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).

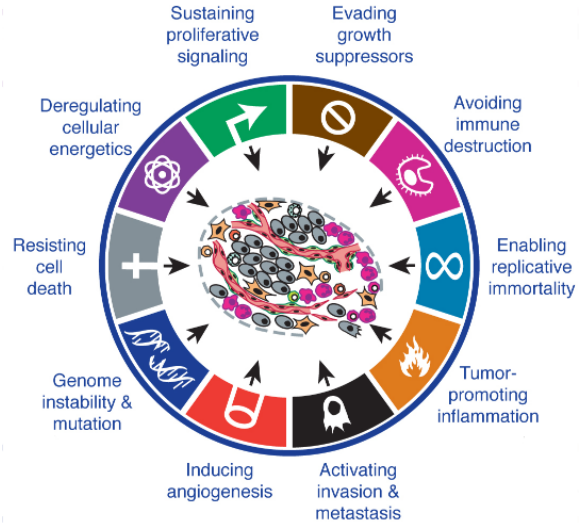

Figure 1.1: **The hallmarks of cancer.** Adapted from [REF 4].

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 to 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 (DNAme), 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 form 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

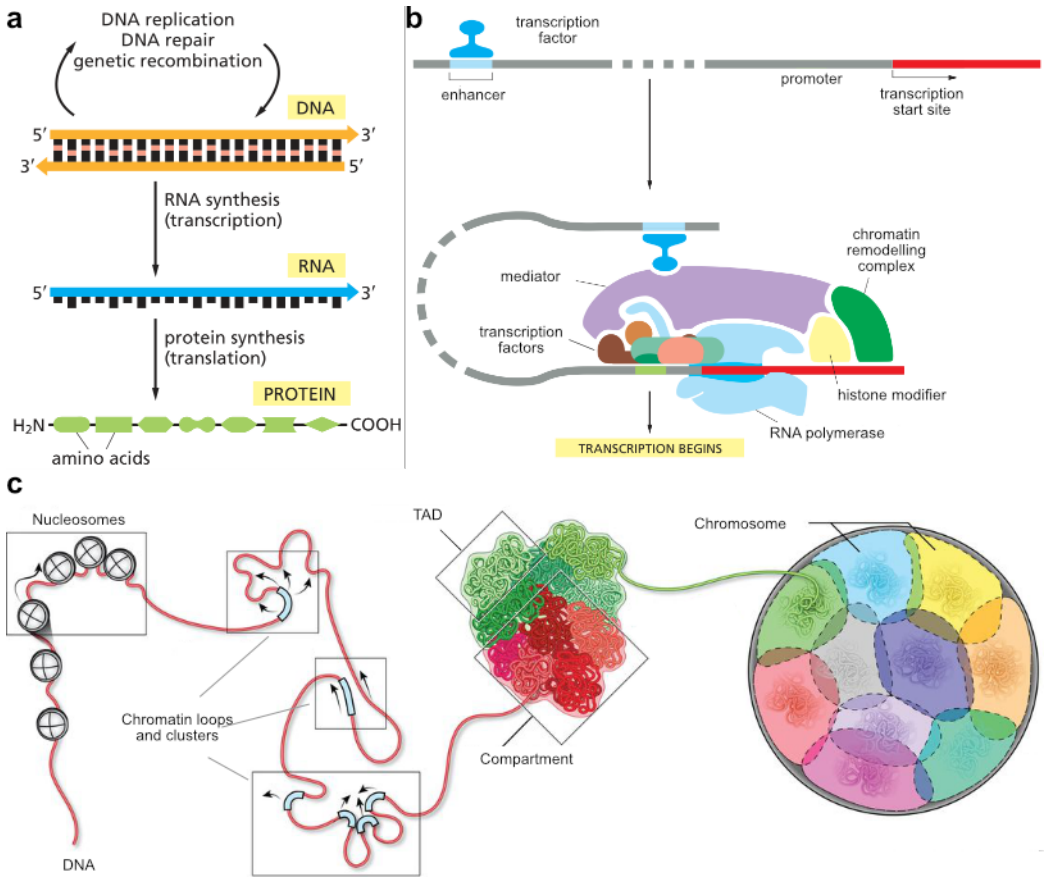

Figure 1.2: **The basics of gene expression inside the nucleus. a.** The central dogma of molecular biology. Adapted from [REF 15]. **b.** Schematic of the transcription machinery to initiate transcription. Adapted from [REF 15]. **c.** The scale of chromatin interactions across length scales. Adapted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

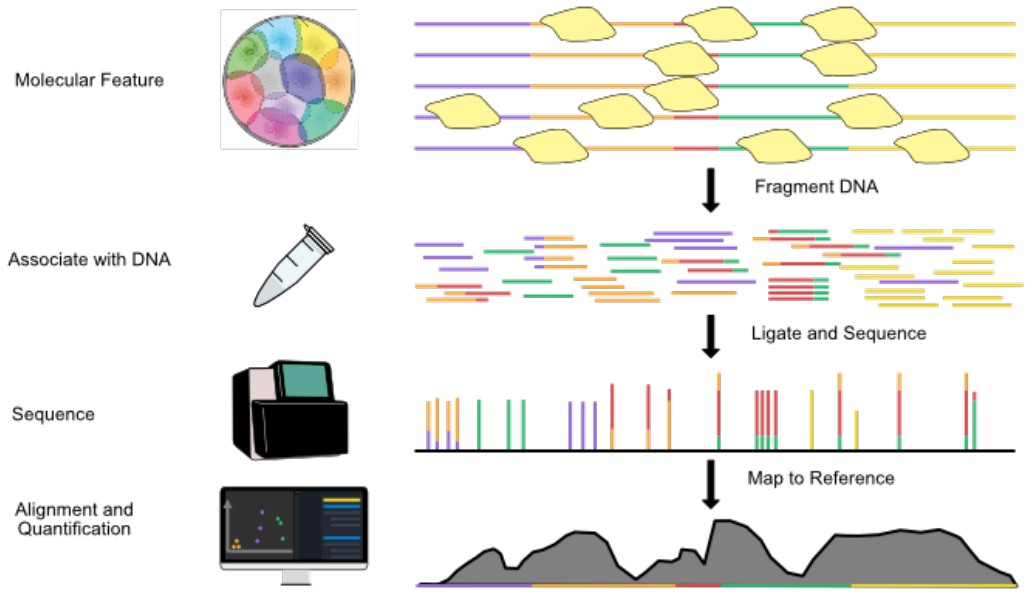

Figure 1.3: **Characterizing functional DNA elements with high throughput sequencing.**

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 micrococcal nuclease sequencing (MNase-seq) protocols, respectively [61–65]. DNAme 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 identify 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 DNAme 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 DNAme (hypomethylation) across the genome and focal increases of DNAme (hypermethylation) have been found across numerous cancers [110, 111]. Importantly, these changes in DNAme 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 DNAme profile [112]. Ependymomas have also been found to display distinct DNAme 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 myelodisplastic 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 DNAme to cancer phenotype are lacking, the impact of DNAme on TF binding has been well-demonstrated. Variable CTCF binding across human cell lines has been shown to vary with DNAme 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 DNAme on three-dimensional genome organization and TF binding, and genetic and non-genetic aberrations can be observed in chromatin contacts and histone modifications.

The affect of chromatin variants on on gene regulation extends beyond DNAme. 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, karyotpying 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 DNAme 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 Chapter 2 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 DNAme 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

Epigenetic dynamics underlying B cell acute lymphoblastic leukemia relapse

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

2.1 Abstract

Relapse of B-ALL remains a significant cause of death in treating the disease. Genomic investigations indicate that relapsed disease often arises from a minor clone of cells present at diagnosis. However, both genetic and epigenetic variation have been observed in B-ALL and other leukemias, and why come cells with particular genetic or epigenetic profiles survive therapy remains unknown. Here, we use targeted genome sequencing, RNA-seq, ATAC-seq, and bisulfite sequencing of patient-matched samples with patient-derived xenografts to investigate the dynamics of the genome and epigenome over B-ALL relapse. We find that DNA methylation profiles most closely resemble ge-

netic clones at diagnosis and relapse. Moreover, we find widespread increases to DNA methylation at relapse, mirroring a more stem-like phenotype. This work suggests that therapy selects for clones with stem-like characteristics, both genetically and epigenetically in B-ALL.

2.2 Introduction

After treatment, relapse of B-ALL occurs in 15 - 25 % of pediatric patients and 40 - 75 % of adult patients [141, 159]. Previous studies of the cells that give rise to primary and relapsed leukemias have identified newly acquired somatic mutations and CNVs targeting cell cycle regulation and B-cell development [154]. These cells predominantly appear to arise from a genetic subclone of cells present at diagnosis, while treatment primarily targets the dominant clone [154, 160–162]. But inactivating mutations are one of many ways in which genes regulating B-cell development, cell cycle, and differentiation can be activated or inactivated. Changes to DNAme play an important role in hematopoietic differentiation [155, 163], and chromatin accessibility signatures in hematopoietic stem and progenitor cells (HSPCs) distinguish phenotypically distinct cell types, even with minimal changes to gene expression patterns [164]. Notably, the binding of the TF CTCF mediates specific focal chromatin interactions that govern cell cycle and self-renewal capacity, and these bindings sites are sensitive to the presence of DNAme [31, 164]. This suggests that non-genetic components of chromatin, including its DNAme and accessibility, can influence B-ALL relapse. Moreover, interactions between genetic mutations and epigenetic aberrations have been observed in other leukemias, such as recurrent inactivating mutations in *TET2* [165–167], *IDH1* and *IDH2* [166, 168], and *DNMT3A* [169, 170], leading to disruption of DNAme genome-wide. In summary, to investigate the origins of B-ALL relapse requires multiomic profiling on diagnosis-relapse matched samples.

Previous studies of B-ALL relapse have primarily focused on genomic and transcriptomic assays [154, 162, 171]. Epigenetic studies of B-ALL relapse have primarily relied on enrichment-based assays or methylation arrays that have limited resolution genome-wide [156, 157, 172]. Further, fewer have investigated the role of chromatin accessibility in B-ALL oncogenesis or relapse [173]. To address the gaps left by these studies, we expand on previously published patient-derived xenografts (PDXs) from 5 patient-matched diagnosis (Dx) and relapse (Rel) samples, as well as relapse-fated genetic subclones that were present at diagnosis (termed disease relapse-initiating (dRI)) [171]. Using total RNA-seq, ATAC-seq for measuring chromatin accessibility, and bisulfite sequencing with DNA methylation capture sequencing (MeCapSeq), we investigate the genetic and epigenetic dynamics of B-ALL relapse.

2.3 Results

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

To investigate the molecular landscape of B-ALL relapse, we profiled gene expression, chromatin accessibility, and DNAme of 3 adult and 2 pediatric B-ALL patients at both Dx and Rel with bulk RNA-seq, ATAC-seq, and MeCapSeq, respectively (Figure 2.1a). These patients' tumours contained $\geq$ 90 % leukemic blasts at diagnosis and were previously profiled using whole exome sequencing (WES) to identify the mutation burden of leukemic driver mutations [171] (see Table D.1; patient numbers used in this study match those from [REF 171]). Matching mutation profiles between Dx and Rel samples allowed for the identification of dRI samples, which are cells present at diagnosis that harbour mutations found at relapse, indicating that these cells are relapse-fated. Comprehensive datasets containing RNA-seq, ATAC-seq, and MeCapSeq were produced for 3 patients, with 2 patients lacking RNA-seq data due to source constraints. While expression, chromatin accessibility, and DNAme are each critical for determining cell phenotype and its role in relapse, we sought to investigate the importance of each dataset in an agnostic manner. To achieve this, similarity scores were calculated between all samples using SNF [123]. For each patient, similarity scores between all samples derived from that patient (both primary and PDX) were calculated, and weighted graphs to cluster samples together were constructed (see Section 2.5.5). This was done for each individual data type, as well as for a fused network comprised of information by considering all data types simultaneously. To determine the importance of each data type, samples were labelled by their disease stage (Dx, dRI, or Rel; Figure 2.1b). For all 3 patients with complete molecular datasets, the combined networks clustered samples based on disease stage more clearly than each individual dataset (Figure 2.1b). This suggests that disease stages can be more clearly identified from multiple molecular components together than a single component alone [123]. The graphs produced from DNAme data more clearly cluster samples by disease stage than gene expression or chromatin accessibility across all patients, suggesting that DNAme may be a clearer marker of relapse. Taken together, we find that B-ALL disease stage can be identified through non-genetic molecular measurements and that DNAme is mostly closely linked to relapse than gene expression and chromatin accessibility.

