

CHROMATIN ARCHITECTURE ABERRATIONS CONTRIBUTE TO PROSTATE CANCER
ONCOGENESIS AND ACUTE LYMPHOBLASTIC LEUKEMIA RELAPSE

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

Chapter 1

Introduction

Cancer is one of the largest causes of death worldwide, ranking in the top ten most frequent causes in over 150 countries and most frequent in over 40 [1]. Disease treatment is complicated by the fact that cancers are a myriad of diseases with unique origins, symptoms, and treatment options, often related to the cell of origin [2]. However, numerous hallmarks of cancers have emerged over the last 50 years to provide understanding about what biological aberrations cause tumours to initiate, how they develop over time, and how they respond to therapeutic interventions [3–6] (Figure 1.1).

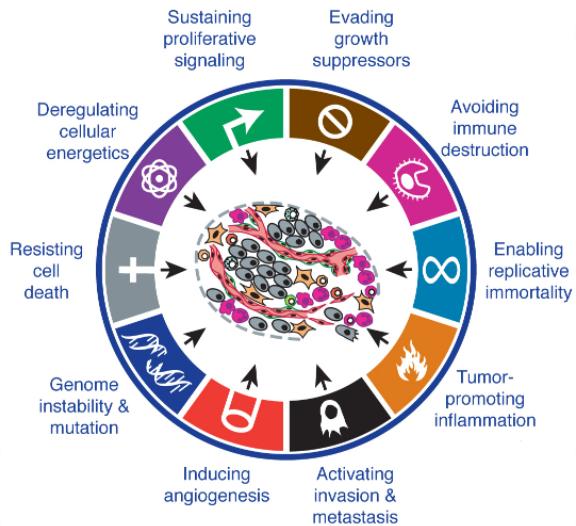


Figure 1.1: **The hallmarks of cancer.** Adapted from [REF 4].

Many of these hallmarks of cancer can be achieved through aberrations to the genome and the molecular machinery that enables cells to function normally [7]. For example, genome instability can be achieved by inhibiting deoxyribonucleic acid (DNA) repair machinery, as is observed with

abnormalities in *MLH1* and *MSH2* repair genes in colorectal cancers [8] or mutations to *BRCA1*, *BRCA2*, and *ATM* genes in prostate cancer (PCa) [9]. Similarly, replicative immortality can be achieved through telomere elongation by over-expression of the *TERT* gene [10]. Mutations to the *TERT* promoter, resulting in its over-expression, were first identified in melanomas [11, 12], but have since been further identified in bladder, thyroid, and brain cancers [10, 13, 14]. But while cancer has long been viewed as a disease of the genome [3, 7], there are many avenues cells can take to arrive these hallmarks resulting from aberrations of how genes are expressed inside the cell nucleus.

1.1 Normal chromatin architecture in mammalian cells

Genes, encoded as DNA, are expressed by being transcribed into ribonucleic acid (RNA) and subsequently translated into proteins in the process known as the Central Dogma of molecular biology [15] (Figure 1.2a). The transcription of genes into messenger RNA (mRNA) requires RNA polymerase to bind at transcription start sites (TSSs) within DNA elements found at the beginning of genes, termed promoters [16]. Promoters are one example of a class of DNA elements, termed *cis*-regulatory elements (CREs) because of their roles in regulating the expression of genes on the same strand of DNA. The recruitment of RNA polymerase is aided by a special class of proteins, termed transcription factors (TFs), that can bind at DNA sequences either close to a gene's promoter, or far from it at other CREs such as enhancers and insulators [17–22] (Figure 1.2b). Together, the binding of TFs to the DNA at specific CREs is fundamental for initiating transcription and expressing genes.

1.1.1 DNA elements and features regulating transcription

The ability of TFs to bind at specific CREs is dependent on multiple features of the DNA. Many TFs bind to DNA at specific sequences, termed motifs [18, 23]. Finding the locations of a given motif in the genome is often the first step in determining the cistrome of a TF, the set of all sites and CREs a TF binds to *in vivo* [24, 25]. The structural protein CCCTC-binding factor (CTCF) has a well-defined motif and binds to this sequence at thousands of locations across the human genome [26, 27]. Mutations to the sequence motif can alter CTCF's binding affinity for DNA, as is the case with many TFs [28–30]. Relying on more than just the genetic sequence, CTCF is also an example of a TF that is sensitive epigenetic features such as DNA methylation (DNAm), the addition of a methyl group to DNA nucleotides [31–35], as are DNA methyltransferases DNMT1, DNMT3A,

and DNMT3B [36, 37]. TF binding to DNA can also be affected by the presence of other proteins at binding sites. TFs can bind in a combinatorial manner at the same location [18, 19, 23] or be blocked from binding altogether by the presence of nucleosomes, protein complexes that DNA winds around to make it compact in three-dimensional space [38, 39]. The collection of DNA, nucleosomes, DNA-bound transcription factors, and chemical modifications is defined as the chromatin, and the presence and density of nucleosomes, as well as DNA coiling, make certain segments of the chromatin more or less accessible for TF binding (euchromatin and heterochromatin, respectively). This can affect normal cellular behaviour such as cell-type-specific gene expression [40, 41] and DNA damage repair in inaccessible regions [42]. Thus, both genetic and epigenetic chromatin features affect how TFs can bind and regulate transcription.

In addition to TF binding, transcription regulation depends on the ability of CREs to localize together in three-dimensional space across large genomic distances [43–45] (Figure 1.2c). Localization of CREs tens to thousands of basepairs (bps) apart from focal interactions is aided by the formation of topologically associated domains (TADs), domains of chromatin whose boundaries are linked by structural proteins, including CTCF and cohesin [22, 46–48]. In addition to TADs which can range in size from $10^4 - 10^6$ bp, chromatin is also organized into active or inactive compartments (A and B compartments, respectively) that range in size from $10^5 - 10^6$ bp [22, 49–51]. These two modes of chromatin organization facilitate the proper localization of CREs and TFs at the right time. While TADs and compartments are largely conserved across cell types [27, 52, 53], focal chromatin interactions can differ up to 45 % between cell types, providing a further mechanism to change chromatin state [50]. Different chromatin states enable cells with the same DNA sequence to express genes differently [17, 19, 46, 54–56], and thus identifying the repertoire of CREs, chromatin interactions, TADs, and compartments are vital in determining the regulation of genes in various cell types.

1.1.2 Methods for identifying DNA elements and chromatin interactions

High throughput sequencing protocols have enabled the characterization of functional elements from across the genome and rely on a similar concept to do so. This concept is to take a molecular feature of interest, be it an RNA transcript or nucleosome position, associate it with a short fragment of DNA, sequence these DNA fragments, and map it to the reference genome to identify where the original molecules came from (Figure 1.3). RNA sequencing (RNA-seq) methods reverse transcribed RNA into DNA that map back to individual genes, with the abundance of fragments indicating how much the gene is expressed [57]. Protein binding sites and histone post-translational modifications

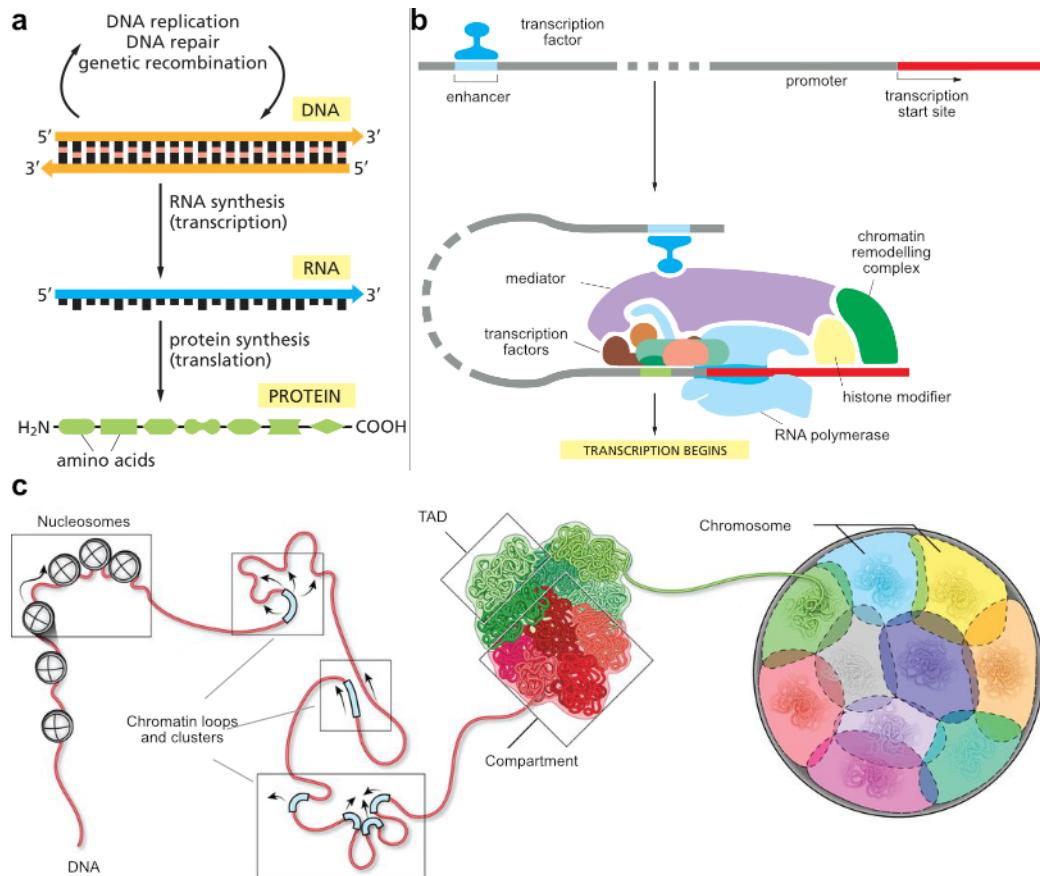


Figure 1.2: **The basics of gene expression inside the nucleus.** **a.** The central dogma of molecular biology. Adapted from [REF 15]. **b.** Schematic of the transcription machinery to initiate transcription. Adapted from [REF 15]. **c.** The scale of chromatin interactions across length scales. Adapted from [REF 48].

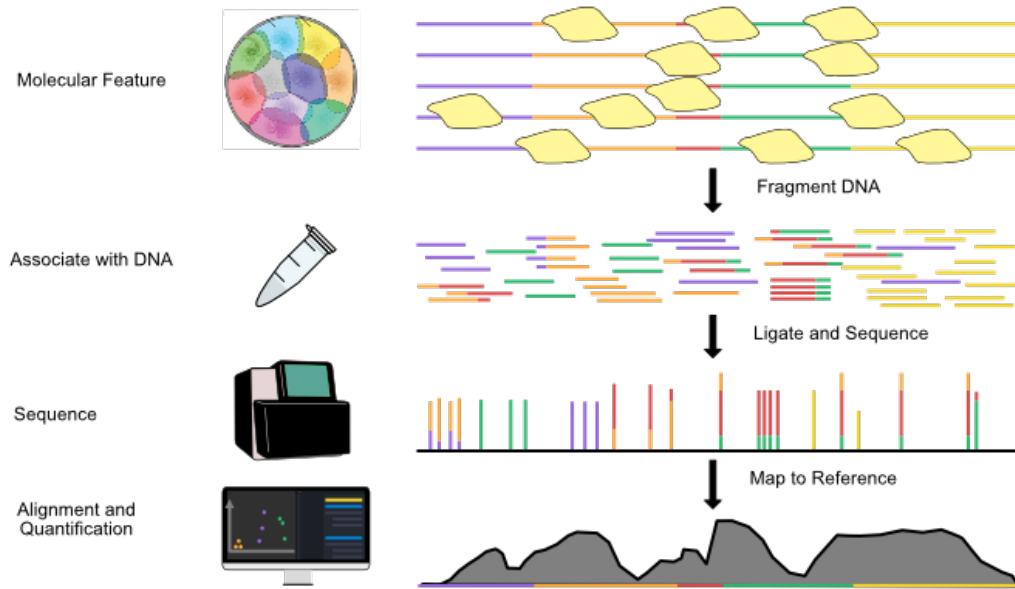


Figure 1.3: Characterizing functional DNA elements with high throughput sequencing.

can be identified by fragmenting DNA around antibodies that bind to these proteins with techniques like chromatin immunoprecipitation sequencing (ChIP-seq) and cleavage under targets and release using nuclease (CUT&RUN) [58–60]. Accessible and inaccessible chromatin can be assessed by the chromatin’s propensity to be cut by enzymes like DNase I, Tn5 transposase, and micrococcal nuclease in DNase I hypersensitive sequencing (DNase-seq), assay for transposase-accessible chromatin sequencing (ATAC-seq), and microccocal nuclease sequencing (MNase-seq) protocols, respectively [61–65]. DNAmc can be measured with bisulfite-sequencing assays [66], and distal chromatin interactions can be identified with chromatin conformation capture (3C) and 3C-based methods such as Hi-C [27, 49, 50, 67, 68]. Yet while these measurements help in identifying candidate CREs and important regions of the genome, determining their function and which target genes they regulate is a further complicating problem.

Varying chromatin states across cell types means that multiple measurements across multiple cell types are necessary to understand the breadth of functions a single CRE may have. In 2007, the ENCODE Project aimed to catalogue all biochemically functional elements in the human genome to better understand all the ways genes are expressed and how they are regulated in different cell types [69, 70]. Using these genome-wide sequencing techniques across a variety of human cell lines and tissues, the ENCODE Project has since catalogued nearly 10^6 candidate CREs, comprising nearly 8 % of the human genome [70]. Interpreting this data requires computational methods to correlate and interpret measurements across samples. Genome segmentation methods such as ChromHMM

[71] and Segway [72, 73] classify genomic regions according to their predicted function which can be validated with *in vitro* or *in vivo* experiments. Many techniques for experimental validation, including clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9, small interfering RNA (siRNA), and small hairpin RNA (shRNA), can interfere with candidate CREs by deleting them from the genome, preventing TFs from binding to the chromatin, or preventing translation of mRNA transcripts into proteins [74, 75]. These same techniques can also be used to screen for candidate CREs themselves, through massively-parallel reporter assays (MRPAs) and CRISPR screens [75], necessitating their own suite of statistical and software tools for analyzing observations. Altogether, a collection of experimental and computational techniques enable the cataloguing and interpretation of thousands of CREs and chromatin interactions across many cell types. These catalogues facilitate understanding how genes are expressed within the complex chromatin architecture in normal cells and, importantly, how aberrations to this architecture can result in disease.

1.2 Aberrations to chromatin architecture in cancer

1.2.1 Genetic aberrations in cancer

Discovery of genetic mutations of oncogenes in tumours nearly 50 years ago spurred the widespread characterization of genetic aberrations in cancers [76–79]. These mutations occur within genic regions that code for proteins, but more than 98 % of somatic mutations acquired in tumours are found in non-coding regions [80]. Single nucleotide variants (SNVs), copy number variants (CNVs), and structural variants (SVs) are found throughout the genome, and interpreting the impact of these mutations on cancer is an active area of research [70, 79, 81, 82]. Analysis of recurrent somatic mutations in tumours led to the identification of *TP53* as a tumour suppressor gene [83], the frequently mutated *SPOP* gene to help define a molecular subtype of prostate tumours [84], and the interpretation of recurrent rearrangements of the proto-oncogene *MYC* in multiple cancers [85]. The impact of a mutation can also be predicted by identifying overlapping regulatory elements or TF binding sites [29, 86, 87]. Grouping CREs by their putative target genes led to the identification of the *ESR1* gene as having its gene regulatory network recurrently mutated in ~10 % breast cancers, resulting in its over-expression, despite the gene itself being mutated in ~1 % of breast cancers [88]. Similarly, the binding sites of the *FOXA1*, *HOXB13*, *AR*, and *SOX9* TFs are enriched with mutations affecting their binding affinities [89] and recurrent amplifications of enhancers near the *AR* and *FOXA1* genes are associated with increased rates of metastasis [90, 91]. Furthermore, mutations that do not

directly target gene bodies or CREs can lead to oncogene over-expression. Multiple non-coding SVs in pediatric medulloblastoma patients were found to bring the *GFI1* and *GFI1B* oncogenes proximal to enhancer clusters, causing the oncogenes to become aberrantly regulated by this enhancer cluster [92]. This mechanism of enhancer hijacking has also been observed in developmental diseases [93, 94]. While this is not an exhaustive list, it is clear that genetic aberrations are abundant in cancers and that integrating genetic information with other components of the chromatin architecture can help identify driver events that promote oncogenesis or aggressive disease.

