Characterizing the genome-wide associations of somatic mutations and chromatin accessibility in cancer

by

Oliver Ocsenas

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Medical Biophysics
University of Toronto

Characterizing the genome-wide associations of somatic mutations and chromatin accessibility in cancer

Oliver Ocsenas

Master of Science

Medical Biophysics

University of Toronto

2020

Abstract

Somatic mutation rates in cancer genomes are associated with chromatin state and informative of tumor tissue of origin. However, earlier studies considered the chromatin states of cell lines and normal tissues, while those of primary tumors remain uncharacterized. We used a machine-learning approach to evaluate tumor-specific chromatin accessibility profiles as predictors of mutation rates in 2,500 whole cancer genomes with 23 million SNVs. Mutation rates were more accurately predicted by chromatin accessibility derived from tumours than healthy cells, suggesting that somatic mutagenesis is largely associated with the chromatin state of tumour cells rather than normal cells. Interestingly, melanoma mutations were better predicted by normal melanocyte chromatin accessibility, suggesting earlier mutational timing. Furthermore, chromatin state was an accurate predictor of carcinogen-induced mutation rates while the mutations of endogenous mutational processes were only weakly statistically associated. Integrative analysis of mutations and chromatin state provides us insight into tumour evolution and heterogeneity.

Acknowledgments

First and foremost, I would like to acknowledge all the patients and their families who have contributed to projects such as TCGA, PCAWG, and ICGC. Thank you to all the organizations that have funded my work and especially the Ontario Institute for Cancer Research.

Thank you to my supervisor Dr. Juri Reimand for the support and mentorship throughout my degree and consistently pushing me to be a better scientist. Thank you to everyone in the Reimand lab for making my experience one I will never forget. I want to a give special thanks to my labmates Karina Isaev, Christian Lee, and Miles Mee for your friendships and support. Finally, thank you to my supervisory committee, Dr. Anne Martel and Dr. Phedias Diamandis for your guidance and scientific expertise.

This experience has been filled with tremendous growth, both on a personal and academic level. I want to specifically thank my friends, my family, my mother Nada, and my partner Alexia, without whose support this could not have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

NGS: Next-generation sequencing

PCAWG: Pan-Cancer Analysis of Whole Genomes

TCGA: The Cancer Genome Atlas

SNV: Single nucleotide variant

NER: Nucleotide excision repair

BER: Base excision repair

MMR: Mismatch repair

WGS: Whole-genome sequencing

TAD: Topologically associated domain

DNase-Seq: DNase I hypersensitive sites sequencing

ATAC-Seq: Assay for Transposase-Accessible Chromatin using sequencing

BP: Base-pair

MB: Megabase-pair

KB: Kilobase-pair

SBS: Single base substitution mutational signature

GBM: Glioblastoma multiforme

LGG: Low-grade glioma

HCC: Hepatocellular carcinoma

CRC: Colorectal carcinoma

Chapter 1

Literature Review

1 Literature Review

1.1 Cancer genetics and biology

1.1.1 Cancer as a genetic disease

Since the completion of the Human Genome Project (Lander et al. 2001), genome sequencing studies have shaped our understanding of many diseases. This is especially true in the case of cancer, where sequencing technology has revolutionized the fields of cancer biology and cancer treatment. First and foremost, the sequencing of a multitude of tumours uncovered that tumours from different parts of the body have distinctly different genetic makeups. Furthermore, tumours of the same tumour type from different patients can demonstrate remarkable molecular heterogeneity. Using genome sequencing data to uncover novel cancer subtypes is an area of active research.

Sequencing experiments for cancer research in the past have mostly focused on the genomic and transcriptomic levels. On the genomic level, positive selection for both gain-of-function and loss-of-function mutations in genes has been demonstrated, with many of these genes thought to be involved in driving cancer-related phenotypes (Martincorena et al. 2017). On the transcriptomic level, the up- or down-regulation of specific genes and pathways has been linked to molecular subtypes and clinical outcome (Sanchez-Vega et al. 2018). Collectively, these experiments have broadened our understanding of cancer biology, prevention, and treatment.

1.1.2 Driver mutations in cancer

Cancer is a disease caused by somatic mutations in cells. Somatic mutations indicate mutations which are passed along from cell to cell, through mitosis, in the somatic cells of an individual. These changes are not inherited from parents and are not passed onto offspring (unlike germline mutations). Cells tend to acquire somatic mutations in every tissue throughout the body and these accumulate throughout the lifetime of an individual. Despite the propensity of somatic cells for genomic alterations, only a minority of these will lead or contribute to tumour evolution. These mutations, known as driver mutations, confer certain hallmark properties to the cell. The hallmarks of cancer were described in a seminal paper by Hanahan and Weinberg in 2000 and again updated in 2011 (Hanahan and Weinberg 2000; Hanahan and Weinberg 2011). The authors argued that all cancers share six hallmark traits which govern oncogenesis. The hallmarks include self-sufficiency in growth signals, evading growth-suppressors, resisting cell death, inducing angiogenesis, enabling replicative immortality, and activating invasion and metastasis. Cancer cells most commonly acquire the hallmarks of cancer through alterations to protein-coding genes. There are two main classes of protein-coding genes relevant to cancer: oncogenes and tumour suppressors.

Oncogenes are originally normally functioning genes ("proto-oncogenes") which can become oncogenic through two distinct methods. The first method involves a significant increase in their expression leading to the upregulation of their original function. This increase can come from mutations to the gene itself, its non-coding promoter region, another protein-coding gene, or even another non-coding regulatory region/gene. One example of such an oncogene is *EGFR* (epidermal growth factor receptor) (Zandi et al. 2007), which is involved in inducing cellular proliferation, and has been shown to be significantly upregulated in both wild-type and mutant forms in many cancers. The second method occurs when proto-oncogenes undergo a gain-of-function mutation, whereby they can become hyperactive and oncogenic. For example, the *Ras* family of proto-oncogenes are involved in cellular signalling leading to cell growth. Mutations to these genes in cancer can lead to proteins which are locked in their active states, leading to uncontrolled cell growth. *Ras* family mutations are the first discovered instance of mutations to oncogenes and 20-30% of all cancer cases involve a mutation to a *Ras* gene (Fernández-Medarde & Santos 2011).

According to Darwinian evolution, gain-of-function mutations would be positively selected for in these genes. Indeed, missense mutations in *NRAS*, *KRAS*, and *HRAS* showed evidence of positive selection in cancer.

The second class of protein-coding genes relevant to cancer consists of tumour-suppressor genes. The human genome has evolved many mechanisms to suppress oncogenic processes, such as genes involved in regulating cell growth, proliferation, and migration. The "two-hit hypothesis" states that both alleles of a tumour suppressor gene need to be mutated for an oncogenic change to the phenotype to occur (Knudson 2001). In many cases, one allele of the gene is mutated in the germline genome, without causing any change in phenotype. The second, wild-type, allele is then mutated later in life contributing to oncogenesis (MacPherson & Dyer 2007). The most important example of a tumour suppressor is the gene *TP53* (tumour protein p53) which is mutated in >50% of human cancers. *TP53* plays an important role in cell cycle progression, DNA repair, and apoptosis. Loss of this gene has been strongly implicated in oncogenesis (Olivier, Hollstein, & Hainaut 2010). Loss-of-function mutations in tumour-suppressor genes also follow Darwinian principles, with positively selected mutations being shown in known tumour-suppressors such as *TP53*, *PTEN (*phosphatase and tensin homolog), and *RB1* (RB transcriptional corepressor 1) in many cancers (Martincorena et al. 2017).

On average, cancer genomes contain 4-5 driver mutations, however, the average number of driver mutations varies greatly between cancer types. Interestingly, there is a subset of tumours with no well-known driver mutations. The 2020 Pan-Cancer Analysis of Whole Genomes (PCAWG) study showed that ~5% of tumours had no mutations in known driver elements (Campbell et al. 2020). Thus, there may be currently unidentified driver elements which require larger-scale studies to identify in the future.

1.1.3 The role of non-coding mutations in cancer

Cancer research focuses primarily on mutations in protein-coding genes. These elements, however, make up less than 2% of the human genome. The function of the non-coding

genome is mostly unclear, but there are some examples of cancer driver mutations occurring in non-coding elements.

The best-characterized of these are mutations to the promoter of the gene *TERT* (telomerase reverse transcriptase). *TERT* codes for a catalytic subunit of the enzyme telomerase which is involved in lengthening the telomere regions of the chromosomes. Telomere lengthening allows cells to escape a postmitotic state and potentially achieve replicative immortality; one of the hallmarks of cancer. Deregulation of *TERT* expression has been linked to oncogenesis in many cancers. The *TERT* promoter sequence regulates the gene's transcription and mutations to it have been shown to enhance its activity two- to four-fold in melanoma (Kim et al. 2016). *TERT* promoter mutations have also been shown to be significantly prognostic of survival in thyroid cancer (Kim et al. 2016).

More recently, work using the PCAWG project identified multiple non-coding driver mutations in cancer. Mutations in the 3′ untranslated regions (UTR's) of the genes *TOB1* (transducer of ERBB2 1), *NFKBIZ* (NFKB inhibitor zeta), and *ALB* (albumin) were shown to be recurrent in specific cancer types (Rheinbay et al. 2020). Furthermore, work in our lab has demonstrated that mutations in several non-coding regulatory elements affect known cancer gene transcription levels through chromatin interactions (Zhu et al. 2020). The known cancer genes *CCNB1IP1* (cyclin B1 interacting protein 1), *ICK* (intestinal cell kinase), and *ZKSCAN3* (zinc finger with KRAB and SCAN Domains 3) are all distally regulated by regulatory elements that were shown to be frequently mutated in cancer.

1.1.4 Somatic mutational processes in normal tissues

Several landmark studies have recently challenged the classical notion of cancer drivers. In classical cancer driver discovery, one sequences all the mutations in a cancer cohort. One then considers the most frequently mutated genes within the cohort relative to a background sequence and labels them as positively selected for and therefore driving cancer. It has been recently shown that these same "driver" mutations exist in many normal tissues around the body and accumulate with age.

Clonal expansion and positive selection underlie the evolution of cancer cells. However, it has been also shown that this process is continuously active in normal cells. In 2015, Martincorena and colleagues showed that normal human skin contains somatic mutations previously thought to be exclusively cancer drivers such as *NOTCH1/3 (*Notch receptor 1/3), *TP53*, *FAT1* (FAT Atypical Cadherin 1), and *RBM10* (RNA binding motif protein 10) (Martincorena et al. 2015). Furthermore, they showed that clonal expansion due to positive selection also occurs in normal cells and significantly increases with age. This phenomenon was not limited to skin, however, as significant driver mutations and age-related clonal expansion were also shown in normal esophageal tissue (Martincorena et al. 2018).

The high frequency of positively selected clones in normal tissues raises the question of the relative rarity of cancer in the population. One possible explanation is that tumour suppressive mechanisms are sufficiently robust, effective, and diverse that these clones can exist for long periods of time without ever undergoing oncogenesis. The presence of cancer associated mutations in normal cells also suggests that cancer is a continuum along which all of our cells lie. The end stage of that continuum is the overt, clinically diagnosed malignancy with which we are familiar. However, many cells can be in a "precancerous" stage without ever leading to cancer (Martincorena 2019).

1.2 Regional mutation rates in cancer genomes

1.2.1 Passenger mutations in cancer

Although cancer research tends to focus on driver mutations, most mutations in a cancer genome are under no evolutionary selection, confer no advantage, and have no effect on cellular phenotypes (Martincorena et al. 2017). These mutations are known as passenger mutations. The total number of passenger mutations greatly exceeds that of driver mutations; however, it is highly variable between cancer types (Lawrence et al 2013). On the higher end, melanoma, due to skin tissue's lifetime of exposure to UV light, has greater than 100,000 somatic single nucleotide variants (SNV's) and insertions/deletions (indels) per genome on average. On the lower end, childhood brain tumours, such as medulloblastoma, have less than 2000 somatic SNV's and indels per genome on average. This is mostly like

due to these tumours lacking both the age and the exposure to exogenous mutagens required to accumulate a high mutation burden.

It has been shown that the regional mutation rates at the level of large genomic windows is highly variable and cancer-type specific. Megabase-scale windows are commonly used to evaluate variability in regional mutation rate. One study was able to predict a patient's cancer type at 92% accuracy using only regional mutation rates and a support vector machine (SVM) model (Salvadores, Mas-Ponte, & Supek 2019). Interestingly, using only driver mutations to classify tumours resulted in an inferior accuracy of 36% in the same study. Furthermore, another study showed that deep learning models trained on regional mutation rates can accurately classify the cancer type of primary tumours and metastatic samples at a 91% rate (Jiao et al., 2020). This is an important observation because in 3% of cancer cases, a patient has a metastatic tumour with no clear primary tumour of origin. Identification of the primary tumour of origin can clarify diagnosis, prognosis, and treatment. The accumulation of mutations in various genomic regions is strongly associated with differential DNA damage and repair processes, chromatin accessibility, and replication timing.

1.2.2 DNA damage and repair

DNA mutations, such as SNV's and small indels, occur due to a complex interplay between DNA damage, repair, and replication. Initially, there may be a DNA damaging agent or process which causes a DNA nucleotide to be chemically modified. These agents and processes may be endogenous (*i.e.*, originating from the cells or the organism) and/or exogenous (*i.e.*, originating from the environment). Some common examples of exogenous DNA damage agents are UV light, smoking, and reactive oxygen species. Endogenous mutational processes are most commonly caused by defects in DNA repair and low-fidelity replication pathways.

DNA damage occurs continually in healthy tissues and single-stranded DNA damage is repaired via one of three major DNA repair pathways. The first of these is the Base Excision Repair (BER) pathway which commonly repairs damaged single bases. The single damaged base is removed from the DNA by glycosylase enzymes which cleave the bond between the

base and its corresponding deoxyribose without disrupting the sugar-phosphate backbone. This creates an apurinic/apyrimidinic site (AP site) which is then corrected by AP endonucleases, DNA polymerase, and DNA ligase (Wallace et al. 2012). The second DNA repair pathway is the Nucleotide Excision Repair (NER) pathway which typically repairs bulky DNA damage due to UV light such as pyrimidine dimerization. NER involves removal of DNA of up to a dozen base pairs both upstream and downstream of the damaged nucleotide after which DNA is resynthesized (Martein et al. 2014).

The third major DNA repair pathway is the Mismatch Repair (MMR) pathway. MMR repairs mismatches which occur spontaneously during DNA replication. It is thought that each time a cell divides, approximately 100,000 polymerase errors occur by chance alone. These are normally corrected through the proofreading mechanisms of the polymerases *epsilon* and *delta*. However, some mismatches escape proofreading and must be corrected by the MMR pathway. MMR pathway proteins detect the mismatch, degrade the mutated stretch of DNA, and initiate re-synthesis. Defects in MMR due to mutations to one or more of the MMR pathway genes are commonly found in cancer and can lead to the microsatellite instability phenotype (MSI) (Baretti & Le 2018). MSI describes a "mutator phenotype" related to high levels of genomic instability and frameshift mutations. MSI has been recorded most prominently in colorectal adenocarcinoma as well as other cancer types such as endometrial adenocarcinoma, stomach adenocarcinoma, adrenocortical carcinoma, breast adenocarcinoma, and others (Bonneville et al. 2017). MMR deficiency has also been shown to be important in the accumulation of mutations related to temozolomide treatment in cancer (Pich et al. 2019).

1.2.3 Mutational signatures in human cancers

As described in the previous section, mutations accumulate as a result of specific endogenous and exogenous mutational processes. Notably, each of these processes are thought to generate a characteristic mutational signature that can be mathematically deciphered. An iterative non-negative matrix factorization (NMF) method has been previously used to discover *de novo* signatures in cancer cohorts with whole-genome sequencing data (Alexandrov et al. 2013). Some of these mutational signatures have been traced back to

specific mutational processes allowing us to better understand which mutational processes are contributing to a set of somatic mutations. Sequencing studies of multiple types of cancer have led to the discovery of more than 40 single-base substitution (SBS) signatures, most with unknown aetiologies (Alexandrov et al. 2020).

SBS signatures consist of probabilities for each substitution type which includes the reference allele, the alternate allele, and the nucleotides adjacent to the mutation in the 5' and 3' directions (96 possible combinations as there are 6 types of substitution x 4 types of 5' base x 4 types of 3' base). Several of these signatures have been attributed to exogenous mutational sources such as tobacco smoke and ultraviolet light as well as endogenous sources such as defective DNA repair and 5-methylcytosine deamination. As an example, SBS4 (attributed to smoking) is dominated by C>A mutations, most commonly in the CCA trinucleotide context (Alexandrov et al. 2013). As opposed to the trinucleotide context based SBS signatures analyzed in this project, signatures involving pentanucleotide context, insertions/deletions, and structural variants have been described more recently.

Multiple mutational processes occurring in a cell will contribute to its mutational landscape. Therefore, the total set of somatic mutations from a single tumour sample often contains mutations from different mutational signatures, most of which have unknown aetiologies. Furthermore, SBS signatures are highly variable between cancer types and cancer samples, as well as in the number of mutations attributed to that signature per cancer sample (Alexandrov et al. 2020).

1.2.4 Scales and covariates of regional mutation rates

Several studies have shown a correlation between regional mutational rates and various other genomic features. The foremost of these include replication timing, histone marks, and chromatin accessibility (Schuster-Böckler & Lehner 2012; Polak et al. 2015). All classes of substitutions are increased in later-replicating regions of DNA, indicating a general mechanism of increased DNA damage and/or decreased DNA repair in these regions. One possible mechanism is the stalling of replication forks in the latter stages of DNA replication, leading to accumulation of single-stranded DNA (ssDNA) regions. ssDNA is more

susceptible to DNA damage, thereby leading to more mutations in these regions (Stamatoyannopoulos et al. 2009).

Variation in regional mutation rates has been observed at multiple genomic scales, ranging from the single base-pair resolution to the megabase-pair resolution (Supek & Lehner 2019). Furthermore, the mechanisms which underlie these mutation rates are different depending on the scale being examined. Starting at the domain scale (10^5-10^6 base pairs), high mutation rates are strongly correlated with repressive histone marks (i.e. H3K9me3) and lower chromatin accessibility. Mounting evidence points to the differential activity of the DNA mismatch repair (MMR) pathway across genomic regions as MMR-deficient tumours were shown to lose regional variation in mutation rates at the domain-scale (Supek & Lehner 2015). Some possible mechanisms are the preferential binding of MMR complexes onto earlier-replicating DNA, the depletion of MMR pathway proteins later in replication, and the reduced accessibility of heterochromatin to MMR repair factors. Increased transcription-coupled DNA repair in open chromatin may also be a contributing factor as RNA polymerases stalled at damaged nucleotides has been shown to recruit NER repair factors (Svejstrup 2002).

When examining regional mutation rates at the gene scale (10^3-10^5 base pairs), the clearest pattern is the asymmetry in mutation rates between the transcribed and non-transcribed strands of DNA. This strand bias results from two processes. The first is that the NER pathway preferentially targets the transcribed strand due to the stalling of RNA polymerase at DNA lesions. The second is an increase in DNA damage at the non-transcribed strand. This is thought to be due to its exposure as vulnerable single-stranded DNA, as the mutation signatures associated with this process correspond to single-strand related mutagenesis (Haradhvala et al. 2016). Mutations in primary tumours caused by platinum-based chemotherapies have also been shown to exhibit transcription-strand asymmetry due to preferential NER on the transcribed strand (Pich et al. 2019).

At the sub-gene scale (10^1-10^3 base pairs), we find that regional mutation rate patterns are highly cell-type specific. For example, there is a high mutation rate at the binding sites of the

protein CTCF and its binding partner cohesin in colon cancer, liver cancer, stomach cancer, and melanoma. Evidence suggests that this may be due to an exclusion of DNA repair factors (MMR and NER) from the binding sites (Kaitanen et al. 2015; Poulos et al. 2016; Guo et al. 2018). Furthermore, regional mutation rates have been shown to be associated with nucleosome (DNA wrapped around an octamer of histone proteins) occupancy as mutation rates follow a 200 bp periodicity matching the inter-nucleosomal distance (Brown et al. 2018). There are also slight variations in mutation rate at a 10bp periodicity which is associated with the major groove/minor groove constraints of DNA as it is wrapped around the nucleosome (Pich et al. 2018).

1.2.5 Whole-genome sequencing datasets of tumours

With the advent of next-generation sequencing, it is possible to document every single point mutation in a cancer genome. In 2008, based on the opportunity that this new technology provided, the cancer genomics community established the International Cancer Genome Consortium (ICGC) with the goal of systematically documenting the somatic alterations that drive a diverse set of cancer types (Hudson et al. 2010; Campbell et al. 2020). The PCAWG collaboration was established to allow for cross-tumour comparisons of cancer whole genome mutation data to be meaningful. All samples in the PCAWG dataset were sequenced and processed using "gold-standard, benchmarked, version-controlled algorithms" (Campbell et al. 2020). In early 2020, the PCAWG papers were released collectively in the journal *Nature* and affiliated journals. The PCAWG dataset contains whole-genome sequencing data from 2,605 primary tumours and 173 metastases from 38 different cancer types. To date, this is the largest-scale, most-diverse, and highest quality dataset of tumour whole-genome mutation data.

