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

Chromatin architecture aberrations contribute to prostate cancer oncogenesis and acute lymphoblastic leukemia relapse

James Hawley
Doctor of Philosophy

Graduate Department of Medical Biophysics
University of Toronto
2022

Abstract

Cancer results from aberrations at the molecular level that enable biological hallmarks. These aberrations can be found within the chromatin architecture of cancer cells that includes the genome, molecular modifications to the genome, and the three-dimensional organization of the chromatin fiber. The majority of genetic variants target non-coding regions of the genome and many genes affected by genetic and epigenetic variants have important roles in chromatin remodelling and maintenance. Thus, understanding the origins of cancer progression requires investigating the targets of these aberrations and how they impact the chromatin architecture.

First, I investigated the impact of non-coding single nucleotide variants that converge on *cis*-regulatory elements for the *FOXA1* gene in primary prostate tumours. We found that deletion and repression of these *cis*-regulatory elements significantly decreases *FOXA1* expression and prostate cancer cell growth by altering the potential of transcription factors to bind at these loci. These results identify *cis*-regulatory elements that control *FOXA1* expression in primary prostate cancer as potential targets for therapeutic intervention.

Secondly, I used chromatin conformation capture of 12 primary prostate cancer tumours and 5 benign prostate tissues to characterize the three-dimensional genome organization. We found that large-scale organization, including topologically associated domains and compartments, is largely stable over oncogenesis but that small-scale focal chromatin interactions change between benign and tumour tissue. We also investigated the impact of structural variants on chromatin organization and identify novel enhancer hijacking events. These results indicate that enhancer hijacking of prostate cancer oncogenes may be a more common driver of disease than previously recognized. Then, I developed a statistical framework for differential gene expression analysis to address the impact of non-recurrent structural variants in our primary prostate tumour cohort. This method improves on conventional gene expression fold change estimates in these unbalanced experimental designs.

Finally, I investigated the genetic and epigenetic changes that underlie B-cell acute lymphoblastic leukemia relapse. I found recurrent loss of DNA methylation in patient-matched relapse samples that

indicate a more stem-like chromatin state. Together, my work investigates the relationship between multiple components of the chromatin architecture, and how aberrations to this architecture connects oncogenesis, disease progression, and relapse.

Acknowledgments

As with all things, there are many people to thank for all of the support and guidance along the way. First and foremost, I would like to thank my fiancée, Alison McNeil, for being a steadfast and supportive partner for all of my endeavours for many years. I would like to thank my parents, Rick and Corrine Hawley, for always supporting me and encouraging my academic pursuits. I would also like to thank my brother, Stephen Hawley, as well as Matt Smart, Ryan Kearns, Radmila Kovac, and Dr. Kathleen Houlahan, for being friends who I always competed with intellectually and helped pushed me to be better.

I would like to thank my supervisor, Dr. Mathieu Lupien, for his encouragement, his guidance, and the freedom he granted me to pursue nearly any academic project of my own choosing. I would like to thank all of the members of my lab I encountered during my tenure: Christopher Arlidge, Bansri Patel, Tina Keshavarzian, Shalini Bahl, Jocelyn Chen, Dr. Giacomo Grillo, Dr. Sarina Cameron, Dr. med Kira Korienko, Ankita Nand, Dr. Stanley Zhou, Dr. Seyed Ali Madani Tonekaboni, Dr. Parisa Mazrooei, Ingrid Kao, Dr. Kinjal Desai, Dr. Qin Wu, Dr. Nergiz Artun Dogan, Dr. Bettina Nadorp, Dr. Aditi Qamra, Dr. Genevieve Deblois, Dr. Paul Guilhamon, Dr. Alexander Murison, Dr. Ken Kron, Dr. Swneke Bailey, Dr. Samah El Ghamrasni, Dr. Elias Orouji, Julie Relatado, Natalia Mukhina, Neda Shokraneh, Romy van Vliet, Aude Gerbaud, Felicia Tran, and Tahmid Mehdi. I would like to thank my Supervisory Committee members Dr. John Dick and Dr. Michael Hoffman for their feedback and guidance over the years, as well as my collaborators: Dr. Helena Boutzen, Dr. Laura Garcia Prat, Dr. Rashim Pal Singh, Dr. Ayesh Seneviratne, Dr. Mingjing Xu, Dr. Rupert Hugh-White, Dr. Theodorus van der Kwast, Dr. Michael Fraser, Dr. Paul Boutros, and Dr. Robert Bristow. I would like to thank members of the Princess Margaret Genomics Centre for their dedicated work, including Nicholas Khuu, Iulia Cirlan, Julissa Tsao, Zhibin Liu, Natalie Stickle, and Carl Virtanen.

Finally, to all the patients whose samples I have have collected and analyzed, I would like to thank them and their families for their sacrifices and willingness to participate in the scientific process. We still live in a world with cancer, but I can only hope that my work here helps us arrive at a world without it.

Contents

List of Tables

List of Figures

List of Abbreviations

3C chromatin conformation capture

ALL acute lymphoblastic leukemia

AML acute myeloid leukemia

ANOVA analysis of variance

APA aggregate peak analysis

AR androgen receptor

ATAC-seq assay for transposase-accessible chromatin sequencing

B-ALL B-cell acute lymphoblastic leukemia

BAM binary alignment map

bp basepair

cDNA complementary DNA

ChIP-seq chromatin immunoprecipitation sequencing

CLL chronic lymphocytic leukemia

CML chronic myeloid leukemia

CMP common myeloid progenitor

CNV copy number variant

CORE cluster of regulatory elements

CPC-GENE Canadian Prostate Cancer Genome Network

CpG CG dinucleotide

crRNA CRISPR RNA

CRE *cis*-regulatory element

CRISPR clustered regularly interspaced short palindromic repeat

CTCF CCCTC-binding factor

CUT&RUN cleavage under targets and release using nuclease

DEPMAP Cancer Dependency Map

DHS DNase I hypersensitive sites

DLBCL diffuse large B-cell lymphoma

DMR differentially methylated region

DNA deoxyribonucleic acid

DNAme DNA methylation

DNase-seq DNase I hypersensitive sequencing

dRI disease relapse-initiating

Dx diagnosis

EGA European Genome-Phenome Archive

EarlyProB early progenitor B cell

FDR false discovery rate

FN false negative

FP false positive

FOX forkhead box

GEO Gene Expression Omnibus

GLM generalized linear model

GMP granulocyte-macrophage progenitor

GO gene ontology

gRNA guide RNA

HSC hematopoietic stem cell

HSPC hematopoietic stem and progenitor cell

ICE iterative correction and eigenvector decomposition

IDH isocitrate dehydrogenase

IID independent and identically distributed

IQR inter-quartile range

ISUP International Society of Urological Pathology

JS James-Stein

KMT histone lysine methyltransferase

KO knockout

LDA limiting dilution assay

LMPP lymphoid-primed multi-potent progenitor

MeCapSeq DNA methylation capture sequencing

MEP megakaryocyte-erythrocyte progenitor

MNase-seq microccocal nuclease sequencing

MSE mean square error

mCRPC metastatic castration-resistant prostate cancer

MDS myelodisplastic syndrome

MLP monocyte-lymphoid progenitor

MPP multi-potent progenitor

MPRA massively-parallel reporter assay

NSG NOD scid gamma

OLS ordinary least squares

mRNA messenger RNA

PBS phosphate-buffered saline

PCa prostate cancer

PDX patient-derived xenograft

PreProB pre-progenitor B cell

ProB progenitor B cell

PSA prostate-specific antigen

REB research ethics board

Rel relapse

RLU relative luciferase unit

RNA ribonucleic acid

RNAi RNA interference

RNA-seq RNA sequencing

shRNA small hairpin RNA

siRNA small interfering RNA

SNV single nucleotide variant

SDS sodium dodecyl sulfate

SRA Sequence Read Archive

SNF similarity network fusion

STR short tandem repeat

SV structural variant

T2E *TMPRSS2-ERG*

TAD topologically associated domain

TCGA The Cancer Genome Atlas

TET ten-eleven translocation

TSS transcription start site

TN true negative

TNM tumour node metastasis

TP true positive

TF transcription factor

tracrRNA trans-activating CRISPR RNA

UHN University Health Network

UTR untranslated region

WES whole exome sequencing

WGBS whole genome bisulfite sequencing

WGS whole genome sequencing

WT wild-type

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 countries [**brayGlobalCancerStatistics2018**]. 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 [**gilbertsonMappingCancerOrigins2011**]. 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 [**hanahanHallmarksCancer2000**, **hanahanHallmarksCancerNext20 flavahanEpigeneticPlasticityHallmarks2017**, **pavlovaEmergingHallmarksCancer2016**] (Figure 1.1).

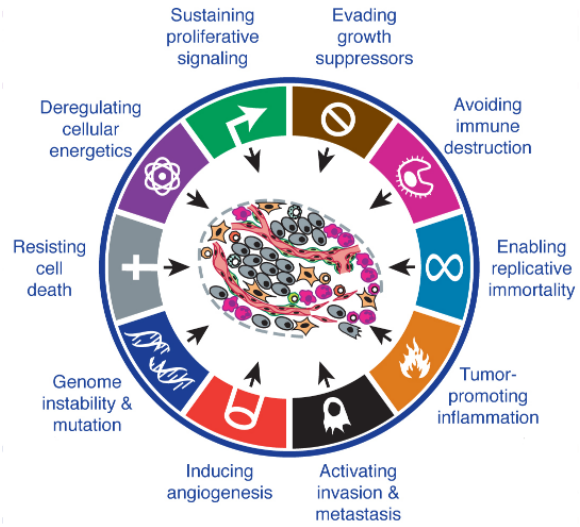

Figure 1.1: **The hallmarks of cancer.** Adapted from [**hanahanHallmarksCancerNext2011**].

Many of these hallmarks of cancer can be achieved through aberrations to the genome and the molecular machinery that enables cells to function normally [**garrawayLessonsCancerGenome2013**]. 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 [**lengauerGeneticInstabilitiesHuman1998**] or mutations to *BRCA1*, *BRCA2*, and *ATM* genes in prostate cancer (PCa) [**abeshouseMolecularTaxonomyPrimary2015**]. Similarly, replicative immortality can be achieved through telomere elongation by over-expression of the *TERT* gene [**vinagreFrequencyTERTPromoter2013**]. Mutations to the *TERT* promoter, resulting in its over-expression, were first identified in melanomas [**huangHighlyRecurrentTERT2013**, **hornTERTPromoterMutations2013**], but have since been further identified in bladder, thyroid, and brain cancers [**vinagreFrequencyTERTPromoter2013**, **nagarajanRecurrentEpimutationsActivate2014**, **sternMutationTERTPromoter2015**]. But while cancer has long been viewed as a disease of the genome [**hanahanHallmarksCancer2000**, **garrawayLessonsCancerGenome2013**], 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 [**albertsMolecularBiologyCell2015**] (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 [**goodrichUnexpectedRolesCore2010**]. 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 [**schoenfelderLongrangeEnhancerPromoter2019**, **spitzTranscriptionFactorsEnhancer2012**, **ongEnhancerFunctionNew2011**, **anderssonDeterminantsEnha**
gasznerInsulatorsExploitingTranscriptional2006, **oudelaarRelationshipGenomeStructure2020**] (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 [**farnhamInsightsGenomicProfiling2009**, **spitzTranscriptionFactorsEnhancer2012**]. 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* [**liuCistromeIntegrativePlatform2011**, **lupienCistromicsHormonedependentCancer2009**]. 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 [**kimAnalysisVertebrateInsulator2007**, **dixonTopologicalDomainsMammalian2012**]. Mutations to the sequence motif can alter CTCF's binding affinity for DNA, as is the case with many TFs [**kasowskiVariationTranscriptionFactor2010**, **mauranoWidespreadSitedependentBuffering2012**, **mauranoLargescaleIdentificationSequence2015**]. 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 [**mauranoRoleDNAMethylation2015**, **wangWidespreadPlasticityCTCF2012**, **wiehleDNAMethy** **xuNascentDNAMethylome2018**, **vinerModelingMethylsensitiveTranscription2016**], as are DNA methyltransferases DNMT1, DNMT3A, and DNMT3B [**gollEukaryoticCytosineMethyltransferases2005**, **listerHumanDNAMethylomes2009**]. 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 [**farnhamInsightsGenomicProfiling2009**, **ongEnhancerFunctionNew2011**, **spitzTranscriptionFactorsEnh** 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 [**henikoffNucleosomeDestabilizationEpigenetic200** **jiangNucleosomePositioningGene2009**]. The collection of DNA, nucleosomes, DNA-bound TFs, 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 [**vierstraGlobalReferenceMapping2020**, **cusanovichSingleCellAtlasVivo2018**] and DNA damage repair in inaccessible regions [**polakCelloforiginChrom** 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 [**zhuTranscriptionFactorsReaders2016**, **fureyChIPSeqNew2012**, **carterEpigeneticBasisCellular2021**] (Figure 1.2c). Localization of

CREs that are tens of 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 [**zhouChartingHistoneModifications2011**, **dekker3DGenomeModerator2016**, **finnMolecularBasisBiological2019**, **oudelaarRelationshipGenomeStr**
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 [**liebermanaidenComprehensiveMappingLongRange2009**, **rao3DMapHuman2014**, **oudelaarRelationshipGenomeStructure2020**, **mirnyTwoMajorMechanisms2019**]. 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 [**dixonTopologicalDomainsMammali**
stergachisConservationTransactingCircuitry2014, **berthelotComplexityConservationRegulatory2017**], focal chromatin interactions can differ up to 45 % between cell types, providing a further mechanism to change chromatin state [**rao3DMapHuman2014**]. Different chromatin states enable cells with the same DNA sequence to express genes differently [**spurrellTiesThatBind2016**, **buenrostroSinglecellChromat**
leeTranscriptionalRegulationIts2013, **ongEnhancerFunctionNew2011**, **schoenfelderLongrangeEnhancer**
zhouChartingHistoneModifications2011], 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 [**conesaSurveyBestPractices2016**]. Protein binding sites and histone post-translational modifications 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) [**robertsonGenomewideProfilesSTAT12007**, **baileyPracticalGuidelinesComprehensive2013**, **skeneTargetedSituGenomewide2018**]. Accessible and inaccessible chromatin can be assessed by

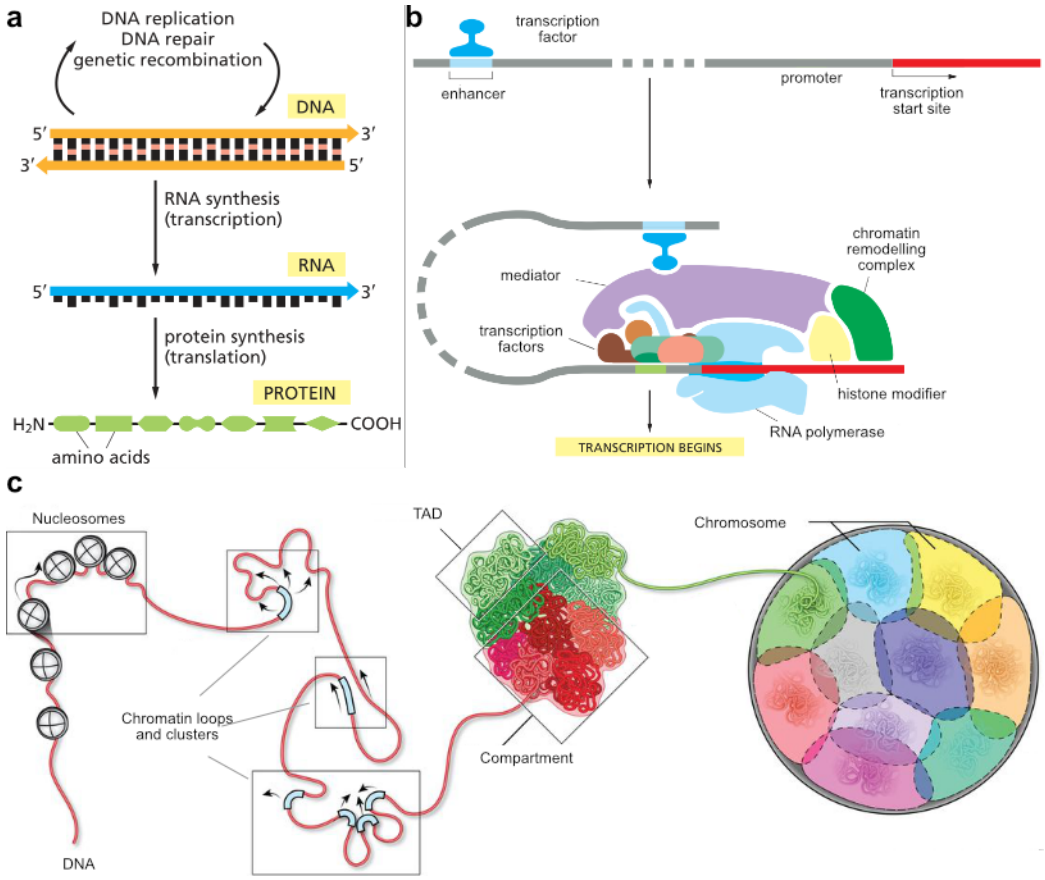

Figure 1.2: **The basics of gene expression inside the nucleus.** **a.** The central dogma of molecular biology. Adapted from [**albertsMolecularBiologyCell2015**]. **b.** Schematic of the transcription machinery to initiate transcription. Adapted from [**albertsMolecularBiologyCell2015**]. **c.** The scale of chromatin interactions across length scales. Adapted from [**finnMolecularBasisBiological2019**].

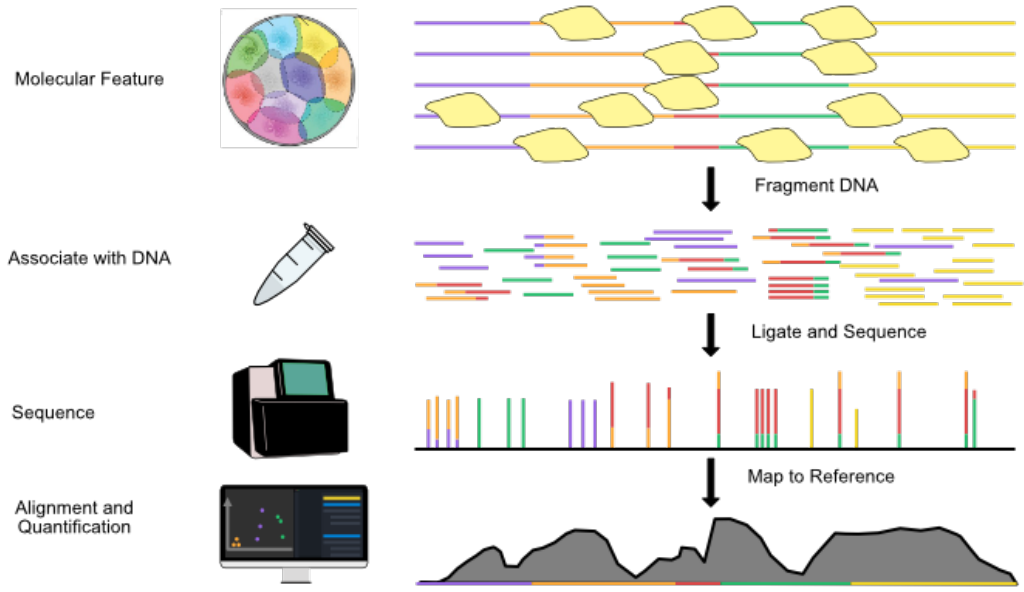

Figure 1.3: **Characterizing functional DNA elements with high throughput sequencing.**

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 [**boyleHighResolutionMappingCharacterization2008**, **buenrostroTranspositionNativeChromatin2013**, **buenrostroATACseqMethodAssaying2015**, **corcesImprovedATACseqProtocol2017**, **schonesDynamicRe** DNAme can be measured with bisulfite-sequencing assays [**lairdPrinciplesChallengesGenomewide2010**], and distal chromatin interactions can be identified with chromatin conformation capture (3C) and 3C-based methods such as Hi-C [**dekkerCapturingChromosomeConformation2002**, **liebermanaidenCompre** **dixonTopologicalDomainsMammalian2012**, **noraSpatialPartitioningRegulatory2012**, **rao3DMapHuma** 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 [**birneyIdentificationAnalysisFunctional2007**, **mooreExpandedEncyclopaediasDNA2020**]. 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 [**mooreExpandedEncyclopaediasDNA2020**]. Interpreting this data requires computational methods to correlate and interpret measurements across samples. Genome segmentation methods such as ChromHMM [**ernstChromHMMAutomatingChromatinstate2012**] and Segway [**hoffmanUnsupervisedPatternDiscovery2012**, **chanSegwayGaussianMixture2018**] 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 [**zhouEmergenceNoncodingCancer2016**, **gasperiniComprehensiveCatalogueValidated2020**]. These same techniques can also be used to screen for candidate CREs themselves, through massively-parallel reporter assays (MPRAs) and CRISPR screens [**gasperiniComprehensiveCatalogueValidated2020**], 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 [**croceOncogenesCancer2008**, **baileyComprehensiveCharacte**
weinsteinCancerGenomeAtlas2013, **PancancerAnalysisWhole2020**]. 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 [**khuranaRoleNoncodingSequence2016**]. 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 [**mooreExpandedEncyclopaediasDNA2020**, **rheinbayAnalysesNoncodingSomatic2020**, **PancancerAnalysisWhole2020**, **zhangHighcoverageWholegenomeAnalysis2020**]. Analysis of recurrent somatic mutations in tumours led to the identification of *TP53* as a tumour suppressor gene [**hollsteinP53MutationsHuman1991**], the frequently mutated *SPOP* gene to help de-

fine a molecular subtype of prostate tumours [**barbieriExomeSequencingIdentifies2012**], and the interpretation of recurrent rearrangements of the proto-oncogene *MYC* in multiple cancers [**meyerReflecting25Years2008**]. The impact of a mutation can also be predicted by identify overlapping regulatory elements or TF binding sites [**mauranoWidespreadSitedependentBuffering2012**, **cowper-sal*lariBreastCancerRisk2012**, **kronEnhancerAlterationsCancer2014**]. 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 % of breast cancers, resulting in its over-expression, despite the gene itself being mutated in ~ 1 % of breast cancers [**baileyNoncodingSomaticInherited2016**]. Similarly, the binding sites of the *FOXA1*, *HOXB13*, *AR*, and *SOX9* TFs are enriched with mutations affecting their binding affinities [**mazrooeiCistromePartitioningReveals2019**] and recurrent amplifications of enhancers near the *AR* and *FOXA1* genes are associated with increased rates of metastasis [**quigleyGenomicHallmarksStructural2018**, **paroliaDistinctStructuralClasses2019**]. 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 [**northcottEnhancerHijackingActivates2014**]. This mechanism of enhancer hijacking has also been observed in developmental diseases [**lupianezDisruptionsTopologica**
allouNoncodingDeletionsIdentify2021]. 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 [**pirozziImplicationsIDHMutations2021**, **imDNMT3AIDHMutations2014**, **issaAcuteMyeloidLeukemia2**
molenaarWildtypeMutatedIDH12018, **shihRoleMutationsEpigenetic2012**]. These mutations often affect the DNAme profiles of tumours and differentiation programs [**pirozziImplicationsIDHMutations**
such as loss of enhancer hydroxymethylation and germinal centre hyperplasia in diffuse large B-cell lymphoma (DLBCL) [**dominguezTET2DeficiencyCauses2018**]. Similarly, mutations to the *EZH2* gene in leukemias can affect the ability of the EZH2 protein to write the H3K27me3 histone mark [**plassMutationsRegulatorsEpigenome2013**, **nikoloskiSomaticMutationsHistone2010**, **ernstInactivatingMutationsHistone2010**, **morinSomaticMutationsAltering2010**] and *EZH2*

over-expression is associated with poor survival in PCa [**varamballyPolycombGroupProtein2002**, **xuEZH2OncogenicActivity2012**, **minOncogeneTumorSuppressor2010**, **kimTargetingEZH2Cancer2016**]. 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 [**jonesCancerepigeneticsComesAge1999**, **jonesFundamentalRoleEpigenetic200**. 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 [**jonesFundamentalRoleEpigenetic2002**, **feinbergHistoryCancerEpigenetics2**. Importantly, these changes in DNAme can be found in the absence of mutations targeting DNA methyltransferases. Analysis of $\sim$ 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 [**zhaoDNAMethylationLandscape2020**]. Ependymomas have also been found to display distinct DNAme profiles in the absence of recurrent mutations across patients [**mackEpigenomicAlterationsDefine2014**] along with acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), glioblastoma, and colorectal, liver, pancreatic, and ovarian cancers [**issaCpGIslandMethylator2004**]. 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 [**schmelzInductionGeneExpression2005**, **azadFutureEpigeneticTherapy2013**, **kellyEpigeneticModificationsTherapeutic2010**]. 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 [**wangWidespreadPlasticityCTCF2012**, **mauranoRoleDNAMethylation2015**]. In gastrointestinal cancer, CTCF binding sites are hypermethylated *SDH*-deficient tumours, resulting in widespread loss of CTCF and increased contact be-

tween the *FGF3* and *FGF4* oncogenes and a nearby enhancer cluster [**flavahanAlteredChromosomalTopology201**
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 TAD boundaries, resulting in aberrant regulation has also been found in T-cell ALL, leading to over-expression of the *TAL1* and *LMO2* oncogenes [**hniszActivationProtooncogenesDisruption2**
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 gene regulation extends beyond DNAme. Cell type differences in nucleosome occupancy can lead to increased rates of mutation across the genome [**pichSomaticGermlineMutatio**
Similarly, TF binding can affect the ability of DNA damage repair complexes to perform local nucleotide excision repair [**sabarinathanNucleotideExcisionRepair2016**, **gonzalez__perezLocalDeterminantsM**
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 [**wangSimilarityNetworkFusion2014**]. 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 [**rappoportMultiomicMultiviewClustering2018**]. 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 [**brayGlobalCancerStatistics2018**, **brennerProjectedEstimatesCancer2020**]. 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 [**rebelloProstateCancer2021**]. 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)[**rebelloProstateCancer2021**]. 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 [**rebelloProstateCancer2021**]. While $\sim$ 93 % of men with localized PCa survive, $\sim$ 70 % of patients with metastatic disease will die within 5 years [**hahnMetastaticCastrationSensitiveProstate2018**, **SEERProstateCancer**], accounting for $\sim$ 10 % of all cancer deaths in men [**brennerProjectedEstimatesCancer2**. 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 [**rawlaEpidemiologyProstateCancer2019**]. 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 [**SEERProstateCancer**]. In addition to age, genetic ancestry is a risk factors for developing the disease. Men of African ancestry are $\sim$ 1.6 times more likely to be diagnosed with PCa than men of western European ancestry, who in turn are $\sim$ 2 times more likely than men of Asian ancestry [**SEERProstateCancer**, **smithAfricanAmericanProstateCancer2017**, **dalleraChangingIncidenceMetastatic2019**]. Men of different ancestries also tend to accumulate different sets of mutations in their tumours. For example, $\sim$ 50 % of men of western European ancestry harbour a fusion of an ETS gene family member [**fraserGenomicHallmarksLocalized2017**],

whereas only ~ 10 % of men of Asian ancestry harboured a similar mutation [**liGenomicEpigenomicAtlas2020**]. 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 [**fraserGenomicHallmarksLocalized2017**, **pcf/su2cinternationalprostatecance** **abeshouseMolecularTaxonomyPrimary2015**]. 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 [**kronTMPRSS2ERGFusion2017**]. 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) [**grassoMutationalLandscapeLethal2012**, **robinsonIntegrativeClinicalGenomics2015**, **quigleyGenomicH** Over-expression of *AR* is associated with castration resistance, reducing the effectiveness of androgen deprivation therapies [**quigleyGenomicHallmarksStructural2018**, **daskivichRecentProgressHormonal2006** Importantly, *FOXA1* is a pioneer TF that regulates *AR* expression, and over-expression of *FOXA1* is also more frequently found in metastatic than primary tumours [**tengPioneerProstateCancer2021**]. 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 [**brayGlobalCancerStatistics2018**, **brennerProjectedEstimatesCancer2020**]. 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 [**quigleyGenomicHallmarksStructural2018**]. B-ALL is an acute clonal expansion of primitive cells restricted to the lymphoid hematopoietic lineage of B-cells and primar-

ily occurs in children [**hungerAcuteLymphoblasticLeukemia2015**]. Currently, overall survival of pediatric B-ALL is $\sim$ 90 % [**hungerAcuteLymphoblasticLeukemia2015**], yet disease relapse after treatment still occurs in 10 - 15 % of patients [**inabaAcuteLymphoblasticLeukaemia2013**, **heikampNextGenerationEvaluationTreatment2018**]. 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 [**inabaAcuteLymphoblasticLeukae**
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 [**hungerAcuteLymphoblasticLeukemia2015**, **inabaAcuteLymphoblasticLeukaemia2013**]. Newly diagnosed patients typically undergo remission-induction therapy, intensification/consolidation therapy, and continuation/maintenance therapy over the span of 2 years [**inabaAcuteLymphoblasticLeukaemia2013**]. 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 [**inabaAcuteLymphoblasticLeukaem**
hungerAcuteLymphoblasticLeukemia2015]. 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*, *KMT2A*, *CRLF2*, and *IKZF1* genes are all associated with poorer outcomes [**inabaAcuteLymphoblasticLeukaemia2013**, **hungerAcuteLymphoblasticLeukemia2015**]. Many of these affected genes regulate B-cell development, such as *PAX5* [**liuPax5LossImposes2014**, **dangPAX5TumorSuppressor2015**, **mullighanGenomewideAnalysisGenetic2007**], *IKZF1* [**mullighanGeno**
and *EBF1* [**bollerDefiningCellChromatin2018**, **nuttTranscriptionalRegulationCell2007**]. 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 [**slanyMLLFusionProteins2020**, **krivtsovMLLTranslocationsHistone2007**, **raoHijackedCancerKMT22015**, **parkPHD3DomainMLL2010**, **liStructuralBasisActivity2016**, **dasBindingHistoneChaperone2014**]. Mutations in these genes are enriched in relapse [**hungerAcuteLymphoblasticLeukemia2015**, **mullighanGenomicAnalysisClonal2008**
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* [**leeGlobalDNAMethylation2012**]. 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 [**leeEpigeneticRemodelingBcell2015**, **nordlundGenomewideSignaturesDifferential2013**]. 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 [**gengIntegrativeEpigenomicAnalysis2012**]. 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 Chapter 2 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 Chapter 3. 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 Chapter 4 and characterize the method.

Given the shared importance of mutations to TFs and epigenetic enzymes in prostate cancer and leukemias, in Chapter 5 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

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

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

zhouNoncodingMutationsTarget2020

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

2.1 Abstract

Prostate cancer is the second most commonly diagnosed malignancy among men worldwide. Recurrently mutated in primary and metastatic prostate tumours, *FOXA1* encodes a pioneer transcription factor involved in disease onset and progression through both androgen receptor-dependent and androgen receptor (*AR*)-independent mechanisms. Despite its oncogenic properties however, the

regulation of *FOXA1* expression remains unknown. Here, we identify a set of six *cis*-regulatory elements in the *FOXA1* regulatory plexus harboring somatic single nucleotide variants in primary prostate tumours. We find that deletion and repression of these *cis*-regulatory elements significantly decreases *FOXA1* expression and prostate cancer cell growth. Six of the ten single nucleotide variants mapping to *FOXA1* regulatory plexus significantly alter the transactivation potential of *cis*-regulatory elements by modulating the binding of transcription factors. Collectively, our results identify *cis*-regulatory elements within the *FOXA1* plexus mutated in primary prostate tumours as potential targets for therapeutic intervention.

2.2 Introduction

Prostate cancer (PCa) is the second most commonly diagnosed cancer among men with an estimated 1.3 million new cases worldwide in 2018 [**brayGlobalCancerStatistics2018**]. Although most men diagnosed with primary PCa are treated with curative intent through surgery or radiation therapy, treatments fail in 30% of patients within 10 years [**boorjianLongTermOutcomeRadical2007**] resulting in a metastatic disease [**litwinDiagnosisTreatmentProstate2017**]. Patients with metastatic disease are typically treated with anti-androgen therapies, the staple of aggressive PCa treatment [**attardProstateCancer2016**]. Despite the efficacy of these therapies, recurrence ultimately develops into lethal metastatic castration-resistant prostate cancer (mCRPC) [**attardProstateCancer2016**]. As such, there remains a need to improve our biological understanding of PCa development and find novel strategies to treat patients. Sequencing efforts identified coding somatic single nucleotide variants (SNVs) mapping to *FOXA1* in up to 9% [**abeshouseMolecularTaxonomyPrimary2015**, **fraserGenomicHallmarksLocalized2017**, **barbieriExomeSequencingIdentifies2012**, **grassoMutationalLan** **paroliaDistinctStructuralClasses2019**, **adamsFOXA1MutationsAlter2019**] and 13% [**paroliaDistinctStru** **adamsFOXA1MutationsAlter2019**, **robinsonIntegrativeClinicalGenomics2015**] of primary and metastatic PCa patients, respectively. These coding somatic SNVs target the Forkhead and transactivation domains of FOXA1 [**robinsonFOXA1MutationsHormonedependent2013**], altering its pioneering functions to promote PCa development [**adamsFOXA1MutationsAlter2019**, **gaoForkheadDomainMutations2019**]. Outside of coding SNVs, whole genome sequencing (WGS) also identified somatic SNVs and indels in the 3' untranslated region (UTR) and C-terminus of FOXA1 in $\sim$ 12% of mCRPC patients [**annalaFrequentMutationFOXA12018**]. In addition to SNVs, the *FOXA1* locus is a target of structural rearrangements in both primary and metastatic PCa tumours, inclusive of duplications, amplifications, and translocations [**paroliaDistinctStructuralClasses2019**,

adamsFOXA1MutationsAlter2019]. Taken together, *FOXA1* is recurrently mutated taking into account both its coding and flanking noncoding sequences across various stages of PCa development.

FOXA1 serves as a pioneer transcription factor (TF) that can bind to heterochromatin, promoting its remodelling to increase accessibility for the recruitment of other TFs [**yangCurrentPerspectivesFOXA12015**]. FOXA1 binds to chromatin at cell-type specific genomic coordinates facilitated by the presence of mono- and dimethylated lysine 4 of histone H3 (H3K4me1 and H3K4me2) histone modifications [**lupienFoxA1TranslatesEpigenetic2008**, **eeckhouteCelltypeSelectiveChromatin2008**]. In PCa, FOXA1 is known to pioneer and reprogram the binding of *AR* alongside HOXB13 [**pomerantzAndrogenRecep** Independent from its role in *AR* signalling, FOXA1 also regulates the expression of genes involved in cell cycle regulation in PCa [**imamuraFOXA1PromotesTumor2012**, **imamuraFOXA1PromotesTumor2012**, **xuAndrogensInduceProstate2006**]. For instance, FOXA1 co-localizes with CREB1 to regulate the transcription of genes involved in cell cycle processes, nuclear division and mitosis in mCRPC [**imamuraFOXA1PromotesTumor2012**, **jinAndrogenReceptorIndependentFunction2013**, **xuAndrogensInduceProstate2006**, **yangFOXA1PotentiatesLineagespecific2016**, **zhangFOXA1DefinesC** **augelloFOXA1MasterSteroid2011**, **sunkelIntegrativeAnalysisIdentifies2017**]. FOXA1 has also been shown to promote feed-forward mechanisms to drive disease progression [**niAmplitudeModulationAndro** **sasseFeedforwardTranscriptionalProgramming2015**]. Hence, FOXA1 contributes to *AR*-dependent and *AR*-independent processes favouring PCa development.

Despite the oncogenic roles of FOXA1, therapeutic avenues to inhibit its activity in PCa are lacking. In the breast cancer setting for instance, the use of cyclin-dependent kinases inhibitors have been suggested based on their ability to block FOXA1 activity on chromatin [**wangHighThroughputChemical2018**]. As such, understanding the governance of *FOXA1* messenger RNA (mRNA) expression offers an alternative strategy to find modulators of its activity. Gene expression relies on the interplay between distal *cis*-regulatory elements (CREs), such as enhancers and anchors of chromatin interaction, and their target gene promoter(s) [**rowleyOrganizationalPrinciples3D2018**]. These elements can lie tens to hundreds of kilobases away from each other on the linear genome but physically engage in close proximity with each other in the three-dimensional space [**vernimmenHierarchyTranscriptionalActivation2015**] By measuring contact frequencies between loci through the use of chromatin conformation capture (3C)-based technologies, it enables the identification of regulatory plexuses corresponding to sets of CREs in contact with each other [**sallariConvergenceDispersedRegulatory2016**, **baileyNoncodingSomaticInherited2016**]. By leveraging these technologies, we can begin to understand the three-dimensional organization of the PCa genome and delineate the *FOXA1* regulatory plexus.

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

2.3 Results

2.3.1 *FOXA1* is essential for prostate cancer proliferation

We interrogated *FOXA1* expression levels across cancer types. We find that *FOXA1* mRNA is consistently the most abundant in prostate tumours compared to 25 other cancer types across patients (Figure 2.1a), ranking in the 95th percentile for 492 of 497 prostate tumours profiled in The Cancer Genome Atlas (TCGA) (Figure A.1a). Using the same dataset we also find that *FOXA1* is the most highly expressed out of 41 other forkhead box (FOX) factors in prostate tumours (Figure A.1b). We next analyzed expression data from Cancer Dependency Map (DEPMAP) and observed *FOXA1* to be most highly expressed in PCa cell lines compared to cell lines of other cancer types (Figure A.2a). Amongst the eight PCa cell lines in the dataset (22Rv1, DU145, LNCaP, MDA-PCa-2B, NCI-H660, PrECLH, PC3, and VCaP), *FOXA1* mRNA abundance is above the 90th percentile in all but one cell line (PrECLH) compared to the $>$ 56,000 protein coding and non-protein coding genes profiled (Figure A.2b). These new results gained from the TCGA and DEPMAP validate previous understanding that *FOXA1* is one of the highest expressed genes in PCa [**tsourlakisFOXA1ExpressionStrong2017**].

