (Dec. 2015).

91. Guo, Y. A. *et al.* Mutation Hotspots at CTCF Binding Sites Coupled to Chromosomal Instability in Gastrointestinal Cancers. en. *Nature Communications* **9,** 1520. ISSN: 2041-1723 (Dec. 2018).

92. Dixon, J. R. *et al.* Integrative Detection and Analysis of Structural Variation in Cancer Genomes. En. *Nature Genetics* **50,** 1388. ISSN: 1546-1718 (Oct. 2018).

93. Akdemir, K. C. *et al.* Disruption of Chromatin Folding Domains by Somatic Genomic Rearrangements in Human Cancer. en. *Nature Genetics,* 1–12. ISSN: 1546-1718 (Feb. 2020).

94. Hnisz, D. *et al.* Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods. en. *Science* **351,** 1454–1458. ISSN: 0036-8075, 1095-9203 (Mar. 2016).

95. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. en. *Cell* **161,** 1012–1025. ISSN: 0092-8674 (May 2015).

96. Allou, L. *et al.* Non-Coding Deletions Identify Maenli lncRNA as a Limb-Specific En1 Regulator. en. *Nature.* ISSN: 0028-0836, 1476-4687 (Feb. 2021).

97. Iyyanki, T. *et al.* Subtype-Associated Epigenomic Landscape and 3D Genome Structure in Bladder Cancer. en. *Genome Biology* **22,** 105. ISSN: 1474-760X (Dec. 2021).

98. Rosen, P. *et al.* Clinical Potential of the ERG Oncoprotein in Prostate Cancer. en. *Nature Reviews Urology* **9,** 131–137. ISSN: 1759-4812, 1759-4820 (Mar. 2012).

99. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. en. *Cell* **153,** 666–677. ISSN: 00928674 (Apr. 2013).

100. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. en. *Science* **326,** 289–293. ISSN: 0036-8075, 1095-9203 (Oct. 2009).

101. Diaz, N. *et al.* Chromatin Conformation Analysis of Primary Patient Tissue Using a Low Input Hi-C Method. en. *bioRxiv,* 372789 (July 2018).

102. Chen, S. *et al.* Widespread and Functional RNA Circularization in Localized Prostate Cancer. en. *Cell* **176,** 831–843.e22. ISSN: 00928674 (Feb. 2019).

103. 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).

104. Johnstone, S. E. *et al.* Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer. en. *Cell* **182,** 1474–1489.e23. ISSN: 00928674 (Sept. 2020).

105. Ho, S. S., Urban, A. E. & Mills, R. E. Structural Variation in the Sequencing Era. en. *Nature Reviews Genetics,* 1–19. ISSN: 1471-0064 (Nov. 2019).

106. Berger, M. F. *et al.* The Genomic Complexity of Primary Human Prostate Cancer. en. *Nature* **470,** 214–220. ISSN: 1476-4687 (Feb. 2011).

107. Gasperini, M. *et al.* A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. en. *Cell* **176,** 377–390.e19. ISSN: 00928674 (Jan. 2019).

108. Tomlins, S. A. *et al.* Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. en. *Science* **310,** 644–648. ISSN: 0036-8075, 1095-9203 (Oct. 2005).

109. Tomlins, S. A. *et al.* Distinct Classes of Chromosomal Rearrangements Create Oncogenic ETS Gene Fusions in Prostate Cancer. en. *Nature* **448,** 595–599. ISSN: 0028-0836, 1476-4687 (Aug. 2007).

110. Wu, S. *et al.* Circular ecDNA Promotes Accessible Chromatin and High Oncogene Expression. en. *Nature* **575,** 699–703. ISSN: 0028-0836, 1476-4687 (Nov. 2019).

111. Kumar, P. *et al.* ATAC-Seq Identifies Thousands of Extrachromosomal Circular DNA in Cancer and Cell Lines. en. *Science Advances* **6,** eaba2489. ISSN: 2375-2548 (May 2020).

112. Morton, A. R. *et al.* Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. en. *Cell* **179,** 1330–1341.e13. ISSN: 00928674 (Nov. 2019).

113. Shoshani, O. *et al.* Chromothripsis Drives the Evolution of Gene Amplification in Cancer. en. *Nature* **591,** 137–141. ISSN: 0028-0836, 1476-4687 (Mar. 2021).

114. Katainen, R. *et al.* CTCF/Cohesin-Binding Sites Are Frequently Mutated in Cancer. *Nature Genetics* **advance on,** 818–21. ISSN: 1061-4036 (2015).

115. Zhao, S. G. *et al.* The DNA Methylation Landscape of Advanced Prostate Cancer. en. *Nature Genetics* **52,** 778–789. ISSN: 1061-4036, 1546-1718 (Aug. 2020).

116. Yu, Y. P. *et al.* Whole-Genome Methylation Sequencing Reveals Distinct Impact of Differential Methylations on Gene Transcription in Prostate Cancer. en. *The American Journal of Pathology* **183,** 1960–1970. ISSN: 00029440 (Dec. 2013).

117. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. en. *Cell Reports* **12,** 1184–1195. ISSN: 22111247 (Aug. 2015).