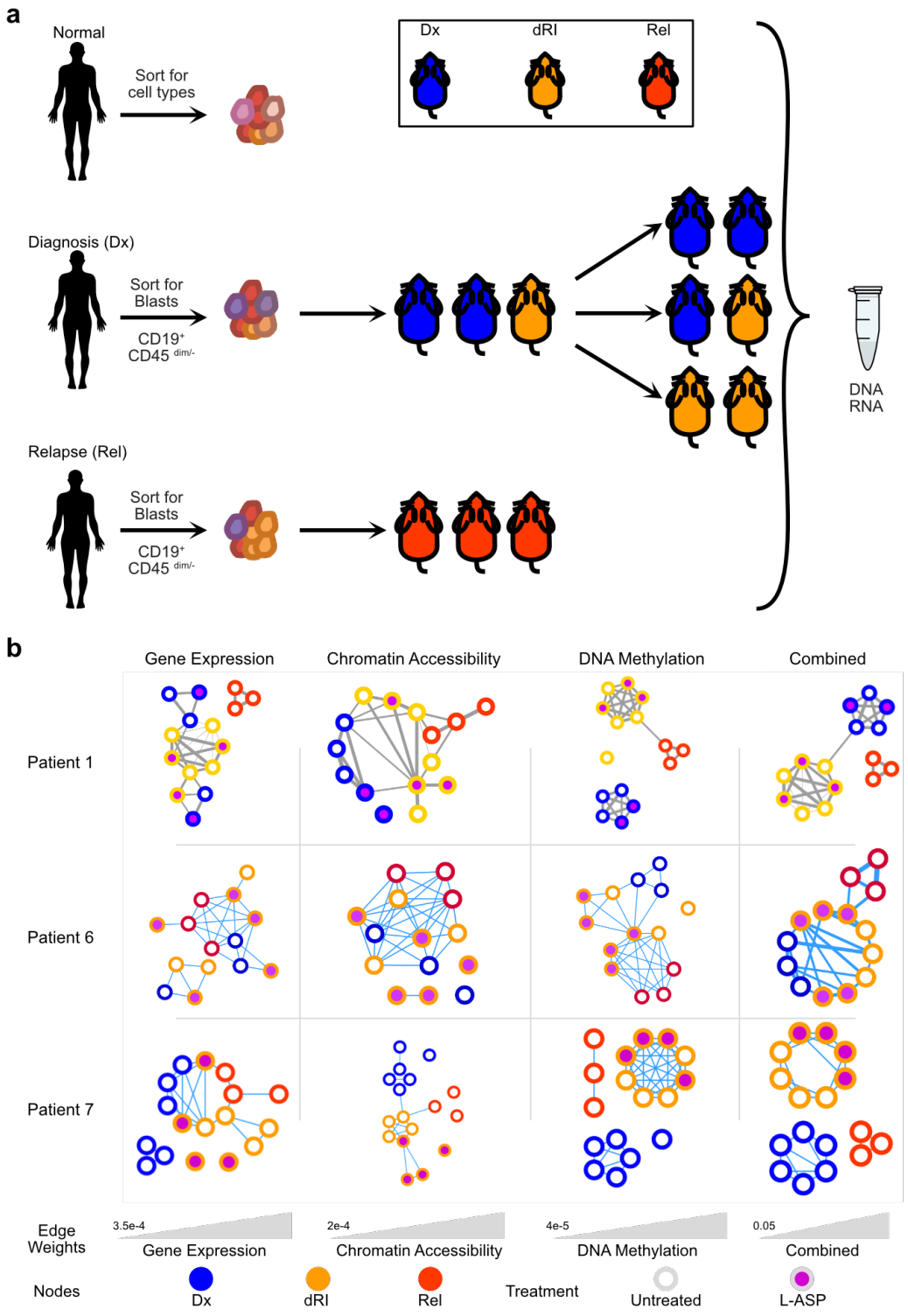

Figure 2.1: **Experimental design and data integration. a.** Experimental design of samples used in this study. Normal samples were obtained from cord blood pools and sorted into various hematopoietic cell types. B-ALL patients who experienced relapse has sorted leukemic blasts collected at Dx and Rel. Based on the mutation profiles from [REF 171] some Dx samples are labelled as dRI. **b.** Individual and fused networks of samples from three patients with complete multiomic profiling. Nodes represent individual samples (either primary or PDX), edges represent similarities between the connected samples.

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

Given the strong correlation between DNAme signal and disease state, we focused on DNAme changes over B-ALL relapse. To understand the dynamic changes to DNAme that happen over the course of B-ALL relapse, we first looked to the hematopoietic hierarchy and DNAme changes over normal B-cell differentiation. Using normal cord blood pools, sorted into B-cells and multiple B-progenitor cell types, we performed MeCapSeq on 8 pools separated into 4 cell types: hematopoietic stem cells (HSCs) and multi-potent progenitors (MPPs); lymphoid-primed multi-potent progenitors (LMPPs) and monocyte-lymphoid progenitors (MLPs); early progenitor B cells (EarlyProBs), pre-progenitor B cells (PreProBs), and progenitor B cells (ProBs) (collectively labelled as ProB); and B-cells (Figure 2.2a; see Table 2.1). Using pairwise comparisons between these cell types, we identified 540 DMRs over the course of B-cell differentiation from HSCs (Figure 2.2b). Significant changes to DNAme occurred in 62 regions from HSC-MPP to LMPP-MLP, 312 regions from LMPP-MLP to ProB, and 166 regions from ProB to fully differentiated B-cells (Figure 2.2c). While roughly equal numbers of loci gained and lost DNAme in the transition from HSCs-MPPs to LMPPs-MLPs, after lymphoid commitment, nearly all regions lost DNAme in later differentiation transitions (Figure B.2c). Overall, 500 (92.6 %) of DMRs identified were loci that became hypomethylated over differentiation. These changes are in agreement with earlier studies profiling DNAme changes over B-cell differentiation using the Illumina 450K arrays [155–157], and provide an expanded set of DMRs with which to track differentiation. Notably, no DMR identified in an earlier transition was found as differentially methylated in a later transition. Regions with altered DNAme in one cell type persisted for all downstream cell type transitions. This suggests that DNAme at these loci can be used as a marker of differentiation. In summary, we find that normal HSCs permanently change DNAme over the course of differentiation, predominantly by losing DNAme.

2.3.3 Recurrent DNA methylation changes identify stem cell pathways in relapse

With the predominant loss of DNAme established in the normal setting, we identified DMRs between Dx and Rel primary and PDX B-ALL samples. When considering all patients together and grouping by disease stage, we found no DMRs remained statistically significant after multiple testing corrections. This result conflicted with previous observations about DNAme changes in B-ALL relapse [156, 157] as well as the earlier SNF analysis. We hypothesized that DNAme changes

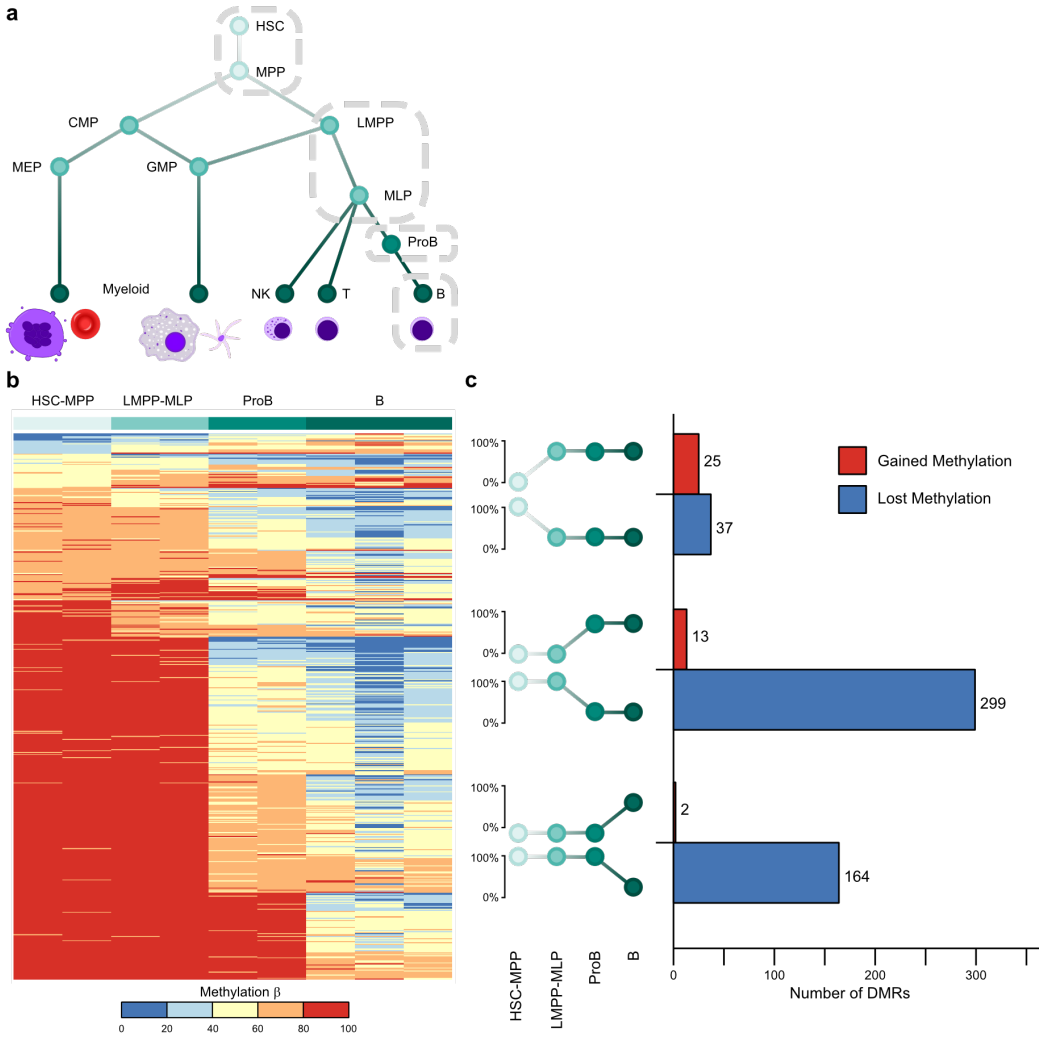

Figure 2.2: **Widespread loss of DNA methylation over B-cell differentiation. a.** Schematic of the hematopoietic hierarchy and the grouping of B-cell progenitors into the groups isolated in this study. **b.** Heatmap of DMRs identified between B lineage cell types. Columns are samples ordered by cell type and rows are DMRs identified in at least one pairwise comparison between cell types (dmrseq, FDR < 0.1). **c.** Bar plot of DMRs classified by which step in differentiation they were identified as significantly changed.

in across patients was heterogeneous, which limited the ability to detect significant changes. Using a patient-oriented approach, we identified DMRs between Dx and Rel for each patient, separately, to track changes over each patient's relapse trajectory. This identified 25 761 DMRs across the cohort of (range 98 - 15 296, median 7 426, $\delta\beta \geq 20$ %, FDR < 0.1, Figure 2.3a). Unlike the process of normal differentiation, most DMRs were hypermethylated at relapse (Figure 2.2b). 18 610 (72.2 %) DMRs were specific to a single patient and did not overlap DMRs from others (Figure 2.3b, left), as expected from the lack of significant DMRs from the cohort-oriented analysis. Notably, the 9 recurrently DMRs in all 5 patients are all in the promoter regions of the following genes: *BARHL2*, *CYP26B1*, *EBF3*, *EN2*, *GDNF*, *HMGA2*, *NKX2-2*, *NR2F2*, and *PAX6*. Using gene ontology (GO) analysis, we find that these genes with nearby recurrent differential methylation are positively associated with differentiation, with the most statistically significant pathway being cell fate determination (Figure 2.3c). For these genes, we find that the promoter regions become hypermethylated at B-ALL relapse (Figure 2.3d). Some genes, like *CYP26B1*, have multiple short DMRs in the promoter and one gains DNAme while the other loses DNAme at relapse, but all these promoters gain DNAme overall. Given the association between hypermethylation in promoter regions and decreased expression [174], these results suggest that these genes are under-expressed at relapse. Taken together, we find that the changes to DNAme over the course of B-ALL relapse is antithetical to the changes seen over normal B-cell differentiation, and that recurrent DNAme changes suggest that B-ALL relapse reverts to a more de-differentiated, stem-like DNAme state.