The tumour genome represents a combination of all the mutational processes mentioned in this section. While whole-genome sequencing (WGS) of tumours gives us a snapshot in time of the effects of these processes, the temporal, spatial, and mechanistic aspects of these processes are largely uncharacterized. Mutations show significant variation between cancer types, samples from the same cancer type, and even between cells within the same tumour. Furthermore, mutational processes show significant spatial variation at the domain-scale,

gene-scale, sub-gene scale, and single base-scale. Understanding these processes will allow us to better understand tumour evolution, improve driver discovery, and develop prognostic and diagnostic tools.

1.3 Epigenetic markings of the genome

1.3.1 Chromatin accessibility and histone marks

Chromatin describes the complex of DNA and protein found in the nucleus of eukaryotic cells. It involves DNA wrapped around an octamer of histone proteins, also known as a nucleosome, and packaged into a highly condensed form. This is crucial, as the DNA strand can be over 3 metres long and must be packaged into a single nucleus with an average diameter of 6 micrometres. Some regions of DNA are less condensed than others, however, and these regions are often associated with transcriptional activity. This is because open chromatin is more accessible to the cell's transcriptional machinery such as transcription factors and RNA polymerase (Flavahan et al. 2017).

Chromatin accessibility is regulated through a variety of post-translational modifications to the tails of histones (also known as histone marks). This regulation occurs through the dynamic activity of enzymes known as histone "writers", "erasers", and "readers". Histone writers deposit methyl, acetyl, and/or phosphoryl groups on specific histone tail residues while erasers remove these modifications. Readers identify the histone modification and exert a downstream effect. While the function of many of these modifications is not fully understood, we have been able to map histone marks to a variety of genomes and associate them with various transcriptional and chromatin states through chromatin immunoprecipitation sequencing (ChIP-Seq). Reader enzymes have multiple reader domains suggesting that they can recognize a combination of multiple marks on neighbouring histones concurrently and that combinations of neighbouring histone marks have variable downstream effects (Gates et al. 2017).

Most of the histone marks occur on lysine residues as they are particularly abundant on histone tails (Tan et al., 2011). In 2015, the Roadmap Epigenomics project released a

mapping of chromatin accessibility (DNase-Seq) and 5 core histone marks **(Table 1)** in 111 reference human epigenomes from various tissues in the body (Roadmap Epigenomics Consortium et al., 2015). These included the mono-methylation of histone 3 lysine 4 (H3K4me1), the tri-methylation of histone 3 lysine 4 (H3K4me3), the tri-methylation of histone 3 lysine 36 (H3K36me3), the tri-methylation of histone 3 lysine 27 (H3K27me3), and the tri-methylation of histone 3 lysine 9 (H3K9me3). H3K4me1 marks are associated with enhancer regions (distal activating elements) while H3K4me3 marks are associated with gene promoters (Heintzman et al., 2007). Both marks are associated with increased transcriptional activity and tend to be near active transcription start sites (Roadmap Epigenomics Consortium et al., 2015). H3K36me3 marks are associated with transcribed regions such as gene bodies while H3K27me3 and H3K9me3 are associated with repressed regions. (Bonasio, Tu, & Reinberg, 2010; Peters et al., 2015).

Table 1: The 5 core histone marks in the Roadmap Epigenomics project and their associations

Histone Mark	**Association with element (transcription)**
H3K4me1	Enhancers (activation)
H3K4me3	Promoters (activation)
H3K36me3	Transcribed gene bodies (activation)
H3K27me3	Polycomb repression (inhibition)
H3K9me3	Heterochromatin (inhibition)

1.3.2 Epigenetic regulation and cell differentiation

A single human genome can lead to hundreds of different cell types, each with varying gene expression patterns. Gene expression patterns are controlled by dynamic epigenetic regulatory mechanisms such as histone modifications, DNA methylation, and nucleosome positioning. These mechanisms control how accessible chromatin is in a specific region thereby regulating the expression of genes contained in that region of DNA (Allis & Jenuwein 2016). However, it is not only the accessibility of chromatin that controls gene

expression. The three-dimensional architecture of DNA is crucial in regulating the interactions between a gene's promoter element (which controls the gene's expression) and other distal regulatory elements such as enhancers and insulators (Dixon, Gorkin, & Ren 2016; Dekker & Misteli 2015). Individual genomic loci are organized into topologically associated domains (TADs) which allow for the coordinated regulation of genes found within various TADs. Genes designed to be active must be accessible to transcription factors, while repressed genes must, at the same time, be sequestered and inaccessible to the same machinery (Flavahan et al. 2017).

A prominent function of chromatin, with respect to cell differentiation, was shown to be the prevention of trans-differentiation that corresponds to the reprogramming of already differentiated cells. In one study, cells from *Drosophila* embryos were modified to be deficient in Polycomb repressors (one family of chromatin regulators) and were found to have the ability to reprogram and switch states (Lee, Maurange, & Paro 2005). The mechanism for this process is that repressive chromatin regulators sequester genes and loci which are unused for the specific lineage, thereby providing a barrier to further differentiation. Interestingly, in undifferentiated pluripotent cells, many loci are in a more dynamic and bivalent state, where genes can go from active to repressed more easily (Flavahan et al. 2017). After differentiation, these restrictive mechanisms become more stable, thereby securing the fate of the cell. The dysregulation of these cell differentiation pathways has been shown to be prominent in tumour development.

1.3.3 Epigenetic regulation in tumour development

In addition to the genetic basis of cancer, epigenetic alterations have been shown to be relevant to tumorigenesis. Cancers are known to be associated with aberrations in gene expression, cellular identity, and environmental response, all of which are regulated by chromatin remodelers. Around 50% of human cancers contain mutations in chromatin remodelling genes (You & Jones 2012). In fact, many tumours demonstrate an aberrant differentiation program indicative of a deregulation of chromatin architecture (Allis & Jenuwein 2016). These aberrant differentiation programs, sometimes caused by mutations to

genes involved in chromatin regulatory pathways, can lead to both increased epigenetic restriction and increased epigenetic plasticity.

In terms of increased epigenetic restriction, it has been shown that the gene *EZH2* frequently accumulates gain-of-function mutations in several lymphoma subtypes as well as melanoma (Gan et al. 2018; Donaldson-Collier 2019). EZH2 is part of the Polycomb repressive complex 2 (PRC2) which is involved in B-cell differentiation, with high activity found in B-cell precursors and a downregulation found in differentiated B-cells. This gain-of-function mutation is shown to induce a restrictive state within B-cells, which prevents differentiation and locks the cell in a proliferative state (Lee & Chang 2019). As previously mentioned, uncontrolled cellular proliferation is one the hallmarks of cancer and is crucial for tumour development.

On the other hand, epigenetic aberrations can lead to increased epigenetic plasticity in tumours. Epigenetic plasticity allows tumour cells to sample various transcriptional states and differentiation programs, some of which may confer oncogenic advantages. One important example of trans-differentiation in cancer is the epithelial-to-mesenchymal transition (EMT) which is seen in many tumours (Suvà, Riggi, & Bernstein 2013). EMT plays a crucial role in metastasis by allowing cancer cells to escape their tissue-specific loci and invade other parts of the body. The reverse process (MET) is then used to settle and colonize the new metastatic niche, developing into a secondary tumour (Roche 2018). Another important example of epigenetic plasticity in cancer is the disruption of oncogene insulation. Gain-of-function mutations in the *IDH1* (isocitrate dehydrogenase (NADP+) 1) gene are shown to be tumour-initiating in glioma, leukaemia, and other tumours (Cairns & Mak 2013). Mutant *IDH1* generates metabolites which inhibit DNA demethylating enzymes, resulting in hypermethylation disrupting the DNA binding of the transcription factor *CTCF* (CCCTC-binding factor). *CTCF* is involved in the establishment of chromosomal loops and TADs thereby insulating oncogenes from potential activating mechanisms. Disruption of *CTCF* binding in *IDH1* mutant gliomas is associated with insulator dysfunction. More specifically, nearby genes normally separated by *CTCF*-mediated insulation show higher correlation of their expression. For example, the glioma oncogene *PDGFRA* (platelet derived

growth factor receptor alpha**)** is upregulated upon the loss of *CTCF*-mediated TAD boundaries as it can interact with a potent neighbouring enhancer.

In summary, alterations to genes involved in epigenetic regulation play significant roles in tumour development by changing the epigenetic state of the cell. These can involve an increase in epigenetic restriction leading to a maturation block and cells being fixed in a proliferative state. It can also involve an increase in epigenetic plasticity which allows the cancer cell to sample states advantageous to its development. Epigenetic alterations may, in fact, be involved in every hallmark of cancer (Hanahan and Weinberg 2011).

1.3.4 Epigenetic datasets of normal tissues and tumours

Based on the increasing breadth of knowledge about the role the epigenome has in normal biology as well as many diseases, epigenomic mapping of human tissues has become paramount. Normal tissue epigenomes have been thoroughly characterized by the NIH Roadmap Epigenomics Mapping Consortium as well as the Encyclopedia of DNA Elements (ENCODE) project. Both groups leverage next-generation sequencing to map DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in human stem cells and primary ex vivo tissues (Roadmap Epigenomics Consortium et al. 2015). In 2015, these groups released a public dataset consisting 127 human cell epigenomes from various tissues (5 of which were cancer cell lines) which were mapped using ChIP-Seq of 5 core histone marks. A subset of these epigenomes have chromatin accessibility (DNase-Seq) and histone mark data.

Until recently, large-scale genome-wide epigenomic mapping of primary tumours from multiple cancer types was not possible due to technical limitations. A study published in 2018 performed Assay for Transposase-Accessible Chromatin using sequencing (ATAC-Seq) experiments on 410 primary tumour samples from 23 different cancer types to map their genome-wide chromatin accessibility landscape (Corces et al. 2018). These tumours were derived from The Cancer Genome Atlas (TCGA) dataset which is a landmark cancer genomics program that molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. The TCGA ATAC-Seq study opened further

possibilities for cancer epigenomics research. This study has revealed that chromatin accessibility peaks are highly cancer type-specific and reveal new functional elements within the cancer genome.

1.4 Random forest method for evaluating the relationship between the epigenome and the cancer genome

1.4.1 Cell-of-origin epigenome defines the mutational landscape of cancer

A study published in *Nature* in 2015 by Polak et al. assessed how well regional mutation rates from various cancer types can be predicted by epigenomic data derived from various primary cells and cell lines (Polak et al. 2015). Firstly, they found that the most predictive epigenome of tumour regional mutation rates comes from the same or similar healthy tissue as the origin of the tumour (*i.e.*, the melanocyte epigenome best predicts melanoma mutation rates as opposed to epigenomes derived from other cells and tissues in the body). Secondly, they found that epigenomes derived from healthy cells of the same tissue better predicted cancer mutation rates than epigenomes derived from cancer cell lines of the same cancer type. The authors concluded that one explanation is that most mutations in cancer are established prior to oncogenesis as they are most associated with the epigenome of the normal cells from which the tumour originated. There are several problems with their second finding and its subsequent conclusions, however.

To test the predictive power of healthy vs. cancer cell epigenomes, they only had access to epigenomes derived from 2 cancer cell lines: melanoma (COLO829) and hepatocellular carcinoma (HEPG2). In addition to the small sample size of cancer cell lines tested, the epigenetics of cultured cell lines can be vastly different from human tissues, normal or tumour. Therefore, extending their hypothesis to primary tumour epigenomes is potentially problematic.

Furthermore, using the machine-learning method of random forests, they were only able to demonstrate that normal liver epigenomic features are more predictive of liver cancer mutation rates than those derived from the liver cancer cell line. For the case of melanoma, they only demonstrated that normal melanocyte epigenomic features were more correlated with melanocyte mutation rates. The random forest prediction accuracies are not shown for melanoma mutation rates in their study. Therefore, these conclusions need to be verified using more cancer types and epigenomes derived from tumour cells.

1.4.2 Tumour mutational landscape is a record of its premalignant state

The group that created the paper in section **1.4.1** later applied their methodology to a larger set of normal tissue epigenomic datasets as well as a larger set of mutational profiles (Kubler et al., 2019). They showed that epigenetic data from the tissue of origin was the best predictor of a cancer-type's mutational rates in 23/32 cancer types tested. They concluded that the mutational landscape of a cancer cell holds the memory of its cell lineage and the cell of origin. They extended their analysis to metastatic tumour samples and found that they could predict the cell of origin by associating the mutational rates of these tumour samples with epigenomic data from normal tissues. They also showed that the mutational landscapes of different subtypes within the same cancer type were best associated with epigenomic tracks from different tissues implying different cells of origin. Finally, they showed that driver mutations were significantly more likely to occur in open chromatin regions of the tumour's cell of origin as the genes in these open chromatin regions tend to play an important functional role. The use of normal cell epigenomes is again a limitation of this study because cancer cells are expected to change their epigenetic state upon transformation and chromatin state is a major correlate of mutation rates at the megabase-scale.

Chapter 2

Hypothesis and research aims

2 Hypothesis and research aims

2.1 Overarching hypothesis

Previous results have demonstrated a strong association between the chromatin landscape of the tumour "cell-of-origin" and the tumour mutational landscape when considering large megabase-scale genomic windows. These studies included chromatin tracks only from normal cells and cancer cell lines and had a more limited WGS mutation dataset. **We hypothesize that regional mutation rates in cancer genomes have a stronger association with chromatin states of tumor tissues rather than normal tissues and cell lines. Therefore, the systematic evaluation of the interactions of genome-wide mutation rates and the chromatin landscapes of tumours and normal cells will reveal insights into tumour evolution, mutational processes, and cancer tissue of origin.**

2.2 Specific research aims

To investigate this hypothesis, we use a machine learning approach to associate genome-wide distributions of cancer mutations with chromatin landscapes. We study 25 types of cancer (plus pan-cancer) with WGS data available to us, 382 chromatin tracks derived from primary tumours, and 53 chromatin tracks derived from normal cells and cancer cell lines. Our project comprises the following two major aims:

Aim 1: The first aim of this project is to evaluate whether each of our 26 cancer-type specific mutational rate tracks are better represented by the chromatin landscapes of tumor cells or normal cells. This allows us to infer the timing of the accumulation of mutations in each of these cancer types and provides us insight into the processes of

mutagenesis and oncogenesis. We can provide further insight into previous work done in the field.

Aim 2: The second aim of this project is to uncover the association between specific mutagenic processes occurring in cancer cells and the chromatin landscape. Using mutations derived from specific mutational signatures, we can evaluate whether the genomic location of mutations due to specific mutational processes are associated with the chromatin landscape of normal or cancer cells.

Chapter 3

Materials and methods

3 Materials and methods

3.1 Overview of approach

To evaluate the relationship between the epigenomic landscape of human cells and the mutational landscape of cancer at the megabase-scale, we used the machine learning method of random forests. The predictor variables to these random forest models were megabase-scale chromatin accessibility tracks from various human normal and cancer cells. These chromatin accessibility tracks were composed of an average chromatin accessibility score (derived from DNase-Seq and ATAC-Seq experiments) for each megabase window in the human genome. The response variable for this model was a cancer-type specific regional mutation rate track. Mutation rate tracks included mutations summed from every sample in the cancer WGS cohort for each megabase window in the human genome. This workflow is demonstrated in **Figure 1.**

Next, we describe the chromatin accessibility and mutation data used in this project and then describe our machine learning approach.

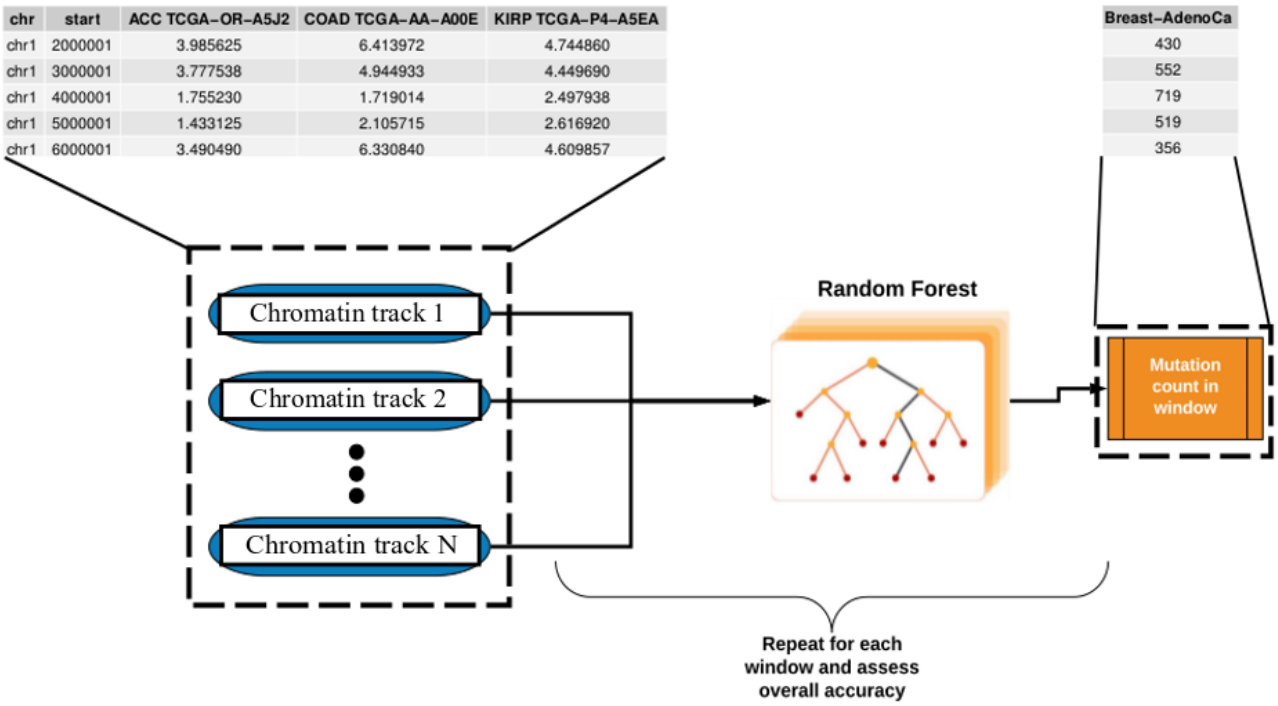


Figure 1: Overview of machine learning workflow. Chromatin accessibility tracks derived from primary tumours, normal cells, and cancer cell lines were split into genomic windows. Each of these chromatin tracks serves as a predictor in our model and each genomic window serves as a sample in our dataset. Within each window, our random forest model uses the chromatin accessibility scores from various tracks to predict total mutations from a cancer cohort within that window. This process is repeated for all genomic windows. Chromatin accessibility within a window is highly associated with mutation count.

3.2 Description of genomic and epigenomic datasets used in this project

3.2.1 Chromatin accessibility tracks from normal tissues and cancer cell lines

Chromatin accessibility tracks were derived from the Roadmap Epigenomics Project, 2015 public data release (Roadmap Epigenomics Consortium et al 2015). This study collected DNase-Seq chromatin accessibility profiles of 53 human cell epigenomes. These were derived from tissues all around the body, multiple cell types, as well as 4 cancer cell lines. The tissues of origin of these chromatin tracks are summarized in **Table 2.**

Genome-wide DNase-Seq tracks had a chromatin accessibility score for every base pair along the human genome (build hg19). These scores were lifted over to the latest human genome build, hg38. The data was then processed into megabase windows representing the average chromatin accessibility score within that window for each chromatin accessibility track. Briefly, DNase-Seq involves the genome-wide mapping of regions sensitive to cleavage by the DNase 1 enzyme. More accessible DNA will be more prone to cleavage by DNase 1 and these cleavage sites are ligated and tagged with a ligand. DNA sequences are amplified using polymerase chain reaction (PCR) and sequenced using next-generation sequencing (NGS). The presence of tags represents a cleavage site and many of these cleavage sites within a specific region represents a highly accessible regulatory element (Song and Crawford 2010).

Table 2: Tissue of origin of DNase-Seq chromatin accessibility tracks from the Roadmap Epigenomics Project

Cell Origin	Number of Samples
Blood	10
Skin	6
Embryonic stem cell derived	4
Lung	4
Muscle	4
Brain	3
Intestine	3
Breast	2
Embryonic stem cell	2
Stomach	2
Induced pluripotent stem cell	2
Adrenal gland	1
Cervix	1
Heart	1
Kidney	1
Liver	1
Leg muscle	1
Ovary	1
Pancreas	1
Placenta	1
Thymus	1
Vascular	1
Total	53

3.2.2 Chromatin accessibility tracks from primary tumours spanning 23 cancer types

TCGA, a major cancer genomics program, molecularly characterized over 10,000 primary cancer samples spanning 33 cancer types (Hoadley et al. 2018). Molecular data for these samples include mapping of the exome, transcriptome, and methylome. In 2018, the first chromatin accessibility study of primary tumours was released comprising ATAC-Seq performed on 410 primary tumour samples spanning 23 cancer types from the TCGA dataset (Corces et al. 2018). The number of samples in each cancer type and their abbreviations are summarized in **Table 3.**

The genome-wide ATAC-Seq tracks from TCGA include an ATAC-Seq insertion score as a proxy for chromatin accessibility for every 100 base pairs in the human genome (hg38 build). ATAC-Seq was performed on 410 tumour samples derived from 404 donors. Two technical replicates were done for 386 out of the 410 samples yielding 796 genome-wide ATAC-Seq tracks (Corces et al. 2018). We processed each track to provide an average ATAC-Seq score for every million base pairs (megabase-pair scale). For each donor, multiple tracks (some donors had multiple samples) were averaged and donors with only single replicates were discarded due to lower confidence in the quality of the sample. This resulted in a megabase-scale tumour chromatin accessibility track for 382 donors from 23 cancer types.

