Convergence of regulatory mutations into oncogenic pathways across multiple tumor types

by

Karthik Murugadoss

B.Tech., Indian Institute of Technology, Madras (2015)

Submitted to the Center for Computational Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Signature redacted

Author . . . .
Center for Computational Engineering
May 12, 2017

Signature redacted

Certified by. . . . .
Manolis Kellis
Professor
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by . . . .
Nicolas Hadjiconstantinou
Professor
Co-Director, Computation for Design and Optimization

Convergence of regulatory mutations into oncogenic pathways across multiple tumor types

by

Karthik Murugadoss

Submitted to the Center for Computational Engineering
on May 12, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Computation for Design and Optimization

Abstract

Cancer sequencing efforts have largely focused on profiling somatic variants in the protein-coding genome and characterizing their functional impact. In this study, we develop a computational pipeline to identify non-coding mutational drivers across multiple tumor types. We describe the non-coding mutational profiles of 854 samples, spread across 15 tumor types, in the context of their respective tissue type-specific reference epigenomes, using recent pan-cancer whole-genome sequencing data. We develop a novel method to detect significantly recurrent non-coding mutations by re-estimating the background mutation density while adjusting for epigenomic covariates. Existing databases on enhancer-gene links allow us to capture the convergence of disparate mutations onto downstream target genes. We then systematically identify key immunomodulatory and tumor-suppressive genes enriched for non-coding mutations in their regulatory neighborhood and evaluate these in a pan-cancer context. Taken together, we show that low-frequency alterations converge into high-frequency recurrent events on downstream targets through tissue-specific regulatory interactions.

Thesis Supervisor: Manolis Kellis
Title: Professor
Department of Electrical Engineering and Computer Science

Acknowledgments

This thesis would not have been possible without the mentorship of Prof. Manolis Kellis whose guidance and thoughful feedback were critical throughout the project. I would like to express my gratitude to Dr. Nicola Rinaldi, Dr. Jose Davilla and other members in the MIT Computational Biology group for the numerous insightful discussions. I would also like to thank my friends and family for their consistent support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Large scale sequencing projects such as the The Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC) [2] have performed molecular characterization of thousands of tumor samples. These projects have primarily focused on unravelling the genomic variation in protein coding sequences owing to the reduced costs of exome sequencing as opposed to whole-genome sequencing. Our understanding of somatic variation in coding regions and their associated functional relevance has dramatically improved resulting in improved patient stratification in the clinic as well as development of targeted therapies in certain cancer types. However, the protein-coding component of the genome accounts for less than 2% of the total sequence length [3]. Additionally, a number of well-studied cancer types contain sub-populations where the observable protein-coding mutations are either nonfunctional or affect genes unrelated to tumor development. These challenges have motivated the exploration of the non-coding cancer genome whose impact on tumor growth and progression is yet to be completely deciphered.

Unravelling the effect of sequence variants in non-coding regions requires functional characterization of genomic regions with regulatory potential. Sequence conservation (similarity) across species is a popular approach to identify functionally relevant genomic segments. More recently, rapid growth of the field of epigenomics has led to the detailed annotation of active and inactive regulatory elements across the genome [4]. Histone modifications and associated chromatin states enable cell-type specific control of gene expression and downstream biological processes.

Numerous experimental techniques have been developed to explore the epigenome. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) assays allow identification of occupied transcription factor (TF) binding sites in addition to pinpointing sites of histone modifications [5]. TFs bind DNA in regions of open (non-nucleosomal) chromatin, which can be identified using DNase I hypersensitivity assays [6]. DNA methylation is a crucial process that modulates activity of genomic segments [7]. DNA methylation can be profiled across the genome through methods such as the 450k array, where over 450,000 methylation sites are analyzed at single-nucleotide resolution, or whole genome bisulfite sequencing, where DNA is treated with bisulfite prior to sequencing [8]. Most of the experimental assays for profiling the epigenome are capable of parallel sequencing of multiple samples and tissue types.

Recent disease association studies have implicated non-coding elements in disease etiology and progression; we therefore expect that non-coding mutations in cancer likewise have a regulatory effect on tumorigenesis [9, 10]. Recurrent mutations in non-coding regions linked to genes relevant for cancer growth support this notion. For instance, a regulatory mutation was identified in the region upstream of the TERT gene (encoding telomerase reverse transcriptase) creating binding motifs for the ETS family of TFs [11, 12]. The ETS family has been previously associated with prostate and breast cancer and is known to be involved in cellular proliferation, differentiation and apoptosis [13]. A SNP in the MDM2 promoter is associated with increased tumor growth across multiple tissue types [14]. SNPs discovered in a gene desert on chromosome 8q24 were found to localize to regions that act as enhancers for the MYC gene in a tissue specific manner [15]. MYC is dysfunctional across a set of tumor types including lung, breast, and colon cancers.

Despite these promising discoveries, heterogeneity in the mutational landscape within and across tumor types makes identifying genome-wide non-coding mutations linked to oncogenesis challenging. Furthermore, the detection of putative driver mutations in cancer, defined as mutations that confer a fitness advantage to the cell, is in itself a difficult task and an active area of research. There are two approaches to flag a potentially oncogenic driver mutation [3, 16]. The first revolves around capturing signals of positive selection that manifest as recurrent mutations. The second technique predicts the functional impact of variants by incorporating protein structural information and evolutionary conservation. Detection of cancer drivers incorporating these frameworks has been widely published. However, these techniques have primarily been limited to discovering protein-coding mutations due to availability of large scale exome sequencing datasets as well as ease of interpretability. Although pan-cancer methods have improved the discovery and analysis of regulatory mutations while avoiding the type I and type II errors made in several tissue-specific cancer projects, the analysis of driver mutations in non-coding elements has not yet been perfected. Long range regulatory links and the three-dimensional folding of the genome, which play a crucial role in establishing the functional potential of a particular variant, add additional complexity [17]. Multiple regulatory loci are found, using these methods, to associate with the same target gene.

1.2 Thesis Overview

In this study, we develop a computational pipeline to identify non-coding mutational drivers in various cancer types. By leveraging recent pan-cancer whole-genome sequencing efforts, we characterize the non-coding mutational profiles of 854 samples, spread across 15 tumor types as described in Chapter 2. Despite widely different mutational burden both within and across tissues, we observe shared patterns of mutational signatures across tumor samples. Following this analysis, Chapter 3 focuses on annotating each sample with the appropriate tissue type-specific reference epigenome which allows us to describe the mutational landscape along with its associated regulatory context. In Chapter 4, we develop a novel framework for detecting significantly recurrent non-coding mutations by re-estimating the background mutation density, factoring in epigenomic covariates such as histone modifications, DNase hypersensitivity, and DNA replication timing. This improved paradigm, centered on an epigenomically-adjusted mutational background, identifies existing as well as novel recurrent non-coding mutations. Additionally, we expand on this model to define active regulatory segments of the genome that harbor significant mutational burden in a tissue-specific manner.

From previous studies and existing databases on enhancer-gene links [18], we extend our framework to comprehensively characterize the impact of distal enhancer regions carrying significant mutational burden on their downstream target genes. This allows us to develop a robust and systematic framework to identify key cancer-related networks enriched for non-coding mutations and evaluate these in a pan-cancer context. Therefore, in Chapter 5, we show that low-frequency alterations converge into high-frequency recurrent events on downstream targets through tissue-specific regulatory interactions. Our proposed model presents an unbiased method to characterize the impact of regulatory mutations on oncogenesis. We anticipate that this approach will improve interpretation of future functional non-coding mutational dissections in tumor-related studies.

Chapter 2

Studying the pan-cancer whole genome mutational landscape

Recent improvements in whole genome sequencing have enabled deep characterization of the mutations in tumors obtained from both solid and liquid biopsies [19, 20]. The cancer genome comprises germline variants that define susceptibility to cancer, as well as somatic mutations that drive cancer growth and proliferation. A genetic predisposition to cancer caused by a germ-line variant exists in an estimated 5-10% of all tumor patients. A well-known example of clinically relevant cancer susceptibility syndrome is the case of breast / ovarian carcinoma caused by alterations in the BRCA1/2 genes [21]. However, identifying similar robust patterns of germ-line predisposition across different tumor types remains challenging. Somatic mutations occur due to spontaneously occurring mutations that accumulate in somatic cells over a person's lifetime. Although these alterations are typically non-functional, a fraction of them can give rise to preferential growth or survival of a cell eventually resulting in a malignant cellular state. The transformation of this cellular clone into a disseminated disease is an intricate process whose complexity is linked to the heterogeneity of the mutational landscape. Recent sequencing studies have revealed that somatic mutations commonly occur in oncogenes and therefore have a direct impact on tumorigenesis [22].

Somatic mutations can arise from DNA damage, incorrectly or incompletely repaired DNA, or from errors in the DNA replication process [23]. DNA damage is commonly caused by external factors such as ultraviolet (UV) light, ionizing radiation, chemicals or reactive oxygen species. Enzymatic action within cells can also result in modification of DNA. Additionally, viruses and endogenous retrotransposons can give rise to insertions in the DNA sequence. These mutational processes occur in a cell type-specific manner and impact regional mutation density to eventually result in unique mutational signatures. Each process will cause a specific type of somatic mutation; for instance, UV light causes C>T mutations at TpC or CpC dinucleotides while the chemical carcinogen aristolochic acid causes A>T substitutions [24, 25]. However, the extent to which a tumor is affected by mutational processes depends on variables such as the age and physiology of the patient, and the existence of germ-line predispositions.