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

Relapse-fated subpopulations of cells present at diagnosis were detected in these patients by their mutations [171]. Yet some Dx samples harboured similar DNAme profiles to the Rel samples (e.g. column 2 for Patient 7 and column 6 for Patient 9 in Figure 2.3a). We hypothesized whether these same populations could be detected by their DNAme profile at diagnosis. By identifying DMRs across all three disease stages (Dx, dRI, and Rel), we identified a median of 2 784 DMRs between disease stages across each patient (range 376 - 4 098; Figure 2.4). There is heterogeneity in DNAme profiles across samples derived from the same patient, and even within the same disease stage (Figure 2.4a). The heterogeneity within disease stages resulted identifying DMRs specific to the dRI samples that were shared between Dx and Rel samples, even when some dRI samples showed similar methylation rates (e.g. leftmost dRI sample for Patient 4). Based on which disease

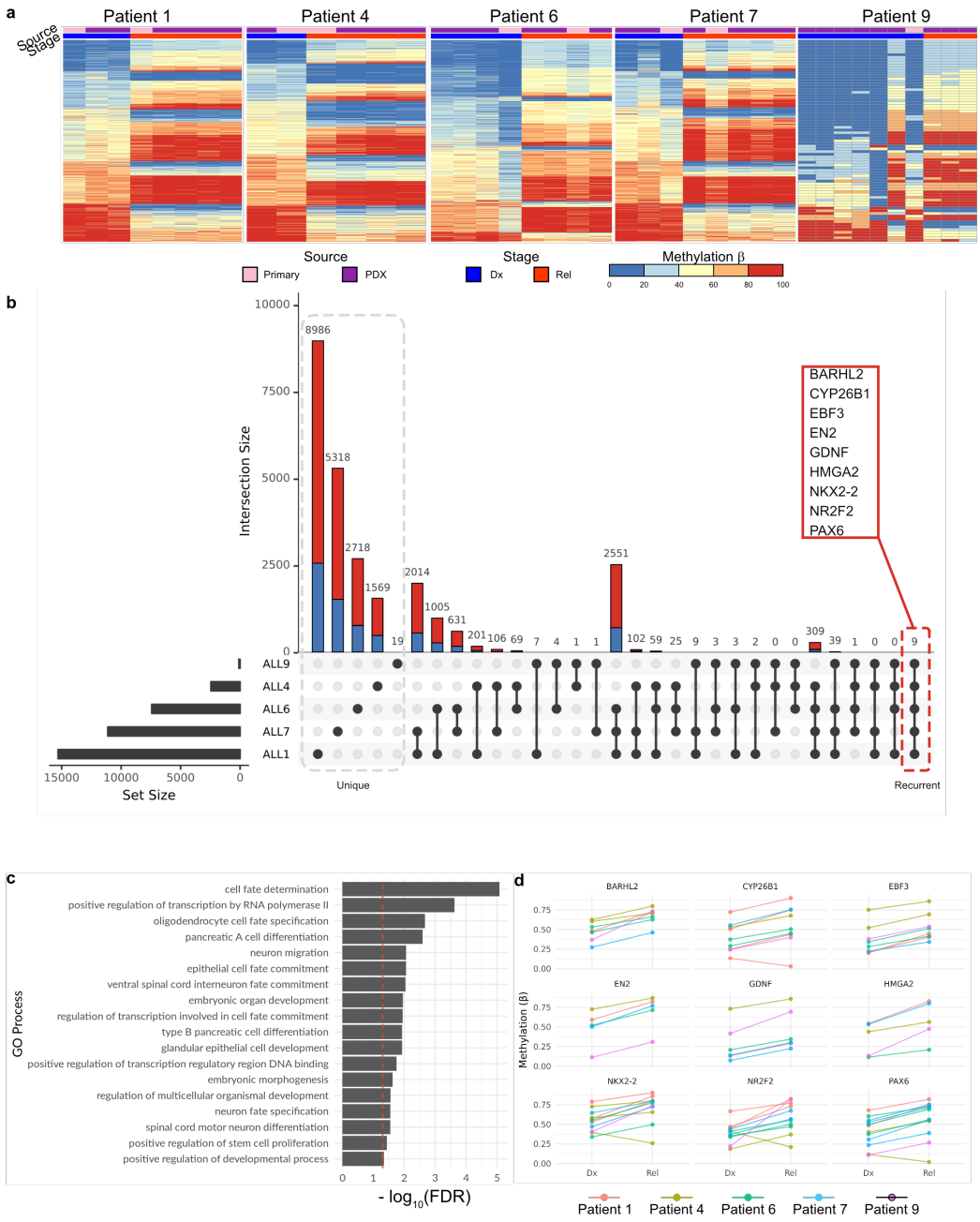

Figure 2.3: **Recurrent relapse DMRs are associated with cell fate decision processes. a.** Heatmaps of DMRs identified between Dx and Rel samples within each patient. **b.** Upset plot showing the shared DMRs between patients. DMRs in the left highlighted block are unique to a single patient, whereas DMRs in the right highlighted block are recurrent changes across all 5 relapse patients. These DMRs are in the promoter regions of the callout genes listed. **c.** GO analysis of genes with recurrently hypermethylated promoters in Rel B-ALL samples. The red dashed line indicates the FDR threshold of 0.05. **d.** Pairwise DNAme changes in each patient at the recurrently hypermethylated loci show increased methylation in all patients.

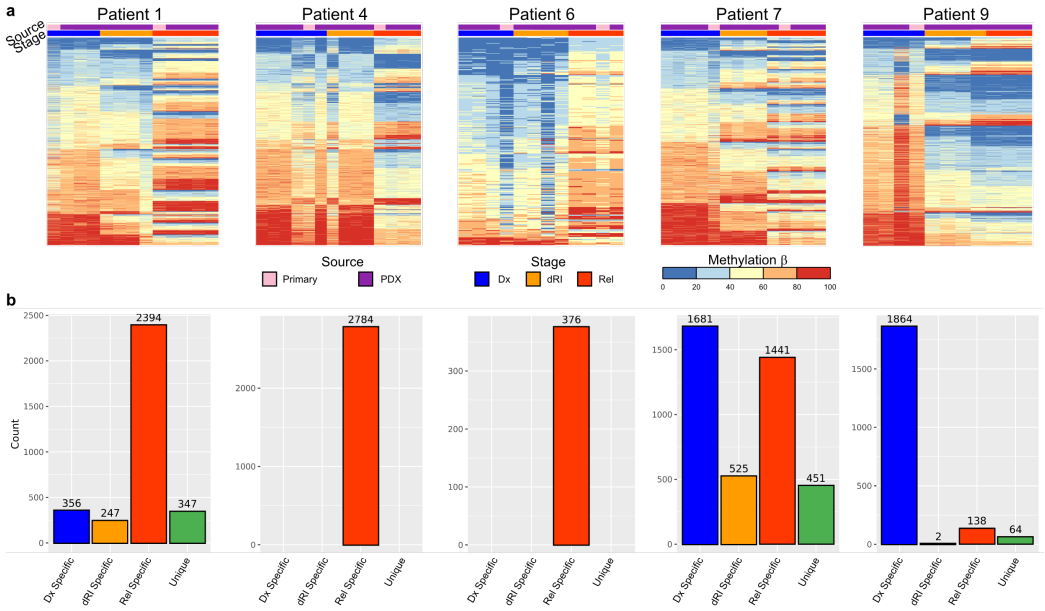

Figure 2.4: **Subpopulations present at diagnosis can harbour relapse-like DNAme profiles. a.** Heatmaps of expanded DMRs identified in dRI PDX samples. **b.** Bar plot showing the number of DMRs classified by which disease stage it is specific to. "Dx Specific" DMRs have shared DNAme between dRI and Rel samples. "DRI Specific" DMRs have shared DNAme between Dx and Rel samples. "Rel Specific" DMRs have shared DNAme between Dx and dRI samples. "Unique" DMRs are regions that have significantly different DNAme at each stage.

stage the DMRs were identified in, each glsdmr was classified as Dx-specific (shared between dRI and Rel), dRI-specific (shared between Dx and Rel), Rel-specific (shared between Dx and dRI), or unique (significantly differentially methylated in all three stages). All patients harboured Rel-specific DMRs, and a majority of DMRs in total were detected in Patients 1, 4, and 6 (range 71.6 %, 100 %, and 100 %, respectively; Figure 2.4b). For these three patients, a majority of the relapse-fated cells shared the DNAme profile of the neighbouring cells, suggesting that these DNAme changes occurred after mutation. Patients 1, 7, and 9, dRI-specific DMRs were found, suggesting that the DNAme profile of cells at diagnosis is not necessarily linked to their mutation status. Further, 41.1 % and 90.1 % of DMRs were found to be Dx-specific for Patients 7 and 9, respectively (Figure 2.4b). Patient 1 also harboured 356 (10.6 %) Dx-specific DMRs. This suggests that some DNAme changes are linked to the mutation status of relapse-fated cells. Taken together, these results suggest that relapse-fated DNAme profiles can be detected at diagnosis, but that the trajectory of DNAme changes over the course of relapse is heterogeneous across patients.

2.4 Discussion

Disease relapse remains a major barrier in treating B-ALL [140, 159, 175]. While the genetic origins of relapse have been characterized, epigenetic aberrations underlying relapse have been less well-studied. In this work, we investigated the epigenetic and transcriptomic changes of 5 B-ALL patients over the course of relapse to identify non-genetic changes in tumours that may lead to relapse. DNAme is more highly correlated with disease stage than RNA or chromatin accessibility and changes to DNAme are antithetical to DNAme changes seen in normal B-cell differentiation. While most DNAme changes are patient-specific, a small number of recurrent changes indicate a more stem-like state at relapse. In some cases, these stem-like DNAme profiles are present at diagnosis, indicating that subclones defined by DNAme may also contribute to B-ALL relapse.

Both genetic and epigenetic aberrations in tumours play important roles in determining disease relapse [156, 157]. Previous reports highlight the frequency that epigenetic regulators are mutated in B-ALL [162] and leukemias more generally, such as *DNMT3A*, *TET2*, *IDH1*, and *IDH2* in AML [176–178] and *CHD2*, *HIST1H1E*, and *ZMYM3* in chronic lymphocytic leukemia (CLL) [179–181]. These findings demonstrate that epigenetic modifications, in conjunction with genetic aberrations, discriminate disease outcomes and can share an evolutionary trajectory in cancer. However, it remains unclear why DNAme changes in B-ALL patients remain mostly patient-specific. One possibility is that the DNAme profile is linked to the genetic profile of the tumour and that this genetic predisposition influences how DNAme changes. Each of the five patients here harbour different genetic mutations, defining different subtypes of B-ALL. While some genetic subtypes are shared between patients (e.g. Patients 1 and 7 both belong to the DUX4 subtype and share > 4000 DMRs), the lack of large sample sizes with a common genetic subtype may confound this relationship. Another possibility is that the selective pressure on cells caused by therapy induces divergent DNAme patterns in a similar fashion to increased mutation rates in cancers after treatment [182]. Stochastic exploration of fitness landscapes through DNAme changes may lead to the type of DMRs observed here; a select few DMRs converge on similar biological pathways to evade therapy surrounded by hundreds or thousands of passenger DMRs that have no effect. Distinguishing between these processes would require genetically identical models to separate the effect of genetic profiles on epigenetic dynamics.

However, it is not the case that genetic and epigenetic states always behave similarly. In this study we found both Dx and dRI PDXs samples that share DNAme profiles with the Rel tumours, suggesting that DNAme states can vary independently of mutations. Moreover, the differences between PDX methylomes derived from the same primary sample demonstrates that subpopulations

of cells can have differing DNAme states while sharing mutations. This decoupling between genome and epigenome has been observed in other tumours, such as pediatric ependymomas, where recurrent DNAme profiles were found in the absence of recurrent mutations and was associated with outcome [113, 183], and glioblastoma, where stem cells are characterized by widespread changes in chromatin accessibility [184] and histone modifications [185]. These studies highlight the role of epigenetic plasticity and intra-tumour heterogeneity in cancers [5]. With similar results found in leukemias that are linked to disease outcome [186–190], it is likely that epigenetic plasticity and heterogeneity are also key factors in therapeutic response and relapse. Taken together, these results suggest that the epigenome can provide mechanisms independent of genetic aberrations, to respond and adapt to therapies, but are often guided by genetic aberrations. This complexity of disease response will need to be addressed to design treatment regimens for patients with an increased propensity towards relapse.

Previous investigations of DNAme aberrations in B-ALL have primarily focused on a select few genes, or single CG dinucleotides (CpGs) in promoter regions [156, 157, 191, 192]. While the recurrent DMRs in this study were found in these same regions, most DMRs were identified in intergenic regions. This suggests that important changes in the epigenetic landscape is currently unidentified, and future studies investigating DNAme aberrations in B-ALL should prioritize genome-wide approaches. The phenotypic impact of focal hypermethylation on engraftment and self-renewal capacity has not been assessed here, so experiments should be conducted to validate these findings (this is a bad sentence but this idea is important). For patients undergoing B-ALL treatment, DNAme has the potential to be used as early indicators of relapse. Moreover, treatment with DNA demethylating agents, such as 5-aza-citidine and 5-aza-2'-deoxycytadine, may be effective at preventing relapse. These treatments have been approved for use in patients with MDS and AML in adult populations and early clinical trials have demonstrated their safety [193, 194], although some toxic effects have been identified in drug combination trials [195]. Taken together, therapeutic targeting of DNAme may be an effective method to prevent B-ALL relapse by preventing the outgrowth of stem-like subpopulations that survive chemotherapy.

2.5 Methods

2.5.1 Patient selection and sample collection

Patient samples were obtained at diagnosis and relapse from patients with B-ALL as previously described [171]. All samples were frozen viably and stored long term at -150 °C. Samples were selected retrospectively based on paired-sample availability.

Human cord blood samples were obtained with informed consent from Trillium and Credit Valley Hospital according to procedures approved by the University Health Network Research Ethics Board, as previously described [171]. Cells were stained with the following antibodies (all from BD Biosciences, unless otherwise stated):

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

Cells were sorted from cord blood cells on the basis of markers listed in Table 2.1, as previously described [196], on a FACSAria III (Becton Dickinson), consistently yielding > 95 % purity.