Table 3: Study names, abbreviations, and number of samples for the TCGA ATAC-Seq primary tumour chromatin accessibility dataset.

Study Name	Study Abbreviation	Number of samples
Breast invasive carcinoma	BRCA	66
Colon adenocarcinoma	COAD	37
Kidney renal papillary cell carcinoma	KIRP	34
Prostate adenocarcinoma	PRAD	26
Lung adenocarcinoma	LUAD	22
Stomach adenocarcinoma	STAD	20
Liver hepatocellular carcinoma	LIHC	17
Kidney renal clear cell carcinoma	KIRC	16
Lung squamous cell carcinoma	LUSC	16
Esophageal carcinoma	ESCA	15
Thyroid carcinoma	THCA	14
Brain Lower Grade Glioma	LGG	11
Skin Cutaneous Melanoma	SKCM	11
Uterine Corpus Endometrial Carcinoma	UCEC	11
Bladder Urothelial Carcinoma	BLCA	10
Adrenocortical carcinoma	ACC	9
Glioblastoma multiforme	GBM	9
Pheochromocytoma and Paraganglioma	PCPG	9
Testicular Germ Cell Tumors	TGCT	9
Head and Neck squamous cell carcinoma	HNSC	7
Mesothelioma	MESO	6
Cholangiocarcinoma	CHOL	5
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	2
TOTAL		382

3.2.3 Catalogue of genome-wide somatic mutations spanning 37 cancer types

To characterize megabase-scale regional mutation rates, we used the PCAWG whole genome mutation dataset. The PCAWG dataset comprises a large and uniformly processed whole genome somatic mutation dataset with whole genome sequencing performed on 2583 tumour samples (Campbell et al. 2020). Their processing of PCAWG involved tumour genomes being compared to the genome of a matched normal from either tumour-adjacent normal tissue, blood, or other skin/lymph tissue to exclude germline variants. Overall, samples were taken from 37 distinct tumour types with over 46 million SNV's called. Tumour types showed considerable differences in mutation burden per patient: from a median of 169 mutations/patient in pilocytic astrocytoma to a median of 70,873 mutations/patient in melanoma.

During our processing, the PCAWG mutations file (MAF) was lifted over from the hg19 build to hg38 to match our chromatin accessibility datasets. We only considered cohorts with greater than 30 patients which was 25 cancer types out of 37. We also considered a pan-cancer dataset consisting of mutations from all 37 cancer types meaning that we analysed 26 regional mutation tracks in total during this study. 66 hypermutated samples (mostly from melanoma), defined as having greater than 90,000 total mutations, were excluded to avoid biasing our mutation rate tracks towards a minority of tumour samples. We were interested in regional mutation rates, which represent an aggregated mutation burden per megabase window within a cohort. Regional mutational rate tracks were defined for each cancer type by summing all the SNV mutations from all the samples within each respective cohort for each megabase window in the genome. Cancer type and sample size data for our processed WGS dataset is shown in **Table 4**.

Table 4: Study names, project codes, and number of samples for the PCAWG whole-genomic sequencing dataset

Study Name	Project Code	Number of samples	Number of mutations	Included/Excluded
Liver Hepatocellular carcinoma	Liver-HCC	314	3764099	Included
Pancreas Adenocarcinoma	Panc-AdenoCA	232	1492798	Included
Prostate adenocarcinoma	Prost-AdenoCA	199	637414	Included
Breast Adenocarcinoma	Breast-AdenoCa	195	1396733	Included
Kidney Renal cell carcinoma	Kidney-RCC	143	897708	Included
Medulloblastoma	CNS-Medullo	141	199596	Included
Ovary Adenocarcinoma	Ovary-AdenoCA	110	965271	Included
Diffuse large B-cell lymphoma	Lymph-BNHL	104	1130214	Included
Esophageal Adenocarcinoma	Eso-AdenoCa	95	2465375	Included
Chronic lymphocytic leukemia	Lymph-CLL	90	217704	Included
Pilocytic Astrocytoma	CNS-PiloAstro	89	22005	Included
Pancreatic Neuroendocrine tumour	Panc-Endocrine	81	252855	Included
Melanoma	Stomach-AdenoCA	66	1048100	Included
Stomach Adenocarcinoma	Skin-Melanoma	65	2274735	Included
Head/Neck Squamous cell carcinoma	Head-SCC	56	876369	Included
Thyroid Adenocarcinoma	Thy-AdenoCA	48	64943	Included
Lung Squamous cell carinoma	Lung-SCC	45	1824949	Included
Colorectal Adenocarcinoma	ColoRect-AdenoCA	43	664599	Included
Kidney Renal cell carcinoma, chromophobe type	Kidney-ChRCC	43	78318	Included
Uterus Adenocarcinoma	Uterus-AdenoCA	42	487894	Included
Bone Osteosarcoma	Bone-Osteosarc	41	157718	Included
Glioblastoma multiforme	CNS-GBM	38	263689	Included
Bone Leiomyosarcoma	Bone-Leiomyo	34	183141	Included
Lung Adenocarcinoma	Lung-AdenoCA	33	725328	Included
Biliary Adenocarcinoma	Biliary-AdenoCA	33	243049	Included
Bladder Transitional cell carcinoma	Bladder-TCC	23	504886	Excluded
Myeloid Myeloproliferative neoplasm	Myeloid-MPN	23	24171	Excluded
Oligodendroglioma	CNS-Oligo	18	48178	Excluded
Cervix Squamous cell carcinoma	Cervix-SCC	18	114186	Excluded
Myeloid Acute myeloid leukemia	Myeloid-AML	13	19265	Excluded
Breast Lobular carcinoma	Breast-LobularCa	13	96983	Excluded
Bone neoplasm, epithelioid	Bone-Epith	11	23424	Excluded
Chondroblastoma	Bone-Cart	9	7846	Excluded
Breast In situ adenocarcinoma	Breast-DCIS	3	6012	Excluded
Lymphoid (Not otherwise specified)	Lymph-NOS	2	26272	Excluded
Cervix Adenocarcinoma	Cervix-AdenoCA	2	8149	Excluded
Myelodysplastic syndrome	Myeloid-MDS	2	1624	Excluded
TOTAL	PANCAN	2517	23215600	

3.2.4 Excluding non-mappable regions of the genome

Certain genomic regions are challenging and error-prone to map to using short read sequencing due to repetitive elements. For these cases, reads can map correctly to multiple genomic regions. These genomic regions show significant loss of signal in our chromatin accessibility and mutation datasets due to their high uncertainty. To address this problem, we used the UMAP tool which provides the coordinates of mappable regions in the genome (Karimzadeh et al. 2018). For our megabase-scale datasets, genomic windows were removed if over 20% of the window was predicted to be unmappable according to UMAP. The datasets originally had 2887 megabase windows and were processed down to 2465 mappable windows.

3.3 Training random forest models

3.3.1 Random forest machine learning method overview

Random forest is a machine learning method used for classification and regression problems that utilizes an ensemble of decision trees (Ho 1995; Ho 1998). A decision tree involves having a series of decision nodes which use input variables to make a decision. This decision can be a classification or regression based on whether the target variable is discrete or continuous, respectively.

A random forest combines many of these decision trees while adding two elements of randomness to the procedure (Amit and Geman 1997). The first element involves a procedure called bootstrap aggregation, or “bagging”. Bagging denotes randomly sampling observations with replacement N times from the dataset to train each tree. The predictions made by all the trees are then aggregated: either averaged in the case of regression or used for majority voting in the case of classification. The second element involves “feature bagging” which means that at every node or decision point, the model can only choose from a random subset of the input features. For our models, we used 250 trees for each forest while the maximum number predictors sampled at each node was the total number of predictors divided by 3.

Aggregation is the major advantage of random forests. Single decision trees that are deep (i.e. have many nodes/decisions) tend to overfit their training sets by learning irregular patterns, mainly noise in the data (Kleinberg 2000). This means that they are not generalizable to new data and serve little use in most applications. By introducing an ensemble of multiple decision trees which are decorrelated by the two steps introducing randomness, the final model is generalizable to new data.

Random forest models have been used extensively in biomedical and bioinformatic research. One important case is the Polak et al., 2015 study where random forest models were used to analyze epigenomic data as predictors of regional mutation rates in cancer. As another example, a 2019 study used random forests to predict drug response in cancer patients using their mutational status of 145 oncogenes (Lind & Anderson 2019). They were able to predict whether a cancer patient would respond to various chemotherapy agents with 87% sensitivity and 87% specificity. In another 2019 study, DNA methylation was used to predict whether prostate tumours would progress to aggressive stages with an AUC (area under the receiver operating characteristic curve) accuracy score of 95% (Toth et al. 2019).

3.3.2 Monte-Carlo cross-validation

Cross-validation is a method to assess how a model will generalize to an independent and previously unseen data set. During cross-validation, the original dataset is first split into training and testing tests (typically 70-80% training set), after which the model is trained on the training set and tested on the testing set. The accuracy of the model regarding the test set is then recorded, and the process is repeated with a different split of the data.

Monte-Carlo cross-validation refers to the method of randomly splitting the dataset into training and testing sets during each round of cross-validation (Dubitzky, Granzow, and Berrar 2007). This is different to K-fold cross-validation which involves splitting the dataset into K equally sized groups after which one group is kept as the testing set and all others as the training set (Hastie, Tibshirani, and Friedman 2009). This process is repeated K times with each group as the testing set once. K-fold cross-validation provides a nearly unbiased

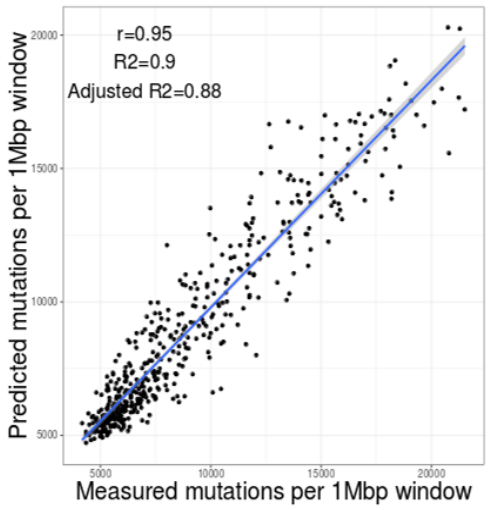
estimate of the model's performance as each data point is in the testing set exactly once. This contrasts with Monte-Carlo cross-validation, in which the same data point can be in the testing set in multiple runs. However, K-fold cross-validation test performance is highly variable relative to Monte-Carlo cross-validation, as all the testing sets are independent. Monte-Carlo cross-validation also provides the ability to test many different combinations of training and testing sets as is computationally practical (Arlot & Celisse 2010).

3.3.3 Assessing model accuracy

R2, also known as "percent variance explained", represents the percent of variance in the response which is explained by the model. R2 is the most common accuracy metric used for random forest regression models. R2 is calculated by taking the Pearson correlation between the predicted and observed values for each observation in the test set, squaring it, and then multiplying by 100%.

R2 tends to increase automatically and spuriously as extra predictors are added to the model. This is because more complex models have more capacity to explain the data by chance and to explain noise in the data. The adjusted R2 metric is an extension of the R2 metric that accounts for the complexity of the model by adjusting the R2 score based on the number of predictors (p) in the model relative to the number of observations (n) **(Fig. 2)** (Shieh 2008). Unlike R2, adjusted R2 increases only when the increase in R2 (due to the inclusion of a new predictor) is more than one would expect to see by chance. Adjusted R2 allows us to make better calibrated comparisons of the accuracy measures of models with differing numbers of predictors.

A



B

$$R^2adj = 1 - \frac{(1 - R^2)(n - 1)}{n - m - 1}$$

where n is the number of samples and
m is the number of independent variables

Figure 2: R2 and Adjusted R2 scores for model accuracy. A) R2 score (or % variance explained) is defined as the square of the Pearson correlation coefficient (r) between the predicted and observed response values. **B)** Adjusted R2 accounts for overfitting due to model complexity by penalizing models with higher number of predictors trained on lower number of samples.

3.3.4 Using random forest models to analyse predictor importance

A predictor's importance metric in a random forest model is calculated by permuting all its values in a test dataset and then making predictions based on that test dataset. The percent increase in the model's mean-squared-error (%IncMSE) after the permutation denotes how important that predictor is.

Model interpretability is a significant advantage of using random forest models compared to more advanced models. One aspect of this interpretability is the ability to calculate how important each predictor is to making predictors on a previously unseen dataset.

3.4 Using chromatin accessibility to predict regional mutation rates in cancer

3.4.1 Assessing the correlation between our chromatin accessibility tracks and cancer type-specific mutation tracks

We used a Spearman correlation between all our chromatin accessibility tracks and all our cancer-type specific regional mutation rate tracks (including pan-cancer). This was an initial, exploratory step to map the associations of chromatin tracks and the mutation frequencies in different cancer types. To account for potential nonlinear associations of mutation rates and chromatin accessibility, the Spearman correlation method, as a non-parametric ranked correlation, addresses nonlinearity more efficiently compared to the Pearson correlation method.

3.4.2 Comparing the predictive power of normal and tumour epigenomes in predicting cancer type-specific mutation tracks

To compare the capacity of normal and tumour epigenomes to predict mutation rates, we trained a random forest model on all our DNase-Seq megabase-resolution tracks derived from normal cells. We also trained a random forest model on all our primary tumour ATAC-Seq megabase-resolution tracks to predict regional mutational rates. All 25 cancer types plus pan-cancer were tested separately and 52 models were run in total. For each of those models, we performed 1000 Monte-Carlo cross-validations using an 80%/20% train/test split and then calculated an accuracy score based on the median Adjusted R2 from the 1000 splits. For each cancer type, we compared the accuracy of the primary tumour model and the accuracy of the normal cell model.

Interestingly, when examining the accuracy scores across the cross-validations, models predicting regional mutation rates in chronic lymphocytic leukemia and B-cell non-Hodgkin's lymphoma showed a high variability in adjusted R2 accuracy metric across cross-validations. Specifically, there is a subset of cross-validation runs which show significantly decreased accuracy. To understand this phenomenon, we examined those cross-validation

runs which were affected and found two specific megabase windows that had over a ten-fold increase in mutations over the median window. When one of or both hypermutated windows were part of the randomly sampled testing set, the model showed significantly decreased testing accuracy. Interestingly, the two hypermutated windows are in the same genomic location in both of our lymphoma and leukemia results (**Fig. 3)**.

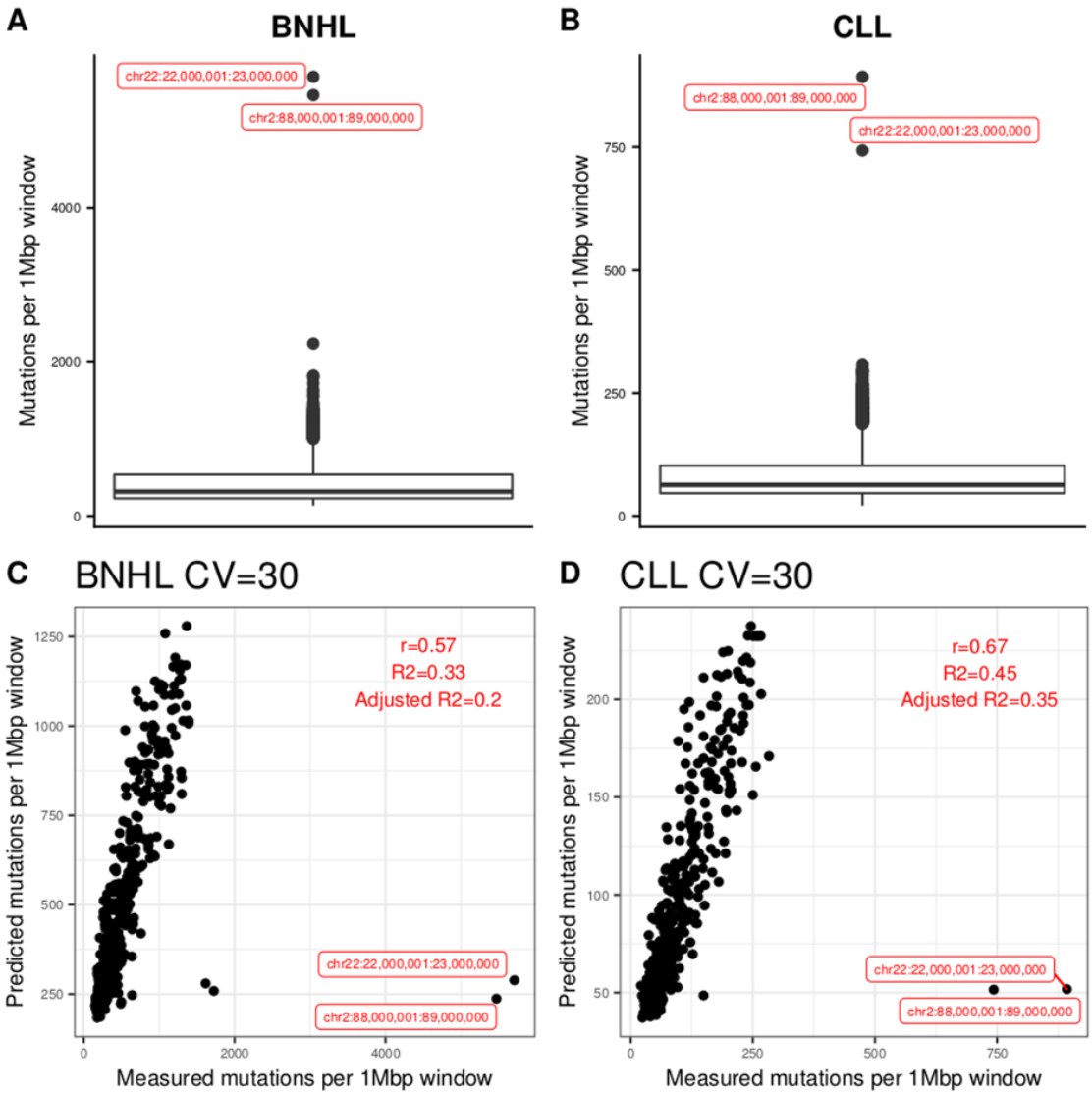


Figure 3: Hypermutated windows in lymphoma and leukemia negatively affect model performance. The same two hypermutated windows show over a 10-fold increase in mutations than the median window in **A)** lymphoma and **B)** leukemia. **C/D)** The inclusion of these windows in the test set during cross-validation results in severely decreased test adjusted R2.

Regional hypermutation has been reported in normal blood cells. Hypermutation in the variable region of immunoglobulin genes has been well-documented as a means of diversifying the immune system's antibodies to be able to bind to a broader set of antigens (Diaz & Casali 2002). The mistargeting of somatic hypermutation has been shown to target oncogenes and is thought to contribute to B-cell related blood cancers (Odegard & Schatz 2006). Interestingly, when examining the 2 hypermutated windows, we found that both windows contained the two genetic loci which code for immunoglobulin light chain genes in humans found on chromosomes 2 and 22. Upon examining the hypermutated region on chromosome 2 more closely, we found that there is a localized enrichment in mutations in the immunoglobulin kappa joining cluster (*IGKJ1-5*) as well as another enrichment ~500 base-pairs downstream not overlapping any genes in both leukemia and lymphoma. The hypermutated region on chromosome 22 showed multiple regional enrichments in mutations overlapping the immunoglobulin lambda gene cluster. Specifically, we observed an enrichment in mutations overlapping the *IGLV3-1* and *IGLL5* genes in both leukemia and lymphoma as well as *IGLC3* in leukemia (**Fig. 4)**.