2.1 Characterization of overall mutational burden across 854 samples

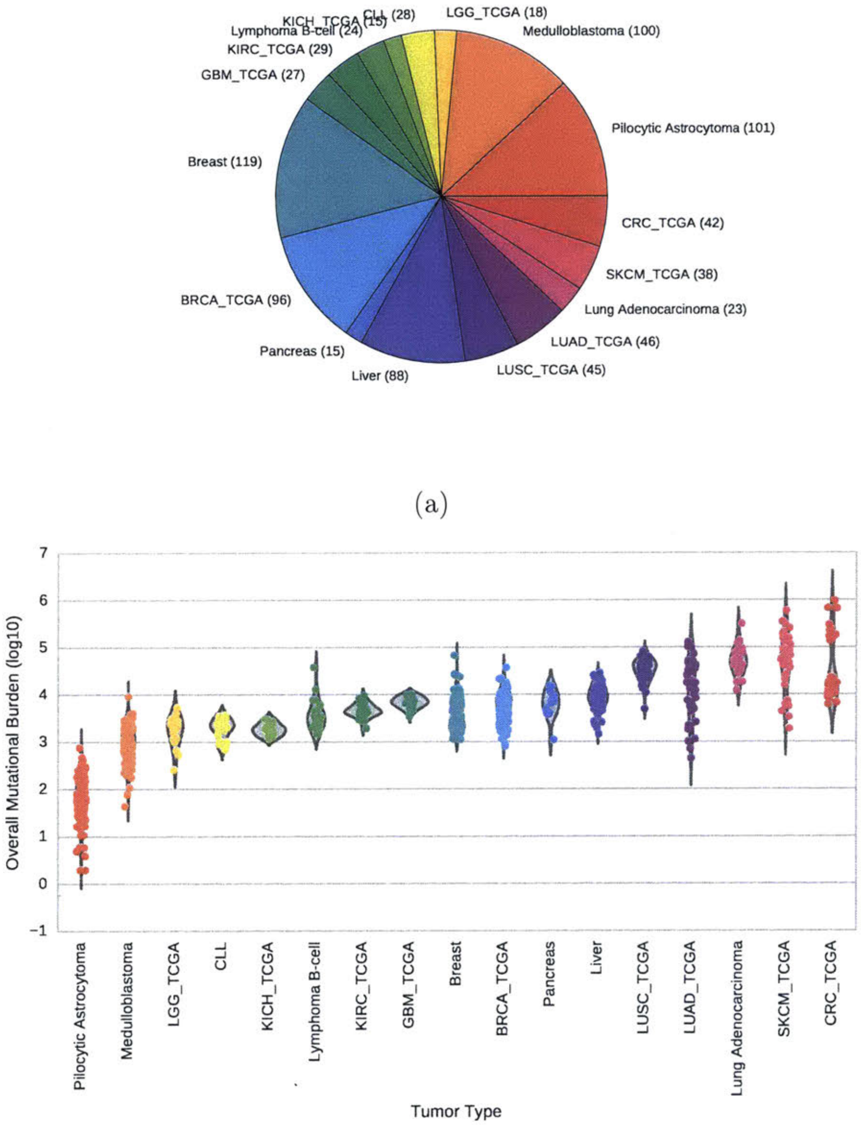
We first systematically evaluate the mutational landscapes of 854 tumor samples across 15 tissue types obtained from two independent cohorts (identified as the TCGA cohort and the Alexandrov cohort, see Supplementary Table B.1). Breast cancer and lung adenocarcinoma samples were available from both datasets and were processed independently for our analysis enabling us to use them as separate discovery and validation sets. Therefore, 17 groups of tumor samples were considered for our study (Fig. 2-1a).

Whole-genome somatic mutation burden is highly variable within and between cancer types (Fig. 2-1b). We observe that the cancer types carrying high mutational burden correspond to tissues that originate from surface epithelia with high rates of turnover such as colorectal cancer, cutaneous melanoma, and lung adenocarcinoma. These tissue types are also commonly subject to exposure by numerous exogenous factors, such as UV rays in the case of melanoma and tobacco in the case of lung cancer. This manifests as a large intra-tissue variance in mutational frequency. Both the colorectal cancer and melanoma cohorts contain a number of hypermutated samples ($\sim$ 600,000 mutations per sample as compared to a mean mutation count of 117,896 and 94,655 respectively), where increased mutation rates could potentially be due to large-scale dysfunction in repair pathways or loss of chromosome integrity [26]. The other end of the spectrum, corresponding to samples with low mutational burden (100-2000 mutations per sample), is represented by a set of brain tumors: pilocytic astrocytoma, medulloblastoma, and lower grade glioma. Additionally, we observe that the breast cancer cohorts obtained from the two studies closely represented each other in mutational burden. However, the lung adenocarcinoma samples from TCGA display significantly higher variance than the samples from the Alexandrov cohort. Mean mutation count and within-tissue standard deviations are reported in Supplementary Table B.2.

2.2 Mutation context analysis and hierarchical clustering

We use whole-genome sequencing data to examine the similarity in mutational signatures across the entire dataset. As described previously [23], all types of base substitution were referred to by the pyrimidine of the mutated Watson-Crick base pair, resulting in six classes of substitution: C>A, C>G, C>T, T>A, T>C, T>G - which together define the mutational spectrum of each sample. We look at the context of each mutation represented by the frequency of A, T, C, and G nucleotides 2 bp 5' and 3' to each variant. The pairwise correlation of mutation context was computed followed by hierarchical clustering across each individual sample for the six mutational categories. We observe that the context analysis around the C>T substitutions results in samples of the same tumor type grouping together despite widely different overall mutational burden (Fig. 2-2). In particular, breast and lung cancers co-cluster with liver and pancreatic cancers while kidney and blood cancers group together. Additionally, samples from colorectal carcinoma fall in the same cluster despite a large variance in mutational burden. Context analysis around the C>G mutation results in robust clustering for some tumor types while samples from other types are dispersed. Analyzing mutational context for the remaining four mutational categories did not yield robust clustering of samples (see Supplementary Figure B-1).

In summary, strong intra- and inter-tissue mutational heterogeneity creates variability in mutational burden that differ by orders of magnitude. Therefore, characterization of genome-wide somatic variants based on overall mutational load alone is a challenging task. Tissue type-specific mutational processes result in unique mutational signatures that enables robust clustering of samples of the same tumor despite diverse mutational burden. To further explore the selective pressure inherent in these mutational processes, functional annotation of the genome would be required, and this forms the subject of the following chapter.



(a)

(b)

Figure 2-1: Description of pan-cancer dataset and overall mutation burden: (a) Tumor types (number of samples) included in this study. (b) Intra- and inter-tissue variability in genome-wide somatic mutation burden observed across 15 distinct tumor types. Tumors are ordered by median mutation burden. Tumor types from TCGA are identified by the '_TCGA' suffix. Expansion of abbreviations are provided in Supplementary Table B.1.

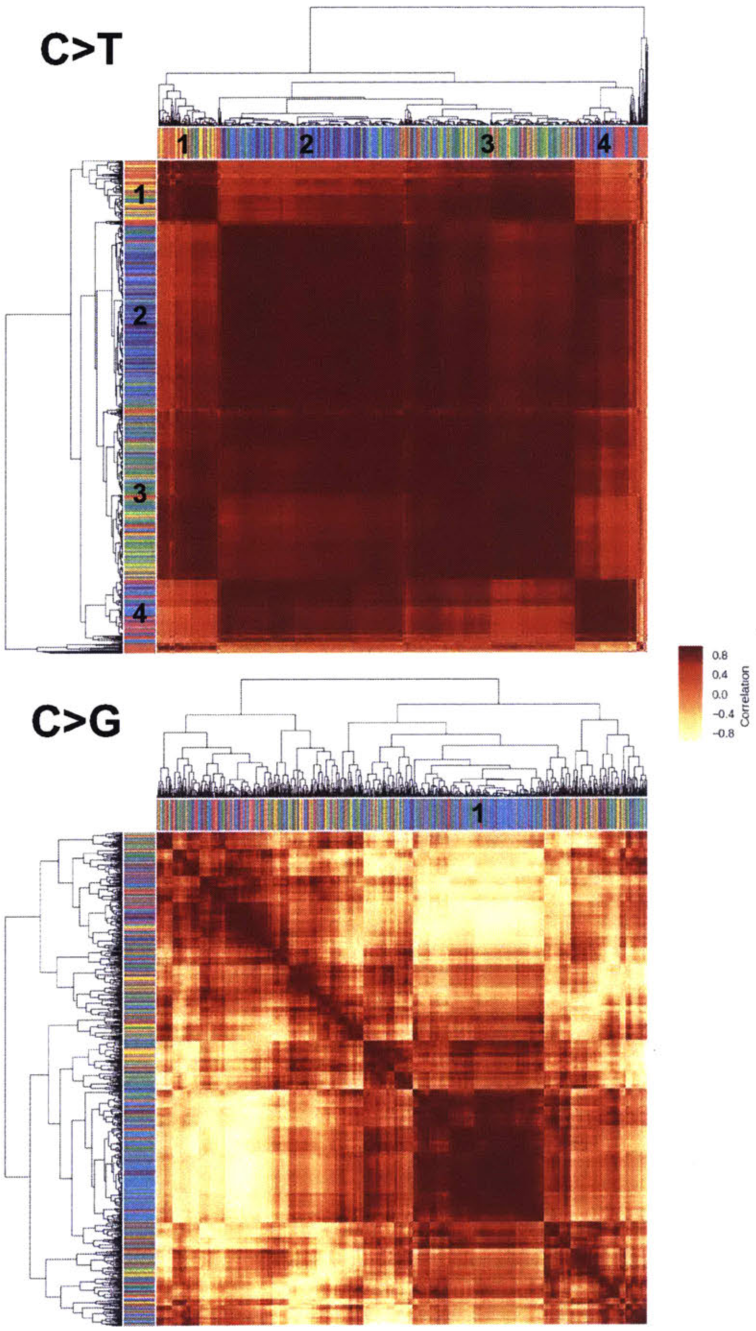


Figure 2-2: Analysis of mutation context: Sequence context analysis for C>T and C>G substitutions across 854 samples. C>T context clustering (left) captures distinct clusters represented by (1) Pilocytic Astrocytoma, Medulloblastoma and Lower grade glioma, (2) Pancreatic, Liver, Breast and Lung cancers, (3) Kidney and blood cell cancers and (4) Melanoma and Colorectal cancers. The C>G clustering (right) reveals (1) Breast and Lung cancers as the only robust group. Color bars along the axis correspond to the same coloring scheme illustrated in 2-1b.