Mutations to DNA methyltransferases and chromatin remodelling proteins are common in cancers, and the impact of these mutations can be observed in their chromatin state. The isocitrate dehydrogenase (*IDH*) enzymes *IDH1*, *IDH2*, and the ten-eleven translocation (*TET*) enzymes *TET1* and *TET2* are frequently mutated in cancers, most often in leukemias and gliomas [95–99]. These mutations often affect the DNAm profiles of tumours and differentiation programs [95], such as loss of enhancer hydroxymethylation and germinal centre hyperplasia in diffuse large B-cell lymphoma (DLBCL) [100]. Similarly, mutations to the *EZH2* gene in leukemias can affect the ability of the *EZH2* protein to write the H3K27me3 histone mark [101–104] and *EZH2* over-expression is associated with poor survival in PCa [105–108]. Together, these findings show that genetic aberrations to genes regulating other aspects of the chromatin architecture are abundant in multiple cancers and can drive specific programs in tumours. These programs can, in turn, affect progression of the disease and treatment strategies for patients. Importantly, the impact of these mutations is dependent on the function of the affected protein or CRE, which varies between different cancers. Thus, understanding how non-genetic aberrations affect tumours can be a vital step in understanding the impact of genetic aberrations.

1.2.2 Non-genetic aberrations in cancer

Non-genetic aberrations to chromatin have long been recognized as important factors in cancer development and progression [109, 110]. Methylation of gene promoters is associated with reduced gene expression and loss of DNAm (hypomethylation) across the genome and focal increases of DNAm (hypermethylation) have been found across numerous cancers [110, 111]. Importantly, these changes in DNAm can be found in the absence of mutations targeting DNA methyltransferases. Analysis of ~200 metastatic PCa patients with matching whole genome sequencing (WGS), RNA-seq, and whole genome bisulfite sequencing (WGBS) identified a subtype of tumours with a distinct DNAm profile [112]. Ependymomas have also been found to display distinct DNAm profiles in the

absence of recurrent mutations across patients [113] along with acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), glioblastoma, and colorectal, liver, pancreatic, and ovarian cancers [114]. Notably, treatment of cancer cells with demethylating agents such as 5-aza-cytidine and 5-aza-2'-deoxycytidine for use in patients with AML and myelodysplastic syndrome (MDS) have shown to significantly increase survival times, demonstrating the clinical relevance of epigenetic marks in treatment strategies [115–117]. Though many causal mechanisms relating DNAm to cancer phenotype are lacking, the impact of DNAm on TF binding has been well-demonstrated. Variable CTCF binding across human cell lines has been shown to vary with DNAm levels, which can affect genome organization [31, 32]. In gastrointestinal cancer, CTCF binding sites are hypermethylated *SDH*-deficient tumours, resulting in widespread loss of CTCF and increased contact between the *FGF3* and *FGF4* oncogenes and a nearby enhancer cluster [118]. Moreover, aberrant contact of *FGF3* and *FGF4* is concomitant with increased H3K27ac modifications, further demonstrating the increased regulation and expression of the oncogenes. Disruptions of CTCF binding sites at TADs boundaries, resulting in aberrant regulation has also been found in T-cell ALL, leading to over-expression of the *TAL1* and *LMO2* oncogenes [119]. Both of these cases mimic the enhancer hijacking mechanism without the need for nearby genetic mutations. Together, these results show the importance of DNAm on three-dimensional genome organization and TF binding, and genetic and non-genetic aberrations can be observed in chromatin contacts and histone modifications.

The effect of chromatin variants on gene regulation extends beyond DNAm. Cell type differences in nucleosome occupancy can lead to increased rates of mutation across the genome [120]. Similarly, TF binding can affect the ability of DNA damage repair complexes to perform local nucleotide excision repair [121, 122]. Thus, cell type differences in chromatin state can influence the frequency and location of DNA damage, which may describe some differences in recurrent mutations across cancer types. Many computational techniques have been developed in an attempt to prioritize the roles of different components of the chromatin architecture. One method, called similarity network fusion (SNF), integrates multiple chromatin measurements together to construct a mathematical graph whereby multiple samples cluster together if they share properties across multiple components [123]. Many similar methods exist that use machine learning-oriented and biology-oriented techniques to integrate multiple data types together to provide a comprehensive view of the chromatin architecture [124]. Taken together, these papers demonstrate the effect of differences in normal cell chromatin architecture on cancer and the multiple computational and experimental methods required to unravel these relationships.

Overall, these non-genetic aberrations of chromatin can be found across multiple cancer types.

But we will continue to focus on two seemingly different cancer types that both display complex relationships between different components of the chromatin architecture: PCa and B-cell acute lymphoblastic leukemia (B-ALL).

1.3 Chromatin architecture of prostate cancer and B-cell acute lymphoblastic leukemia

1.3.1 Prostate cancer

Diagnosis, treatment, and risk factors

PCa is the second most commonly diagnosed cancer in men globally, with an estimated 23 300 men being diagnosed with the disease in Canada in 2020 [1, 125]. Diagnosis typically begins with the detection of prostate-specific antigen (PSA) in the blood, followed by a digital rectal exam for an enlarged prostate and a core needle biopsy to rule out benign prostate hyperplasia [126]. Once diagnosed, patients are typically grouped into one of several risk categories based on factors including PSA levels, histopathological assessment (i.e. Gleason or International Society of Urological Pathology (ISUP) scores), and medical imaging to detect for distal metastases (tumour node metastasis (TNM) staging)[126]. PCa patients assessed to have a low mortality risk often undergo active surveillance to monitor for changes in the disease that pose a risk to the patient. Patients with high mortality risks often undergo one of multiple treatment regimens, including surgery, androgen deprivation therapy, chemotherapy, and radiotherapy [126]. While ~93 % of men with localized PCa survive, ~70 % of patients with metastatic disease will die within 5 years [127, 128], accounting for ~10 % of all cancer deaths in men [125]. This highlights the need for accurate risk assessment at diagnosis and knowledge of what aberrations lead to aggressive, metastatic disease.

Risk of developing PCa is associated with age and the median age at diagnosis is 66 years old [129]. While developing PCa at a young age is rare, younger men who are diagnosed typically have a more aggressive disease and relatively poorer survival rates [128]. In addition to age, genetic ancestry is a risk factors for developing the disease. Men of African ancestry are ~1.6 times more likely to be diagnosed with PCa than men of western European ancestry, who in turn are ~2 times more likely than men of Asian ancestry [128, 130, 131]. Men of different ancestries also tend to accumulate different sets of mutations in their tumours. For example, ~50 % of men of western European ancestry harbour a fusion of an *ETS* gene family member [132], whereas only ~10 % of

men of Asian ancestry harboured a similar mutation [133]. Inherited germline mutations are also a risk factor for PCa, as men with *BRCA1* and *BRCA2* mutations are ~2 times more likely to develop PCa than those without. Studies identifying these risks demonstrate that familial history, in addition to age and genetic ancestry, are important factors for developing PCa.

Chromatin aberrations in prostate cancer

Large cohort studies of prostate tumours have identified numerous driver mutations for the disease. These driver mutations include, but are not limited to, coding mutations to the *BRCA1*, *BRCA2*, *CHD1*, *IDH1*, *MYC*, *NKX3-1*, *PTEN*, *RB1*, *SPOP*, and *TP53* genes, as well as *ETS*, *FOX*, *HOX*, *KLK*, and *KMT* factors [9, 132, 134]. *ETS* factor mutations, such as the *TMPRSS2-ERG* (T2E) fusion, can lead to a globally *cis*-regulatory landscape, affecting TF binding genome-wide and *NOTCH* signalling [135]. Metastatic tumours are enriched for amplifications to the *FOXA1* and androgen receptor (*AR*) genes compared to primary tumours, as well as mutations targeting epigenetic regulators, such as histone lysine methyltransferases (KMTs) [90, 136, 137]. Over-expression of *AR* is associated with castration resistance, reducing the effectiveness of androgen deprivation therapies [90, 138]. Importantly, *FOXA1* is a pioneer TF that regulates *AR*, and over-expression of *FOXA1* is also more frequently found in metastatic than primary tumours [139]. Together, these two genes, their CREs, and their cistromes constitute important regions of chromatin that impact the progression of low-risk, localized PCa into high-risk metastatic PCa.

1.3.2 B-cell acute lymphoblastic leukemia

Diagnosis, treatment, and risk factors

Leukemia is the 15th most commonly diagnosed cancer globally, with an estimated 6 900 individuals being diagnosed with the disease in Canada in 2020 [1, 125]. Leukemias, generally, result from an overgrowth of undifferentiated blast cells that do not exhibit the same behaviours as fully differentiated cells in the hematopoietic hierarchy [90]. B-ALL is an acute clonal expansion of primitive cells restricted to the lymphoid hematopoietic lineage of B-cells and primarily occurs in children [140]. Currently, overall survival of pediatric B-ALL is ~90 % [140], yet disease relapse after treatment still occurs in 10 - 15 % of patients [141, 142]. Diagnosis of B-ALL typically begins with the detection of over-abundant lymphoblasts by microscopy and immunophenotypic assessment of cell surface markers indicating lineage commitment and developmental stage [141]. After diagnosis, mortality risk is assessed based on factors including age and white blood cell counts. Patients under 2 or

over 10 years of age have worse prognoses than patients of other ages, as do patients with $\geq 50 \times 10^3$ cells / mL [140, 141]. Newly diagnosed patients typically undergo remission-induction therapy, intensification/consolidation therapy, and continuation/maintenance therapy over the span of 2 years [141]. Risk factors for developing the disease include sex, genetic ancestry, and chromosomal rearrangements, with men, African or Hispanic ancestry, and Down's syndrome all associated with an increased risk [140, 141]. Risk factors for disease relapse remain elusive; however, karyotyping and high throughput sequencing technologies are helping to identify new biomarkers.

Chromatin aberrations in B-cell acute lymphoblastic leukemia

B-ALL is commonly classified according to the presence of recurrent mutations. Hyperploidy and the presence of the fusion of the *ETV6* and *RUNX1* genes are associated with favourable outcomes, whereas hypoploidy with < 44 chromosomes, fusion of the *BCR* and *ABL1* genes, and mutations affecting the *PAX5*, *EBF1*, *MLL/KMT2A*, *CRLF2*, and *IKZF1* genes are all associated with poorer outcomes [140, 141]. Many of these affected genes regulate B-cell development, such as *PAX5* [143–145], *IKZF1* [145], and *EBF1* [146, 147]. Similarly, *KMT2A* and *CREBBP* are histone writers, depositing methyl groups to the histone H3 lysine 4 residue and acetyl groups to the histone H3 lysine 56 residue, respectively [148–153]. Mutations in these genes are enriched in relapse [140, 154], suggesting that not only do epigenetic regulators play a key role in oncogenesis, but that they also promote relapse.

Aberrant changes to DNAme may also play a role in B-ALL relapse. DNAme has been shown to change across B-cell differentiation, with differentially methylated regions (DMRs) found in the cistromes of TFs that regulate differentiation, including *EBF1* and *PAX5* [155]. Additionally, the DNAme profile of B-ALL cells differ at thousands of loci across the genome, compared to normal B-cells, primarily in bivalent CREs and promoter regions [156, 157]. These findings suggest that aberrant DNAme pattern in B-ALL may be affecting B-cell differentiation through TF binding. Moreover, hypomethylation of the *IL2RA* gene is associated with a worse prognosis, as is aberrant DNAme in the presence of *E2A-PBX1* or *KMT2A* fusions [158]. This suggests that specific DNAme changes may cooperate with mutated epigenetic regulators to promote aggressive disease that is more likely to relapse after treatment. Overall, numerous genetic and epigenetic alterations in primary B-ALL and relapsed B-ALL suggest that multiple chromatin aberrations impact the development and progression of this disease.

While cellular phenotypes and treatment strategies for PCa and B-ALL do not resemble each other, PCa oncogenesis, PCa metastases, and B-ALL relapse all harbour aberrations to different

components of the chromatin architecture that interact with each other. Thus, to mitigate, or even prevent, these processes from occurring, this thesis investigates mutations targeting CREs of important TFs, the relationship between three-dimensional genome organization and SVs, and the effect of DNAm changes over the course of relapse.

1.4 Thesis structure

I begin with ?? by exploring the *cis*-regulatory landscape of PCa and delineating the CREs of the prostate oncogene *FOXA1*. I demonstrate the essentiality of *FOXA1* for prostate tumours, identify putative CREs based on integration of multiomic datasets in PCa cell lines, and assess the functional impact of recurrent PCa SNVs on *FOXA1* expression and TF binding.

With the *cis*-regulatory network of *FOXA1* established in PCa, I attempt to construct the *cis*-regulatory landscape genome-wide in PCa with 3C mapping in ?. Using Hi-C, I characterize the three-dimensional chromatin organization of PCa and investigate changes to this structure over oncogenesis, and explore the relationship between chromatin organization, SVs, and CRE hijacking.

In assessing the impact of SVs on chromatin organization, I uncovered a statistical problem stemming from the lack of recurrent SVs across PCa patients, leading to unbalanced experimental comparisons. To address this problem, I developed a statistical method for reducing error in gene expression fold-change estimates from unbalanced experimental designs in ?? and characterize the method.

Given the shared importance of mutations to TFs and epigenetic enzymes in prostate cancer and leukemias, in ?? I explore the epigenetic landscape of B-ALL and its relapse after treatment. I characterize molecular changes to B-ALL tumours over the course of disease relapse and identify important changes to DNAm that indicate the reversion to a stem-like phenotype, often present in a subpopulation of cells at diagnosis.

Together, this thesis investigates the multiple layers of the chromatin architecture that contribute to oncogenesis and cancer progression. I demonstrate that aberrations to the genome, epigenome, and three-dimensional organization of chromatin play important roles individually, and together, in the orchestration of the disease.

Chapter 2

Discussion & Future Directions

Each of the previous chapters have presented a story interrogating multiple components of the chromatin architecture, how they interact with each other, and the plethora of computational and experimental methods required to unravel this architecture. ?? identifies and validates CREs of the *FOXA1* gene, a critical TF that regulates PCa development and regulates *AR* expression to control disease progression. ?? expands on these ideas to investigate how the three-dimensional genome organization impacts gene regulatory networks and how genetic aberrations can alter this organization to promote oncogenesis. ?? develops a mathematical and computational framework to reduce uncertainty about how individual aberrations in chromatin architecture impact gene expression. Finally, ?? identifies the strong relationship between genetic and epigenetic profiles in B-ALL relapse and investigates how DNAm changes and revision to a more stem-like chromatin state may underlie disease relapse. Together, the work presented in this thesis demonstrates that different components of the chromatin architecture, the genome, molecular chromatin modifications, and three-dimensional organization, can all individually contribute to cancer development and progression. Moreover, this thesis demonstrates that aberrations in these components work together to drive disease. These multiple components of the chromatin architecture need to be studied in tandem to understand the origins of cancer and how to develop curative treatments for it.

2.1 Implications of non-coding single nucleotide variants targeting a single gene

In ??, I used gene essentiality screening data from multiple cell lines to prioritize the *FOXA1* TF as a critical factor across PCa cell lines. I also made use of the concept that SNVs converge on CREs of important genes in a given tumour type to predict how these mutations may impact candidate CREs for the *FOXA1* gene. *FOXA1* is also an important TF in breast cancers Figure A.3. Similar investigations into the impact of SNVs in breast tumours may identify the impact of aberrations to the CREs of *FOXA1*. Identifying important genes in this manner is not limited to *FOXA1* and breast and prostate tumours. Critical genes may be identified in other cancer types using CRISPR screens or MRPA. Similarly SNVs are not the only chromatin aberrations that can affect TF binding or gene regulation. Other chromatin aberrations may accumulate in CREs of important genes in a similar fashion. Complex SVs, changes in DNAm, or histone modifications may only need to accumulate in the set of CREs for a given gene, rather than be recurrent in a single element, to affect its expression. Interpreting chromatin aberrations in cancer in light of this plexus-based approach may aid in identifying driver events for cancer by aggregating previously unrelated events together. These approaches are not limited to prostate tumours and can serve as a starting point to identify important genes in other cancers, more generally.

2.2 Implications of three-dimensional organization and enhancer hijacking in prostate cancer

In ??, my co-authors optimized a low-input Hi-C method to interrogate genome organization in cryo-preserved prostate tissue slides. I then demonstrated that this could produce a high quality Hi-C library and helped produce that largest collection of genome organization data in prostate tumours to date. This technological step forward opens the door for profiling the three-dimensional genome in cancer patients without relying in cell lines or other models, and may be a critical step in moving personalized medicine forward. We add to existing evidence that SVs can, but rarely, alter 3D structure in disease [22, 159–165]. Elucidating when and how SVs impact genome organization, then, is still an area that requires investigation. Developments in statistical methods, such as those discussed in ??, may help identify the effects of individual, non-recurrent SVs. Subclonality of SVs may interfere with the ability to detect rearranged domains in bulk Hi-C measurements. Thus,

developments in high throughput sequencing and microscopy measurements in single cells, such as ORCA [166] and STORM [167], as well as organoid or explant models that recapitulate the chromatin state of the original tumour, may help in identifying the effect of such events [168]. This work also adds to our ability to detect chromatin interactions between promoters and enhancers in patient samples, allowing for better characterization of gene regulatory networks for each and every gene [22, 75, 169]. Given the benefits of plexus-based approaches to interpreting aberrations in the chromatin architecture, this work serves as a foundation on which to integrate gene regulatory networks with chromatin aberrations in cancers more generally. This foundation can be extended to studying the evolution of these networks, their genome organization, and their resiliency between species or over time as tumours respond to therapeutic interventions.