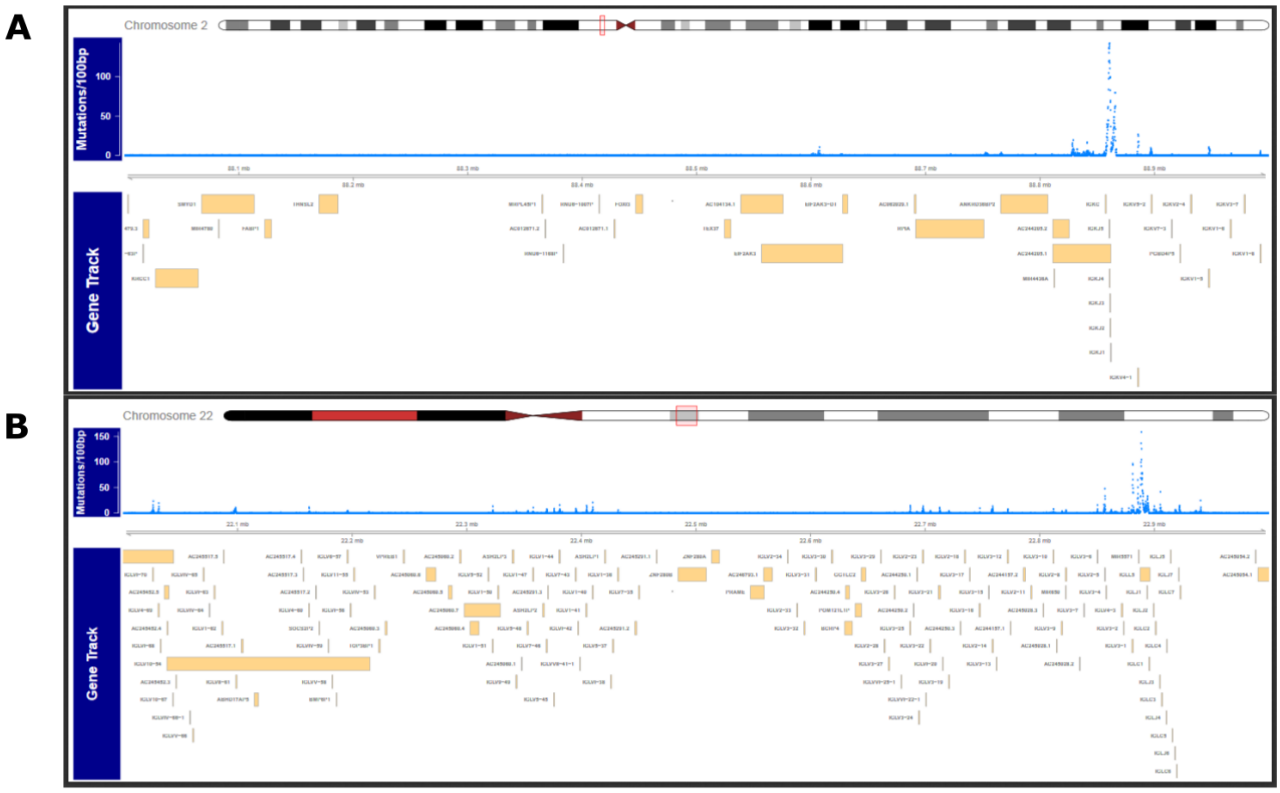


Figure 4: Hypermutated regions in lymphoma overlap the human immunoglobulin light chain genes. A) Lymphoma mutation density plot is shown along with overlapping genes in the hypermutated megabase window on chromosome 2. We see an enrichment of mutations overlapping the immunoglobulin kappa joining cluster. **B)** Lymphoma mutation density plot is shown along with overlapping genes in the hypermutated megabase window in chromosome 22. We see an enrichment of mutations overlapping the immunoglobulin lambda gene cluster.

To confirm and avoid the effect of these hypermutated windows, the random forest experiment was repeated in leukemia and lymphoma after excluding these 2 windows. The exclusion of the 2 hypermutated windows led to significantly increased model accuracies and decreased variance in accuracy (**Fig. 5)**, suggesting that the random forest method may be sensitive to few outlier samples with high deviation from the values observed in other samples. Excluding the outlier genomic windows leads to a more conservative analysis throughout this study.

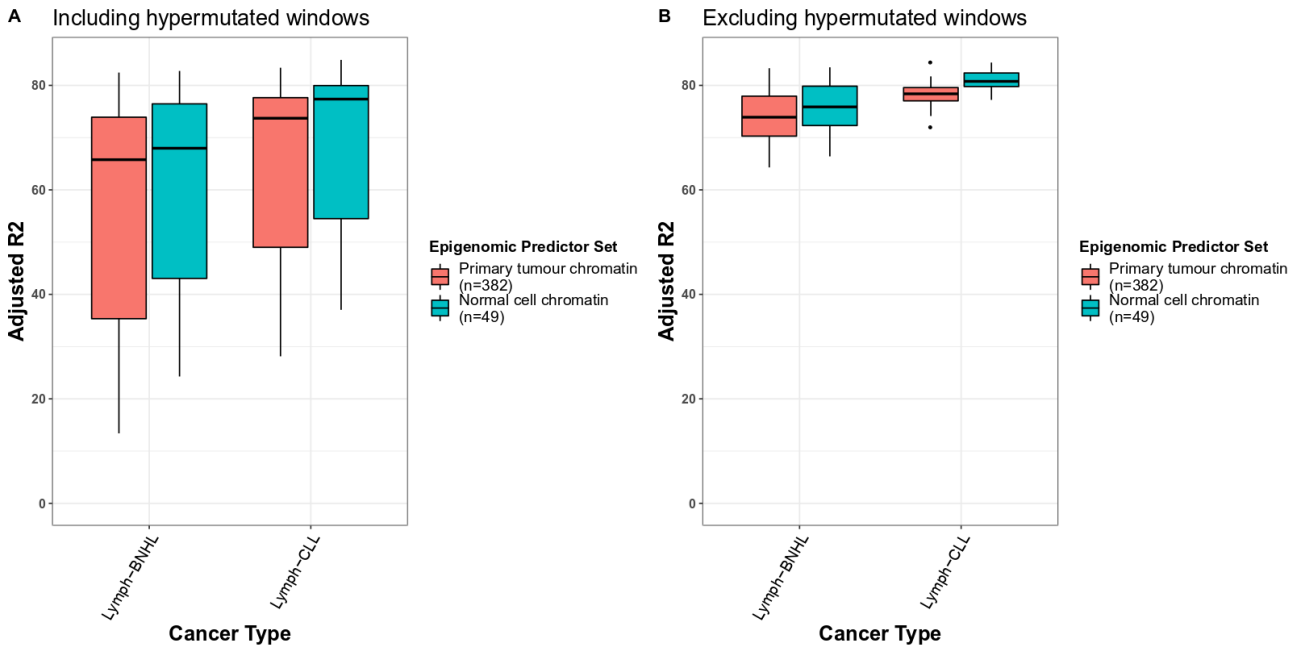


Figure 5: Excluding two hypermutated windows in leukemia and lymphoma resolves model performance and variability A) Model performance before exclusion of hypermutated windows shows high variability due to hypermutated windows being included in the test set during cross-validation. **B)** Model performance after exclusion of the two hypermutated windows shows increased performance and decreased variability.

3.4.3 Training models on chromatin tracks derived only from matching normal/tumour tissue

We examined the nine cancer types with available chromatin accessibility tracks derived from both normal and primary tumours of the same tissue. These nine cancer types included breast adenocarcinoma, glioblastoma multiforme, colorectal adenocarcinoma, renal cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, melanoma, stomach adenocarcinoma, and uterus adenocarcinoma. As an exception, uterus adenocarcinoma had no precisely matching chromatin accessibility track available and a cervical cancer cell line was used instead. We randomly sampled one primary tumour chromatin accessibility track from each cancer type we were examining during each cross-validation run (1000 runs, 80/20 train/test split). This approach was used because there are considerably more predictors from primary tumours than normal cells and this sampling method allows us to account for the bias of comparing two sets of predictors of vastly different sizes.

In conclusion, each of our nine random forest models were trained on the same 18 predictors, nine of which were derived from matching DNase-Seq (normal cells with one exception) tracks and nine of which were derived from matching ATAC-Seq (primary tumours) tracks. Therefore, for each model, each of the nine cancer types had one matching normal predictor and one matching primary tumour predictor.

3.4.4 Training models on all tumour chromatin tracks and analyzing predictor importance

For this experiment, we trained 26 different random forest models to predict 26 different regional mutation rate tracks (25 cancer types plus pan-cancer). Each model was trained on all 382 primary tumour chromatin accessibility tracks. We then analyzed which cancer types contributed the most predictive chromatin tracks by using the median importance score to compare the contributions of various cancer types for each cancer mutational landscape.

3.4.5 Training individual models trained on predictors derived from each cancer type

We used a complementary method to study the relationship between the primary tumour epigenome and mutational landscapes. In contrast to training models on all our primary tumour chromatin tracks as described above, we trained individual models on only primary tumour chromatin tracks derived from one tumour type. We then compared the accuracy of the models trained on the various cancer types to determine which epigenomic landscape is best associated with a cancer type's regional mutation rates.

3.4.6 Using chromatin tracks to predict mutations derived from specific mutational signatures

Here we used a precomputed dataset where each mutation from our WGS dataset was attributed to a specific mutational signature using the SigProfiler tool (Alexandrov et al. 2020). SigProfiler is a computational framework based on non-negative matrix factorization (NMF) used to 1) discover mutational signatures from mutational catalogues and 2) classify

mutations as one of these signatures. Single-base substitutions can be biologically categorized by their trinucleotide context which includes the base change itself (i.e. C>T) and two flanking nucleotides (i.e. ACG). SBS signatures consists of percentages for each substitution type; there are 96 possible combinations as there are 6 types of substitution (mapped to the pyrimidine base) x 4 types of 5' base x 4 types of 3' base. Mutational signatures are linear combinations of these mutation types with each coefficient describing the probability of that signature producing that mutation type.

We used this preprocessed dataset to create cancer type and mutations signature-specific regional mutation rate tracks using only mutations from specific cohorts and signatures. To analyze the relationship between the mutational processes associated with these signatures and the chromatin landscape, we used our total set of chromatin predictors to predict these mutation tracks. Cancer types were kept with >30 samples (same 25 cancer types plus pan-cancer) and mutational signatures were kept if they included greater than 3% of the total mutations in the cohort.

3.4.7 Examining the effect of using 100KB windows in the random forest experiments

Somatic mutation rates in cancer genomes show regional variation that is apparent at multiple scales of resolution. The major contribution of these regional mutation rates likely comes from functionally neutral passenger mutations since driver mutations only make up a small minority of all somatic mutations in a cancer genome (Vogelstein et al. 2013). To assess the impact of changing the spatial scales, we repeated our random forest modelling to compare the prediction accuracy of models trained on chromatin accessibility profiles of primary tumours and normal cells at the 100KB scale (window size). We found that the models trained on primary tumour chromatin tracks were more predictive of regional mutation rates in all 26 cancer types (including pan-cancer) than the models trained on normal cell chromatin tracks at the 100KB scale **(Fig. 6)**. We also found that the adjusted R2 accuracies of our models were higher at the megabase-scale in nearly every cancer type tested.

Megabase-scale regional mutation rates, as opposed to other resolutions of the genome, are associated with replication timing and chromatin accessibility (Supek & Lehner 2019). Furthermore, it makes random forest experiments more viable technically (less windows and therefore less computational training time) as well as more mutations per window making our analysis better powered. Our observed variations in megabase-scale mutation rates are consistent with earlier studies (Polak et al. 2015, Kubler et al., 2019). However, confirming these variations at a smaller scale demonstrates the robustness of our observations.

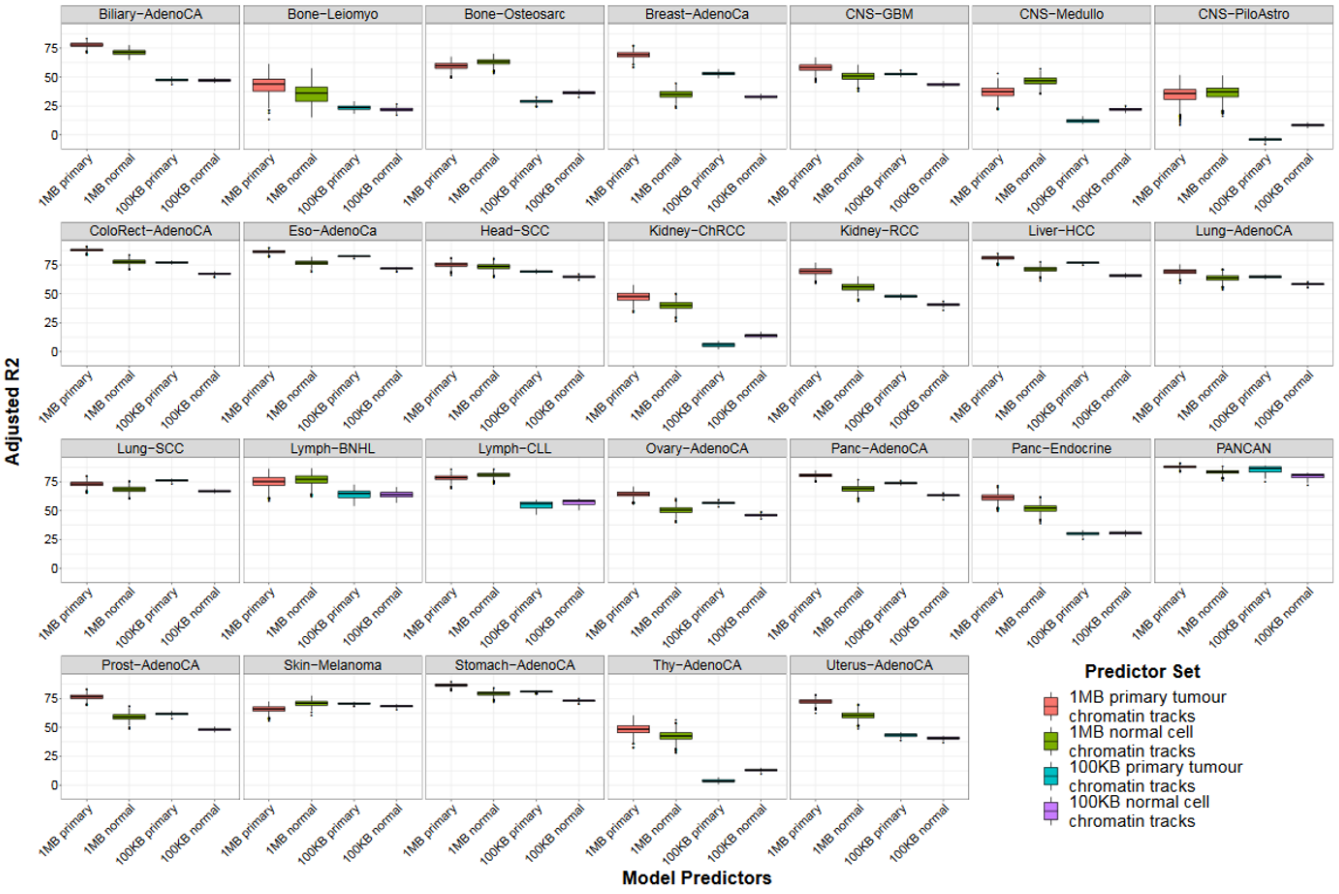


Figure 6: The relationship between the chromatin and mutational landscapes are concordant at the 100KB and the 1MB scale. Boxplots show predictive accuracy of models trained on chromatin tracks derived from primary tumours vs. normal cells at a megabase- and 100 kilobase-scale. Megabase-scale models are more accurate in predicting cancer regional mutation rates. Also, primary tumour chromatin tracks are more predictive of regional mutation rates than normal cell chromatin tracks at the 100 kilobase-scale.

3.4.8 Comparing the predictive power of histone marks to chromatin accessibility at predicting cancer regional mutation rates

As discussed in section **1.3.1**, histone mark tracks contain information complementary to chromatin accessibility about the epigenetic state of the cell. In Polak et al. 2015, histone mark tracks from various normal cells were used to predict the cell-of-origin of multiple tumours based on their capacity to predict regional mutation rates. For our analysis, however, no large-scale dataset containing ChIP-Seq of histone marks in primary tumours is available so we cannot use the existing ChIP-Seq of normal cells to address our hypothesis. We can, however, compare the predictive power of chromatin accessibility tracks to that of histone mark tracks to understand if histone marks contain additional information about the epigenome which can help us predict regional mutation rates. If so, repeating our analysis using histone mark data from primary tumours, once available, will provide us more insight into the cancer epigenome as well as the mutational landscape of cancer.

After training separate random forest models on DNase-Seq and the 5 core histone marks from the Roadmap Epigenomics Project from the same 53 tissues and cells, we compared their adjusted R2 accuracy scores in predicting regional mutation rates **(Fig. 7)**. We found that the models trained on the H3K9me3 mark showed modest increases in accuracy relative to the other models in several cancer types. Chromatin accessibility (DNase-Seq) showed negligible differences in terms of accuracy relative to the other models in most cancer types. Therefore, to complement the available chromatin accessibility tracks from primary tumours, we chose to focus on only chromatin accessibility from normal cells in this study.

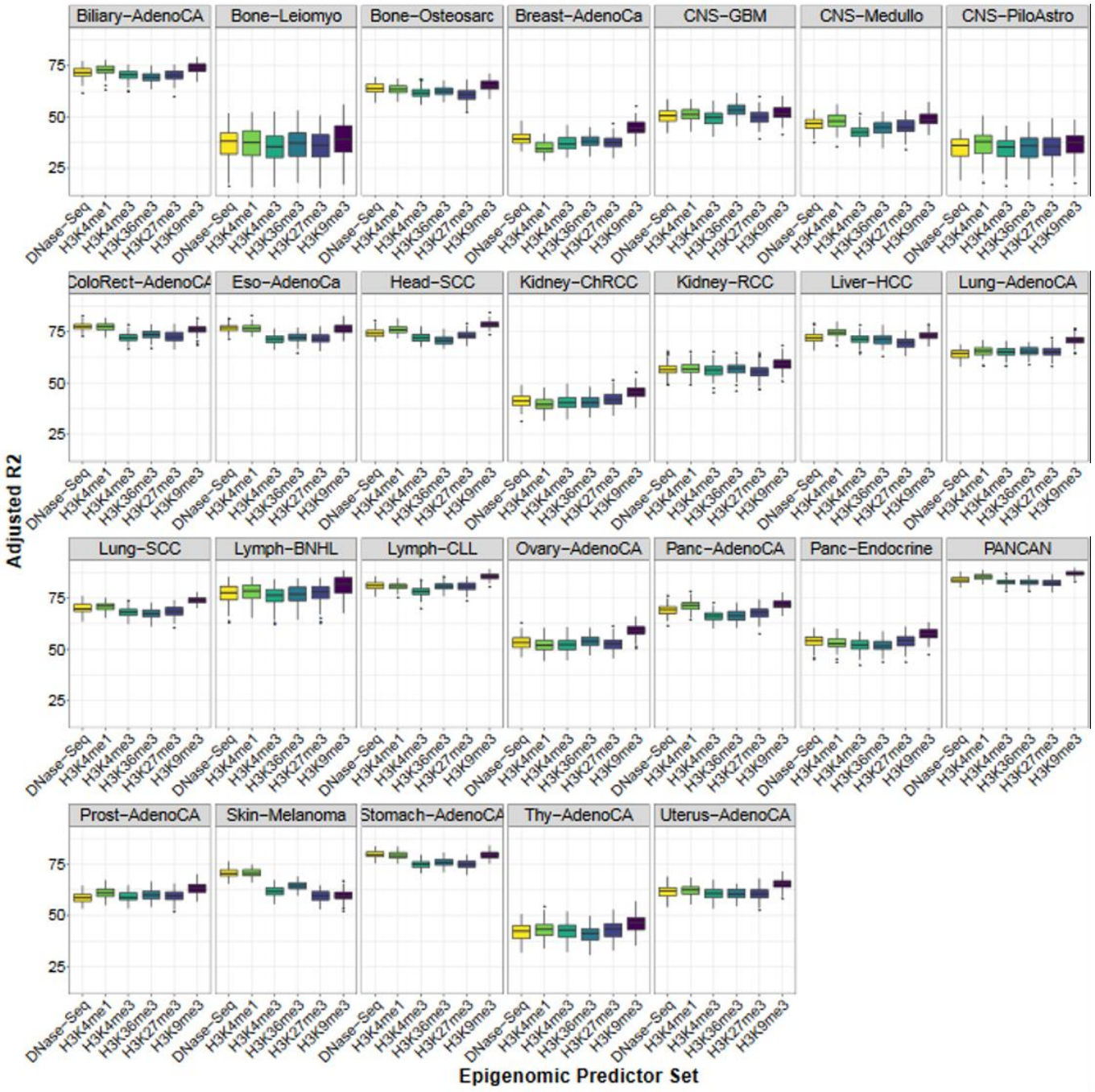


Figure 7: Boxplots of performances of models (1000 cross-validations) trained on various epigenomic tracks to predict regional mutation rates in 26 cancer types. Epigenomic tracks were derived from genome-wide DNase-Seq, H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3 mapping on the same 53 tissues/cells. Models trained on each of these predictor sets are shown on the x-axis and their test adjusted R2 scores (%) are shown on the y-axis.

Chapter 4

Results

4 Results

4.1 Correlating primary tumour chromatin accessibility and cancer regional mutation rates

4.1.1 Primary tumours demonstrate the classical negative relationship between chromatin accessibility and regional mutation rates

Chromatin accessibility tracks derived from normal cells and cancer cell lines were previously shown to be negatively correlated with regional mutation rates derived from cancer cells at a megabase-scale (Schuster-Böckler & Lehner 2012). We have confirmed that this relationship also exists when deriving chromatin accessibility from primary tumours. We tested the Spearman correlation of our megabase-scale primary tumour chromatin tracks with regional mutation rates derived from 26 different cancer types from PCAWG (including pan-cancer) **(Fig. 8a)**. We found that all of our primary tumour chromatin tracks had a negative correlation with the 26 different regional mutation rate tracks (**Fig. 8b)**, with the exception of one chromatin track from a low-grade glioma sample which showed a slight positive correlation (rho~0.3) (**Fig. 8c)**. It is unclear why this sample is an outlier, as the patient it is derived from has no exceptional clinical characteristics and may in fact be a technical artifact. These results demonstrate that the well-characterized negative association between chromatin accessibility and lower mutational density are also present in primary tumours.