Following up on *FOXA1* mRNA expression levels, we interrogated the essentiality of *FOXA1* for PCa cell growth. RNA interference (RNAi)-mediated essentiality screens compiled in DEPMAP show that *FOXA1* lies in the 94th percentile across 6 of the 8 available PCa cell lines: 22Rv1, LNCaP, MDA PCa 2B, NCI-H660, PC3, and VCaP cells (Figure 2.1b-c). The median RNAi-mediated essentiality score for all prostate cell lines is significantly lower than all other cell lines, suggesting that *FOXA1* is especially essential for PCa cell proliferation (permutation test, $p = 1 \times 10^{-6}$, see Section 2.5; Figure A.3a). Growth assays in LNCaP and VCaP cells following *FOXA1* knockdown using two independent small interfering RNA (siRNA)s (Figure 2.1d, Figure A.3b) show significant growth inhibition in LNCaP (siRNA #1: 4-fold, siRNA #2: 3.35-fold) and VCaP (siRNA #1: 8.7-fold, siRNA #2: 2-fold) cells five days post-transfection (Mann-Whitney U Test, $p < 0.05$;

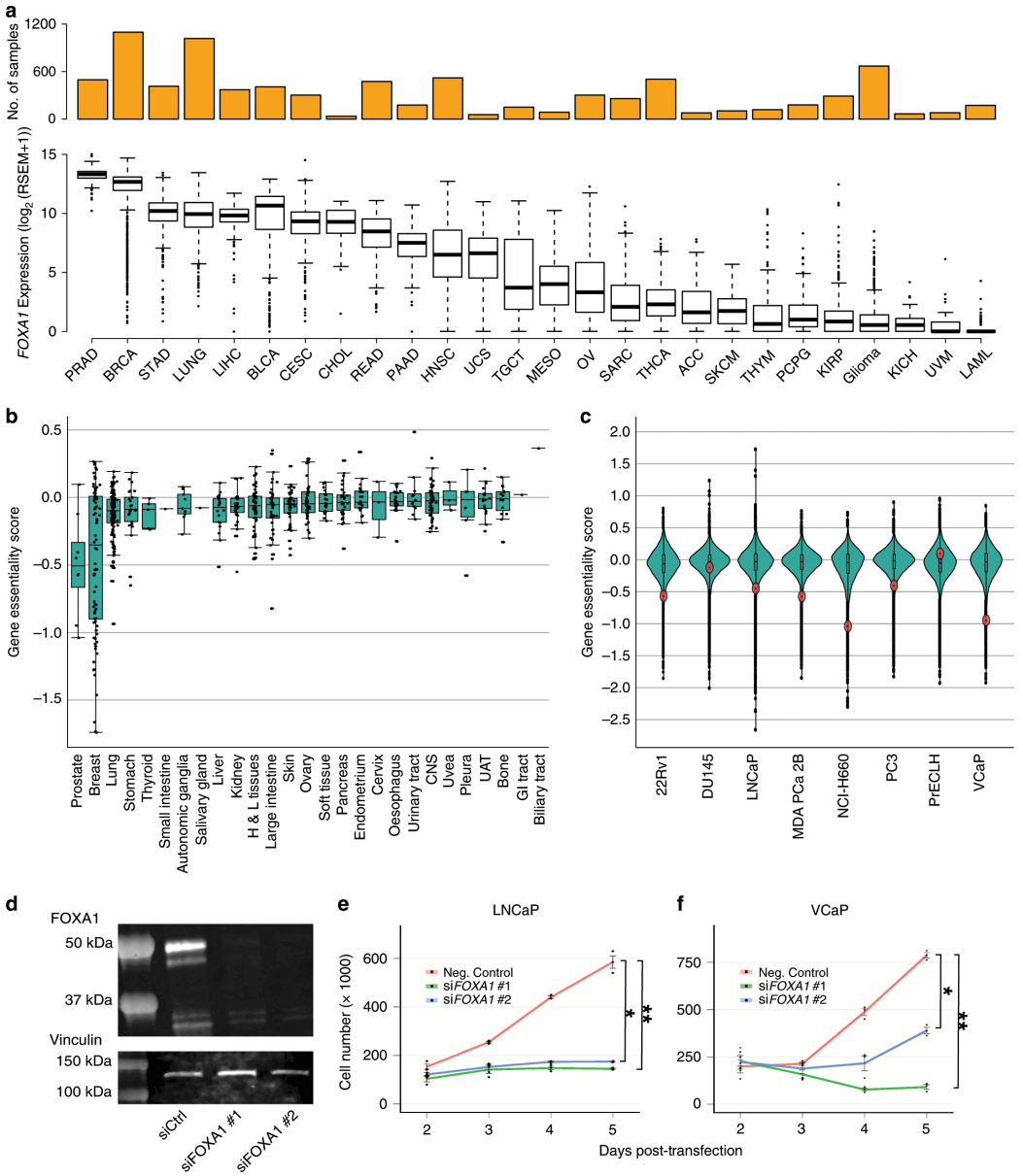

Figure 2.1: ***FOXA1* is highly expressed in PCa and essential for PCa cell proliferation..** **a.** The mRNA expression of *FOXA1* across tumour types ($n = 26$) from RNA-seq data of TCGA. **b.** *FOXA1* essentiality mediated through RNAi across various cell lines ($n = 707$) from DEPMAP. Gene essentiality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

Figure 2.1e-f). In accordance with previous reports, our results using essentiality datasets followed by knockdown validation reveals that *FOXA1* is oncogenic and essential for PCa cell proliferation.

2.3.2 Identifying putative *FOXA1* CREs

The interweaving of distal CREs with target gene promoters establishes regulatory plexuses with some to be ascribed to specific genes [**sallariConvergenceDispersedRegulatory2016**, **baileyNoncodingSomati** Regulatory plexuses stem from chromatin interactions orchestrated by various factors including ZNF143, YY1, CTCF and the cohesin complex [**phillipsCTCFMasterWeaver2009**, **weintraubYY1StructuralR** **baileyZNF143ProvidesSequence2015**]. Motivated by the oncogenic role of *FOXA1* in PCa, we investigated its regulatory plexus controlling its expression. According to chromatin contact frequency maps generated from Hi-C assays performed in LNCaP PCa cells, *FOXA1* lies in a 440 kbp topologically associated domain (TAD) (chr14: $37720001 - 38160000 \pm 40$ kbp adjusting for resolution; Figure 2.2a). By overlaying DNase-seq data from LNCaP PCa cells, there are a total of 123 putative CREs reported as DNase I hypersensitive sitess (DHSs) that populate this TAD (Figure 2.2a). We next inferred the regulatory plexus of *FOXA1* using the C3D method [**mehdiC3DToolPredict2019**]. C3D aggregates and draws correlation of DHS signal intensities between the cell line of choice and the DHS signal across all systems in a collection of cell lines and tissues [**mehdiC3DToolPredict2019**]. Anchoring our analysis to the *FOXA1* promoter and using accessible chromatin regions defined in LNCaP PCa cells identified 55 putative CREs to the *FOXA1* regulatory plexus ($r > 0.7$; Figure 2.2b).

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

To delineate the CREs that could be actively involved in the transcriptional regulation of *FOXA1*, we annotated the DHS from LNCaP cells with available chromatin immunoprecipitation sequencing (ChIP-seq) data for histone modifications and TFs conducted in LNCaP, 22Rv1, VCaP PCa cell lines and primary prostate tumours (Figure 2.2b) [**pomerantzAndrogenReceptorCistrome2015**, **kronTMPRSS2ER** Close to 60% (33/55) of the putative *FOXA1* plexus CREs are marked by H3K27ac profiled in primary prostate tumours [**kronTMPRSS2ERGFusion2017**], indicative of active CREs in tumours (Figure 2.2b) [**creyghtonHistoneH3K27acSeparates2010**]. Next, considering that noncoding SNVs can target a set of CREs that converge on the same target gene in cancer [**baileyNoncodingSomaticInherited** we overlapped the somatic SNVs called from WGS across 200 primary prostate tumours to the 33 H3K27ac-marked DHS predicted to regulate *FOXA1* [**fraserGenomicHallmarksLocalized2017**,

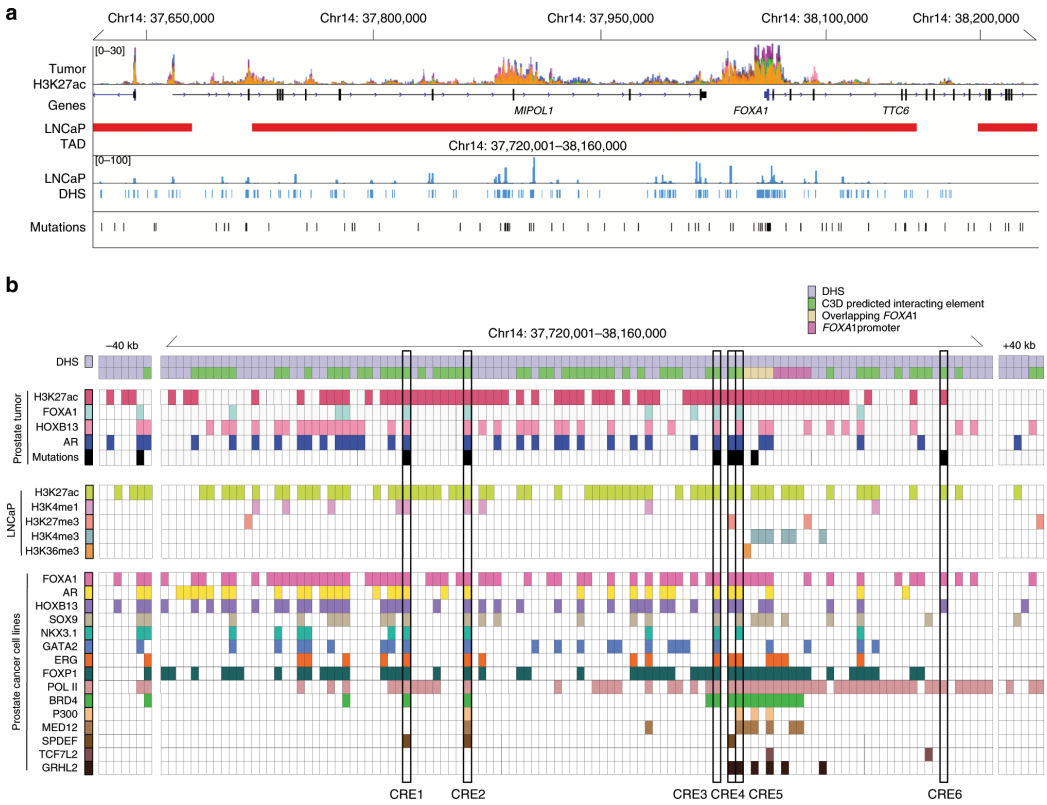

Figure 2.2: **Epigenetic annotation of 14q21.1 locus and identification of *FOXA1* CREs. a.** Overview of *cis*-regulatory landscape surrounding *FOXA1* on the 14q21.1 locus. H3K27ac signal track is the ChIP-seq signal overlay of 19 primary prostate tumours. LNCaP Hi-C depicts the TAD structure around *FOXA1*. LNCaP signal values with peak calls below indicate the accessibility of the chromatin around *FOXA1*. Mutations indicate SNVs identified in 200 primary prostate tumours. **b.** Functional annotation of putative *FOXA1* CREs using TF and histone modification ChIP-seq conducted in primary tumours and PCa cell lines. Annotated in the matrix are all DHS within the TAD and ± 40 kbp resolution left and right of the TAD. Putative *FOXA1* CREs targeted by noncoding SNVs for downstream validation are boxed. Coordinates are 1-based in the GRCh37 reference genome.

espirituEvolutionaryLandscapeLocalized2018]. This analysis identified 6 out of the 33 DHS marked with H3K27ac (18.2%) harboring one or more SNVs (10 total SNVs called from 9 tumours; Figure 2.2b). We observe that these 6 CREs can be bound by multiple TFs in PCa cells, including *FOXA1*, AR and *HOXB13* (Figure 2.2b, Figure A.4). The Hi-C data from the LNCaP PCa cells corroborates the C3D predictions as demonstrated by the elevated contact frequency between the region harboring the *FOXA1* promoter and where the 6 CREs are located, compared to other loci in the same TAD (Figure 2.3a). The 6 CREs lie in intergenic or intronic regions (Figure 2.3b-h). Together, histone modifications, TF binding sites and noncoding SNVs support that these 6 putative CREs are active in primary PCa. The Hi-C and C3D predictions suggest that they regulate *FOXA1* expression.

2.3.4 Disruption of CREs reduces *FOXA1* mRNA expression

We next assessed the role of CREs toward *FOXA1* expression using LNCaP and 22Rv1 clones stably expressing the wild-type (WT) Cas9 protein (Figure 2.4a-b). Guide RNAs (gRNAs) designed against the *FOXA1* gene (exon 1 and intron 1) served as positive controls while an outside-TAD region (termed Chr14 (-)), a region on a different chromosome (the human *AAVS1* safe-harbor site at the *PPP1R12C* locus [**kronTMPRSS2ERGFusion2017**, **dekelverFunctionalGenomicsProteomics2010**]), and three regions within the TAD predicted to be excluded from the *FOXA1* plexus served as negative controls. Individual deletion of the *FOXA1* plexus CREs through transient transfection of gRNAs into the LNCaP cells (see Section 2.5) led to significantly decreased *FOXA1* mRNA expression (ΔCRE1 $\sim$ 29.3 $\pm$ 8.3%, ΔCRE2 $\sim$ 40.1 $\pm$ 11.8%, ΔCRE3 $\sim$ 30.6 $\pm$ 9.1%, ΔCRE4 $\sim$ 23.6 $\pm$ 8.2%, ΔCRE5 $\sim$ 25.3 $\pm$ 6.6%, ΔCRE6 $\sim$ 24.5 $\pm$ 10.2% and Δ*FOXA1* (exon 1 and intron 1) $\sim$ 87.4 $\pm$ 8.8% reduction relative to basal levels; Figure 2.4c, Figure A.5a-f). In contrast, deletion of several negative control regions within the same TAD did not significantly reduce *FOXA1* mRNA level (Figure 2.4c, Figure A.5g-i). Similar results were observed in 22Rv1 PCa cells (Figure 2.4d). As each clone expressed Cas9 protein at different levels, there may be a difference between genome editing efficiencies between the clones. We compared the CRISPR/Cas9 on-target genome editing efficiency across the five LNCaP cell line-derived clones with the relative *FOXA1* mRNA levels, and indeed observe a significant inverse correlation across all CREs (Pearson's correlation $r = 0.49, p < 0.005$; Figure A.6a) and agreeing trends for each individual CRE (Figure A.6b).

Complementary to our findings using the WT CRISPR/Cas9 system, we next generated four LNCaP and four 22Rv1 cell line-derived dCas9-KRAB fusion protein expressing clones (Figure 2.4e-

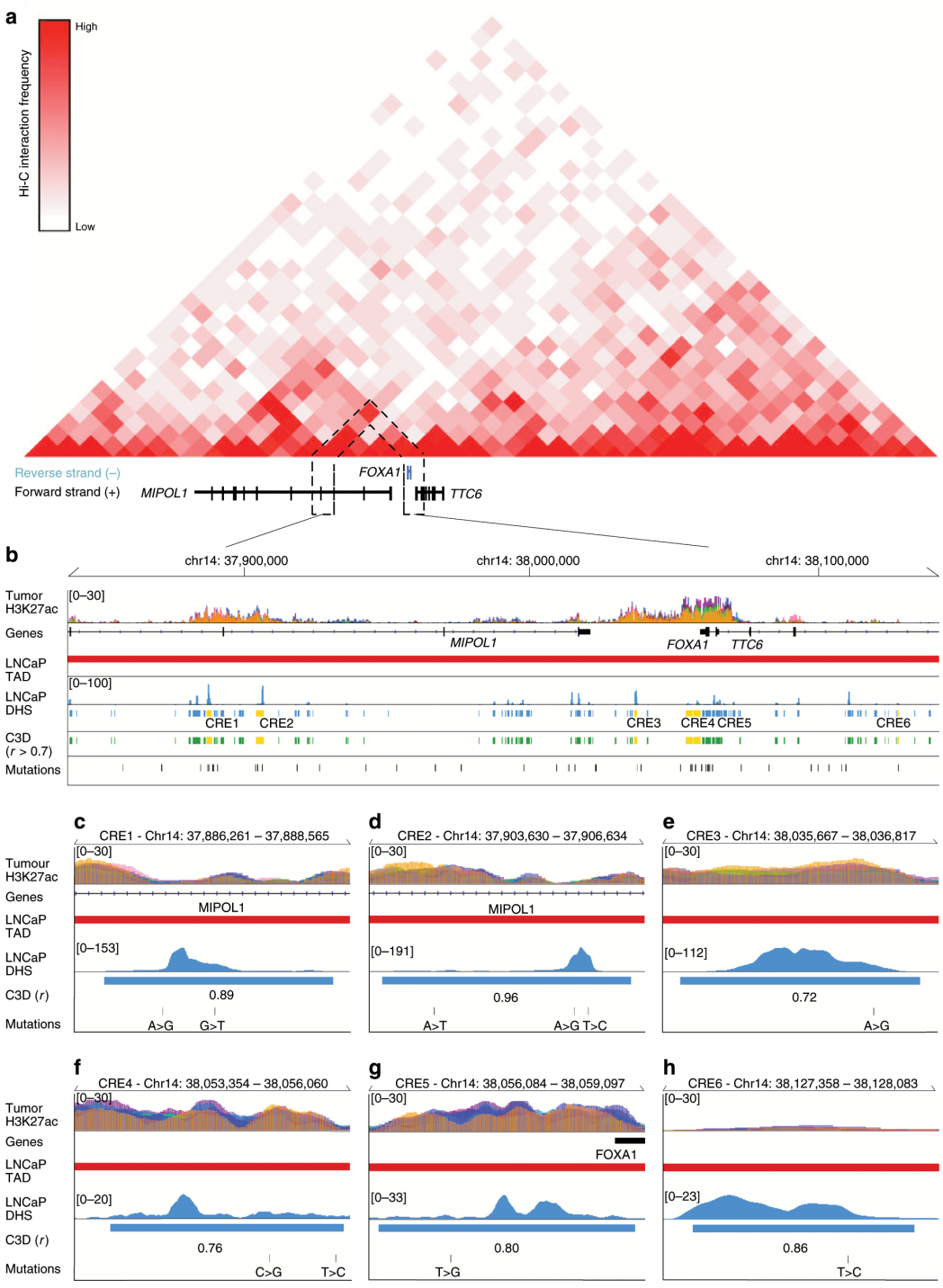

Figure 2.3: **Putative CREs predicted to interact with *FOXA1* promoter. a.** Hi-C contact matrix conducted in LNCaP cells indicating physical interactions between putative *FOXA1* CREs and the *FOXA1* promoter. Hi-C resolution is 40 kbp. **b.** Epigenome annotations around the *FOXA1* locus. The six putative *FOXA1* CREs are coloured in yellow. **c-h.** Zoom-in of each individual putative *FOXA1* CRE. C3D value is the Pearson correlation of DHS signal between LNCaP and the DHS reference matrix. Coordinates are 1-based in the GRCh37 reference genome.

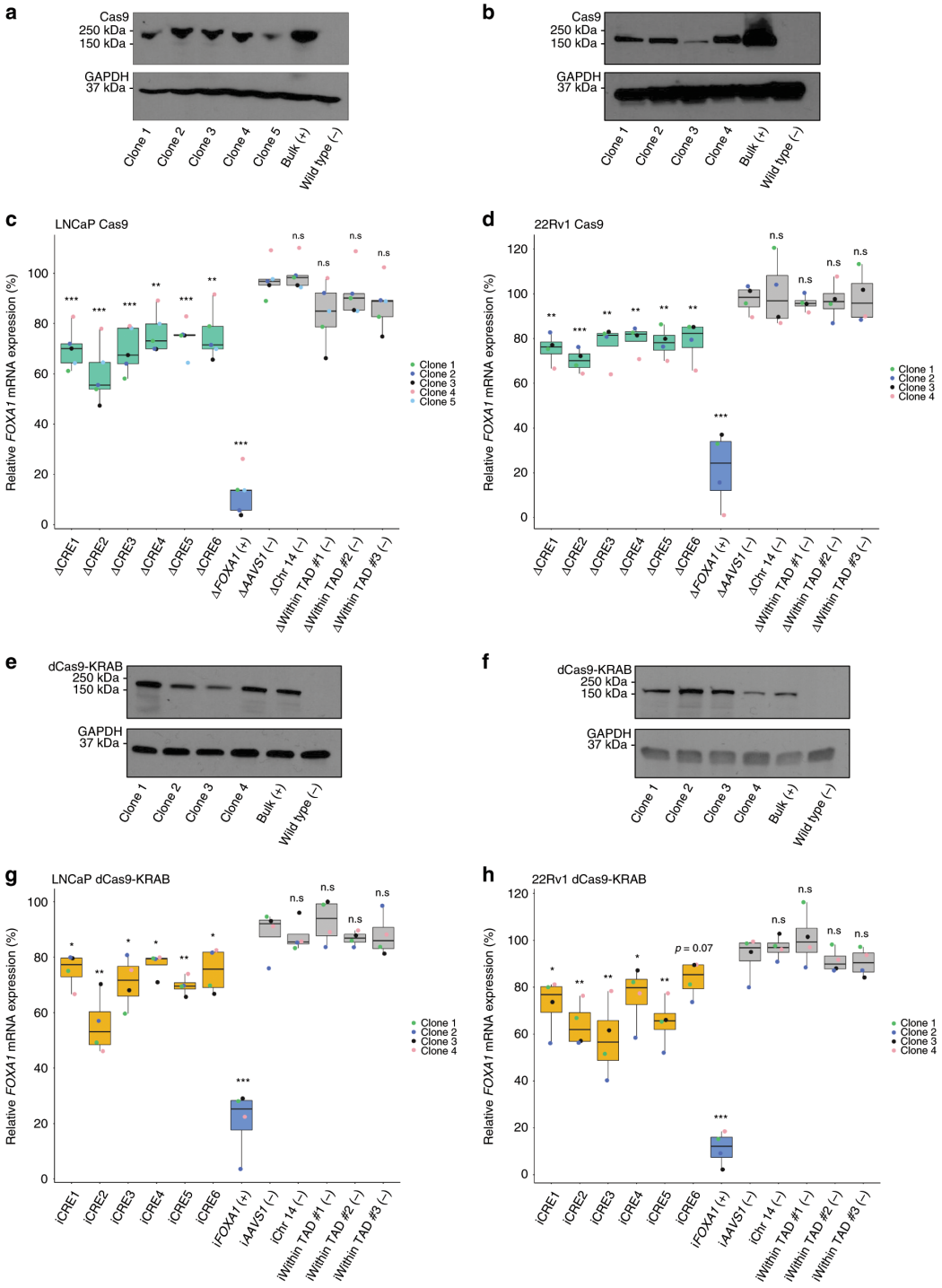

Figure 2.4: **Functional dissection of putative *FOXA1* CREs.** (Continued on the following page)

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

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

We further assessed the regulatory activity of the six *FOXA1* plexus CREs by testing the consequent mRNA expression on other genes within the same TAD, namely *MIPOL1* and *TTC6*. ΔCRE1 and ΔCRE2 significantly reduced *MIPOL1* mRNA expression by $\sim$ 38.4 $\pm$ 6.4% and $\sim$ 48.4 $\pm$ 9%, respectively relative to basal levels, whereas deletion of the other four CREs did not result in any significant *MIPOL1* expression changes (Student's t-test, $p < 0.05$, Figure A.7a). On the other hand, deletion of CREs each significantly reduced *TTC6* mRNA expression relative to its basal levels (ΔCRE1 $\sim$ 52.9% $\pm$ 6.4%, ΔCRE2 $\sim$ 66 $\pm$ 11.3%, ΔCRE3 $\sim$ 55.5 $\pm$ 12.8%, ΔCRE4 44.9 $\pm$ 10.6%, ΔCRE5 43.1 $\pm$ 11.9% and ΔCRE6 52.2 $\pm$ 7.3% reduction (Student's t-test, $p < 0.05$, Figure A.7b), in agreement with the fact that *TTC6* shares its promoter with *FOXA1* as both genes are transcribed on opposing strands (Figure A.7c).

Reduction in *FOXA1* mRNA expression resulting from the deletion of *FOXA1* plexus CREs may

also impact gene expression downstream of *FOXA1*, we assessed the mRNA expression of several *FOXA1* target genes, namely *SNAI2*, *ACPP*, and *GRIN3A*. Deletion of CREs resulted in significant change in *SNAI2* (up-regulation; ΔCRE1 ∼ 190%, ΔCRE2 ∼ 162.8%, ΔCRE3 ∼ 147.5%, ΔCRE4 ∼ 133.3%, ΔCRE5 ∼ 137.3%, ΔCRE6 ∼ 120.8%, Δ*FOXA1* ∼ 266.7%), *ACPP* (down-regulation; ΔCRE1 ∼ 73.5%, ΔCRE2 ∼ 62.5%, ΔCRE3 ∼ 69.6%, ΔCRE4 ∼ 75.6%, ΔCRE5 ∼ 70.9%, ΔCRE6 ∼ 74.6%, Δ*FOXA1* ∼ 52.2%) and *GRIN3A* expression (up-regulation; ΔCRE1 ∼ 138.2%, ΔCRE2 ∼ 168.8%, ΔCRE3 ∼ 144.6%, ΔCRE4 ∼ 132.1%, ΔCRE5 ∼ 131.4%, ΔCRE6 ∼ 127%, Δ*FOXA1* ∼ 228%; Student's t-test, $p < 0.05$, Figure A.7d-f). Collectively, our results support the restriction of most *FOXA1* plexus CREs towards *FOXA1* and its target genes.

2.3.5 *FOXA1* CREs collaborate to regulate its expression

Expanding on the idea that multiple CREs can converge to regulate the expression of a single target gene [**sallariConvergenceDispersedRegulatory2016**, **baileyNoncodingSomaticInherited2016**, **pennacchioEnhancersFiveEssential2013**], we asked whether the CREs we identified collaboratively regulate *FOXA1* mRNA expression. Here, we applied a transient approach that delivers Cas9 protein:gRNA as a ribonucleoprotein (RNP) complex formed prior to transfection that would avoid the heterogeneity of Cas9 protein expression across the PCa cell clones (see Section 2.5). We first validated this system through single CRE deletions, where we transiently transfected a set of gRNA targeting the CRE of interest. In accordance with data from our PCa cell clones stably expressing WT Cas9 and dCas9-KRAB, individual CRE deletion resulted in a significant reduction in *FOXA1* mRNA expression: (ΔCRE1 ∼ 29.3 ± 7.3%, ΔCRE2 ∼ 36 ± 11.8%, ΔCRE3 ∼ 30.6 ± 12.7%, ΔCRE4 ∼ 24.5 ± 6.1%, ΔCRE5 ∼ 23.7 ± 13.2%, ΔCRE6 ∼ 26.8 ± 14.2% and Δ*FOXA1* ∼ 96.2 ± 1.4% reduction; Student's t-test, $p < 0.05$, Figure 2.5a, Figure A.8a-f). Next for combinatorial deletions, we prioritized the CREs that harbor more than 1 SNV (i.e. CRE1, CRE2, CRE4), and transiently transfected RNP complexes that target both CREs in various combinations (i.e. CRE1 + CRE2, CRE1 + CRE4, CRE2 + CRE4), and assessed *FOXA1* mRNA expression. Compared to negative control regions, the combinatorial deletion of ΔCRE1 + ΔCRE2, ΔCRE1 + ΔCRE4, and ΔCRE2 + ΔCRE4 resulted in a significant ∼ 48.5 ± 4.5%, ∼ 50.4 ± 2.9% and ∼ 45.2 ± 5.5% reduction in *FOXA1* mRNA expression, respectively (Student's t-test, $p < 0.05$, Figure 2.5b, Figure A.9a-f) a fold reduction greater than single CRE deletions (Student's t-test, Figure A.10, $p < 0.05$). These results together demonstrate that these CREs collaboratively contribute to the establishment and regulation of *FOXA1* expression in PCa.

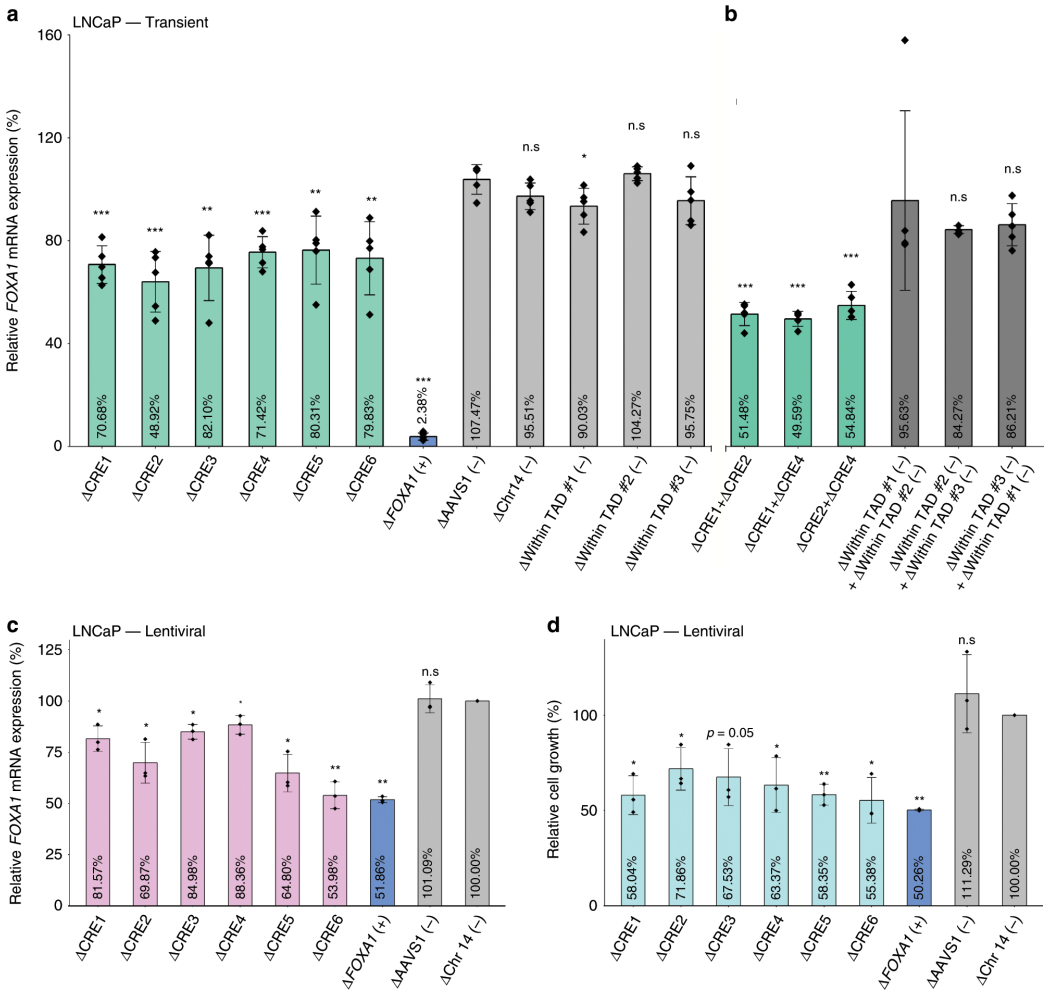

Figure 2.5: ***FOXA1* CREs collaborate to regulate its expression and are critical for PCa cell proliferation. a.** *FOXA1* mRNA expression normalized to housekeeping *TBP* mRNA expression upon transient transfection-based CRISPR/Cas9-mediated deletion of CRE1, CRE2, CRE4, and sequential deletion combinations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

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

As *FOXA1* is essential for PCa growth (Figure 2.1b-e), we next sought to assess the importance of the six CREs in the *FOXA1* plexus towards PCa cell growth. We adapted a lentiviral-based approach that expressed both the Cas9 protein and two gRNA that target each CRE for deletion (see Section 2.5). Upon lentiviral transduction with subsequent selection, we separated LNCaP PCa cells for RNA, DNA and for cell proliferation. We first tested the system by measuring *FOXA1* mRNA expression, and independently observed significant reductions of *FOXA1* mRNA expression (ΔCRE1 $\sim$ 18%, ΔCRE2 $\sim$ 30%, ΔCRE3 $\sim$ 15%, ΔCRE4 $\sim$ 12%, ΔCRE5 $\sim$ 35%, ΔCRE6 $\sim$ 46% and Δ *FOXA1* (exon 1 and intron 1) $\sim$ 48% reduction (Student's t-test, $p < 0.05$, Figure 2.5c, Figure A.11a-f). We then seeded these cells at equal density. Six days post-seeding, we harvested the cells and observed a significant reduction in cell growth upon deleting any of the six *FOXA1* plexus CREs (ΔCRE1 $\sim$ 42%, ΔCRE2 $\sim$ 28%, ΔCRE3 $\sim$ 33%, ΔCRE4 $\sim$ 27%, ΔCRE5 $\sim$ 42%, ΔCRE6 $\sim$ 44% and Δ *FOXA1* (exon 1 and intron 1) $\sim$ 50% reduction (Student's t-test, $p < 0.05$, Fig 5d). These results suggest that the six *FOXA1* plexus contribute to PCa etiology, in agreement with their ability to regulate *FOXA1* expression and the essentiality of this gene in PCa cell growth.

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

SNVs can alter the transactivation potential of CREs [**baileyNoncodingSomaticInherited2016**, **rheinbayRecurrentFunctionalRegulatory2017**, **zhangIntegrativeFunctionalGenomics2012**, **huangHighlyRecurrentTERT2013**, **hornTERTPromoterMutations2013**, **fuxmanbassHumanGeneCente** **zhouEmergenceNoncodingCancer2016**, **feiginRecurrentNoncodingRegulatory2017**, **khuranaRoleNonco** **cowper-sal*lariBreastCancerRisk2012**]. In total, we found 10 SNVs called from 9 out of the 200 tumours that map to the six *FOXA1* plexus CREs (Figure 2.6a). To assess the impact of these noncoding SNVs, we conducted luciferase assays comparing differential reporter activity between the variant and the WT allele of each CRE (Figure 2.6b-k). We found that the variant alleles of 6 of the 10 SNVs displayed significantly greater luciferase reporter activity when compared to the WT alleles (Mann-Whitney U test, $p < 0.05$). Specifically, we observed the following fold changes: chr14:37,887,005 A > G (1.65-fold), chr14:37,904,343 A > T (1.35-fold), chr14:37,905,854 A > G (1.28-fold), chr14:37,906,009 T > C (1.71-fold), chr14:38,036,543 A > G (1.44-fold), chr14:38,055,269 C > G (1.39-fold; Figure 2.6b, d-h). These results indicate that these SNVs can alter the transactivation potential of *FOXA1* plexus CREs in PCa cells.

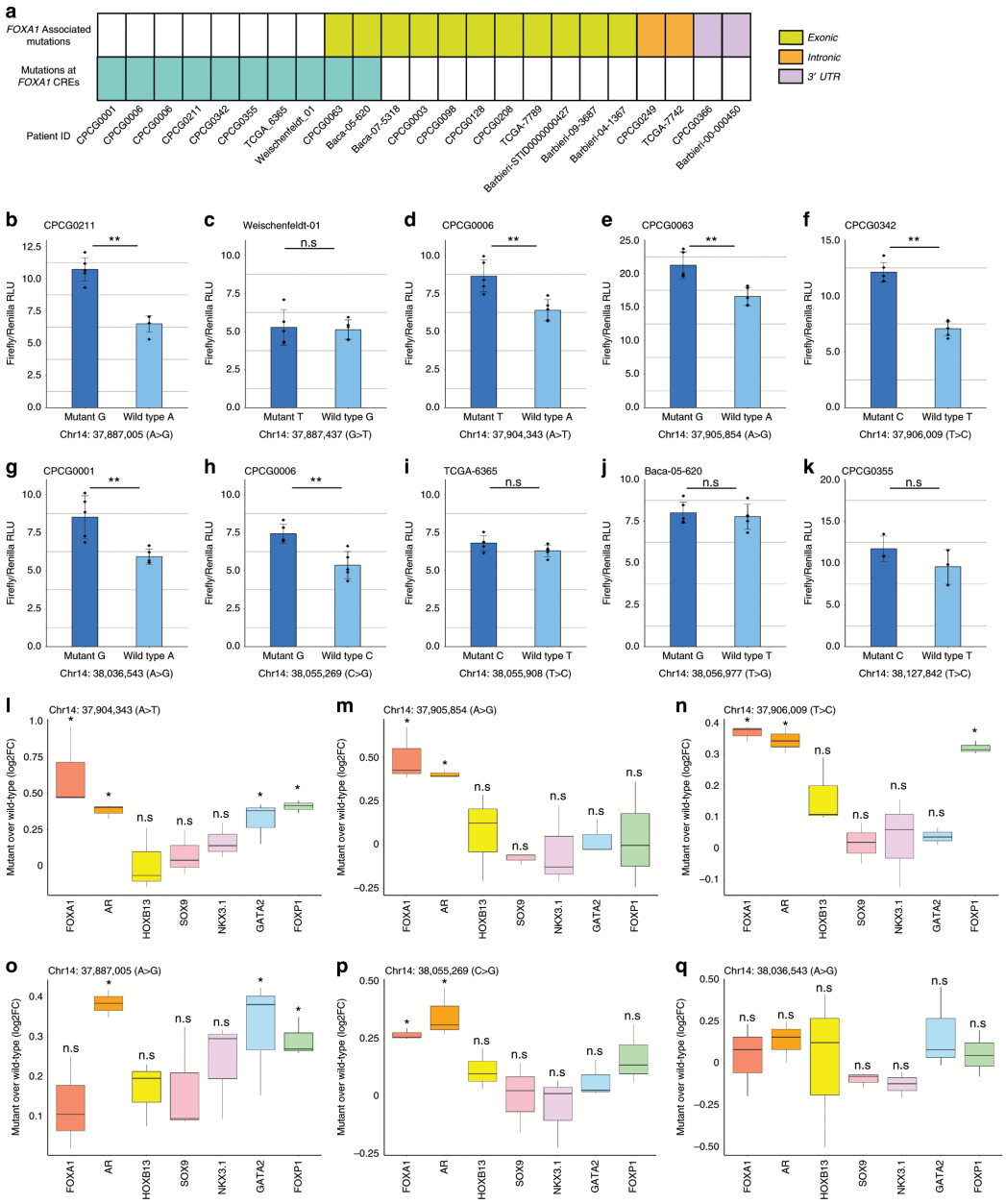

Figure 2.6: **A subset of noncoding SNVs mapping to the *FOXA1* CREs are gain-of-function. a.** Matrix showcasing the patients from the CPC-GENE dataset that harbour SNVs at the *FOXA1* CREs, exons, introns, and the 3′ UTR of *FOXA1*. **b-k.** Luciferase assays are conducted in LNCaP cells. Bar plot showcases the mean firefly luciferase activity normalized by *Renilla* luciferase activity in RLUs. Error bars indicate ± s.d. $n = 5$ independent experiments for all CREs except for chr14:38,127,842 T > C where $n = 3$. Each diamond represents an independent experiment. Hypothesis testing done with Mann-Whitney U test. **l-q.** Allele-specific ChIP-qPCR conducted on plasmids carrying the WT or variant sequence upon transient transfection in PCa cells. Data is presented as $\log_2$ fold change of variant sequence upon comparison to WT sequence ($n = 3$ independent experiments per ChIP). Hypothesis testing done with Student's t-test, n.s. not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Coordinates are 1-based in the GRCh37 reference genome.

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

We next assessed if the changes in transactivation potential induced by noncoding SNVs related to changes in TF binding to CREs by allele-specific ChIP-qPCR [**baileyNoncodingSomaticInherited2016**, **zhangIntegrativeFunctionalGenomics2012**, **cowper-sal*lariBreastCancerRisk2012**] in LNCaP PCa cells. We observed differential binding of FOXA1, AR, HOXB13, GATA2, and FOXP1 for the chr14:37887005 (A > G) SNV found in CRE1; the chr14:37904343 (A > T), chr14:37905854 (A > G) and chr14:37906009 (T > C) SNVs found in CRE2; and the chr14:38055269 (C > G) SNV found in CRE4 (Student's *t*-test, $p < 0.05$, Figure 2.6l-p). In contrast, SOX9 and NKX3.1 binding was unaffected by these SNVs (Figure 2.6l-q). Compared to the WT sequence, chr14:37,887,005 A > G significantly increased AR binding (1.31-fold increase), GATA2 binding (1.25-fold increase) and FOXP1 binding (1.23-fold increase); chr14:37,904,343 A > T significant increased AR binding (1.30-fold increase), GATA2 (1.25-fold increase) and FOXP1 (1.33-fold increase); chr14:37,905,854 A > G significantly increased FOXA1 binding (1.41-fold increase) and AR binding (1.33-fold increase); chr14:37,906,009 T > C significantly increased the binding of FOXA1 (1.29-fold increase), AR (1.31-fold increase), HOXB13 (1.13-fold increase) and FOXP1 (1.25-fold increase); and chr14:38,055,269 C > G significantly increased FOXA1 binding (1.20-fold increase). Notably all six SNVs increased the binding of the TFs known to bind at these CREs. In contrast, none of the SNVs significantly decreased the binding of these TFs. Our observations suggest that gain-of-function populate the *FOXA1* plexus CREs.