Table 2.1: **Cell surface markers used to isolate cell populations from cord blood pools.**

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

2.5.2 Patient-derived xenograft generation and limiting dilution assays

PDXs were generated as previously described [171]. Clinical samples were stained with the following antibodies:

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

Each sample was sorted on a FACSAria III (BD Biosciences) for leukemic blasts ($CD19^{+}CD45^{dim/-}$) and T cells ($CD3^{+}CD45^{hi}$). NOD scid gamma (NSG) mice were bred according to protocols established and approved by the Animal Care Committee at the University Health Network. 8-to-12-week-old mice were sublethally irradiated at 225 cGy 24 hours prior to transplants. Only female mice were used. Intra-femoral injections of 10 to 250 000 sorted leukemic blasts were performed as previously described [197]. Mice were sacrificed 20-to-30 weeks post-transplant or at the onset of disease symptoms. Human cell engraftment in the injected femur, bone marrow (non-injected bones, left tibia, right tibia, left femur), spleen, and central nervous system were assessed using human-specific antibodies for CD45 (PE-Cy7, BD Biosciences, clone HI30; v500 BD Biosciences, clone HI30), CD44 (PE, BD Biosciences, clone 515; FITC, BD Biosciences, clone L178), CD3 (APC, BD Biosciences, clone UCHT1), CD19 (PE-Cy5, Beckman Coulter, clone J3-119), CD33 (PE-Cy7,

BD Biosciences, clone P67-6; APC, BD Biosciences, clone P67-6), and CD34 (APC-Cy7, BD Biosciences, clone 581) analyzed on an LSRII (BD Biosciences). Mice were considered to be engrafted when > 0.1 % of cells in the injected femur were positive for one or more human B-ALL-specific cell surface marker (CD45, CD44, CD19, and CD34). Confidence intervals for the frequency of leukemia initiating cells was calculated using ELDA [198].

2.5.3 Human cell isolation from patient-derived xenografts

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

2.5.4 Primary and patient-derived xenograft sample sequencing

RNA sequencing

RNA-seq was performed as previously described [171]. Briefly, amplified complementary DNA (cDNA) was sequenced as paired-end libraries on an Illumina HiSeq2000. The libraries were sequenced as 2×75 bp for the adult and 2×100 bp for the pediatric samples.

DNA methylation capture sequencing

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

Assay for transposase-accessible chromatin sequencing

Library preparation for ATAC-seq was performed with the Nextera DNA Sample Preparation Kit (FC-121-1030, Illumina), according to a previously reported protocol [62]. ATAC-seq libraries were sequenced with an Illumina HiSeq 2500 sequencer to generate single-end 50 bp reads.

2.5.5 Sequencing data analysis

Differential gene expression analysis

The methods are described in [REF 171]. Briefly, RNA-seq reads were aligned against the GRCh38 reference human genome with STAR (v2.5.2b) [199] and annotated with the Ensembl reference (v90). Default parameter were used with the following exceptions: chimeric segments were screened with a minimum size of 12 bp, junction overlap of 12 bp, and maximum segment reads gap of 3 bp; splice junction overlap of 10 bp; maximum gap between aligned mates of 100 000 bp; maximum aligned intron of 100 000; and alignSJstitchMismatchNmax of 5 1 5 5. Transcript counts were obtained with HTSeq (v0.7.2) [200]. Data was library size normalized using the RLE method, followed by a variance stabilizing transformation using DESeq2 (v1.22.1) [201]. Principal component analysis plots were generated on a per sample basis using the top 1 000 variable genes. For downstream analysis, the mean expression of each sample clone condition was used. For per-patient analyses, differentially expressed genes were identified between disease stage and clone status using DESeq2. Genes with an FDR < 0.05 and absolute $\log_2$(fold change) > 1 were considered significant.

Identification of accessible chromatin peaks

ATAC-seq reads were aligned against the GRCh38 reference human genome with Bowtie2 (v2.0.5) [202] with default parameters. Accessible peaks were identified with MACS2 (v2.0.10) [203] with the following command:

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

A catalogue of peaks from all samples was collected with a custom R script. ATAC-seq signal was mapped from each sample to this catalogue using Bedtools [204] for downstream analysis.

Bisulfite sequencing pre-processing

Sequencing read qualities were assessed with FastQC (v0.11.8) [205]. Low quality bases were trimmed with Trim Galore! (v0.6.3) [206] with the following command:

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

Trimmed reads were aligned to the GRCh38 reference human genome with Bismark (v0.22.1) [207] with default parameters. Duplicates were removed from the resulting alignment file with the following command:

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

The deduplicated BAM file was sorted by position with sambamba (v0.7.0) [208]. M-biases were calculated with MethylDackel (v0.4.0) [209], and methylation β values were extracted from the BAM files with the following command:

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

Both M and β values were for each CpG were used in downstream analyses.

Similarity network fusion

Preprocessed data from each sample was collected with the following features: normalized gene expression abundance for all genes, chromatin accessibility signal within previously identified accessible peaks, and mean β value for all CpGs listed in the manifest for targeted bisulfite sequencing kit. These features and sample labels were processed with the SNFtool R package [123] to perform the similarity network fusion analysis. Graphs were constructed for all samples deriving from a single patient where each node is a sample and each edge is weighted according to the determined similarity between the samples. Edges whose weights were below specific thresholds were removed from the graph. The threshold weight for the fused graph was 0.05. Similar graphs were constructed using the individual components for each sample (e.g. using just the similarity in RNA-seq data), and the component graphs were compared to the fused graph, to compare the importance of each feature. Threshold weights for these individual graphs were determined to be 6×10^{-5} for DNAme, 4×10^{-4} for gene expression, and 2×10^{-4} for chromatin accessibility.

Differentially methylated region identification

DMRs were identified using the dmrseq R package (v1.3.8) [210] with an absolute filtering cutoff value of 0.05 and using the sequencing batch as an adjustment covariate. Normal samples from all donors were compared pairwise based on their sorted cell type. B-ALL samples were compared by their designated disease stage (Dx, DRI, or Rel), and were compared both across all patients (e.g. all Dx samples against all Rel samples), or within a single patient (e.g. all Dx samples from Patient 1 against all Rel samples from Patient 1). A multiple testing correction with the FDR method was performed [211]. Regions with an FDR < 0.1 were determined to be significant.

Gene ontology enrichment analysis

Gene ontology enrichment analysis was performed using the PANTHER classification system (database version 2019-10-08) [212]. Gene symbols for the genes whose promoter regions contained the recurrently hyper-methylated regions in all B-ALL patient samples were supplied, with the entire human genome as the background. An over-representation Fisher test for biological processes was performed with an FDR correction. Biological processes at the top of the hierarchy with an FDR < 0.05 were determined to be significant.

Appendix A

Supplementary Material for Chapter 2

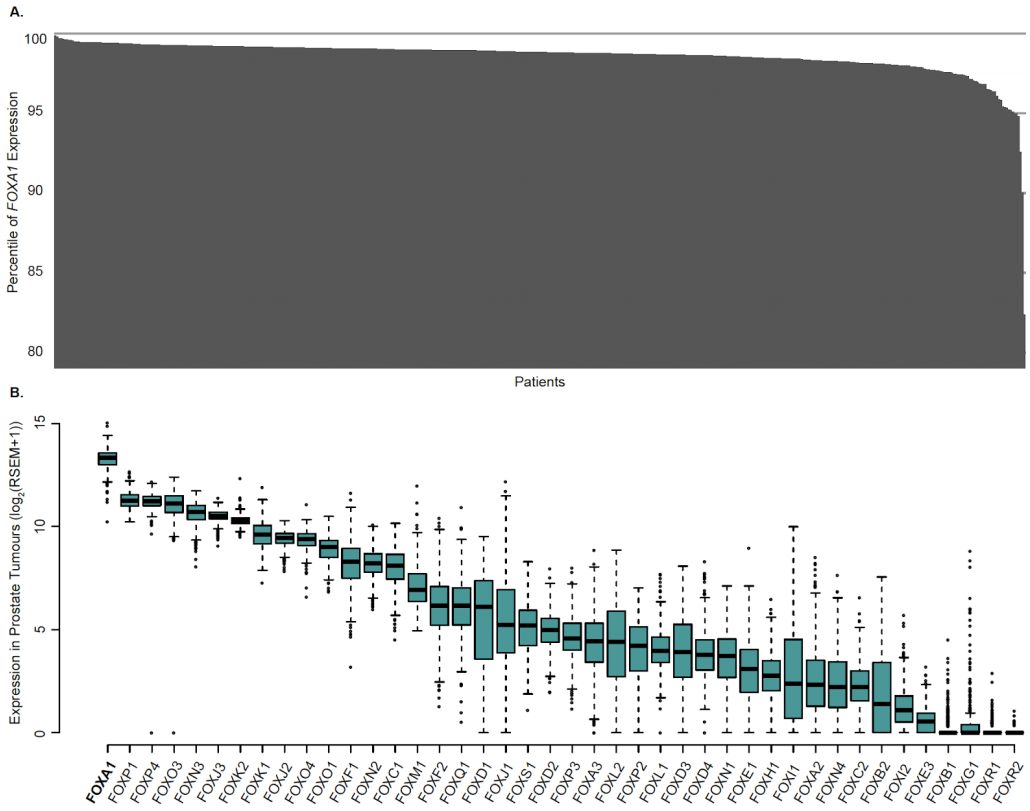

Figure A.1: ***FOXA1* mRNA expression in prostate tumours. a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

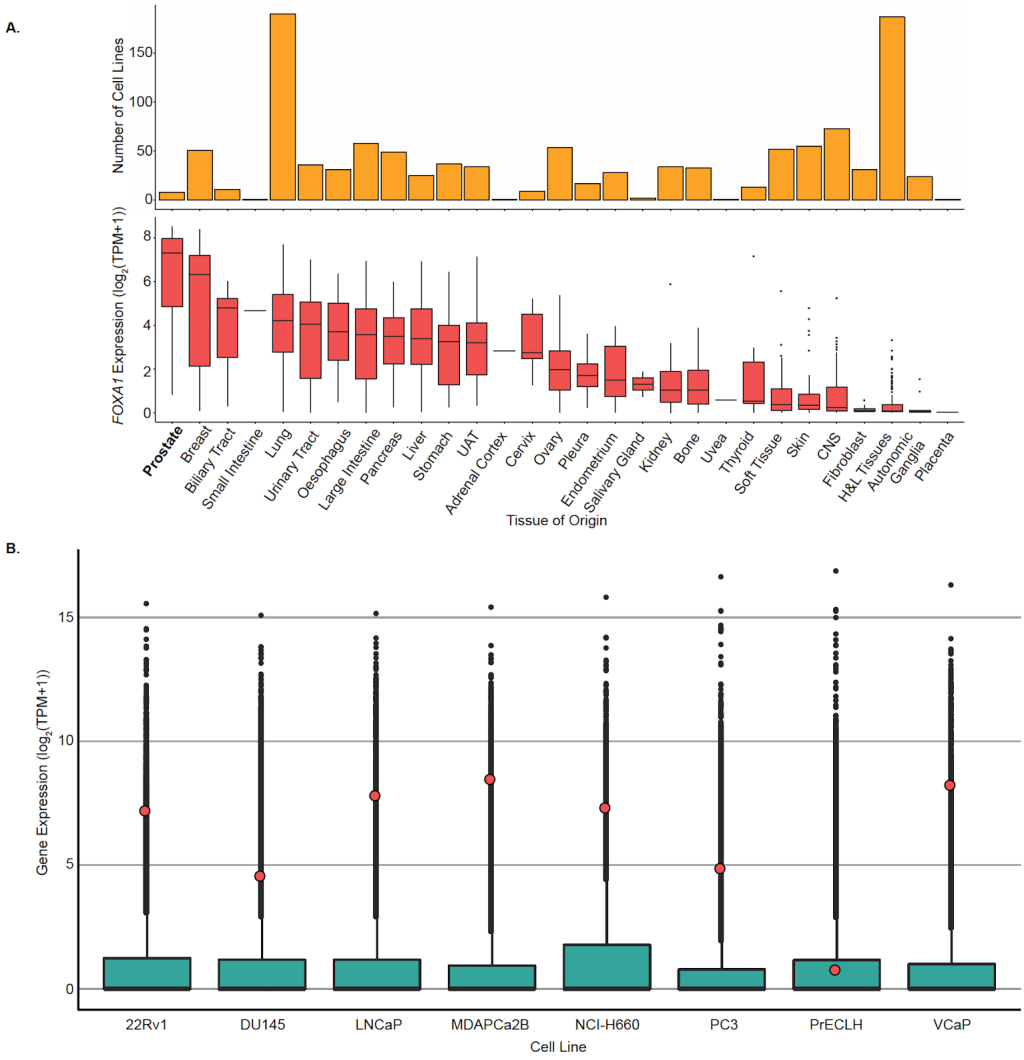

Figure A.2: ***FOXA1* mRNA expression across PCa cell lines. a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Methods). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA expression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