Chapter 3

Pan-cancer epigenomic annotation and relationship with regional mutational burden

The conventional theory of evolution assumes a uniform distribution of mutations across the genome. The longevity of a genomic alteration is determined through natural selection. While selective pressure remains a factor in structuring the genomic landscape, the notion that DNA variants occur uniformly and randomly has become outdated. Mutation density across the genome is highly heterogeneous [27, 28]. Local nucleotide context, chromatin architecture and replication timing have been shown to influence the influence of single nucleotide changes. Of these covariates, chromatin organization is the dominant factor in determining regional variation in mutational density [29, 30]. For instance, regions in open chromatin can show dramatically lower mutational background due to higher accessibility to DNA repair machinery.

Given the non-uniform regional distribution of mutational occurrence, a number of functional regulatory sites across the genome have been found to harbor mutations subject to positive selection. Evidence for this observation arises from the identification of mutations in the TERT promoter region and in the linked regulatory regions of PLEKHS1, WDR74 and SDHD [11, 31]. Additionally, recurrent germline mutations have been reported in the CCND1 enhancer [32]. Recently, GNAS, BCL11B, ANKRD11 and NEDD4L, also cancer-associated genes, were found to harbor mutations in their regulatory loci [33].

3.1 Chromatin state-specific sample-specific mutational enrichment

Characterizing the genome-wide incidence of somatic mutations is a challenging task. The chromatin organization of each cell type is distinct, thereby resulting in unique mutational landscapes and specific patterns of gene expression. The Roadmap Epigenomics Consortium has produced reference epigenomes that reveal key insights on active functional elements that control gene expression and cellular identity in 127 human tissues and cell types including adult and embryonic tissues in both diseased and healthy individuals [4]. In addition to deeply profiled epigenomic marks, the Roadmap project includes chromatin state annotations for each cell type characterized by the ChromHMM method [34]. The ChromHMM framework is a multivariate Hidden Markov Model that integrates multiple ChiP-seq datasets to model the occurrence of each chromatin mark. Additionally, the Roadmap study highlighted that different histone marks show distinct levels of DNA methylation and accessibility, and were predictive of gene expression changes. The dynamic nature of the histone marks across multiple chromatin states presented an opportunity for a data-driven approach to infer biologically meaningful relationships specific to each cell type. Therefore, the Roadmap project has comprehensively characterized the epigenomic landscape and its associated heterogeneity across multiple cell types and significantly improves our understanding of the regulatory genome.

Epigenomic architecture has been known to impact the mutational landscape of different tumor types [35]. Megabase scale variation in mutation density across the genome is strongly associated with regional variation in chromatin organization [30]. In particular, repressive chromatin features are indicative of regions harboring high mutational burden while epigenomic features linked to transcription and active chromatin are associated with low mutation density. A recent study has determined that epigenomic features derived from the cell type of origin are the strongest determinants of cancer mutations and display higher predictive accuracy than matched cancer cell lines [29]. There are two possible explanations for this observation. First, somatic mutations may arise before the epigenetic landscape changes. Second, over the course of development, the epigenome of individual tumors may transform uniquely thus separating them from other tumors of the same tissue type.

For this study, we select the most closely matching primary cell type of origin for each tumor type from the Roadmap epigenomics project (see Supplementary Table B.1) [4]. The 15 chromatin state model as described previously is utilized. This model encompasses 8 active states and 7 inactive states (see Supplementary Table B.3). Mutational enrichment per megabase normalized by the overall mutation burden per sample is computed for the 15 states across each of the 854 samples (Fig. 3-1). Hierarchical clustering on the enrichment profiles reveals tissue type-specific clustering and dispersion patterns. In particular, we notice that hypermutated samples of melanoma or colorectal cancer (Fig. 3-1, indicated by 'a') showed strong similarities in their chromatin state-specific mutational enrichment. Samples with a low mutation rate corresponding to pilocytic astrocytoma or lower grade glioma (Fig. 3-1, indicated by 'b') also display highly correlated enrichment profiles. The TxFlnk (transcription at gene 5' and 3'), BivFlnk (flanking bivalent transcription start site / Enhancer) and TssBiv (Bivalent / poised transcription start site) chromatin states on average harbor the least mutational burden across tumor types whereas the Het (Heterochromatin) and the ZNF/Rpts (ZNF genes, repeats) states show higher enrichment for mutations on average.

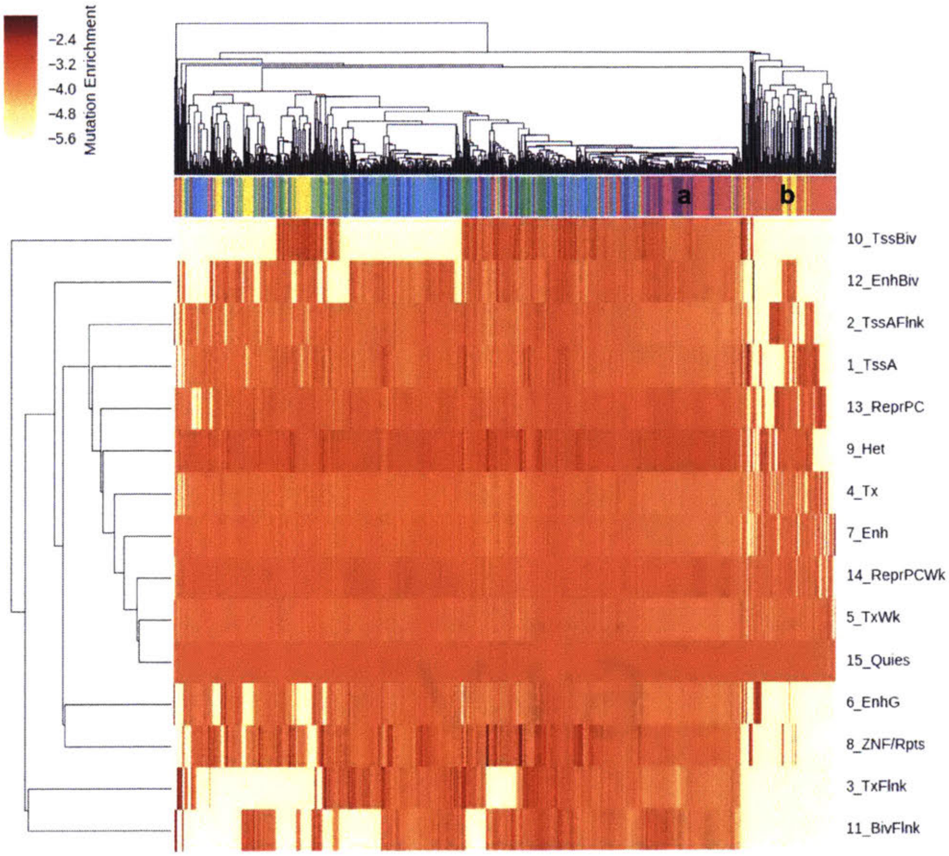


Figure 3-1: Chromatin state-specific sample-specific mutational enrichment. Mutation enrichment is calculated for each chromatin state (columns) as defined by the Roadmap Epigenomics project and for each tumor sample (rows) as the number of somatic mutations localizing to a specific chromatin state in that sample normalized by (1) the overall sequence length of that chromatin state, and (2) the overall mutational burden of that sample.

3.2 Correlation of regional mutation density with replication timing

The temporal control of the DNA replication is referred to as the replication-timing program. There is increasing evidence that sequential changes in replication timing often accompany cancer growth and progression potentially acting in positive feedback to direct the cell away from its normal state [36]. Although DNA replication is a highly regulated and stable process in normal cells, cancer cells exhibit asynchronous replication. Additionally, this asynchrony in replication has been described for multiple oncogenes across the genome, indicating genome-wide dysregulation. The effects of aberrant replication timing are multi-fold. In addition to alterations in the epigenome or downstream gene expression, distinct genomic events have been observed. For instance, localized changes in replication timing near translocation breakpoints lead to localized mutagenesis, termed 'kataegis' in that region. A number of recent studies have verified this correlation between regional mutation densities and replication timing [27, 30, 37]. In particular, later-replicating regions have been found to harbor increased mutational burden [38]. Therefore, the correlation between regionally stratified replication timing profile and the local mutation density is of value when exploring the mutational landscape across tumor types.