2.4 Discussion

Modern technologies and understanding of the epigenome allow the possibility of probing CREs involved in regulating genes implicated in disease. Despite *FOXA1* being recurrently mutated [**abeshouseMolecularTaxonomyPrimary2015**, **fraserGenomicHallmarksLocalized2017**, **barbieriExomeS** **grassoMutationalLandscapeLethal2012**, **robinsonIntegrativeClinicalGenomics2015**] and playing potent oncogenic roles in PCa etiology [**paroliaDistinctStructuralClasses2019**, **adamsFOXA1Mutatio** **gaoForkheadDomainMutations2019**], the CREs involved in its transcriptional regulation are poorly understood. Understanding how *FOXA1* is expressed can provide a complementary strategy to antagonize *FOXA1* in PCa.

We used the DHSs profiled in PCa cells to identify putative *FOXA1* CREs through annotating these regions with five different histone modifications, TF binding sites and noncoding SNVs profiled

in PCa cells and primary prostate tumours. Our efforts identified and validated a set of six active CREs involved in *FOXA1* regulation, agreeing with a recent report where a subset of our CREs map to loci suggested to be in contact with the *FOXA1* promoter [**rhieHighresolution3DEpigenomic2019**]. The disruption of these six distal CREs each significantly reduced *FOXA1* mRNA levels, similar to what has been demonstrated for *ESR1* in luminal breast cancer [**baileyNoncodingSomaticInherited2016**], *MLH1* in Lynch syndrome [**liuDisruption35Kb2018**], *MYC* in lung adenocarcinoma and endometrial cancer [**zhangIdentificationFocallyAmplified2016**], and *AR* in mCRPC [**takedaSomaticallyAcquiredEnh viswanathanStructuralAlterationsDriving2018**]. Through combinatorial deletion of two CREs, *FOXA1* mRNA levels were further reduced in comparison with single CRE deletions, raising the possibility of CRE additivity [**osterwalderEnhancerRedundancyProvides2018**]. The deletion of the FOXA1 plexus CREs also significantly reduced PCa cell proliferation at levels comparable to what has been reported upon deletion of the amplified CRE upstream of the *AR* gene in mCRPC [**takedaSomaticallyAcquiredEnhancer2018**], suggestive of onco-CREs as reported in lung [**zhangIdentificationFocallyAmplified2016**] and prostate [**takedaSomaticallyAcquiredEnhancer2018**] cancer.

More than 90% of SNVs found in cancer map to the noncoding genome [**meltonRecurrentSomaticMutations20 mazrooeiCistromePartitioningReveals2019**] with a portion of these SNVs mapping to CREs altering their transactivation potential [**baileyNoncodingSomaticInherited2016**, **zhangIntegrativeFunctionalG huangHighlyRecurrentTERT2013**, **hornTERTPromoterMutations2013**] and/or downstream target gene expression [**zhouEmergenceNoncodingCancer2016**, **meltonRecurrentSomaticMutations2015**, **weinholdGenomewideAnalysisNoncoding2014**]. We extended this concept with SNVs identified from primary prostate tumours mapping to FOXA1 plexus CREs. We observed that a subset of these SNVs can alter transactivation potential by modulating the binding of specific TFs whose cistromes are preferentially burdened by SNVs in primary PCa [**mazrooeiCistromePartitioningReveals2019**]. Our findings complement recent reports of SNVs found in the noncoding locus of *FOXA1* that could affect its expression [**annalaFrequentMutationFOXA12018**, **camcapstudygroupSequencingProstateCancer** The *FOXA1* plexus CREs we identified here are also reported to be target of structural variants (SVs) in both the primary and metastatic settings [**paroliaDistinctStructuralClasses2019**, **quigleyGenomicHallmarksStructural2018**], including tandem duplication in $\sim$ 14% (14/101) mCRPC tumours over CRE2 [**quigleyGenomicHallmarksStructural2018**], amplification, duplication and translocation over CRE3, CRE4, and CRE5 [**paroliaDistinctStructuralClasses2019**]. Notably, the translocation and duplication defining the FOXMIND enhancer driving *FOXA1* expression reported in primary and metastatic settings harbors the CRE3 element we characterized

[**paroliaDistinctStructuralClasses2019**]. Collectively, these studies combined with our discoveries reveal the fundamental contribution of the *FOXA1* plexus in PCa etiology. As a whole, our findings in conjunction with recent reports suggest that CREs involved in the transcriptional regulation of *FOXA1* may be hijacked in prostate tumours through various types of genetic alterations.

Despite initial treatment responses from treating aggressive primary and metastatic PCa through castration to suppress *AR* signalling [**attardProstateCancer2016**], resistance ensues as 80% of mCRPC tumours harbor either *AR* gene amplification, amplification of a CRE upstream of *AR*, or activating *AR* coding mutations [**robinsonIntegrativeClinicalGenomics2015**, **takedaSomaticallyAcquiredEnha** **quigleyGenomicHallmarksStructural2018**]. Given the *AR*-dependent [**yangCurrentPerspectivesFOXA120** **pomerantzAndrogenReceptorCistrome2015**] and *AR*-independent [**sunkelIntegrativeAnalysisIdentifies20** oncogenic activity of *FOXA1* in PCa, its inhibition is an appealing alternative therapeutic strategy. Our dissection of the *FOXA1 cis*-regulatory landscape complement recent findings through revealing loci that are important for the regulation of *FOXA1*. Theoretically, direct targeting of the CREs regulating *FOXA1* would down-regulate *FOXA1* and could therefore serve as a valid alternative to antagonize its function.

Taken together, we identified *FOXA1* CREs targeted by SNVs that are capable of altering transactivation potential through the modulation of key PCa TFs. The study supports the importance of considering CREs not only as lone occurrences but as a team that works together to regulate their target genes, particularly when considering the impact of genetic alterations. As such, our work builds a bridge between the understanding of *FOXA1* transcriptional regulation and new routes to *FOXA1* inhibition. Aligning with recent reports [**paroliaDistinctStructuralClasses2019**, **adamsFOXA1MutationsAlter2019**, **gaoForkheadDomainMutations2019**], our findings support the oncogenic nature of *FOXA1* in PCa. Gaining insight on the *cis*-regulatory plexuses of important genes such as *FOXA1* in PCa may provide new avenues to inhibit other drivers across various cancer types to halt disease progression.

2.5 Methods

2.5.1 Cell Culture

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

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

2.5.2 Prostate tumours and cancer cell lines expression

Cancer cell line mRNA abundance data were collected from DEPMAP (`https://depmap.org/portal/`; RNA sequencing (RNA-seq) TPM values from 2018q4 version with all 5 non-cancer cell lines were removed) [**thecancercelllineencyclopediaconsortiumPharmacogenomicAgreementTwo2015**] projects. Prostate tumour mRNA abundance data was collected from TCGA prostate cancer (TCGA-PRAD) project via the Xena Browser (`https://xenabrowser.net/`; dataset description: TCGA prostate adenocarcinoma gene expression by RNA-seq (polyA+ Illumina HiSeq; RSEM)).

2.5.3 Prostate cancer cell line gene essentiality

Essentiality scores were collected from the DEPMAP Project [**mcfarlandImprovedEstimationCancer2018**]. To compare gene essentiality between PCa cell lines and others, essentiality scores for *FOXA1* were collected from all available cell lines ($n = 707$). To perform a permutation test, the median expression of 8 randomly selected cell lines was calculated one million times to generate a background distribution of essentiality scores across all cell types available. The median essentiality score from the 8 PCa cell lines was calculated and its percentile within the background distribution is reported.

2.5.4 siRNA knockdown and cell proliferation assay

300,000 LNCaP cells (Day 0) were reverse transfected with siRNA (siFOXA1 using Lipofectamine mRNAmax reagent; ThermoFisher Scientific, Cat No. 13778150). Cells were counted using Countess automated cell counter (Invitrogen). Whole cell lysates LNCaP cells after siRNA-mediated *FOXA1* knockdown was collected at 96 h post-transfection in RIPA buffer. Protein concentrations were determined through the bicinchoninic acid method (ThermoFisher Scientific, Cat No. 23225). Then 25 μg of lysate was subjected to sodium dodecyl sulfate (SDS)-PAGE. Upon completion of SDS-PAGE, protein was transferred onto PVDF membrane (Bio-Rad, Cat No. 1704156). The membrane was blocked with 5% non-fat milk for 1 h at room temperature with shaking. After blocking, anti-*FOXA1* (Abcam Cat No. 23737) in 2.5% non-fat milk was added, and was incubated at 4 °C

overnight. Next day, the blot was washed and incubated with IRDye 800CW Goat Anti-Rabbit IgG secondary antibody (LI-COR, Cat No. $925-32211$) at room temperature for 1 h. The blot was then washed and assessed with the Odyssey CLX imaging system (LI-COR).

2.5.5 Identifying putative *FOXA1* CREs

Putative *FOXA1* CREs were identified through the use of C3D method based on DNase I Hypersensitivity [**mehdiC3DToolPredict2019**]. Predicted interacting DHS with a Pearson's correlation above 0.7 [**thurmanAccessibleChromatinLandscape2012**] were kept for downstream analysis.

2.5.6 Hi-C and TADs in LNCaP cells

Hi-C and TADs conducted and called, respectively, in LNCaP cells are publicly available off ENCODE portal (experiment accession ENCSR346DCU; FASTQ file accessions: ENCFF726LGW, ENCFF550SKU, ENCFF950CQX, and ENCFF411TZJ; TADs file accession: ENCFF139JCA). Visualization of the Hi-C dataset is available on the Hi-C Browser [**wang3DGenomeBrowser2018**].

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

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

Lentiviral particles were generated in 293FT cells (ThermoFisher) using the pMDG.2 and psPAX2 packaging plasmids (Addgene; #12259 and #12260, a gift from Didier Trono) alongside the Lenti-dCas9-KRAB-blast plasmid (Addgene #89567, a gift from Gary Hon) and collected 72 hrs post transfection. LNCaP and 22Rv1 cells were then transduced for $24-48$ h with equal amounts of virus followed by selection with media containing blasticidin (7.5 μg/mL for LNCaP cells, 6 μg/mL

for 22Rv1 cells). Upon selection, clones were derived by serial dilution with subsequent single cell seeding into 96-well plates containing selection media. dCas9-KRAB protein expression for each clone was then assessed through Western blotting (1 °Ms-Cas9 (Cell Signalling Technology, Cat No. #14697) 1:1000, Ms-GAPDH 1:5000 (Santa Cruz Biotechnology, Cat No. #sc47724) in 5% non-fat milk; 2 °HRP-linked Anti-Mouse IgG (Cell Signalling Technology, Cat No. #7076S) 1:10 000 in 2.5% non-fat milk. The full unprocessed blot is in the Source Data File.

For gRNA design, five to six unique CRISPR RNA (crRNA) molecules (Integrated DNA Technologies) were designed to tile across the region of interest using the CRISPOR tool (`http://crispor.tefor.net/`) [**haeusslerEvaluationOfftargetOntarget2016**] and the Zhang lab CRISPR Design tools (`http://crispr.mit.edu/`) [**hsuDNATargetingSpecificity2013**]. See published manuscript for gRNA sequences. Each crRNA and trans-activating CRISPR RNA (tracrRNA) (Integrated DNA Technologies) were duplexed according to company supplier protocol to a concentration of 50 μM. Upon generation of the clones, six guides (crRNA-tracrRNA duplexes) for each region of interest were pooled into a single tube (1 μL each guide, 6 μL per reaction) (Integrated DNA Technologies). Lastly, 1 μL (100 μM) of electroporation enhancer (Integrated DNA Technologies) was added to the mix (7 μL total) prior to transfection. The entire transfection reaction was transfected into 350,000 cells through Nucleofection (SF Solution EN120 - 4D Nucleofector, Lonza). Cells were then harvested 24 h post-transfection for RNA and DNA for RT-PCR and confirmation of deletion, respectively.

2.5.8 Transient Cas9-mediated disruption of CREs

Deletion of elements through this method were achieved through the transfection of Cas9 nuclease protein complexed with the crRNA (Integrated DNA Technologies). Briefly, five to six unique crRNA molecules (Integrated DNA Technologies) were designed to tile across the region of interest using the CRISPOR tool (`http://crispor.tefor.net/`) [**haeusslerEvaluationOfftargetOntarget2016**] and the Zhang lab CRISPR Design tools (`http://crispr.mit.edu/`) [**hsuDNATargetingSpecificity2013**]. Each crRNA and tracrRNA (Integrated DNA Technologies) were duplexed according to company supplier protocol to a concentration of 50 μM. The six crRNA-tracrRNA duplexes were pooled into a single tube (6 μL per reaction), prior to adding 1 μL (5 μg) of Alt-R S.p. HiFi Cas9 Nuclease 3NLS (Integrated DNA Technologies). The reaction was incubated at room temperature for 10 min for ribonucleoprotein (RNP) complex formation. Lastly, 1 μL (100 μM) of electroporation enhancer (Integrated DNA Technologies) was added to the mix prior to transfection. The entire transfection

reaction was transfected into 350 000 cells through Nucleofection (SF Solution EN120 - 4D Nucleofector, Lonza). Cells were then harvested 24 h post-transfection for RNA and DNA for RT-PCR and confirmation of deletion, respectively. For double deletions, two sets of gRNA-RNP complex (10 μg of Alt-R S.p. HiFi Cas9 Nuclease 3NLS) were transfected and harvested 24 h post-transfection for RNA and DNA for RT-PCR and confirmation of deletion, respectively. To control for double deletions, two negative control regions within the TAD were also compounded. Due to size, see published manuscript for primers.

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

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

2.5.10 Confirmation of Cas9-mediated deletion of CREs

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

2.5.11 Cell proliferation upon deletion of *FOXA1* CREs

Pairs of gRNAs flanking the CREs of interest, *FOXA1* promoter and control regions were designed using CRISPOR (`http://crispor.tefor.net/`) and Zhang lab CRISPR Design tool (`http://crispr.mit.edu/`) (due to size, see published manuscript). Each pair of gRNAs were cloned into the lentiCRISPRv2 (Addgene; a gift from Feng Zhang #52961) and the lentiCRISPRv2-Blast (Addgene; a gift from Feng Zhang #83480) plasmid as previously described [**sanjanaImprovedVectorsGenomewide2014**]. Lentiviral particles were generated in 293FT cells (ThermoFisher) using the pMDG.2 and psPAX2 packaging plasmids (Addgene; #12259 and #12260, a gift from Didier Trono), and collected 72

hrs post transfection. LNCaP cells were transduced for $24 - 48$ h with equal amounts of virus, followed by selection with media containing puromycin (3.5 μg/mL, ThermoFisher) and blasticidin (7 μg/mL, Wisent). Cells were harvested upon selection for RNA and DNA for RT-PCR and confirmation of DNA cleavage, respectively. For cell proliferation, cells were seeded at equal density per well (on a 96-well plate; Day 1) upon puromycin and blasticidin selection. Growth of the cells were monitored through cell counting using Countess automated cell counter (Invitrogen). Cell numbers were calculated as a percentage compared to negative control. Statistical significance was calculated using Student's t-test.

2.5.12 Luciferase reporter assays

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

2.5.13 Allele-specific ChIP-qPCR

Briefly, pGL3 plasmids containing the WT sequence and the mutant sequence used in the luciferase reporter assay were transfected into 7 million cells (2 μg per allele, per 1 million cells) using Lipofectamine 2000 (ThermoFisher Scientific), per manufacturer's instructions. Next day, each antibody (*FOXA1* 5 μg, Abcam, ab23738; *AR* 5 μg, Abcam, ab1083241; HOXB13 5 μg, Abcam, ab201682; SOX9 5 μg, Abcam, ab3697; GATA2 5 μg, Abcam, ab22849; FOXP1 5 μg, Abcam, ab16645; NKX3.1 10 μl, Cell Signalling Technology, #83700) was conjugated with 10 μL of each Dynabeads A and G (Thermo Fisher Scientific) for each ChIP for 6 h with rotation at 4 °C. When antibody-beads conjugates were ready for use, cells were lifted using trypsin and fixed by resuspending with 300 μL of 1% formaldehyde in phosphate-buffered saline (PBS) for 10 min at room temperature. 2.5 M Glycine was added to quench excess formaldehyde (final concentration 0.125 M).

Cells were then washed with cold PBS and lysed using 300 μL of Modified RIPA buffer (10 mM Tris-HCl, pH 8.0; 1 mM EDTA; 140 mM NaCl; 1% Triton X-100; 0.1% SDS; 0.1% sodium deoxycholate) supplemented with protease inhibitor. The lysate was subject to 25 cycles of sonication (30s ON 30s OFF) using Diagenode Bioruptor Pico (Diagenode). 15 μL of sonicated lysate was set aside as input with the rest used for chromatin pulldown through addition of antibody-beads conjugates and overnight incubation at 4 °C with rotation. Next day, the beads were washed once with Modified RIPA buffer, washed once with Modified RIPA buffer + 500 mM NaCl, once with LiCl buffer (10 mM TrisHCl, pH 8.0; 1 mM EDTA; 250 mM LiCl; 0.5% NP-40; 0.5% sodium deoxycholate) and twice with Tris-ETDA buffer (pH 8). After washes, beads and input were de-crosslinked by addition of 100 μL de-crosslinking buffer and incubation at 65 °C for 6 h. Samples were then purified and eluted. ChIP and input DNA were then used for allele-specific ChIP-qPCR using MAMA primers as described previously. fold change significance was calculated using Student's t-test in R.

All analyses were done using GRCh37 [**internationalhumangenomesequencingconsortiumFinishingEuchro**
reference genome coordinates.

2.6 Code and data availability

Genomic and epigenomic data sets used to support this study can be found from the following accession codes: primary tumors—H3K27ac ChIP-seq (GSE96652), SNVs called from primary tumors (`https://dcc.icgc.org/projects/PRAD-CA`), *FOXA1*, *AR*, and *HOXB13* ChIP-seq in primary prostate tumors is available under the following accession code: GSE137527 and EGAS00001003928, TF ChIP-seq data were from public databases of ReMap and ChIP-Atlas. All code for pre-processing, analyses, and plotting can be found on GitHub (`https://github.com/LupienLab/Zhou_FOXA1_ncomm_2020`).

All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding author upon request.

Chapter 3

Reorganization of the 3D genome pinpoints non-coding drivers of primary prostate tumors

This chapter is a version of the paper published in *Cancer Research* as follows:

hawleyReorganization3DGenome2021

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

3.1 Abstract

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

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

3.2 Introduction

The human genome is organized into hubs of chromatin interactions within the nucleus, setting its three-dimensional topology [**finnMolecularBasisBiological2019**]. Two classes of higher-order topology, topologically associated domains (TADs) and compartments, define clusters of contacts between DNA elements that are linearly distant from each other, such as *cis*-regulatory elements (CREs) and their target gene promoters [**dixonTopologicalDomainsMammalian2012**, **noraSpatialPartitioningRegulatory2012**]. Insulating these hubs to prevent ectopic interactions are TAD boundaries, maintained by CCCTC-binding Factor (CTCF) and the cohesin complex [**pomboThreedimensionalGenomeArchitecture2015**]. Disruption of TAD boundaries through genetic or epigenetic variants can activate oncogenes, as observed in medulloblastoma [**northcottEnhancerHijackingActivates2014**], acute myeloid leukemia [**groschelSingleOncogenicEnhancer20**
gliomas [**flavahanInsulatorDysfunctionOncogene2016**], and salivary gland acinic cell carcinoma [**hallerEnhancerHijackingActivates2019**]. However, recent studies depleting CTCF or the cohesin complex produced little effect on gene expression despite global changes to the three-dimensional chromatin organization [**oudelaarRelationshipGenomeStructure2020**, **despangFunctionalDissec**
williamsonDevelopmentallyRegulatedShh2019]. In contrast, CRE hijacking caused by genetic alterations can result in large changes to gene expression, despite having little impact on the higher-order chromatin organization [**northcottEnhancerHijackingActivates2014**, **zhouEmergenceNoncodingCanc**
These contrasting observations raise questions about the interplay between components of the genetic architecture, namely, how genetic alterations, chromatin states, and the three-dimensional genome cooperate to misregulate genes in disease. Understanding the roles that chromatin organization and *cis*-regulatory interactions play in gene regulation is crucial for understanding how their disruption can promote oncogenesis.

The roles of noncoding mutations targeting CREs in cancer are becoming increasingly clear [**zhouEmergenceNoncodingCancer2016**, **rheinbayAnalysesNoncodingSomatic2020**, **liPatternsSomaticS**

Mutations to the TERT promoter, for example, lead to its over-expression and telomere elongation in multiple cancer types [**vinagreFrequencyTERTPromoter2013**, **huangHighlyRecurrentTERT2013**, **sternMutationTERTPromoter2015**]. Similarly, mutations in the CREs of the *ESR1* and *FOXA1* oncogenes in breast and prostate tumours, respectively, lead to their sustained over-expression [**baileyNoncodingSomaticInherited2016**, **zhouNoncodingMutationsTarget2020**, **paroliaDistinctStructu** which is associated with resistance to hormonal therapies [**jeselsohnESR1MutationsMechanism2015**, **robinsonFoxA1KeyMediator2012**, **fuFOXA1OverexpressionMediates2016**, **fuFOXA1UpregulationPro** Point mutations have the potential to alter three-dimensional chromatin organization, albeit indirectly, by modifying transcription factor (TF) or CTCF binding sites [**mauranoLargescaleIdentificationSequence** **guoMutationHotspotsCTCF2018**]. structural variants (SVs), on the other hand, are large rearrangements of chromatin that can directly impact its structure [**dixonIntegrativeDetectionAnalysis2018**, **akdemirDisruptionChromatinFolding2020**]. This can establish novel CRE interactions from separate TADs or chromosomes, as has been observed in leukemia [**hniszActivationProtooncogenesDisruption20** and multiple developmental diseases [**lupianezDisruptionsTopologicalChromatin2015**, **allouNoncodingDelet** But how prevalent and to what extent these rearrangements affect the surrounding chromatin remains largely unstudied in primary tumours [**akdemirDisruptionChromatinFolding2020**, **liPatternsSomaticSt** **iyyankiSubtypeassociatedEpigenomicLandscape2021**]. Hence, to understand gene misregulation in cancer, it is critical to understand how SVs impact three-dimensional chromatin organization and CRE interactions in primary tumours.

SVs play an important role in prostate cancer (PCa), both for oncogenesis and progression. An estimated 97% of primary tumours contain SVs [**liPatternsSomaticStructural2020**, **fraserGenomicHallmarksLo** and translocations and duplications of CREs for oncogenes such as *AR* [**takedaSomaticallyAcquiredEnhancer201** *ERG* [**rosenClinicalPotentialERG2012**], *FOXA1* [**quigleyGenomicHallmarksStructural2018**, **paroliaDistinctStructuralClasses2019**] and *MYC* [**paroliaDistinctStructuralClasses2019**] are highly recurrent. While coding mutations of *FOXA1* are found in $\sim 10\%$ of metastatic castration-resistant prostate cancer (mCRPC) patients, SVs that target *FOXA1* CREs are found in over 25% of metastatic prostate tumours [**paroliaDistinctStructuralClasses2019**]. In addition to oncogenic activation, SVs in prostate tumours disrupt and inactivate key tumour suppressor genes including *PTEN*, *BRCA2*, *CDK12*, and *TP53* [**quigleyGenomicHallmarksStructural2018**, **abeshouseMolecularTaxono** Furthermore, over 90% of prostate tumours contain complex SVs, including chromothripsis and chromoplexy events [**bacaPunctuatedEvolutionProstate2013**], making it a prime model to study the effects of SVs. However, despite large-scale tumour sequencing efforts, investigating the impact of SVs on three-dimensional prostate genome remains difficult, owing to constraints from chromatin

conformation capture-based assays. In this work, we build on recent technological advances in Hi-C protocols to investigate the three-dimensional chromatin organization of the prostate from primary benign and tumour tissues. Using patient-matched whole genome sequencing (WGS), RNA-seq, and chromatin immunoprecipitation sequencing (ChIP-seq) data, we show that SVs in PCa repeatedly hijack CREs to disrupt the expression of multiple genes with minimal impact to higher-order three-dimensional chromatin organization.

3.3 Results

3.3.1 Three-dimensional genome organization is stable over oncogenesis

Chromatin conformation capture (3C) technologies enable the measurement of three-dimensional genome organization. These assays, however, are often limited to cell lines, animal models and liquid tumours due to the amount of input required [**liebermanaidenComprehensiveMappingLongRange2009**]. Here, we optimized and conducted low-input Hi-C [**diazChromatinConformationAnalysis2018**] on 10 μm thick cryosections from 12 primary prostate tumours and 5 primary benign prostate sections (see Section 3.4, Figure 3.1a, Figure B.1a). The 12 tumours were selected from the Canadian Prostate Cancer Genome Network (CPC-GENE) cohort previously assessed for WGS [**fraserGenomicHallmarksLocalized2017**], RNA-seq [**chenWidespreadFunctionalRNA2019**], and H3K27ac ChIP-seq [**kronTMPRSS2ERGFusion2017**, **mazrooeiCistromePartitioningReveals2019**] (Supplementary Table 1). All 12 of these PCa patients previously underwent radical prostatectomies and 6 of our 12 samples (50%) harbour the *TMPRSS2-ERG* (T2E) fusion found in approximately half of the primary PCa patients [**fraserGenomicHallmarksLocalized2017**]. The total percent of genome altered ranges from 0.99%−18.78% (Supplementary Table 1) [**fraserGenomicHallmarksLocalized2017**]. The 12 tumour samples were histopathologically assessed to have $\geq$ 70% prostate cellularity and $\geq$ 60% for our group of 5 normal prostate samples. Upon Hi-C sequencing, we reached an average of 9.90×10^8 read pairs per sample (range $5.84 \times 10^8 - 1.49 \times 10^9$ read pairs) with minimal sequencing duplication rates (range $10.6\% - 20.8\%$) (Supplementary Table 2). Pre-processing resulted in an average of 6.23×10^8 (96.13%) valid read pairs per sample (range $3.95 \times 10^8 - 9.01 \times 10^8$, or $82.42\% - 99.22\%$; Supplementary Table 2). Hence, we produced a high depth, high quality Hi-C library on 17 primary prostate tissue slices.

To characterize the higher-order organization of the primary prostate genome, we first identified TADs. Across the 17 primary tissue samples, we observed an average of 2,305 TADs with a me-

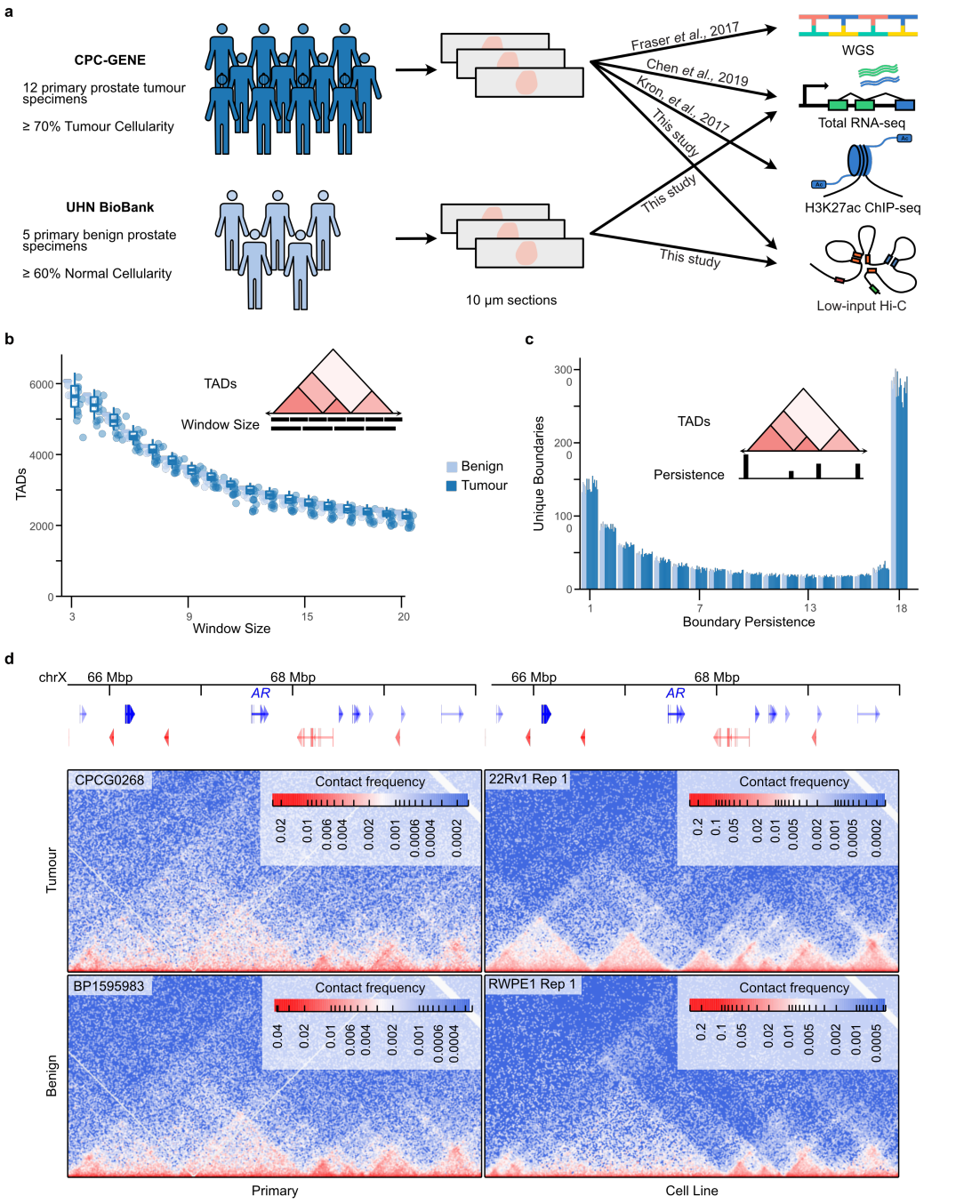

Figure 3.1: **Topologically associated domains are stable over prostate oncogenesis. a.** The sample collection and data usage of primary prostate samples in this study. 10 μm sections from 6 tumours previously identified as T2E+ and 6 T2E− were used for Hi-C sequencing. 5 additional 10 μm sections were collected from benign prostate specimens in the UHN BioBank. **b-c.** A comparison of the number of TADs detected at multiple window sizes (**b**) and boundary persistence (**c**) in each patient sample, with inset schematics. Boxplots show the first, second, and third quartiles. Whiskers extend to the median ±1.5× the IQR. **d.** ICE-normalized contact matrices around the AR gene in primary samples and cell lines. Hi-C data for 22Rv1 and RWPE1 cell lines obtained from [**rhieHighresolution3DEpigenomic2019**]. Coordinates are listed according to the GRCh38 reference genome.

dian size of 560 kbp (TopDom [**shinTopDomEfficientDeterministic2016**], window size $w = 20$, $FDR < 0.05$; Supplementary Tables 3-4). However, when considering all hierarchical levels of TAD organization, we did not observe significant differences in the number of TADs identified across length scales (Figure 3.1b), nor in the persistence of their boundaries, the number of window sizes the boundary is identified within (Figure 3.1c). This suggests few, if any, differences in three-dimensional genome organization at the TAD level between benign and tumour tissue. Notably, we observed differences in organization around essential genes for PCa between primary tissue and previously profiled cell lines. For example, chromatin around the androgen receptor (AR) gene that was previously found enriched in the 22Rv1 compared to RWPE1 prostate cell lines [**rhieHighresolution3DEpigenomic2019**] were not recapitulated in either benign or tumour primary samples (Figure 3.1d). Moreover, when compared to other Hi-C datasets, the primary prostate samples clustered separately from cell lines (Figure B.1b), despite similar enrichment of CTCF binding sites near TAD boundaries (Figure B.1c). These results suggest that TADs are constrained over oncogenesis and that cell line models may not harbour disease-relevant three-dimensional genome organization.

We next investigated compartmentalization changes, the second class of higher-order three-dimensional genome organization. Recurrent changes to segments nearly the size of chromosome arms showed differential compartmentalization in multiple tumour samples compared to benign samples, such as compartment B-to-A transitions on 19q and A-to-B transitions on chromosome Y (Figure B.2a-c). Only two transcripts on chromosome 19 were differentially expressed between the 8 tumours with benign-like compartmentalization and the other 4 (Figure B.2d). Similarly, no genes on chromosome Y were differentially expressed between the 4 tumours with benign-like compartmentalization and the remaining samples (Figure B.2e). Both arms on chromosome 3 show differential mean compartmentalization, but this appears to be driven by one tumour sample (CPCG0255) and one benign sample (BP1664855) for each arm and is not recurrent (Figure B.2f). Collectively, these results suggest that phenotypic differences between benign and tumour tissues do not stem from differences in higher-order three-dimensional genome organization alone.

3.3.2 Focal chromatin interactions shift over oncogenesis

Changes to focal chromatin interactions have been observed in the absence of higher-order chromatin changes [**takayamaTransitionQuiescentActivated2021**, **johnstoneLargeScaleTopologicalChanges20**
and we hypothesized that this may be the case in PCa. We detected chromatin interactions, identi-

fying a median of 4,395 interactions per sample (range $1,286 - 6,993$; Figure B.3a, Supplementary Table 5). Among these detected interactions, we identified known contacts in PCa such as those between two distal CREs on chromosome 14 and the *FOXA1* promoter [**zhouNoncodingMutationsTarget2020**] (Figure B.3b), and CREs upstream of *MYC* on chromosome 8 that are frequently duplicated in metastatic disease [**quigleyGenomicHallmarksStructural2018**] (Supplementary Table 5). 16,474 unique chromatin interactions were identified in at least one sample (Figure 3.2a), reaching an estimated $\sim 80\%$ saturation of detection (Figure B.3c) in the entire cohort and a median of 57 % saturation per sample (Figure B.3c-d). This suggests that identification of focal interactions may be boosted when assessed in a cohort of samples, rather than single samples. Restricting our analysis to the 8,486 interactions present in at least two samples (51.5% of all interactions) yielded 1,405 tumour- and 273 benign-specific interactions, suggesting focal changes in three-dimensional genome organization occur over oncogenesis. Aggregate peak analysis revealed Hi-C contact enrichment at all detected interactions in all samples (Figure 3.2b-c), demonstrating that tumors- and benign-specific interactions are not binary. Rather, the contacts at "tumour-specific" loci are more enriched than those at "benign-specific" loci in tumour samples (Figure 3.2b). Similarly, the contacts at "benign-specific" loci are more enriched than those at "tumour-specific" loci in benign samples (Figure 3.2c). Together, these results suggest that more focal changes to chromatin interactions are present in prostate oncogenesis despite the stable higher-order organization.

3.3.3 Cataloguing structural variants from Hi-C data

In prostate tumours, SVs populate the genome to aid disease onset and progression [**fraserGenomicHallmarksLo**
quigleyGenomicHallmarksStructural2018]. Advances in computational methods now enable the identification of SVs from Hi-C datasets [**dixonIntegrativeDetectionAnalysis2018**, **hoStructuralVariationS**
Applying Hi-C Breakfinder, an SV caller that uses Hi-C data [**dixonIntegrativeDetectionAnalysis2018**], to our primary prostate tumour Hi-C dataset, yielded a total of 317 unique breakpoints with a median of 15 unique breakpoints per tumour (range $3-95$; Figure 3.3a; Supplementary Table 6). As an example, we found evidence of the T2E fusion spanning the 21q22.2-3 locus in 6/12 (50%) patients (CPCG0258, CPCG0324, CPCG0331, CPCG0336, CPCG0342, and CPCG0366) (Figure 3.3b), in accordance with previous WGS findings [**fraserGenomicHallmarksLocalized2017**]. Combining unique breakpoint pairs into rearrangement events yielded 7.5 total events on average per patient (range $1 - 36$, Figure B.4a-b). We also identified more inter-chromosomal breakpoint pairs with the Hi-C data in 11 of 12 tumours (Figure 3.3b), including a novel translocation event that

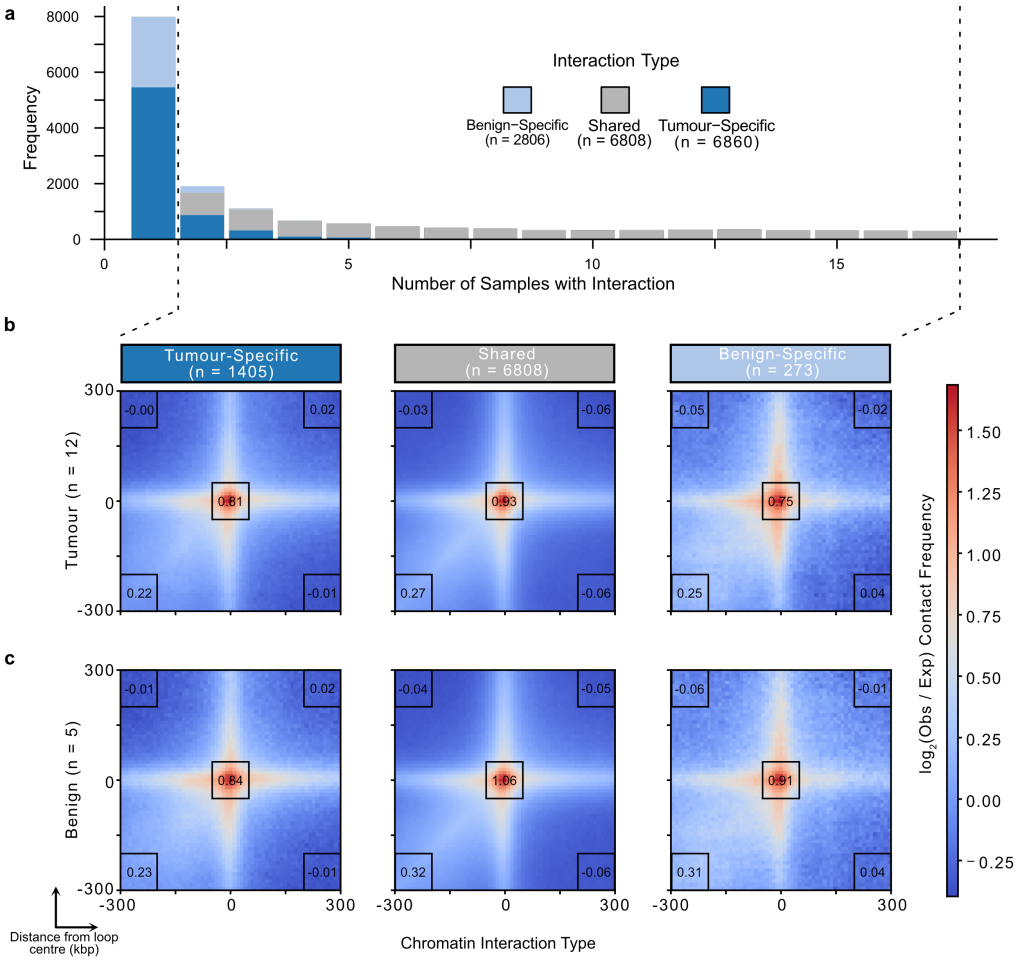

Figure 3.2: **Focal chromatin interactions display subtle differences between benign and tumour tissue. a.** Stacked bar plots of the number of samples that chromatin interactions were identified in. **b-c.** APA of tumour (**b**) or benign (**c**) contacts in tumour-specific, benign-specific, and shared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

encompasses the deleted region between *TMPRSS2* and *ERG* into chromosome 14 (Figure 3.4c). Few loci contained SV breakpoints recurrent between patients (Figure B.4c) and fewer inversions were detected by Hi-C (Figure 3.4a). These numbers are smaller than previously reported from matched WGS data [**fraserGenomicHallmarksLocalized2017**]; however, the median distance between breakpoints on the same chromosome was much larger at 31.6 Mbp for Hi-C-identified breakpoints, compared to 1.47 Mbp from WGS-identified breakpoints (Figure 3.3c). This is consistent with the inherent nature and resolution of the Hi-C method to detect larger, inter-chromosomal events [**dixonIntegrativeDetectionAnalysis2018**]. No SVs were detected in the 5 primary benign prostate tissue samples from Hi-C data. While this does not rule out the presence of small rearrangements undetectable by Hi-C limited by its resolution, the absence of large and inter-chromosomal SVs further supports a difference in genome stability between benign and tumour tissues [**fraserGenomicHallmarksLocalized2017**, **bergerGenomicComplexityPrimary2011**, **bacaPunctuated**
mazrooeiCistromePartitioningReveals2019]. Collectively, Hi-C defines a valid method to interrogate for the presence of SV in tumour samples, compatible with the detection of intra- and inter-chromosomal interactions otherwise missed in WGS analyses.