2.3 Implications of DNA methylation changes in relapse

?? identifies DNAm as an epigenetic marker that can mirror that mutational profile of B-ALL cells. The DNAm changes observed over the course of B-ALL relapse have the potential to become a biomarker predicting relapse, although variation in DNAm changes across patients necessitates larger sample sizes before recurrent events may be robustly identified. Our patient-oriented approach to identify recurrent changes to DNAm boosted our discovery of recurrent DMRs over a cohort-oriented approach, and thus may be a beneficial strategy for similar studies. Inter-tumour heterogeneity in tumours is a well-studied phenomenon [45, 170–175], so patient-oriented discovery approaches may be advantageous when assessing molecular trajectories of relapse and therapeutic response, as well. Changes in DNAm can be detected through blood draws [176–178] and offers the potential for a non-invasive biomarker to predict relapse in B-ALL patients. Given the widespread hypermethylation observed at relapse in all patients, it is possible that B-ALL patients undergoing continuation/maintenance therapy may additionally benefit from demethylating agents such as 5-aza-cytidine and 5-aza-2'-deoxycytidine to prevent relapse. Finally, the widespread hypermethylation of B-ALL at relapse reflects revision of malignant blasts to a more stem-like phenotype, a characteristic typically observed in other leukemias such as AML [149, 179–182]. This suggests that the role of leukemic stem cells and the impact that the cell-of-origin has on therapeutic response, should be prioritized in future research of B-ALL relapse.

2.4 Implications for functional genomics and cancer patients

The studies presented here highlight the importance of non-genetic properties of chromatin and their role in cancer development and progression. I used multiple sequencing-based assays to identify chromatin elements and structures, such as CREs and TADs, associate these elements and structures with potential target genes, and measure how these elements and structures affect gene expression when they are perturbed. This has multiple implications for how functional genomics and cancer research can be performed. Firstly, considering the entire regulatory plexus of genes may help pinpoint drivers of oncogenesis or cancer progression by offering alternative methods to identify candidate drivers. Similarly, viewing CRE hijacking as a common mechanism of aberrant gene expression may simultaneously increase the number of potential driver events found and offer a biological mechanism of action. These two perspectives may help resolve previously identified risk SNVs with no known mechanism of action. Considering the entire regulatory plexus as a functional unit of chromatin can also have therapeutic implications. Genes important for cancer progression, such as *FOXA1* and *AR* in PCa, have multiple regulatory elements. Each element is a potential option for targeted therapies, and combination therapies targeting multiple elements may offer more robust and controllable options than targeting the gene alone or an individual element. Developing targeted therapies for the multiple validated CREs discovered in ?? may help control or prevent castration resistance in PCa. Regarding the regulatory plexus as a functional unit may also help predict the evolutionary dynamics that tumours undergo during treatment. Therapeutic resistance via tumour evolution remains a clinical concern, and much study of evolutionary dynamics over cancer progression has focused on identifying genetic subclonal populations of tumour initiating cells [183, 184]. The evolution of gene regulatory networks have long been studied in model organisms such as *Arabidopsis thaliana* [185] and *Drosophila melanogaster* [186]. Bridging these fields of research together by including CREs, epigenetic marks, and genome organization as important players in the fitness and evolution of tumours may help inform how to prevent or counteract therapeutic resistance.

Secondly, I have shown that valuable information about tumour regulatory networks cannot be captured by sequencing DNA alone. Assaying more than solely the DNA from cancer patients is required to understand the origin and potential trajectories of their tumours. Studying these regulatory networks in cancer patients requires robust characterization using multiple genome-wide assays and validation experiments. Given the differences observed between cell line models and primary patient samples observed in ??, as well as the inter-patient heterogeneity observed in ??, it is vi-

tal that these networks are constructed on a per-patient basis. Currently, chromatin-based assays are not prioritized for clinical samples in the same manner as genome-based assays are. Further, profiling multiple facets of the chromatin architecture from primary patient samples requires technological development in genome-wide assays. Optimizing for small amounts of chromatin obtained from samples that are fresh or stored as flash-frozen paraffin embedded or cryo-preserved material are clear next steps to more comprehensively characterize the chromatin of tumours. In ??, my colleagues and I pushed technological limits in this area in the case of the Hi-C assay. If more comprehensive characterization of per-patient chromatin architecture is pursued, adopting patient-specific discovery strategies such as those presented in ???? may help guide personalized therapeutic strategies. Alternatively, developing patient-derived models that recapitulate the entire chromatin architecture, as mentioned previously, can help test and validate potential therapies for each patient, such as targeted DNAm experiments proposed in ??.

2.5 Limitations

The main limitations in the studies presented here stem from cancer models, cohort sizes, and correlative measurements. In ??, we used essentiality data from PCa cell lines that serve as models of primary prostate tumours. Further, our dissection of CREs relied on encyclopedias of DNA elements derived from cell lines, as did our validation experiments for assessing the impact of mutations seen in primary prostate tumours. Cell line models are not necessarily representative of primary tumours observed in patients [187, 188]. Moreover, different cell lines may have different regulatory networks, as recent studies of chromatin organization and CRE contacts in PCa cell lines have highlighted [189]. This separation between the models we study and the patients we aim to serve may prohibit translation of discoveries from our models to the clinic. In ??, we were able to profile the genome organization of 12 PCa patients and 5 benign prostate samples. Similarly, in ??, we used patient-derived xenografts (PDXs) originating from 5 patients with B-ALL. In light of the large degree of inter-patient variation seen in tumours from the same tissue, and the intra-patient variation seen in PDXs (??), observations in these small sample sizes may not generalize to patients with these diseases as a whole. In both cases, we used patient-matched data to corroborate findings from other components of the chromatin architecture, where possible, but validation experiments for some observations remain difficult. Using CRISPR technologies to knock in SNVs or inactivate individual CREs is possible, but replicating complex SVs or selectively (de)methylating CREs to validate some findings remains technically challenging [169, 190, 191]. Finally, despite the small sample sizes in

some of these studies, the amount of high throughput sequencing data generated was larger than many computational methods could handle efficiently, particularly with Hi-C data. Some methods used here had to be adapted from their original publications, and other methods that could have been used would have required extensive re-engineering to function properly, limiting some analyses.

2.6 Summary and concluding remarks

Diagnosis, treatment, and the foundational understanding of cancer has been revolutionized by high throughput sequencing technologies and the ability to detect and interpret genetic aberrations in tumours. When combined with information about CREs, the essentiality of genes in multiple cell types, and the three-dimensional genome organization, genetic aberrations can provide a significant amount of explanatory power for the hallmarks of cancer and what causes them. But genetic aberrations are not the only mechanism cancer cells can use to arrive at these hallmarks to drive disease. Identifying aberrations across multiple components of the chromatin architecture can uncover complex mechanisms through which cancer cells turn off the expression of tumour suppressor genes, activate oncogenes, or activate pathways. Interrogating the role of mutations in a gene's set of CREs (??), identifying stable regions of genome organization and shifting chromatin interactions over oncogenesis (??), and comparing genetic and epigenetic states over disease progression (??) are all important approaches to better understand the aberrations, origins, and trajectories of cancers in patients. This multi-pronged approach requires computational, statistical, and molecular, methods, optimized for low-input samples and small cohort sizes (????), and methodological developments in these areas are still required. This thesis focuses on PCa and B-ALL, but the methodologies employed here are not restricted to a single disease. In summary, measuring multiple facets of the chromatin architecture directly from patients and viewing aberrations in light of the regulatory networks that this architecture describes, we can better understand how cancer arises and develop better, more targeted therapies for patients.

Appendix A

Supplementary Material for Chapter 2

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

Table A.2 guide RNA (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

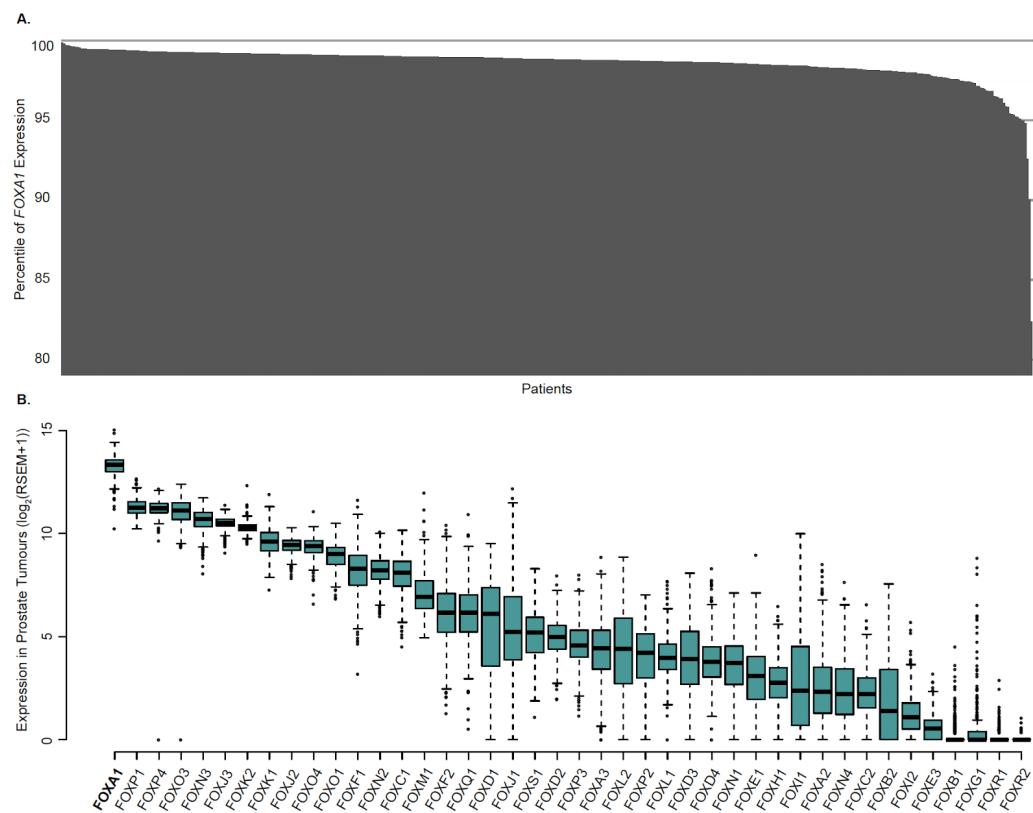


Figure A.1: ***FOXA1* mRNA expression in prostate tumours.** **a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

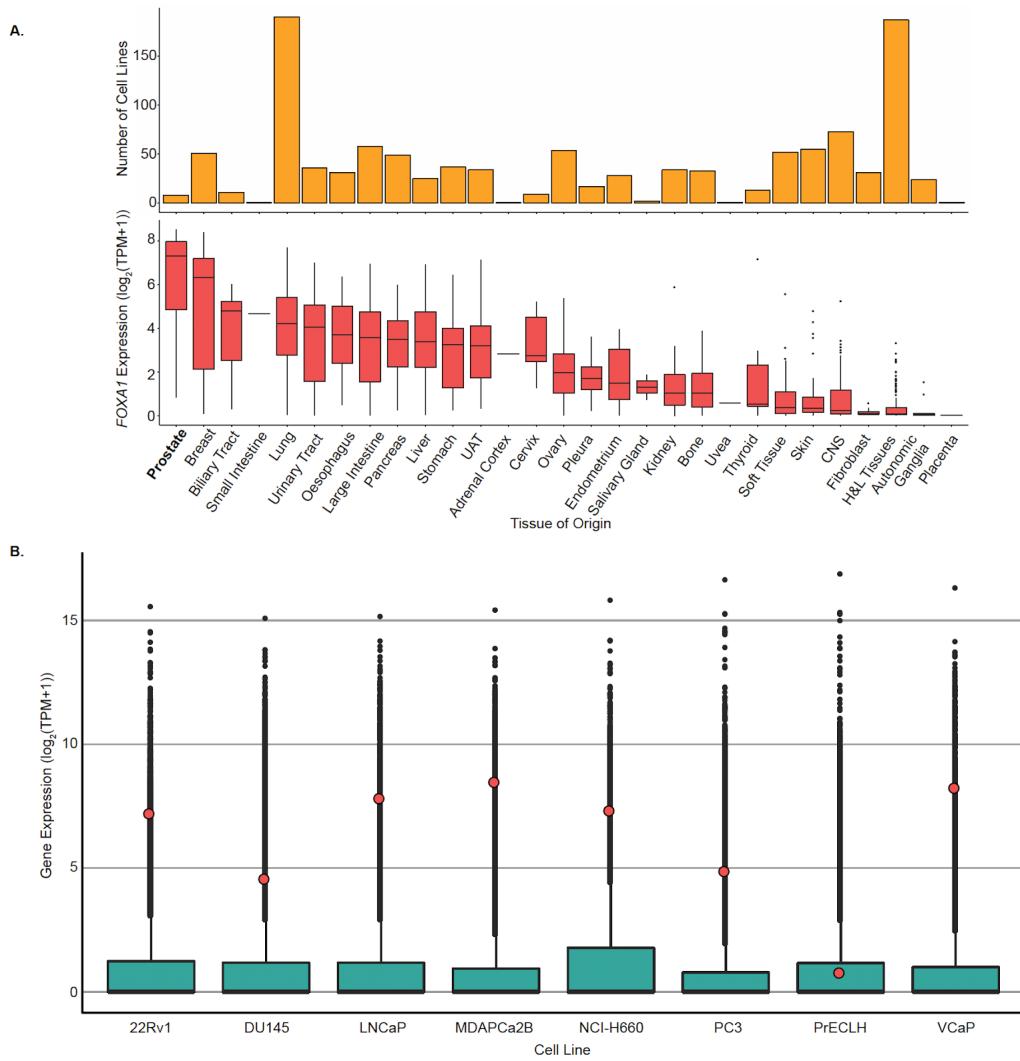


Figure A.2: **FOXA1** mRNA expression across PCa cell lines. **a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Methods). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA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

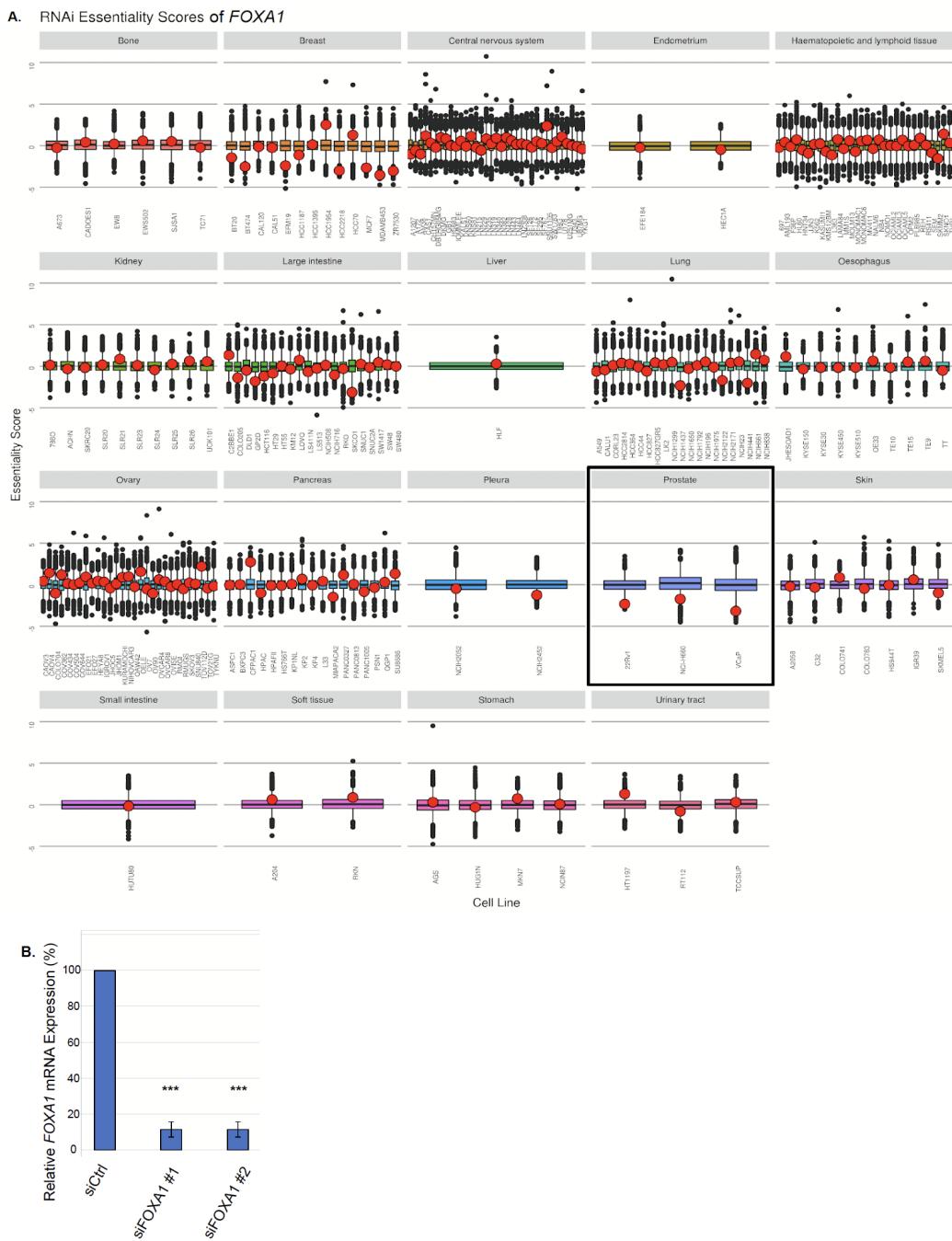


Figure A.3: Essentiality of *FOXA1* across cancer cell lines of various cancer types. **a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot 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 \pm s.d., Student's *t*-test, *** $p < 0.001$.