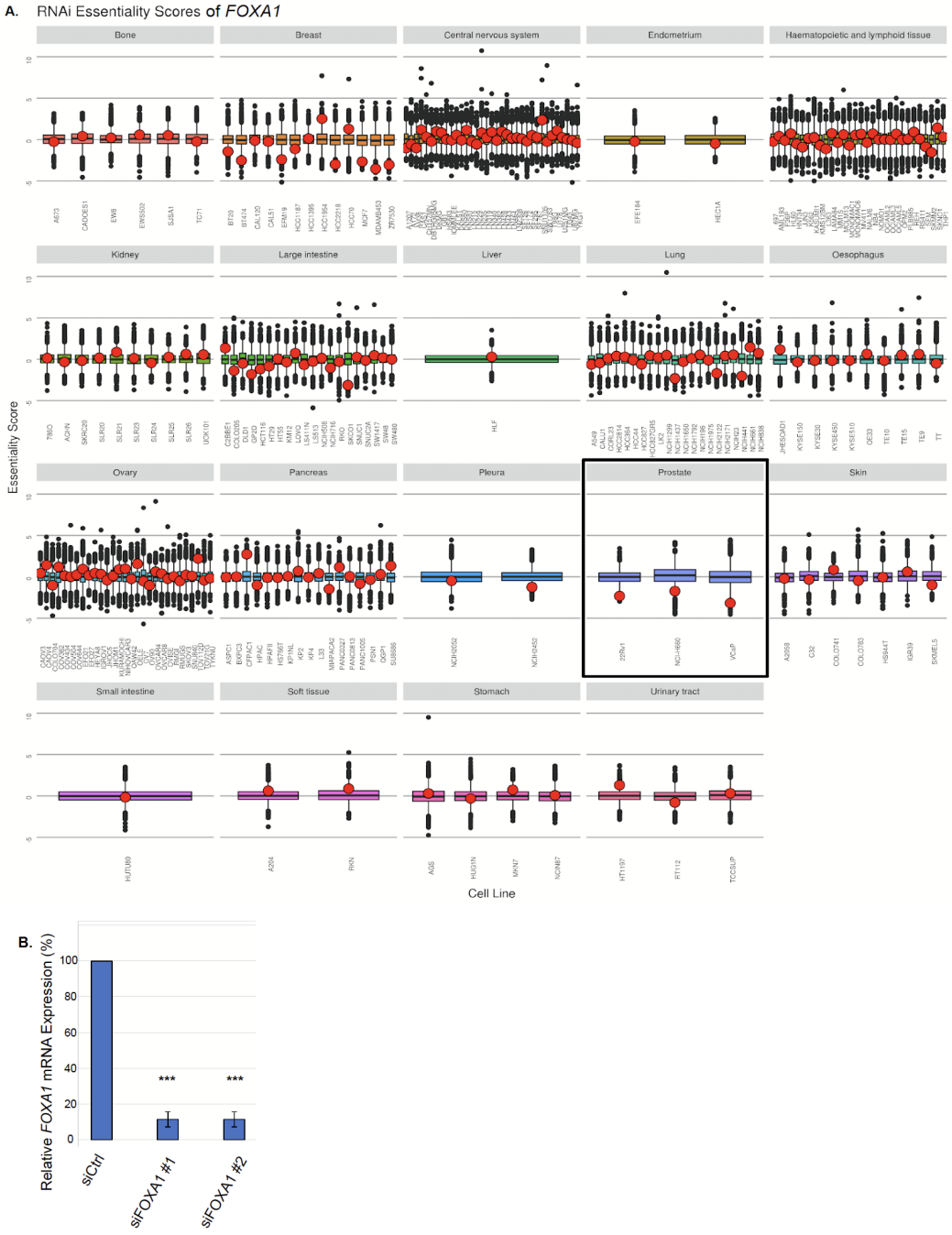

Figure A.3: **Essentiality of *FOXA1* across cancer cell lines of various cancer types. a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot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

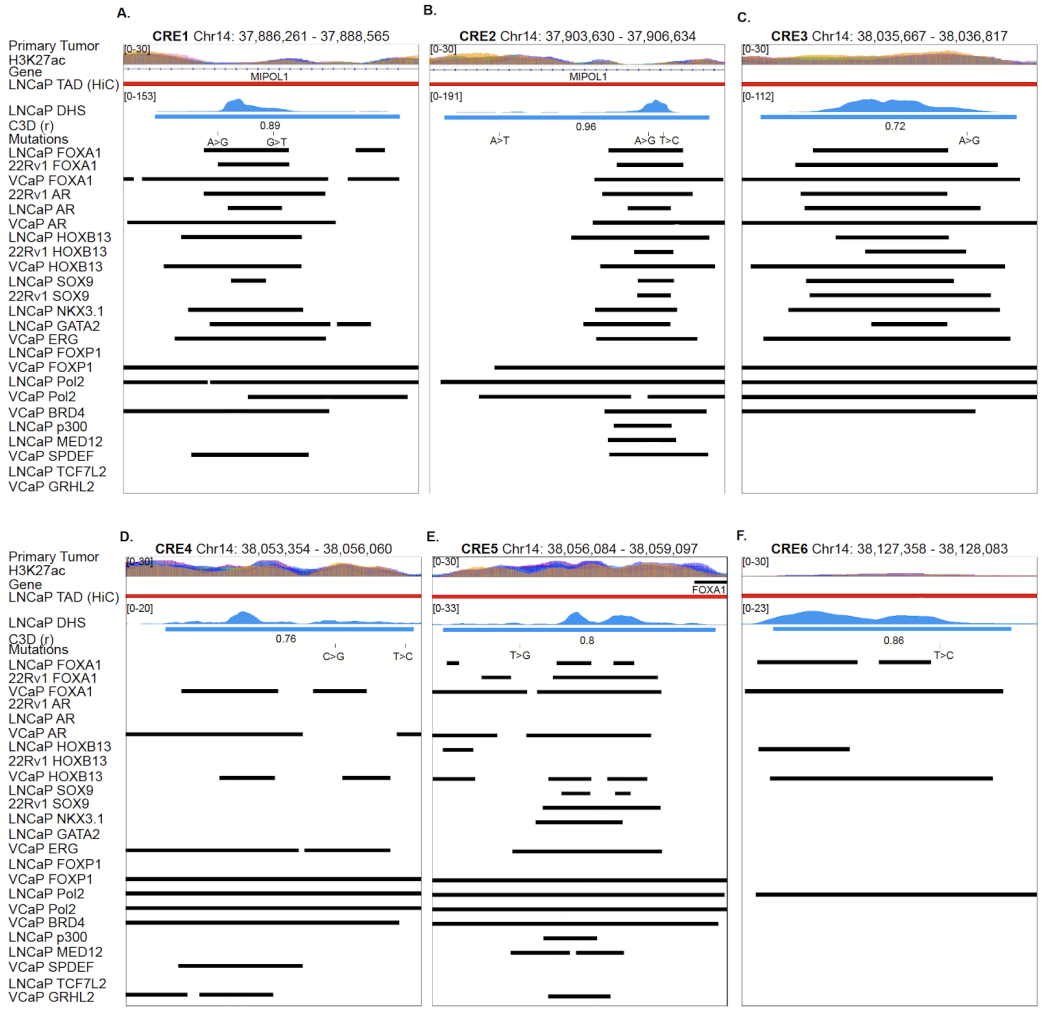

Figure A.4: **Visualization of the functional annotation of the six *FOXA1* CREs. a-f.** Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

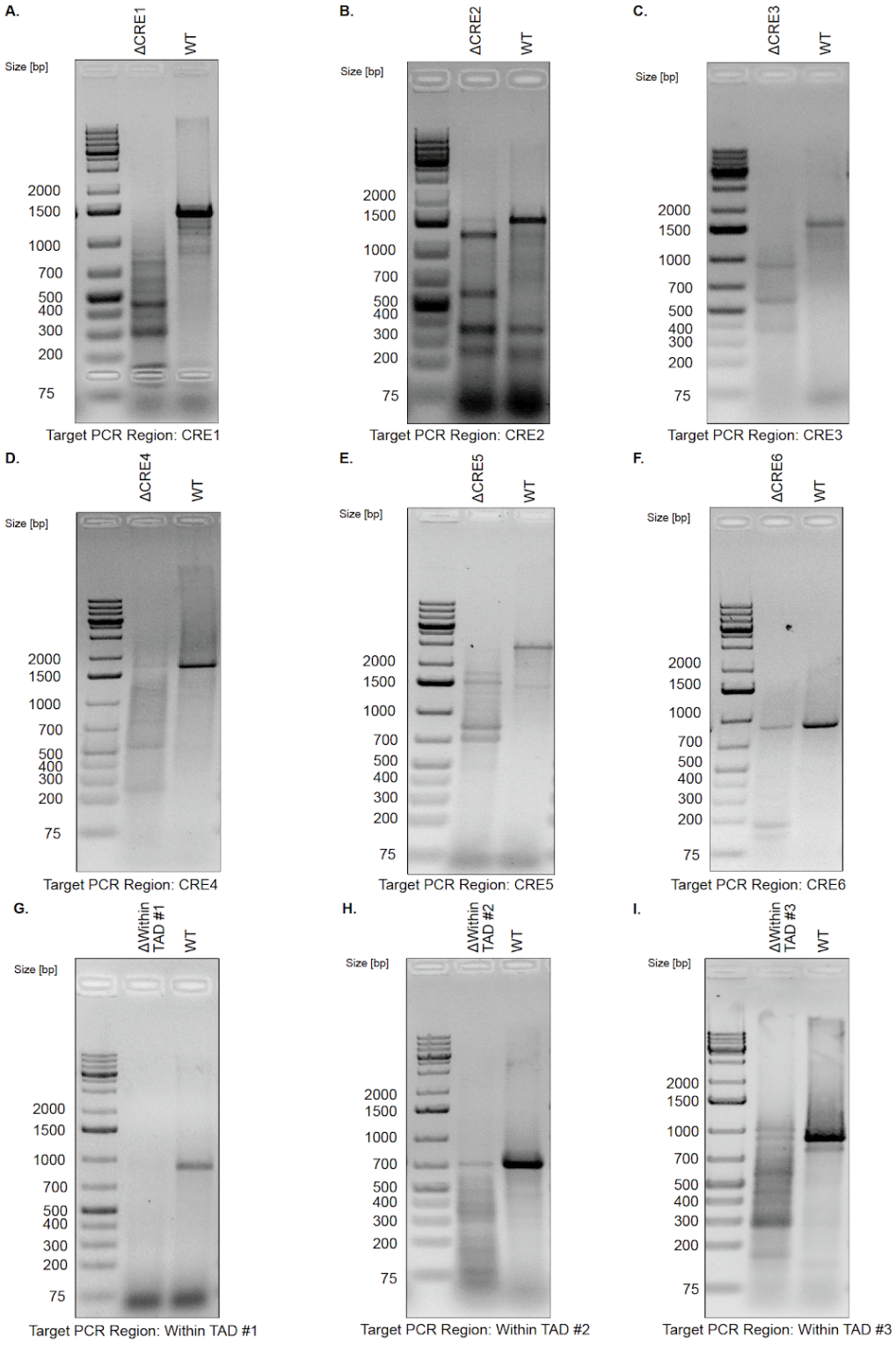

Figure A.5: **Validation of clonal Cas-mediated deletions of CREs. a-f**. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

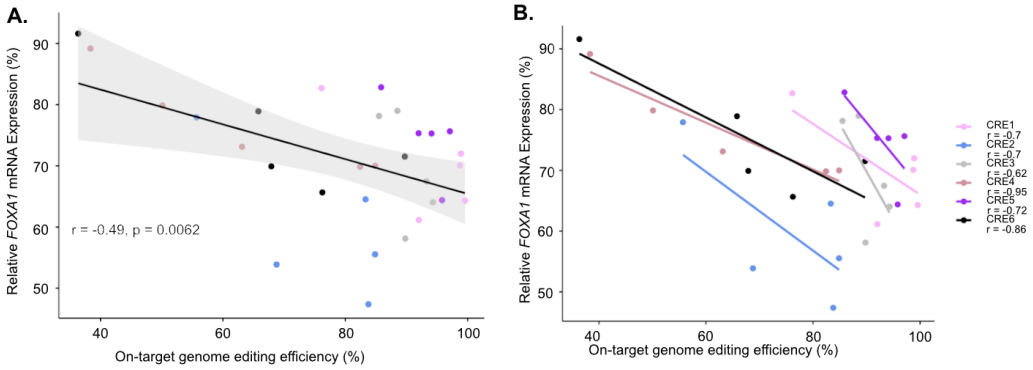

Figure A.6: **Genome editing efficiency (%) is inversely correlated with *FOXA1* mRNA expression. a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all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