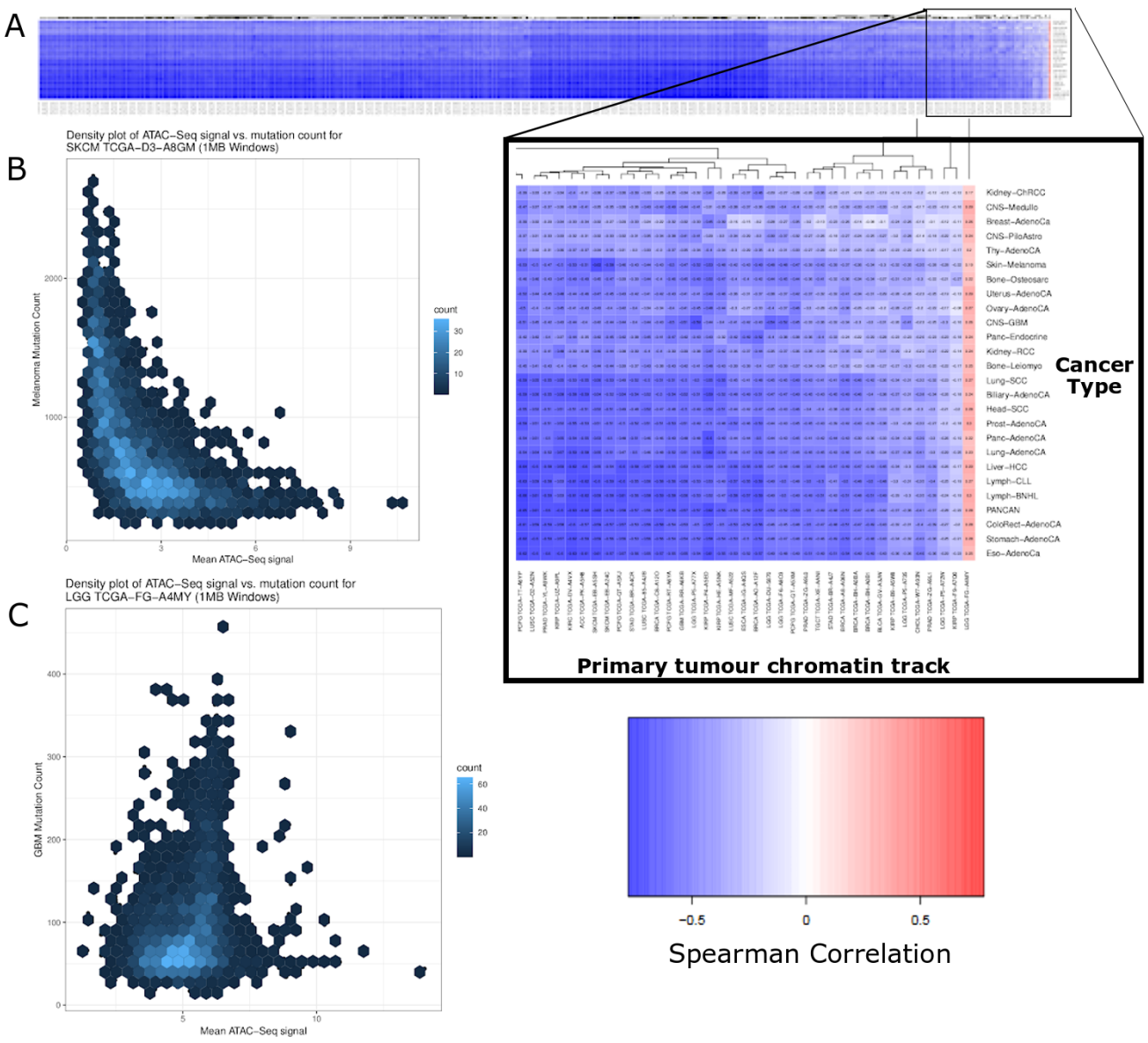


Figure 8: Consistent negative correlation between cancer mutations and primary tumour chromatin accessibility. A) Heatmap showing correlation between all our megabase-scale primary tumour chromatin tracks and megabase-scale regional mutation rates from 26 cancer cohorts including pan-cancer. **B)** Density plot showing megabase-scale melanoma cohort mutation count vs. chromatin acccessibility derived from a melanoma sample. **C)** Density plot showing megabase-scale GBM cohort mutation count vs. chromatin accessibility derived from outlier LGG sample. This is the only chromatin track that has a positive correlation with cancer mutation counts.

4.2 Tumour epigenomes are more predictive of regional mutation rates than epigenomes of normal cells in most cancer types

We trained random forest models on the 382 megabase-scale primary tumour chromatin accessibility tracks derived from TCGA ATAC-Seq. Models were used to predict megabase-scale mutation rates derived from 25 different cancer types and pan-cancer. We also trained random forest models on chromatin accessibility tracks derived from 53 normal cells (Roadmap Epigenomics Project) to predict mutation rates in the same 25 cancer types plus pan-cancer. Therefore, we have two models for each of our 26 regional mutation rate tracks, one trained on primary tumour and one trained on normal cell chromatin accessibility. By comparing the two models, we can describe whether the regional mutation rate of a specific cancer type is more associated with a normal or tumour epigenome.

4.2.1 Primary tumour chromatin tracks out-predict the regional mutation rates of most cancer types when compared to healthy cell lines

Polak et al. 2015 described that normal epigenomes define the mutational landscape of cancer. This study found that in 20/26 cancer types (including pan-cancer), the model trained on primary tumour chromatin accessibility outperformed the model trained on normal cell chromatin accessibility in predicting regional mutation rates in cancer. **(Fig. 9).** The accuracy measures of both the primary tumour and normal cell epigenome models were strongly correlated with the log-transformed average mutation burden in a cancer type as these datasets were better powered. This result indicates that our previous understanding of the mutational landscape of cancer needs refinement. Although as the chromatin profiles of normal cells are informative of regional mutation rates, the predictive power of tumor chromatin profiles is consistently larger. However, our analysis only indicates statistical significance suggesting biological associations and follow-up studies.

The 2 cancer types which were tested by Polak et al., 2015 were liver cancer and melanoma (in both cases, epigenomic data was derived from cancer cell lines). For Liver-HCC, we found that primary tumour chromatin tracks outperformed normal cell chromatin tracks in

predicting Liver-HCC regional mutation rates (Adj. R2= 0.82 and 0.71, respectively). For melanoma, however, we found that normal epigenomes better predicted mutation rates when compared to primary tumour epigenomes (Adj. R2= 0.75 and 0.71, respectively). Melanoma was again seen as an exception due to a strong association of mutation rates and chromatin accessibility of normal melanocytes.

The other five cancer types for which mutation rates were more strongly associated with the chromatin tracks of normal cells included chronic lymphocytic leukemia, bone osteosarcoma, B-cell non-Hodgkin's lymphoma, medulloblastoma, and pilocytic astrocytoma. Interestingly, none of these cancer types are carcinomas (which make up over 90% of cancer cases). The predictor set consisting of primary tumour chromatin tracks may be less predictive because it does not contain chromatin tracks from any blood cancers or sarcomas. Another possibility is that only carcinoma mutation rate tracks demonstrate a stronger association with primary tumour chromatin tracks because carcinomas are all derived from the same general cell type: epithelial cells. We know that the epigenome is the determinant of cell type and therefore other cancer cell types (sarcomas, blood cancer cells), may show a stronger association between their mutational landscapes and normal cell chromatin landscapes.

The limitation of this analysis is that only some cancer types that we tested had chromatin accessibility tracks derived from both matching primary tumour and normal cells. Most of these will have matching chromatin accessibility coming from only primary tumour or normal cells and some have no matching chromatin accessibility tracks whatsoever. We can address this by examining only the cancer types for which we have primary tumour chromatin accessibility, normal cell chromatin accessibility, and mutation rates.

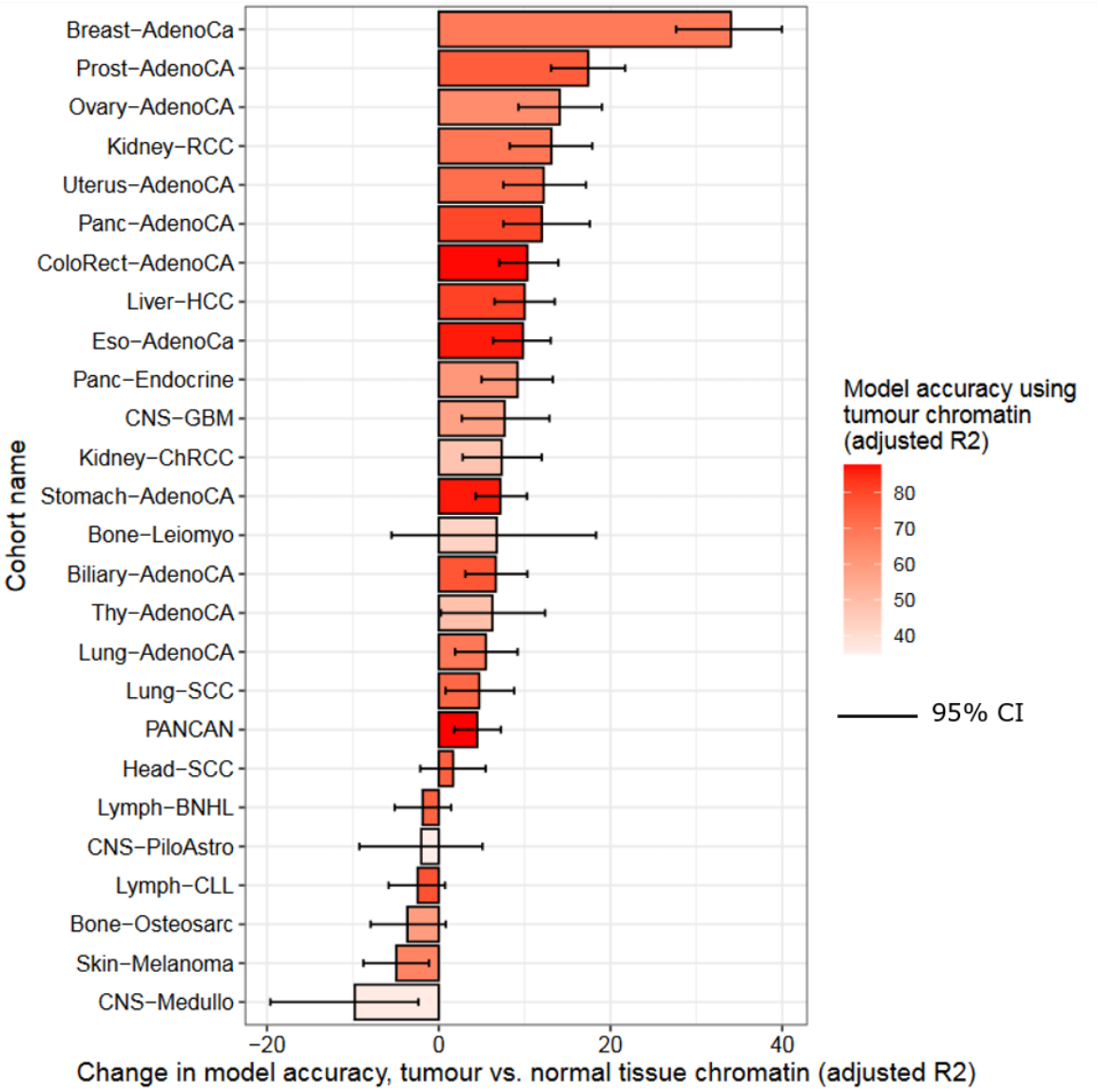


Figure 9: Primary tumour chromatin tracks better predict regional mutational rates in 20/26 cancer types than normal chromatin tracks. Barplots demonstrate the mean difference in adjusted R2 accuracies of random forest models trained on primary tumour chromatin tracks over normal chromatin tracks (1000 cross-validations) in predicting cancer regional mutation rates (cancer types shown on y-axis). Error bars represent 95% confidence intervals based on 1000 cross-validations.

4.3 Matched-tissue predictors reveal underlying relationships between the cancer genome and epigenome

There were eight cancer types for which we had datasets on primary tumour chromatin accessibility, normal cell chromatin accessibility, and regional mutation rates. An additional cancer type used in this analysis, uterine adenocarcinoma, also had a chromatin tracks from a cervical carcinoma cell line instead of a normal cell chromatin track (defined in section **3.4.3**). For this analysis, we developed a unified set of predictors on which we trained each model to predict regional mutation rates. This set of predictors included one predictor for the primary tumour and one predictor for the normal tissue or cancer cell line matching to each of the nine cancer types (18 predictors in total). As there were multiple primary tumour chromatin tracks for each cancer type, one track was randomly sampled during each Monte-Carlo cross-validation over a series of 1000 samplings. We used this strategy of random sampling of tracks, as opposed to choosing only one tumour from each cancer type or averaging multiple tumours, to account for the heterogeneity of tumour epigenomes from the same cancer type. Nine models were trained on these predictors to predict regional mutation rates in the nine cancer types and the top predictors were modeled using their importance metric **(Fig. 10)**.

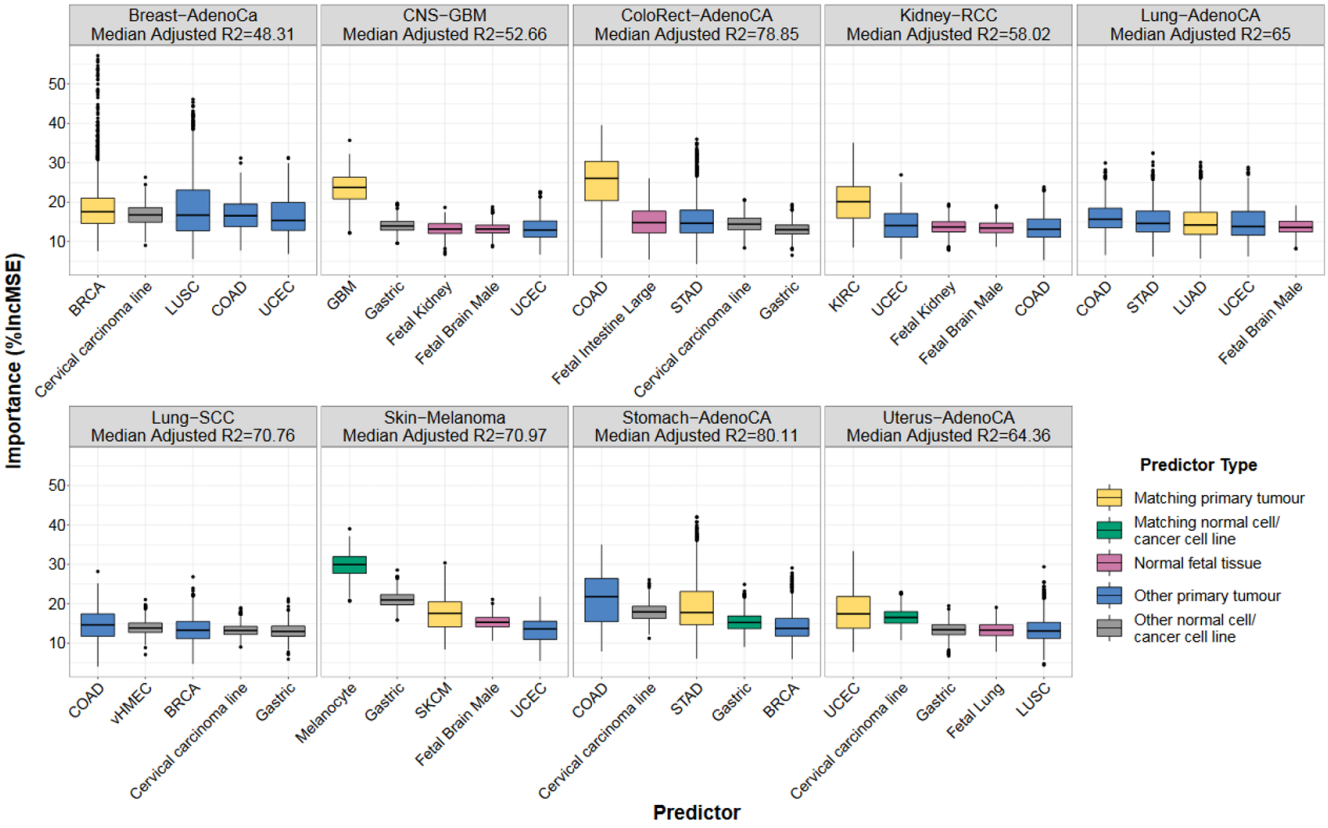


Figure 10: Importance metric of top 5 chromatin track predictors of regional mutation rates in 9 cancer types. Primary tumour chromatin tracks from matching tissues are the top predictors of regional mutation rates in 5/9 cancer types. Melanoma is the only cancer type for which regional mutational rates are best predicted by its matching normal cell chromatin track.

4.3.1 Tumor epigenomes are the strongest predictors of mutation rates in most cancer types

Based on median predictor importance, five out of the nine cancer types had their matching primary tumour chromatin accessibility track as their top predictor: breast adenocarcinoma, glioblastoma multiforme (GBM), colorectal adenocarcinoma, renal cell carcinoma, and endometrial adenocarcinoma. This suggests that the majority of the mutational landscape of these cancers is established after oncogenesis as it is most associated with an epigenome that appears to be post-oncogenic in origin.

In colorectal cancer, the later mutational timing has been supported by WGS of normal colorectal tissue from middle-aged individuals. These normal tissues demonstrated a three- to seven-fold decrease in total substitutions compared to colorectal cancer (excluding hypermutated CRC samples) (Lee-Six et al., 2019). This may be due to genome instability found in colorectal cancers, as 20% of CRC samples exhibit mutational signatures associated with DNA mismatch repair deficiency and/or mutations in DNA polymerase ε or δ (Alexandrov et al. 2020).

Later mutational timing in endometrial adenocarcinoma is also supported by a WGS study of normal endometrial cells. They showed that normal endometrial cells demonstrate a ~5-fold decrease in base substitutions (Moore et al. 2018). They attributed this to a subset of endometrial adenocarcinoma patients having mutational signatures associated with DNA mismatch repair deficiency and/or mutations in DNA polymerase ε or δ.

As a current limitation, WGS has not yet been performed on the corresponding normal tissues of the remaining four cancer types discussed in this section (breast adenocarcinoma, glioblastoma multiforme, kidney renal cell carcinoma, and stomach adenocarcinoma). Based on our models and the supporting evidence for the other cancer types, we infer that the mutational landscapes of these 4 cancer types are also mostly established after oncogenesis. This inference is further supported by evidence of defective DNA repair in all 3 of the 4 cancer types. Both kidney renal cell carcinoma and stomach adenocarcinoma are characterized by genomic instability and defective DNA repair pathways (Pilie et al. 2017; Sohn et al. 2017). WGS of breast adenocarcinoma also demonstrated the presence of a mutational signature related to defective homologous repair pathways (Alexandrov et al. 2020). Deficient NER and BER repair pathways have also been reported in breast cancer (Anurag et al. 2018). Glioblastoma multiforme, however, does not demonstrate a mutational signature related to genomic instability or defective DNA repair and will be discussed more closely in section **4.4.1**.

4.3.2 Epigenetic profiles of non-lung cancers are the top predictors of mutation rates in lung cancers

For the two types of lung cancers (adenocarcinoma and squamous cell carcinoma), the top predictors came from non-matching primary tumours (stomach and colon adenocarcinoma, respectively). In both cancer types, the matching primary tumour was the fourth best predictor and four of the top five predictors were derived from primary tumours. We did not observe considerable differences between the median importance metrics of the top predictors, suggesting that the associations of the epigenome and the megabase mutation rates in lung cancer derived from this analysis remain unclear. Interestingly, we did not observe the stronger association of regional mutation rates in lung cancer with the chromatin accessibility landscapes of cells derived from either the normal or tumour lung tissue.