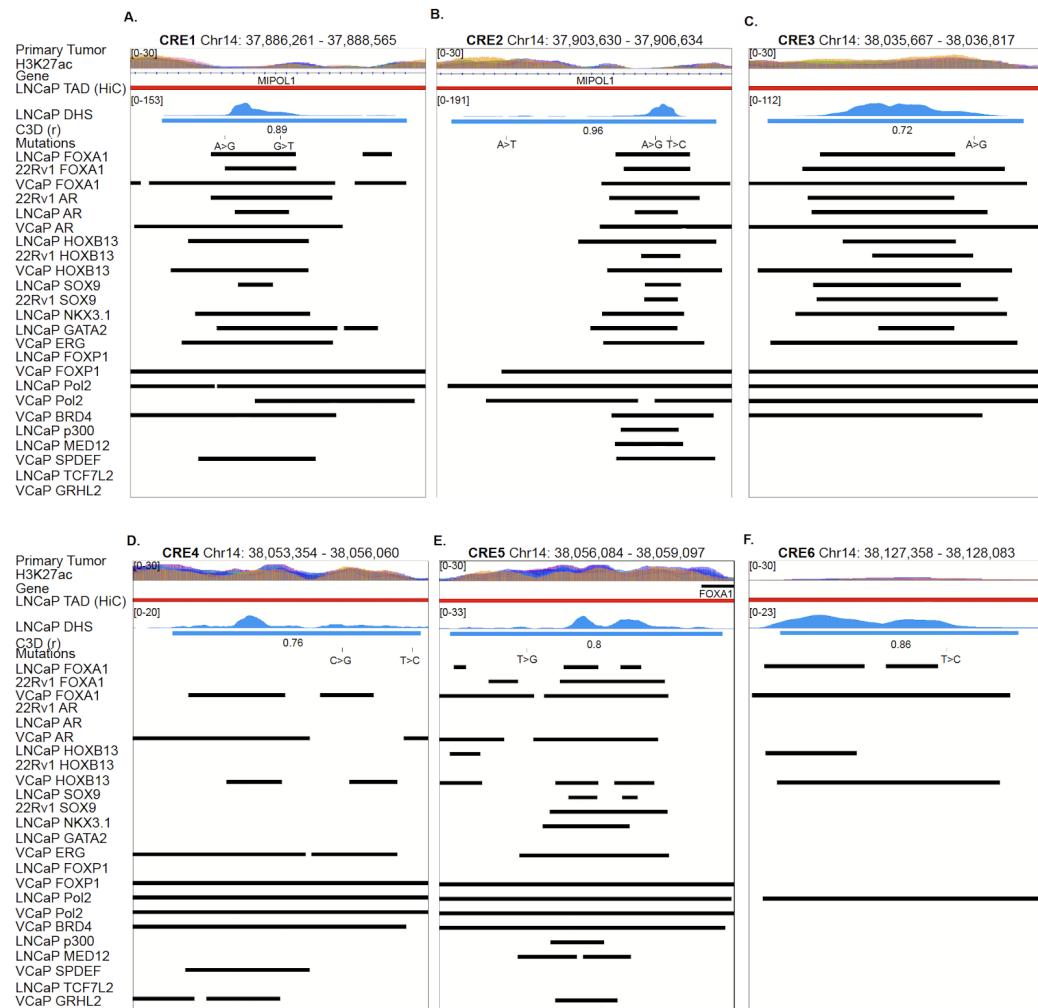


Figure A.4: Visualization of the functional annotation of the six *FOXA1* CREs. a-f. Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

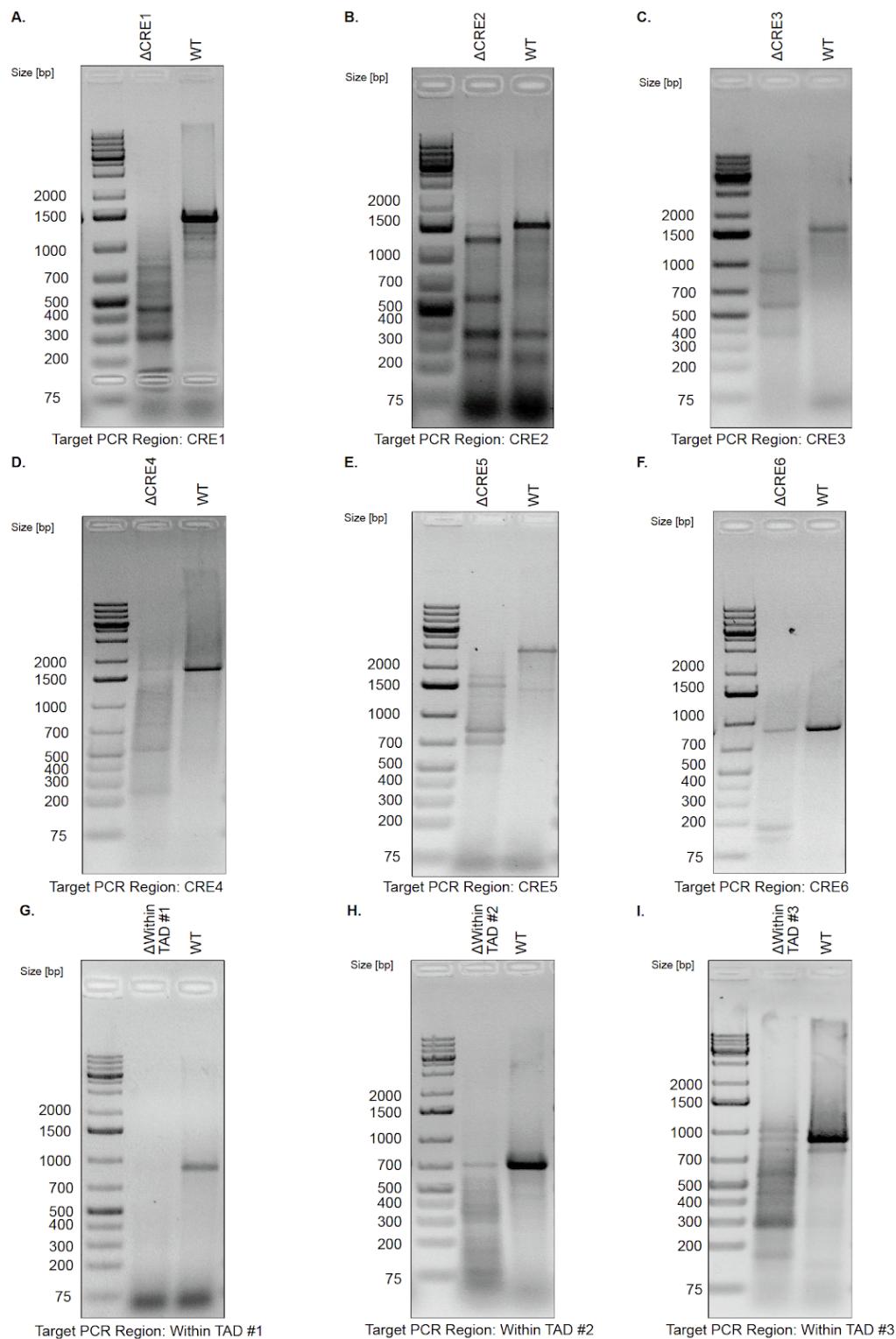


Figure A.5: Validation of clonal Cas-mediated deletions of CREs. a-f. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

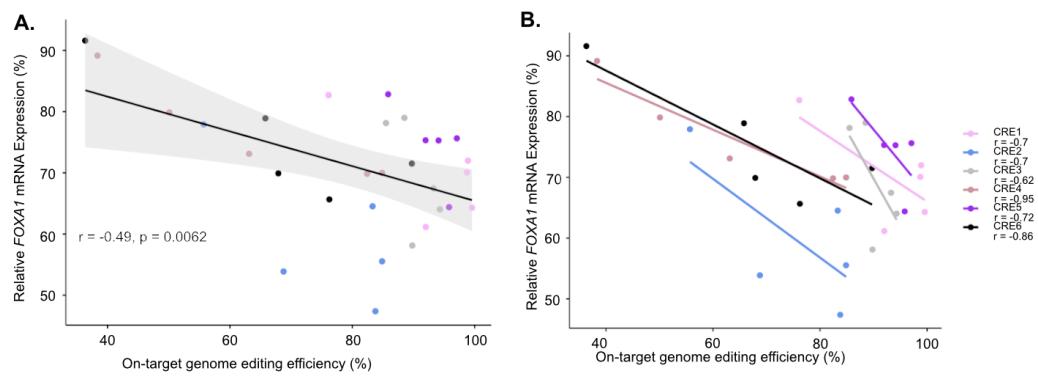


Figure A.6: Genome editing efficiency (%) is inversely correlated with *FOXA1* mRNA expression. **a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all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

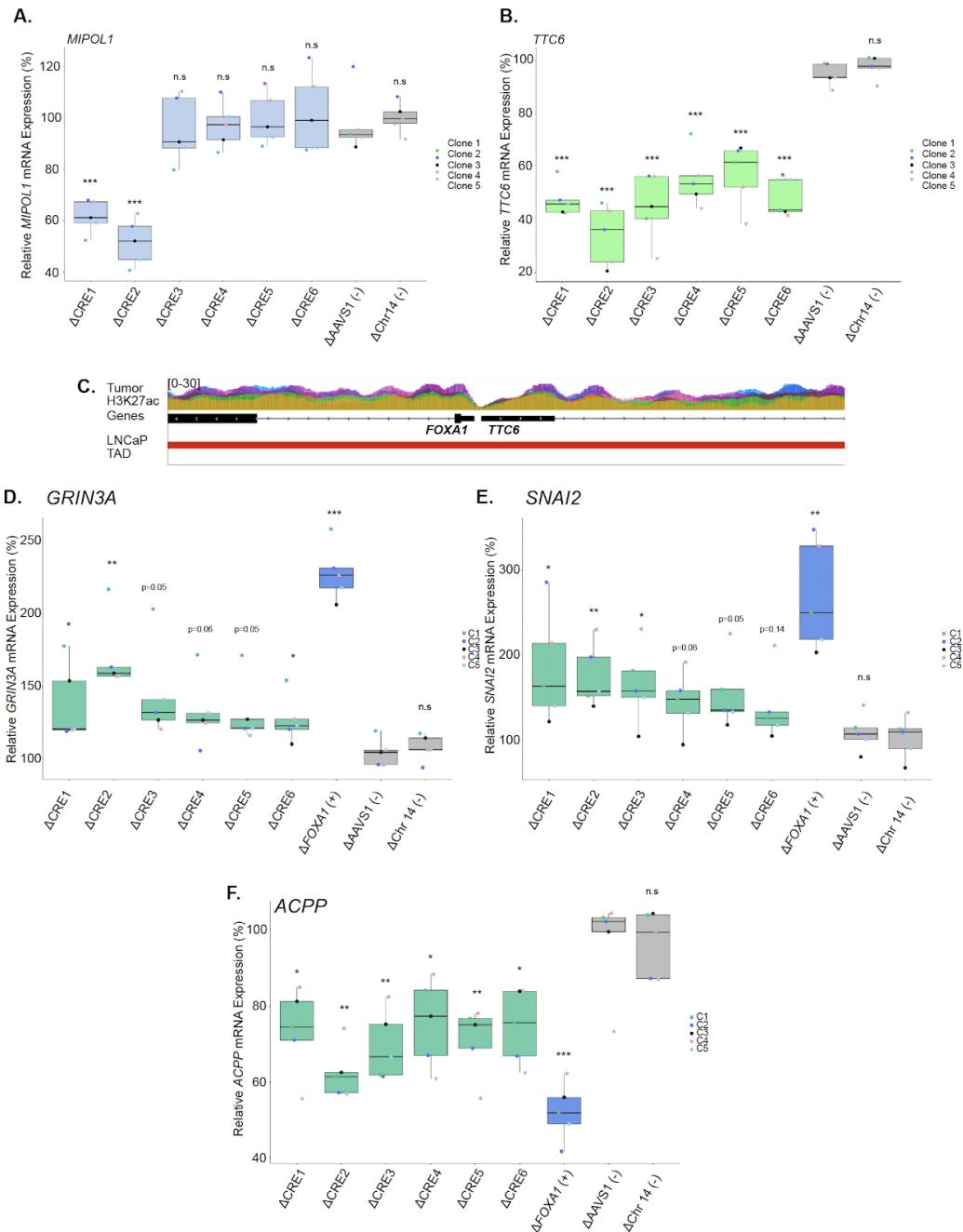


Figure A.7: Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a. *MIPO1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. b. *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each CRE. 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 \pm s.d. Student's *t*-test, * $p < 0.05$, ** $p < 0.01$, * $p < 0.001$.**

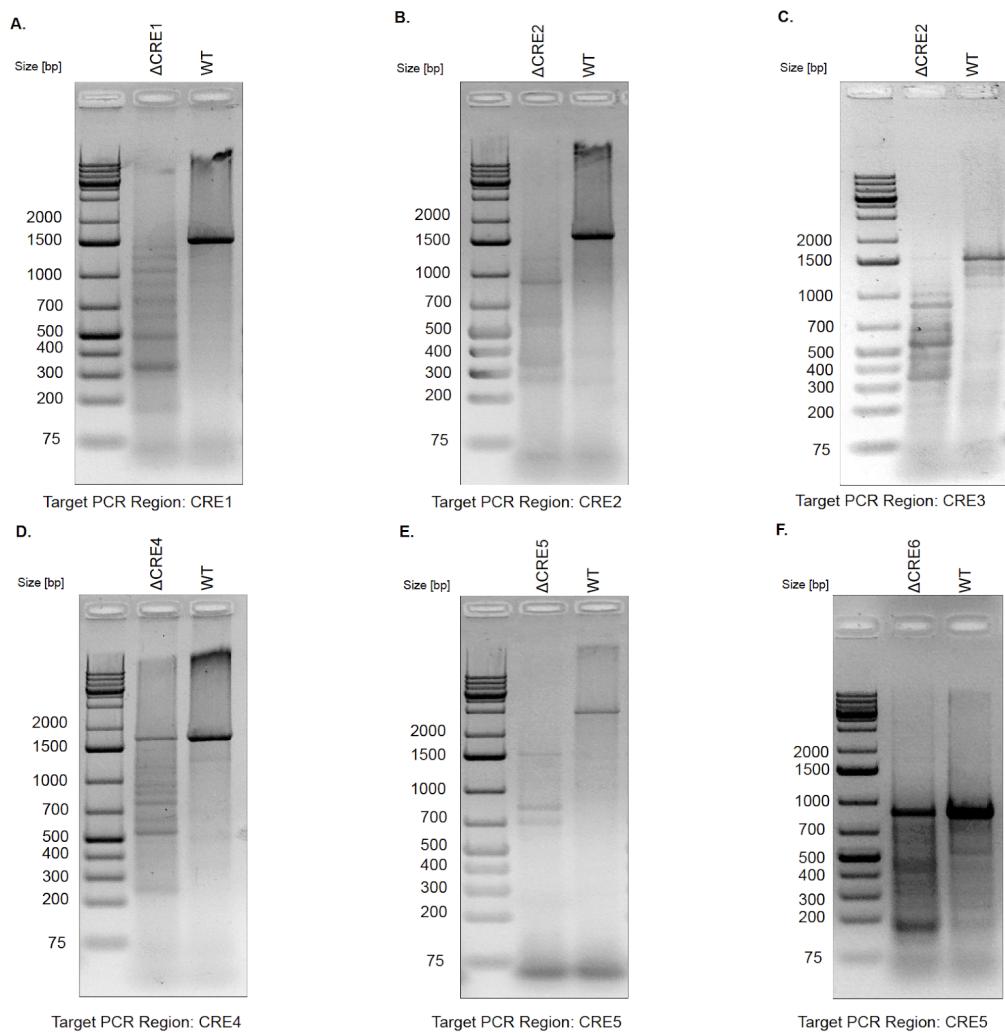


Figure A.8: Validation of transient Cas9-mediated single deletion of CREs. a-f. Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