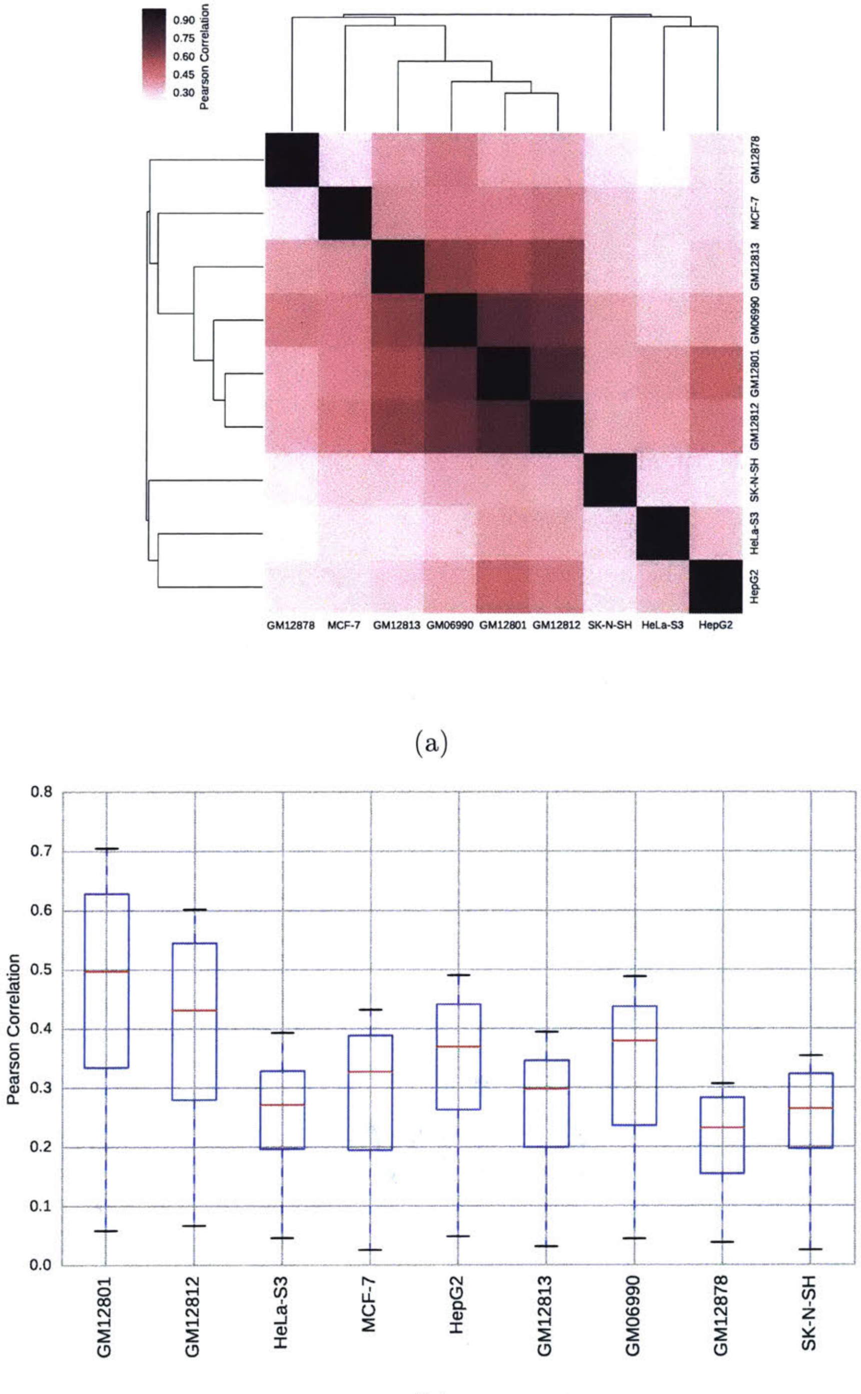
Repli-Seq is a genome-scale method that maps temporally ordered replicating DNA and enables analysis of replication timing patterns specific to cell-cycle time and genomic position. In this study, Repli-Seq data from 9 ENCODE cell lines are used to determine if significant changes in regional mutation rate across the 15 tumor types were associated with replication timing. However, since Repli-Seq profiles are not available for each tumor type, we choose a single representative cell type to define the genome-wide variation in replication timing. We observe that genome-wide replication timing for the ENCODE cell lines display on average good correlation (with a Pearson correlation mean of 0.46 and variance of 0.05) with each other (Fig 3-2a). The mean correlation of the Repli-Seq signal of GM12801 (Lymphoblastoid cell line) with the tumor type-specific 1Mb mutation densities is higher than that for other ENCODE cell lines (Fig 3-2b). Therefore, for further analysis we utilize the Repli-Seq profile from GM12801 as a proxy for replication timing across cancer types.

3.3 Feature importance analysis of epigenomic profiles through Random Forest regression

The availability of diverse epigenomic features (histone marks, DNase hypersensitivity, and replication timing) allows us to determine which features capture the most variation in regional mutational burden in a tissue type-specific manner. We employ Random Forest regression, a non-parametric framework that integrates the result of an ensemble of regression trees to infer the value of a continuous response variable. Tree-based models have been shown to approximate functions with any shape, whereas linear models can only fit combinations of linear functions [39, 40]. Multiple regression trees make the model robust to outliers and noise in the input data. A subset of the observations used for training is drawn from the dataset, with replacement, for each regression tree. The remaining data corresponds to the test set and are used to calculate the mean squared prediction error of the tree, which is then averaged over all trees to provide the error for each observation. The difference between the mean squared error computed on the test set and the mean squared error computed upon random permutation of each of the predictor variables is averaged over all trees and normalized by the standard error. This represents the raw importance score for each feature, indicating in this case how closely the particular feature is associated with mutational density across the genome.

The importance scores for available epigenomic features are computed for each tissue type (Fig 3-3). These features are constructed by counting the number of peaks in the NarrowPeak profile over discrete 1Mb bins. H3K9me3 and the replication timing are found to be the most explanatory across tumor types. Lysine 9 in histone H3 plays an activating as well as repressive role by switching on genes when acetylated and silencing them when methylated. H3K9me3 is known to cause binding of heterochromatin protein 1 (HP1) to constitutive heterochromatin [41]. HP1 has been functionally linked to transcriptional repression. Therefore, association of H3K9me3 with mutational density in select tumor types implies a larger incidence of mutations in inactive regions. These mutations could potentially result in de-repression of normally silenced genes. Similar mechanisms revolving around epigenetic activation of oncogenes have been previously reported in literature [35, 42]. We also observe that, for most tumor types, either H3K9me3 or replication timing are strongly explanatory. In the case of breast cancers from either cohort, H3K9me3 possesses high importance score whereas other histone profiles have low contribution. This could indicate potentially distinct mechanisms of action for cancer growth or progression. Of particular interest is the case of lung adenocarcinoma, where the TCGA cohort shows greater explanatory capacity by H3K9me3 whereas the Alexandrov cohort shows a higher importance score for replication timing despite the tumor type as well as the cell-of-origin epigenome (E096) being identical.

A generalization of this analysis would be to infer the reference epigenome and its associated cell type instead of imposing a cell-of-origin specificity for the epigenome. This framework could potentially be useful in identifying the tissue of origin of cancers whose biopsies have been obtained from a metastatic site. The method considers a unified feature set of all histone marks, DNase hypersensitivity, and replication timing, and selects a cell type based on its representation in the overall ranked correlation list of each of the epigenomic features. As before, the target variable is the binned mutation density. We observe that the performance of this framework was not consistent across all tumor types and in particular was poor in cases of low overall mutational burden. This dependence of the regression fit on genome-wide mutational load has been reported earlier [29]. In particular, the decline in model performance in a highly sparse mutational landscape also made the interrogation of feature importances and cell-of-origin resemblance on a sample-specific basis challenging. Random forest regression has enabled the prioritization of epigenomic profiles that are more strongly linked to mutation incidence. However, we do not utilize regression coefficients or feature importances in our model for computing the epigenomically adjusted mutational background described in the following chapter.



(a)

(b)

Figure 3-2: Analysis of replication timing profiles (Repli-Seq) from 9 cell types from ENCODE. (a) Correlation of 1Mb binned genome-wide replication timing for 9 cell lines. (b) 1Mb binned genome-wide correlation between replication timing and tumor-specific mutation density.

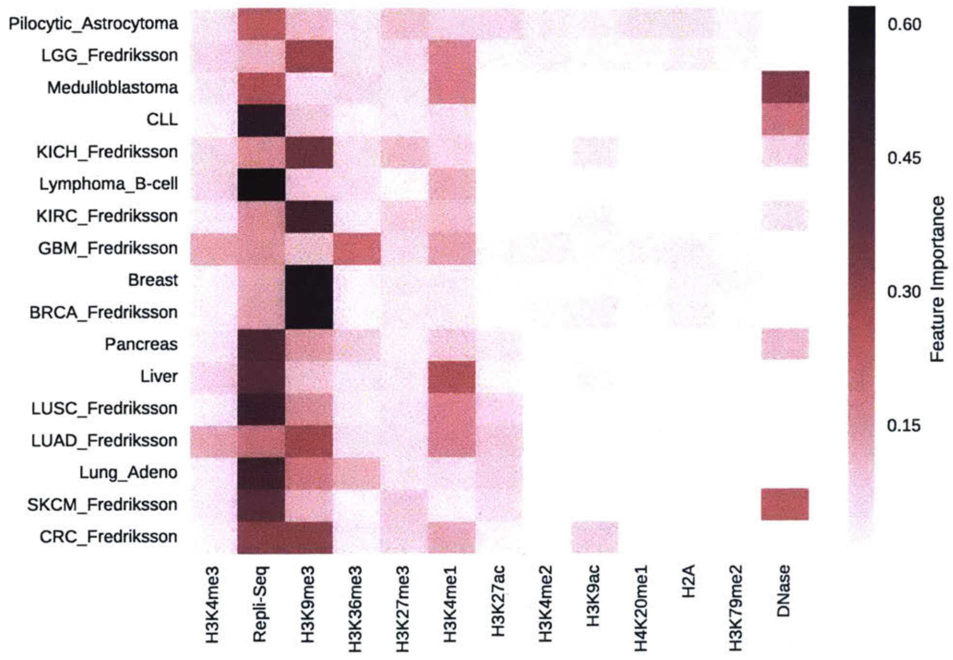


Figure 3-3: Contribution of different epigenomic features towards mutation density across tumor types: H3K9me3 and GM12801 replication timing (Repli-Seq) are most explanatory of the 1Mb binned mutational background across tumors. We also observe tumor specific features of high importance such as H3K4me1 in the case of liver cancer or DNase hypersensitivity in the case of Melanoma or Medulloblastoma. A blank cell indicates unavailability of the epigenomic profile corresponding to the particular cell type

Chapter 4

Mutational driver detection framework

Although the mutational landscape of tumors is diverse and varies from one sample to another, not all mutations contribute to tumorigenesis [43, 44]. Comprehensive genomic and clinical characterization of different tumor types have motivated the development of computational methods to identify molecular alterations that are advantageous to the fitness of the cell. These variants or abnormalities are known as cancer drivers. Bioinformatics tools have been developed to evaluate genes or genomic regions that have driver potential by harboring a higher number of somatic mutations than expected by chance. This is equivalent to a higher mutation density than the background mutation density. Accounting for the mutational background is important since it is likely that mutations occurring in regions of high density are simply a consequence of genomic instability and do not possess functional implications. However, mutations observed in a section of the genome with low overall mutational density are likely to have been positively selected and are hence functionally relevant.