Among SVs detected in primary prostate tumours, we identified both simple and complex chains of breakpoints. While simple SVs correspond to fusion between two distal DNA sequences, complex chains are evidence of chromothripsis and chromoplexy [**bacaPunctuatedEvolutionProstate2013**]. These genomic aberrations affecting multiple regions of the genome are known to occur in both primary and metastatic PCa [**bacaPunctuatedEvolutionProstate2013**, **fraserGenomicHallmarksLocalized201**
liPatternsSomaticStructural2020]. The chains can be pictured as paths connecting breakpoints in the contact matrix (Figure B.4d). 8 of the 12 (66.7%) tumour samples contained these chains, including one patient (CPCG0331) harbouring 11 complex events and three patients (CPCG0246, CPCG0345, and CPCG0365) each harbouring a single complex event. We observed a median of 1 complex event per patient (range $0 - 11$) consisting of a median of 3 breakpoints (range $3 - 7$) spanning a median of 2 chromosomes per event (range $1 - 4$, Supplementary Table 7, Figure B.4b). Patient CPCG0331 had 11 complex events, including a 6-breakpoint event spanning 3 chromosomes (Figure B.4b). A highly rearranged chromosome 3 was also found in the same patient (Figure 3.3d). The most common type of complex event involved 3 breakpoints and spanned 2 chromosomes, occurring 9 times across 5 of the 8 patients with complex events. In summary, using Hi-C, we detected both simple and complex SVs in primary prostate tumours not previously identified using WGS-based methods. We were able to identify known observations, such as a highly mutated region on chromosome 3 and subtype-specific differences in abundance, as well as find novel inter-chromosomal

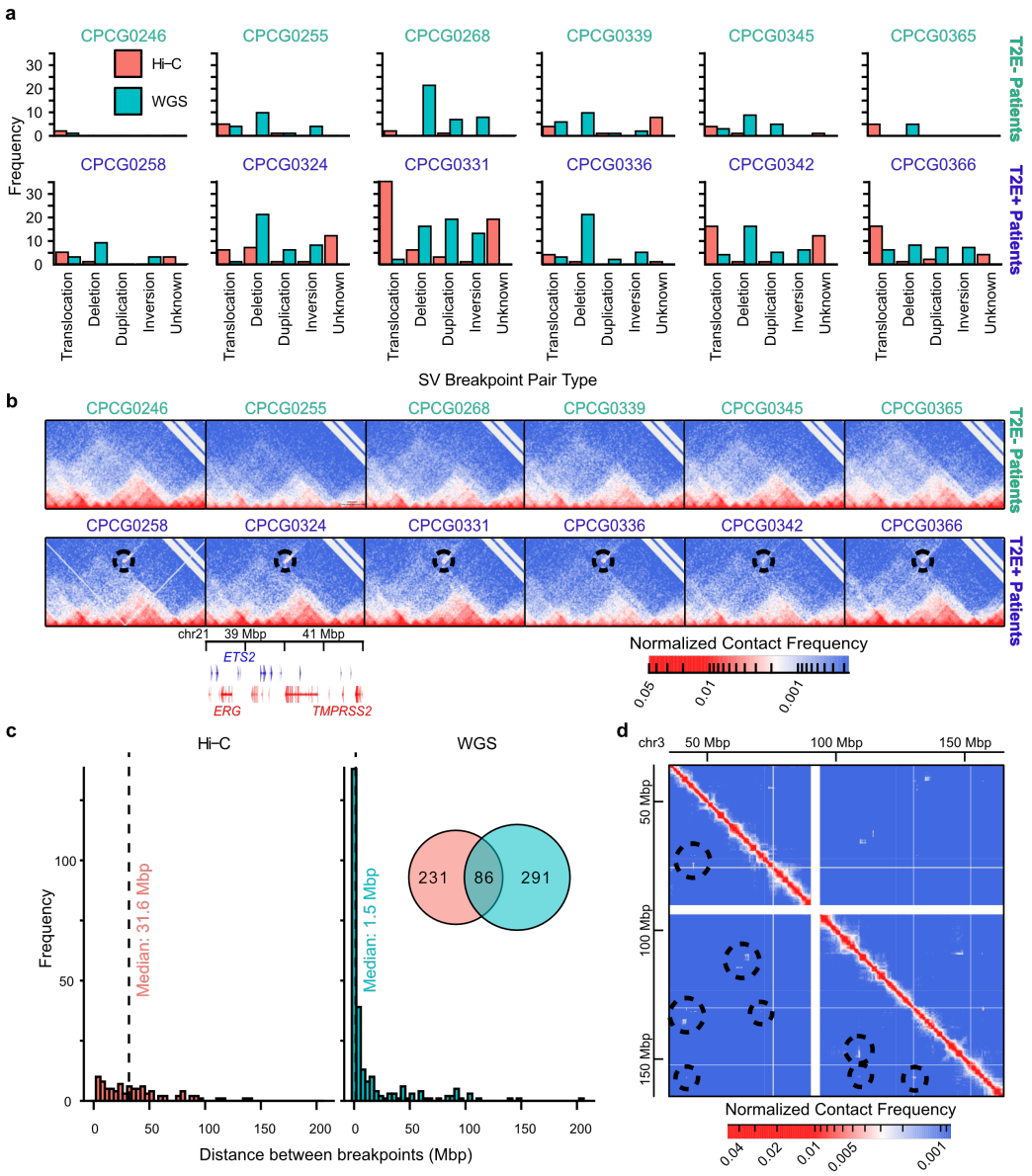

Figure 3.3: **SVs are identified in primary tissue through chromatin conformation capture.** **a.** Bar plot of SV breakpoint pairs identified by Hi-C and WGS on matched samples. **b.** Contact matrices of the chr21:37-42 Mbp locus harbouring the *TMPRSS2* and *ERG* genes. Circles indicate increased contact between *TMPRSS2* and *ERG* in the T2E+ tumours. **c.** Histogram showing the distance between breakpoints on the same chromosome detected by Hi-C (left) versus WGS (right). The inset Venn diagram shows the number of unique or shared breakpoint calls from both methods. **d.** An example of a complex set of rearrangements spanning both arms of chromosome 3 in a single patient. Contact matrices are ICE-normalized. Coordinates are listed according to the GRCh38 reference genome.

events not previously reported.

3.3.4 SVs alter gene expression independently of intra-TAD contacts

Using combined WGS called SVs with those from Hi-C data, we next systematically examined the impact of SVs on TAD structure. This led us to look at the intra-TAD and inter-TAD interactions around each breakpoint. We observed that only 18 of the 260 (6.9%) TADs containing SV breakpoints were associated with decreased intra-TAD or increased inter-TAD interactions (Figure 3.4a). 12 of 18 (66.7%) occurrences were within T2E+ tumours. We found no evidence that simple versus complex SVs were a factor in determining whether a TAD was altered (Pearson's χ^2 test, $\chi^2 = 0.0166$, $p = 0.897$, $df = 1$). Similarly, the type of SV (deletion, inversion, duplication, or translocation) was not predictive of whether the TAD would be altered (Pearson's χ^2 test, $\chi^2 = 4.7756$, $p = 0.3111$, $df = 4$). Overall, we find that SVs are associated with higher-order topological changes in a small percentage of cases and that the presence of an SV breakpoint is not predictive alone of an altered TAD.

Despite the evidence that SVs rarely impact higher-order chromatin topology, we evaluated whether SVs affected the expression of genes within the TADs surrounding the breakpoint using patient-matched RNA-seq data [**chenWidespreadFunctionalRNA2019**]. We found that 23 of 260 breakpoints (8.8%) are associated with significant changes to local gene expression (Figure 3.4b). Notably, only one of these breakpoints with local gene expression appeared within an altered TAD (Figure 3.4b). Complex events can have opposite effects at each breakpoint. For example, while the T2E fusion across all tumours leads to over-expression of *ERG* and under-expression of *TMPRSS2* [**fraserGenomicHallmarksLocalized2017**, **kronTMPRSS2ERGFusion2017**], the deleted locus between these two genes was inserted into chromosome 14 as part of a complex translocation event in one patient (Figure 3.4c-f). This inserted fragment positions *ERG* towards the $5'$ end of the *RALGAPA1* gene and *TMPRSS2* towards the $3'$ end (Figure 3.4c) resulting in a significant drop in intra-TAD contacts at the *RALGAPA1* locus on chromosome 14 (two-sample unpaired t-test, $t = 6.38$, $p = 1.04 \times 10^{-9}$; Figure 3.4d). Despite the significant topological change on chromosome 14, no significant changes to expression was detectable across genes within the same TAD on chromosome 14 (Figure 3.4e). Conversely, TAD alterations are not required changes to gene expression. As part of a complex SV involving the *RIMBP2* gene (Figure 3.4g-j), both ends of the gene contain breakpoints (Figure 3.4g). This rearrangement is not associated with changes to intra-TAD contacts (two-sample unpaired t-test, $t = 0.8101$, $p = 0.4183$; Figure 3.4h). However, RIMBP2 is

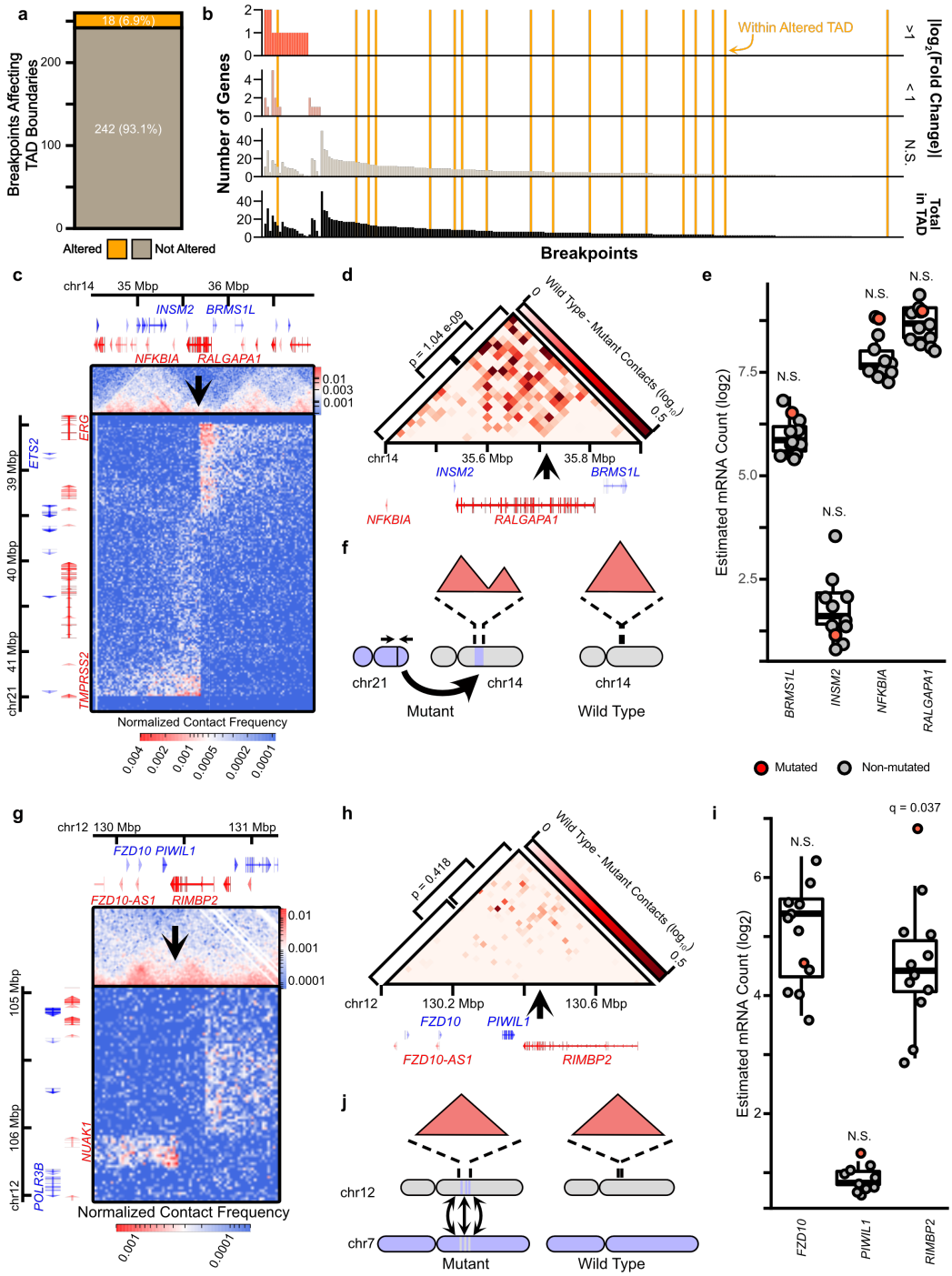

Figure 3.4: **SVs can alter TADs or gene expression around breakpoints, but rarely alters both.** (Continued on the following page)

Figure 3.4: **a.** A count of the number of SV breakpoints associated with altered TAD boundaries. **b.** Bar plot showing the number of genes differentially expressed around SV breakpoints. **c-f.** An example of an SV that alters intra-TAD contacts without significantly affecting the expression of the nearby genes. **c.** Contact matrix showing a translocation of the *TMPRSS2-ERG* locus into chromosome 14 in the *RALGAPA1* gene. **d.** Differential contact matrix between the tumour containing this translocation and another tumour without it (two-sample unpaired t-test). **e.** Gene expression scatterplot and boxplot of genes within the affected TAD for each sample. Boxplots show the first, second, and third quartiles of WT tumours. Whiskers extend to the median $\pm 1.5\times$ the IQR. Red dots represent the tumour with the example SV, grey dots represent the tumours without. **f.** A schematic representation of the translocation and its affect on TADs. **g-j.** An example of an SV that does not alter intra-TAD contacts but does alter the expression of a nearby gene. **g.** Contact matrix showing a complex rearrangement around the *RIMBP2* gene. **h.** Differential contact matrix between the tumour containing this rearrangement and another tumour without it (two-sample unpaired t-test). **e.** Gene expression scatterplot and boxplot of genes within the affected TAD for each sample. **j.** A schematic representation of the rearrangement. Coordinates are listed according to the GRCh38 reference genome.

over-expressed in this patient (Figure 3.4i). More generally, only a single breakpoint was observed with both TAD contact and gene expression changes, although we did not find evidence to suggest these are dependent events (Pearson's χ^2 test, $\chi^2 = 6.31 \times 10^3$, $p = 0.9367$, $df = 1$). For TADs where at least one gene was differentially expressed, 19 (83%) of them had at least one gene with doubled or halved expression. Notably, we found that inter-chromosomal translocations are associated with altering the expression of genes nearby their breakpoints compared to intra-chromosomal breakpoints (Pearson's χ^2 test, $\chi^2 = 7.01$, $p = 8.11 \times 10^{-3}$, $df = 1$; Figure B.5). Taken together, these results suggest that while SVs can alter contacts within TADs, this is neither necessary nor sufficient to alter gene expression.

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

Mutations in PCa have previously been found to converge on active CREs [**mazrooeiCistromePartitioningRev**

To assess if SVs function in a similar fashion, we investigated the convergence of SV breakpoints in active CREs. We find that SV breakpoints are enriched in the catalogue of CREs captured by H3K27ac ChIP-seq from our 12 primary prostate tumours compared to the rest of the genome (one-sided permutation z-test, $z = 25.591$, $p = 0.0099$, $n = 100$; Figure 3.5a-b). This is similar to the enrichment of point mutations in CREs active in PCa [**mazrooeiCistromePartitioningReveals2019**], suggesting that SVs which alter gene expression may do so by recurrently targeting CREs. Since individual CREs can regulate multiple genes [**gasperiniGenomewideFrameworkMapping2019**], we suspected that SVs that do alter gene expression may predominantly affect multiple genes at the same

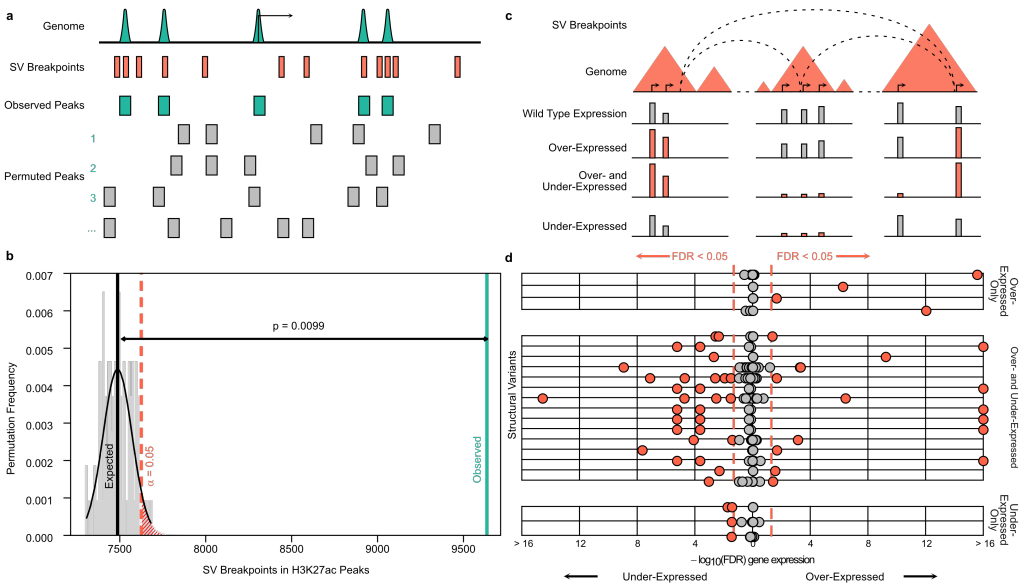

Figure 3.5: **SV breakpoints are enriched in active CREs and repeatedly alter the expression of multiple genes. a.** Schematic of permutation testing for the overlap between SV breakpoints in all CPC-GENE prostate tumours and the catalogue of active CREs in the 12 tumour samples in this study. Peaks above the genome represent putative active CREs, identified by H3K27ac ChIP-seq data. **b.** Histogram of permutation test results in grey. The vertical black and green bars refer to the expected and observed overlap of SV breakpoints and CREs, respectively. P-value is obtained from the permutation test, $n = 100$. **c.** Schematic of how the expression of genes within TADs containing SV breakpoints are compared between mutant and WT tumours. **d.** Scatterplot of FDR values obtained from differential gene expression analysis as outlined in **c.** Red dots are differentially expressed genes ($FDR < 0.05$), grey dots are genes not differentially expressed between the mutant and WT tumours.

time, instead of single genes. In agreement, when considering all SVs associated with altered gene expression near a breakpoint we find 16 of the 22 (72.7%) SVs are associated with altered expression of multiple genes (Figure 3.5c-d). Notably, 15 of these 16 SVs (93.8%) are associated with both over- and under-expression of genes, instead of genes all being either over-expressed or under-expressed (Figure 3.5d). 12 of these 15 (80%) SVs are associated with expression changes at SV antipodes, opposite ends of a breakpoint pair (Figure B.6). The recurrent targeting of active CREs, combined with the opposite gene expression changes at SV antipodes, suggests that SVs may repeatedly alter expression by CRE hijacking.

The fusion of *PMEPA1* and *ZNF516* is an example of CRE hijacking resulting in opposite differential gene expression. Specifically, the fusion results in the *PMEPA1* promoter being hijacked to the 5′ end of the *ZNF516* gene. This is concomitant with the over-expression of *ZNF516* and under-expression of *PMEPA1* (Figure 3.6a-c). In addition to hijacking the *PMEPA1* promoter to the *ZNF516* gene, this fusion also coincides with gains in H3K27ac over the *ZNF516* gene body and

of H3K27ac histone hypo-acetylation over the 3′ end of *PMEPA1*'s gene body. This mirrors the creation of a cluster of regulatory elements (CORE) reported for the T2E fusion, reflective of new CREs enabling *ERG* over-expression and the concomitant under-expression of *TMPRSS2* (Figure B.6) [**kronTMPRSS2ERGFusion2017**, **tomlinsRecurrentFusionTMPRSS22005**, **tomlinsDistinctClassesChro**
Putative CRE hijacking is also observed with inter-chromosomal rearrangements such as seen at the SV connecting chromosomes 7 and 19, creating 2 fusion products (termed C2B and B2C; Figure 3.6d). This SV separates the 3′ end of *BRAF* from its promoter and upstream enhancers on chromosome 7 (C2B; Figure 3.6d), fusing it to the 3′ end of *CYP4F11* (Figure 3.6e). Focal chromatin interactions between *BRAF* and multiple active CREs are only observed in the fusion on chromosome 19 (Figure 3.6e). Using matched RNA-seq data, we observe an estimated 5 fold increase in expression for the 3′ exons of *BRAF* in the mutated tumour compared to others (fold change = 4.976, FDR = 0.0181; Figure 3.6f). Collectively, over-expression of the oncogenes, such as *ERG* and *BRAF*, and suppression of the tumour suppressor *PMEPA1* demonstrates recurrence of CRE hijacking mediated by SVs and the disease-relevance of this recurrent mechanism in primary prostate tumours. These SVs result in changes to focal chromatin interactions and overshadow the effect on higher-order topology in primary prostate cancer.

3.3.6 Discussion

Genetic alterations that subvert the higher-order chromatin organization to allow for aberrant focal interactions may be more common in cancer than previously recognized. In this work we demonstrated that CRE hijacking by SVs is often associated with opposing gene expression changes at SV antipodes, whereby genes on one flank of the breakpoint are up-regulated while genes on the other flank are repressed. Complex SVs, such as chromoplexy and chromothripsis, are found in numerous cancer types [**bacaPunctuatedEvolutionProstate2013**, **liPatternsSomaticStructural2020**], providing many opportunities for widespread effects on gene expression and CRE hijacking. This is in addition to many known cancer drivers that alter CRE interactions, including the *AR* and FOXA1 enhancer amplifications in primary and metastatic prostate tumours [**paroliaDistinctStructuralClasses2019**, **quigleyGenomicHallmarksStructural2018**, **takedaSomaticallyAcquiredEnhancer2018**, **zhouNoncodingM**
kronTMPRSS2ERGFusion2017, **viswanathanStructuralAlterationsDriving2018**]. More recent findings also fit this model, such as accumulation of extra-chromosomal circular DNA activating oncogenes that would otherwise be constrained by chromatin topology [**wuCircularEcDNAPromotes2019**, **kumarATACseqIdentifiesThousands2020**, **mortonFunctionalEnhancersShape2019**, **shoshaniChromothr**

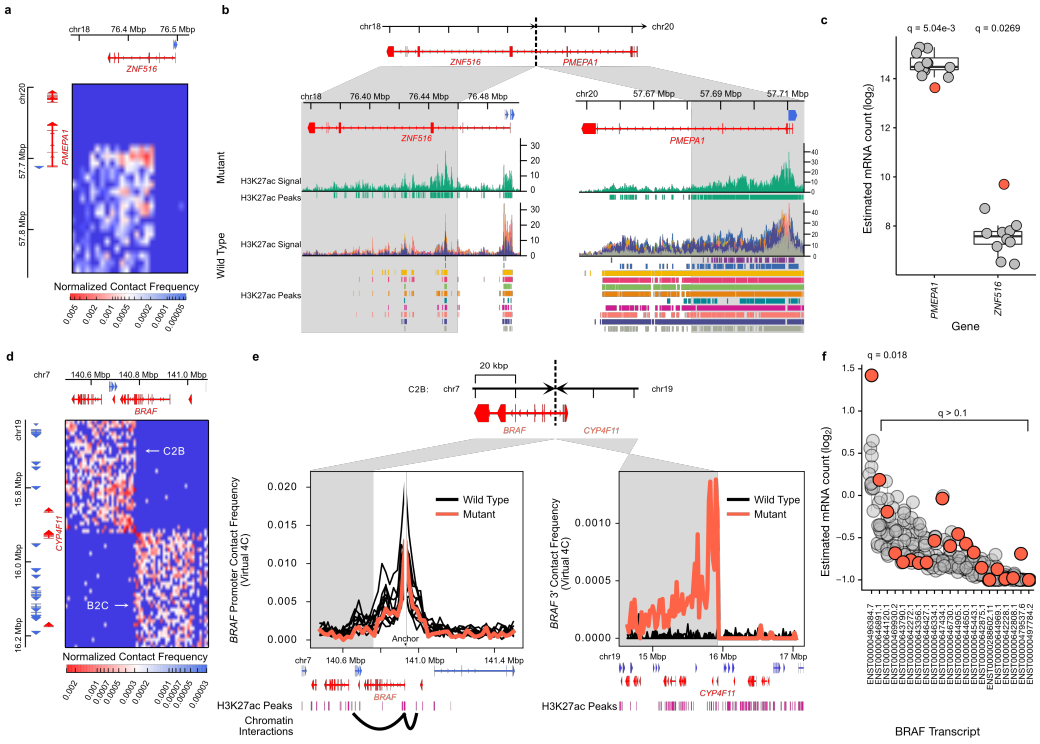

Figure 3.6: **SVs altering gene expression by rewiring focal chromatin interactions. a.** Contact matrix of the deletion between *PMEPA1* and *ZNF516*. **b.** Genome tracks of H3K27ac ChIP-seq signal around the *ZNF516* and *PMEPA1* genes with the rearrangement. Grey regions are loci brought into contact by the SV. **c.** Gene expression of *PMEPA1* and *ZNF516* in all tumour samples. Boxplots represent the first, second, and third quartiles of WT patients (grey dots). Red dots are the gene expression for the mutated patient. **d.** The contact matrix of an inter-chromosomal break between chromosome 7 and chromosome 19. Upper left contacts correspond to the C2B fusion, and lower-left contacts correspond to the B2C fusion. **e.** Contact frequencies of the *BRAF* promoter on chromosome 7 (left) and the 3′ end of *BRAF* on chromosome 19 (right). SV-associated contacts between the 3′ end of *BRAF* on chromosome 19 (right) are focally enriched at H3K27ac peaks downstream of *CYF4P11*. Bar plot of SVs categorized by how differentially expressed genes are altered. Contact matrices are ICE-normalized. Coordinates are listed according to the GRCh38 reference genome.

These insights stress the importance of investigating all ends of an SV to assess the biological impact of these mutations on the *cis*-regulatory landscape as a whole, as opposed to focusing on CREs or SV breakpoints as single entities.

Changes to the three-dimensional genome reported in disease onset or development are often inferred from alterations in TAD boundaries [**oudelaarRelationshipGenomeStructure2020**, **akdemirDisruptionChromatinFolding2020**]. For instance, CTCF activity is targeted by somatic mutations that enrich at its binding sites in colorectal, esophageal, and liver cancers [**katainenCTCFCohesinb** **guoMutationHotspotsCTCF2018**]. Furthermore, gains in DNA methylation (DNAme) at CTCF binding sites are linked to altered TAD structures in gliomas [**flavahanInsulatorDysfunctionOncogene2016**]. In primary PCa 97% of differentially methylated regions genome-wide in primary PCa are losses of DNAme [**zhaoDNAMethylationLandscape2020**, **yuWholeGenomeMethylationSequencing2013**], an epigenetic process previously shown to have limited impact on CTCF chromatin binding [**mauranoRoleDNAMet** This suggests that aberrant CTCF binding at TAD boundaries is not a hallmark of prostate oncogenesis. Our observation of stable chromatin organization supports this model. Notably, stable TAD structures observed in these primary tissues contrast previous reports of chromatin organization in cell lines derived from prostate cells [**taberlayThreedimensionalDisorganizationCancer2016**, **rhieHighresolution3DEpigenomic2019**], highlighting the necessity of low-input protocols and primary tissues [**diazChromatinConformationAnalysis2018**]. While some differences in large-scale organization between primary tissues and cell lines may arise from differences in sequencing depths, recent findings demonstrate differential focal interactions between 22Rv1 and LNCaP cells are associated with differential expression of the *MYC* oncogene [**ahmedCRISPRiScreensReveal2021**]. Thus, the differences we observe may indicate differential regulation of the *AR* gene between primary tissues and prostate cell lines. Our findings further support recent reports of shared higher-order chromatin organization among phenotypically distinct cell types in model organisms [**rao3DMapHuman2014**, **dixonTopologicalDomainsMammalian2012**, **ing-simmonsIndependenceChroma** **ghavi-helmHighlyRearrangedChromosomes2019**, **iyyankiSubtypeassociatedEpigenomicLandscape2021** **akdemirDisruptionChromatinFolding2020**]. Taken together, this body of evidence suggests that large disruptions to TADs and compartments may constrain the transformation of normal to cancer cells or the divergent subtyping within prostate tumours. Instead, changes to focal chromatin interactions seem to reflect alterations in the genetic architecture leading to cancer development. Investigating these focal chromatin interactions may provide insights on the relationship between CREs, such as between enhancers and their target gene promoter [**gasperiniComprehensiveCatalogueValidated2** **nasserGenomewideEnhancerMaps2021**] to better understand the etiology of disease.

In conclusion, by bypassing technical limitations to characterize the three-dimensional genome organization across benign and tumour prostate tissue, our work reveals the predominant stable nature of genome topology across prostate oncogenesis. Instead, alterations to discrete chromatin interactions populate the PCa genome. These impact the function of CREs, such as we report for SV-mediated CRE hijacking events. Considering the contribution of SVs across human cancers [**hanahanHallmarksCancerNext2011**], our collective work presents a framework inclusive of genetics, chromatin state, and three-dimensional genome organization to understand the genetic architecture across individual primary tumours.

3.4 Methods

3.4.1 Patient selection criteria

Patients were selected from the CPC-GENE cohort of ≥ 200 Canadian men with indolent PCa, Gleason scores of 3+3, 3+4, and 4+3 [**fraserGenomicHallmarksLocalized2017**]. All primary human material was obtained with written informed consent with approval of our institutional research ethics board (REB) (University Health Network (UHN) 11-0024). The intersection of previously published data for WGS [**fraserGenomicHallmarksLocalized2017**], RNA abundance [**chenWidespreadFunctionalRNA2019**], and H3K27ac ChIP-seq [**kronTMPRSS2ERGFusion2017**] led to 25 samples having data for all assays. 11 of these were positive for the T2E fusion and 14 were not. To accurately represent the presence of this subtype of PCa in the disease generally, and to ensure minimum read depths required to perform accurate analysis on chromatin conformation data, we selected approximately half of these remaining samples (6 T2E+ and 6 T2E−).

3.4.2 Patient tumour in situ low-input Hi-C sequencing

We followed the general in situ low input Hi-C protocol from [**diazChromatinConformationAnalysis2018**] with our own re-optimization for solid tumour tissue sections. It is worth noting that throughout the protocol, the pellet would be hardly visible and would require careful pipetting. The specific modifications of the protocol are described below.

Tumour tissue preparation

Twelve cryopreserved-frozen PCa tumour tissue specimens were obtained from primary PCa patients as part of the CPC-GENE effort [**fraserGenomicHallmarksLocalized2017**]. Written

informed consent was obtained from all patients with REB approval (UHN 11-0024). These tumour specimens were sectioned into 10 μm sections. Sections before and after the sections used for Hi-C were stained with hematoxylin and eosin (H&E) and assessed pathologically for $\geq$ 70% PCa cellularity. The percentage of infiltrating lymphocytes was also estimated by pathological assessment to be $\leq$ 3%. Stratification into T2E+ or T2E− was determined through either WGS detection of the fusion, immunohistochemistry, or messenger RNA (mRNA) expression microarray data [**fraserGenomicHallmarksLocalized2017**].

Normal tissue preparation

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

Fixation and lysis

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

Enzyme digestion and overhang fill-in

The samples stored at −80 °C were thawed on ice and spun down at 300 × g for 5 min at 4 °C. The samples were then re-suspended in 125 μL of ice-cold 10X NEB2 Buffer (New England Biolabs), and again spun down at 13000 × g for 5 min at 4 °C. The pellet was then re-suspended in

25 μL of 0.4% sodium dodecyl sulfate (SDS) and incubated at 65 °C for 10 min without agitation for permeabilization. To quench the SDS, 10% Triton X-100 in water (12.5 μL + 75 μL water) was then added to the samples and incubated at 37 °C for 45 min at 650 rpm. For enzymatic digestion, 35 μL of 10X NEB2.1 buffer (New England Biolabs) was added to each sample, followed by the addition of 50 U of MboI and 90 min incubation at 37 °C with gentle agitation (add 30 U first, incubate 45 min, followed by the addition of another 20 U and another 45 min of incubation). Upon digestion, the MboI enzyme was inactivated by incubating at 62 °C for 20 min. The overhangs generated by the MboI enzyme was then filled-in by adding a mix of dNTPs and DNA Polymerase I Klenow Fragment directly to each sample (10 μL of 0.4 mM biotin-14-dCTP, 0.5 μL of 10 mM dATP, 0.5 μL of 10 mM dGTP, 0.5 μL of 10 mM dTTP, 4 μL of 5U/μL DNA Polymerase I Klenow Fragment). The samples were then mixed by gentle pipetting followed by incubation at 37 °C for 90 min with gentle agitation.

Proximity ligation and de-crosslinking

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

DNA extraction

Phenol-chloroform extraction columns were spun down at 17000 $\times$ g for 1 min at 4 °C to get gel down to the bottom of the tube. The samples incubated overnight were then added to the column. Next, an equal volume ($\sim$ 325 μL) of phenol-chloroform-isoamyl alcohol mixture (25:24:1) (Sigma) was also added to the column. The column was then inverted for thorough mixing and spun down at 17000 $\times$ g for 5 min at 4 °C. The surface layer on top of the gel upon spinning contains the sample and is transferred to a clean 1.5 mL tube ($\sim$ 325 μL). Each sample was mixed with 31.5 μL of 3 M sodium acetate, 2 μL of GlycoBlue (ThermoFisher Scientific), and 504 μL of 100% ethanol for DNA precipitation. The samples were inverted several times for mixing and incubated at -80 °C for 20

min, followed by a centrifuge spin at 17000 $\times$ g for 45 min at 4 °C. The supernatant was carefully discarded and the pellet was washed with 800 μL of ice-cold 70% ethanol followed by a centrifuge spin at 17000 $\times$ g for 5 min at 4 °C. The supernatant was then discarded and the tube was air-dried until no traces of ethanol was left prior to dissolving the DNA pellet with 30 μL of Elution Buffer (Qiagen PCR Clean-Up Kit). 1 μL of RNase A (ThermoFisher Scientific) was added to each sample followed by incubation at 37 °C for 15 min. A mix of 5 μL of 10X NEB2.1 buffer (New England Biolabs), 1.25 μL of 1 mM dATP, 1.25 μL of 1 mM dCTP, 1.25 μL of 1 mM dGTP, 1 mM of dTTP, 0.5 μL of 10 mg/mL BSA, 5 μL of water, 3.5 μL of 3 U/μL T4 DNA Polymerase (New England Biolabs) was added to each sample. The samples were mixed thoroughly by gentle pipetting, and then incubated at 20 °C for 4 h.

Fragmentation and biotin pull-down

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

Library preparation and size selection

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

XP beads and sent for BioAnalyzer analysis prior to sequencing.

3.4.3 Hi-C Sequencing and data pre-processing

Sequencing

The Hi-C libraries for each tumour sample were sent for shallow paired-end 150 bp sequencing ($\sim 10 - 15 \times 10^6$ reads per sample) on an Illumina NextSeq 500 at the Princess Margaret Genomics Centre. Upon confirming library quality and low duplication rates ($< 2\%$), samples were sent for deep paired-end 150 bp sequencing with the aim of 800 million raw read pairs per sample on an Illumina NovaSeq 6000.

Sequence alignment and Hi-C artefact removal

Paired-end FASTQ files were pre-processed with HiCUP (v0.7.2) [**wingettHiCUPPipelineMapping2015**]. Reads were truncated at MboI ligation junction sites prior to alignment with `hicup_digester`. Each mate was independently aligned to the GRCh38 genome and were then paired and assigned to MboI restriction sites by `hicup_map`. `hicup_map` uses Bowtie2 (v2.3.4) [**langmeadFastGappedreadAlignment2012**] as the underlying aligner which has the following parameters: `--very-sensitive --no-unal --reorder`. Reads that reflect technical artefacts were filtered out with `hicup_filter`. Duplicate reads were removed with `hicup_deduplicator`.

Reads that came from different sequencing batches were then aggregated for each tumour sample at this stage using `sambamba merge` (v0.6.9) [**tarasovSambambaFastProcessing2015**]. This resulted in an average of 1.12×10^9 read pairs per tumour sample (Supplementary Table 2).

Contact matrix generation and balancing

Aggregated binary alignment map (BAM) files were converted to the pairs format using pairtools (v0.2.2) [**goloborodkoMirnylabPairtoolsV02019**] and then the cooler format using the cooler package (v0.8.5) [**abdennurCoolerScalableStorage2020**]. The pairs files were generated with the following command:

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

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

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

The raw contact matrices stored in the cooler file format were balanced using cooler's implementation of the iterative correction and eigenvector decomposition (ICE) algorithm [**imakaevIterativeCorrectionHiC20** using the `cooler balance` command. Contact matrices at different resolutions were created with the `cooler zoomify` command.

3.4.4 Hi-C data analysis

TAD identification

Contact matrices were binned at a resolution of 40 kbp. To remove sequencing depth as a confounding factor, contact matrices for all samples were first downsampled to match the sequencing depth of the shallowest sample. For comparisons including cell lines, this was 120×10^6 contacts. For comparisons only involving primary samples, this was 300×10^6 contacts. This was achieved with Cooltools (v0.3.2) [**venevMirnylabCooltoolsV02020**] with the following command:

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

TADs were identified using TopDom [**shinTopDomEfficientDeterministic2016**] on the downsampled, ICE-normalized contact matrices. To identify domains at multiple length scales, similar in concept to Artamus' gamma parameter [**filippovaIdentificationAlternativeTopological2014**], TopDom was run multiple times per sample, with the window size parameter set at values between 3 and 40, inclusive (corresponding to 120 kbp and 1.6 Mbp). The lower bound for the window size parameter allowed for the identification of domains multiple megabases in size at the upper end and domains < 100 kbp at the lower end without being dominated by false calls due to sparsity of the data. Despite TopDom being more resistant to confounding by sequencing depth than other TAD calling tools [**forcatoComparisonComputationalMethods2017**], biases in boundary persistence were evident between samples of different sequencing depth. Downsampling contact matrices to similar depths resolved these biases.

Given the stochasticity of Hi-C sequencing, boundaries called at one window size may not correspond to the exact same location at a different window size. To attempt to resolve these different boundary calls and leverage power from multiple window sizes, boundaries for a given patient were considered at all window sizes. Boundaries within one bin (40 kbp) of each other and called at

different window sizes were marked as conflicting calls. If only two boundaries were in conflict and all the window sizes where the first boundary was called are smaller than the window sizes where the second boundary was called, the second boundary was selected since larger smoothing windows are less sensitive to small differences in contact counts. If only two boundaries were in conflict but there is no proper ordering of the window sizes, the boundary that was identified most often between the two was selected. If three boundaries are in conflict, the middle boundary was selected. If four or more boundaries were in conflict, the boundary that was identified most often was selected.