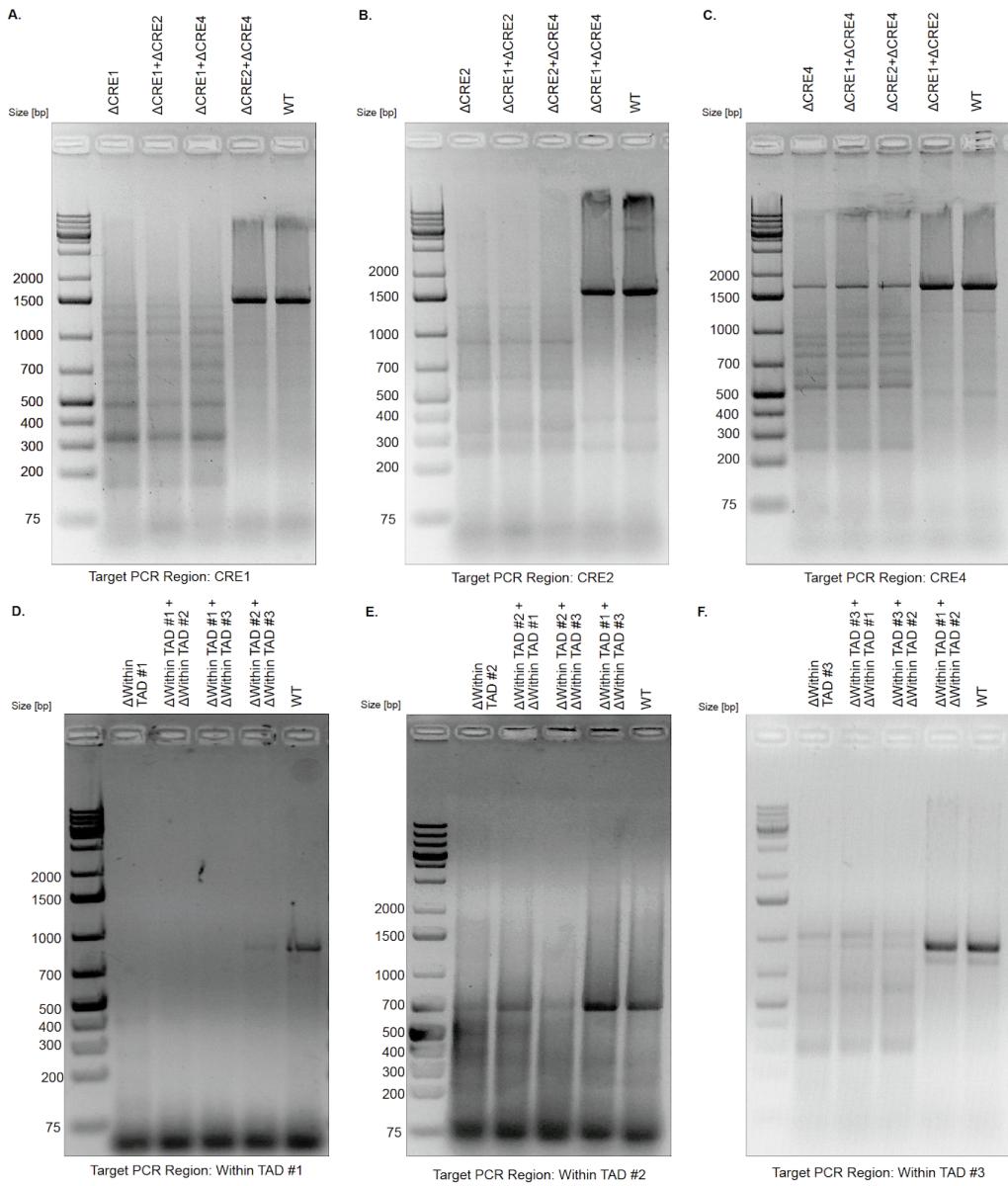


Figure A.9: Validation of transient Cas9-mediated double deletion of CREs. a-f. Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

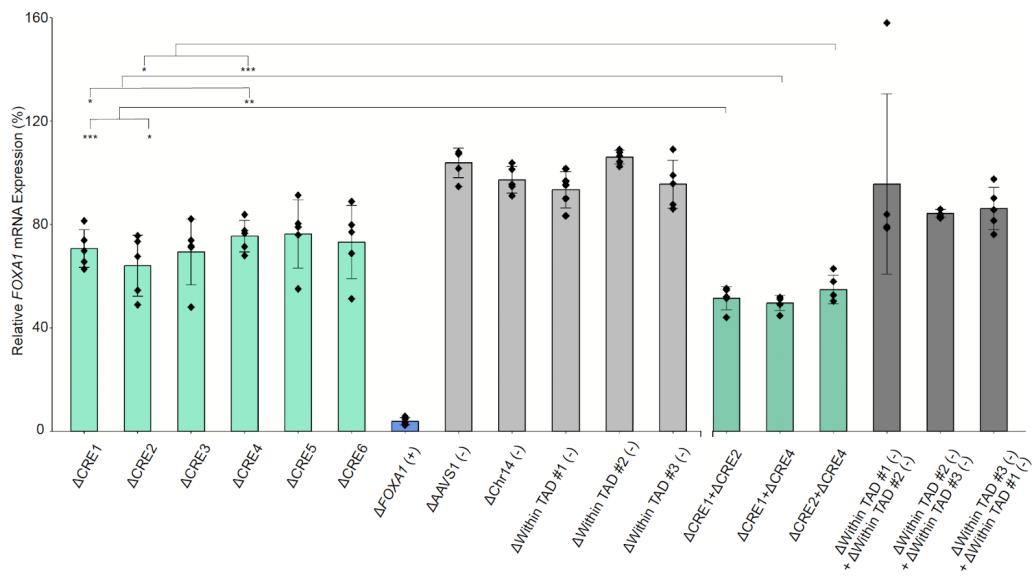


Figure A.10: Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s). *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate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

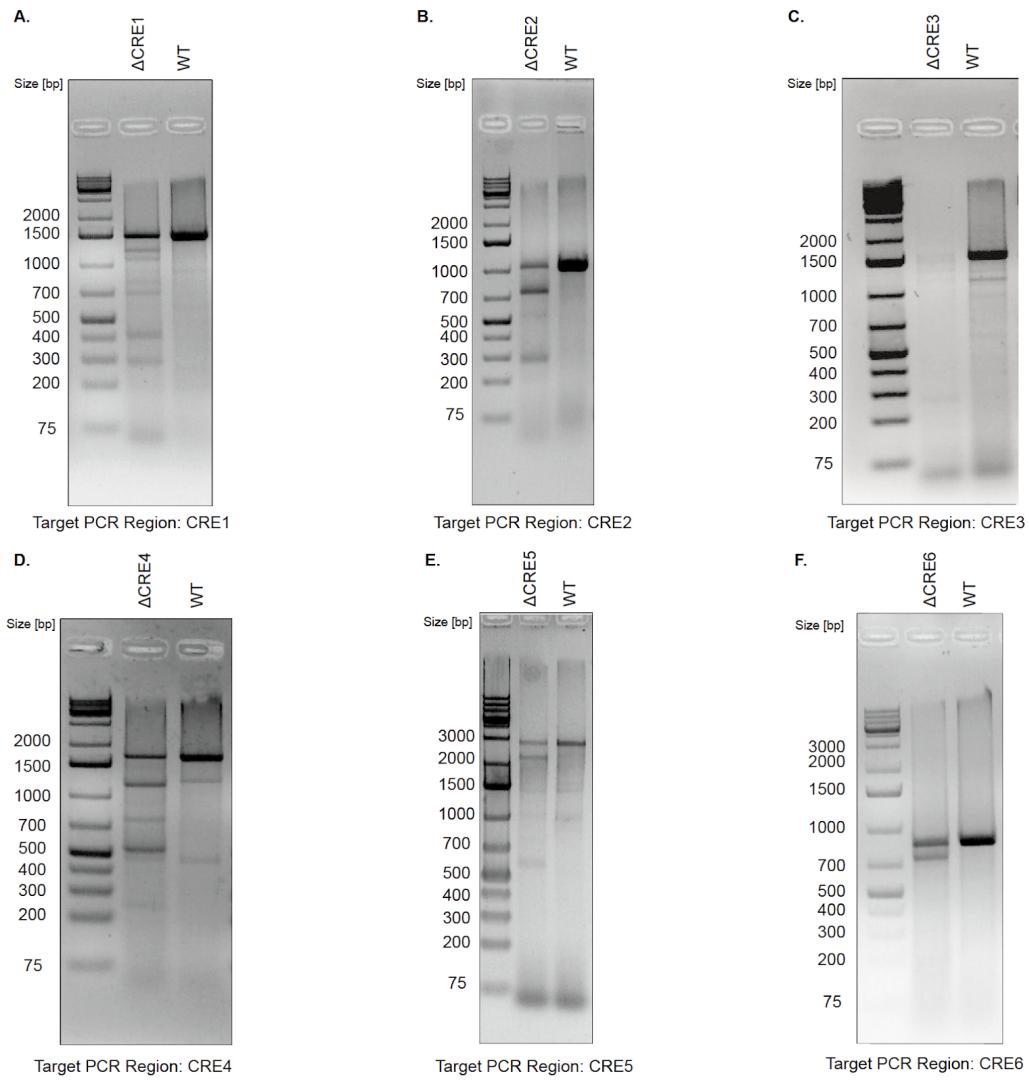


Figure A.11: Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f. Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs followed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

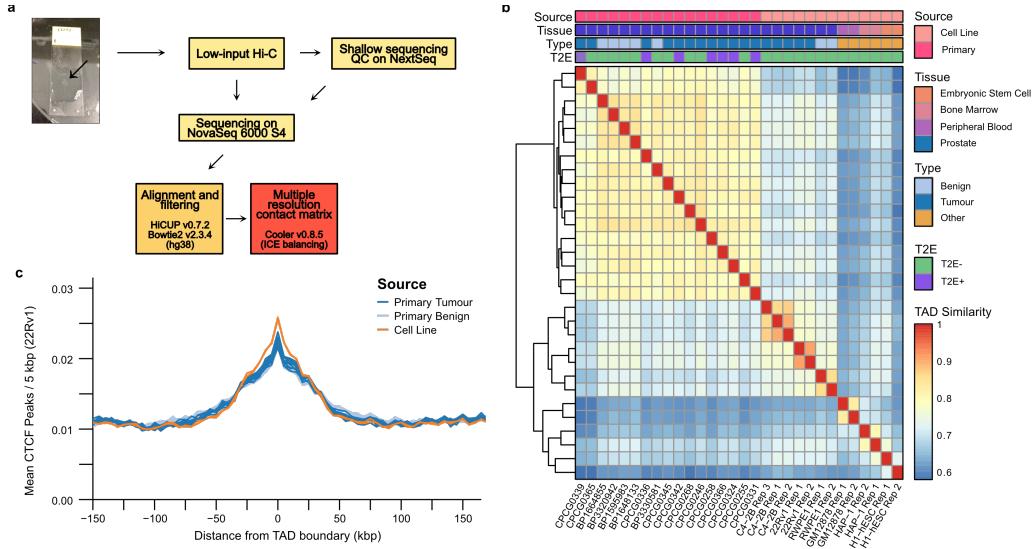


Figure B.1: Sample processing and TAD similarity between samples. **a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate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

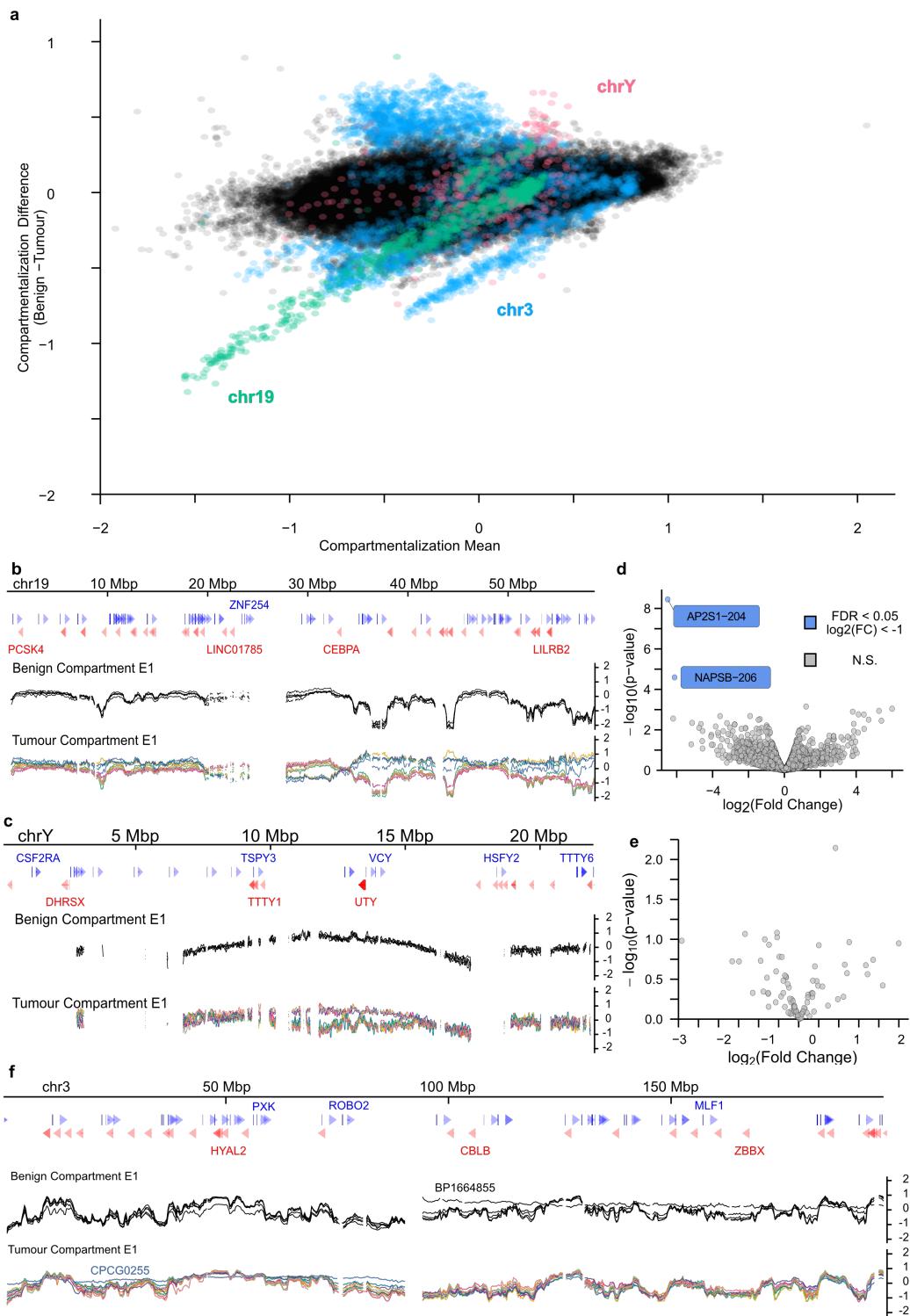


Figure B.2: Compartmentalization changes in tumours is not associated with widespread differential gene expression. (Continued on the following page)

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.