4.1 Background to driver detection methods

Numerous methods have been developed to detect driver genes or mutations and distinguish them from non-functional counterparts [16, 45]. However, limited genomic alterations have been confidently established to confer selective advantage. In the absence of a comprehensive gold standard, comparison of existing predictive methods and subsequent selection becomes a difficult process. A few of the most commonly employed driver detection frameworks are detailed below.

The MuSiC pipeline aims to incorporate standardized sequence-based inputs along with multiple clinical characteristics to infer correlations between sites of mutations and associated genes and pathways [46]. ActiveDriver is gene-centric framework built on a generalized linear regression model that incorporates information about a given gene and its phosphosite region [47]. TUSON is a computational pipeline that characterizes the patterns of mutational signatures across tumor types to determine whether a particular gene is likely to be an oncogene or a tumor suppressor [48]. MutSigCV uses genome- or exome-wide variant calls from multiple samples, with coverage and covariate information to identify genes that are harbor more mutational burden than expected by chance [22].

Multiple driver detection tools have been developed to determine the functional impact of missense mutations. These methods typically incorporate information about the site of the mutation by evaluating existing knowledge from evolutionary conservation, sequence context, or physiochemical properties of the resultant proteins. For example, SIFT and MutationAssessor utilize features derived from evolutionary stability of different segments of the genome [49, 50]. Similarly, CHASM is a random forest classifier that includes evolutionary aspects and protein domain information, as well as tissue-type specific factors, to prioritize missense mutations whose resultant functional alterations are linked to tumor progression [51]. Other methods include SSA-ME and ParsSNP. SSA-ME is a network-based method to identify potential cancer drivers through scoring small subnetworks for mutual exclusivity using a reinforced learning approach [52]. ParsSNP, on the other hand, is an unsupervised predictor of functional impact that employs an expectation-maximization framework to find mutational drivers [53].

Amongst the described techniques for mutational driver detection, the MutsigCV method is most closely linked to the notion of identifying drivers through recurrence and the necessary adjustment of the background mutation density. MutsigCV has shown that identification of significantly mutated genes in cancer is improved by accounting for patient-specific attributes of mutation frequency and signatures, and gene-specific background mutation densities incorporating genomic covariates such as expression level and replication time. This is a consequence of the fact that while it is possible to accurately compute the background mutation density for larger regions and genes, it is relatively difficult to do so for shorter regions or for tumor types with low mutational burden. MutsigCV incorporates a technique based on binning where genes are grouped based on expression level and the average mutation rate is calculated for each bin individually.

4.2 Method description

In this study, we develop a novel driver detection framework by factoring in genome-wide variation in tissue-specific epigenomes (Fig. 4-1). In particular, the availability of cell type-specific reference profiles of diverse epigenomic marks from the Roadmap Epigenomics Consortium allows us to uniquely model the mutational landscape and distinguish driver alterations from non-functional passengers for each tumor type. We leverage the various histone profiles as well as DNase hypersensitivity tracks as covariates in correcting the mutational background in a robust manner across tumor types with varying mutational burden. We incorporate the genome-wide Repli-Seq profile for GM12801 as an additional covariate. Altogether, this allows us to effectively re-estimate the local background mutation density uniformly across the genome for coding as well as non-coding regions.

In this framework, we first establish the set of unique mutations for the 17 tumor types considered. A 1kb genomic window is constructed around each mutation, for which we compute mean signal enrichment for every corresponding histone mark, DNase hypersensitivity track, and replication timing profile. This allows us to project each mutation as a point in an N-dimensional epigenomic covariate space (where N is defined by the number of available features for the cell type in question as indicated in Fig. 3-3). This representation of the mutational landscape enables the identification of mutations that share epigenomic context unique to each tissue. We re-estimate the mutational density in the 1kb window around a particular mutation of interest by averaging over the background mutation rates of proximal mutations. This method inherently corrects for epigenomic covariates in calculating the mutational background. The local neighborhood is defined as the 50 nearest neighbors of a mutation of interest, where distance between two mutations is computed by Euclidean distance in the N-dimensional space. Our motivation for considering only mutated segments of the genome and not the unaltered regions is two-fold. First, our method should be uniformly applicable across tumor types. The mutational burden and the number of regions lacking mutation vary across cancers from different tissue types. Secondly, our representation of the mutational landscape is based on the assumption that the samples in the cohort capture the entire space of possible mutations for each tumor type. Therefore, accounting for non-mutated regions would only have the effect of decreasing the mutation density.

4.3 Mutations identified based on site-specific recurrence

The calculated background mutation density specified at the location of a given mutation defines the position-specific probability of mutation for each tumor type. Given this probability p and the total number of samples n, the binomial distribution B(n,p) represents the number of samples in which the given site is mutated. Since we know the actual number of samples harboring that mutation (observed), we can define a p-value to capture the statistical significance of mutational occurrence at the concerned site. We then perform false discovery rate (FDR) correction to obtain a ranked list of recurrent mutations for each tumor type.

Significance analysis of recurrent mutations reveals unique patterns of mutational incidence in a tissue-specific context. For instance, the most significant recurrent mutations in breast cancer across both TCGA and Alexandrov cohorts are found to localize in the quiescent chromatin state. This was indicated earlier through the strong explanatory potential of H3K9me3 in the both breast cancer datasets (Fig. 3-3), thereby suggesting de-repression may promote oncogenesis. In addition to the quiescent state, other highly recurrent mutations localize to the weak transcription (5_TxWk) chromatin state and upon closer inspection are found to be located in intronic regions. In contrast to breast cancer, the significant mutations in liver cancer are present in diverse chromatin states. In particular, liver-specific enhancers are found to harbor three of the top twenty identified recurrent mutations. Chr5:1295228 lies within a cancer susceptibility locus (5p15.33) and is immediately downstream of the TERT gene. Additionally, the enhancer is immediately upstream of the CRR9p gene, whose over-expression is known to confer resistance to apoptosis caused by genotoxic agents. A recurrent promoter mutation is found in B-cell lymphoma upstream of the IGLL5 gene, an immune-related gene known to be involved in the B-cell receptor signalling pathway. In addition to identifying potential regulatory drivers, our model also reveals recurrent exonic mutations. For instance, B-cell lymphomas harbors a mutation in the RHOA gene, member of the Ras superfamily, consisting of proteins involved in cell cycle regulation. Additionally, a protein-coding mutation in the IDH1 gene is the top ranked hit in the lower grade glioma (LGG) cohort from TCGA. Disruption of IDH1 function is known to cause widespread changes in histone and methylation landscapes that potentially promote oncogenesis.

4.4 Recovery of potential drivers and identification of false positives

Previous studies on non-coding mutations in cancer have reported that limited sample size presents a challenge to characterizing functional drivers [31]. In particular, site-specific mutational recurrence in the regulatory genome, as illustrated in the previous section, may correspond to only a small number of downstream target genes, especially when the overall cohort size is limited [17]. This is primarily because non-coding mutations occurring at the same site are typically rare. Our analysis pipeline is capable of recovering mutations normally classified as non-significant when considering the local mutational background, but with potential functional implications. This is enabled by comparing the statistical significance of mutational recurrence computed using the local mutational background with the recurrence using the epigenomically-adjusted mutational background. Mutations that are identified as significant in the latter case but not the former can be considered as 'recovered' mutations. Such instances typically arise when the adjusted mutation density is lower than the original, and when the actual number of occurrences of the mutation is less. While it is possible that such discoveries do not carry true functional value, it is likely that these mutations will be identified as recurrent upon augmenting the cohort with additional samples [17]. Alternatively, a mutation identified as recurrent using the local mutational background could potentially be a false positive when validated with the model using the adjusted background mutation density.

Recurrence-based driver detection methods that return a p-value or similar measure of statistical significance usually involve defining a significance threshold to separate functional drivers (associated with low p-values) from non-functional passenger mutations (with high p-values). However, a common challenge faced when segregating groups based on statistical significance is defining an appropriate cutoff. In particular, diverse mutational landscapes across tumor types make it difficult to choose a uniform threshold without mis-classified mutations. To characterize the effect of different significance thresholds, we plot the variation in the number of mutations with p-value lower than the threshold as a function of the cutoff value for both the local mutation density framework and the adjusted mutation density framework for two tumor types as an example (Lower-grade glioma and Renal clear cell carcinoma, Fig 4-2). We observe that across all tumor types, the number of mutations classified as potential drivers was lower when using the epigenomically-adjusted mutational background than with the local mutational background. This reduction is more prominent in cancers with overall low mutational burden. This indicates our method for re-estimating the mutation density has the effect of restricting the space of potentially functional mutations. As verification, we explore the distribution of chromatin states harboring mutations crossing the significance threshold. With more stringent thresholds, we notice a reduction in unique chromatin states across multiple tumor types, suggesting preferential state-specific incidence of recurrent mutations.