4.3.3 Regional mutation rates in melanoma are best predicted by a normal melanocyte chromatin track

The top predictor of regional mutation rates in melanoma was derived from a healthy melanocyte skin cell. This is consistent with the Polak et al., 2015 study, demonstrating that the mutational landscape of melanoma is more associated with its earlier, potentially pre-oncogenic epigenetic state, rather than the later oncogenic epigenetic state. This finding is supported by the study of Martincorena et al. 2015, where they performed WGS on 4 adult skin samples. They found that the mutation burden of these skin samples was within the lower end of the range found in melanomas and greater than what is found in many adult solid tumours. The likely explanation is that sun-exposed skin is an exceptional tissue due to a lifetime of exposure to ultraviolet light and its mutagenic effects. No other normal tissue has been shown to carry such a high mutation burden without first undergoing oncogenesis (Martincorena et al. 2019)

4.4 Models trained on all primary tumour chromatin tracks inform of the epigenetic determinants of mutation rates

We wanted to assess the relationship between the cancer types of the most predictive primary tumour chromatin tracks and the cancer types of the regional mutation rate tracks they were predicting. We asked whether the chromatin tracks from the same cancer type would be most predictive of mutation rates. We also wanted to investigate the top predictors of regional mutation rates in cancer types with no chromatin tracks from the same cancer type available. Therefore, we trained a model on all primary tumour chromatin accessibility tracks to predict regional mutation rates for all 26 cancer types (including pan-cancer). 14 of these regional mutation tracks had chromatin accessibility tracks from tumours of the same cancer type. We then used the predictor importance metric of each of the predictors to determine which cancer type's chromatin tracks were the most important to predicting regional mutation rates in each cancer type **(Fig. 11)**

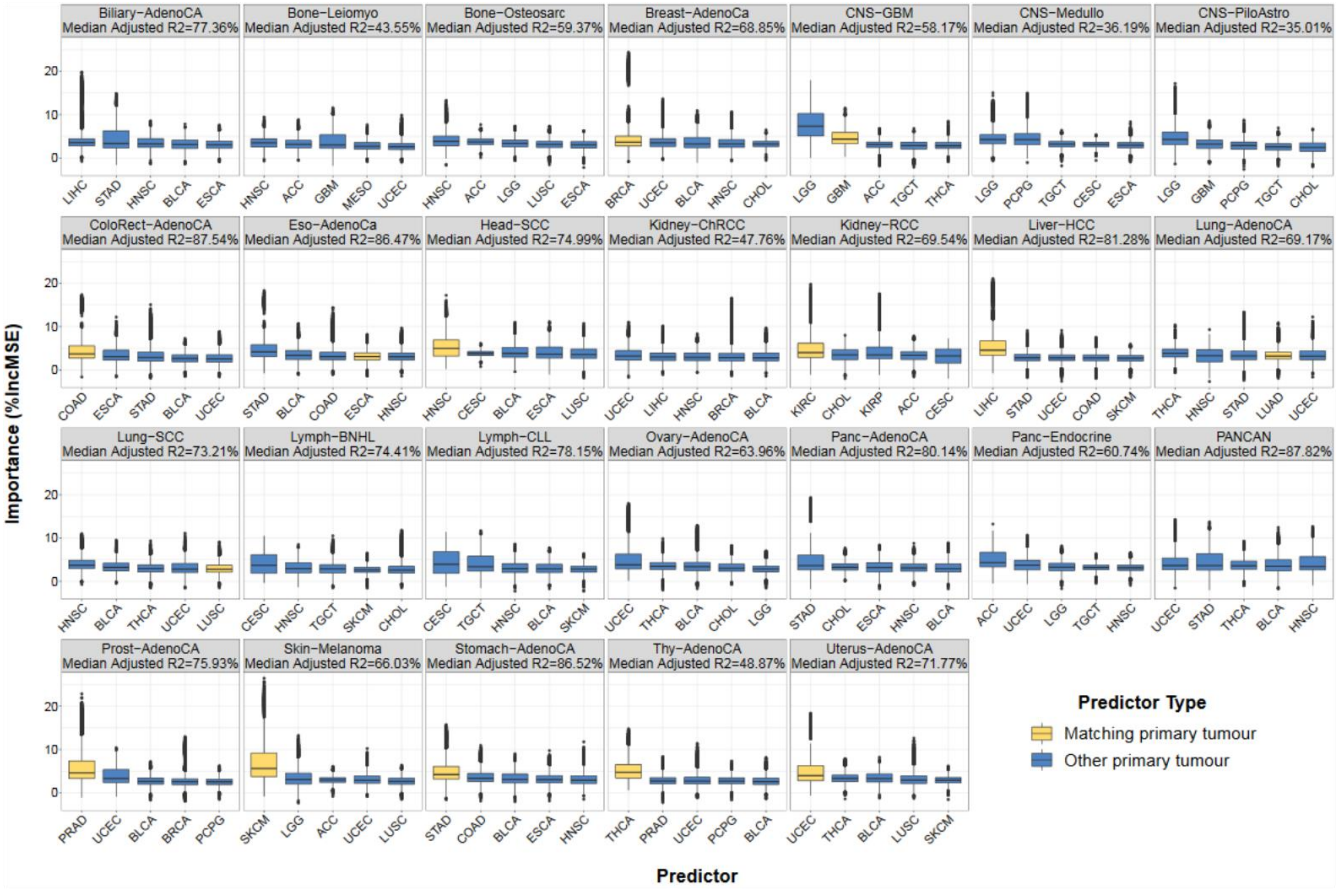


Figure 11: Importance of chromatin tracks in predicting cancer type-specific regional mutation rates. Boxplots demonstrate results of experiment where all primary tumour chromatin tracks were used to predict regional mutation rates. Boxplots are ranked in terms of chromatin tracks belonging to the 5 most predictive tumour types. 10/14 mutation rate tracks with chromatin tracks derived from the same tumour type (yellow), have those chromatin tracks as the top predictors. The exceptions include GBM, esophageal adenocarcinoma, lung SCC, and lung adenocarcinoma mutation rate tracks.

4.4.1 Most cancer type-specific regional mutation rates are best predicted by chromatin tracks from the same cancer type

Regarding the experiment described in **Fig. 11** where we trained models on all primary tumour predictors; 14/26 cancer type-specific mutation rate tracks had a chromatin track from the same cancer type. We found that the matching cancer type contributed the most predictive chromatin tracks in 10 out of these 14 cancer types. These included breast,

colorectal, head and neck, kidney, liver, prostate, stomach, thyroid, endometrial cancer, and melanoma.

Unexpectedly, the most predictive chromatin tracks of GBM mutation rates were derived not from GBM, but from low-grade glioma (LGG). LGG is often a precursor of secondary GBM which comprise 10% of glioblastomas (Mansouri, Karamchandani, & Das 2017). LGG chromatin tracks represent an epigenome which is more like normal glial epigenomes in terms of somatic evolution than GBM chromatin tracks. We may even view LGG epigenomes as a proxy of normal glial cell epigenomes, as we do not have any chromatin accessibility data for these cell types. LGG chromatin tracks being the best predictors of GBM mutation rates suggests that the mutational landscape of GBM is established as an earlier event in tumour evolution than expected.

Another exception involves regional mutation rates derived from esophageal adenocarcinoma, for which the top predictors were derived from stomach adenocarcinoma chromatin tracks. This makes sense, however, as the esophageal cancer chromatin tracks include both adenocarcinomas and squamous cell carcinomas. Therefore, the chromatin track of a stomach adenocarcinoma may be the closest match to regional mutation rates of esophageal adenocarcinoma in terms of cell type. Furthermore, both esophageal and stomach adenocarcinomas share a prominent mutational process caused by oxidative damage due to acid reflux (Tomkova et al. 2018).

The final exceptions involve the two lung cancers, which do not have chromatin tracks from matching tissues as their top predictors, as was observed in several of our analyses. In this analysis, chromatin tracks derived from thyroid adenocarcinoma were the most predictive of lung adenocarcinoma regional mutation rates. Chromatin tracks derived from head and neck cancer were the most predictive of regional mutation rates in lung squamous cell carcinoma. Interestingly, the top predictors (thyroid and head and neck cancer) were derived from the same tumour cell type as the mutation tracks (lung adenocarcinoma and lung squamous cell carcinoma), potentially because chromatin accessibility is highly cell type specific.

4.4.2 Top predictors of regional mutation rates without chromatin tracks from the same cancer type reveal cell type-specific associations between the cancer genome and epigenome

When examining the other 11 cancer types with no matched primary tumour chromatin tracks (excluding pan-cancer), we saw interesting trends in terms of top predictors. Regarding the two mutation rate tracks derived from brain tumours without chromatin tracks from the same cancer type, medulloblastoma and pilocytic astrocytoma, we saw the strongest association with chromatin tracks derived from the nervous system. Medulloblastoma regional mutation rates were best predicted by chromatin tracks derived from pheochromocytoma and paraganglioma (PCPG) and low-grade glioma (LGG) samples. PCPG tumours occur in the sympathetic nervous system, most commonly in neuroendocrine cells found in the adrenal medulla. LGG, as previously discussed, is a lower-grade glioma with a lower mutation burden than glioblastomas. Regional mutation rates in pilocytic astrocytoma were also best predicted by chromatin tracks derived from LGG. Both cancer types being examined here are childhood brain tumours and are two of the three lowest cancer types in terms of mutation burden in our dataset. It is possible that their mutational landscapes are associated with an earlier state in tumour development, however, chromatin accessibility from more CNS tissues and cell types are needed to confirm these results.

In biliary adenocarcinoma, regional mutation rates were best predicted by chromatin tracks derived from liver hepatocellular carcinoma. Although these two cancer types are derived from different cell types, they do interact functionally and are proximal to each other in the body. Both diseases are associated liver cirrhosis caused by hepatitis B, hepatitis C, and alcohol abuse (Heidelbaugh & Bruderly 2006; Shaib et al. 2005). The association between the chromatin landscape of biliary adenocarcinoma and the mutational landscape of HCC may point to a shared mechanism of mutagenesis occurring within these 2 tumour types.

4.4.3 Individual models trained on chromatin tracks from only one cancer type support relationship between the cancer genome and epigenome

We looked to validate our results from sections **4.4.1** and **4.4.2** using a different methodology. As opposed to training models on all primary tumour chromatin track predictors, we trained models on chromatin tracks derived from only one cancer type and repeated this procedure for each cancer type in our chromatin track dataset **(Fig. 12)**.

Overall, we found that this method agreed with the previous results in most cancer types. The previous findings showed that LGG chromatin tracks were the top predictors of GBM regional mutation rates and stomach adenocarcinoma chromatin tracks were the top predictors of regional mutation rates in esophageal adenocarcinoma. These results were replicated in this experiment as the most predictive model of GBM regional mutation rates was trained on LGG chromatin tracks and the most predictive model of regional mutation rates in esophageal adenocarcinoma was trained on stomach adenocarcinoma chromatin tracks. Contrary to previous results, we found that the most predictive model of the lung adenocarcinoma mutation track was trained on lung adenocarcinoma chromatin tracks. In the previous experiments, lung cancer regional mutation rates showed no preferential association with chromatin tracks derived from the same cancer type indicating there is a complementary relationship between the two methods.

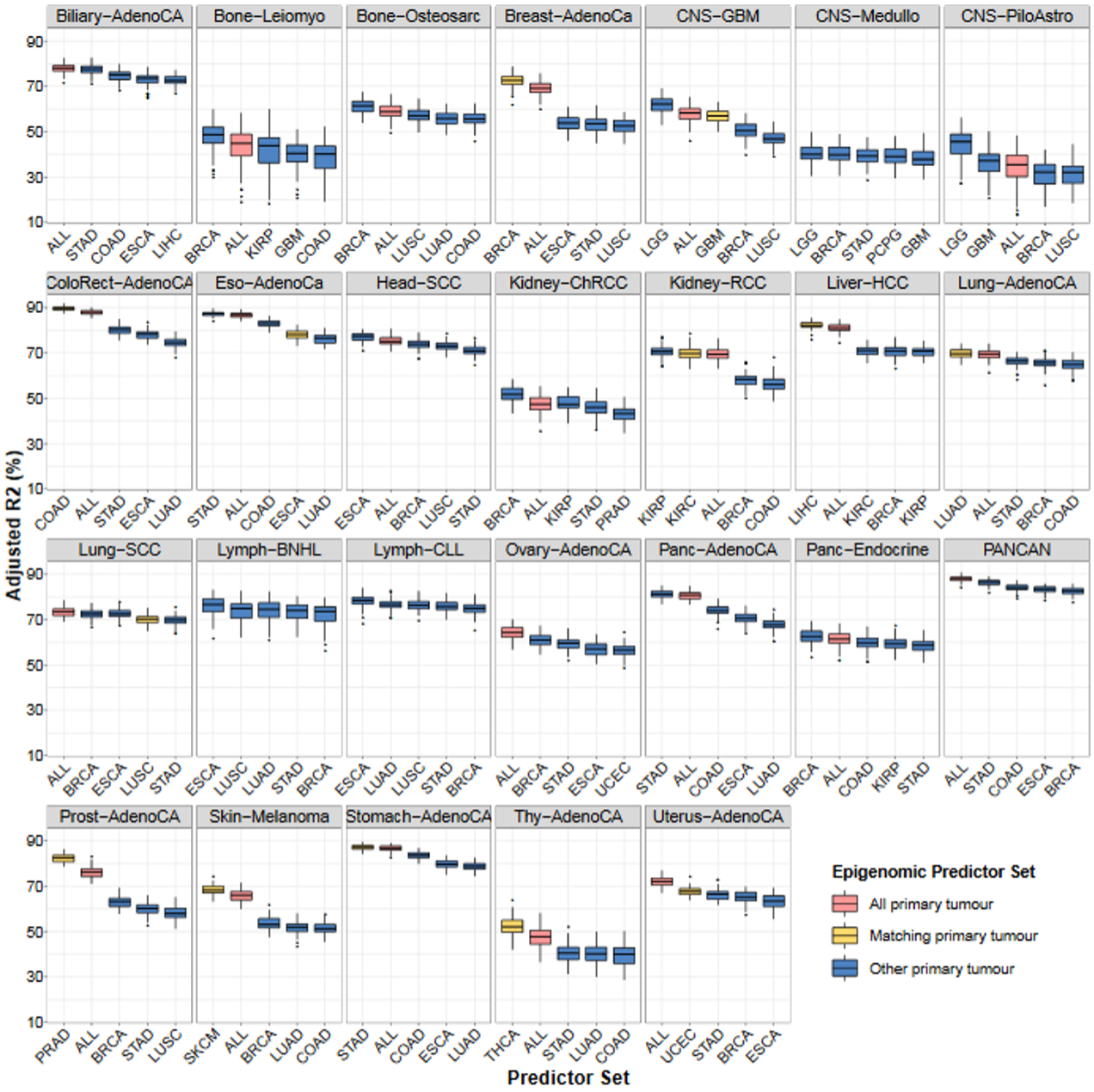


Figure 12: Accuracy of models trained on chromatin tracks from differing cancer types. Boxplots show adjusted R2 (y-axis) of models trained on primary tumour chromatin tracks (x-axis) to predict regional mutation rates in 26 cancer types. Boxplots are coloured based on whether a model was trained on all primary tumour chromatin tracks (pink), primary tumour chromatin tracks derived from the same cancer type as the mutation track (yellow), or primary tumour chromatin tracks derived from another cancer type (blue).

4.5 Mutational signatures reveal underlying relationships between the cancer genome and epigenome

Somatic mutations occur in genomes of individual cells due to mutational processes caused by specific endogenous or exogenous mutagens, deficient DNA repair processes, and/or DNA replication. Previous research has shown that different mutations grouped by their trinucleotide context are associated with these endogenous and exogenous processes, using algorithms that discover mutational signatures (Petljak et al. 2019; Alexandrov et al. 2020). It is unclear, however, what role the chromatin landscape plays in determining the genome-wide activity of mutational processes and the related genomic distribution of mutational signatures.

To uncover such relationships, we trained our models on our combined set of normal and tumour chromatin tracks to predict regional mutation rates consisting of mutations from only one signature in one cancer type. We predicted the role of each signature in the cancer type-specific relationship between the genome and the epigenome.

4.5.1 Specific mutational signatures drive the association between chromatin landscape and the mutational landscapes of cancer

After examining the models trained to predict regional mutation rates from single mutational signatures, several trends became evident **(Fig. 13)**. The foremost of these is that the performance of models trained to predict mutations from some signatures nearly matched the performance of those trained to predict all mutations in a cohort. For example, we see this trend with signatures SBS18 and SBS40, both of unknown aetiology, in colorectal adenocarcinoma. The adjusted R2 scores for models trained to predict SBS18 and SBS40 (Adj. R2=85.4 and 86.4, respectively) mutations in the colorectal cancer cohort nearly match that of the model trained to predict all the mutations in the cohort (Adj. R2=87.3). We saw this trend occur in multiple signatures and cancer types, such as SBS40 in several cancer types, SBS7a (UV) in melanoma, and SBS4 (smoking) in lung cancers. However, we also found that models trained to predict mutations attributed to several signatures showed considerable decreases in accuracy such as SBS1 and SBS13 in several cancer types. These

results indicate that the regional mutation rates due to some mutational signatures are highly associated with the chromatin landscape of the cell while some others are not.

We also observed that regional mutation rates from several mutational signatures of unknown aetiology were highly predicted in related cancer types. This was most evident with SBS18 in esophageal, stomach, and colorectal cancer as well as SBS17a/b in esophageal and stomach cancer, possibly suggesting that these mutational signatures represent a common mutagenic exposure occurring in the esophageal tract and the stomach.

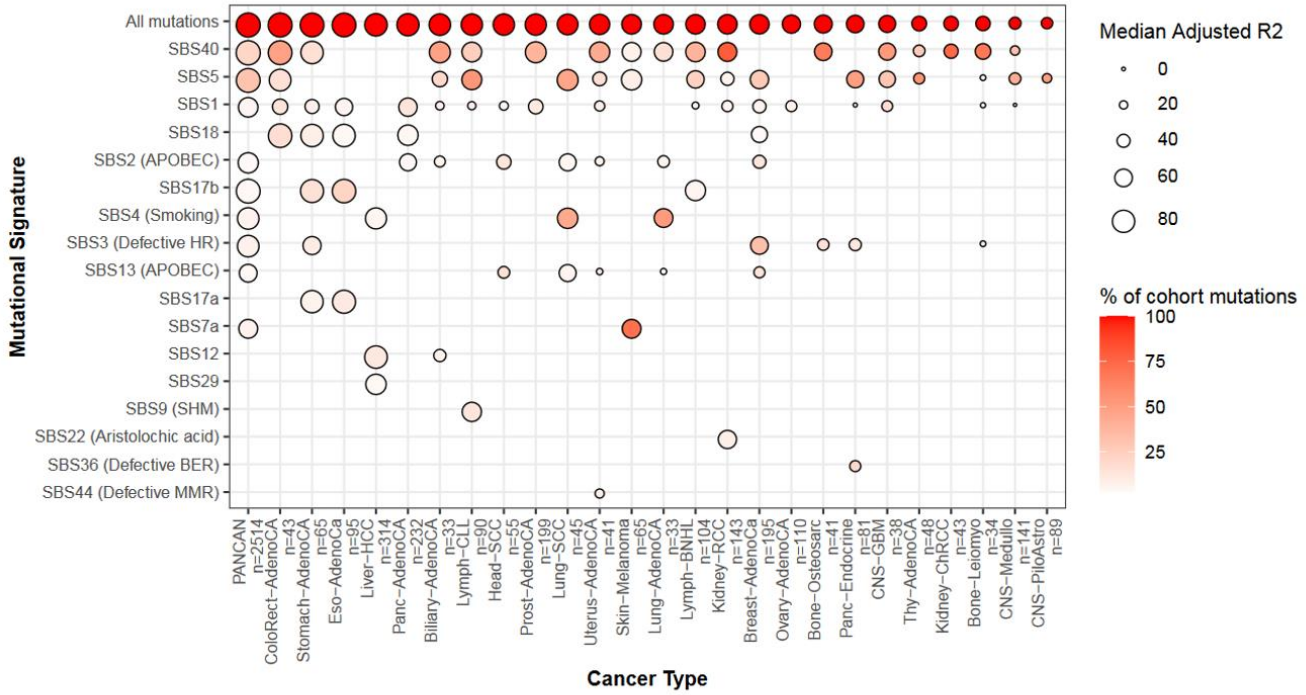


Figure 13: Overview of chromatin tracks predicting cancer-type and mutational signature-specific regional mutation rates. Each dot represents a model trained on all of our normal and primary tumour chromatin tracks to predict regional mutation rates from a specific signature and cancer type. Colour of dots represents the % of total mutations in that cancer type which that mutational signature represents. Size of dot represents predictive power of the model (adjusted R2 %).

4.5.2 Primary tumour and normal chromatin tracks are highly predictive of regional mutation rates related to exogenous mutational processes

Upon further examination of our results **(Fig. 14)**, we found that the models trained to predict regional mutation rates from exogenous sources were similar in performance to the models trained to predict all mutations in the cohort. Models trained to predict endogenous signatures tended to show more significant decreases in model accuracy suggesting that the location of these mutations is less dependent on the chromatin landscape. Finally, several signatures of unknown aetiology showed little to no decrease in model accuracy suggesting that mutations due to these signatures may in fact be of exogenous origin.

In terms of exogenous mutational processes, we found that models trained to predict regional mutation rates attributed to SBS7a (UV light) in melanoma showed a similar performance to models trained to predict all melanoma mutations (Adj. R2=65.4 and 68.6, respectively) with SBS7a mutations representing 71% of total melanoma mutations in our dataset. We found that models trained to predict mutations attributed to SBS4 (tobacco smoking) showed similar performances to models trained to predict all mutations in lung squamous cell carcinoma (Adj. R2=72.1 and 73.5 respectively), lung adenocarcinoma (Adj. R2=65.0 and 68.3, respectively), and liver HCC (Adj. R2 74.6 & and 80.8%, respectively). Finally, the SBS29 signature associated with tobacco chewing showed an 8.8% decrease in model accuracy. These results imply that the genomic regional variation in mutations due to exogenous signatures is highly dependent on the chromatin landscape of the cell.