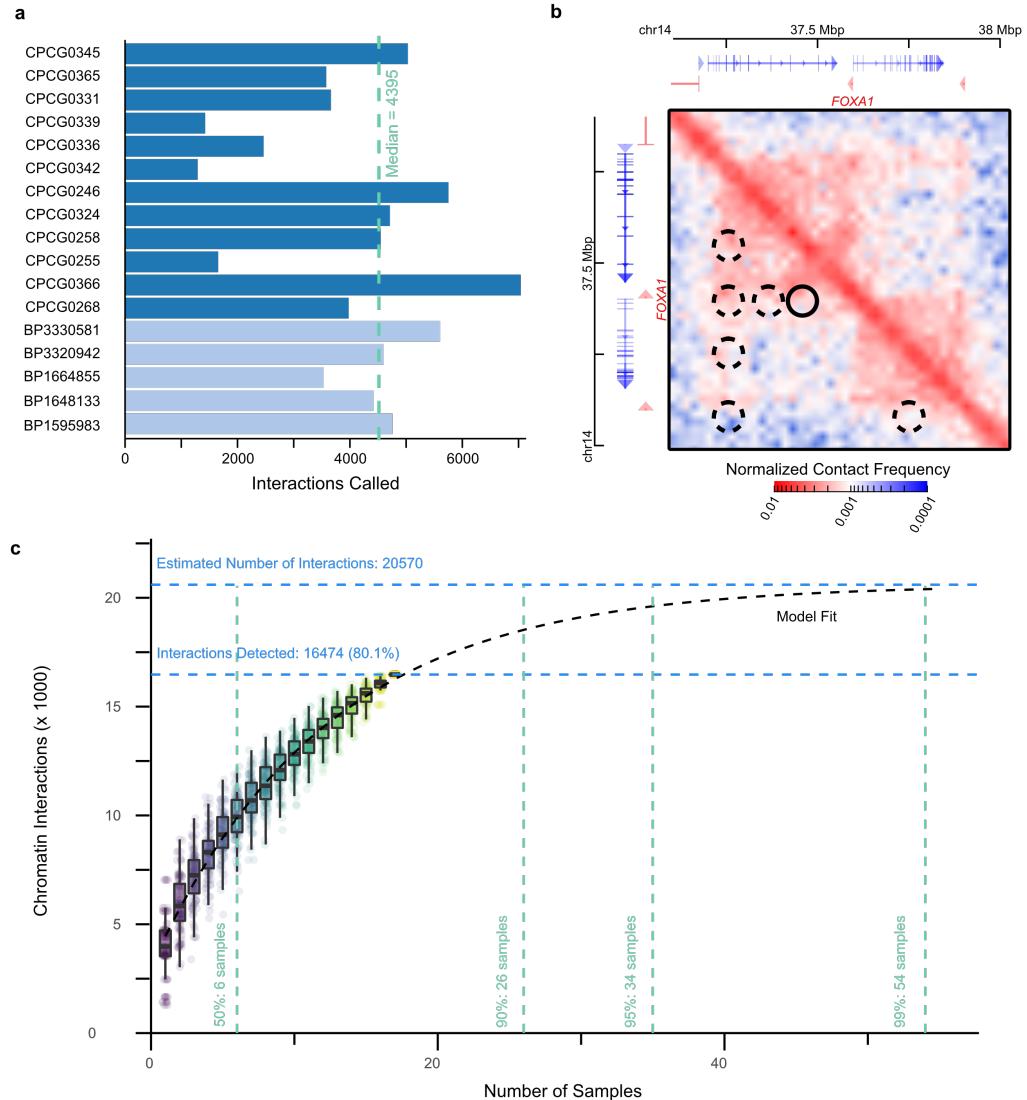


Figure B.3: Characterization of chromatin interactions in benign and tumour tissue.

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.

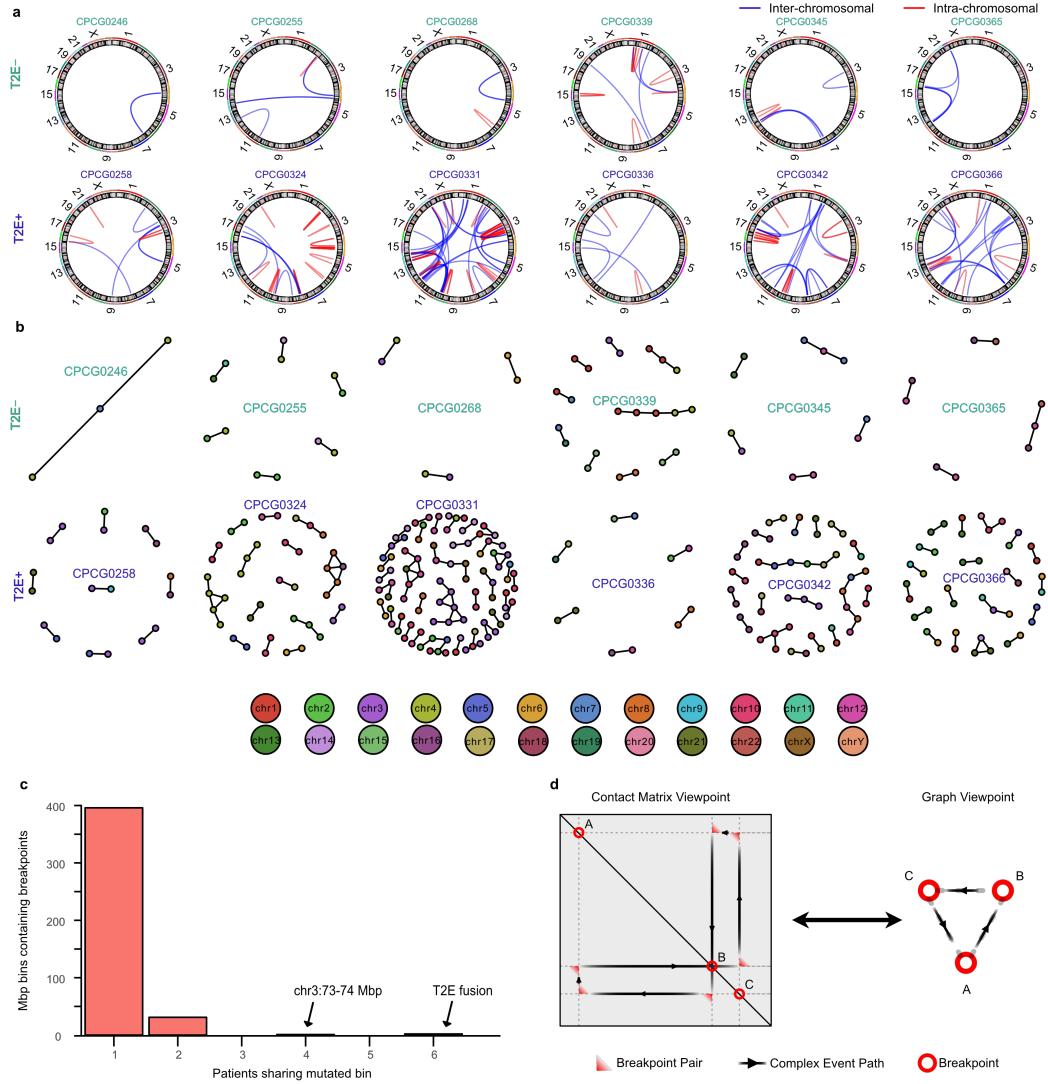


Figure B.4: Structural variant detection from Hi-C data. **a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1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

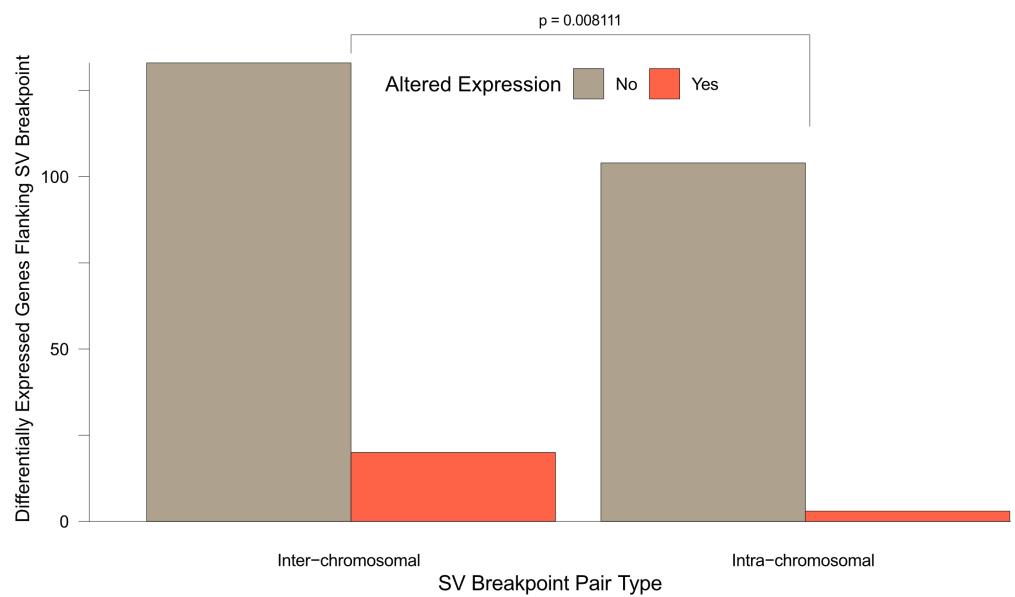


Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes.

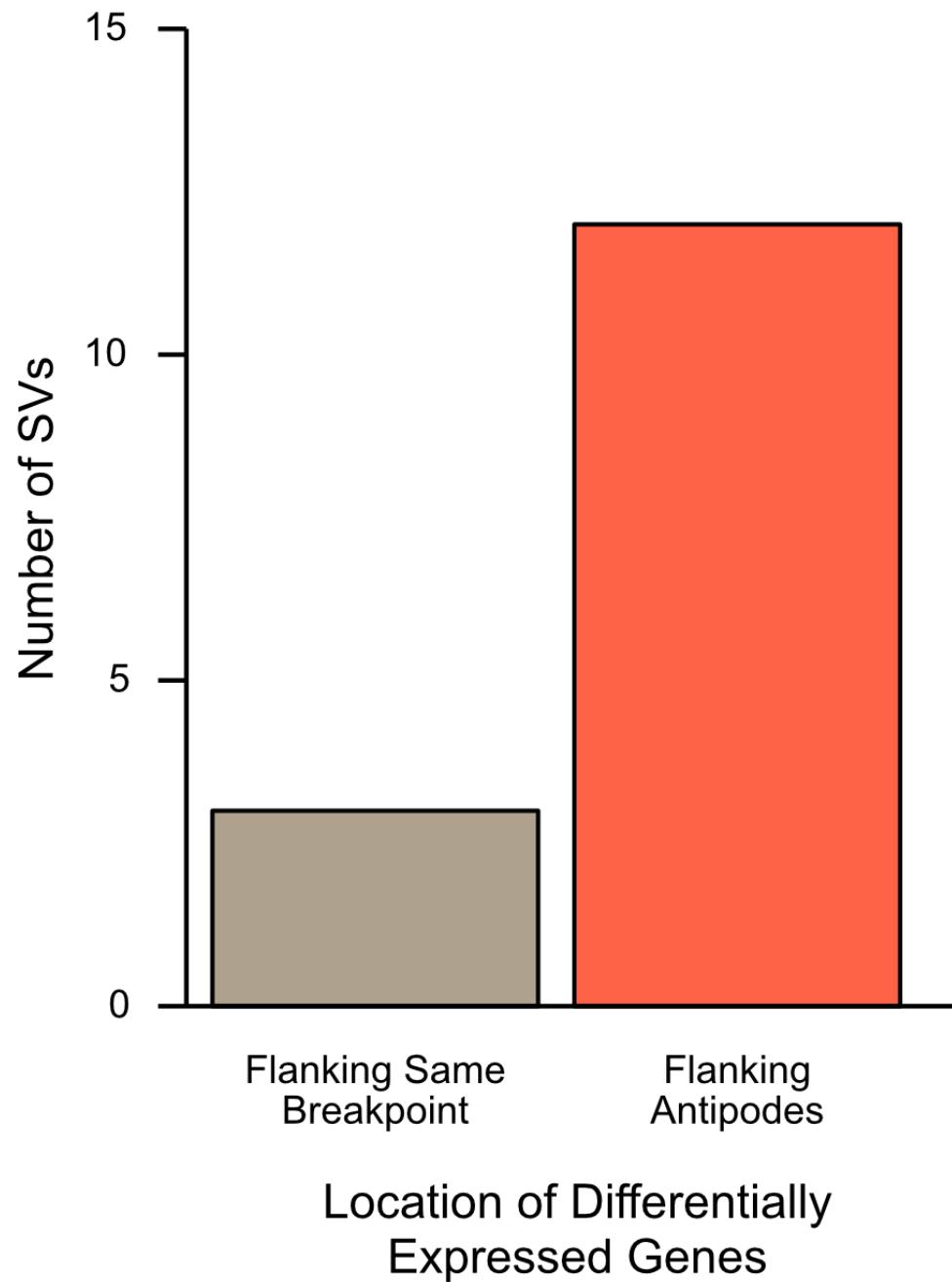


Figure B.6: **Location of differentially expressed genes around SV breakpoints.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

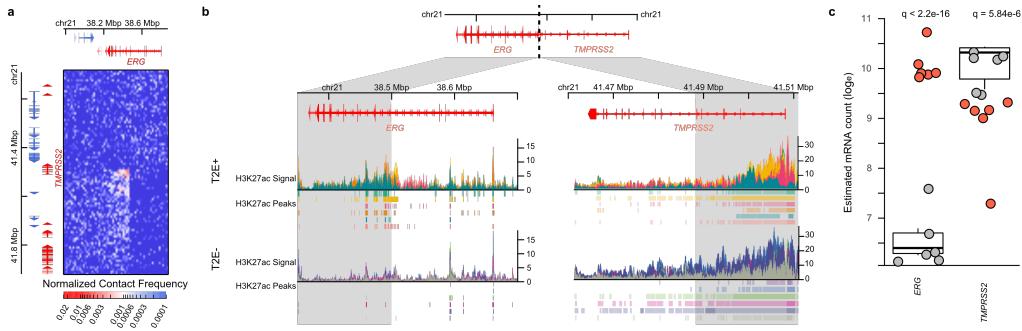


Figure B.7: Chromatin organization of the *TMPRSS2-ERG* fusion. **a.** Contact matrix of the deletion between *TMPRSS2* and *ERG*. **b.** Genome tracks of H3K27ac ChIP-seq signal in T2E+ and T2E- patients. The grey region highlights the loci that come into contact as a result of the deletion. **c.** 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) [192, 193] 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

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

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:

$$\text{Cov}[\epsilon_{si}, \epsilon_{rj}] = \sigma_s^2 \delta_{i,j} \delta_{s,r}$$

$$\text{Cov}[\zeta_{si}, \zeta_{rj}] = \tau_s^2 \delta_{i,j} \delta_{s,r}$$

$$\text{Cov}[\epsilon_{si}, \zeta_{rj}] = 0$$

$$\forall s, r \forall i, j$$

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

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

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

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

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

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

where d_s is the observed abundances given by

$$d_{si} = \ln \left(\frac{k_{si}}{\hat{f}_i} + 0.5 \right)$$

$$\hat{f}_i = \underset{s \in S^*}{\text{median}} \frac{k_{si}}{\sqrt[N]{\prod_{j=1}^N k_{sj}}}$$

where k_{si} is the estimated read count from the Kallisto package (v0.46.1) [194] 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 192], the estimator is unbiased, Namely

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

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

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

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 \mathbf{B}_1 to approximate the expected value of $\hat{\mathbf{B}}_1^{(JS)}$. Consider:

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

where

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

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

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

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

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

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 $\|u - a\|^2 \ll \|B_1\|^2$), the Taylor approximation is close to

$$\mathbb{E} [\hat{B}_1^{(JS)}] \approx \left(1 - \frac{c}{B_1^T \Sigma^{-1} B_1}\right) B_1 \quad (C.6)$$

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

C.3.2 Variance of the James-Stein estimator

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

$$\mathbb{E} [\|\hat{B}_1^{(JS)} - B_1\|^2] = \sum_{s \in S} \mathbb{E} [\left(\hat{B}_{1,s}^{(JS)} - B_{1,s}\right)^2] = \sum_{s \in S} \mathbb{V} [\hat{B}_{1,s}^{(JS)}]$$

By [REF 195], $\mathbb{E} [\|\hat{B}_1^{(JS)} - B_1\|^2] \leq \mathbb{E} [\|\hat{B}_1^{(OLS)} - B_1\|^2]$. However, this does not imply that $\mathbb{V} [\hat{B}_{1,s}^{(JS)}] \leq \mathbb{V} [\hat{B}_{1,s}^{(OLS)}] \forall s \in S$. 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.

$$\begin{aligned} \mathbb{V} [\hat{B}_1^{(JS)}] &\approx \mathbb{E} [\hat{B}_1^{(JS)} (\hat{B}_1^{(JS)})^T] - \left(1 - \frac{c}{B_1^T \Sigma^{-1} B_1}\right)^2 B_1 B_1^T \\ &= \Sigma^{1/2} \mathbb{E} \left[uu^T - \frac{2c}{u^T u} uu^T + \left(\frac{c}{u^T u}\right)^2 uu^T \right] \Sigma^{1/2} - \left(1 - \frac{c}{B_1^T \Sigma^{-1} B_1}\right)^2 B_1 B_1^T \end{aligned}$$

where, again, $u = \Sigma^{-1/2} \hat{B}_1^{(OLS)}$. Expanding about $a = \Sigma^{-1/2} B_1$ yields:

$$\mathbb{V} [\hat{B}_1^{(JS)}] = \left(1 - \frac{2c}{B_1^T \Sigma^{-1} B_1}\right) \Sigma - \frac{2c}{(B_1^T \Sigma^{-1} B_1)^2} B_1 B_1^T + \mathcal{O}(\|u - a\|^4)$$

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

$$\mathbb{V} \left[\hat{B}_1^{(JS)} \right] \approx \left(1 - \frac{2c}{B_1^T \Sigma^{-1} B_1} \right) \Sigma - \frac{2c}{(B_1^T \Sigma^{-1} B_1)^2} B_1 B_1^T \quad (C.7)$$

Since the diagonal elements of $\frac{2c}{(B_1^T \Sigma^{-1} B_1)^2} B_1 B_1^T$ are all ≥ 0 and $0 \leq \left(1 - \frac{2c}{B_1^T \Sigma^{-1} B_1} \right) \leq 1 \forall c > 0$, 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:

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

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

The coefficient of $\hat{B}_{1,s}^{(OLS)}$ 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 $B_1 B_1^T$ 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 $\frac{2c}{(B_1^T \Sigma^{-1} B_1)^2}$ coefficient. While rarely true in practice, this statistical dependence can affect the results of statistical inference, in theory. For most purposes, is not expected to have a large effect on the results of statistical inference.