4.5 Active epigenomic regions

Variants in the regulatory genome functionally impact downstream gene expression to promote oncogenesis through a wide range of mechanisms such as disrupting TF binding sites or non-coding RNA function. In particular, mutations that localize to promoter or enhancer elements can create altered motifs and loss-of-function events that affect the binding affinity of TFs [3, 54]. For all mutations occurring in a particular tumor type, we can use the position-specific significance of mutational recurrence to derive a test statistic that yields information regarding the driver potential of the region of interest. For this study, we focus on enhancer and promoter elements defined in a tissue-specific manner by the Roadmap project.

Convolving the effective recurrence potential of individual variants across genomic segments reveals multiple novel active non-coding regions with driver potential that are linked in their regulatory context with genes impacting tumor progression. For instance, a 2kb enhancer region in Medulloblastoma was the most significant in terms of mutational burden and recurrence and governs the expression of the EIFA4A2 gene. EIF4A2 has been studied in the context of breast cancer, non-small cell lung cancer, and melanoma and is linked to translational repression and miRNA-mediated gene regulation [55]. Among the breast cancer samples from the Alexandrov cohort, the top-ranking enhancer region was in the 17p13.1 locus. This enhancer has been linked to a number of genes including SOX15, POLR2A, PLSCR2. SOX15 has been well characterized as a tumor suppressor, PLSCR2 has been linked to apoptosis, and POLR2A is known to experience co-deletion with TP53, a well-known tumor suppressor [56, 57].

Although this framework allows us to perform preliminary prioritization of regulatory regions that harbor recurrent mutations, characterizing the functional impact of each variant will require computational or experimental follow-up and is necessary for a complete understanding of the dysregulated pre-transcriptional machinery in cancer. An example of a computational technique developed with this goal is the Intragenomic Replicates (IGR) method [58]. IGR describes the binding affinity of a given TF for both reference and variant alleles of the mutation of interest while accounting for the local nucleotide context around the variant.

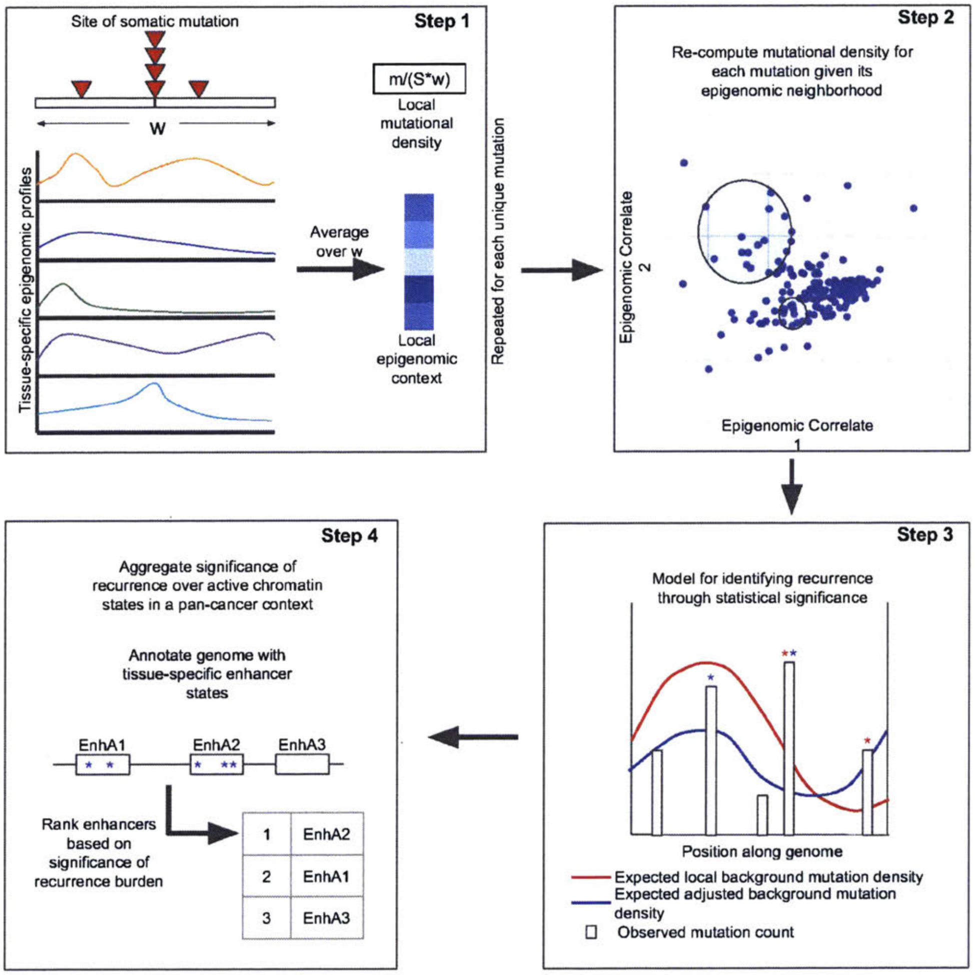

Site of somatic mutation
Step 1
m/(S*w)
Local mutational density
W
Tissue-specific epigenomic profiles
Average over w
Local epigenomic context
Repeated for each unique mutation
Step 2
Re-compute mutational density for each mutation given its epigenomic neighborhood
Epigenomic Correlate 2
Epigenomic Correlate 1
Step 3
Model for identifying recurrence through statistical significance
Position along genome
Expected local background mutation density
Expected adjusted background mutation density
□ Observed mutation count
Step 4
Aggregate significance of recurrence over active chromatin states in a pan-cancer context
Annotate genome with tissue-specific enhancer states
EnhA1
EnhA2
EnhA3
Rank enhancers based on significance of recurrence burden
1 EnhA2
2 EnhA1
3 EnhA3

Figure 4-1: Computational framework for identifying recurrent non-coding mutations. In Step 1, we compute enrichment of different histone marks, DNase hypersensitivity tracks, and replication timing (N unique features) in a 1kb window around a somatic mutation to derive a vector of averaged epigenomic signal for each mutation. In Step 2, mutations are localized in the N-dimensional epigenomic space (N=2 in figure), the neighborhood of each mutation is computed, and the background mutation density is updated. In Step 3, we identify statistically significant recurrent mutations (asterisks) using both the original and the adjusted mutational background. In Step 4, we aggregate significance measures for mutations present in the same active chromatin state to derive the 'recurrence potential' of a regulatory locus.

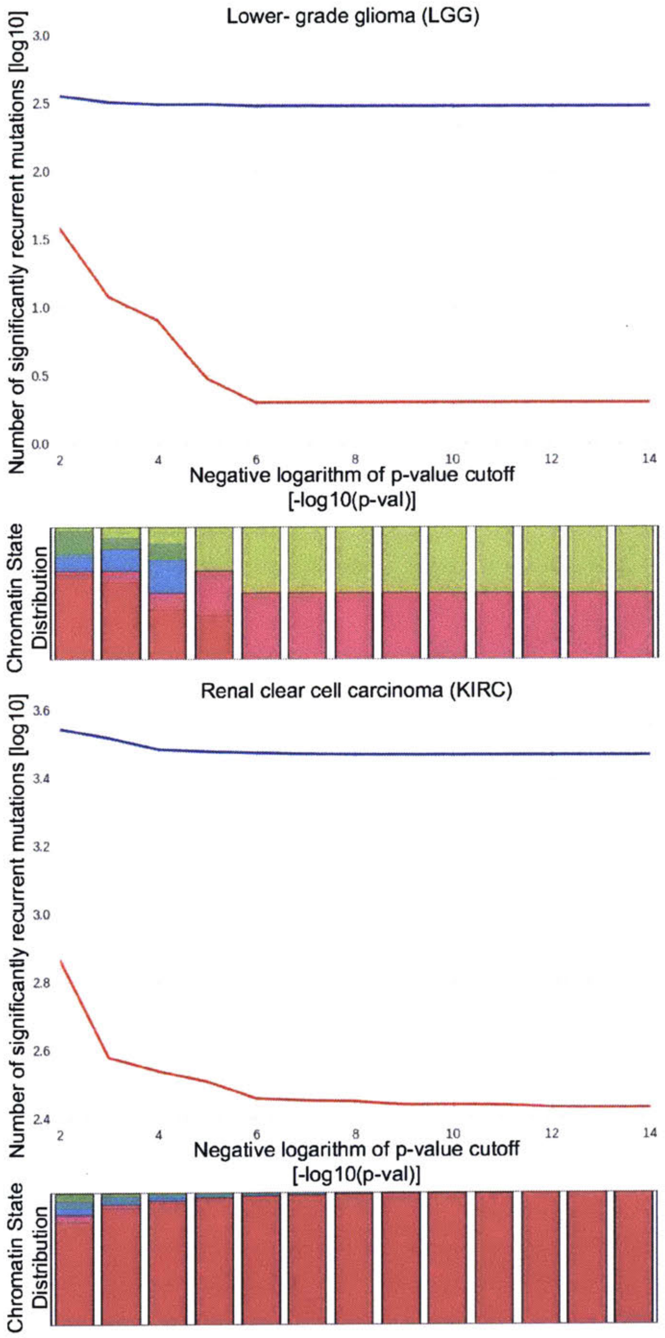


Figure 4-2: Analysis of recurrent mutations with varying significance threshold: Top plots indicate the number of recurrent mutations that meet the significance threshold (Y axis) for decreasing p-value cutoff (X axis) for the two models for background mutation density - the local mutational background (blue) and the epigenomically adjusted mutational background (red). The stacked bars at the bottom represent the variation in chromatin state distribution of recurrent mutations for different p-value cutoffs - weak repressed Polycomb state 14_ReprPCWk, lime green), transcribed state (4_Tx, violet), quiescent state15_Quies, dark pink).