To determine the maximum window size for TAD calls, TAD calls were compared across window sizes for the same patient using the BPscore metric [**zaborowskiBPscoreEffectiveMetric2019**]. TAD calls are identical when the BPscore is 0, and divergent when 1. The cut-off window size for a single patient was determined when the difference between TAD calls at consecutive window sizes was < 0.005, twice in a row. The maximum window size was determined by the maximum window size cut-off across all samples in a comparison. For comparisons involving only primary samples, the maximum window size was determined to be $w = 20 \times 40$ kbp. For comparisons involving cell lines, this was $w = 32 \times 40$ kbp.

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

Sample clustering by TADs

Using the TAD calls at the window size $w = 32 \times 40$ kbp, the similarity between samples was calculated with BPscore. The resulting matrix, containing the similarity between any two samples, was used as the distance matrix for unsupervised hierarchical clustering with `ward.D2` linkage [**murtaghWardHierarchicalAgglomerative2014**].

Compartment identification

Contact matrices were binned at a resolution of 40 kbp, similarly to TAD identification. To remove sequencing depth as a confounding factor, contact matrices for all samples were first downsampled to match the sequencing depth of the shallowest sample. Contact matrix eigenvectors were calculated with Cooltools. To standardize the sign of each eigenvector, the GC content of the reference genome, binned at 40 kbp, was used as a phasing reference track. This reference track was calculated with the `frac_gc` function from the Bioframe Python package (v0.0.12) [**nezarabdennurMirnylabBioframeV02020**]. The first eigenvector was used to identify compartments with the following command:

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

Identification of significant chromatin interactions

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

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

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

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

Chromatin interaction saturation analysis

To estimate the detection of all chromatin interactions across all samples, a nonlinear regression on an asymptotic model was performed. This is similar in method to peak saturation analysis used to assess peaks detected in ChIP-seq experiments from a collection of samples [**kronTMPRSS2ERGFusion2017**]. Bootstrapping the number of unique interactions detected in a random selection of n samples was calculated for n ranging from 1 to 17. 100 iterations of the bootstrapping process were performed. An exponential model was fit against the mean number of unique interactions detected in n samples using the `nls` and `SSaymp` functions from the stats R package (v3.6.3). The model was fit to the following equation:

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

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

Structural variant breakpoint pair detection

Breakpoint pairs for each patient were called on the merged BAM files [**liSequenceAlignmentMap2009**] using `hic_breakfinder` (commit 30a0dcc6d01859797d7c263df7335fd2f52df7b8) [**dixonIntegrativeDetectionAnaly** Pre-calculated expected observation files for the GRCh38 genome were downloaded from the `hic_breakfinder` git repository on July 24, 2019, as per the instructions. Breakpoints were explicitly called with the following command:

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

For the T2E fusion, only one patient had the deletion identified by `hic_breakfinder` with default parameters (CPCG0336). Difficulties identifying SVs with `hic_breakfinder` have been previously noted [**hoStructuralVariationSequencing2020**]. After adjusting the detection threshold, we were able to identify the fusion in other samples. To ensure the T2E+ tumours were effectively stratified for future analyses, the fusion was annotated using the same coordinates for the other T2E+ samples. No other additions to breakpoint calls were made. Certain breakpoints that appeared to be artefacts were removed, as described below.

Structural variant annotation and graph construction

The contact matrix spanning 5 Mbp upstream and downstream around the breakpoint pairs were plotted and annotated according to previously published heuristics (Supplementary Figure 4 for [**dixonIntegrativeDetectionAnalysis2018**]). Breakpoint pairs that were nearby other breakpoints or did not match the heuristics in this figure were labelled as `UNKNOWN`. These annotations were matched against the annotations identified from the previously published WGS SVs [**fraserGenomicHallmarksLocalized2017**]. Breakpoint pairs matching the following criteria were considered as detection artefacts and were ignored.

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

To identify unique breakpoints that were identified in multiple breakpoint pairs, breakpoints that were within 50 kbp of each other were considered as possibly redundant calls. This distance

was considered as the resolution of the non-artefactual calls is 100 kbp. Plotting the contact matrix 5 Mbp around the breakpoint, breakpoints calls were considered the same breakpoint if the sharp edge of each breakpoint was equal to within 5 kbp. Similar in concept to the `ChainFinder` algorithm [**bacaPunctuatedEvolutionProstate2013**], we consider each breakpoint as a node in a graph. Nodes are connected if they are detected as a pair of breakpoints by `hic_breakfinder`. Simple SVs are connected components in the breakpoint graph containing only two nodes, and complex variants those with greater than two nodes. A visual representation of these graphs can be found in Figure B.4b. Graphs are displayed with a spring-force layout, adjusted using the Kamada Kawai optimization [**kamadaAlgorithmDrawingGeneral1989**] from the NetworkX Python package (v2.4) [**hagbergExploringNetworkStructure2008**].

Determination of SV breakpoints altering intra-TADs contacts

Patients are assigned into one of two groups using hierarchical clustering (complete linkage) with the matrix of pairwise BPscore [**zaborowskiBPscoreEffectiveMetric2019**] values as a distance matrix. If the clustering equals the mutated samples from the non-mutated samples (i.e. the clustering matches the mutation status in this locus), then the local topology was classified as `altered` because of the SV.

Virtual 4C

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

Complex structural variant assembly

NeoLoopFinder [**wangGenomewideDetectionEnhancerhijacking2021**] was used to estimate copy number aberrations from Hi-C data with the following command:

```
calculate-cnv -g hg38 -e MboI -H {input_cooler} --output {cnv_estimates}
```

This estimation was subsequently used to perform copy number-aware normalization of 10 kbp resolution contact matrices with the following command:

```
correct-cnv -H {input_cooler} --cnv-file {cnv_estimates} --nproc 8
```

Complex SVs were assembled using SV breakpoints identified by Hi-C Breakfinder [**dixonIntegrativeDetectionA** and the copy number-aware normalized contact matrices with the following command:

```
assemble-complexSVs --breakpoints {sample_SV_breakpoints} --hic {
    input_cooler} --nproc 8
```

Assembled contact matrices were plotted with a custom Python script using the Cooler and NeoLoopFinder Python packages.

3.4.5 Patient tumour tissue H3K27ac ChIP-seq

ChIP-seq against H3K27ac was previously published for these matching samples in [**kronTMPRSS2ERGFusion** Sequencing data was processed similarly to the previous publication of this data [**kronTMPRSS2ERGFusion2017**]; however, the GRCh38 reference genome was used instead of GRCh37 [**internationalhumangenomesequencingcon**

Sequence alignment

FASTQ files from single-end sequencing were aligned to the GRCh38 genome using Bowtie2 (v2.3.4) with the following command:

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

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

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

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

Peak calling

Peak calling was performed using MACS2 (v2.1.2) [**zhangModelbasedAnalysisChIPSeq2008**] with the following command:

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

ENCODE GRCh38 blacklist regions were then removed from the narrow peaks [**amemiyaENCODEBlacklistIde**
Peaks calls are in Supplementary Table 8.

Differential acetylation analysis

Unique peak calls and de-duplicated pull-down and control BAM files from tumour samples were loaded into R with the DiffBind package (v2.14.0) [**starkDiffBindDifferentialBinding2011**] using DESeq2 (v1.26.0) [**loveModeratedEstimationFold2014**] as the differential analysis model. 3 of the 12 samples had low quality peak calls compared to the other 9 and were not considered when calculating differential acetylation (CPCG0268, CPCG0255, and CPCG0336). We considered each unique breakpoint one at a time in the remaining 9 samples. Samples were grouped by their mutation status (i.e. a design matrix where the mutation status is the only covariate) and DiffBind's differential binding analysis method was performed to identify all differentially acetylated regions between the two groups. Acetylation peaks outside of the TADs overlapping the breakpoint were filtered out. Multiple test correction with the Benjamini-Hochberg FDR method [**benjaminiControllingFalseDiscovery1995**] was performed on all peaks after all breakpoints were considered, due to similar group stratifications depending on the breakpoint under consideration.

Structural variant breakpoint enrichment

Structural variant breakpoint coordinates from WGS data from the CPC-GENE cohort were obtained from the International Cancer Genome Consortium (structural somatic mutations from the PRAD-CA dataset, release 28, `https://dcc.icgc.org/releases/release_28/Projects/PRAD-CA`) [**zhangInternationalCancerGenome2019**]. Breakpoint coordinates were lifted over to GRCh38 coordinates using the liftOver function from the rtracklayer R package (v1.46.0) [**lawrenceRtracklayerPackageInte**
Permutation tests were performed with the regioneR R package (v1.18.0) [**gelRegioneRBioconductorPackage201**
selecting randomized regions from the GRCh38 genome, excluding the ENCODE blacklist regions

[**amemiyaENCODEBlacklistIdentification2019**] and masked loci. 100 permutations were calculated and a one-sided permutation z-test was used to calculate statistical significance.

3.4.6 Primary tissue RNA data analysis

Tumour sample RNA sequencing

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

RNA sequencing data pre-processing

RNA-seq FASTQ files were pseudo-aligned to the GRCh38 genome using Kallisto (v0.46.1) [**brayNearoptimalProbabilisticRNAseq2016**] with the following command:

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

Differential gene expression analysis

To assess whether SVs were associated with local gene expression changes, we considered each unique breakpoint one at a time. For each breakpoint, we compared the gene expression between the mutated and non-mutated tumour samples using Sleuth (v0.30.0) [**pimentelDifferentialAnalysisRNAseq2017**, **yiGenelevelDifferentialAnalysis2018**] with a linear model where the mutation status was the only covariate. To reduce the chance of falsely identifying genes as differentially expressed, only genes located within the TADs (window size $w = 20$) containing breakpoints were considered. Fold change estimates of each transcript were assessed for significance using a Wald test. Transcript-level p-values are combined to create gene-level p-values using the Lancaster aggregation method provided by the Sleuth package [**yiGenelevelDifferentialAnalysis2018**]. Correcting for multiple tests was then performed with the Benjamini-Hochberg FDR correction for all genes that were potentially altered in the mutated sample(s).

3.5 Code and data availability

Whole genome and RNA-seq data are available in the European Genome-Phenome Archive (EGA) under accession number EGAS00001000900. H3K27ac ChIP-seq data are available in EGA under accession number EGAS00001002496. Hi-C sequencing data are under submission to the EGA (EGAS00001005014). TADs and chromatin interactions are available in the Gene Expression Omnibus (GEO) archive under accession number GSE164347. Hi-C sequencing data from cell lines was obtained from GEO under accession number GSE118629 (22Rv1, RWPE1, and C4-2B cell lines) and from teh 4D Nucleome under accession numbers 4DNFI6HDY7WZ (H1-hESC Rep 1), 4DNFITH978XV (H1-hESC Rep 2), 4DNFIT64Q7A3 (HAP-1 Rep 1), 4DNFINSKEZND (HAP-1 Rep 2), 4DNFIIV4M7TF (GM12878 Rep 1), and 4DNFIXVAKX9Q (GM12878 Rep 2).

Processed data and code for processing, analysis, and plotting can be found on CodeOcean (`https://codeocean.com/capsule/5232537/tree`) and in GitHub (`https://github.com/LupienLab/3d-reorganization-prostate-cancer`).

Chapter 4

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

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

4.1 Abstract

Differential analysis in RNA sequencing assays requires controlling biological and technical sources of noise to make accurate statistical inferences. The use of biological replicates in experiments can help reduce the impact these sources of noise, but are not always readily available, such as in patients with complex mutations or rare diseases. These non-ideal experimental designs increase the uncertainty of differential analysis of RNA sequencing data. To combat this uncertainty, we introduce a statistical method for suboptimal experimental designs using the James-Stein estimator for a normal distribution. Compared to conventional methods, we find a 7-23 % error reduction in gene expression fold change estimates. We propose that this approach to reducing noise in differential expression analysis can be useful in single-cell analyses and studying rare diseases.

4.2 Introduction

In statistical modelling, discrepancies between model predictions and observations can be categorized into either the variance or bias of the model, known as the "bias-variance tradeoff". When evaluating statistical models of gene expression data against high throughput RNA sequencing (RNA-seq) data, many methods have prioritized reducing variance to control errors in gene expression fold changes [**robinsonEdgeRBioconductorPackage2010**, **loveModeratedEstimationFold2014**, **hansenRemovingTechnicalVariability2012**, **trapnellDifferentialAnalysisGene2013**, **liRSEMAccurateTra** **hardcastleBaySeqEmpiricalBayesian2010**, **ritchieLimmaPowersDifferential2015**, **lawVoomPrecisionWei** **lengEBSeqEmpiricalBayes2013**, **liModelingAnalysisRNAseq2018**, **rissoNormalizationRNAseqData201** **bullardEvaluationStatisticalMethods2010**, **pimentelDifferentialAnalysisRNAseq2017**, **yiGenelevelDiffe** Using empirical Bayesian [**hardcastleBaySeqEmpiricalBayesian2010**, **lengEBSeqEmpiricalBayes2013**] models, quantile normalization [**hansenRemovingTechnicalVariability2012**], and generalized linear models (GLMs) with negative binomial distributions have greatly reduced the false discovery rates in RNA-seq analyses. Moreover, controlling for confounding variables such as library size [**bullardEvaluationStatisticalMethods2010**], batch effects [**leekSvaPackageRemoving2012**], and gene length [**bullardEvaluationStatisticalMethods2010**, **loveModeratedEstimationFold2014**, **robinsonEdgeRBioconductorPackage2010**, **oshlackTranscriptLengthBias2009**] have also reduced errors in quantifying gene expression differences. One of the simplest ways to reduce the impact of variance on differential analysis is to increase the number of biological replicates in each experimental condition and to balance the size of experimental groups. Under ideal conditions, it has been shown that RNA-seq experiments should contain $6-12$ biological replicates of each condition [**schurchHowManyBiological2016**]. However, this is rarely possible due to cost limitations of RNA-seq assays and sample availability. For example, the Orphanet Database of rare diseases estimates that $\sim 4\,500$ diseases have a prevalence $< 10^{-6}$ [**nguengangwakapEstimatingCumulativePoint2020**], many of which are genetic in nature. This means individuals identified with the disease will likely not have biological replicates available for differential analysis. Similarly, in Chapter 3, nearly all structural variants (SVs) were not found in more than a single patient. Estimating the impact of a given mutation only found in a single individual is difficult, and recapitulating complex SVs often found in primary prostate tumours in cell line models remains challenging [**nakamuraCRISPRTechnologiesPrecise2021**, **pickar-oliverNextGenerationCRISPR2019**, **wangEngineering3DGenome2021**].

To address this problem of unreplicated biological samples in differential gene expression analy-

ses, we attempt to reduce the mean square error (MSE) of gene expression fold change estimates from models of RNA-seq data by addressing the bias (Figure 4.1a). Using the James-Stein (JS) estimator for the mean of normal distribution based on a single observation [**steinInadmissibilityUsualEstimator1956**, **bockMinimaxEstimatorsMean1975**, **steinEstimationMeanMultivariate1981**], we reduce the MSE of fold change estimates by simultaneously increasing the bias and decreasing the variance of the statistical model (Figure 4.1b). This method works by combining variation across dimensions for a multivariate observation and uses this combined information to reduce the estimate error (Figure 4.1c). First, we use the Sleuth model [**pimentelDifferentialAnalysisRNAseq2017**, **yiGenelevelDifferentialAnalysis2018**] for differential analysis to derive the JS estimator fold gene expression fold change and relate it to the ordinary least squares (OLS) estimator. Then, using random samplings from a highly replicated RNA-seq experiment [**gierlinskiStatisticalModelsRNAseq2015**], we compare the differences in statistical inferences between the JS and OLS estimators. Finally, we investigate how the number of transcripts under consideration affects the reduction in MSE, suggesting how this method can be best used in practice.

4.3 Derivation of the James-Stein fold change estimator

This section will define variables and symbols as they arise. For a complete list of conventions and notation used in this section, see Appendix C.1.

For a p-variate normal distribution with mean μ and covariance matrix equal to the identify matrix, denoted as $\mathcal{N}_p(\mu, I_p)$, if the mean is unknown, the following theorem holds [**steinInadmissibilityUsualEstimat**

Theorem 1 *Consider the distribution $\mathcal{N}_p(\mu, I_p)$ with unknown mean, μ, from which a single sample, Z, is drawn. Consider an estimator for the mean, $\hat{\mu}^{(OLS)}$, that is equal to the single observation (i.e. $\hat{\mu}^{(OLS)} = Z$). The estimator $\hat{\mu}^{(OLS)}$ does not minimize the MSE, $\mathbb{E}\left[(\mu - \hat{\mu}^{(OLS)})^2\right]$ over all possible estimators. Namely, the estimator $\hat{\mu}^{(JS)} = \left(1 - \frac{b}{a+\|Z\|^2}\right) Z$ has a smaller MSE than $\hat{\mu}^{(OLS)}$ for a sufficiently small coefficient b and sufficiently large coefficient a.*

This result was generalized to non-singular covariance matrices that were not necessarily the identity matrix [**jamesEstimationQuadraticLoss1961**, **bockMinimaxEstimatorsMean1975**]:

Theorem 2 *Consider the distribution $\mathcal{N}_p(\mu, \Sigma)$ where the mean, μ, is unknown and the covariance matrix, Σ, is known. Consider a single sample, Z, from this distribution and let $\hat{\mu}^{(JS)} = \left(1 - \frac{c}{Z^{\mathrm{T}}\Sigma^{-1}Z}\right) Z$ where c is some scaling coefficient be an estimator for μ. If $p \geq 3$, $\mathbb{T}[\Sigma] \geq 2\lambda$*

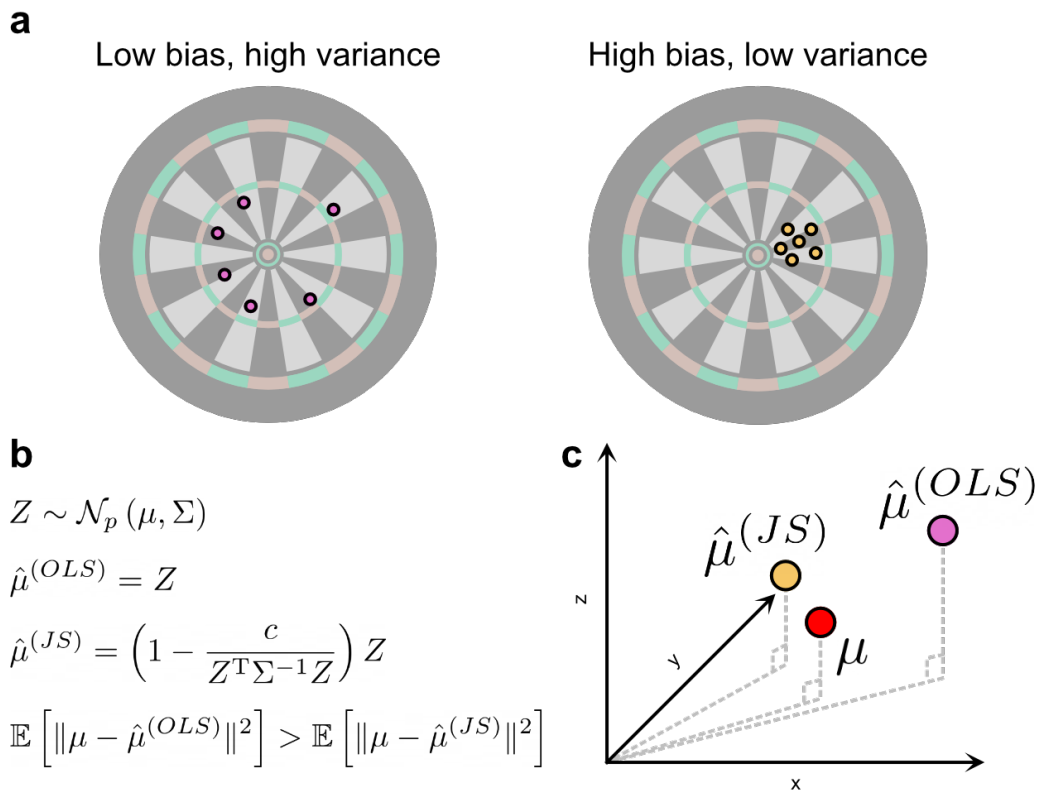

Figure 4.1: **Reducing the bias-variance tradeoff by combining information across multiple features. a.** Schematic of the bias-variance tradeoff for assessing model performance. Dartboard on the left shows low bias of the darts (mean is close to the bullseye) but a large variance. Dartboard on the right shows a high bias of the darts (mean is off-centre), but a small variance. **b.** For a p-variate random variable, Z, with a normal distribution with mean μ and covariance matrix Σ from which a single observation is made, the OLS estimator for the mean, $\hat{\mu}^{(OLS)}$, has a higher MSE than the JS estimator, $\hat{\mu}^{(JS)}$. **c.** A schematic showing how the JS estimators work in theory. The OLS estimate of a random variable with 3 dimensions from a single observation will give an estimate that is some distance from the truth. Combining information from the three variables in the JS estimate can produce an estimate that is closer to the truth.

(where $\mathbb{T}[\cdot]$ is the trace of a matrix, λ is the largest eigenvalue of the covariance matrix, Σ), and $0 \leq c \leq 2\left(\frac{\mathbb{T}[\Sigma]}{\lambda} - 2\right)$, then $\hat{\mu}^{(JS)}$ is the estimator that minimizes the MSE.

We can demonstrate how the Sleuth statistical model for performing differential expression analysis [**pimentelDifferentialAnalysisRNAseq2017**, **yiGenelevelDifferentialAnalysis2018**] can be transformed to fit the criteria for Theorem 2.

Consider an experiment with n_{WT} wild-type (WT) samples and 1 mutant sample where RNA-seq is performed. Sequencing reads from each sample are mapped to each transcript in the organism's transcriptome, S. For each of the n_{WT} WT samples, RNA-seq read counts can be modelled as:

$$D_s \sim \mathcal{N}\left(\beta_{0s}, \sigma_s^2 + \tau_s^2\right) \tag{4.1}$$

where D_s is the abundance of transcript $s \in S$ and β_{0s} is the mean abundance for transcript s. The variance in this model is decomposed into two components: biological noise arising from differences between biological replicates and library preparation, and technical noise arising from the stochastic nature of sequencing measurements and computational analysis of sequecing reads [**pimentelDifferentialAnalysisRNAseq2017**]. σ_s^2 denotes the biological variance of counts for transcript s and τ_s^2 denotes the inferential variance for transcript s. For details about this variance decomposition, see [**pimentelDifferentialAnalysisRNAseq2017**] and Appendix C.

The model for the single mutated sample is slightly modified, with a parameter β_{1s} representing the abundance fold change associated with the mutation [**pimentelDifferentialAnalysisRNAseq2017**, **loveModeratedEstimationFold2014**]:

$$D_s \sim \mathcal{N}\left(\beta_{0s} + \beta_{1s}, \sigma_s^2 + \tau_s^2\right) \tag{4.2}$$

Under this model it is assumed that both the biological and inferential variances are the same between the mutated and WT samples. All of the parameters in this model are unknown and will be estimated from the data. While β_{0s} can be estimated from the WT samples, the fold change β_{1s} is can only be estimated from the single mutated sample. By reparameterizing this model to consider every transcript, Equation (4.1) can be re-written as an $|S|$-dimensional random vector:

$$\Delta \sim \mathcal{N}_{|S|}(\mathrm{B}_0, \Sigma) \tag{4.3}$$

and Equation (4.2) can be similarly re-written as:

$$\nabla \sim \mathcal{N}_{|S|}(\mathrm{B}_0 + \mathrm{B}_1, \Sigma) \tag{4.4}$$

where Δ is the vector of abundances for all transcripts in a WT sample; ∇ is the same but for the mutated sample; B_0 is the $|S|$-dimensional vector of baseline abundances for each transcript; B_1 is the $|S|$-dimensional vector of abundance fold changes in each transcript associated with the mutation; and Σ is the covariance matrix. Mathematically, we can express these parameters in terms of parameters from Equation (4.2):

$$\mathrm{B}_{0s} = \beta_{0s} \forall s \in S$$

$$\mathrm{B}_{1s} = \beta_{1s} \forall s \in S$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \tau_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{|S|}^2 + \tau_{|S|}^2 \end{bmatrix}$$

The Sleuth statistical model can be used derive estimates for B_0 and Σ solely from the n_{WT} WT samples (hereafter denoted with the $\hat{}$ symbol). If we treat the estimates $\hat{\mathrm{B}}_0$ and $\hat{\Sigma}$ derived from the WT samples as plug-in parameters for the mutated model, then Equation (4.2) for a single mutated sample meets the criteria for the JS estimators. This is not necessarily an accurate assumption, since these estimates may be biased, inaccurate for small n_{WT}, or affected by batch effects that are not shared between the WT and mutant samples. However, this assumption is made out of necessity due to the constraints of measuring a single sample. With sufficiently large n_{WT} and quality control in RNA-seq library preparation, this may not be a practical concern.

Under these assumptions, we have the following distribution from which we are drawing a single observation:

$$\nabla - \hat{\mathrm{B}}_0 \sim \mathcal{N}_{|S|}\left(\mathrm{B}_1, \hat{\Sigma}\right) \tag{4.5}$$

By defining $Z = \nabla - \hat{\mathrm{B}}_0$, we can see that this satisfies the criteria for Theorem 2. A JS estimator for the unknown fold change, B_1, can be constructed:

$$\hat{\mathrm{B}}_1^{(JS)} = \left(1 - \frac{c}{(\nabla - \hat{\mathrm{B}}_0)^{\mathrm{T}} \hat{\Sigma}^{-1} (\nabla - \hat{\mathrm{B}}_0)}\right)\left(\nabla - \hat{\mathrm{B}}_0\right) \tag{4.6}$$

4.3.1 Comparison between the OLS and JS estimators

To demonstrate how the JS estimator, $\hat{\mathrm{B}}_1^{(JS)}$, compares to conventional approaches for estimating gene expression fold change that do not make use of Theorem 2, we can consider the OLS estimator. For the experimental design described above, the OLS estimator for B_1, $\hat{\mathrm{B}}_1^{(OLS)}$, is given by:

$$\hat{\mathrm{B}}_1^{(OLS)} = \nabla - \hat{\mathrm{B}}_0 \tag{4.7}$$

Substituting this into Equation (4.6) yields a simplified form of the JS estimator for B_1:

$$\hat{\mathrm{B}}_1^{(JS)} = \left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \hat{\mathrm{B}}_1^{(OLS)}}\right) \hat{\mathrm{B}}_1^{(OLS)} \tag{4.8}$$

From this definition, one can see that the JS estimate is colinear with the OLS estimate but uniformly shrunk towards 0. We can summarize the above with the following theorem.

Theorem 3 *For an experiment containing n_{WT} WT samples and a single mutated sample, an estimate for the expression fold change of $|S|$ transcripts can be given by the JS estimator Equation* (4.8) *where $\hat{\Sigma}$ is the covariance matrix for all $|S|$ transcripts estimated from all n_{WT} WT samples; $\hat{\mathrm{B}}_1^{(OLS)}$ is the OLS estimator for expression fold change given by Equation* (4.7)*; and c is some scaling coefficient. When $|S| \geq 3$, $\mathbb{T}\left[\hat{\Sigma}\right] \geq 2\lambda$ (where λ is the largest eigenvalue of $\hat{\Sigma}$), and c satisfies Equation* (C.15)*, then the MSE of the JS estimator, $\hat{\mathrm{B}}_1^{(JS)}$, is smaller than the MSE of the OLS estimator, $\hat{\mathrm{B}}_1^{(OLS)}$.*

It can be shown that the JS estimator is biased towards 0 with a smaller variance than the OLS estimator (see Appendix C.5). In theory, this may increase the error of some transcripts, but will decrease the MSE for a set of transcripts in aggregate (see Appendix C.5).

There are two parameters of this model that will affect the amount of biasing: the scaling coefficient, c, and the transcripts being considered for differential expression, S. Firstly, the scaling coefficient can be manually specified, and the largest biasing occurs when c is its maximum value, $2\left(\frac{\mathbb{T}[\Sigma]}{\lambda} - 2\right)$. Secondly, the transcripts under consideration can also be manually specified, which will affect the value of the denominator, $\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}} \hat{\Sigma}^{-1} \hat{\mathrm{B}}_1^{(OLS)}$, and thus the amount of biasing. The more transcripts under consideration, the larger expected value for the denominator, and so the smaller the biasing effect. Taken together, we have produced a high-bias, low-variance fold change estimator that has a lower MSE than the OLS estimator and two tunable parameters.

4.4 Results

To measure the performance of this method in practice, we make use of a highly replicated RNA-seq experiment involving Δ*Snf2* knockout (KO) and WT *Saccharomyces cerevisiae* cells [**gierlinskiStatisticalModel** This dataset contains 48 biological replicates of each condition, an infeasible sample size for most RNA-seq experiments. We randomly select N total samples from the full dataset with an optimal even split between the two groups (i.e. $N/2$ Δ*Snf2* and $N/2$ WT) or a suboptimal $(N-1)$-vs-1 split (i.e. $N-1$ Δ*Snf2* and 1 WT or 1 Δ*Snf2* and $N-1$ WT). To measure the effect of total sample size, this random selection is repeated for multiple values of N. Experiments with smaller sample sizes can be compared to the full dataset to estimate the number of true and false detections for a given experimental design and a given method. For the purposes of this comparison, we assume the fold change estimates obtained from the full dataset are accurate without bias and that all determinations of significant differential expression are true. This assumption is made to focus on the potential model improvements provided by the JS estimator over the OLS model. This is also due to the practical considerations of designing RNA-seq experiments and that this experimental data is as close to ideal as can be found in the field.

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

To investigate where the changes in statistical inferences come from, we can view a representative sample (Figure 4.3). In the full dataset with 48 biological replicates, *Snf2* is the most under-expressed gene with an estimated 99 % reduction in expression (Figure 4.3a). Three other example genes, *PHO12*, *TIS11*, and *TYE7*, are also under-expressed in the Δ*Snf2* cells. Using 4 samples in

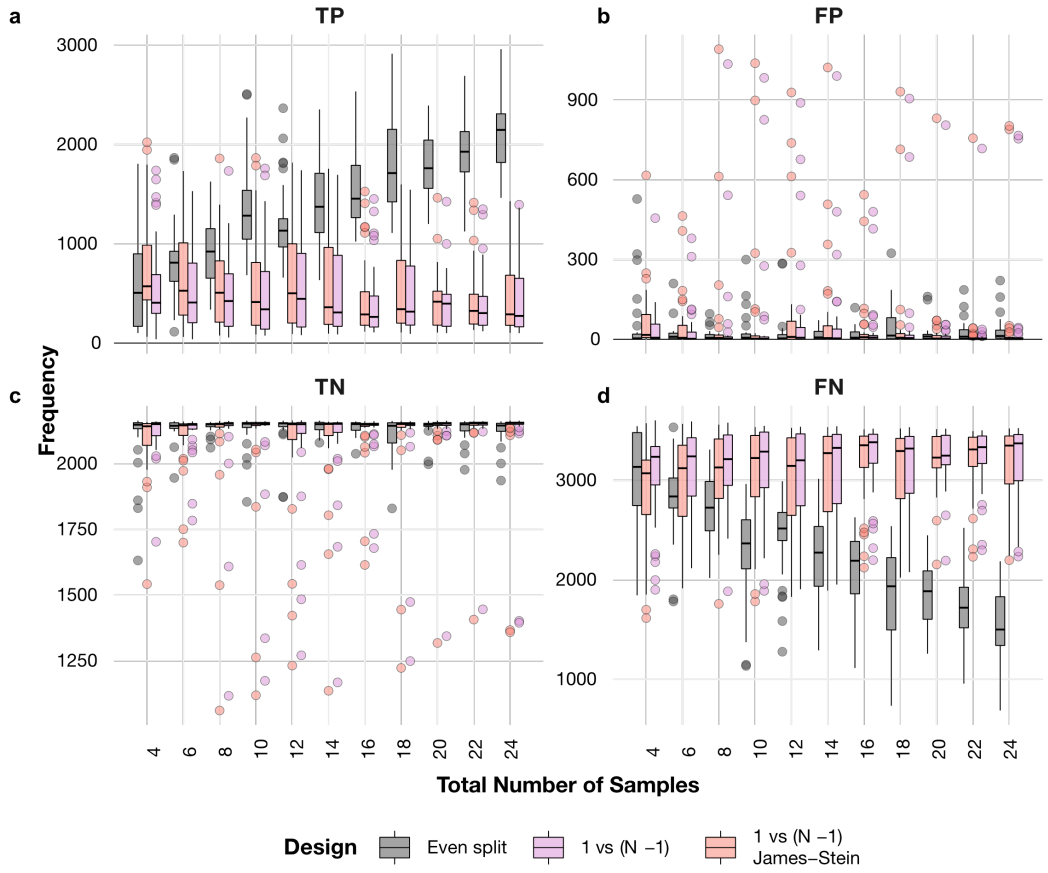

Figure 4.2: **Differential gene expression analysis of the entire yeast transcriptome with differently sized experimental designs.** Random samplings ($n = 30$) were compared to the full dataset of 48 Δ*Snf2* vs 48 WT to calculate TP (**a**), FP (**b**), TN (**c**), and FN (**d**).

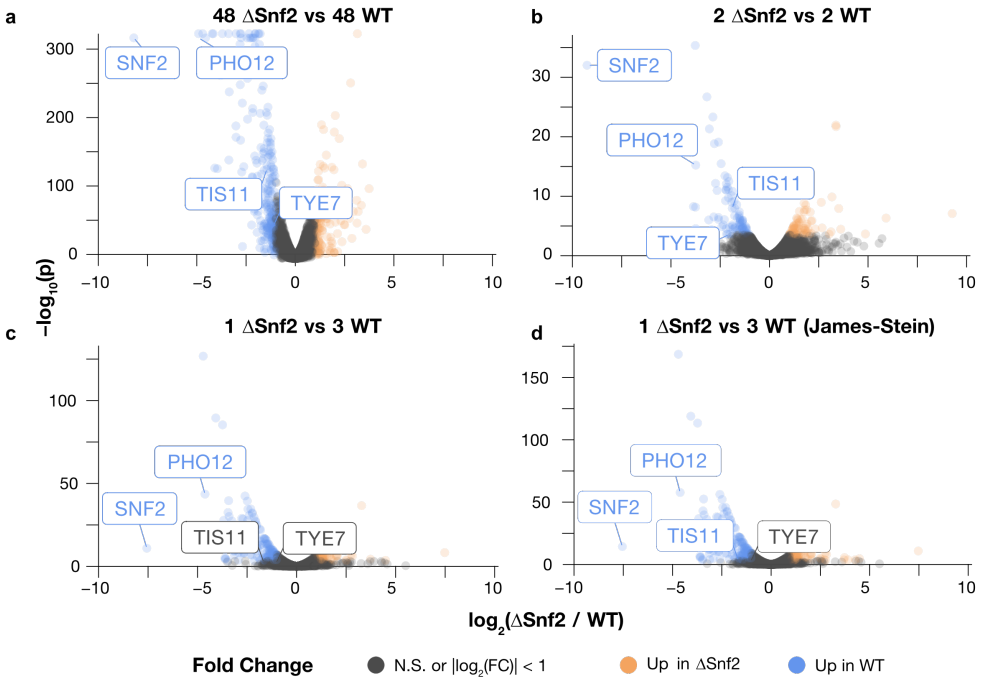

Figure 4.3: **Differential gene expression analysis of Δ*Snf2* vs WT yeast cells using different sample sizes and experimental designs. a.** Volcano plot of differential expression results with OLS estimates in a highly replicated experiment consisting of 48 biological replicates of each condition. **b.**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

total, evenly split between the two groups, all four genes remain detected as differentially expressed (Figure 4.3b). Using a suboptimal design with 1 Δ*Snf2* and 3 WT samples, *TIS11* and *TYE7* are no longer detected as differentially expressed using the OLS method (Figure 4.3c). However, the *TIS11* gene is detected as differentially expressed using the JS method (Figure 4.3d). The fold change estimates in Figure 4.3c-d are similar, since the biasing is small effect. Thus, the differences in statistical inference result from the smaller variance of the JS estimator, as predicted.

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

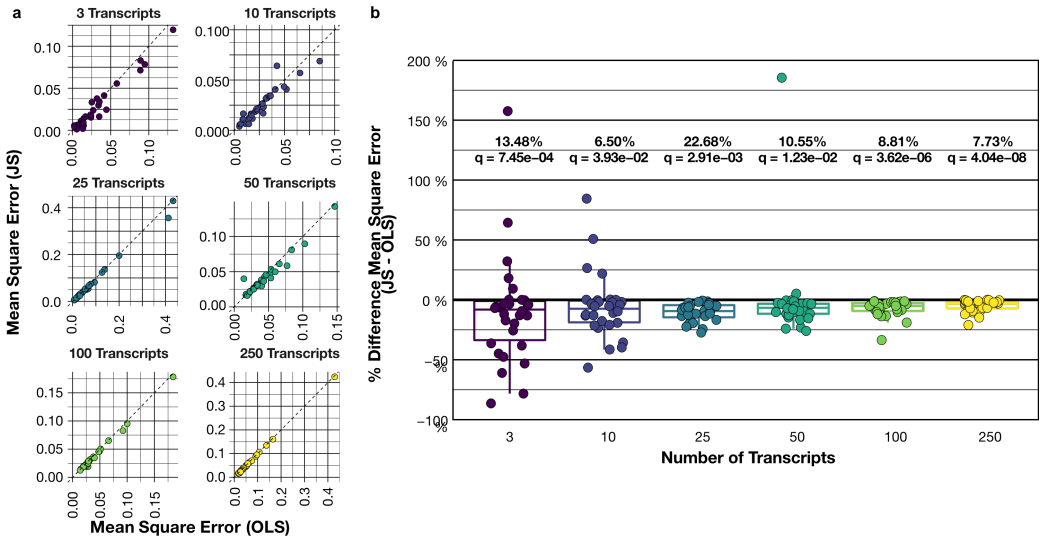

Figure 4.4: **Differential gene expression analysis focusing on a subset of trancsripts, not the entire transcriptome.** All experiments use 1 Δ*Snf2* vs 5 WT samples (or vice versa). **a.** Comparison of the MSE of the JS estimates (y-axis) against the OLS estimates (x-axis). The total number of transcripts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

However, the effect is also more variable when fewer transcripts are considered, as some samplings with 3 transcripts resulted in an 150 % increase in error (Figure 4.4b). Taken together, we find that the JS method significantly reduces the fold change MSE, with greater reductions found by considering a smaller number of transcripts.

4.5 Discussion

In this work, we developed a statistical framework for differential expression analysis where unreplicated conditions are found. We found that its larger bias and smaller variance, compared to conventional OLS estimators, can improve statistical inferences, particularly when focusing on ≤ 100 transcripts. In cases with unreplicated conditions, the DESeq2 [**loveModeratedEstimationFold2014**], edgeR [**robinsonEdgeRBioconductorPackage2010**], and Sleuth [**pimentelDifferentialAnalysisRNAseq2017 yiGenelevelDifferentialAnalysis2018**] frameworks group all samples, both case and control, before performing normalization and performing variance shrinkage. The method presented here builds on these shrinkage methods to reduce uncertainty in gene expression fold change estimates for when biological replicates are missing. This method can help improve differential gene expression analyses for singular observations, such as unique chromatin aberrations or rare diseases.

There are some limitations to this work. The WT estimates, $\hat{\mathrm{B}}_0$ and $\hat{\Sigma}$, are used as plug-in

estimates for Equation (4.5). Plug-in estimates under-estimate the variation in this distribution and thus may lead to more frequent false discoveries [**efronIntroductionBootstrap1993**]. Evaluation of the sensitivity of the JS estimator to this effect was not performed here. Future studies using simulated data with known expression fold changes should be performed to assess this sensitivity.