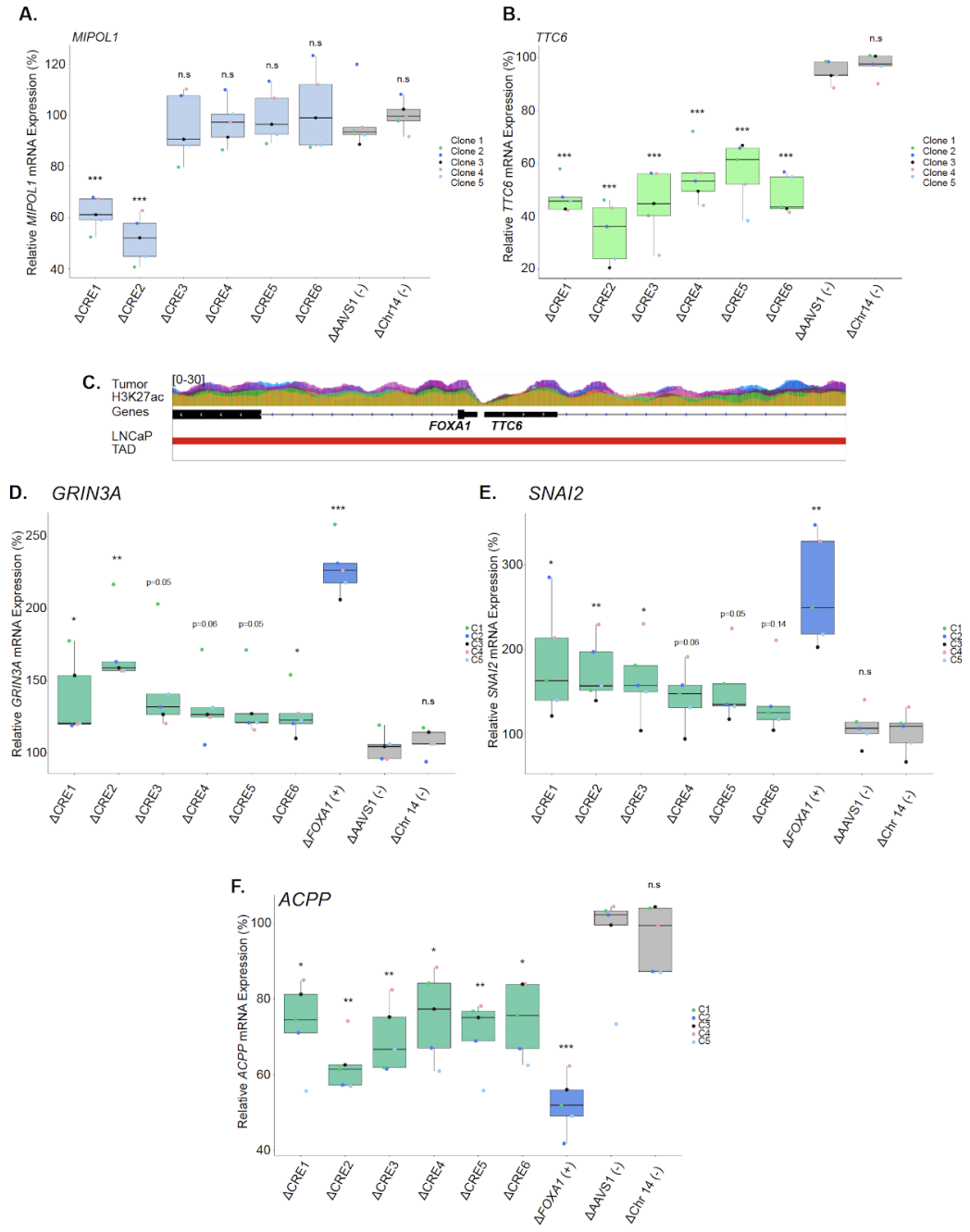

Figure A.7: **Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a.** *MIPOL1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. **b.** *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of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

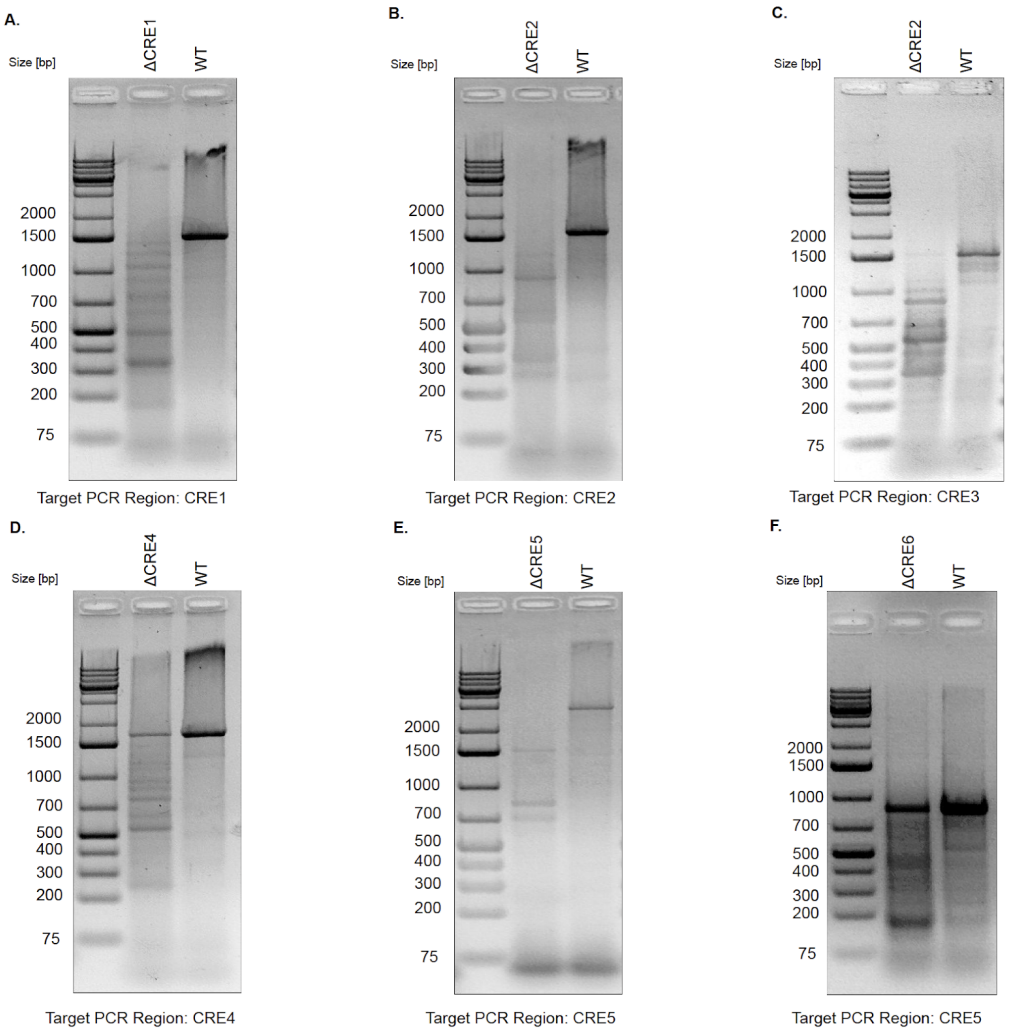

Figure A.8: **Validation of transient Cas9-mediated single deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

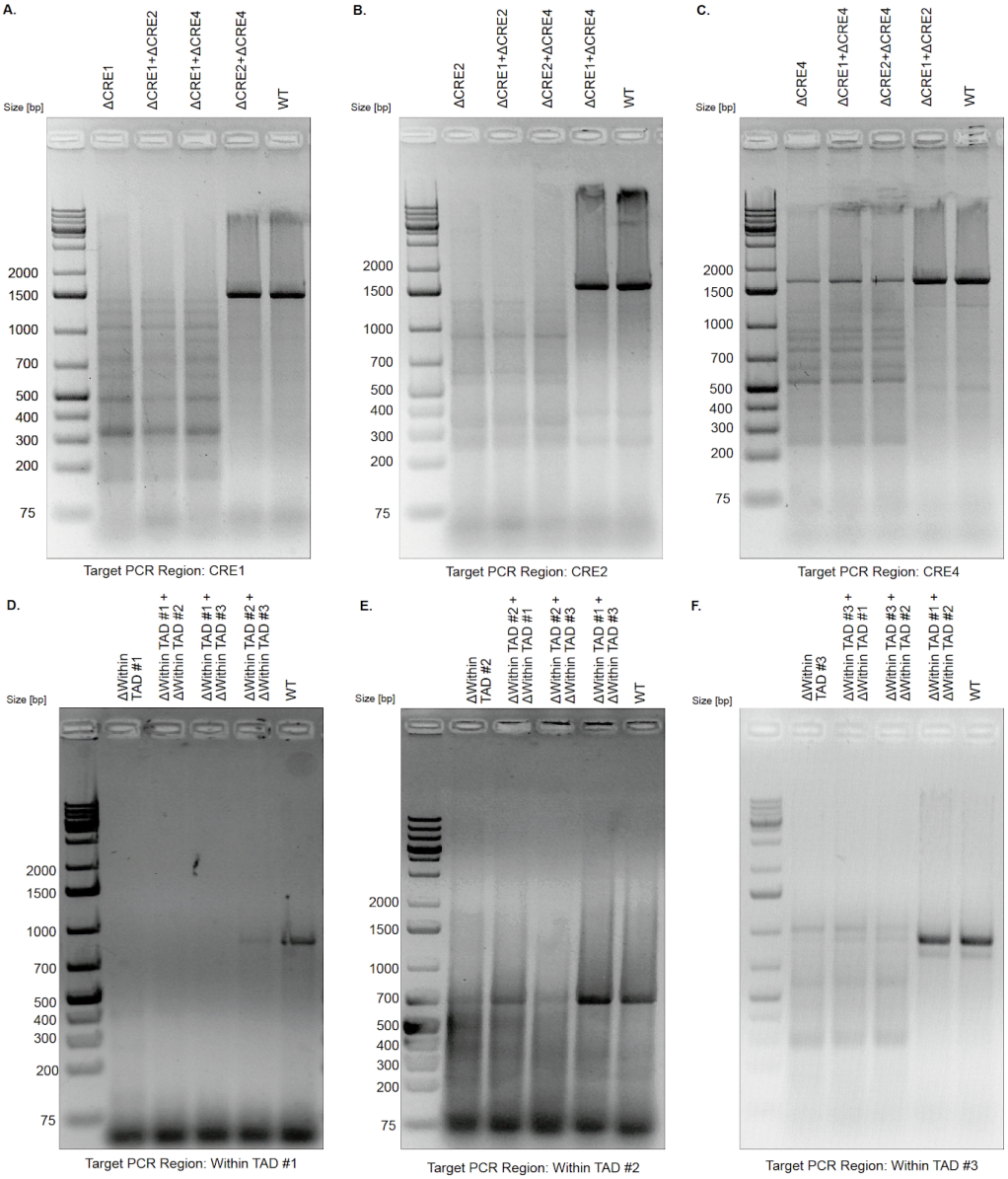

Figure A.9: **Validation of transient Cas9-mediated double deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

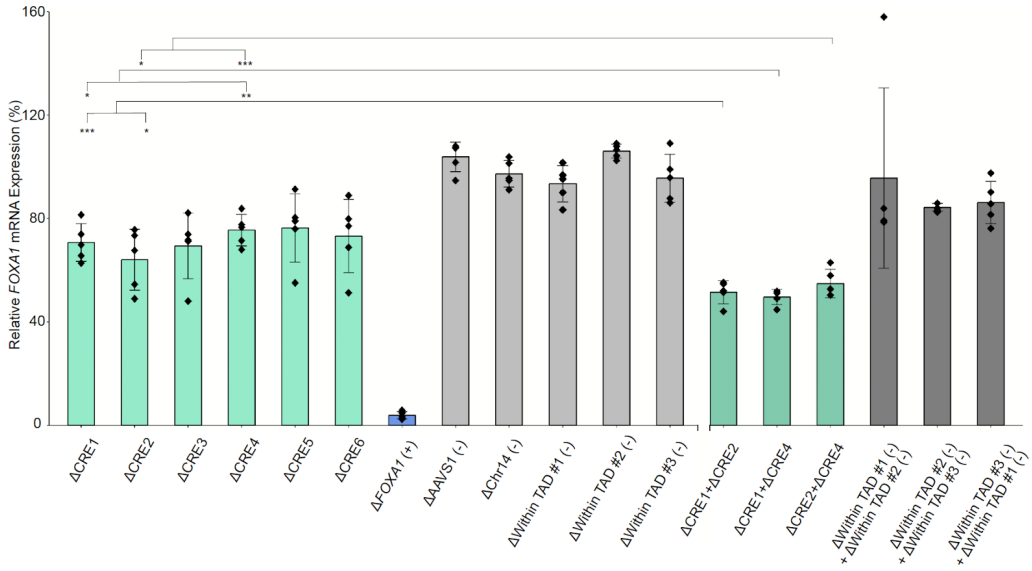

Figure A.10: **Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s).** *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate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

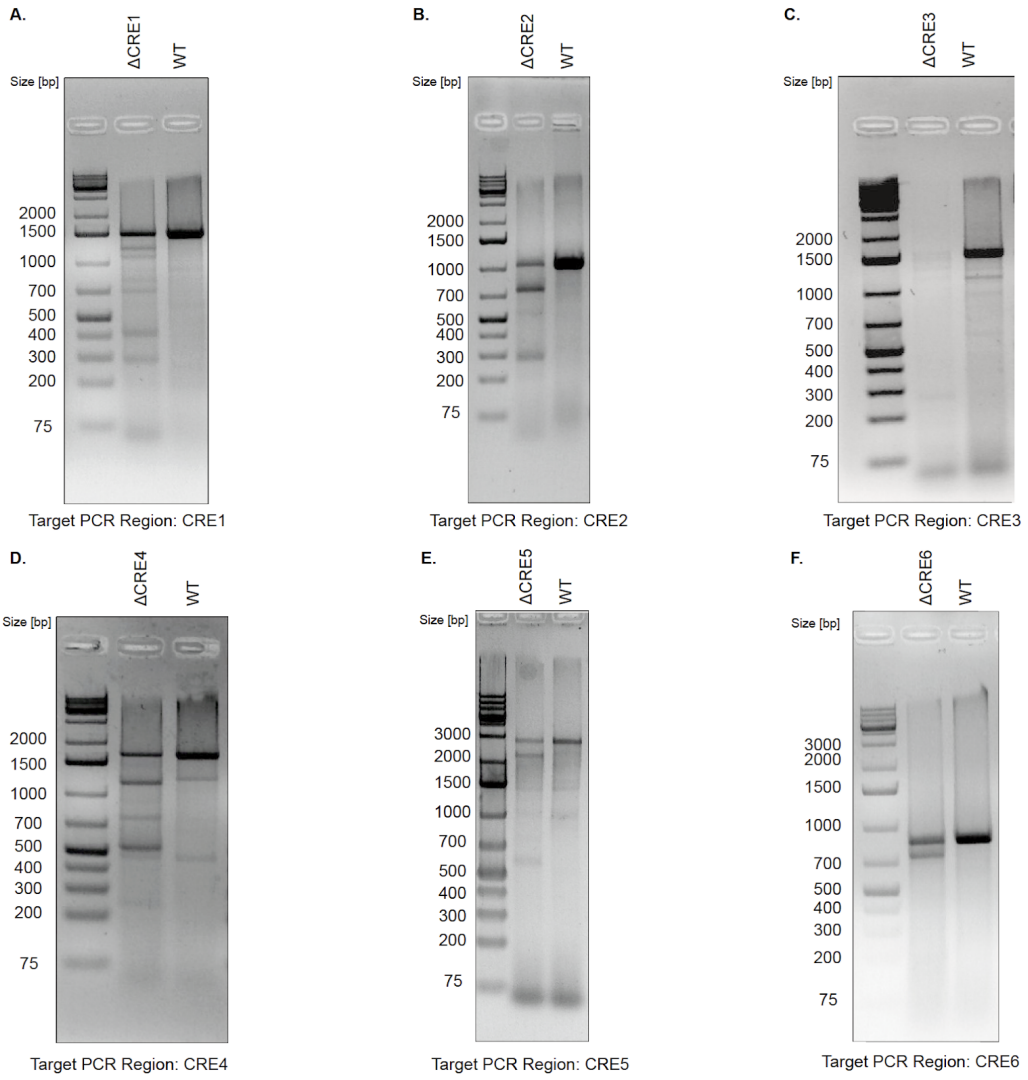

Figure A.11: **Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f.** Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs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