Chapter 5

Regulatory convergence of genomic regions carrying mutational burden onto common target genes

Multiple distinct regulatory regions often impact the same downstream target gene and can exhibit diverse patterns of mutational burden in a sample-specific manner (Fig. 5-1). Additionally, the regulatory loci linked with a particular gene vary from one tissue to another and this linking exhibits added complexity in cancer by virtue of genomic instability and copy number variation. Modeling the regulatory neighborhood of each gene and identifying recurrent non-coding mutations will enhance our understanding of regulatory drivers. In this chapter, we describe a statistical model of regulatory convergence that characterizes the cumulative effect of genomic alterations aggregated across regulatory regions linked to a common target gene. Epigenomic interaction data, in tandem with carefully curated databases on enhancer-gene links, allow for comprehensive identification of distal enhancer regions and their targets. This in turn enables integration of individual low-frequency alterations into high-frequency recurrent events across different tumor types.

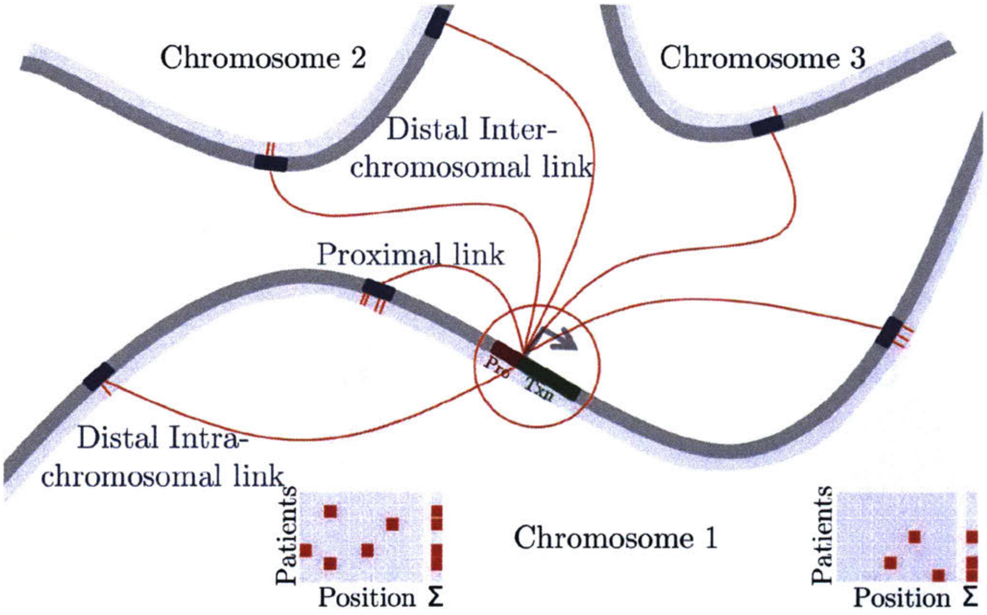


Figure 5-1: Framework of convergence analysis: The regulatory neighborhood of a gene of interest can contain proximal and distal regulatory elements corresponding to different chromatin states. Mutational incidence across these regulatory loci is heterogeneous and disparate. The idea of convergence implies that individual low-frequency non-coding mutations can result in the same functional consequence when aggregated across the regulatory neighborhood of a target gene.

5.1 Defining a tissue-specific regulatory neighborhood for each gene

Regulatory loci often lie far from their downstream target along the one dimensional genome. However, three dimensional chromatin conformation results in active regulatory sites and target genes in physical proximity [59]. Therefore, accurate and robust methods integrating epigenomic context and chromatin organization, among other features, are necessary to describe the pre-transcriptional regulatory interactome.

For this study, we leverage an existing database of enhancer-gene links defined in a tissue-specific manner to construct tissue-specific reference interactomes for each gene [18]. The framework for mapping enhancers to genes is based on correlations in activity profiles defined for each enhancer, where distance-based information and association between gene expression and genome-wide histone signal intensity tracks were combined. The active regulatory landscape was complemented with gene expression maps of potential downstream targets, sequence motif enrichment, and expression of transcription factors binding each motif. This enabled probabilistic interpretation of the pre-transcriptional regulatory linking surrounding each enhancer. A classifier based on logistic regression was implemented to distinguish real enhancer gene links from control groups where expression values were randomly permuted across the gene set. This framework, developed originally for ENCODE cell lines, was then extended to encompass all the 127 cell types from the Roadmap project.

Enhancer-gene databases generated based on epigenomic correlations tend to result in a smaller number of regulatory links than HiC or other chromatin interaction-based approaches that consider extended segments of the genome to be in physical proximity. However, there are some challenges with using HiC databases to inform the regulatory neighborhood in the context of this study. First, HiC profiling has not yet been performed uniformly across multiple tissue types. Second, HiC databases are available in different contact resolutions and the choice of resolution defines the domain of regulatory linking. To address the challenge of under-representation of the regulatory neighborhood from enhancer-gene databases, we present a novel method toward augmenting the set of loci associated to a target gene in the following section.

5.2 Expanding regulatory neighborhood by factoring epigenomic plasticity

Previous studies have selected reference interactomes based on the closest matched epigenomes for each tumor type. In this setting, mutations that disrupt chromatin interactions are inherently accounted for. However, genomic alterations that create de novo tumor-specific regulatory links will not be captured as relevant. Here, we describe a novel framework to identify mutations with regulatory potential that occur in inactive regions of the genome in the matched reference cell type (from the Roadmap Epigenomics Project).

We leverage chromatin state annotations from the Roadmap project across 127 cell types. The genome is binned into segments of 200 bp and each segment can belong to one of 15 distinct chromatin states, defined using the ChromHMM method. Information concerning the genomic location of various histone modifications is used to estimate the distribution of the 15 chromatin states across 127 distinct cell types. For each 200 bp segment we compute the Shannon entropy of the state distribution, which enables us to quantify the uncertainty of the distribution and acts as a proxy for chromatin state plasticity at that region. We also define a *Repressive Index* for each 200 bp segment, which is the probability of occurrence of a repressive state at that location considering independence in the occurrence of different chromatin states across cell types. Therefore, this functions as an empirical estimation of the chromatin state distribution.

To capture out-of-context de-repression of genomic segments as a result of somatic mutations, this analysis focusses on the repressed or quiescent states across the reference epigenomes for each tumor type (Fig. 5-2a). Genomic segments of primary interest are those that are repressed in the reference cell type but active to some degree in other cell types. In other words, the manifestation of the phenotypic impact of a variant can be masked by repressed state in original normal cell while that repressive state becomes active in cancer. We define the search space as non-overlapping 200 bp segments repressed in the cell type of interest, but not repressed across all cell types. To characterize the effect of repressive index in a segment of the genome on the local mutational density we plot the corresponding mutational enrichment for each tumor type (Fig. 5-2b). To compute mutational enrichment, we determine the number of mutations across all samples of a tumor type, in 200 bp windows with a given value of repressive index. The mutation count is normalized first by the total number of genomic segments with that repressive index, then by the overall average mutational burden for that tumor type. We observe an increase in mutational enrichment with repressive index, along with a decrease in variance. We hypothesize that this pattern is due to the effect of DNA repair machinery which is likely to be more active at regions with low repressive index thereby resulting in lesser mutational incidence.

5.3 Identifying genes with regulatory burden for each cancer type as well as genes shared across cancer types

We discover genes harboring non-coding mutations in their regulatory neighborhood by exploring their linked regulatory regions and aggregating over statistically significant recurrent mutations. The enhancer-gene interaction set is augmented by de-repressed regions characterized by low repressive index in a tissue-specific manner as described in the previous section. The includes de-repressed segments that do not contain linking information in the tissue of interest. However, we can assign a target gene for these regions using our enhancer-gene database in the event that these regions function as active enhancers in other tissues. Upon ranking genes in decreasing order of the significance of recurrent mutations in their regulatory neighborhood, we find that the top genes represent convergence events in lung adenocarcinoma, melanoma and colorectal carcinoma. Many of these genes are linked with oncogenic or immunomodulatory biological processes. For instance, the genes DUSP22 and IRF4 identified in lung adenocarcinoma are both known to impact immune pathways. DUSP22 is a JNK pathway-associated phosphatase known to inhibit T-cell receptor signalling and autoimmunity by inactivating Lck [60]. IRF4 has been linked to cellular signalling leading to differentiation of dendritic cells in lung tissue [61]. In melanoma, our framework for regulatory convergence reveals a number of significant genes linked with immune signalling. SPSB1 was the top hit in this case and is known to down-regulate the TGF-beta signaling pathway targeting the type II receptor [62]. The occulocutaneous albinism 2 (OCA2) gene is involved in small molecule transport of tyrosine and is known to influence melanoma susceptibility [63]. The most significant convergence events in colorectal carcinoma are genes also related with immune function. SMAD3 plays a crucial role in TGF-beta signalling and is important for cell fate and differentiation [64]. NFATC2 is a member of the nuclear factors of the activated T cells transcription complex and is known to enhance cell motility for metastasis in breast and colon cancers [65].

Tumor types with intermediate and low mutational load are also found to possess genes experiencing significant regulatory burden. The EHF gene, discovered in pancreatic cancer, functions as a tumor-suppressor and plays an important role in carcinogenesis through changes in subcellular localization [66]. Also significant for regulatory convergence in pancreatic cancer is ZBTB16, known to influence the unique characteristics of NKT cells such as antigen response [67]. In the breast cancer cohort from the TCGA study, we identify the FGFBP1 gene, which is known to contribute to miRNA induced angiogenesis, and the TRIM33 gene, which acts as a tumor suppressor and regulates the TGF-beta receptor [68, 69]. In chronic lymphocytic leukemia (CLL), IKZF3, which is a hematopoietic-specific transcription factor, is found to harbor regulatory mutations and is known to be over-expressed in CLL [70].