In contrast, when examining signatures related to endogenous sources of mutations, we find more significant decreases in model accuracy. In terms of mutations related to defective DNA repair pathways (SBS3, SBS36, and SBS44), we observed at least a 20% decrease in Adjusted R2 between the model predicting all mutations with a specific cohort and the model predicting regional mutation rates from only these signatures. This decrease in model accuracy was found in six of seven cancer type-signature combinations. The one exception was SBS3 in breast adenocarcinoma which only showed a 7.7% reduction in adjusted R2. Interestingly, SBS3 is associated with germline and somatic BRCA1 and BRCA2 mutations as well as BRCA1 promoter methylation in breast, pancreatic, and ovarian cancer (Nik-

Zainal et al. 2012). In terms of mutational signatures related to APOBEC enzyme activity of mutagenesis (signatures SBS2 and SBS13), we observed at least a 20% decrease in adjusted R2 between the model predicting all mutations with a specific cohort and the model predicting mutations only from these signatures in 10 out of 12 cancer type-signature combinations. The two exceptions were SBS2 and SBS13 in lung squamous cell carcinoma which both showed a 16% decrease in adjusted R2 (SBS2 and SBS13 are highly associated with each other). This suggests that certain endogenous mutational processes are less associated with the chromatin landscape, potentially because these mutational processes act uniformly on DNA, regardless of chromatin accessibility.

Two signatures are known to be associated with total number of stem cell replications and as a proxy of patient age – SBS1 and SBS5. These signatures showed a clear dichotomy between SBS5, for which regional mutation rates were accurately predicted and SBS1, for which model performance was significantly decreased. Aside from being highly correlated with age, SBS5 aetiology is mostly unknown. Some associations have been noted in bladder cancer samples with *ERCC2* (*ERCC* excision repair 2) mutations and in many cancer types due to tobacco smoking (Kim et al. 2016). SBS1 is thought to be caused by the spontaneous enzymatic deamination of 5-methylcytosine to thymine. SBS1 mutation rates are highly different between tissues and are correlated with the tissue-specific rate of cellular division suggesting that it is a stem cell division or a mitotic clock.

Most of the signatures with unknown aetiology we tested in this study demonstrated little to no decrease in model performance, with only one of 29 cancer type-signature combinations showing over a 20% decrease in model accuracy relative to the model predicting all mutations. Signatures from which regional mutation rates were highly predicted by chromatin tracks include SBS40, SBS18, SBS17a/b, and SBS12. Previously, we found that exogenous signatures were highly associated with the chromatin landscape while endogenous were not, suggesting that mutations due to these signatures may be of exogenous origin. Several of these unknown signatures are prominent in cancer types derived from related tissues, such as SBS18 in colorectal, stomach, esophageal, and pancreatic adenocarcinoma, SBS17a/b in stomach and esophageal adenocarcinoma, and SBS12 in liver HCC and biliary

adenocarcinoma (**Fig. 13)**. This suggests that these signatures are associated with a shared process of mutagenesis in these related cancer tissues. The exception to this trend is SBS40 which is found across most cancer types and across unrelated cancer tissues. Previously, SBS17 was shown to be associated with replication timing (which is also associated with chromatin accessibility) and it is thought to be caused by oxidative damage due to gastro-oesophageal and duodena-gastric reflux as it was also found in Barrett's oesophagus (Tomkova et al. 2018).

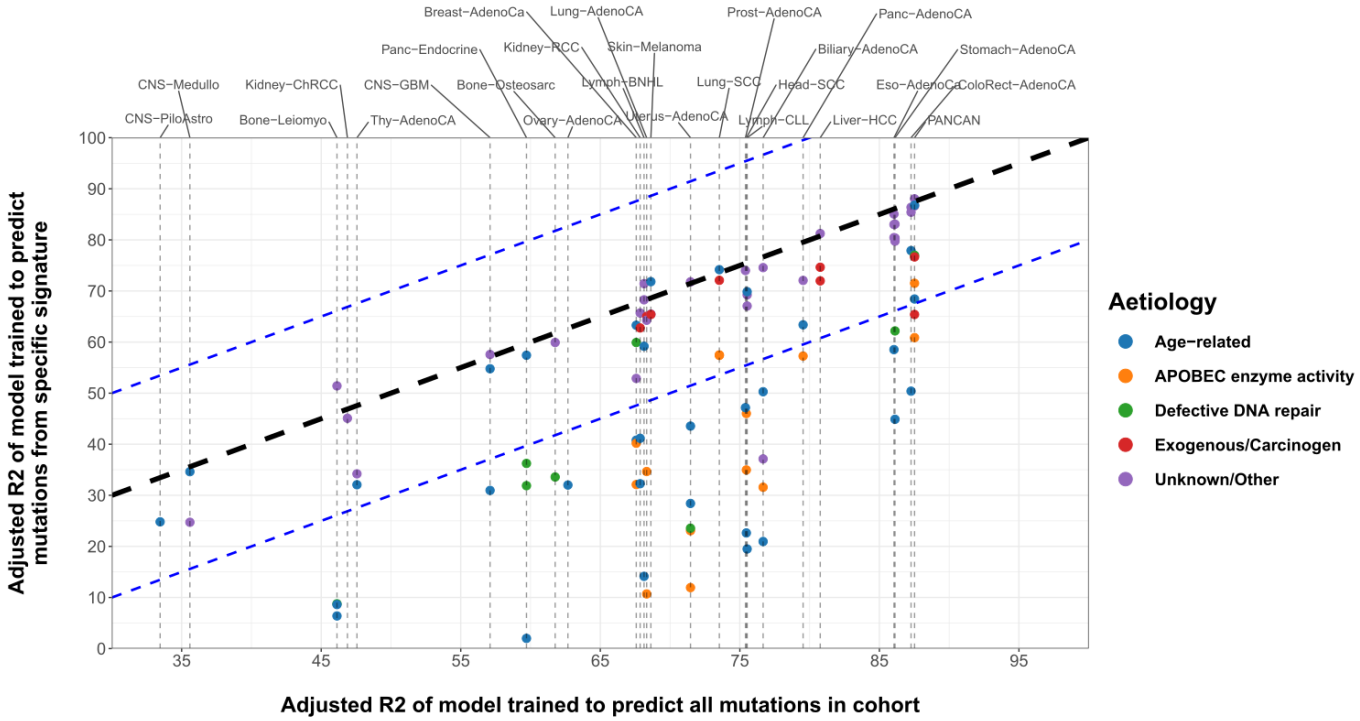


Figure 14: Mutational aetiology influences association between mutational and chromatin landscape. Scatterplot represents the performance of models trained to predict mutations from a specific mutational signature (y-axis) vs. the performance of the model trained to predict all mutations (x-axis) within the same cancer type. Black diagonal line represents no change in model accuracy while blue diagonal lines represent a 20% increase/decrease in model accuracy. Points on vertical segments represent models trained to predict mutations from various signatures within the same cancer type. Points are coloured according to aetiology of mutational signature of which mutations are being predicted. Exogenous mutational signatures (red) show minimal decreases in model accuracy, with most being within 20%, while endogenous signatures (blue, orange, and green) show considerable decreases. Several signatures of unknown aetiology (purple) also show minimal decreases in model accuracy suggesting possible exogenous aetiology.

4.5.3 Regional mutation rates from most mutational signatures have consistent most predictive chromatin tracks within the same cancer type

To examine which chromatin tracks were the most predictive of mutational signature- and cancer type-specific regional mutation rates, we performed the same analysis as in **section 4.3.** We examined the same nine cancer types where mutational profiles were available with chromatin profiles of tumor cells and normal cells. For each of these cancer types, we used the same set of unified predictors, however in addition to predicting all mutations in a cohort, we predicted mutations belonging to specific signatures only **(Fig. 15)**. We observed that most signatures had consistent top predictors within the same cancer type. This result demonstrates that for most signatures and cancer types, primary tumour epigenomes are better predictors of regional mutation rates, regardless of the underlying mutational process or signature examined in a given analysis.

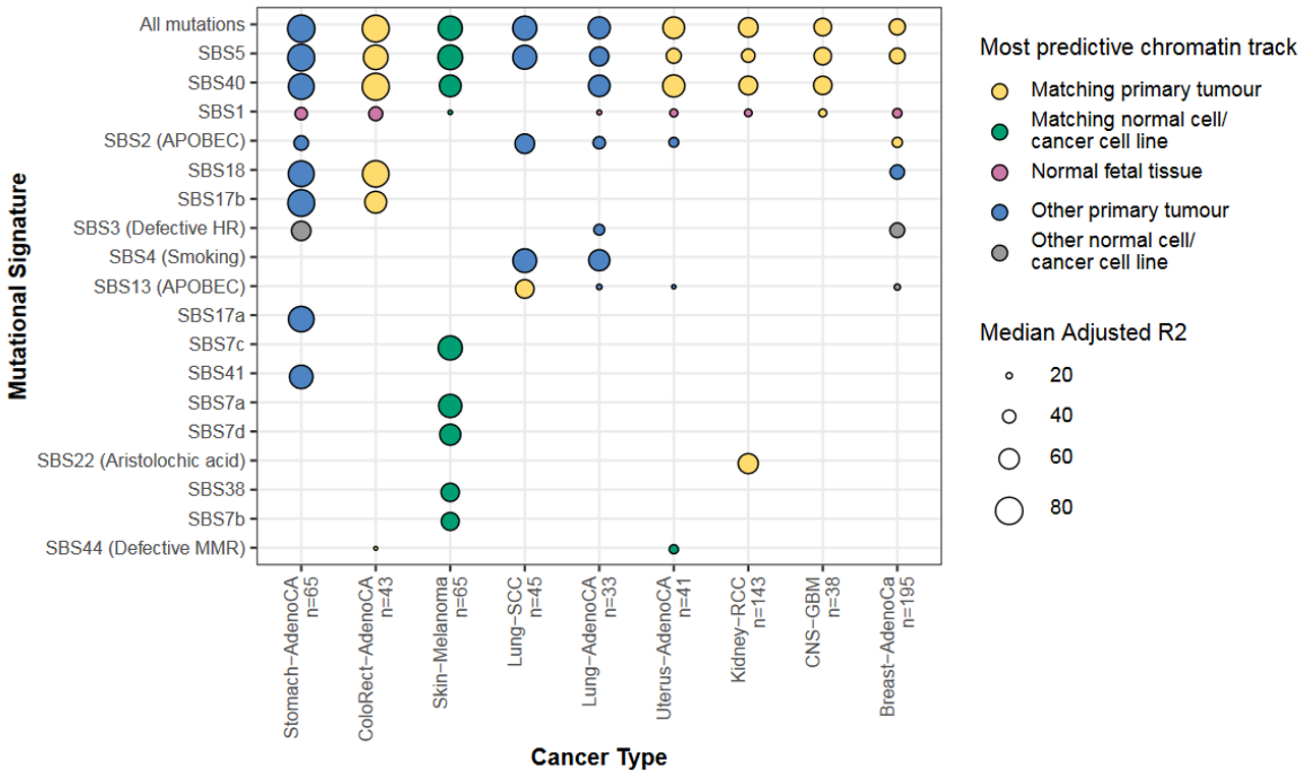


Figure 15: Most predictive chromatin tracks of mutational signature and cancer type-specific mutation tracks. Size of dot indicates model accuracy and colour dot indicates classification of the model's most predictive chromatin track. Most predictive chromatin tracks are consistent across signatures within the same cancer type but not vice-versa.

4.5.4 Chromatin tracks derived from fetal tissues are the best predictors of SBS1 regional mutation rates across most cancer types

Interestingly, fetal tissues were observed to be the most predictive chromatin tracks of SBS1 mutation rates in multiple cancer types. In fact, SBS1 was best predicted by a chromatin track derived from fetal brain tissue in seven out of nine cancer types tested **(Fig. 16)**. Furthermore, there were chromatin tracks derived from other fetal tissues such as kidney and intestine in the top five predictors of regional mutation rates of all nine cancer types. SBS1 has been previously linked with stem cell division rate in multiple cancers and has been proposed as a mitotic clock (Alexandrov et al. 2020). Furthermore, the mutational landscape of fetal intestinal stem cells has been shown to be dominated by SBS1 mutations (Kuijk et al. 2019). We speculate that fetal chromatin tracks are most predictive of the SBS1 mutational

landscape because fetal tissues have increased cellular replication rates and are therefore exposed to higher levels of replication-related mutations such as SBS1. Furthermore, fetal tissues are not exposed to the same set of exogenous mutagens as adult tissues and so the fetal chromatin landscape may be a proxy of the SBS1 and the replication-related mutational landscape. However, SBS1 mutations are only a minority of mutations in the cancer types we considered (10-15%) and the performances of our random forest models are mostly substandard (Adj. R2 15-45%). Therefore, these findings should be treated with caution. With the availability of chromatin and mutation profiles from fetal and adult tissues, further work can elucidate this link.

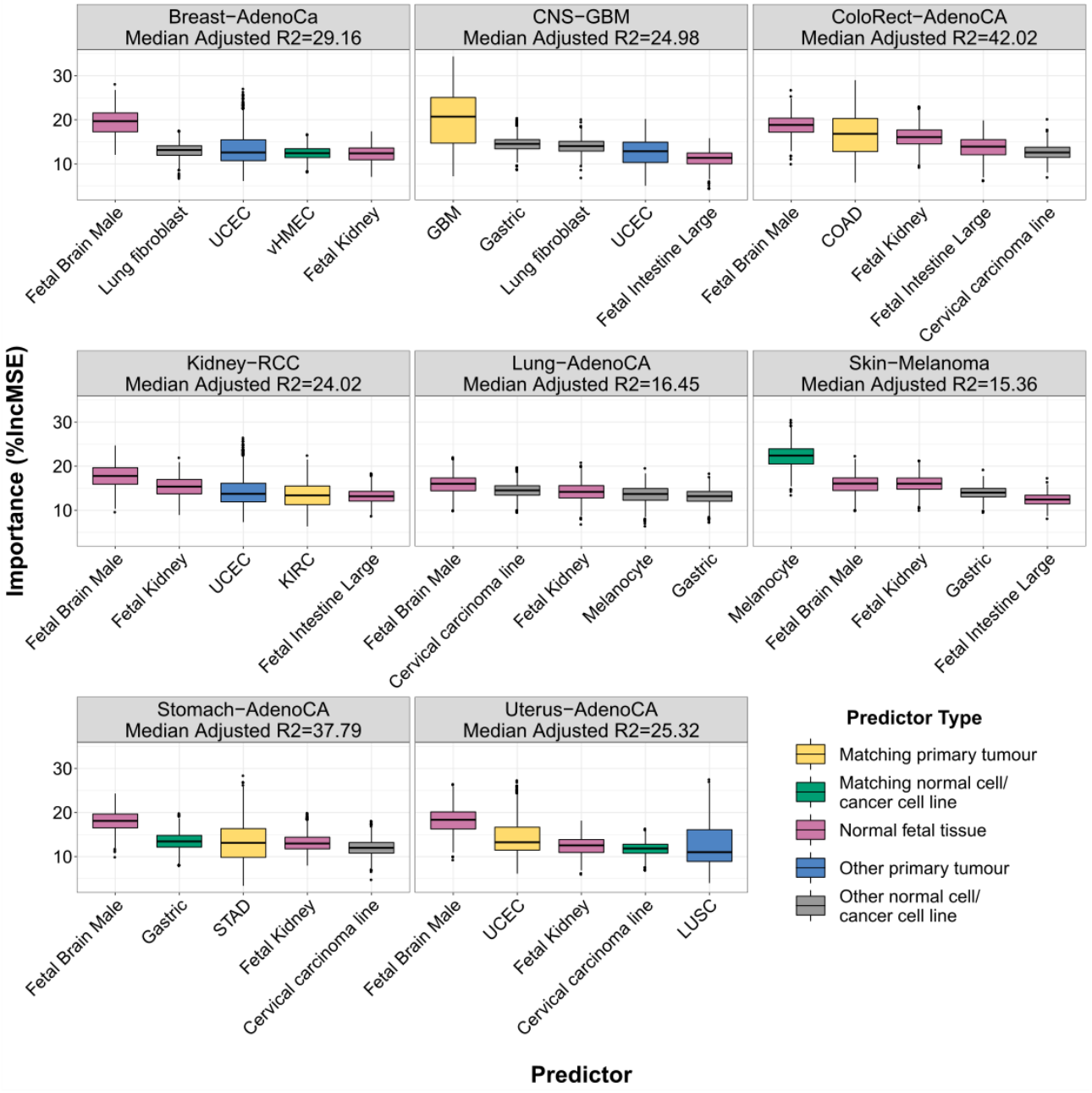


Figure 16: Most predictive chromatin tracks of regional mutation rates attributed to SBS1. Boxplots represent importance scores for the top 5 most predictive chromatin tracks of SBS1 mutations in 9 cancer types. In 7/9 cancer types, chromatin tracks derived from fetal brain tissues are the top predictor. Chromatin tracks derived from fetal tissues are present in the top 5 predictors in all 9 cancer types.

Chapter 5

Discussion

5 Discussion

5.1 Summary of findings

To date, few studies have examined the relationship between the mutational landscape of cancer and the chromatin accessibility landscape of primary tumours. This is mostly due to large-scale chromatin accessibility datasets derived from primary tumours only recently becoming available. Previous studies have used chromatin accessibility from normal cells and cancer cell lines as a proxy of the epigenetic landscape of primary tumors, however, this approach has profound limitations. Firstly, the chromatin landscapes of normal cells evolve dramatically after transformation. Epigenetic changes have been observed to be associated with all hallmarks of cancer. Secondly, cancer cell lines represent a genetically and epigenetically distinct entity to cells making up primary tumours due to *in vitro* culturing and passaging effects. Thirdly, the normal cell chromatin accessibility dataset only contains a few samples sequenced for each cell type thereby reducing the statistical power of such an analysis. Previous studies suggest that the associations of mutation rates and chromatin state are the strongest across the same cell type (*i.e.* melanocyte chromatin best predicts melanoma regional mutation rates) (Polak et al. 2015). With the availability of chromatin accessibility profiles from primary tumours rather than only cell lines and normal tissues, we can obtain a more accurate understanding of the associations of mutagenesis and chromatin state. Systematic evaluation of the interactions of genome-wide mutation rates and the chromatin landscapes of primary tumours and normal cells will reveal insights into tumour evolution, mutational processes, and cancer tissue of origin.

Firstly, we demonstrated that chromatin tracks of primary tumour tissues are more predictive of regional mutation rates than chromatin tracks of normal cells in 20/26 cancer types (including pan-cancer). This would suggest contradicting evidence to the Polak et al. 2015

study, as we observe that primary tumour chromatin landscape is more strongly associated with the mutational landscape of cancer than normal cells (Polak et al. 2015).

Upon examining the most predictive chromatin tracks of each cancer type-specific mutation track, we found several interesting trends. Firstly, in five out of nine cancer types, we observed that the most predictive chromatin tracks were derived from primary tumours of the same cancer type as the mutation track which they predicted. This was shown in breast adenocarcinoma, glioblastoma multiforme, colorectal adenocarcinoma, kidney renal cell carcinoma, and uterine adenocarcinoma. This result, in addition to previous work, implies that the mutational landscape in cells of these cancer types is established as a later event in tumour evolution. Further support is derived from defective DNA repair mutational signatures found in many of these late mutation-timing cancer types (Pilie et al. 2017; Sohn et al. 2017; Anurag et al. 2018; Alexandrov et al. 2020). This suggests that DNA repair may have been compromised during or after oncogenesis, after which most of the mutations occurred due to the lack of DNA repair. Importantly, these results allow us to infer mutational timing in cancer types for which WGS has not been performed in the corresponding normal tissue. This is only an indirect approach, however, and more direct evidence such as WGS experiments of more normal and tumour tissues are needed.

Interestingly, we found that in melanoma, the top predictor of regional mutation rates was derived from the matching normal cell chromatin track of normal melanocytes. In various computational experiments, regional mutation rates in melanoma were most strongly associated with the chromatin landscape of normal melanocytes, rather than the chromatin landscape of melanoma cells. We speculate that while the mutational landscape in most of these cancer types is established after oncogenesis, most of the mutations in melanoma occur prior to oncogenesis. This is supported by WGS mutation data revealing high levels of mutation burden in normal skin tissue, within the range of mutation burden found in melanoma cells (Martincorena et al. 2015).

In our next analysis, we observed that out of 14 cancer types with both mutation rate tracks and chromatin tracks from the same cancer type, 10 had these chromatin tracks as the top

predictors. This indicates high cell-type specificity in the genome-epigenome relationship. Interestingly, we found that regional mutation rates in GBM were best predicted by chromatin tracks from low-grade glioma (LGG). LGG is often a precursor to GBM and therefore represents an earlier stage in tumour development. This suggests that the mutational landscape of GBM may be established earlier than expected during tumour evolution, perhaps during the low-grade phase of glioma development, however, whole-genome sequencing data of lower-grade gliomas is lacking in our dataset. Further analysis of WGS and epigenetic data from a cohort of LGGs and GBMs may reveal further insights into the genome evolution of these heterogeneous tumors.