Appendix B

Supplementary Material for Chapter 3

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

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

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

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

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

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

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

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

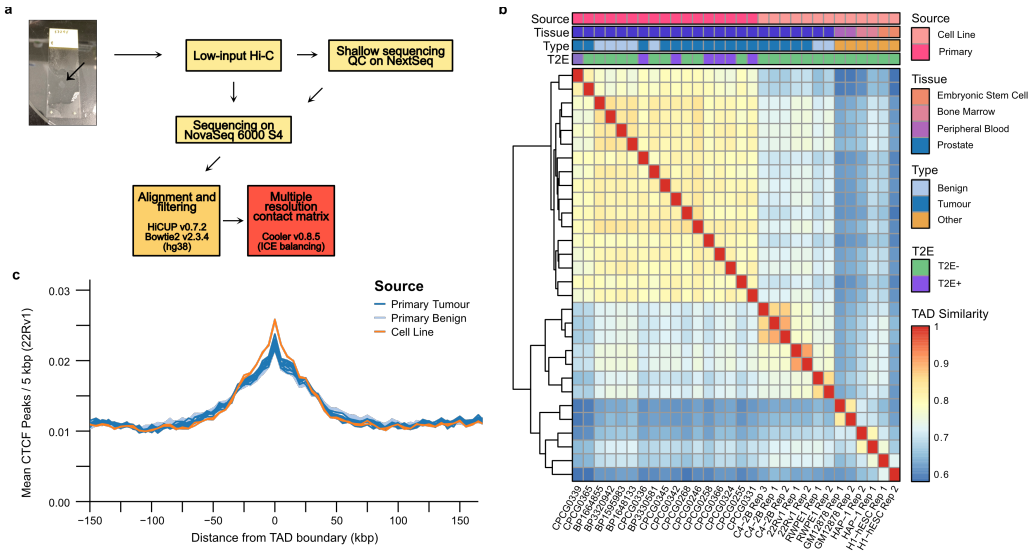

Figure B.1: **Sample processing and TAD similarity between samples. a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate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

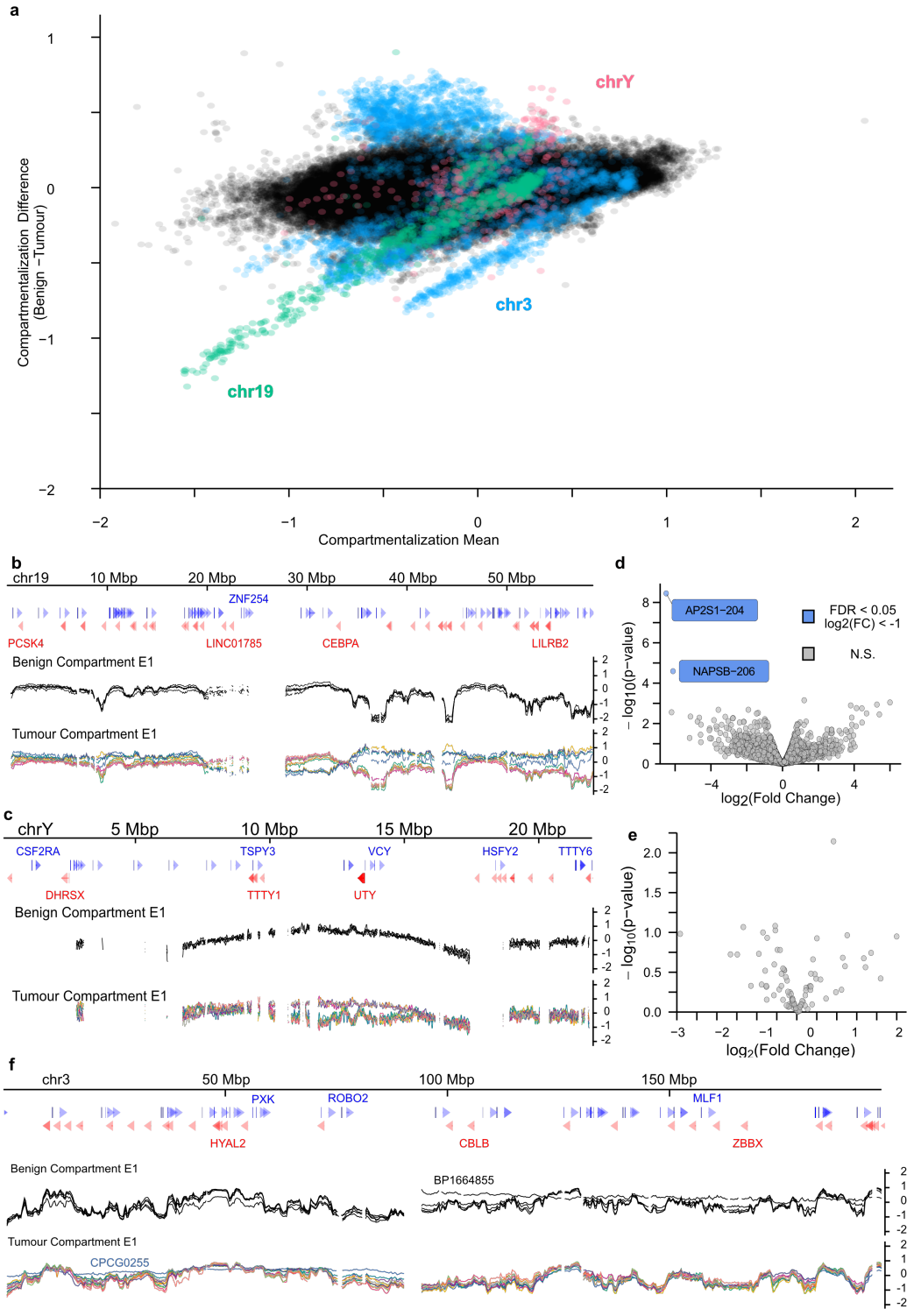

Figure B.2: **Compartmentalization changes in tumours is not associated with widespread differential gene expression.** (Continued on the following page)

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

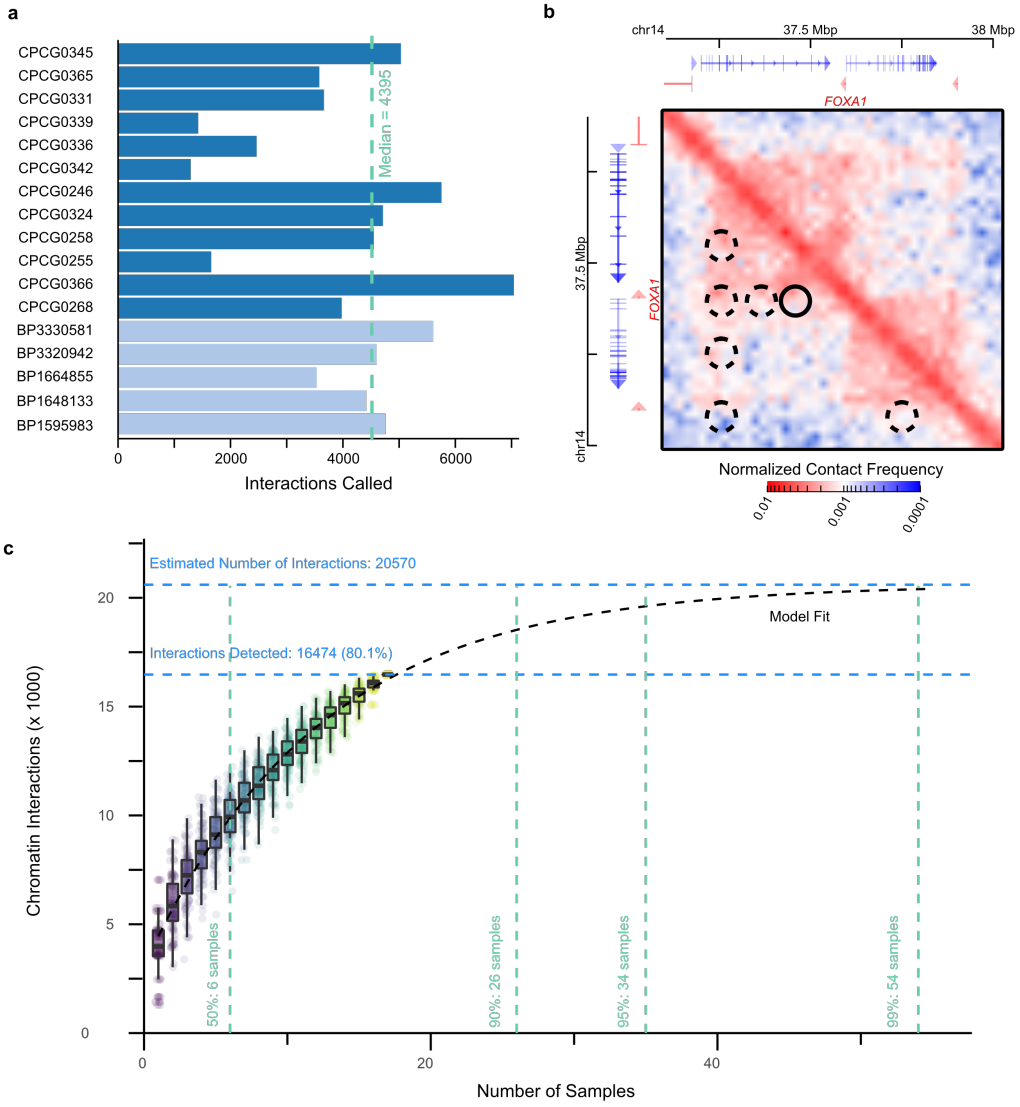

Figure B.3: **Characterization of chromatin interactions in benign and tumour tissue. a.** Bar plot of the number of significant chromatin interactions identified in each of the primary prostate samples. **b.** A snapshot of significant chromatin interactions called around the *FOXA1* gene. Identified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

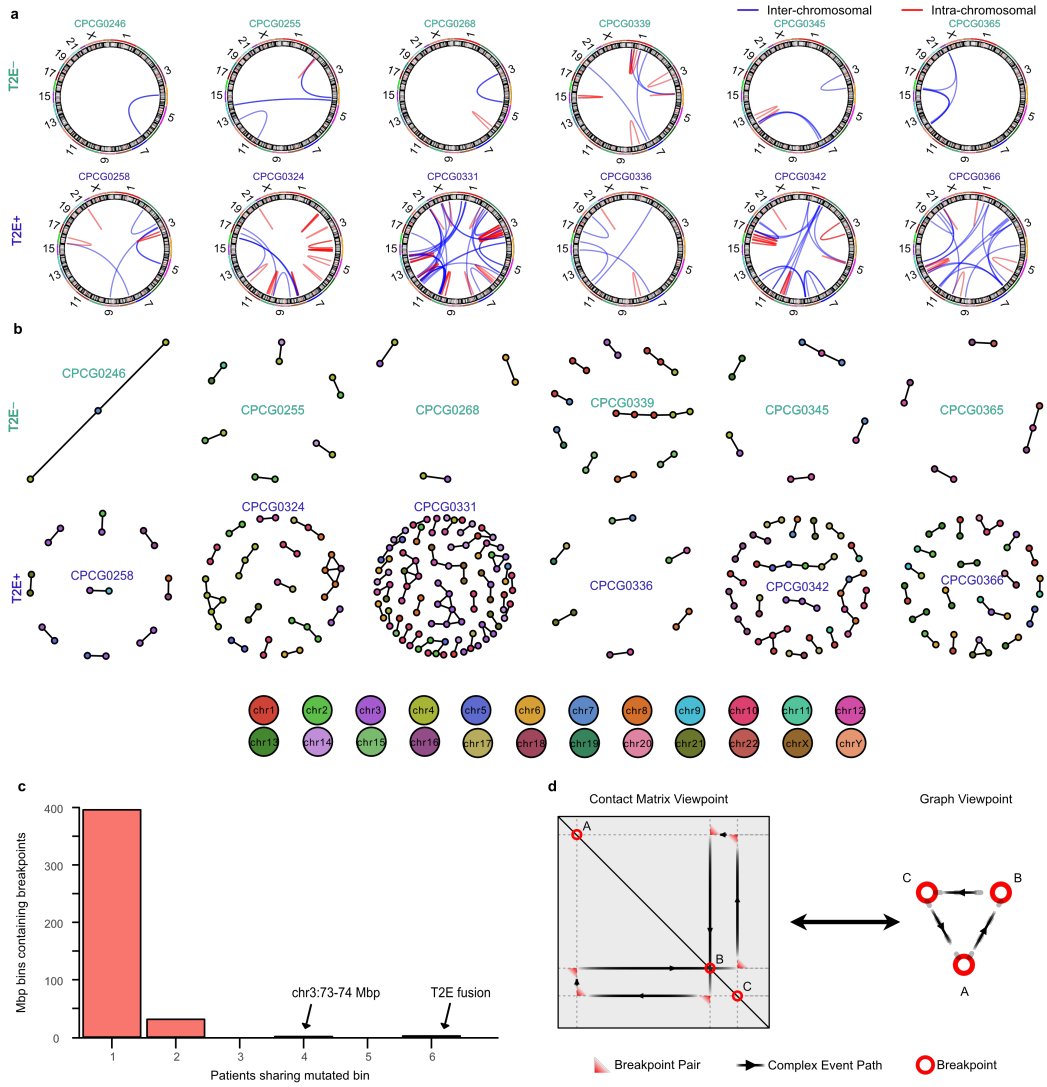

Figure B.4: **Structural variant detection from Hi-C data. a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1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