Appendix D

Supplementary Material for Chapter 5

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

Patient	Subtype	Age	Sex	Bone Marrow Blast Count	Time to Relapse (months)
1	DUX4	> 18	M	90%	9.00
4	B-other	> 18	M	90%	6.30
6	B-other	> 18	M	90%	33.97
7	DUX4	< 18	F	92%	39.60
9	B-other	< 18	M	96%	48.12

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. Gilbertson, R. J. Mapping Cancer Origins. English. *Cell* **145**, 25–29. ISSN: 0092-8674, 1097-4172 (Apr. 2011).
3. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (Jan. 2000).
4. 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).
5. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic Plasticity and the Hallmarks of Cancer. *Science* **357**, eaal2380–eaal2380 (July 2017).
6. Pavlova, N. N. & Thompson, C. B. The Emerging Hallmarks of Cancer Metabolism. en. *Cell Metabolism* **23**, 27–47. ISSN: 1550-4131 (Jan. 2016).
7. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. English. *Cell* **153**, 17–37. ISSN: 0092-8674, 1097-4172 (Mar. 2013).
8. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic Instabilities in Human Cancers. *Nature* **396**, 643–649 (Dec. 1998).
9. Abeshouse, A. *et al.* The Molecular Taxonomy of Primary Prostate Cancer. en. *Cell* **163**, 1011–1025. ISSN: 00928674 (Nov. 2015).
10. Vinagre, J. *et al.* Frequency of TERT Promoter Mutations in Human Cancers. en. *Nature Communications* **4**, 2185. ISSN: 2041-1723 (Oct. 2013).
11. 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).

12. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. en. *Science* **339**, 959–961. ISSN: 0036-8075, 1095-9203 (Feb. 2013).
13. Nagarajan, R. P. *et al.* Recurrent Epimutations Activate Gene Body Promoters in Primary Glioblastoma. *Genome Research* **24**, 761–774 (May 2014).
14. 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).
15. Alberts, B. *Molecular Biology of the Cell* Sixth edition. ISBN: 978-0-8153-4524-4 (Garland Science, Taylor and Francis Group, New York, NY, 2015).
16. Goodrich, J. A. & Tjian, R. Unexpected Roles for Core Promoter Recognition Factors in Cell-Type-Specific Transcription and Gene Regulation. en. *Nature Reviews Genetics* **11**, 549–558. ISSN: 1471-0056, 1471-0064 (Aug. 2010).
17. Schoenfelder, S. & Fraser, P. Long-Range Enhancer–Promoter Contacts in Gene Expression Control. En. *Nature Reviews Genetics*, 1. ISSN: 1471-0064 (May 2019).
18. Spitz, F. & Furlong, E. E. M. Transcription Factors: From Enhancer Binding to Developmental Control. en. *Nature Reviews Genetics* **13**, 613–626. ISSN: 1471-0064 (Sept. 2012).
19. Ong, C.-T. & Corces, V. G. Enhancer Function: New Insights into the Regulation of Tissue-Specific Gene Expression. en. *Nature Reviews Genetics* **12**, 283–293. ISSN: 1471-0064 (Apr. 2011).
20. Andersson, R. & Sandelin, A. Determinants of Enhancer and Promoter Activities of Regulatory Elements. en. *Nature Reviews Genetics* **21**, 71–87. ISSN: 1471-0064 (Feb. 2020).
21. Gaszner, M. & Felsenfeld, G. Insulators: Exploiting Transcriptional and Epigenetic Mechanisms. en. *Nature Reviews Genetics* **7**, 703–713. ISSN: 1471-0064 (Sept. 2006).
22. Oudelaar, A. M. & Higgs, D. R. The Relationship between Genome Structure and Function. en. *Nature Reviews Genetics*. ISSN: 1471-0056, 1471-0064 (Nov. 2020).
23. Farnham, P. J. Insights from Genomic Profiling of Transcription Factors. en. *Nature Reviews Genetics* **10**, 605–616. ISSN: 1471-0064 (Sept. 2009).
24. Liu, T. *et al.* Cistrome: An Integrative Platform for Transcriptional Regulation Studies. en. *Genome Biology* **12**, 1–10. ISSN: 1474-760X (Aug. 2011).
25. Lupien, M. & Brown, M. Cistromics of Hormone-Dependent Cancer. *Endocrine-Related Cancer* **16**, 381–389. ISSN: 1351-0088, 1479-6821 (June 2009).

26. Kim, T. H. *et al.* Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. en. *Cell* **128**, 1231–1245. ISSN: 0092-8674 (Mar. 2007).
27. 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).
28. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. en. *Science* **328**, 232–235. ISSN: 0036-8075, 1095-9203 (Apr. 2010).
29. Maurano, M. T., Wang, H., Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread Site-Dependent Buffering of Human Regulatory Polymorphism. *PLoS Genetics* **8**. ISSN: 1553-7404 (Electronic)\n1553-7390 (Linking) (2012).
30. 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).
31. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. en. *Cell Reports* **12**, 1184–1195. ISSN: 22111247 (Aug. 2015).
32. Wang, H. *et al.* Widespread Plasticity in CTCF Occupancy Linked to DNA Methylation. *Genome Research* **22**, 1680–1688. ISSN: 1549-5469 (Electronic)\n1088-9051 (Linking) (Sept. 2012).
33. Wiehle, L. *et al.* DNA (de)Methylation in Embryonic Stem Cells Controls CTCF-Dependent Chromatin Boundaries. en. *Genome Research*, gr.239707.118. ISSN: 1088-9051, 1549-5469 (Apr. 2019).
34. Xu, C. & Corces, V. G. Nascent DNA Methylome Mapping Reveals Inheritance of Hemimethylation at CTCF/Cohesin Sites. *Science* **359**, 1166–1170 (2018).
35. Viner, C. *et al.* Modeling Methyl-Sensitive Transcription Factor Motifs with an Expanded Epigenetic Alphabet. en. *bioRxiv*, 043794 (Mar. 2016).
36. Goll, M. G. & Bestor, T. H. Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry* **74**, 481–514. ISSN: 0066-4154 (June 2005).
37. Lister, R. *et al.* Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences. *Nature* **462**, 315–322. ISSN: 1476-4687 (Electronic)\n0028-0836 (Linking) (Nov. 2009).
38. Henikoff, S. Nucleosome Destabilization in the Epigenetic Regulation of Gene Expression. en. *Nature Reviews Genetics* **9**, 15–26. ISSN: 1471-0064 (Jan. 2008).

39. Jiang, C. & Pugh, B. F. Nucleosome Positioning and Gene Regulation: Advances through Genomics. en. *Nature Reviews Genetics* **10**, 161–172. ISSN: 1471-0064 (Mar. 2009).
40. Vierstra, J. et al. Global Reference Mapping of Human Transcription Factor Footprints. en. *Nature* **583**, 729–736. ISSN: 1476-4687 (July 2020).
41. Cusanovich, D. A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. en. *Cell* **174**, 1309–1324.e18. ISSN: 0092-8674 (Aug. 2018).
42. Polak, P. et al. Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer. en. *Nature* **518**, 360–364. ISSN: 1476-4687 (Feb. 2015).
43. Zhu, H., Wang, G. & Qian, J. Transcription Factors as Readers and Effectors of DNA Methylation. *Nature Reviews Genetics* **17**, 551–565 (Aug. 2016).
44. Furey, T. S. ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions. en. *Nature Reviews Genetics* **13**, 840–852. ISSN: 1471-0064 (Dec. 2012).
45. Carter, B. & Zhao, K. The Epigenetic Basis of Cellular Heterogeneity. en. *Nature Reviews Genetics* **22**, 235–250. ISSN: 1471-0064 (Apr. 2021).
46. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting Histone Modifications and the Functional Organization of Mammalian Genomes. *Nature Reviews Genetics* **12**, 7–18 (Jan. 2011).
47. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. English. *Cell* **164**, 1110–1121. ISSN: 0092-8674, 1097-4172 (Mar. 2016).
48. 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).
49. 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).
50. 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).
51. Mirny, L. A., Imakaev, M. & Abdennur, N. Two Major Mechanisms of Chromosome Organization. en. *Current Opinion in Cell Biology. Cell Nucleus* **58**, 142–152. ISSN: 0955-0674 (June 2019).
52. Stergachis, A. B. et al. Conservation of Trans-Acting Circuitry during Mammalian Regulatory Evolution. *Nature* **515**, 365–370. ISSN: 1476-4687 (Electronic)\r0028-0836 (Linking) (2014).

53. Berthelot, C., Villar, D., Horvath, J., Odom, D. T. & Flieck, P. Complexity and Conservation of Regulatory Landscapes Underlie Evolutionary Resilience of Mammalian Gene Expression. *bioRxiv*, 1–31 (2017).
54. Spurrell, C. H., Dickel, D. E. & Visel, A. The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome. English. *Cell* **167**, 1163–1166. ISSN: 0092-8674, 1097-4172 (Nov. 2016).
55. Buenrostro, J. D. *et al.* Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. en. *Nature* **523**, 486–490. ISSN: 0028-0836, 1476-4687 (June 2015).
56. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. English. *Cell* **152**, 1237–1251. ISSN: 0092-8674, 1097-4172 (Mar. 2013).
57. Conesa, A. *et al.* A Survey of Best Practices for RNA-Seq Data Analysis. *Genome Biology* **17**, 13–13. ISSN: 1474-760X (Electronic)\r1474-7596 (Linking) (Dec. 2016).
58. Robertson, G. *et al.* Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing. en. *Nature Methods* **4**, 651–657. ISSN: 1548-7091, 1548-7105 (Aug. 2007).
59. Bailey, T. *et al.* Practical Guidelines for the Comprehensive Analysis of ChIP-Seq Data. en. *PLOS Computational Biology* **9**, e1003326. ISSN: 1553-7358 (Nov. 2013).
60. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in Situ Genome-Wide Profiling with High Efficiency for Low Cell Numbers. en. *Nature Protocols* **13**, 1006–1019. ISSN: 1754-2189, 1750-2799 (May 2018).
61. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. English. *Cell* **132**, 311–322. ISSN: 0092-8674, 1097-4172 (Jan. 2008).
62. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nature Methods* **10**, 1213–8 (Dec. 2013).
63. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology* **6**, 356–372. ISSN: 6314442508 (2015).
64. Corces, M. R. *et al.* An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues. en. *Nature Methods* **14**, 959–962. ISSN: 1548-7105 (Oct. 2017).
65. Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. English. *Cell* **132**, 887–898. ISSN: 0092-8674, 1097-4172 (Mar. 2008).

66. Laird, P. W. Principles and Challenges of Genome-Wide DNA Methylation Analysis. *Nature Reviews Genetics* **11**, 191–191. ISSN: 1471-0064 (Electronic)\r1471-0056 (Linking) (Mar. 2010).
67. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. en. *Science* **295**, 1306–1311. ISSN: 0036-8075, 1095-9203 (Feb. 2002).
68. 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).
69. Birney, E. *et al.* Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project. en. *Nature* **447**, 799–816. ISSN: 1476-4687 (June 2007).
70. Moore, J. E. *et al.* Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. en. *Nature* **583**, 699–710. ISSN: 1476-4687 (July 2020).
71. Ernst, J. & Kellis, M. ChromHMM: Automating Chromatin-State Discovery and Characterization. en. *Nature Methods* **9**, 215–216. ISSN: 1548-7105 (Mar. 2012).
72. Hoffman, M. M. *et al.* Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation. en. *Nature Methods* **9**, 473–476. ISSN: 1548-7091, 1548-7105 (May 2012).
73. Chan, R. C. W. *et al.* Segway 2.0: Gaussian Mixture Models and Minibatch Training. en. *Bioinformatics* **34** (ed Birol, I.) 669–671. ISSN: 1367-4803, 1460-2059 (Feb. 2018).
74. 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).
75. 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).
76. Croce, C. M. Oncogenes and Cancer. *New England Journal of Medicine* **358**, 502–511. ISSN: 0028-4793 (Jan. 2008).
77. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. English. *Cell* **173**, 371–385.e18. ISSN: 0092-8674, 1097-4172 (Apr. 2018).
78. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. en. *Nature Genetics* **45**, 1113–1120. ISSN: 1546-1718 (Oct. 2013).

79. Pan-Cancer Analysis of Whole Genomes. en. *Nature* **578**, 82–93. ISSN: 1476-4687 (Feb. 2020).
80. 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).
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. Zhang, Y. *et al.* High-Coverage Whole-Genome Analysis of 1220 Cancers Reveals Hundreds of Genes Deregulated by Rearrangement-Mediated Cis -Regulatory Alterations. en. *Nature Communications* **11**, 1–14. ISSN: 2041-1723 (Feb. 2020).
83. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. P53 Mutations in Human Cancers. en. *Science* **253**, 49–53. ISSN: 0036-8075, 1095-9203 (July 1991).
84. 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).
85. Meyer, N. & Penn, L. Z. Reflecting on 25 Years with MYC. en. *Nature Reviews Cancer* **8**, 976–990. ISSN: 1474-1768 (Dec. 2008).
86. 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).
87. Kron, K. J., Bailey, S. D. & Lupien, M. Enhancer Alterations in Cancer: A Source for a Cell Identity Crisis. en. *Genome Medicine* **6**, 1–12. ISSN: 1756-994X (Dec. 2014).
88. 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).
89. 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).
90. 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).
91. 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).
92. Northcott, P. A. *et al.* Enhancer Hijacking Activates GFI1 Family Oncogenes in Medulloblastoma. en. *Nature* **511**, 428–434. ISSN: 1476-4687 (July 2014).

93. 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).
94. 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).
95. Pirozzi, C. J. & Yan, H. The Implications of IDH Mutations for Cancer Development and Therapy. en. *Nature Reviews Clinical Oncology*, 1–17. ISSN: 1759-4782 (June 2021).
96. Im, A. P. *et al.* DNMT3A and IDH Mutations in Acute Myeloid Leukemia and Other Myeloid Malignancies: Associations with Prognosis and Potential Treatment Strategies. en. *Leukemia* **28**, 1774–1783. ISSN: 1476-5551 (Sept. 2014).
97. Issa, G. C. & DiNardo, C. D. Acute Myeloid Leukemia with IDH1 and IDH2 Mutations: 2021 Treatment Algorithm. en. *Blood Cancer Journal* **11**, 1–7. ISSN: 2044-5385 (June 2021).
98. Molenaar, R. J., Maciejewski, J. P., Wilmink, J. W. & van Noorden, C. J. F. Wild-Type and Mutated IDH1/2 Enzymes and Therapy Responses. en. *Oncogene* **37**, 1949–1960. ISSN: 1476-5594 (Apr. 2018).
99. Shih, A. H., Abdel-Wahab, O., Patel, J. P. & Levine, R. L. The Role of Mutations in Epigenetic Regulators in Myeloid Malignancies. en. *Nature Reviews Cancer* **12**, 599–612. ISSN: 1474-1768 (Sept. 2012).
100. Dominguez, P. M. *et al.* TET2 Deficiency Causes Germinal Center Hyperplasia, Impairs Plasma Cell Differentiation, and Promotes B-Cell Lymphomagenesis. en. *Cancer Discovery* **8**, 1632–1653. ISSN: 2159-8274, 2159-8290 (Dec. 2018).
101. Plass, C. *et al.* Mutations in Regulators of the Epigenome and Their Connections to Global Chromatin Patterns in Cancer. en. *Nature Reviews Genetics* **14**, 765–780. ISSN: 1471-0064 (Nov. 2013).
102. Nikoloski, G. *et al.* Somatic Mutations of the Histone Methyltransferase Gene EZH2 in Myelodysplastic Syndromes. en. *Nature Genetics* **42**, 665–667. ISSN: 1546-1718 (Aug. 2010).
103. Ernst, T. *et al.* Inactivating Mutations of the Histone Methyltransferase Gene EZH2 in Myeloid Disorders. en. *Nature Genetics* **42**, 722–726. ISSN: 1546-1718 (Aug. 2010).
104. Morin, R. D. *et al.* Somatic Mutations Altering EZH2 (Tyr641) in Follicular and Diffuse Large B-Cell Lymphomas of Germinal-Center Origin. en. *Nature Genetics* **42**, 181–185. ISSN: 1546-1718 (Feb. 2010).