By using our chromatin tracks to predict regional mutation rates containing only mutations from specific mutational signatures, we made several observations to better understand the relationship between the epigenome and specific mutational processes. We first found that the genome-wide distributions of certain mutational signatures were better predicted by chromatin accessibility than others. This implies that mutations of these well-predicted signatures are highly associated with the cell's epigenome. First and foremost, signatures associated with exogenous or carcinogenic aetiologies such as SBS7a (UV) in melanoma and SBS4 (smoking) in lung and liver cancer were associated with the chromatin landscape. In contrast, endogenous signatures related to defective DNA repair and APOBEC enzyme activity showed significantly weaker associations with the chromatin landscape. This may due to these endogenous mutational processes affecting DNA uniformly, regardless of chromatin accessibility. Several signatures of unknown aetiology (namely SBS12, SBS17a/b, SBS18, and SBS40) were highly associated with the chromatin landscape, suggesting that these signatures are of exogenous or carcinogenic origin, based on their associations with chromatin accessibility. Interestingly, SBS17 has been shown to be potentially associated with oxidative damage due to acid reflux in gastric and esophageal tissue, which our results support. Further work into uncovering the origin of these mutational signatures can yield insight into novel carcinogenic processes and agents and their associations with chromatin state.

Upon examining the most predictive chromatin track of each signature, we found that only one signature shows a common top predictor across cancer types; SBS1. SBS1 is a signature attributed to stem cell division rate and has been described as a mitotic clock (Alexandrov et al. 2015). The top predictor of SBS1 regional mutation rates in seven of nine cancer types was derived from normal fetal brain tissue. This is noteworthy as the fetal tissue has a high rate of stem cell division and is less exposed to exogenous sources of mutations. These results suggest that the fetal chromatin landscape is a proxy for the SBS1 mutational landscape. Performance of these models is substandard, however, so these results should be treated with caution. Further work using a larger set of SBS1 mutations and fetal chromatin tracks is required to elucidate these findings.

In conclusion, we found that by analyzing the relationship between the cancer genome and epigenome, we gain insight into tumour evolution, mutational timing, and mutational processes. It is documented that the chromatin state undergoes significant changes while normal cells transform into tumour cells (Perdigoto 2019). The chromatin state is known to be associated with mutation rates in both normal and tumour cells. Therefore, if the chromatin state of tumour cells is more predictive of somatic regional mutation rates than the chromatin state of normal cells, the tumour chromatin state has had a stronger impact in shaping genome-wide mutation rates. This would suggest that most mutations have occurred after oncogenic transformation. We found this to be the case in most cancer types as primary tumour chromatin tracks were most predictive of cancer mutation rates in most cancer types. We found several exceptions, however, such as melanoma and to a lesser extent, GBM. Most of the mutations in melanoma are established prior to oncogenesis due to a lifetime of UV exposure and this was supported by our results from multiple analyses. Furthermore, the GBM mutational landscape was most associated with an epigenome earlier in tumour development (namely the LGG epigenome) rather than epigenomes derived from GBM tumours. This suggests that most of the mutations are established earlier than expected in tumour evolution. Finally, we examined the relationship between the chromatin landscape and the mutational landscape of specific mutational signatures/processes. We found that exogenous signatures were highly associated with the chromatin landscape while endogenous signatures were not. We also found evidence for the exogenous aetiology of several

signatures with previously unknown aetiologies. Finally, we showed that regional mutation rates due to the mutational signature SBS1 in cancer, associated with stem cell replication rates, were best predicted by chromatin tracks derived from fetal tissue, although the general model accuracy was relatively lower than in other analyses.

5.2 Considerations and challenges

Several challenges exist for bulk tumour sequencing datasets and their applications. First, bulk tumour sequencing datasets represent the average of a heterogenous composition of cells existing within the tumour microenvironment and may contain non-cancer cells such as immune infiltrates as well as cells in varying cell states. The primary tumour itself may contain several subclonal populations of cells with considerably heterogenous genomic and epigenomic compositions (Gerstung et al. 2020). In this project, it is important to consider that the mutational or chromatin landscapes used may be confounded by the contribution of one or more subclones as well as immune cells and other cells in the microenvironment that were included in the sequencing analysis. The emergence of single-cell sequencing as an alternative to bulk tumour sequencing will enable us to elucidate the molecular compositions of various tumour subclones as well as other cells within the tumour microenvironment.

Next-generation sequencing of primary tumours also has several technical limitations. The length of reads being used as well as sequencing depth have a significant impact on the ability to call somatic variants or align reads to a reference genome. Short-read sequencing (75-300 bps) is the most common method as it is more cost-effective. However, reassembling the genome from short reads can be challenging, as many fragments look highly similar without additional context and are therefore difficult to align to the reference genome. Long-read sequencing (>10,000 bps) is an emerging technology that allows for increased overlap between reads and the ability to call variants in more repetitive regions of the genome. On the other hand, sequencing depth, or coverage, represents the number of reads sequenced for each nucleotide. The higher the sequencing depth, the more confidence there is that a variant call is in fact due to a variant and not a technical error. Higher sequencing depth is becoming more available as sequencing technology advances.

In terms of chromatin accessibility mapping, several limitations of DNase-Seq and ATAC-Seq have been described. It has been shown that at the base pair scale, both methods have sequence biases (Calviello et al. 2019). However, we found no evidence that these sequence biases affected chromatin accessibility at the megabase pair scale, as all our chromatin tracks derived from DNase-Seq and ATAC-Seq showed a concordant negative correlation with regional mutation rates. Furthermore, peaks from these DNase-Seq and ATAC-Seq tracks were previously shown to be highly correlated (Corces et al. 2018). Importantly, the epigenomic landscape of a cell, unlike the genomic landscape, is highly dynamic while sequencing studies only capture one time point. The epigenome changes in response to endogenous stimuli such as hormones and exogenous stimuli such as chemicals, diet, and stress. Evidence indicates that epigenetic changes due to these stimuli can permanently alter the epigenetic state of an individual's cells (Kanherkar et al. 2014).

Regarding our tumour sequencing datasets, the most obvious limitation is the number of tumour samples in addition to the range of tissues and cancer types sequenced. It is important to note that we used datasets available to us at the time of the study, while the field is rapidly expanding with new studies and datasets being published regularly. For the WGS mutation dataset, we currently lack the statistical power to accurately describe the mutational landscapes of tumour types with lower mutational burdens such as childhood brain tumours. To elucidate the mutational landscapes of these cancer types, more samples are needed. In terms of our chromatin accessibility datasets, an increase in the number of samples and tissues being sequenced will lead to a better understanding of the human normal and tumour chromatin landscapes. With respect to our project, the desired combinations comprising chromatin accessibility mapping from matching types of normal and tumour tissues (*i.e.*, melanocyte and melanoma) as well as WGS mutation data from the matching cancer type is only available for nine tissues. With the availability of more WGS and chromatin accessibility data published in the future, these experiments can be performed for more human tissues to better understand mutational processes, mutational timing, and tissue of origin across a wider spectrum of cancer types.

There are several limitations to our machine learning framework as well. Firstly, our mutation tracks were derived from the aggregated mutations within a cohort and then split into genomic windows of fixed width of one megabase. Cancer types with higher mutation burden such as melanoma are bound to provide better-powered signals compared to those with more silent genomes such as paediatric brain cancers. Secondly, our chromatin tracks were derived from the binned average of chromatin accessibility scores within each window. Several other methods have been used in the past such as counting number of accessible peaks, but the binned average method was used for its improved consistency in comparing chromatin accessibility across two experiment types such as DNase-Seq and ATAC-Seq. Thirdly, the random forest model best represented the non-linear relationship between the chromatin and mutation landscapes analysed in this project. However, with the availability of larger sequencing datasets, more complex machine learning models such as deep neural networks may become more relevant.

5.3 Future directions

The availability of larger and more diverse datasets of cancer genomics and epigenomics will enable the application of our analytical framework to more tissues with higher statistical power. Aside from increased data, however, there are two clear directions which can be taken to advance this project.

First, our method can be applied to WGS datasets of metastatic tumours to elucidate the relationship between the mutational landscape of metastatic tumours and the chromatin landscape of normal cells and primary tumours (Priestly et al. 2019). Using the mutational landscape as a proxy, we can gain insights into the chromatin landscapes of metastatic tumours, for which literature is lacking. This will allow us to elucidate mutational timing, tissue of origin, and mutational processes in metastatic tumours and their response to cancer therapies. Additionally, we can better understand the role of the epigenome in the tumour cell's transition from a localized primary tumour to a metastasis. In addition to advancing our understanding of the disease, this could have significant clinical impact as metastasis is the most common cause of death for cancer patients.

Second, although we have studied the regional variation in mutations and chromatin accessibility, work can be done to elucidate the local effects. Local effects indicate which regions of the genomes demonstrate a high association between the chromatin and mutational landscapes and which regions do not. This can provide insight into genomic regions, genes, and regulatory elements in which epigenetic deregulation may be the cause of oncogenic mutations. Further work into understanding these exceptional regions can lead to novel cancer prognostics, diagnostics, and therapies.

5.4 Significance

Here we demonstrated the association between the chromatin landscapes of normal cells and primary tumours and the mutational landscape of cancer. To our knowledge, this is the first study to analyze this relationship using chromatin accessibility derived from primary tumours. Using our chromatin tracks, we inferred mutational timing in multiple cancer types. We found that the mutational landscapes of most cancer types are more strongly associated with the primary tumour chromatin landscape compared to that of normal cells. These results suggest that most mutations in these cancer types are more likely to have occurred later in tumour evolution. Exceptionally, we found that the mutational landscape of melanoma is most associated with the chromatin landscape of normal skin cells suggesting a pre-oncogenic accumulation of mutations, which has been confirmed by WGS experiments of normal skin tissue. Furthermore, we found that the mutational landscape of GBM is most associated with the chromatin landscape of LGG, a less aggressive tumour within the same cell type, suggesting that the mutational landscape was mostly established earlier than expected during tumour evolution, possibly in the low-grade glioma phase. Finally, we showed that exogenous mutational processes are highly associated with the chromatin landscape and demonstrated evidence for the exogenous aetiologies of several previously unclassified mutational processes. We also showed that mutation rates of signature SBS1 in cancer are most associated with the fetal chromatin landscape, suggesting that these mutations dominate the fetal genome. In conclusion, integrative analysis of mutations and chromatin state allows us to learn about tumour evolution, mutational processes, and cancer tissue of origin.

References

1. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* (2015) doi:10.1038/ng.3441.

2. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* (2013) doi:10.1016/j.celrep.2012.12.008.

3. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* (2020) doi:10.1038/s41586-020-1943-3.

4. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* (2016) doi:10.1038/nrg.2016.59.

5. Amit, Y. & Geman, D. Shape Quantization and Recognition with Randomized Trees. *Neural Comput.* (1997) doi:10.1162/neco.1997.9.7.1545.

6. Anurag, M. *et al.* Comprehensive profiling of DNA repair defects in breast cancer identifies a novel class of endocrine therapy resistance drivers. *Clin. Cancer Res.* (2018) doi:10.1158/1078-0432.CCR-17-3702.

7. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* (2010) doi:10.1214/09-SS054.

8. Artandi, S. E. & DePinho, R. A. Telomeres and telomerase in cancer. *Carcinogenesis* (2009) doi:10.1093/carcin/bgp268.

9. Baretti, M. & Le, D. T. DNA mismatch repair in cancer. *Pharmacology and Therapeutics* (2018) doi:10.1016/j.pharmthera.2018.04.004.

10. Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* (2010) doi:10.1126/science.1191078.

11. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis. Oncol.* (2017) doi:10.1200/po.17.00073.

12. Brown, A. J., Mao, P., Smerdon, M. J., Wyrick, J. J. & Roberts, S. A. Nucleosome positions establish an extended mutation signature in melanoma. *PLoS Genet.* (2018) doi:10.1371/journal.pgen.1007823.

13. Brown, A. L., Li, M., Goncearenco, A. & Panchenko, A. R. Finding driver mutations in cancer: Elucidating the role of background mutational processes. *PLoS Comput. Biol.* **15**, e1006981 (2019).

14. Cairns, R. A. & Mak, T. W. Oncogenic isocitrate dehydrogenase mutations: Mechanisms, models, and clinical opportunities. *Cancer Discovery* (2013) doi:10.1158/2159-8290.CD-13-0083.

15. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* (2020) doi:10.1038/s41586-020-1969-6.

16. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science (80-.).* **362**, (2018).

17. Dekker, J. & Misteli, T. Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.* (2015) doi:10.1101/cshperspect.a019356.

18. Diaz, M. & Casali, P. Somatic immunoglobulin hypermutation. *Current Opinion in Immunology* (2002) doi:10.1016/S0952-7915(02)00327-8.

19. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* (2016) doi:10.1016/j.molcel.2016.05.018.

20. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature Reviews Cancer* (2003) doi:10.1038/nrc969.

21. Dubitzky, W., Granzow, M. & Berrar, D. *Fundamentals of data mining in genomics and proteomics*. *Fundamentals of Data Mining in Genomics and Proteomics* (2007). doi:10.1007/978-0-387-47509-7.

22. Fernández-Medarde, A. & Santos, E. Ras in cancer and developmental diseases. *Genes and Cancer* (2011) doi:10.1177/1947601911411084.

23. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* (2017) doi:10.1126/science.aal2380.

24. Gan, L. *et al.* Epigenetic regulation of cancer progression by EZH2: From biological insights to therapeutic potential. *Biomarker Research* (2018) doi:10.1186/s40364-018-0122-2.

25. Gates, L. A., Foulds, C. E. & O'Malley, B. W. Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends in Biochemical Sciences* (2017) doi:10.1016/j.tibs.2017.10.004.

26. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* (2020) doi:10.1038/s41586-019-1907-7.

27. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).

28. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* (2018) doi:10.1038/s41467-018-03828-2.

29. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* (2000) doi:10.1016/S0092-8674(00)81683-9.

30. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* (2011) doi:10.1016/j.cell.2011.02.013.

31. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* (2016) doi:10.1016/j.cell.2015.12.050.

32. Hastie, T., Tibshirani, R. & Friedman, J. *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. *Springer* (2009). doi:10.1007/b94608.

33. Heidelbaugh, J. J. & Bruderly, M. Cirrhosis and chronic liver failure: Part I. Diagnosis and evaluation. *American Family Physician* (2006).

34. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* (2007) doi:10.1038/ng1966.

35. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* (1998) doi:10.1109/34.709601.

36. Ho, T. K. Random decision forests. in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (1995). doi:10.1109/ICDAR.1995.598994.

37. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* (2018) doi:10.1016/j.cell.2018.03.022.

38. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* (2010) doi:10.1038/nature08987.

39. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* (2020) doi:10.1038/s41586-020-1969-6.

40. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* (2020) doi:10.1038/s41467-019-13825-8.

41. Kanherkar, R. R., Bhatia-Dey, N. & Csoka, A. B. Epigenetics across the human lifespan. *Frontiers in Cell and Developmental Biology* (2014) doi:10.3389/fcell.2014.00049.

42. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* (2019) doi:10.1186/s13059-019-1654-y.

43. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: Quantifying genome and methylome mappability. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky677.

44. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* (2015) doi:10.1038/ng.3335.

45. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* (2016) doi:10.1038/ng.3557.

46. Kim, T. H. *et al.* TERT promoter mutations and long-term survival in patients with thyroid cancer. *Endocr. Relat. Cancer* (2016) doi:10.1530/ERC-16-0219.

47. Kleinberg, E. M. On the algorithmic implementation of stochastic discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.* (2000) doi:10.1109/34.857004.

48. Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* (2001) doi:10.1038/35101031.

49. Kübler, K. *et al.* Tumor mutational landscape is a record of the pre-malignant state. *bioRxiv* (2019) doi:10.1101/517565.

50. Kuijk, E. *et al.* Early divergence of mutational processes in human fetal tissues. *Sci. Adv.* (2019) doi:10.1126/sciadv.aaw1271.

51. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* (2001) doi:10.1038/35057062.

52. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* (2013) doi:10.1038/nature12213.

53. Lee, N., Maurange, C., Ringrose, L. & Paro, R. Suppression of Polycomb group proteins by JNK signalling induces transdetermination in Drosophila imaginal discs. *Nature* (2005) doi:10.1038/nature04120.

54. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* (2019) doi:10.1038/s41586-019-1672-7.

55. Li, B. & Chng, W. J. EZH2 abnormalities in lymphoid malignancies: Underlying mechanisms and therapeutic implications. *Journal of Hematology and Oncology* (2019) doi:10.1186/s13045-019-0814-6.

56. Lind, A. P. & Anderson, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS One* (2019) doi:10.1371/journal.pone.0219774.

57. MacPherson, D. & Dyer, M. A. Retinoblastoma: From the two-hit hypothesis to targeted chemotherapy. *Cancer Research* (2007) doi:10.1158/0008-5472.CAN-07-0276.

58. MANSOURI, A., KARAMCHANDANI, J. & DAS, S. Molecular Genetics of Secondary Glioblastoma. in *Glioblastoma* (2017). doi:10.15586/codon.glioblastoma.2017.ch2.

59. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology* (2014) doi:10.1038/nrm3822.

60. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Medicine* (2019) doi:10.1186/s13073-019-0648-4.

61. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science (80-.).* (2018) doi:10.1126/science.aau3879.

62. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

63. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* (2015) doi:10.1126/science.aaa6806.

64. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *bioRxiv* 505685 (2018) doi:10.1101/505685.

65. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* (2012) doi:10.1016/j.cell.2012.04.024.

66. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nature Reviews Immunology* (2006) doi:10.1038/nri1896.

67. Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology* (2010) doi:10.1101/cshperspect.a001008.

68. Peters, A. H. F. M. *et al.* Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin. *Mol. Cell* (2003) doi:10.1016/S1097-2765(03)00477-5.

69. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* (2019) doi:10.1016/j.cell.2019.02.012.

70. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0525-5.

71. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* (2018) doi:10.1016/j.cell.2018.10.004.

72. Pilie, P. G. *et al.* Genomic instability and DNA damage repair in clear cell renal cell carcinoma. *J. Clin. Oncol.* (2017) doi:10.1200/jco.2017.35.15_suppl.4581.

73. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

74. Poulos, R. C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Rep.* (2016) doi:10.1016/j.celrep.2016.11.055.

75. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* (2019) doi:10.1038/s41586-019-1689-y.

76. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* (2020) doi:10.1038/s41586-020-1965-x.

77. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

78. Roche, J. The epithelial-to-mesenchymal transition in cancer. *Cancers* (2018) doi:10.3390/cancers10020052.

79. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* (2019) doi:10.1371/journal.pcbi.1006953.

80. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* (2018) doi:10.1016/j.cell.2018.03.035.

81. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

82. Shaib, Y. H., El-Serag, H. B., Davila, J. A., Morgan, R. & Mcglynn, K. A. Risk factors of intrahepatic cholangiocarcinoma in the United States: A case-control study. *Gastroenterology* (2005) doi:10.1053/j.gastro.2004.12.048.

83. Shaikhina, T. *et al.* Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed. Signal Process. Control* (2019) doi:10.1016/j.bspc.2017.01.012.

84. Shieh, G. Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organ. Res. Methods* (2008) doi:10.1177/1094428106292901.

85. Sohn, B. H. *et al.* Clinical significance of four molecular subtypes of gastric cancer identified by The Cancer Genome Atlas project. *Clin. Cancer Res.* (2017) doi:10.1158/1078-0432.CCR-16-2211.

86. Song, L. & Crawford, G. E. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* (2010) doi:10.1101/pdb.prot5384.

87. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* (2009) doi:10.1038/ng.363.

88. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* (2015) doi:10.1038/nature14173.

89. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* (2019) doi:10.1016/j.dnarep.2019.102647.

90. Suvà, M. L., Riggi, N. & Bernstein, B. E. Epigenetic reprogramming in cancer. *Science* (2013) doi:10.1126/science.1230184.

91. Svejstrup, J. Q. Mechanisms of transcription-coupled DNA repair. *Nature Reviews Molecular Cell Biology* (2002) doi:10.1038/nrm703.

92. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* (2011) doi:10.1016/j.cell.2011.08.008.

93. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* (2018) doi:10.1186/s13059-018-1509-y.

94. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* (2013) doi:10.1126/science.1235122.

95. Wallace, S. S., Murphy, D. L. & Sweasy, J. B. Base excision repair and cancer. *Cancer Letters* (2012) doi:10.1016/j.canlet.2011.12.038.

96. Xinarianos, G. *et al.* hMLH1 and hMSH2 expression correlates with allelic imbalance on chromosome 3p in non-small cell lung carcinomas. *Cancer Res.* (2000).

97. You, J. S. & Jones, P. A. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* (2012) doi:10.1016/j.ccr.2012.06.008.

98. Zandi, R., Larsen, A. B., Andersen, P., Stockhausen, M. T. & Poulsen, H. S. Mechanisms for oncogenic activation of the epidermal growth factor receptor. *Cellular Signalling* (2007) doi:10.1016/j.cellsig.2007.06.023.

99. Zhu, H. *et al.* Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. *Mol. Cell* (2020) doi:10.1016/j.molcel.2019.12.027.