Many statistical methods exist to similarly reduce uncertainty. However, as has been shown here, balanced experimental designs, where the number of case and control samples are equal, outperformed both the OLS and JS methods. This is similarly the case with other differential analysis methods [**schurchHowManyBiological2016**, **gierlinskiStatisticalModelsRNAseq2015**]. Complex statistical methods often cannot overcome limitations in sample size. Thus, technological development for biological models that recapitulate the chromatin state of tissues, such as organoids and cell lines, may be more advantageous for experimental validation [**zanoniModelingNeoplasticDisease2020**]. The statistical method presented here may be useful in rare disease settings, where a single patient's data may be compared against a large reference database consisting of thousands of individuals to address the potential issue of small n_{WT}. The JS estimator could be applied to the entire transcriptome, consisting of thousands of transcripts, but the method may have limited benefit in this naive application due to the large number of dimensions. To take advantage of the benefits provided by the JS estimator, one could refine the search space to transcripts involved with a gene that is known to undergo differential splicing in a related disease. Alternatively, the scenario in Chapter 3 could be addressed by only considering the transcripts of genes within a topologically associated domain (TAD) containing an SV breakpoint. This statistical method may similarly apply to single-cell RNA-seq differential analysis, where single cells are often clustered together to improve differential expression analysis inferences [**chenAssessmentComputationalMethods2019**], or other epigenomic assays, such as chromatin immunoprecipitation sequencing (ChIP-seq) [**rotemSinglecellChIPseqReveals2015**], cleavage under targets and release using nuclease (CUT&RUN) [**hainerProfilingPluripotencyFactors2019**], or assay for transposase-accessible chromatin sequencing (ATAC-seq) [**buenrostroSinglecellChromatinAccessibil**
These applications require special consideration for inflated zero counts in single-cell data [**chenAssessmentComput**
and require further investigation before adopting ideas presented here.

4.6 Methods

4.6.1 RNA sequencing data collection and pre-processing

RNA-seq reads were obtained from [**gierlinskiStatisticalModelsRNAseq2015**] via the Sequence Read Archive (SRA) [**leinonenSequenceReadArchive2011**] (Project Accession PRJEB5348). Briefly, the Δ*Snf2* KO and WT cells were both sequenced as single-end, 50 bp reads, split across seven lanes of a single Illumina HiSeq 2000 flow cell. The quality of the sequencing reads was assessed with FastQC [**andrewsFastQCQualityControl2010**]. The *Saccharomyces cerevisiae* R64-1-1 reference transcriptome index from Ensembl [**cunninghamEnsembl20192019**] (v96) was downloaded from `https://github.com/pachterlab/kallisto-transcriptome-indices/` (commit `944fbcc972ceca3b7c643c82670a04d20b0bd443`). RNA-seq reads from all technical replicates of a single biological replicate were quantified against the reference transcriptome with Kallisto (v0.46.2) [**brayNearoptimalProbabilisticRNAseq2016**] with the following command:

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

4.6.2 Differential expression analysis

Differential expression analysis was conducted in R (v4.0.2) [**rcoreteamLanguageEnvironmentStatistical201**
with the Sleuth package (v0.30.0) [**pimentelDifferentialAnalysisRNAseq2017**, **yiGenelevelDifferentialAnaly**
Statistical significance for the fold change of transcripts from the mutation status was determined with a two-sided Wald test. Multiple testing correction was performed using the Benjamini-Hochberg FDR method [**benjaminiControllingFalseDiscovery1995**]. Transcripts were determined to be significantly differentially expressed if $FDR < 0.01$.

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

4.6.3 Random sampling procedure

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

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

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

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

Differential expression analysis was calculated with the Sleuth package [**pimentelDifferentialAnalysisRNAseq yiGenelevelDifferentialAnalysis2018**]. For the JS calculations, the scaling coefficient, c, was set to its largest value of $2\left(\frac{\mathrm{Tr}(\Sigma)}{\lambda_L}-2\right)$ to reduce error as much as possible. For each sampling, the number of TP, FP, TN, and FN calls was calculated by comparing the classification of a given transcript as significantly differentially expressed or not (i.e. if $FDR < 0.01$ for a given transcript) in the sampling to the full experiment of 48 Δ*Snf2* vs 48 WT. With this confusion matrix, derived rates, such as the true positive rate, were then calculated.

4.6.4 Random sampling of smaller numbers of transcripts

In a similar fashion to above, 30 iterations of randomly selected samples were chosen to match the same three scenarios as above, with a fixed $N = 6$ (15 iterations with 1 Δ*Snf2* vs 5 WT and 15 iterations with 5 Δ*Snf2* vs 1 WT). A subset of transcripts, S, were randomly selected from those transcripts with ≥ 10 estimated reads in all 6 of the randomly selected samples. Fold change estimates for the three scenarios were calculated as above. The number of transcripts selected was chosen from $|S| \in \{3, 10, 25, 50, 100, 250\}$.

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

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

4.7 Code and data availability

RNA-seq data was obtained from the SRA [**leinonenSequenceReadArchive2011**] (Project Accession PRJEB5348). All code for pre-processing, analyses, and plotting can be found on GitHub (`https://github.com/jrhawley/n1diff`).

Chapter 5

Epigenetic dynamics underlying B-cell acute lymphoblastic leukemia relapse

This chapter is a version of the following manuscript in preparation:

hawleyEpigeneticDynamicsUnderlying2021

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

5.1 Abstract

Relapse of B-cell acute lymphoblastic leukemia 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-cell acute lymphoblastic leukemia and other leukemias, and why come cells with particular genetic or epigenetic profiles survive therapy remains unknown. Here, we use targeted genome sequencing, RNA sequencing, assay for transposase-accessible chromatin sequencing, and bisulfite sequencing of patient-matched samples with patient-derived xenografts to investigate the dynamics of the genome and epigenome over B-cell acute lymphoblastic leukemia relapse. We find that DNA methylation profiles most closely resemble genetic 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-cell acute lymphoblastic leukemia.

5.2 Introduction

After treatment, relapse of B-cell acute lymphoblastic leukemia (B-ALL) occurs in 15 - 25 % of pediatric patients and 40 - 75 % of adult patients [**inabaAcuteLymphoblasticLeukaemia2013**, **formanMythSecondRemission2013**]. Previous studies of the cells that give rise to primary and relapsed leukemias have identified newly acquired somatic mutations and copy number variants (CNVs) targeting cell cycle regulation and B-cell development [**mullighanGenomicAnalysisClonal2008**]. These cells predominantly appear to arise from a genetic subclone of cells present at diagnosis, while treatment primarily targets the dominant clone [**oshimaMutationalLandscapeClonal2016**, **oshimaMutationalL** **maRiseFallSubclones2015**, **mullighanGenomicAnalysisClonal2008**]. 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 [**leeGlobalDNAMethylation2012**, **izzoDNAMethylationDisruption2020**], and chromatin accessibility signatures in hematopoietic stem and progenitor cells (HSPCs) distinguish phenotypically distinct cell types, even with minimal changes to gene expression patterns [**takayamaTransitionQuie** Notably, the binding of the transcription factor (TF) CCCTC-binding factor (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 [**takayamaTransitionQuiescentActivated2021**, **mauranoRoleDNAMethylation2015**]. 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* [**hirschConsequencesMutantTET22018**, **shihCombinationTargetedThe** **duyRationalTargetingCooperating2019**], *IDH1* and *IDH2* [**shihCombinationTargetedTherapy2017**,

figueroaLeukemicIDH1IDH22010], and *DNMT3A* [**luEpigeneticPerturbationsArg882Mutated2016**, **yangDNMT3ALossDrives2016**], 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 [**mullighanGenomicAnalysisClonal2008**, **maRiseFallSubclones2015**, **dobsonRelapseFatedLatentDiagnos**
Epigenetic studies of B-ALL relapse have primarily relied on enrichment-based assays or methylation arrays that have limited resolution genome-wide [**hoganIntegratedGenomicAnalysis2011**, **nordlundGenomewideSignaturesDifferential2013**, **leeEpigeneticRemodelingBcell2015**]. Further, fewer have investigated the role of chromatin accessibility in B-ALL oncogenesis or relapse [**diedrichProfilingChromatinAccessibility2021**]. 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)) [**dobsonRelapseFatedLatentDiagnosis2020**]. 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.

5.3 Results

5.3.1 Multi-omic 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 5.1a). 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 [**dobsonRelapseFatedLatentDiagnosis2020**] (see Table D.1; patient numbers used in this study match those from [**dobsonRelapseFatedLatentDiagnosis2020**]) 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 similarity network fusion (SNF) [**wangSimilarityNetworkFusion2014**]. 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 5.5.5). 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 5.1b). For all 3 patients with complete molecular datasets, the combined networks clustered samples based on disease stage more clearly than each individual dataset (Figure 5.1b). This suggests that disease stages can be more clearly identified from multiple molecular components together than a single component alone [**wangSimilarityNetworkFusion2014**]. 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.

5.3.2 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 multipotent 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 5.2a; see Table 5.1). Using pairwise comparisons between these cell types, we identified 540 differentially methylated regions (DMRs) over the course of B-cell differentiation from HSCs (Figure 5.2b). 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

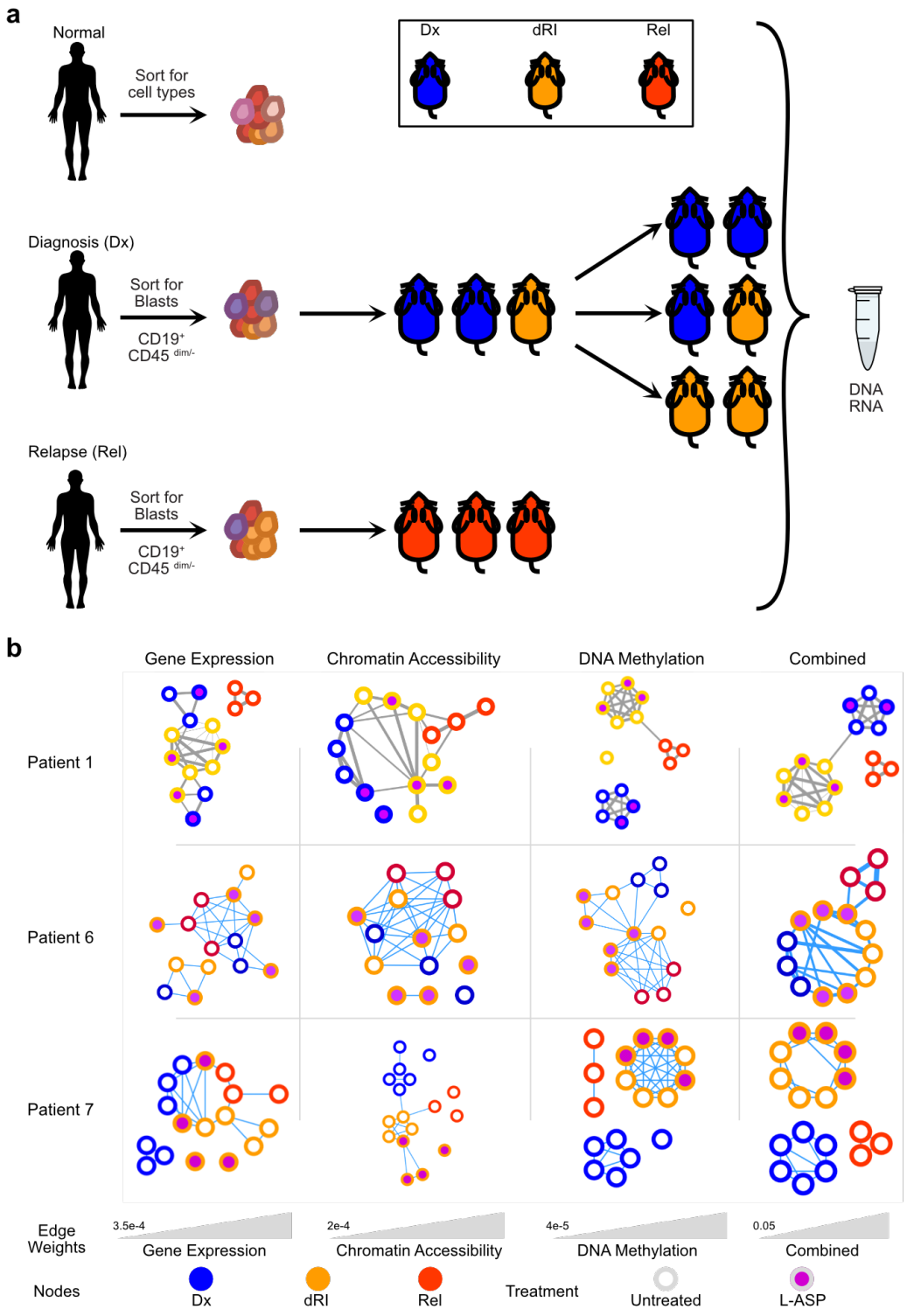

Figure 5.1: **Experimental design and data integration. a.** Experimental design of samples used in this study. Normal samples were obtained from cord blood pools and sorted into various hematopoietic cell types. B-ALL patients who experienced relapse has sorted leukemic blasts collected at Dx and Rel. Based on the mutation profiles from [**dobsonRelapseFatedLatentDiagnosis2020**] 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.

to fully differentiated B-cells (Figure 5.2c). 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 [**leeGlobalDNAMethylation2012**, **leeEpigeneticRemodelingBcell2015**, **nordlundGenomewideSignaturesDifferential2013**], 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.

5.3.3 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 [**leeEpigeneticRemodelingBcell2015**, **nordlundGenomewideSignaturesDifferential2013**] as well as the earlier SNF analysis. We hypothesized that DNAme changes in across patients was heterogeneous, which limited the ability to detect significant changes. Using a patient-oriented approach, we identified DMRs between Dx and Rel for each patient, separately, to track changes over each patient's relapse trajectory. This identified 25 761 DMRs across the cohort of (range 98 - 15 296, median 7 426, $\delta\beta \geq 20$ %, false discovery rate (FDR) < 0.1, Figure 5.3a). Unlike the process of normal differentiation, most DMRs were hypermethylated at relapse (Figure 5.2b). 18 610 (72.2 %) DMRs were specific to a single patient and did not overlap DMRs from others (Figure 5.3b, 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

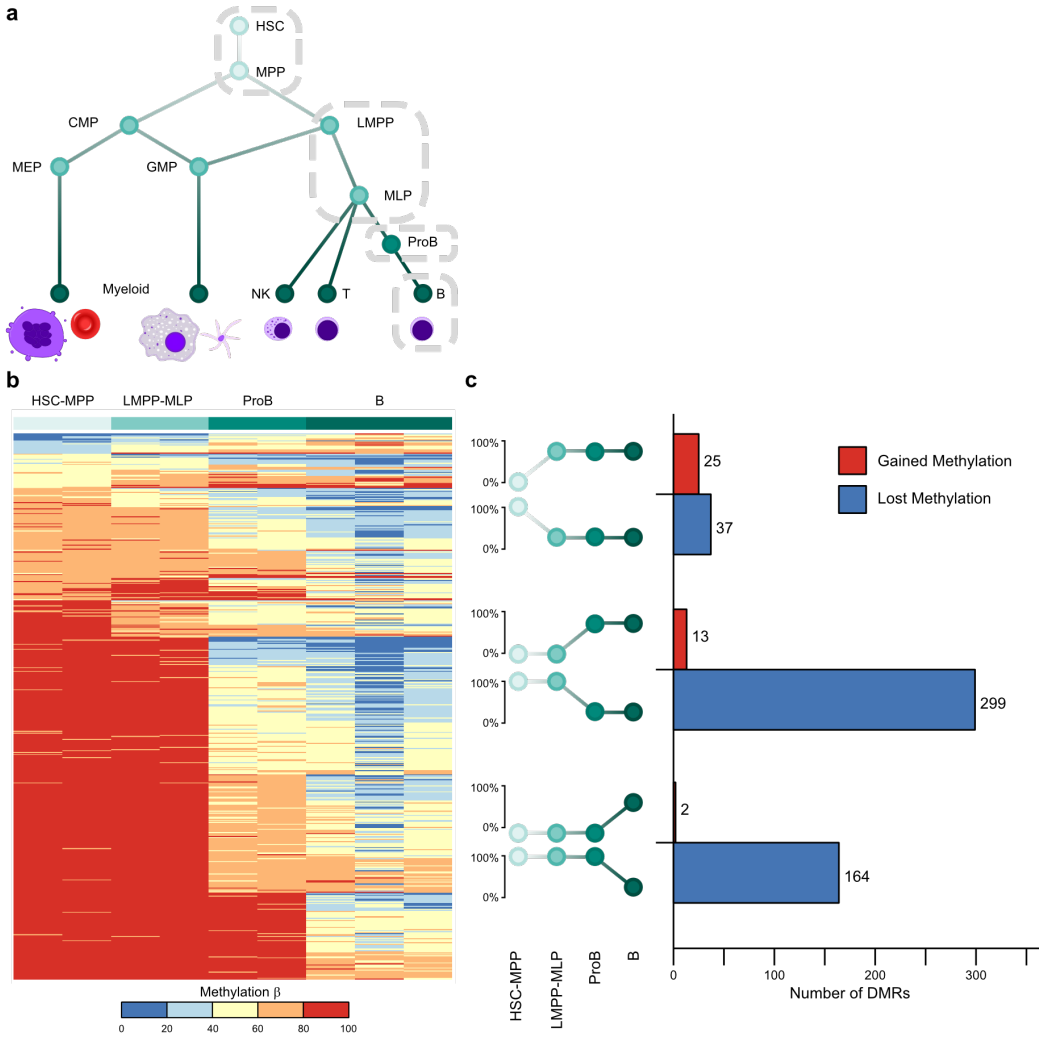

Figure 5.2: **Widespread loss of DNA methylation over B-cell differentiation. a.** Schematic of the hematopoietic hierarchy and the grouping of B-cell progenitors into the groups isolated in this study. **b.** Heatmap of DMRs identified between B lineage cell types. Columns are samples ordered by cell type and rows are DMRs identified in at least one pairwise comparison between cell types (dmrseq [**korthauerDetectionAccurateFalse2018**], FDR < 0.1). **c.** Bar plot of DMRs classified by which step in differentiation they were identified as significantly changed.

fate determination (Figure 5.3c). For these genes, we find that the promoter regions become hypermethylated at B-ALL relapse (Figure 5.3d). 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 [**jonesFunctionsDNAMethylation2012**], 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.

5.3.4 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 [**dobsonRelapseFatedLatentDiagnosis2020**]. 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 5.3a). 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 5.4). There is heterogeneity in DNAme profiles across samples derived from the same patient, and even within the same disease stage (Figure 5.4a). The heterogeneity within disease stages resulted identifying DMRs specific to the dRI samples that were shared between Dx and Rel samples, even when some dRI samples showed similar methylation rates (e.g. leftmost dRI sample for Patient 4). Based on which disease stage the DMRs were identified in, each glsdmr was classified as Dx-specific (shared between dRI and Rel), dRI-specific (shared between Dx and Rel), Rel-specific (shared between Dx and dRI), or unique (significantly differentially methylated in all three stages). All patients harboured Rel-specific DMRs, and a majority of DMRs in total were detected in Patients 1, 4, and 6 (range 71.6 %, 100 %, and 100 %, respectively; Figure 5.4b). 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 5.4b). Patient 1 also harboured 356 (10.6 %) Dx-specific DMRs. This

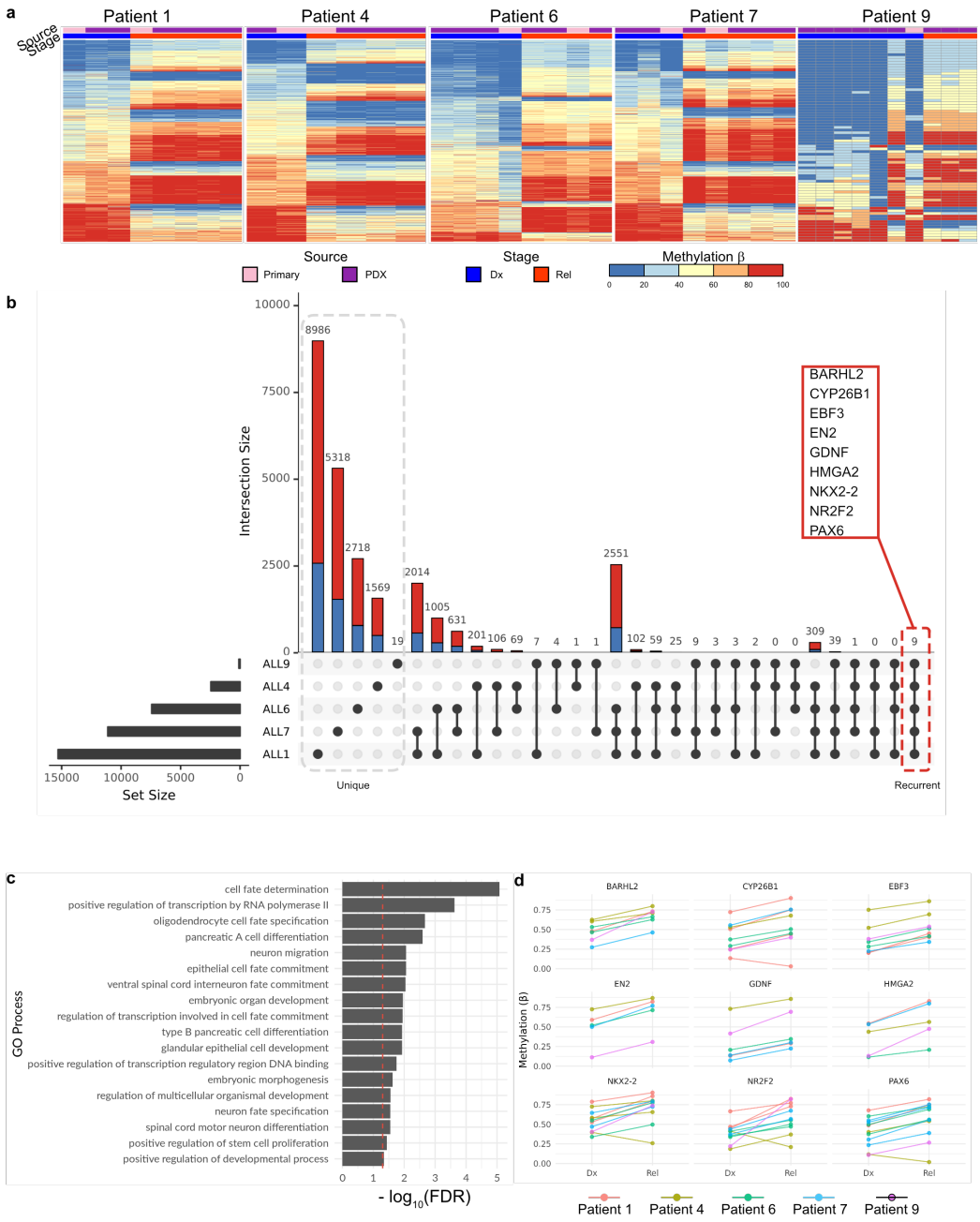

Figure 5.3: **Recurrent relapse DMRs are associated with cell fate decision processes. a.** Heatmaps of DMRs identified between Dx and Rel samples within each patient. **b.** Upset plot showing the shared DMRs between patients. DMRs in the left highlighted block are unique to a single patient, whereas 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.

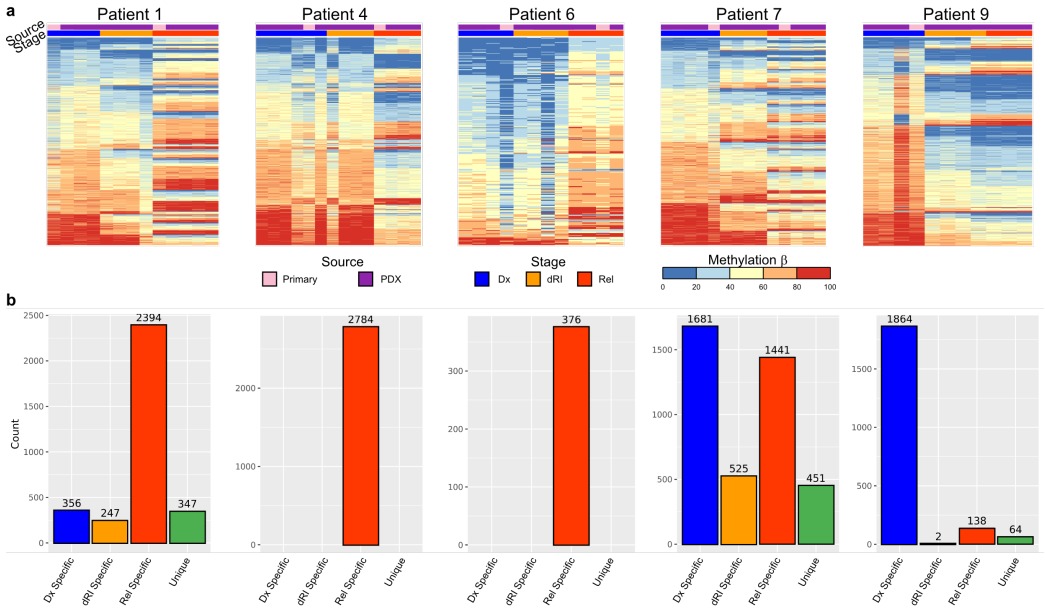

Figure 5.4: **Subpopulations present at diagnosis can harbour relapse-like DNAme profiles. a.** Heatmaps of expanded DMRs identified in dRI PDX samples. **b.** Bar plot showing the number of DMRs classified by which disease stage it is specific to. "Dx Specific" DMRs have shared DNAme between dRI and 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.

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.

5.4 Discussion

Disease relapse remains a major barrier in treating B-ALL [**formanMythSecondRemission2013**, **liewOutcomesAdultPatients2012**, **hungerAcuteLymphoblasticLeukemia2015**]. 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 [**nordlundGenomewideSignaturesDifferential2013**, **leeEpigeneticRemodelingBcell2015**]. Previous reports highlight the frequency that epigenetic regulators are mutated in B-ALL [**maRiseFallSubclones201** and leukemias more generally, such as *DNMT3A*, *TET2*, *IDH1*, and *IDH2* in acute myeloid leukemia (AML) [**kishtagariDriverMutationsAcute2020**, **papaemmanuilGenomicClassificationPrognosis2016**, **leyDNMT3AMutationsAcute2010**] and *CHD2*, *HIST1H1E*, and *ZMYM3* in chronic lymphocytic leukemia (CLL) [**billotDeregulationAiolosExpression2011**, **landauChronicLymphocyticLeukemia201** **landauEvolutionImpactSubclonal2013**]. 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 [**russoAdaptiveMutabilityColorectal2019**]. 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.

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 [**mackEpigenomicAlterationsDefine2014**, **pajtlerMolecularClassificationEpendymal2015**], and glioblastoma, where stem cells are characterized by widespread changes in chromatin accessibility [**guilhamonSinglecellChromatinAccessibility2021**] and histone modifications [**liauAdaptiveChromatinRemo** These studies highlight the role of epigenetic plasticity and intra-tumour heterogeneity in cancers

[**flavahanEpigeneticPlasticityHallmarks2017**]. With similar results found in leukemias that are linked to disease outcome [**pastoreCorruptedCoordinationEpigenetic2019**, **landauLocallyDisorderedMethy** **gaitiEpigeneticEvolutionLineage2019**, **namIntegratingGeneticNongenetic2021**, **liDistinctEvolutionDyn** 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 [**nordlundGenomewideSignaturesDifferential2013** **leeEpigeneticRemodelingBcell2015**, **garcia-maneroDNAMethylationMultiple2002**, **garcia-maneroAber** While the recurrent DMRs in this study were found in these same regions, most DMRs were identified in intergenic regions. This suggests that important changes in the epigenetic landscape is currently unidentified, and future studies investigating DNAme aberrations in B-ALL should prioritize genome-wide approaches. The phenotypic impact of focal hypermethylation on engraftment and self-renewal capacity has not been assessed here, so experiments should be conducted to validate these findings (this is a bad sentence but this idea is important). For patients undergoing B-ALL treatment, DNAme has the potential to be used as early indicators of relapse. Moreover, treatment with DNA demethylating agents, such as 5-aza-citidine and 5-aza-2′-deoxycytadine, may be effective at preventing relapse. These treatments have been approved for use in patients with myelodisplastic syndrome (MDS) and AML in adult populations and early clinical trials have demonstrated their safety [**bentonSafetyClinicalActivity2014**, **nationalcancerinstitutenciGroupwidePilotStudy2021**], although some toxic effects have been identified in drug combination trials [**therapeuticadvancesinchildhoodleuke** 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.

5.5 Methods

5.5.1 Patient selection and sample collection

Patient samples were obtained at diagnosis and relapse from patients with B-ALL as previously described [**dobsonRelapseFatedLatentDiagnosis2020**]. 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 [**dobsonRelapseFatedLatentDiagnosis2020**]. 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 5.1, as previously described [**nottaIsolationSingleHuman2011**], on a FACSAria III (Becton Dickinson), consistently yielding > 95 % purity.

Table 5.1: **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-

5.5.2 Patient-derived xenograft generation and limiting dilution assays

PDXs were generated as previously described [**dobsonRelapseFatedLatentDiagnosis2020**]. 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 h prior to transplants. Only female mice were used. Intra-femoral injections of 10 to 250 000 sorted leukemic blasts were performed as previously described [**mazurierRapidMyeloerythroidRepopulation2003**]. 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 [**huELDAExtremeLimiting2009**].

5.5.3 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).

5.5.4 Primary and patient-derived xenograft sample sequencing

RNA sequencing

RNA-seq was performed as previously described [**dobsonRelapseFatedLatentDiagnosis2020**]. 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 [**buenrostroTranspositionNativeChromatin**]. ATAC-seq libraries were sequenced with an Illumina HiSeq 2500 sequencer to generate single-end 50 bp reads.

5.5.5 Sequencing data analysis

Differential gene expression analysis

The methods are described in [**dobsonRelapseFatedLatentDiagnosis2020**]. Briefly, RNA-seq reads were aligned against the GRCh38 reference human genome with STAR (v2.5.2b) [**dobinSTARUltrafastUn**] 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) [**andersHTSeqPythonFramework2015**]. Data was library size normalized using the RLE method, followed by a variance stabilizing transformation using DESeq2 (v1.22.1) [**loveModeratedEstimationFold2014**]. 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) [**langmeadFastGappedreadAlignment2012**] with default parameters. Accessible peaks were identified with MACS2 (v2.0.10) [**zhangModelbasedAnalysisChIPSeq2008**] 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 [**quinlanBEDToolsSwissArmyTool2014**] for downstream analysis.

Bisulfite sequencing pre-processing

Sequencing read qualities were assessed with FastQC (v0.11.8) [**andrewsFastQCQualityControl2010**]. Low quality bases were trimmed with Trim Galore! (v0.6.3) [**kruegerTrimGalore2012**] 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) [**kruegerDNAMethylomeAnalysis2012**] with default parameters. Duplicates were removed from the resulting alignment file with the following command:

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

The deduplicated binary alignment map (BAM) file was sorted by position with sambamba (v0.7.0) [**tarasovSambambaFastProcessing2015**]. M-biases were calculated with MethylDackel (v0.4.0) [**ryanMethylDackel2019**], 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 [**wangSimilarityNetworkFusion2014**] 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) [**korthauerDetectionAccurateFalse2018**] 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 [**benjaminiControllingFalseDiscovery1995**]. 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) [**miLargescaleGeneFunction2013**]. 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.

Chapter 6

Discussion & Future Directions

Each of the previous chapters have presented a story interrogating multiple components of the chromatin architecture, how they interact with each other, and the plethora of computational and experimental methods required to unravel this architecture. Chapter 2 identifies and validates *cis*-regulatory elements (CREs) of the *FOXA1* gene, a critical transcription factor (TF) that regulates prostate cancer (PCa) development and regulates androgen receptor (*AR*) expression to control disease progression. Chapter 3 expands on these ideas to investigate how the three-dimensional genome organization impacts gene regulatory networks and how genetic aberrations can alter this organization to promote oncogenesis. Chapter 4 develops a mathematical and computational framework to reduce uncertainty about how individual aberrations in chromatin architecture impact gene expression. Finally, Chapter 5 identifies the strong relationship between genetic and epigenetic profiles in B-cell acute lymphoblastic leukemia (B-ALL) relapse and investigates how DNA methylation (DNAme) changes and revision to a more stem-like chromatin state may underlie disease relapse. Together, the work presented in this thesis demonstrates that different components of the chromatin architecture, the genome, molecular chromatin modifications, and three-dimensional organization, can all individually contribute to cancer development and progression. Moreover, this thesis demonstrates that aberrations in these components work together to drive disease. These multiple components of the chromatin architecture need to be studied in tandem to understand the origins of cancer and how to develop curative treatments for it.

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

In Chapter 2, I used gene essentiality screening data from multiple cell lines to prioritize the *FOXA1* TF as a critical factor across PCa cell lines. I also made use of the concept that single nucleotide variants (SNVs) converge on CREs of important genes in a given tumour type to predict how these mutations may impact candidate CREs for the *FOXA1* gene. *FOXA1* is also an important TF in breast cancers Figure A.3. Similar investigations into the impact of SNVs in breast tumours may identify the impact of aberrations to the CREs of *FOXA1*. Identifying important genes in this manner is not limited to *FOXA1* and breast and prostate tumours. Critical genes may be identified in other cancer types using clustered regularly interspaced short palindromic repeat (CRISPR) screens or massively-parallel reporter assays (MPRAs). Similarly SNVs are not the only chromatin aberrations that can affect TF binding or gene regulation. Other chromatin aberrations may accumulate in CREs of important genes in a similar fashion. Complex structural variants (SVs), changes in DNAme, or histone modifications may only need to accumulate in the set of CREs for a given gene, rather than be recurrent in a single element, to affect its expression. Interpreting chromatin aberrations in cancer in light of this plexus-based approach may aid in identifying driver events for cancer by aggregating previously unrelated events together. These approaches are not limited to prostate tumours and can serve as a starting point to identify important genes in other cancers, more generally.

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

In Chapter 3, my co-authors optimized a low-input Hi-C method to interrogate genome organization in cryo-preserved prostate tissue slides. I then demonstrated that this could produce a high quality Hi-C library and helped produce that largest collection of genome organization data in prostate tumours to date. This technological step forward opens the door for profiling the three-dimensional genome in cancer patients without relying in cell lines or other models, and may be a critical step in moving personalized medicine forward. We add to existing evidence that SVs can, but rarely, alter 3D structure in disease [**ghavi-helmHighlyRearrangedChromosomes2019**, **oudelaarRelationshipGenomeStructure2020**, **despangFunctionalDissectionSox92019**, **williamsonDevelo**

dixonIntegrativeDetectionAnalysis2018, **akdemirDisruptionChromatinFolding2020**, **liPatternsSomatic iyyankiSubtypeassociatedEpigenomicLandscape2021**]. Elucidating when and how SVs impact genome organization, then, is still an area that requires investigation. Developments in statistical methods, such as those discussed in Chapter 4, may help identify the effects of individual, non-recurrent SVs. Subclonality of SVs may interfere with the ability to detect rearranged domains in bulk Hi-C measurements. Thus, developments in high throughput sequencing and microscopy measurements in single cells, such as ORCA [**mateoVisualizingDNAFolding2019**] and STORM [**batesStochasticOpticalReconstruction2013**], as well as organoid or explant models that recapitulate the chromatin state of the original tumour, may help in identifying the effect of such events [**zanoniModelingNeoplasticDisease2020**]. This work also adds to our ability to detect chromatin interactions between promoters and enhancers in patient samples, allowing for better characterization of gene regulatory networks for each and every gene [**gasperiniComprehensiveCatalogueValidated2020**, **oudelaarRelationshipGenomeStructure2020**, **wangEngineering3DGenome2021**]. Given the benefits of plexus-based approaches to interpreting aberrations in the chromatin architecture, this work serves as a foundation on which to integrate gene regulatory networks with chromatin aberrations in cancers more generally. This foundation can be extended to studying the evolution of these networks, their genome organization, and their resiliency between species or over time as tumours respond to therapeutic interventions.

6.3 Implications of DNA methylation changes in relapse

Chapter 5 identifies DNAme as an epigenetic marker that can mirror that mutational profile of B-ALL cells. The DNAme changes observed over the course of B-ALL relapse have the potential to become a biomarker predicting relapse, although variation in DNAme changes across patients necessitates larger sample sizes before recurrent events may be robustly identified. Our patient-oriented approach to identify recurrent changes to DNAme boosted our discovery of recurrent differentially methylated regions (DMRs) over a cohort-oriented approach, and thus may be a beneficial strategy for similar studies. Inter-tumour heterogeneity in tumours is a well-studied phenomenon [**marusykTumorHeterogeneityCauses2010**, **sottorivaCatchMyDrift2017**, **huangGeneticNongeneticInsta mcgranahanBiologicalTherapeuticImpact2015**, **landauChronicLymphocyticLeukemia2013**, **ben-davidGeneticTranscriptionalEvolution2018**, **carterEpigeneticBasisCellular2021**], so patient-oriented discovery approaches may be advantageous when assessing molecular trajectories of relapse and therapeutic response, as well. Changes in DNAme can be detected through blood draws

[**peterDynamicsCellfreeDNA2020**, **shenSensitiveTumourDetection2018**, **nassiriDetectionDiscriminatio**
and offers the potential for a non-invasive biomarker to predict relapse in B-ALL patients. Given the widespread hypermethylation observed at relapse in all patients, it is possible that B-ALL patients undergoing continuation/maintenance therapy may additionally benefit from demethylating agents such as 5-aza-cytidine and 5-aza-2'-deoxycytidine to prevent relapse. Finally, the widespread hypermethylation of B-ALL at relapse reflects reprogramming of malignant blasts to a more stem-like phenotype, a characteristic typically observed in other leukemias such as acute myeloid leukemia (AML) [**krivtsovMLLTranslocationsHistone2007**, **liClinicalImplicationsGenomewide2017**, **kresoEvolutionCanc**
shlushIdentificationPreleukaemicHaematopoietic2014, **shlushTracingOriginsRelapse2017**]. This suggests that the role of leukemic stem cells and the impact that the cell-of-origin has on therapeutic response, should be prioritized in future research of B-ALL relapse.

6.4 Implications for functional genomics and cancer patients

The studies presented here highlight the importance of non-genetic properties of chromatin and their role in cancer development and progression. I used multiple sequencing-based assays to identify chromatin elements and structures, such as CREs and topologically associated domains (TADs), associate these elements and structures with potential target genes, and measure how these elements and structures affect gene expression when they are perturbed. This has multiple implications for how functional genomics and cancer research can be performed. Firstly, considering the entire regulatory plexus of genes may help pinpoint drivers of oncogenesis or cancer progression by offering alternative methods to identify candidate drivers. Similarly, viewing CRE hijacking as a common mechanism of aberrant gene expression may simultaneously increase the number of potential driver events found and offer a biological mechanism of action. These two perspectives may help resolve previously identified risk SNVs with no known mechanism of action. Considering the entire regulatory plexus as a functional unit of chromatin can also have therapeutic implications. Genes important for cancer progression, such as *FOXA1* and *AR* in PCa, have multiple regulatory elements. Each element is a potential option for targeted therapies, and combination therapies targeting multiple elements may offer more robust and controllable options than targeting the gene alone or an individual element. Developing targeted therapies for the multiple validated CREs discovered in Chapter 2 may help control or prevent castration resistance in PCa. Regarding the regulatory plexus as a functional unit may also help predict the evolutionary dynamics that tumours undergo during treatment. Therapeutic resistance via tumour evolution remains a clinical concern,

and much study of evolutionary dynamics over cancer progression has focused on identifying genetic subclonal populations of tumour initiating cells [**pogrebniakHarnessingTumorEvolution2018**, **maleyClassifyingEvolutionaryEcological2017**] The evolution of gene regulatory networks have long been studied in model organisms such as *Arabidopsis thaliana* [**nowickLineagespecificTranscriptionFactors2** and *Drosophila melanogaster* [**peterEvolutionGeneRegulatory2011**]. Bridging these fields of research together by including CREs, epigenetic marks, and genome organization as important players in the fitness and evolution of tumours may help inform how to prevent or counteract therapeutic resistance.