118. Taberlay, P. C. *et al.* Three-Dimensional Disorganization of the Cancer Genome Occurs Coincident with Long-Range Genetic and Epigenetic Alterations. en. *Genome Research* **26,** 719–731. ISSN: 1088-9051, 1549-5469 (June 2016).

119. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. English. *Cell* **159,** 1665–1680. ISSN: 0092-8674, 1097-4172 (Dec. 2014).

120. Ing-Simmons, E. *et al.* Independence of Chromatin Conformation and Gene Regulation during Drosophila Dorsoventral Patterning. en. *Nature Genetics* **53,** 487–499. ISSN: 1061-4036, 1546-1718 (Apr. 2021).

121. Ghavi-Helm, Y. *et al.* Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression. En. *Nature Genetics,* 1. ISSN: 1546-1718 (July 2019).

122. Gasperini, M., Tome, J. M. & Shendure, J. Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers. en. *Nature Reviews Genetics,* 1–19. ISSN: 1471-0064 (Jan. 2020).

123. Nasser, J. *et al.* Genome-Wide Enhancer Maps Link Risk Variants to Disease Genes. en. *Nature.* ISSN: 0028-0836, 1476-4687 (Apr. 2021).

124. Hanahan, D. & Weinberg, R. A. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674. ISSN: 1097-4172 (Electronic)\r0092-8674 (Linking) (Mar. 2011).

125. Wingett, S. W. *et al.* HiCUP: Pipeline for Mapping and Processing Hi-C Data. en. *F1000Research* **4,** 1310. ISSN: 2046-1402 (Nov. 2015).

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

127. 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).

128. Goloborodko, A., Abdennur, N., Venev, S., Hbbrandao & Gfudenberg. *Mirnylab/Pairtools: V0.2.2* Zenodo. Jan. 2019.

129. Abdennur, N. *et al. Mirnylab/Cooler: V0.8.5* Zenodo. Apr. 2019.

130. Imakaev, M. *et al.* Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization. en. *Nature Methods* **9,** 999–1003. ISSN: 1548-7105 (Oct. 2012).

131. Venev, S. *et al. Mirnylab/Cooltools: V0.3.2* Zenodo. May 2020.

132. Shin, H. *et al.* TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes. en. *Nucleic Acids Research* **44,** e70–e70. ISSN: 0305-1048 (Apr. 2016).

133. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of Alternative Topological Domains in Chromatin. en. *Algorithms for Molecular Biology* **9,** 14. ISSN: 1748-7188 (2014).

134. Forcato, M. *et al.* Comparison of Computational Methods for Hi-C Data Analysis. en. *Nature Methods* **14,** 679–685. ISSN: 1548-7091, 1548-7105 (July 2017).

135. Zaborowski, R. & Wilczyński, B. BPscore: An Effective Metric for Meaningful Comparisons of Structural Chromosome Segmentations. en. *Journal of Computational Biology* **26,** 305–314. ISSN: 1557-8666 (Apr. 2019).

136. Abdennur, N., Goloborodko, A., Abraham, S. & Aafkevandenberg. *Mirnylab/Bioframe: V0.0.12-Doi* Zenodo. June 2020.

137. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: Multi-Scale Detection of Chromatin Loops from Hi-C and Micro-C Maps Using Scale-Space Representation. en. *Genome Biology* **21,** 256. ISSN: 1474-760X (Dec. 2020).

138. Kamada, T. & Kawai, S. An Algorithm for Drawing General Undirected Graphs. en. *Information Processing Letters* **31,** 7–15. ISSN: 00200190 (Apr. 1989).

139. Hagberg, A. A., Schult, D. A. & Swart, P. J. *Exploring Network Structure, Dynamics, and Function Using NetworkX* in *Proceedings of the 7th Python in Science Conference* (Pasadena, California, USA, 2008), 11–15.

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

141. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. En. *Scientific Reports* **9,** 9354. ISSN: 2045-2322 (June 2019).

142. Stark, R. & Brown, G. DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data. en, 34.

143. 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).

144. Lawrence, M., Gentleman, R. & Carey, V. Rtracklayer: An R Package for Interfacing with Genome Browsers. en. *Bioinformatics* **25,** 1841–1842. ISSN: 1367-4803, 1460-2059 (July 2009).

145. Gel, B. *et al.* regioneR: An R/Bioconductor Package for the Association Analysis of Genomic Regions Based on Permutation Tests. en. *Bioinformatics,* btv562. ISSN: 1367-4803, 1460-2059 (Sept. 2015).

146. 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).

147. 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).

148. 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).

149. Gierliński, M. *et al.* Statistical Models for RNA-Seq Data Derived from a Two-Condition 48-Replicate Experiment. en. *Bioinformatics* **31,** 3625–3630. ISSN: 1367-4803, 1460-2059 (Nov. 2015).

150. Stein, C. *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution* en. in *Contribution to the Theory of Statistics* **3** (University of California Press, Berkeley, California, USA, Dec. 1956), 197–206. ISBN: 978-0-520-31388-0.

151. 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).

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

153. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2013.