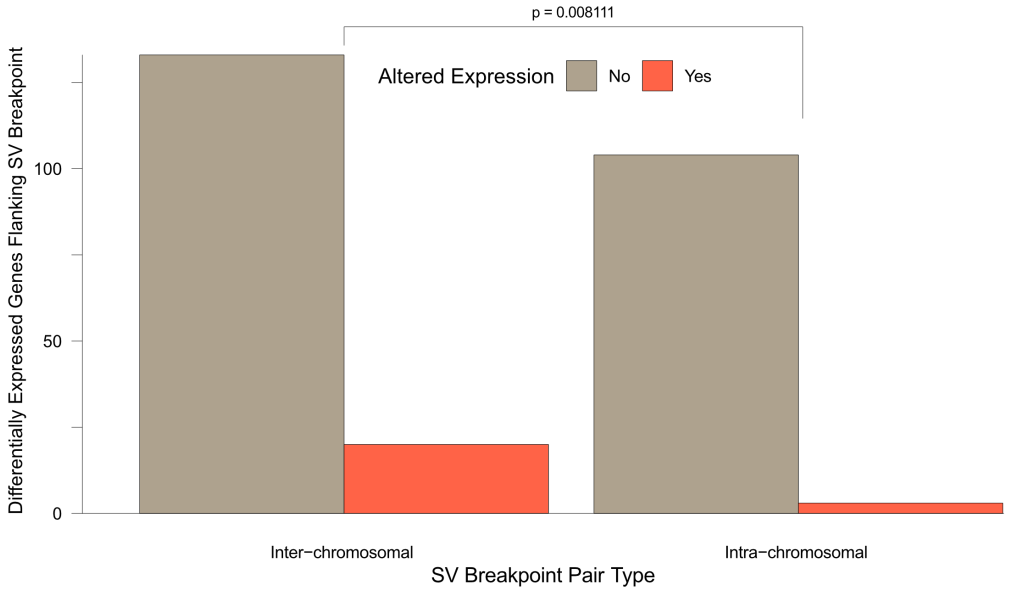

Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes.

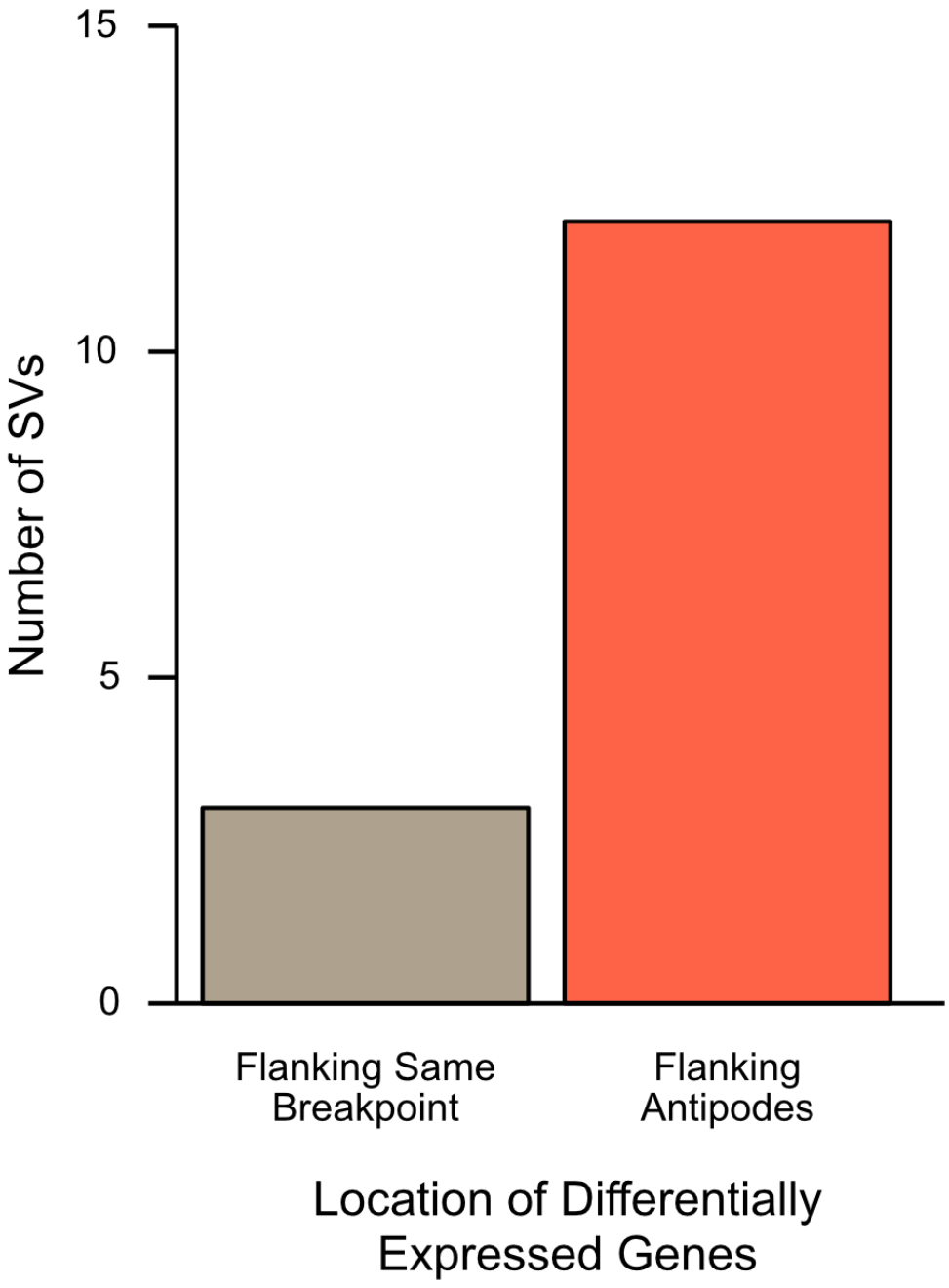

Figure B.6: **Location of differentially expressed genes around SV breakpoints.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

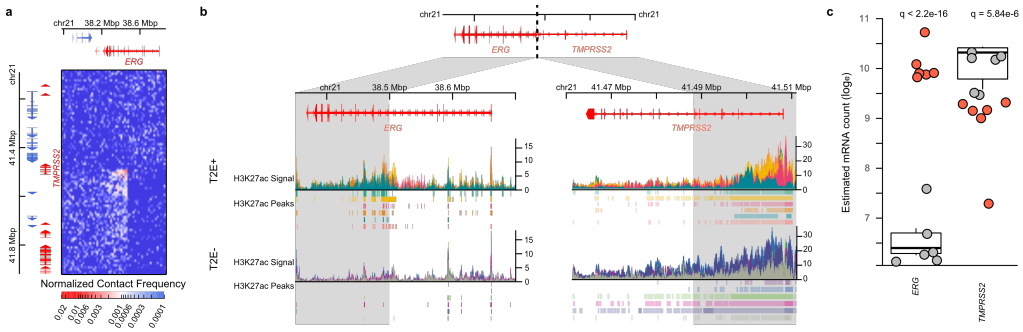

Figure B.7: **Chromatin organization of the *TMPRSS2-ERG* fusion. a.** Contact matrix of the deletion between *TMPRSS2* and *ERG*. **b.** Genome tracks of H3K27ac ChIP-seq signal in T2E+ and T2E- patients. The grey region highlights the loci that come into contact as a result of the deletion. **c.**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

Appendix C

Supplementary Material for Chapter 4

C.1 Differential expression analysis with Sleuth

The differential expression model employed in the Sleuth (v0.30.0) [213, 214] can be described as follows. Consider a set of transcripts, S, measured in N samples with an experimental design matrix, $X \in \mathbb{R}^{N \times p}$, where p is the number of covariates considered. Let Y_{si} be the natural log of the abundance of transcript s in sample i. Given the design matrix

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where d_s is the observed abundances given by

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

where k_{si} is the estimated read count from the Kallisto package (v0.46.1) [215] 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 213], the estimator is unbiased, Namely

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

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

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

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

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

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

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

The variance of the OLS estimator is then

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

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

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

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

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

C.3 Statistical moments of the James-Stein estimator

C.3.1 Expected value of the James-Stein estimator

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

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

where

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

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

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

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

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

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

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

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

C.3.2 Variance of the James-Stein estimator

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

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

By [REF 216], $\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] \leq \mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(OLS)} - \mathrm{B}_1\|^2\right]$. However, this does not imply that $\mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(JS)}\right] \leq \mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(OLS)}\right] \forall s \in S$. Some transcripts may have larger variances than the OLS estimator, but all transcripts in aggregate will have a smaller MSE. This is still desirable if the goal is to find if there is an effect on any transcripts in the set S, instead of a particular one within the set.

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

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

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

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

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

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

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

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

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

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

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

Appendix D

Supplementary Material for Chapter 5

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. & Flicek, 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 Coordinately Activating Ras and Nuclear Factor-κB. 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-Opts 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. Forman, S. J. & Rowe, J. M. The Myth of the Second Remission of Acute Leukemia in the Adult. en. *Blood* **121,** 1077–1082. ISSN: 0006-4971 (Feb. 2013).

160. Oshima, K. *et al.* Mutational Landscape, Clonal Evolution Patterns, and Role of RAS Mutations in Relapsed Acute Lymphoblastic Leukemia. *Proceedings of the National Academy of Sciences* **113,** 11306–11311 (Oct. 2016).

161. Oshima, K. *et al.* Mutational and Functional Genetics Mapping of Chemotherapy Resistance Mechanisms in Relapsed Acute Lymphoblastic Leukemia. en. *Nature Cancer* **1,** 1113–1127. ISSN: 2662-1347 (Nov. 2020).

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

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

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

165. Hirsch, C. *et al.* Consequences of Mutant TET2 on Clonality and Subclonal Hierarchy. *Leukemia* (2018).

166. Shih, A. H. *et al.* Combination Targeted Therapy to Disrupt Aberrant Oncogenic Signaling and Reverse Epigenetic Dysfunction in IDH2- and TET2-Mutant Acute Myeloid Leukemia. *Cancer Discovery* **7,** 494–505. ISSN: 6469626726 (2017).

167. Duy, C. *et al.* Rational Targeting of Cooperating Layers of the Epigenome Yields Enhanced Therapeutic Efficacy against AML. en. *Cancer Discovery* **9,** 872–889. ISSN: 2159-8274, 2159-8290 (July 2019).

168. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* **18,** 553–567. ISSN: 1878-3686 (Electronic)\r1535-6108 (Linking) (2010).

169. Lu, R. *et al.* Epigenetic Perturbations by Arg882-Mutated DNMT3A Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development. *Cancer Cell* **30,** 92–107. ISSN: 1535-6108 (July 2016).

170. Yang, L. *et al.* DNMT3A Loss Drives Enhancer Hypomethylation in FLT3-ITD-Associated Leukemias. en. *Cancer Cell* **29,** 922–934. ISSN: 15356108 (June 2016).

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

172. Hogan, L. E. *et al.* Integrated Genomic Analysis of Relapsed Childhood Acute Lymphoblastic Leukemia Reveals Therapeutic Strategies. en. *Blood* **118,** 5218–5226. ISSN: 0006-4971 (Nov. 2011).

173. Diedrich, J. D. *et al.* Profiling Chromatin Accessibility in Pediatric Acute Lymphoblastic Leukemia Identifies Subtype-Specific Chromatin Landscapes and Gene Regulatory Networks. en. *Leukemia,* 1–14. ISSN: 1476-5551 (Mar. 2021).

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

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

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

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

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

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

180. 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).

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

182. Russo, M. *et al.* Adaptive Mutability of Colorectal Cancers in Response to Targeted Therapies. en. *Science* **366,** 1473–1480. ISSN: 0036-8075, 1095-9203 (Dec. 2019).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

213. 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).

214. 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).

215. 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).

216. 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).