Secondly, I have shown that valuable information about tumour regulatory networks cannot be captured by sequencing DNA alone. Assaying more than solely the DNA from cancer patients is required to understand the origin and potential trajectories of their tumours. Studying these regulatory networks in cancer patients requires robust characterization using multiple genome-wide assays and validation experiments. Given the differences observed between cell line models and primary patient samples observed in Chapter 3, as well as the inter-patient heterogeneity observed in Chapter 5, it is vital that these networks are constructed on a per-patient basis. Currently, chromatin-based assays are not prioritized for clinical samples in the same manner as genome-based assays are. Further, profiling multiple facets of the chromatin architecture from primary patient samples requires technological development in genome-wide assays. Optimizing for small amounts of chromatin obtained from samples that are fresh or stored as flash-frozen paraffin embedded or cryo-preserved material are clear next steps to more comprehensively characterize the chromatin of tumours. In Chapter 3, my colleagues and I pushed technological limits in this area in the case of the Hi-C assay. If more comprehensive characterization of per-patient chromatin architecture is pursued, adopting patient-specific discovery strategies such as those presented in Chapters 4 and 5 may help guide personalized therapeutic strategies. Alternatively, developing patient-derived models that recapitulate the entire chromatin architecture, as mentioned previously, can help test and validate potential therapies for each patient, such as targeted DNAme experiments proposed in Chapter 5.

6.5 Limitations

The main limitations in the studies presented here stem from cancer models, cohort sizes, and correlative measurements. In Chapter 2, we used essentiality data from PCa cell lines that serve as models of primary prostate tumours. Further, our dissection of CREs relied on encyclopedias of DNA elements derived from cell lines, as did our validation experiments for assessing the impact

of mutations seen in primary prostate tumours. Cell line models are not necessarily representative of primary tumours observed in patients [**domckeCompetitionDNAMethylation2015**, **ghandiNextgenerationCharacterizationCancer2019**]. Moreover, different cell lines may have different regulatory networks, as recent studies of chromatin organization and CRE contacts in PCa cell lines have highlighted [**ahmedCRISPRiScreensReveal2021**]. This separation between the models we study and the patients we aim to serve may prohibit translation of discoveries from our models to the clinic. In Chapter 3, we were able to profile the genome organization of 12 PCa patients and 5 benign prostate samples. Similarly, in Chapter 5, we used patient-derived xenografts (PDXs) originating from 5 patients with B-ALL. In light of the large degree of inter-patient variation seen in tumours from the same tissue, and the intra-patient variation seen in PDXs (Chapter 5), observations in these small sample sizes may not generalize to patients with these diseases as a whole. In both cases, we used patient-matched data to corroborate findings from other components of the chromatin architecture, where possible, but validation experiments for some observations remain difficult. Using CRISPR technologies to knock in SNVs or inactivate individual CREs is possible, but replicating complex SVs or selectively (de)methylating CREs to validate some findings remains technically challenging [**nakamuraCRISPRTechnologiesPrecise2021**, **pickar-oliverNextGenerationCRISPR2019**, **wangEngineering3DGenome2021**]. Finally, despite the small sample sizes in some of these studies, the amount of high throughput sequencing data generated was larger than many computational methods could handle efficiently, particularly with Hi-C data. Some methods used here had to be adapted from their original publications, and other methods that could have been used would have required extensive re-engineering to function properly, limiting some analyses.

6.6 Summary and concluding remarks

Diagnosis, treatment, and the foundational understanding of cancer has been revolutionized by high throughput sequencing technologies and the ability to detect and interpret genetic aberrations in tumours. When combined with information about CREs, the essentiality of genes in multiple cell types, and the three-dimensional genome organization, genetic aberrations can provide a significant amount of explanatory power for the hallmarks of cancer and what causes them. But genetic aberrations are not the only mechanism cancer cells can use to arrive at these hallmarks to drive disease. Identifying aberrations across multiple components of the chromatin architecture can uncover complex mechanisms through which cancer cells turn off the expression of tumour suppressor genes, activate oncogenes, or activate pathways. Interrogating the role of mutations in a gene's set of CREs

(Chapter 2), identifying stable regions of genome organization and shifting chromatin interactions over oncogenesis (Chapter 3), and comparing genetic and epigenetic states over disease progression (Chapter 5) are all important approaches to better understand the aberrations, origins, and trajectories of cancers in patients. This multi-pronged approach requires computational, statistical, and molecular, methods, optimized for low-input samples and small cohort sizes (Chapters 3 and 4), and methodological developments in these areas are still required. This thesis focuses on PCa and B-ALL, but the methodologies employed here are not restricted to a single disease. In summary, measuring multiple facets of the chromatin architecture directly from patients and viewing aberrations in light of the regulatory networks that this architecture describes, we can better understand how cancer arises and develop better, more targeted therapies for patients.

Appendix A

Supplementary Material for Chapter 2

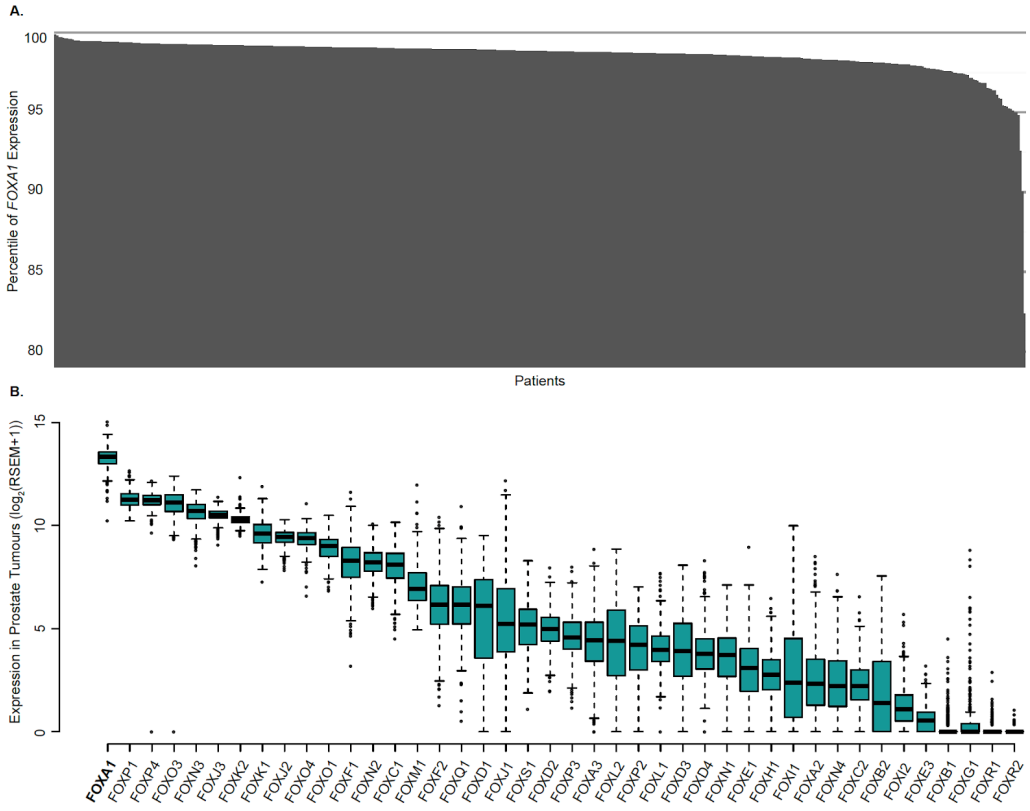

Figure A.1: ***FOXA1* mRNA expression in prostate tumours. a.** The ranking of *FOXA1* mRNA expression across 497 primary prostate tumours profiled in TCGA. **b.** mRNA expression of all genes coding for FOX TFs across 497 primary prostate tumours profiled in TCGA.

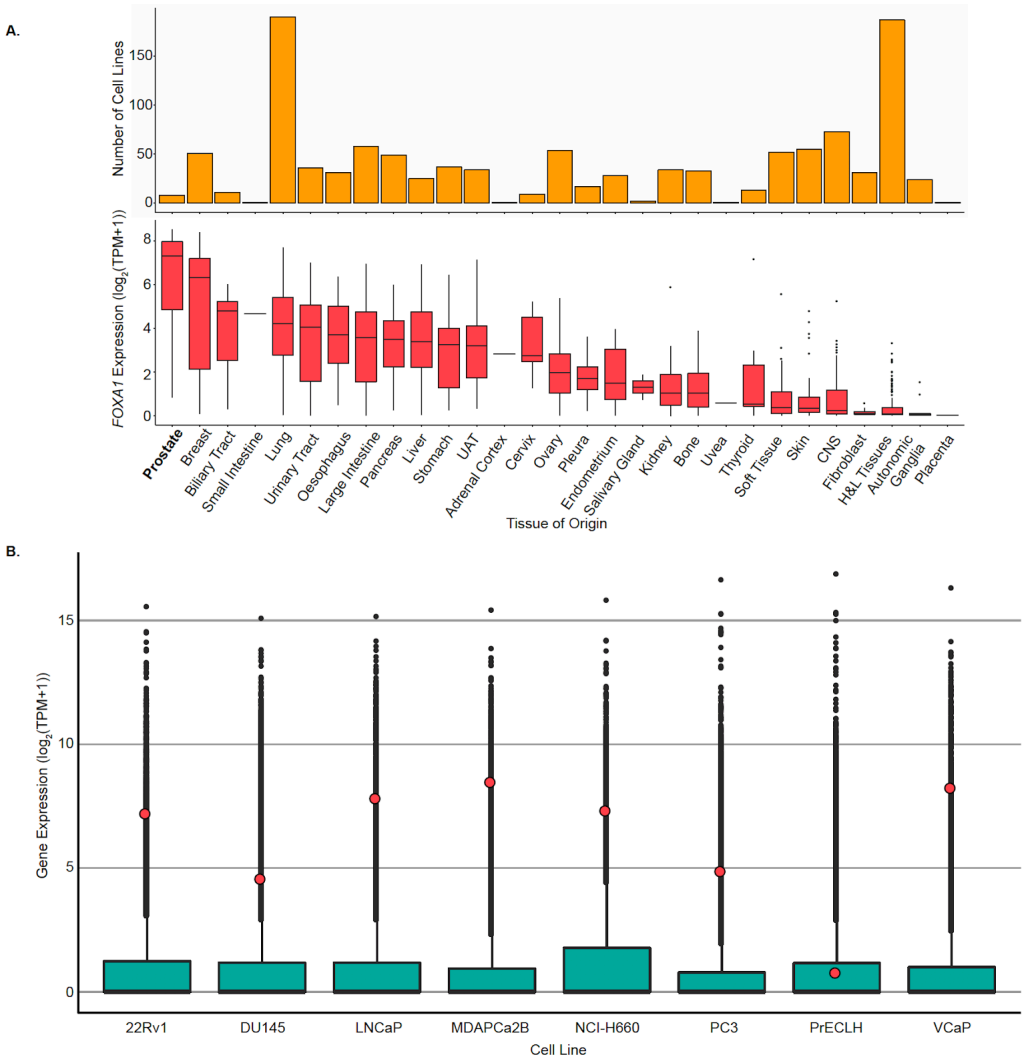

Figure A.2: ***FOXA1* mRNA expression across PCa cell lines. a.** *FOXA1* mRNA expression across all cancer cell lines from DEPMAP, profiled by RNA-seq (see Section 2.5). UAT = Upper Aerodigestive Tract, CNS = Central Nervous System, H&L Tissues = Hematopoietic and Lymphoid Tissues. **b.** *FOXA1* mRNA expression across eight PCa cell lines from DEPMAP, profiled by RNA-seq (see Section 2.5). Red dots indicate *FOXA1*.

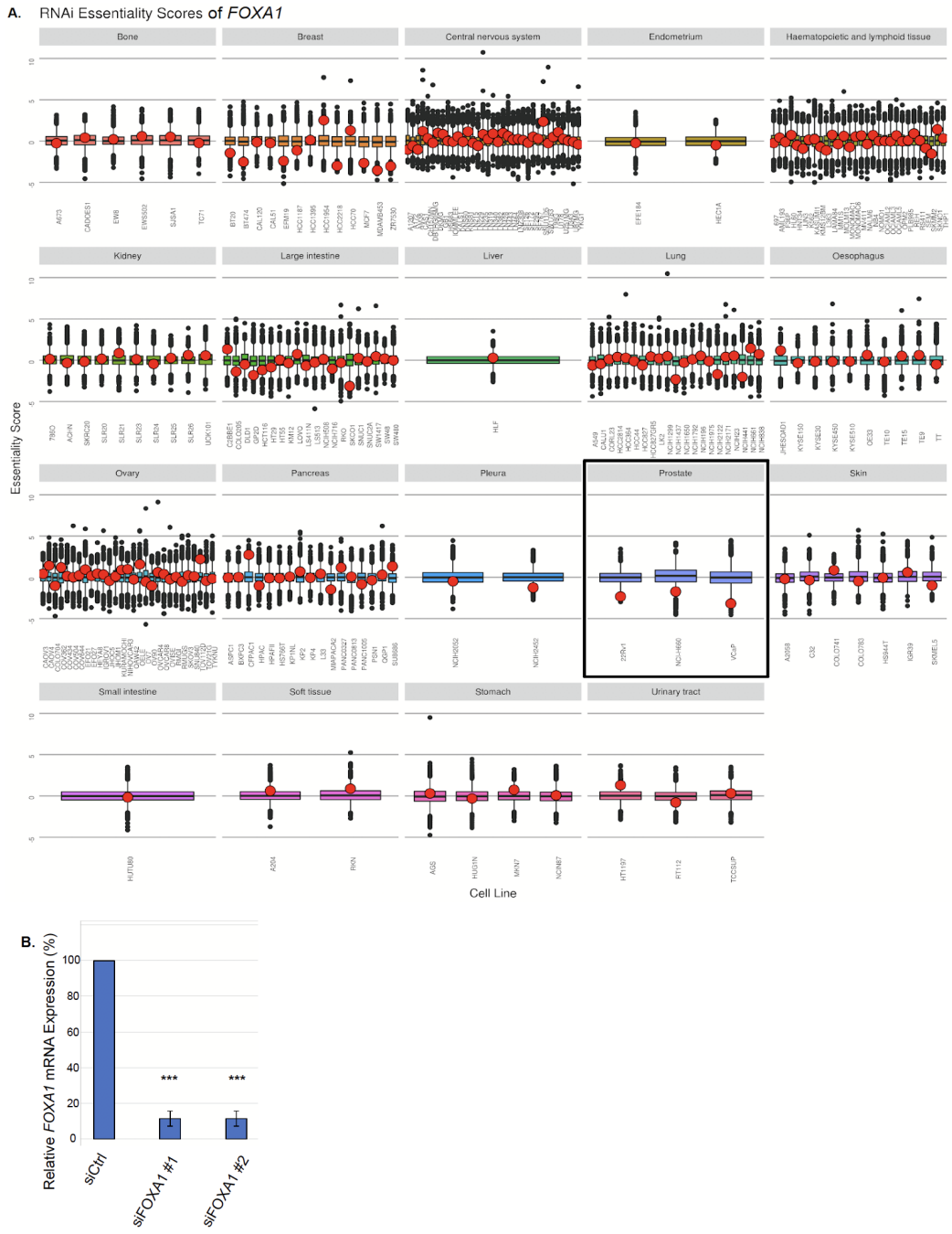

Figure A.3: **Essentiality of *FOXA1* across cancer cell lines of various cancer types. a.** Gene essentiality screen mediated through shRNA/mRNA across various cancer cell lines ($n = 707$). Higher score indicates less essential, and lower score indicates more essential for cell proliferation. Red dot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

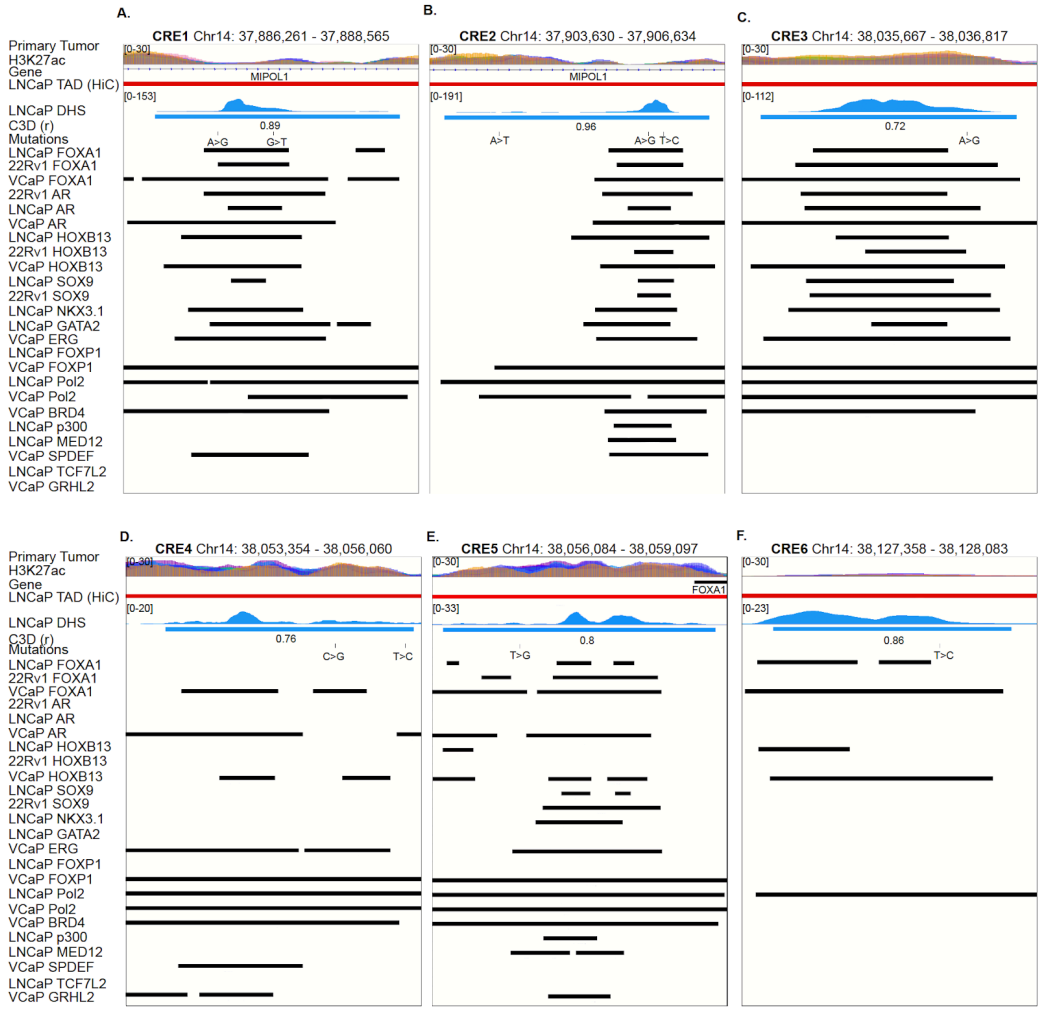

Figure A.4: **Visualization of the functional annotation of the six *FOXA1* CREs. a-f.** Visualization of Functional annotation of the six FOXA1 CREs using public and in-house ChIP-seq datasets.

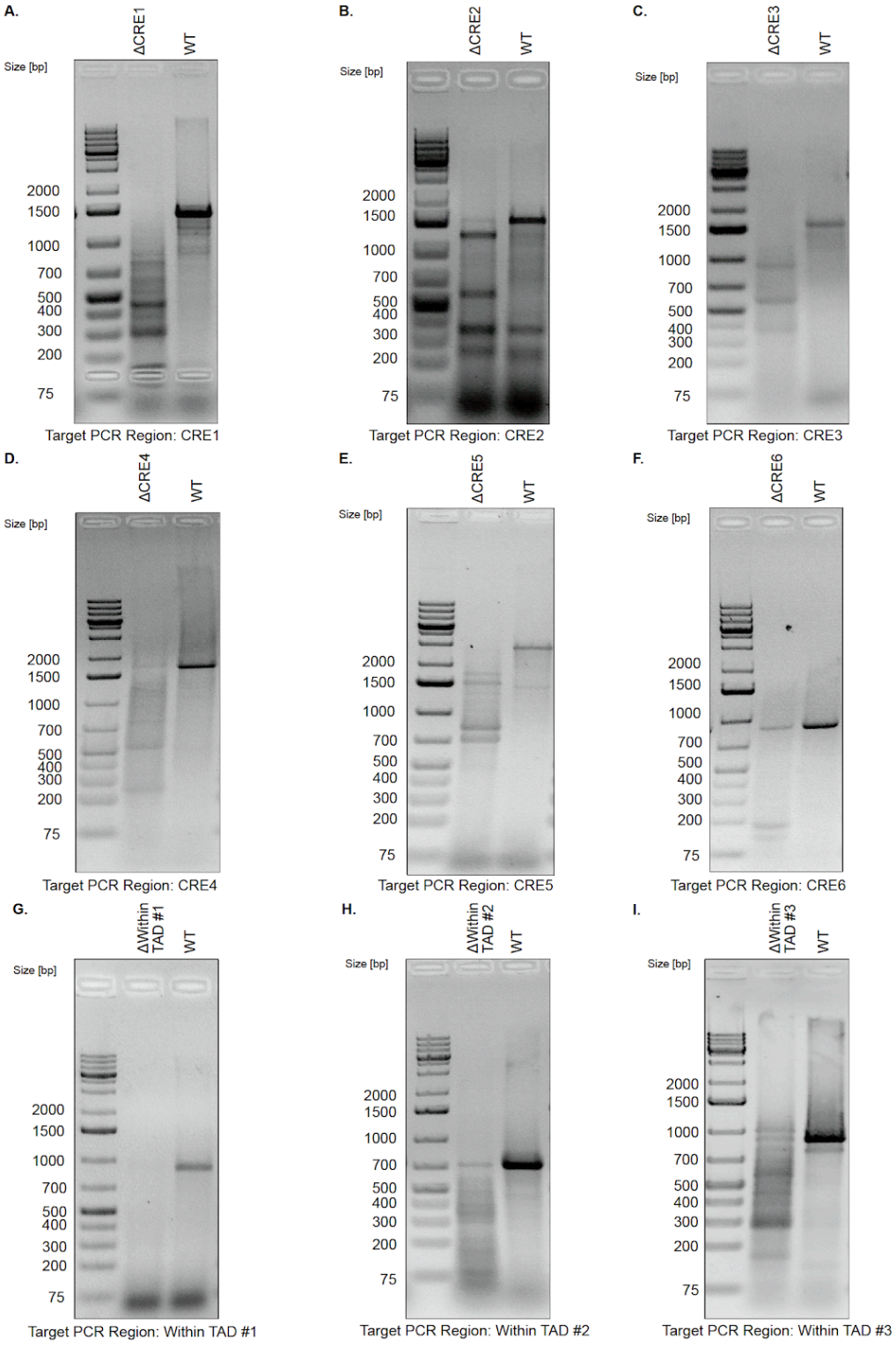

Figure A.5: **Validation of clonal Cas-mediated deletions of CREs. a-f**. Representative agarose gels from LNCaP clonal CRISPR/Cas9-mediated deletion products or WT product from PCR amplification of intended CRE, followed by T7 Endonuclease I assay.

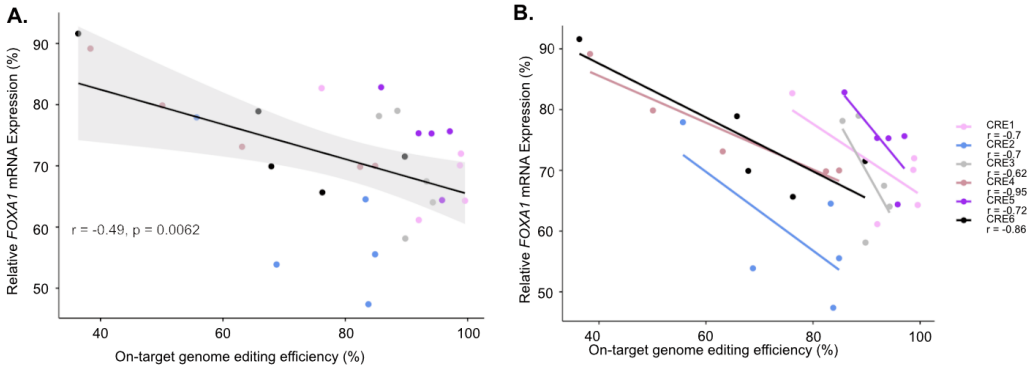

Figure A.6: **Genome editing efficiency is inversely correlated with *FOXA1* mRNA expression. a.** Pearson's correlation to investigate the relationship between genome editing efficiency mediated by CRISPR/Cas9 and *FOXA1* mRNA expression in LNCaP cells. The Pearson's correlation here is across all of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

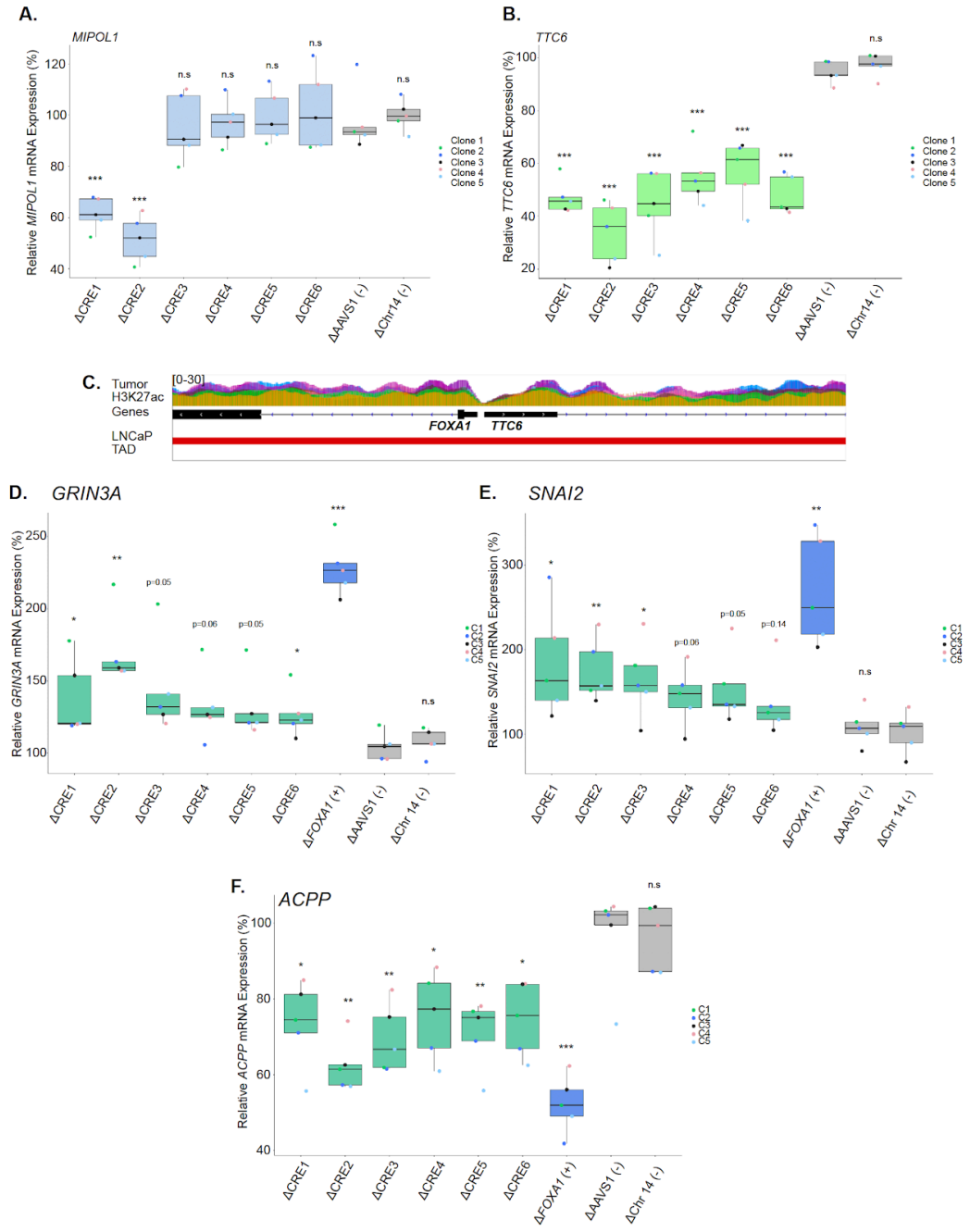

Figure A.7: **Intra-TAD genes and *FOXA1* downstream genes are significantly changed upon deletion of CREs. a.** *MIPOL1* mRNA expression normalized to housekeeping gene *TBP* upon deletion of each region of interest. **b.** *TTC6* mRNA expression normalized to housekeeping gene *TBP* upon deletion of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

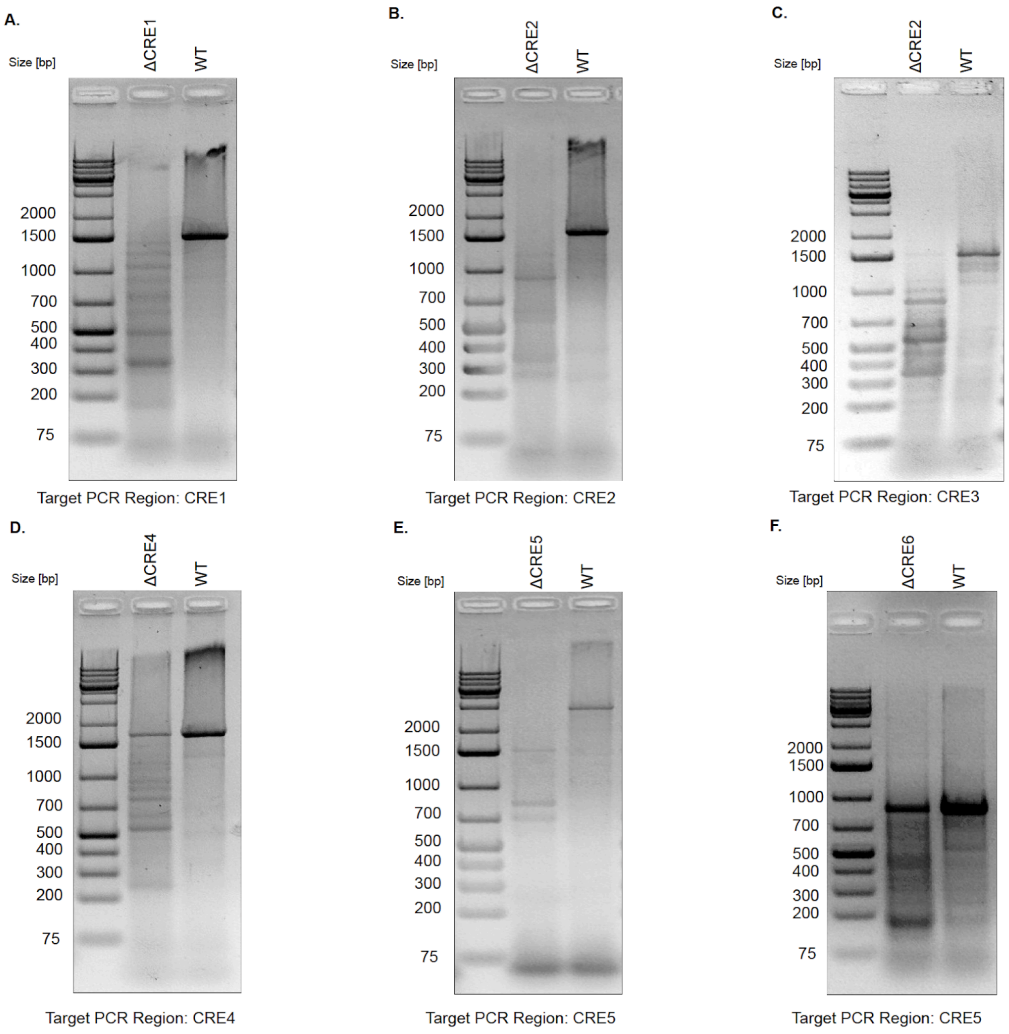

Figure A.8: **Validation of transient Cas9-mediated single deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CRE followed by T7 Endonuclease I assay.

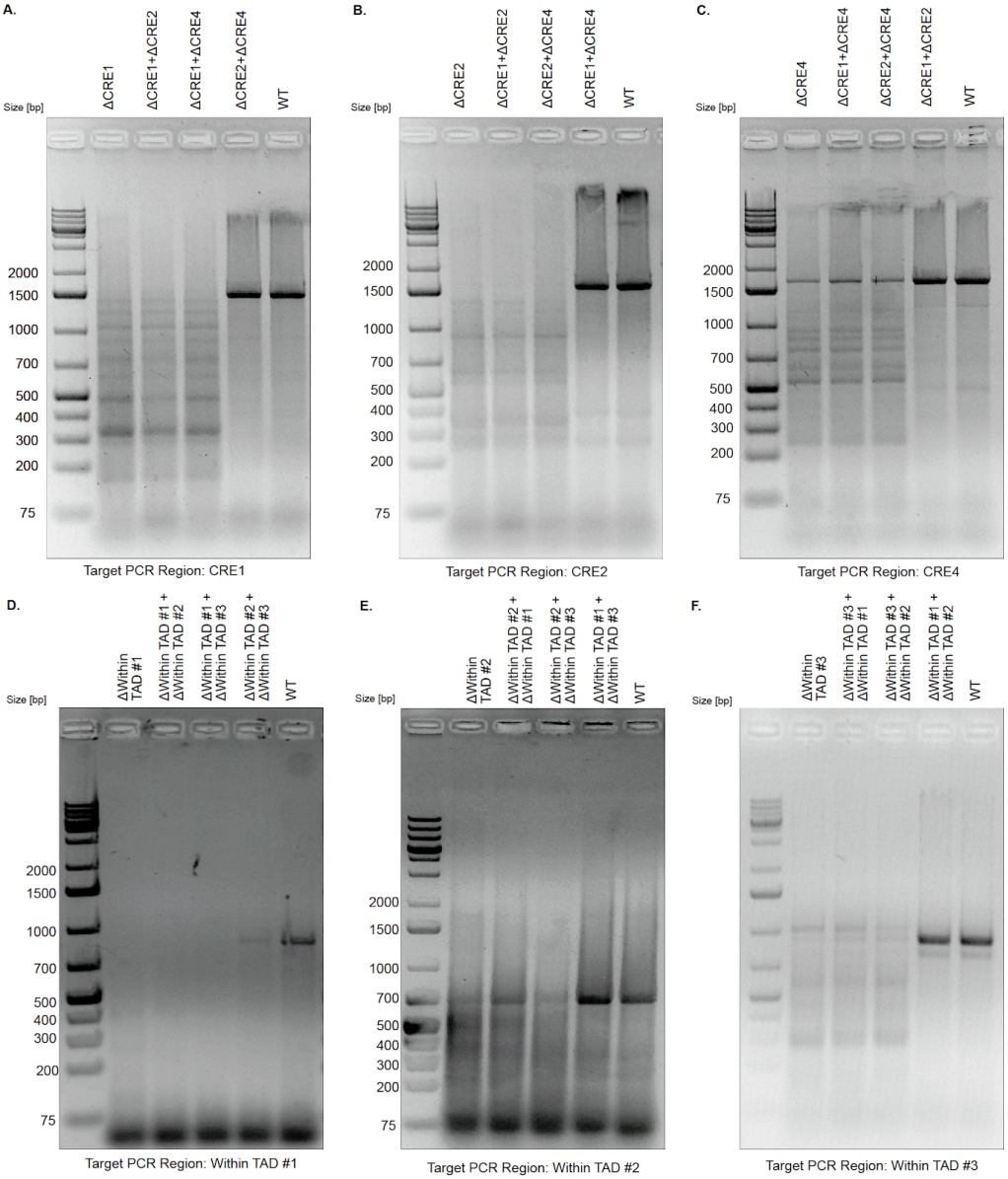

Figure A.9: **Validation of transient Cas9-mediated double deletion of CREs. a-f.** Agarose gel of transient transfection RNP-based Cas9-mediated deletion product from PCR amplification of intended CREs followed by T7 Endonuclease I assay.