105. Varambally, S. *et al.* The Polycomb Group Protein EZH2 Is Involved in Progression of Prostate Cancer. en. *Nature* **419**, 624–629. ISSN: 1476-4687 (Oct. 2002).
106. Xu, K. *et al.* EZH2 Oncogenic Activity in Castration-Resistant Prostate Cancer Cells Is Polycomb-Independent. en. *Science* **338**, 1465–1469. ISSN: 0036-8075, 1095-9203 (Dec. 2012).
107. Min, J. *et al.* An Oncogene–Tumor Suppressor Cascade Drives Metastatic Prostate Cancer by Coordinate Activating Ras and Nuclear Factor- κ B. en. *Nature Medicine* **16**, 286–294. ISSN: 1546-170X (Mar. 2010).
108. Kim, K. H. & Roberts, C. W. M. Targeting EZH2 in Cancer. en. *Nature Medicine* **22**, 128–134. ISSN: 1546-170X (Feb. 2016).
109. Jones, P. A. & Laird, P. W. Cancer-Epigenetics Comes of Age. en. *Nature Genetics* **21**, 163–167. ISSN: 1546-1718 (Feb. 1999).
110. Jones, P. A. & Baylin, S. B. The Fundamental Role of Epigenetic Events in Cancer. en. *Nature Reviews Genetics* **3**, 415–428. ISSN: 1471-0064 (June 2002).
111. Feinberg, A. P. & Tycko, B. The History of Cancer Epigenetics. *Nature Reviews Cancer* **4**, 143–153. ISSN: 1474-175X (Print)\r1474-175X (Linking) (Feb. 2004).
112. 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).
113. Mack, S. C. *et al.* Epigenomic Alterations Define Lethal CIMP-Positive Ependymomas of Infancy. en. *Nature* **506**, 445–450. ISSN: 0028-0836, 1476-4687 (Feb. 2014).
114. Issa, J.-P. CpG Island Methylator Phenotype in Cancer. en. *Nature Reviews Cancer* **4**, 988–993. ISSN: 1474-1768 (Dec. 2004).
115. Schmelz, K. *et al.* Induction of Gene Expression by 5-Aza-2'-Deoxycytidine in Acute Myeloid Leukemia (AML) and Myelodysplastic Syndrome (MDS) but Not Epithelial Cells by DNA-Methylation-Dependent and -Independent Mechanisms. en. *Leukemia* **19**, 103–111. ISSN: 1476-5551 (Jan. 2005).
116. Azad, N., Zahnow, C. A., Rudin, C. M. & Baylin, S. B. The Future of Epigenetic Therapy in Solid Tumours—Lessons from the Past. en. *Nature Reviews Clinical Oncology* **10**, 256–266. ISSN: 1759-4782 (May 2013).
117. Kelly, T. K., De Carvalho, D. D. & Jones, P. A. Epigenetic Modifications as Therapeutic Targets. en. *Nature Biotechnology* **28**, 1069–1078. ISSN: 1546-1696 (Oct. 2010).

118. Flavahan, W. A. *et al.* Altered Chromosomal Topology Drives Oncogenic Programs in SDH-Deficient GIST. en. *Nature*, 1–1. ISSN: 1476-4687 (Oct. 2019).
119. 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).
120. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. en. *Cell* **175**, 1074–1087.e18. ISSN: 0092-8674 (Nov. 2018).
121. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide Excision Repair Is Impaired by Binding of Transcription Factors to DNA. en. *Nature* **532**, 264–267. ISSN: 1476-4687 (Apr. 2016).
122. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. English. *Cell* **177**, 101–114. ISSN: 0092-8674, 1097-4172 (Mar. 2019).
123. Wang, B. *et al.* Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. en. *Nature Methods* **11**, 333–337. ISSN: 1548-7091, 1548-7105 (Mar. 2014).
124. Rappoport, N. & Shamir, R. Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark. *Nucleic Acids Research* **46**, 10546–10562. ISSN: 0305-1048 (Nov. 2018).
125. Brenner, D. R. *et al.* Projected Estimates of Cancer in Canada in 2020. en. *CMAJ* **192**, E199–E205. ISSN: 0820-3946, 1488-2329 (Mar. 2020).
126. Rebello, R. J. *et al.* Prostate Cancer. en. *Nature Reviews Disease Primers* **7**, 1–27. ISSN: 2056-676X (Feb. 2021).
127. Hahn, A. W., Higano, C. S., Taplin, M.-E., Ryan, C. J. & Agarwal, N. Metastatic Castration-Sensitive Prostate Cancer: Optimizing Patient Selection and Treatment. *American Society of Clinical Oncology Educational Book*, 363–371. ISSN: 1548-8748 (May 2018).
128. Institute, N. C. *SEER Cancer Stat Facts: Prostate Cancer* <https://seer.cancer.gov/statfacts/html/prost.html>.
129. Rawla, P. Epidemiology of Prostate Cancer. *World Journal of Oncology* **10**, 63–89. ISSN: 1920-4531 (Apr. 2019).
130. Smith, Z. L., Eggener, S. E. & Murphy, A. B. African-American Prostate Cancer Disparities. en. *Current Urology Reports* **18**, 81. ISSN: 1534-6285 (Aug. 2017).
131. Dall'Era, M. A., deVere-White, R., Rodriguez, D. & Cress, R. Changing Incidence of Metastatic Prostate Cancer by Race and Age, 1988–2015. en. *European Urology Focus* **5**, 1014–1021. ISSN: 2405-4569 (Nov. 2019).

132. Fraser, M. *et al.* Genomic Hallmarks of Localized, Non-Indolent Prostate Cancer. en. *Nature* **541**, 359–364. ISSN: 1476-4687 (Jan. 2017).
133. Li, J. *et al.* A Genomic and Epigenomic Atlas of Prostate Cancer in Asian Populations. en. *Nature*, 1–7. ISSN: 1476-4687 (Mar. 2020).
134. PCF/SU2C International Prostate Cancer Dream Team *et al.* The Long Tail of Oncogenic Drivers in Prostate Cancer. en. *Nature Genetics* **50**, 645–651. ISSN: 1061-4036, 1546-1718 (May 2018).
135. Kron, K. J. *et al.* TMPRSS2–ERG Fusion Co-Opt Master Transcription Factors and Activates NOTCH Signaling in Primary Prostate Cancer. en. *Nature Genetics* **49**, 1336–1345. ISSN: 1546-1718 (Sept. 2017).
136. 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).
137. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. en. *Cell* **161**, 1215–1228. ISSN: 00928674 (May 2015).
138. Daskivich, T. J. & Oh, W. K. Recent Progress in Hormonal Therapy for Advanced Prostate Cancer. en-US. *Current Opinion in Urology* **16**, 173–178. ISSN: 0963-0643 (May 2006).
139. Teng, M., Zhou, S., Cai, C., Lupien, M. & He, H. H. Pioneer of Prostate Cancer: Past, Present and the Future of FOXA1. en. *Protein & Cell* **12**, 29–38. ISSN: 1674-8018 (Jan. 2021).
140. Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. en. *New England Journal of Medicine* **373** (ed Longo, D. L.) 1541–1552. ISSN: 0028-4793, 1533-4406 (Oct. 2015).
141. Inaba, H., Greaves, M. & Mullighan, C. G. Acute Lymphoblastic Leukaemia. English. *The Lancet* **381**, 1943–1955. ISSN: 0140-6736, 1474-547X (June 2013).
142. Heikamp, E. B. & Pui, C.-H. Next-Generation Evaluation and Treatment of Pediatric Acute Lymphoblastic Leukemia. English. *The Journal of Pediatrics* **203**, 14–24.e2. ISSN: 0022-3476, 1090-123X (Dec. 2018).
143. Liu, G. J. *et al.* Pax5 Loss Imposes a Reversible Differentiation Block in B-Progenitor Acute Lymphoblastic Leukemia. en. *Genes & Development* **28**, 1337–1350. ISSN: 0890-9369, 1549-5477 (June 2014).
144. Dang, J. *et al.* PAX5 Is a Tumor Suppressor in Mouse Mutagenesis Models of Acute Lymphoblastic Leukemia. *Blood* **125**, 3609–3617. ISSN: 0006-4971 (June 2015).

145. Mullighan, C. G. *et al.* Genome-Wide Analysis of Genetic Alterations in Acute Lymphoblastic Leukaemia. en. *Nature* **446**, 758–764. ISSN: 1476-4687 (Apr. 2007).
146. Boller, S., Li, R. & Grosschedl, R. Defining B Cell Chromatin: Lessons from EBF1. en. *Trends in Genetics* **34**, 257–269. ISSN: 0168-9525 (Apr. 2018).
147. Nutt, S. L. & Kee, B. L. The Transcriptional Regulation of B Cell Lineage Commitment. en. *Immunity* **26**, 715–725. ISSN: 1074-7613 (June 2007).
148. Slany, R. K. MLL Fusion Proteins and Transcriptional Control. eng. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms* **1863**, 194503. ISSN: 1876-4320 (Mar. 2020).
149. Krivtsov, A. V. & Armstrong, S. A. MLL Translocations, Histone Modifications and Leukaemia Stem-Cell Development. en. *Nature Reviews Cancer* **7**, 823–833. ISSN: 1474-1768 (Nov. 2007).
150. Rao, R. C. & Dou, Y. Hijacked in Cancer: The KMT2 (MLL) Family of Methyltransferases. en. *Nature Reviews Cancer* **15**, 334–346. ISSN: 1474-1768 (June 2015).
151. Park, S. *et al.* The PHD3 Domain of MLL Acts as a CYP33-Regulated Switch between MLL-Mediated Activation and Repression. eng. *Biochemistry* **49**, 6576–6586. ISSN: 1520-4995 (Aug. 2010).
152. Li, Y. *et al.* Structural Basis for Activity Regulation of MLL Family Methyltransferases. en. *Nature* **530**, 447–452. ISSN: 1476-4687 (Feb. 2016).
153. Das, C. *et al.* Binding of the Histone Chaperone ASF1 to the CBP Bromodomain Promotes Histone Acetylation. eng. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E1072–1081. ISSN: 1091-6490 (Mar. 2014).
154. Mullighan, C. G. *et al.* Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia. en. **322**, 4 (2008).
155. Lee, S.-T. *et al.* A Global DNA Methylation and Gene Expression Analysis of Early Human B-Cell Development Reveals a Demethylation Signature and Transcription Factor Network. en. *Nucleic Acids Research* **40**, 11339–11351. ISSN: 0305-1048 (Dec. 2012).
156. Lee, S.-T. *et al.* Epigenetic Remodeling in B-Cell Acute Lymphoblastic Leukemia Occurs in Two Tracks and Employs Embryonic Stem Cell-like Signatures. en. *Nucleic Acids Research* **43**, 2590–2602. ISSN: 1362-4962, 0305-1048 (Mar. 2015).
157. Nordlund, J. *et al.* Genome-Wide Signatures of Differential DNA Methylation in Pediatric Acute Lymphoblastic Leukemia. *Genome Biology* **14**, r105. ISSN: 1474-760X (Sept. 2013).

158. Geng, H. *et al.* Integrative Epigenomic Analysis Identifies Biomarkers and Therapeutic Targets in Adult B-Acute Lymphoblastic Leukemia. *Cancer Discovery* **2**, 1004–1023 (Nov. 2012).
159. 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).
160. 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).
161. Williamson, I. *et al.* Developmentally Regulated *Shh* Expression Is Robust to TAD Perturbations. en. *Development* **146**, dev179523. ISSN: 0950-1991, 1477-9129 (Oct. 2019).
162. 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).
163. 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).
164. Li, Y. *et al.* Patterns of Somatic Structural Variation in Human Cancer Genomes. en. *Nature* **578**, 112–121. ISSN: 1476-4687 (Feb. 2020).
165. Iyyanki, T. *et al.* Subtype-Associated Epigenomic Landscape and 3D Genome Structure in Bladder Cancer. en. *Genome Biology* **22**, 105. ISSN: 1474-760X (Apr. 2021).
166. Mateo, L. J. *et al.* Visualizing DNA Folding and RNA in Embryos at Single-Cell Resolution. En. *Nature* **568**, 49. ISSN: 1476-4687 (Apr. 2019).
167. Bates, M., Jones, S. A. & Zhuang, X. Stochastic Optical Reconstruction Microscopy (STORM): A Method for Superresolution Fluorescence Imaging. en. *Cold Spring Harbor Protocols* **2013**, pdb.top075143. ISSN: 1940-3402, 1559-6095 (June 2013).
168. Zanoni, M. *et al.* Modeling Neoplastic Disease with Spheroids and Organoids. en. *Journal of Hematology & Oncology* **13**, 1–15. ISSN: 1756-8722 (Dec. 2020).
169. Wang, H., Han, M. & Qi, L. S. Engineering 3D Genome Organization. en. *Nature Reviews Genetics* **22**, 343–360. ISSN: 1471-0064 (June 2021).
170. Marusyk, A. & Polyak, K. Tumor Heterogeneity: Causes and Consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1805**, 105–117. ISSN: 0006-3002 (Print)\r0006-3002 (Linking) (Jan. 2010).

171. Sottoriva, A., Barnes, C. P. & Graham, T. A. Catch My Drift? Making Sense of Genomic Intra-Tumour Heterogeneity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. Evolutionary Principles - Heterogeneity in Cancer?* **1867**, 95–100. ISSN: 0304-419X (Apr. 2017).
172. Huang, S. Genetic and Non-Genetic Instability in Tumor Progression: Link between the Fitness Landscape and the Epigenetic Landscape of Cancer Cells. *Cancer and Metastasis Reviews* **32**, 423–448. ISSN: 0167-7659 (Dec. 2013).
173. McGranahan, N. & Swanton, C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* **27**, 15–26 (Jan. 2015).
174. Landau, D. A. & Wu, C. J. Chronic Lymphocytic Leukemia: Molecular Heterogeneity Revealed by High-Throughput Genomics. en. *Genome Medicine* **5**, 47. ISSN: 1756-994X (May 2013).
175. Ben-David, U. *et al.* Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response. en. *Nature*, 1. ISSN: 1476-4687 (Aug. 2018).
176. Peter, M. R. *et al.* Dynamics of the Cell-Free DNA Methylome of Metastatic Prostate Cancer during Androgen-Targeting Treatment. *Epigenomics* **12**, 1317–1332. ISSN: 1750-1911 (Aug. 2020).
177. Shen, S. Y. *et al.* Sensitive Tumour Detection and Classification Using Plasma Cell-Free DNA Methylomes. en. *Nature* **563**, 579–583. ISSN: 1476-4687 (Nov. 2018).
178. Nassiri, F. *et al.* Detection and Discrimination of Intracranial Tumors Using Plasma Cell-Free DNA Methylomes. en. *Nature Medicine* **26**, 1044–1047. ISSN: 1546-170X (July 2020).
179. Li, Y. *et al.* Clinical Implications of Genome-Wide DNA Methylation Studies in Acute Myeloid Leukemia. *Journal of Hematology & Oncology* **10**, 41–41 (Dec. 2017).
180. Kreso, A. & Dick, J. E. Evolution of the Cancer Stem Cell Model. *Cell Stem Cell* **14**, 275–291. ISSN: 1875-9777 (Electronic) (2014).
181. Shlush, L. I. *et al.* Identification of Pre-Leukaemic Haematopoietic Stem Cells in Acute Leukaemia. *Nature* **506**, 328–333. ISSN: 9788578110796 (Feb. 2014).
182. Shlush, L. I. *et al.* Tracing the Origins of Relapse in Acute Myeloid Leukaemia to Stem Cells. *Nature* **547**, 104–108 (June 2017).
183. Pogrebniak, K. L. & Curtis, C. Harnessing Tumor Evolution to Circumvent Resistance. en. *Trends in Genetics* **34**, 639–651. ISSN: 0168-9525 (Aug. 2018).

184. Maley, C. C. *et al.* Classifying the Evolutionary and Ecological Features of Neoplasms. en. *Nature Reviews Cancer* **17**, 605–619. ISSN: 1474-1768 (Oct. 2017).
185. Nowick, K. & Stubbs, L. Lineage-Specific Transcription Factors and the Evolution of Gene Regulatory Networks. *Briefings in Functional Genomics* **9**, 65–78. ISSN: 2041-2649 (Jan. 2010).
186. Peter, I. S. & Davidson, E. H. Evolution of Gene Regulatory Networks Controlling Body Plan Development. en. *Cell* **144**, 970–985. ISSN: 0092-8674 (Mar. 2011).
187. Domcke, S. *et al.* Competition between DNA Methylation and Transcription Factors Determines Binding of NRF1. en. *Nature* **528**, 575–579. ISSN: 1476-4687 (Dec. 2015).
188. Ghandi, M. *et al.* Next-Generation Characterization of the Cancer Cell Line Encyclopedia. en. *Nature* **569**, 503–508. ISSN: 1476-4687 (May 2019).
189. Ahmed, M. *et al.* CRISPRi Screens Reveal a DNA Methylation-Mediated 3D Genome Dependent Causal Mechanism in Prostate Cancer. en. *Nature Communications* **12**, 1781. ISSN: 2041-1723 (Mar. 2021).
190. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR Technologies for Precise Epigenome Editing. en. *Nature Cell Biology* **23**, 11–22. ISSN: 1476-4679 (Jan. 2021).
191. Pickar-Oliver, A. & Gersbach, C. A. The next Generation of CRISPR–Cas Technologies and Applications. en. *Nature Reviews Molecular Cell Biology* **20**, 490–507. ISSN: 1471-0080 (Aug. 2019).
192. 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).
193. 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).
194. 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).
195. 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).