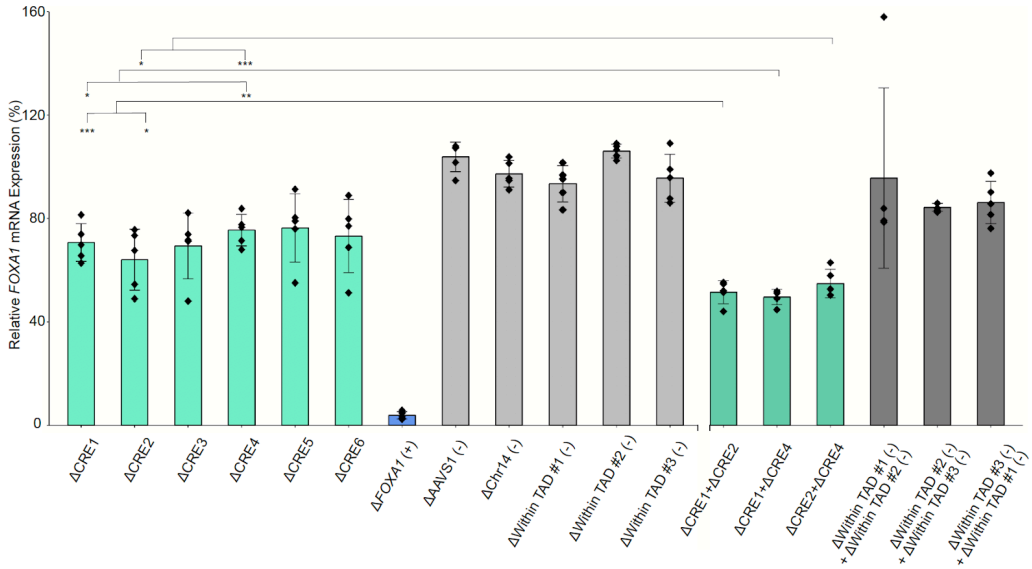

Figure A.10: **Comparison of *FOXA1* mRNA expression upon double versus single deletion of CRE(s).** *FOXA1* mRNA expression normalized to housekeeping gene *TBP* upon single or double deletion of target CREs. Δ indicates CRISPR/Cas9-mediated deletion ($n = 5$ independent experiments). Error bars indicate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

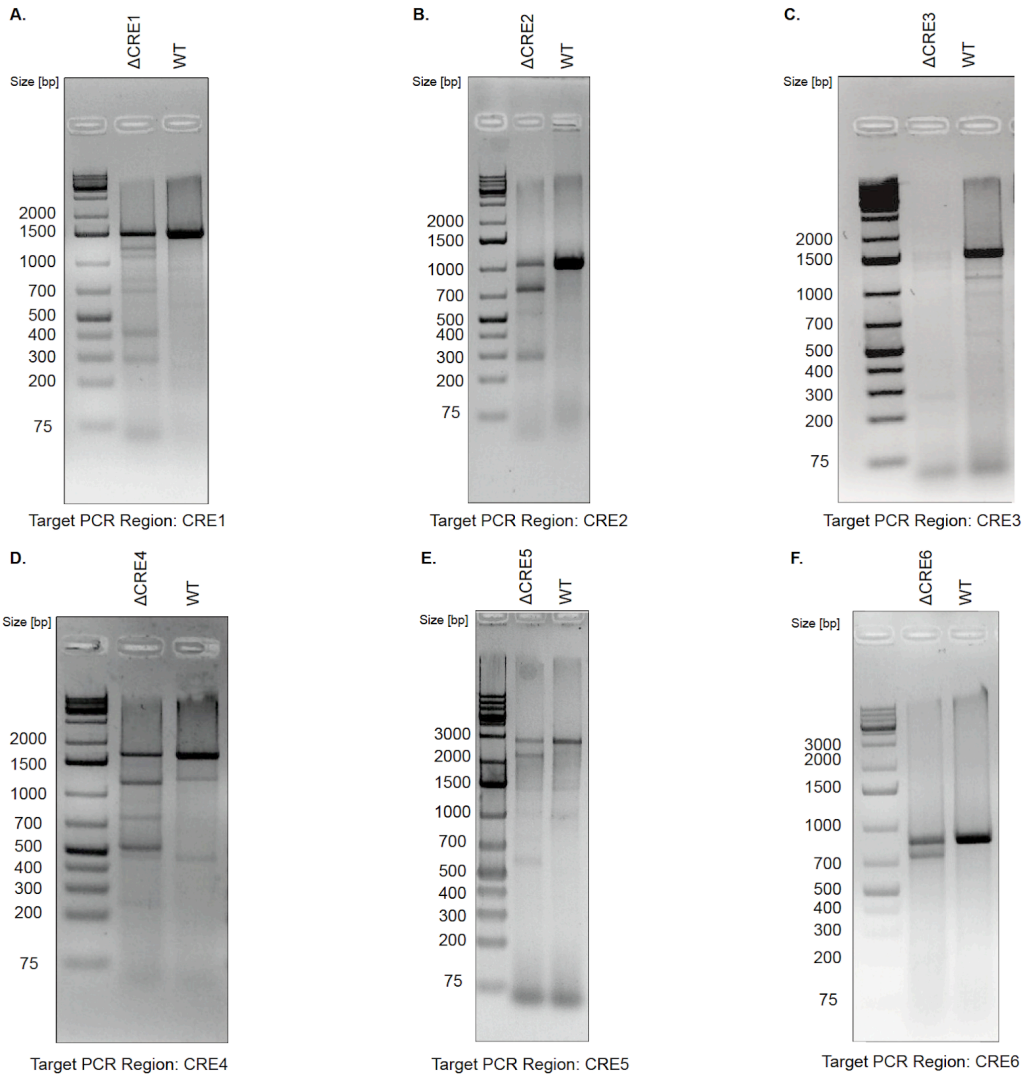

Figure A.11: **Validation of Cas9-mediated deletion of CREs from lentiviral system expressing both Cas9 protein and gRNA for cell proliferation assays. a-f.** Agarose gel of lentiviral-based (expression of Cas9 protein and two gRNA) Cas9-mediated deletion product from PCR amplification of intended CREs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

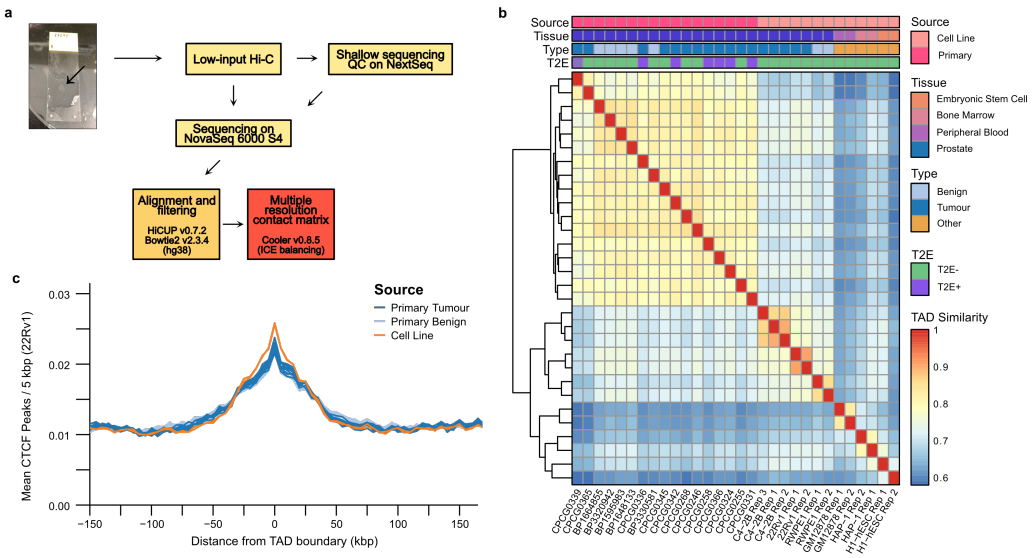

Figure B.1: **Sample processing and TAD similarity between samples. a.** Schematic representation of the protocol and data pre-processing pipeline used in this study to obtain Hi-C sequencing data. **b.** Heatmap of TAD similarities between primary prostate samples, prostate cell lines, and non-prostate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

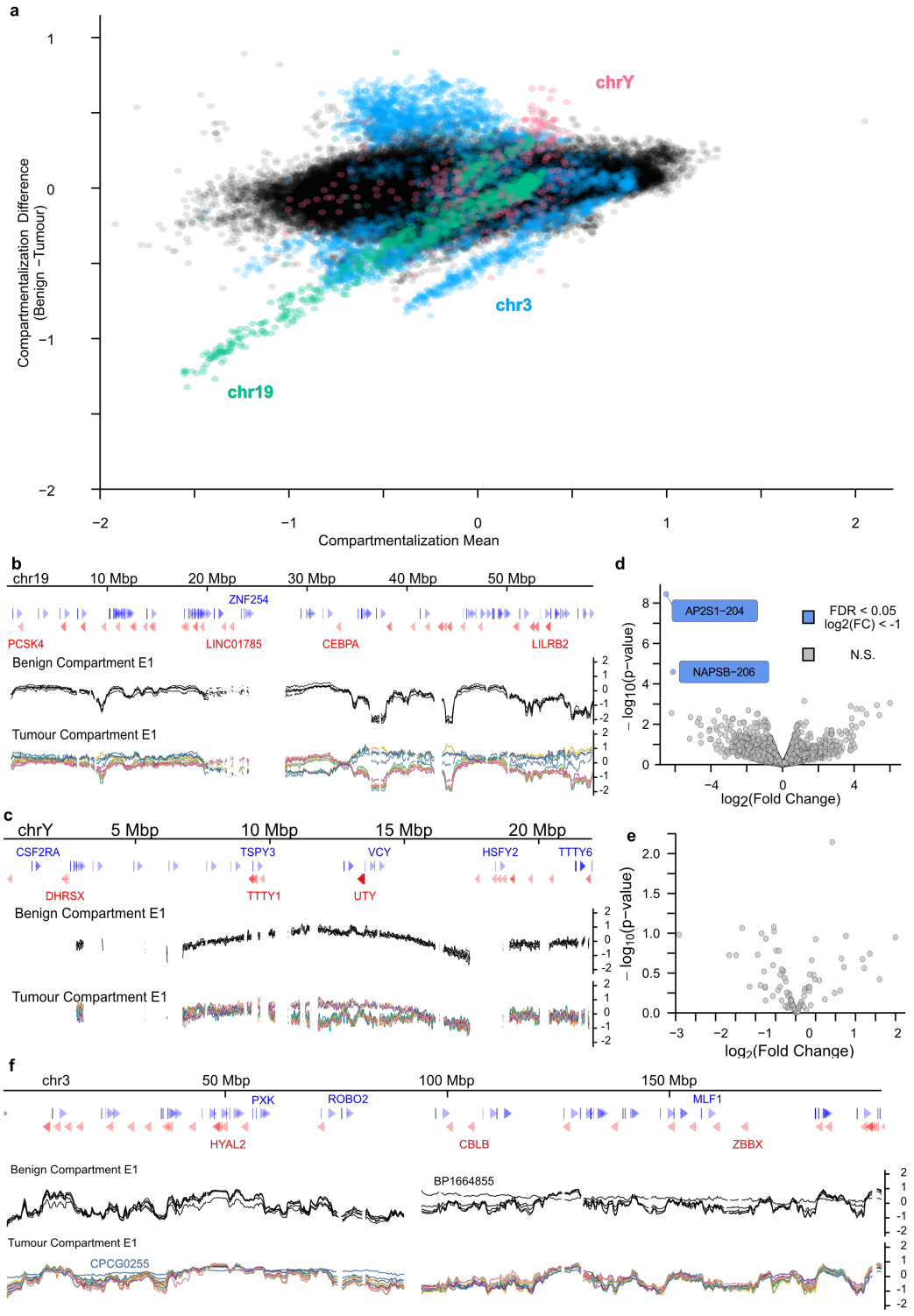

Figure B.2: **Compartmentalization changes in tumours is not associated with widespread differential gene expression.** (Continued on the following page)

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. Coordinates are listed according to the GRCh38 reference genome.

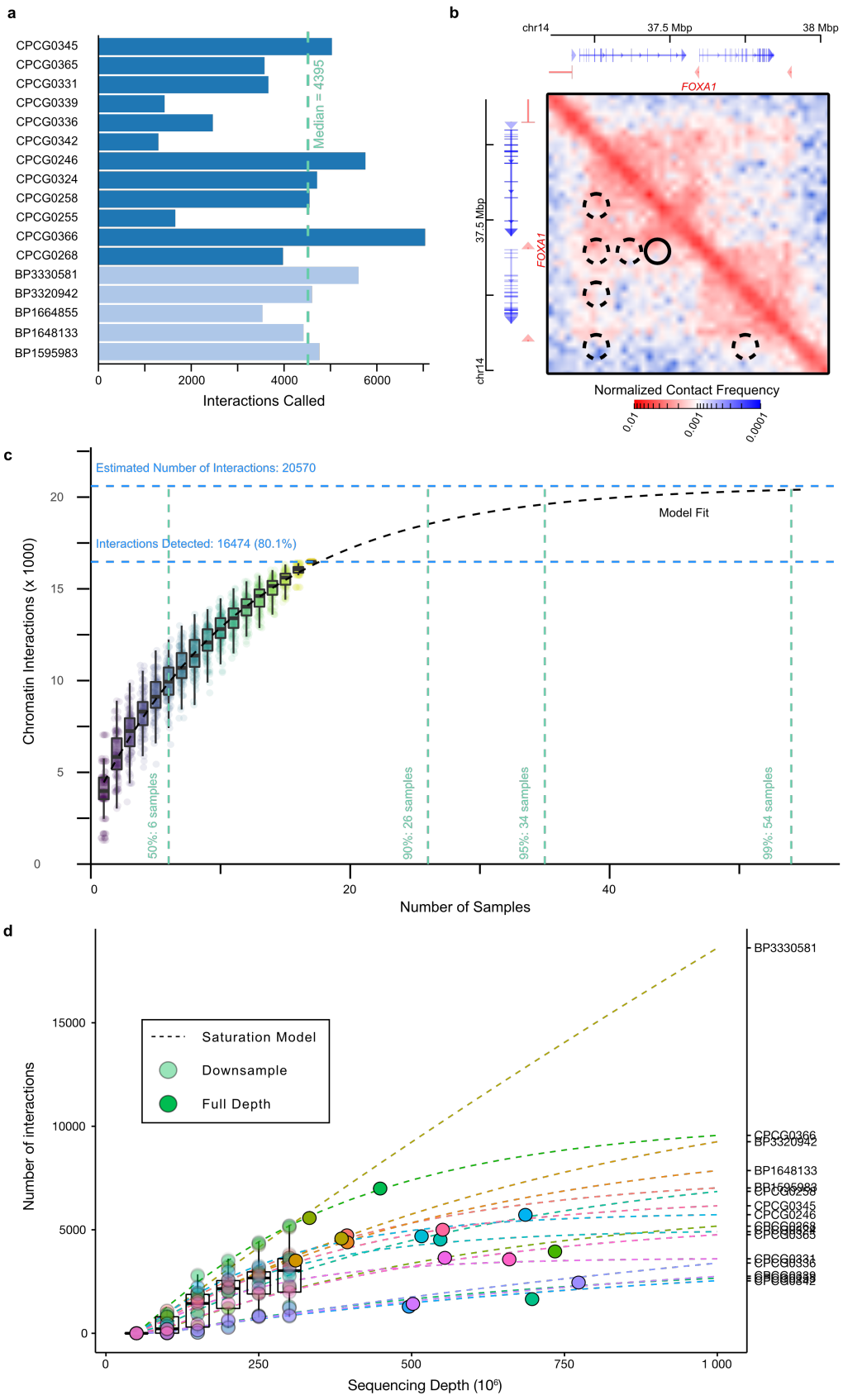

Figure B.3: **Characterization of chromatin interactions in benign and tumour tissue.** (Continued on the following page)

Figure B.3: **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 [**zhouNoncodingMutationsTarget2020**] (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. **d.** Saturation analysis as in (**c**), but performed via downsampling sequencing reads from each individual sample.

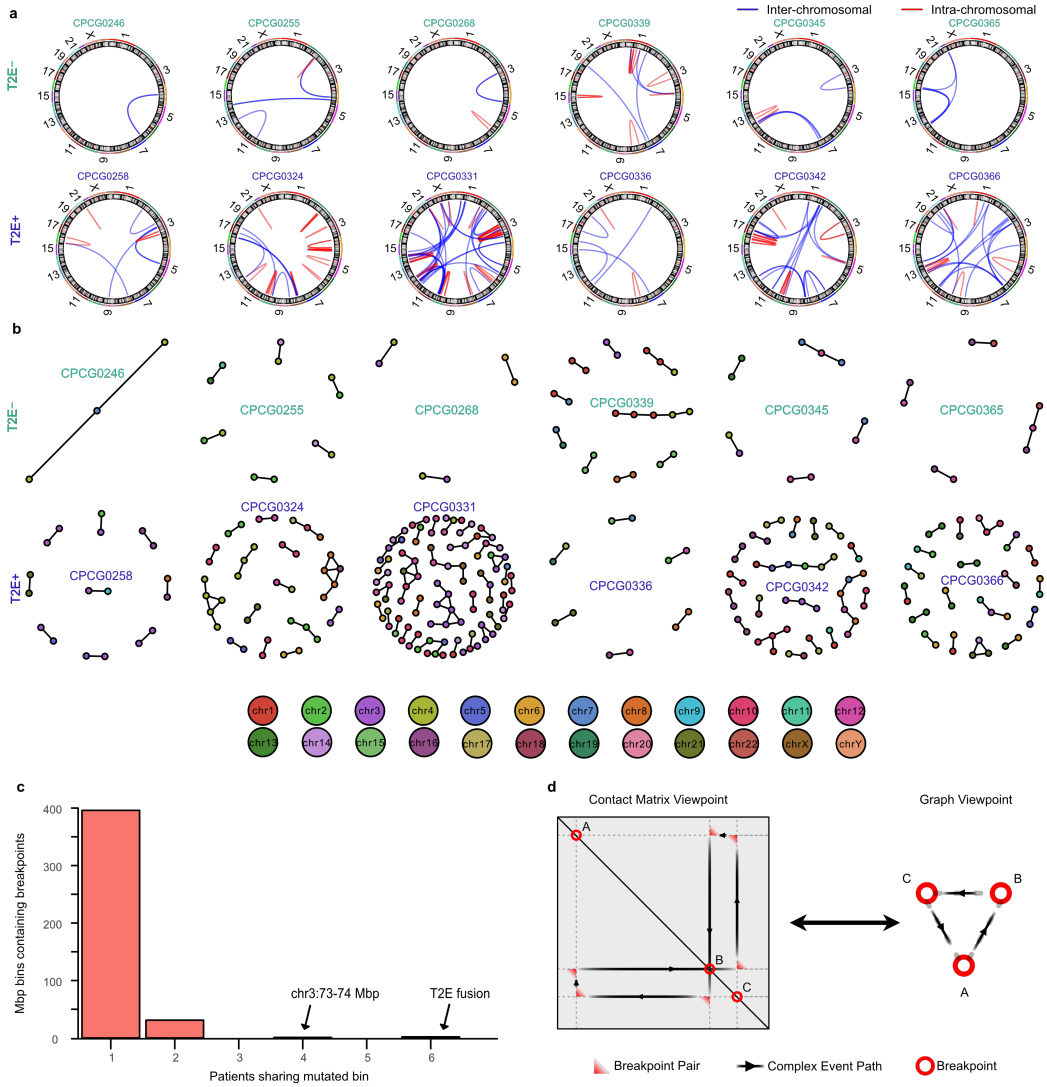

Figure B.4: **Structural variant detection from Hi-C data. a.** Circos plots of SVs identified in the 12 primary prostate tumours. **b.** Graph reconstructions of the simple and complex SVs in all 12 tumours. The node colour corresponds to the chromosome of origin. **c.** Bar plot of the number of 1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

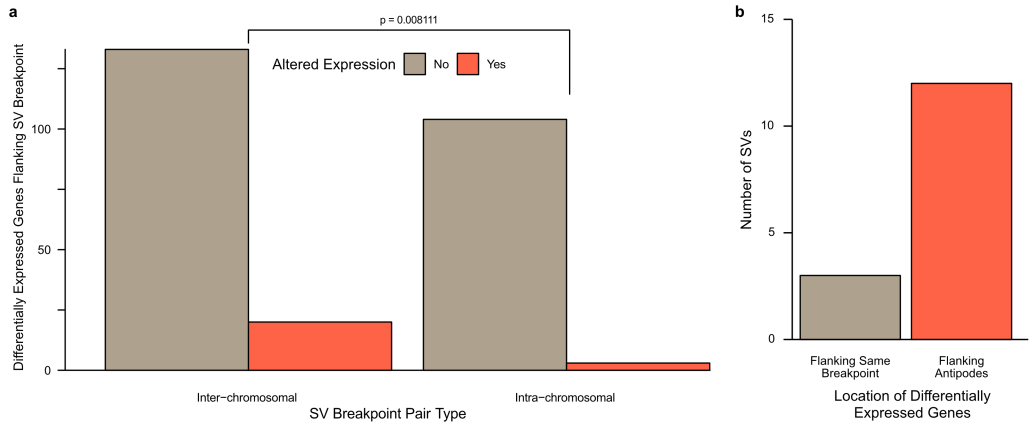

Figure B.5: **Relationship between inter-chromosomal rearrangements and differential gene expression. a.** Bar plot of the number of differentially expressed genes and whether they are involved in SVs spanning multiple chromosomes. Pearson's χ^2 test, $\chi^2 = 7.0088$, $p = 8.11 \times 10^{-3}$, $df = 1$. **b.** Bar plot of all 15 SVs associated with both over- and under-expression, categorized by which breakpoints the differentially expressed genes flank.

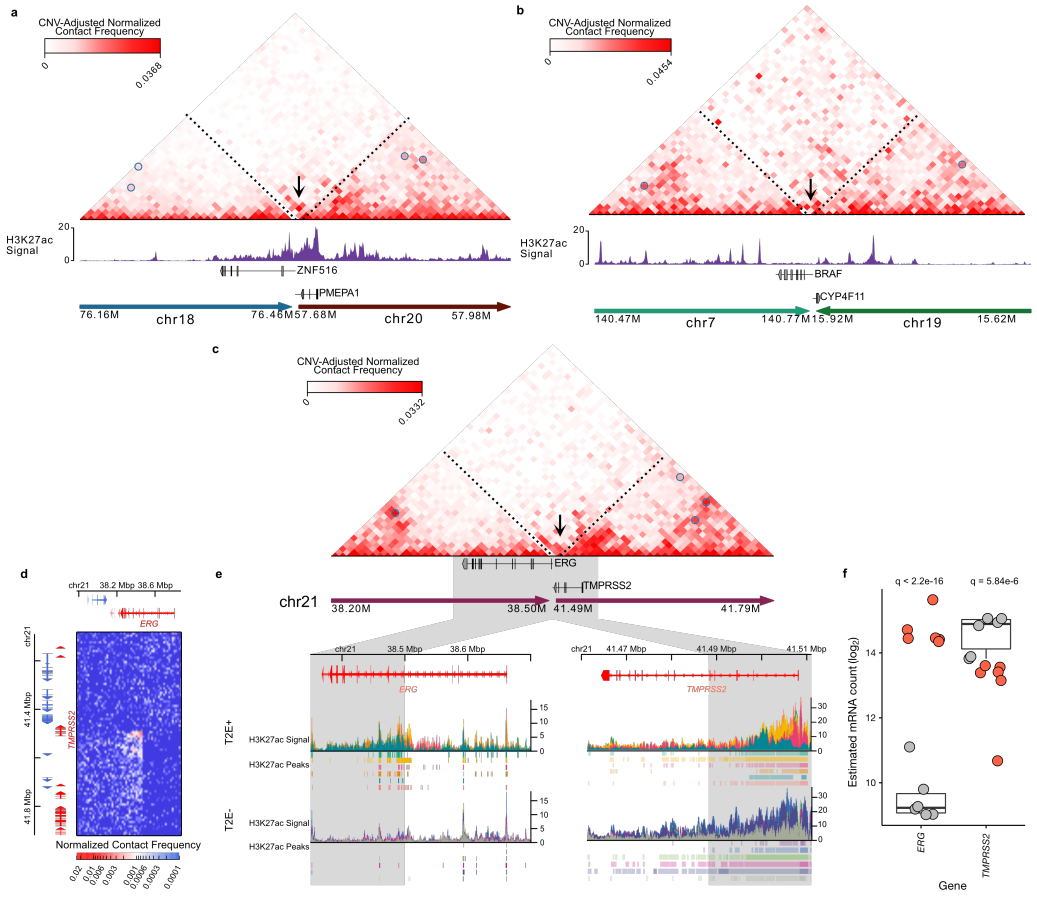

Figure B.6: **Assembly of structural variants involving enhancer-hijacking events. a.** Re-assembled contact matrix from the *ZNF516-PMEPA1* fusion using NeoLoopFinder [**wangGenomewideDetectionEnhancerhijacking2021**]. Dashed lines in the ICE-normalized contact matrix indicate SV breakpoints and black arrows indicate the increased contact frequency between newly-fused segments. **b.** Re-assembled contact matrix from the C2B fusion. **c.** Re-assembled contact matrix of the T2E fusion. **d.** Original contact matrix of the T2E fusion from a single patient. **e.** 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. **f.** 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 Notation

Below is a summary of the notation and symbols that are used throughout this appendix and Chapter 4.

Table C.1: **Operators used throughout Appendix C and Chapter 4.**

Name	Notation	Definition
Expectation	$\mathbb{E}[\cdot]$	Expectation of a random variable.
Covariance	$\mathbb{V}[\cdot, \cdot]$	Covariance of two random variables.
Variance	$\mathbb{V}[\cdot]$	Variance of a random variable.
Trace	$\mathbb{T}[\cdot]$	Trace of a matrix (or the matrix representation of a random variable).

Table C.2: **Random variables used throughout Appendix C and Chapter 4.**

Name	Notation	Units	Definition
Latent transcript abundance	Y	None	Abundance of sequencing reads assigned to each transcript in an organism's transcriptome.
Observed transcript abundance	D	None	Observed abundance of sequencing reads assigned to each transcript in an organism's transcriptome.
Biological noise	ϵ	None	Differences in expression between samples and variance from library preparation methods [**pimentelDifferentialAna**
Inferential noise	ζ	None	Differences in expression arising from computational inference of reads to transcripts and sequencing stochasticity [**pimentelDifferentialAnalysisRNAseq**
OLS mean estimator	$\hat{\mu}^{(OLS)}$	None	Estimator of mean fold change using the OLS method.
JS mean estimator	$\hat{\mu}^{(JS)}$	None	Estimator of mean fold change using the JS method.

Table C.3: **Symbols used throughout Appendix C and Chapter 4.**

Name	Notation	Definition
Estimate	$\hat{\cdot}$	Estimation of some constant or parameter.
Transpose	$\cdot^{\mathrm{T}}$	Transpose of a matrix (or the matrix representation of a random variable).
Normal distribution	$\mathcal{N}_p(\mu, \Sigma)$	A p-dimensional normal distribution with mean μ and covariance matrix Σ.
Identity matrix	I_n, $I_{r\times c}$	A square $n \times n$, or rectangular $r \times c$, identity matrix.

Table C.4: **Parameters and constants used throughout Appendix C and Chapter 4.**

Name	Notation	Definition
Design matrix	X	Matrix of covariates as columns and samples as rows. All p covariates for sample i are denoted by the p-dimensional column vector x_i.
Set of transcripts	S	All transcripts from an organism's transcriptome that are considered in a differential expression analysis.
Wild type sample size	n_{WT}	The number of samples in the "wild type" group.
Mutant sample size	n_{Mut}	The number of samples in the "mutant" group. In this application, $n_{\text{Mut}} = 1$.
Total sample size	n_{Tot}	Total sample size in a differential expression analysis experiment. In this application, $n_{\text{Tot}} = n_{\text{WT}} + n_{\text{Mut}}$.
Variance of inferential noise	τ^2	Inferential noise, ζ, is distributed as $\mathcal{N}_{\lvert S\rvert}\left(0, \tau^2\right)$. By assumption τ^2 is a diagonal matrix where the i-th diagonal element is denoted as τ_i^2.
Variance of biological noise	σ^2	Biological noise, ϵ, is distributed as $\mathcal{N}_{\lvert S\rvert}\left(0, \sigma^2\right)$. By assumption σ^2 is a diagonal matrix where the i-th diagonal element is denoted as σ_i^2.
Shrinkage-based variance of biological noise	$\tilde{\sigma}^2$	A shrunken estimate of biological variance. This is used to reduce the instability of variances estimators in experiments with small sample sizes [**pimentelDifferenti**
Estimated count	k_{si}	Read count of transcript s in sample i as estimated by pseudo-alignment in Kallisto.
Sample scale factor	f_i	Read count scale factor for sample i to adjust for differences in sequencing depth between samples.
Coefficient	β_s	Effect of each covariate on the expression of transcript s. In this application, β_{s0} is the baseline abundnace and β_{s1} is the fold change between mutant and WT samples.

C.2 Sleuth model for differential expression analysis

The differential expression model employed in the Sleuth (v0.30.0) [**pimentelDifferentialAnalysisRNAseq201**
yiGenelevelDifferentialAnalysis2018] can be described as follows. Consider a set of transcripts, S, that are each measured in n_{Tot} samples with an experimental design matrix, $X \in \mathbb{R}^{n_{\text{Tot}} \times p}$, 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:

$$X = \begin{bmatrix} x_1^{\text{T}} \\ x_2^{\text{T}} \\ ... \\ x_n^{\text{T}} \end{bmatrix} \tag{C.1}$$

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

$$Y_{si} = x_i^{\text{T}} \beta_s + \epsilon_{si} \tag{C.2}$$

where $\epsilon_{si} \sim \mathcal{N}(0, \sigma_s^2)$ is the biological noise of transcript s in sample i and $\beta_s \in \mathbb{R}^p$ 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:

$$D_{si}|Y_{si} = Y_{si} + \zeta_{si} \tag{C.3}$$

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:

$$\mathbb{V}\left[\epsilon_{si}, \epsilon_{rj}\right] = \begin{cases} \sigma_s^2 & i = j, s = r \\ 0 \end{cases} \tag{C.4}$$

$$\mathbb{V}\left[\zeta_{si}, \zeta_{rj}\right] = \begin{cases} \tau_s^2 & i = j, s = r \\ 0 \end{cases} \tag{C.5}$$

$$\mathbb{V}\left[\epsilon_{si}, \zeta_{rj}\right] = 0 \tag{C.6}$$

where $s, r \in S$, and $i, j \in \{1, ..., n_{\text{Tot}}\}$. The abundances for transcript s in all n_{Tot} samples can then

modelled as a multivariate normal distribution:

$$D_s \sim \mathcal{N}_{n_{\text{Tot}}}(X\beta_s, (\sigma_s^2 + \tau_s^2) I_{n_{\text{Tot}}}) \tag{C.7}$$

The OLS estimate for the differential effect is given by:

$$\hat{\beta}_s = (X^{\text{T}} X)^{-1} X^{\text{T}} D_s \tag{C.8}$$

D_s is defined by a transformation of read counts to correct for biases such as transcript length and sequencing depth between different samples:

$$D_{si} = \ln\left(\frac{k_{si}}{f_i} + 0.5\right) \tag{C.9}$$

$$f_i = \underset{s \in S^*}{\text{median}} \frac{k_{si}}{\sqrt[N]{\prod_{j=1}^{N} k_{s,j}}} \tag{C.10}$$

where k_{si} is the estimated read count from the Kallisto model [**brayNearoptimalProbabilisticRNAseq2016**] for transcript s in sample i and f_i is the read count scale factor for sample i. The read count scale factor is calculated from the set of all transcripts that pass initial filtering, S^*. This is typically the entire transcriptome, S, but excludes transcripts with read counts below some threshold (see [**pimentelDifferentialAnalysisRNAseq2017**] for details).

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

C.3 Plug-in estimators derived from wild-type samples

Using the Sleuth model described above on the n_{WT} WT samples provides the following estimates:

$$\hat{\text{B}}_0 = \frac{1}{n_{\text{WT}}} \sum_{i=1}^{n_{\text{WT}}} \Delta^{(i)} \tag{C.11}$$

$$\hat{\Sigma} = \begin{bmatrix} \max\{\hat{\sigma}_1^2, \tilde{\sigma}_1^2\} + \hat{\tau}_1^2 & & 0 \\ & \ddots & \\ 0 & & \max\{\hat{\sigma}_{|S|}^2, \tilde{\sigma}_{|S|}^2\} + \hat{\tau}_{|S|}^2 \end{bmatrix} \tag{C.12}$$

where $\Delta^{(i)}$ is the abundance for WT sample i; $\hat{\sigma}_s^2$ is the raw estimate of the biological variance for transcript s; $\tilde{\sigma}_i^2$ is the shrunken estimate of the biological variance for transcript s made through aggregating data across transcripts; and $\hat{\tau}_s^2$ is the estimate of the inferential variance for transcript s [**pimentelDifferentialAnalysisRNAseq2017**].

With these estimators used as plug-in values for Equation (4.5), a JS estimator for the fold change can be calculated according to Theorem 2. Doing so yields the following relations:

$$\mathbb{T}\left[\hat{\Sigma}\right] = \sum_{s\in S} \max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2 \tag{C.13}$$

$$\lambda = \max_{s\in S} \left\{\max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2\right\} \tag{C.14}$$

$$0 \leq c \leq 2\left(\frac{\sum_{s\in S} \max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2}{\max_{s\in S}\left\{\max\{\hat{\sigma}_s^2, \tilde{\sigma}_s^2\} + \hat{\tau}_s^2\right\}} - 2\right) \tag{C.15}$$

C.4 Statistical moments of the OLS estimator

As shown in Supplementary Note 2 of [**pimentelDifferentialAnalysisRNAseq2017**], the OLS estimator is unbiased:

$$\mathbb{E}\left[\hat{\beta}_s^{(OLS)}\right] = \beta_s \tag{C.16}$$

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

$$\mathbb{V}\left[\hat{\beta}_s^{(OLS)}\right] = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\Sigma X(X^{\mathrm{T}}X)^{-1} \tag{C.17}$$

In the case where the covariance matrix is diagonal (i.e. $\Sigma = (\sigma_s^2 + \tau_s^2)I_{n_{\mathrm{Tot}}}$), this reduces to:

$$\mathbb{V}\left[\hat{\beta}_s{}^{(OLS)}\right] = (\sigma_s^2 + \tau_s^2)(X^{\mathrm{T}}X)^{-1} \tag{C.18}$$

Now consider the simple experimental design from Chapter 4, where the only covariate of interest is the presence of a mutation, n_{Mut} samples have this mutation, and n_{WT} samples do not. Then the

design matrix looks like so:

$$X = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{n_{\text{Tot}} \times 2} \tag{C.19}$$

where the first column corresponds to the mean expression (i.e. the intercept, in statistical terms) and the second column corresponds to the mutation status. The variance of the OLS estimator is then:

$$\mathbb{V}\left[\hat{\beta_s}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)}{n_{\text{Mut}} n_{\text{WT}}} \begin{bmatrix} n_{\text{Mut}} & -n_{\text{Mut}} \\ -n_{\text{Mut}} & n_{\text{Mut}} + n_{\text{WT}} \end{bmatrix} \tag{C.20}$$

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

$$\mathbb{V}\left[\beta_{s,\text{Mut}}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)(n_{\text{Mut}} + n_{\text{WT}})}{n_{\text{Mut}} n_{\text{WT}}} \tag{C.21}$$

When there is only a single mutated sample (i.e. $n_{\text{Mut}} = 1$), this reduces to:

$$\mathbb{V}\left[\beta_{s,\text{Mut}}^{(OLS)}\right] = \frac{(\sigma_s^2 + \tau_s^2)(1 + n_{\text{WT}})}{n_{\text{WT}}} \tag{C.22}$$

C.5 Statistical moments of the JS estimator

C.5.1 Expected value of the JS estimator

We can use a Taylor expansion around B_1 to approximate the expected value of $\hat{\text{B}}_1^{(JS)}$. Using the definition of $\hat{\text{B}}_1^{(JS)}$ from Equation (4.8), let $u = \hat{\Sigma}^{-1/2}\hat{\text{B}}_1^{(OLS)}$. Then we have the following:

$$\|u\|^2 = \left(\hat{\text{B}}_1^{(OLS)}\right)^{\text{T}} \hat{\Sigma}^{-1} \hat{\text{B}}_1^{(OLS)} \tag{C.23}$$

$$\hat{\text{B}}_1^{(JS)} = \left(1 - \frac{c}{\|u\|^2}\right) \hat{\Sigma}^{1/2} u \tag{C.24}$$

$$\mathbb{E}\left[\hat{\text{B}}_1^{(JS)}\right] = \mathbb{E}\left[\hat{\text{B}}_1^{(OLS)}\right] - c\hat{\Sigma}^{1/2}\mathbb{E}\left[\frac{u}{\|u\|^2}\right] \tag{C.25}$$

$$= \text{B}_1 - c\hat{\Sigma}^{1/2}\mathbb{E}\left[\frac{u}{\|u\|^2}\right] \tag{C.26}$$

Let $a = \hat{\Sigma}^{-1/2}\mathrm{B}_1$. Then:

$$\|a\|^2 = \mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1 \tag{C.27}$$

Expanding $\frac{u}{\|u\|^2}$ around gives:

$$\mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right] = \mathrm{B}_1 - c\hat{\Sigma}^{1/2}\mathbb{E}\left[\frac{a}{\|a\|^2} + \left(\frac{1}{\|a\|^2} - \frac{2}{\|a\|^4}aa^{\mathrm{T}}\right)(u-a) + \mathcal{O}(\|u-a\|^2)\right] \tag{C.28}$$

$$= \left(1 - \frac{c}{\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1}\right)\mathrm{B}_1 + \mathcal{O}(\|u-a\|^2) \tag{C.29}$$

For sufficiently small $|S|$ and sufficiently small coefficient of variation (i.e. that $\|u-a\|^2 \ll \|\mathrm{B}_1\|^2$), the Taylor approximation is approximately:

$$\mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right] \approx \left(1 - \frac{c}{\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1}\right)\mathrm{B}_1 \tag{C.30}$$

Thus, the JS estimator is an estimate of B_1 that is biased towards 0.

C.5.2 Variance of the JS 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 [**bockMinimaxEstimatorsMean1975**], $\mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(JS)} - \mathrm{B}_1\|^2\right] \leq \mathbb{E}\left[\|\hat{\mathrm{B}}_1^{(OLS)} - \mathrm{B}_1\|^2\right]$. This does not imply that $\mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(JS)}\right] \leq \mathbb{V}\left[\hat{\mathrm{B}}_{1,s}^{(OLS)}\right] \forall s \in S$, however. Some transcripts may have larger variances than the OLS estimator. 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 compute the variance of $\hat{\mathrm{B}}_1^{(JS)}$, we take a similar approach as above with Taylor expansions. Again, let $u = \hat{\Sigma}^{-1/2}\hat{\mathrm{B}}_1^{(OLS)}$. Then:

$$\hat{\mathrm{B}}_1^{(JS)}\left(\hat{\mathrm{B}}_1^{(JS)}\right)^{\mathrm{T}} = \hat{\Sigma}^{1/2}\left(1 - \frac{c}{\|u\|^2}\right)^2 uu^{\mathrm{T}}\hat{\Sigma}^{1/2} \tag{C.31}$$

$$= \hat{\Sigma}^{1/2}\left[uu^{\mathrm{T}} - \frac{2c}{\|u\|^2}uu^{\mathrm{T}} + \left(\frac{c}{\|u\|^2}\right)^2 uu^{\mathrm{T}}\right]\hat{\Sigma}^{1/2} \tag{C.32}$$

The variance is then given by:

$$\mathbb{V}\left[\hat{\mathrm{B}}_1^{(JS)}\right] = \mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\left(\hat{\mathrm{B}}_1^{(JS)}\right)^{\mathrm{T}}\right] - \mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right]\mathbb{E}\left[\hat{\mathrm{B}}_1^{(JS)}\right]^{\mathrm{T}} \tag{C.33}$$

$$\approx \hat{\Sigma}^{1/2}\mathbb{E}\left[uu^{\mathrm{T}} - \frac{2c}{\|u\|^2}uu^{\mathrm{T}} + \left(\frac{c}{\|u\|^2}\right)^2 uu^{\mathrm{T}}\right]\hat{\Sigma}^{1/2} - \left(1 - \frac{c}{\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1}\right)^2 \mathrm{B}_1\mathrm{B}_1^{\mathrm{T}} \tag{C.34}$$

Expanding the expectation about $a = \hat{\Sigma}^{-1/2}\mathrm{B}_1$ yields:

$$\mathbb{V}\left[\hat{\mathrm{B}}_1^{(JS)}\right] \approx \left(1 - \frac{2c}{\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1}\right)\hat{\Sigma} - \frac{2c}{\left(\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^{\mathrm{T}} \tag{C.35}$$

Since the diagonal elements of $\frac{2c}{\left(\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^{\mathrm{T}}$ are all ≥ 0 and $0 \leq \left(1 - \frac{2c}{\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1}\right) \leq 1 \forall c > 0$, the variance than of the JS estimators are smaller than the OLS estimators.

C.6 Wald test statistics for the OLS and JS 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:

$$W_s^{(OLS)} = \frac{\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2}{\hat{\Sigma}_{s,s}} \tag{C.36}$$

$$W_s^{(JS)} = \frac{\left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right)^2 \left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2}{\left(1 - \frac{2c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right)\hat{\Sigma}_{s,s} - \frac{2c}{\left(\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mathrm{B}}_1^{(OLS)}\right)^2}\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2} \tag{C.37}$$

The coefficient for $\hat{\mathrm{B}}_{1,s}^{(OLS)}$ in the numerator

$$\left(1 - \frac{c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right)^2 \tag{C.38}$$

is larger than the coefficient of $\hat{\Sigma}$ in the denominator

$$\left(1 - \frac{2c}{\left(\hat{\mathrm{B}}_1^{(OLS)}\right)^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{\mathrm{B}}_1^{(OLS)}}\right) \tag{C.39}$$

since $1-2d+d^2 > 1-2d \forall d \in \mathbb{R}$. The denominator is also made smaller by the second term involving $\left(\hat{\mathrm{B}}_{1,s}^{(OLS)}\right)^2$. These two terms imply that the Wald test statistics will tend to be larger for the JS estimator than for the OLS estimator. Thus the JS method will tend produce more positive calls, in general, than the OLS method. This can increase the true positive rate of the JS method, while also potentially increasing the false positive rate.

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 $\mathrm{B}_1\mathrm{B}_1^{\mathrm{T}}$ 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, $\hat{\Sigma}$, 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 $\frac{2c}{\left(\mathrm{B}_1^{\mathrm{T}}\hat{\Sigma}^{-1}\mathrm{B}_1\right)^2}\mathrm{B}_1\mathrm{B}_1^{\mathrm{T}}$ term. While often ignored in practice, this statistical dependence can affect 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