In addition to evaluating downstream gene targets of regulatory loci for each tumor type, our model allows us to examine tissue-specific uniqueness of non-coding mutational burden as well as shared features of dysregulation across tumor types.

For instance, the IGLL5 gene was found to harbor regulatory mutations uniquely in B-cell lymphoma. IGLL5 encodes an immunoglobulin lambda-like polypeptide and is homologous to IGLL1, which is critical for B-cell development [71]. Similarly, we capture potential dysregulation of ARF6 in breast cancer, whose overexpression is linked to invasion and metastasis [72]. Our framework also uncovers genes exhibiting regulatory burden across multiple tumor types. In this case, the most significant hit is the MYH9 gene, which is an established tumor suppressor and is known to regulate TP53, and exhibits a recurrently mutated regulatory neighborhood across 11 distinct tumor types [73]. Additionally, dysregulation in EGFR and AMOTL2 is revealed across 10 tumor types. While EGFR is a well-characterized oncogene, AMOTL2 promotes angiogenesis and is linked to invasiveness of tumors [74].

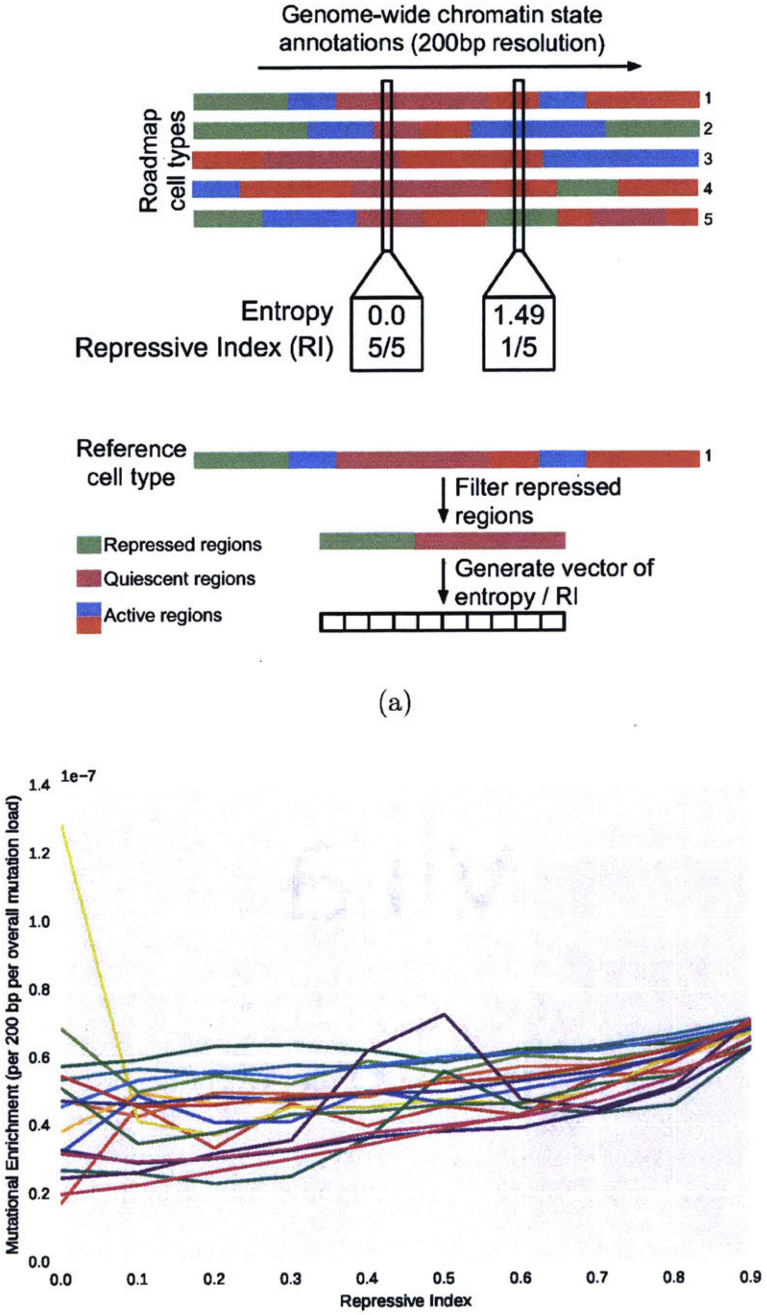
5.4 Correlation with gene expression for TCGA tissues

For genes with significant regulatory burden, we partition the set of samples from each tissue type into two categories: those possessing at least one mutation in the regulatory neighborhood of the gene of interest, and those with zero mutations. We find that aggregating somatic variants across regulatory loci accounts for dysregulation of the downstream target gene in a larger fraction of samples than that observed by considering regulatory loci independently. Additionally, we validate pre-transcriptional dysregulation on a gene-by-gene basis by comparing gene expression profiles across the two partitions of samples.

We identify tissue-specific relationships between the presence of a regulatory mutation and subsequent gene expression patterns for the same gene across tissue types. For instance, although AMOTL2 harbors mutations in its regulatory neighborhood across 10 tumor types, a significant correlation between mutational incidence and expression profiles is seen only in lung squamous cell carcinoma (LUSC). Similarly, regulatory mutations in SPSB1 cause differential expression in lung adenocarcinoma

(Fig. 5-3).

Although a number of genes relevant to cancer growth and progression are discovered to have statistically significant regulatory burden, we observe that only a fraction of those non-coding mutations aggregated over regulatory loci result in distinct differential gene expression patterns. This can be attributed to low overall expression (in both mutated and unmutated samples) resulting in weaker comparisons. Tissue-specific expression and activity of transcription factors can also explain the varying impact regulatory burden has on gene expression across different tissue types.



(a)

(b)

Figure 5-2: Analysis of chromatin state plasticity and mutation burden: (a) Computational framework for measuring chromatin state plasticity across the genome. The entropy and repressive index for the distribution over 15 chromatin states and 127 cell types are calculated. Genome-wide entropy and repressive index (RI) vectors are derived for cell type-specific repressed regions (b) Mutational enrichment variation for 200 bp repressed segments of the genome for each tumor type.

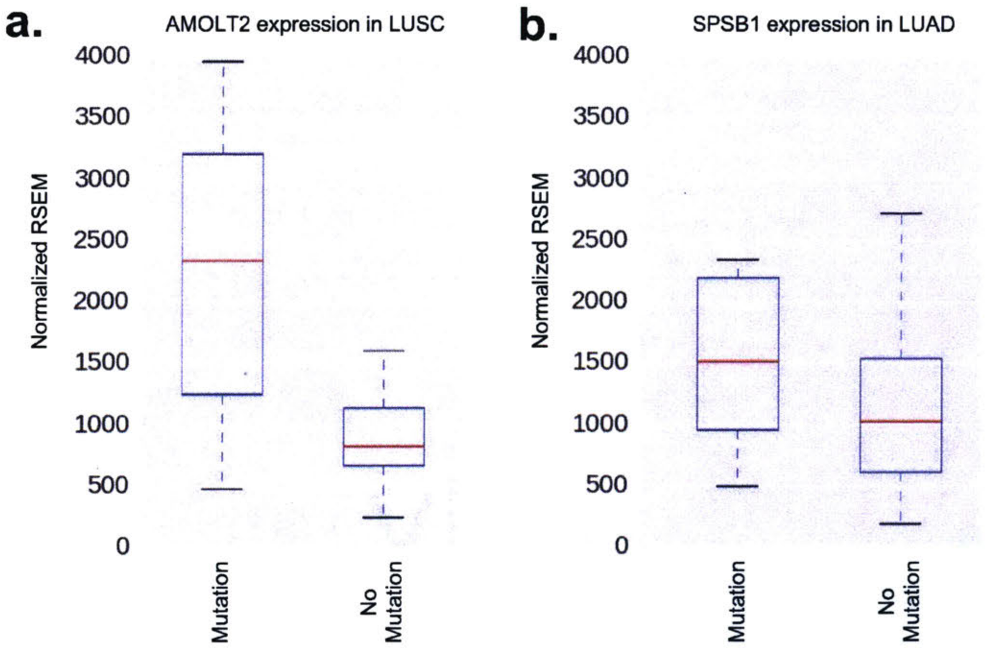


Figure 5-3: Differential gene expression between sets of tumor samples with and without at least one regulatory mutation for the gene of interest: Differential gene expression is quantified through the normalized RSEM (RNA-Seq by Expectation Maximization) measure. (a) AMOLT2 expression in lung squamous carcinoma (T-test p-value: 8.39e-5). (b) SPSB1 expression in lung adenocarcinoma (T-test p-value: 0.02)

Chapter 6

Conclusions and future work

The diverse mutational landscape across the cancer genome makes it challenging to characterize the functional potential of each variant. Here, we present a framework to establish recurrence of non-coding mutations augmented by tissue-specific epigenomic context. We leverage existing databases on tissue-specific enhancer-gene interactions to aggregate variants that are individually rare across cancer cohorts into high-order statistically significant recurrent events. Additionally, the proposed methodology for non-coding mutational driver discovery was found to be applicable across tumor types of diverse mutational burden. Novel convergence events shed light on tumor type-specific cancer driver genes that were previously uncharacterized in the tissue(s) of interest. We also discover genes with patterns of regulatory burden shared across multiple tumor types. The identified genes take part in tumor suppression, angiogenesis and immune evasion. While recurrent protein-coding mutations are known to activate oncogenes or silence tumor suppressors, the functional convergence of non-coding mutations towards multiple hallmarks of cancer highlights their relevance to cancer growth and progression.

We recognize the limitations of our model. The assumption of the cell-type-of-origin epigenome as the closest reference epigenome may restrict our ability to capture histone modification or methylation changes brought about due to genomic instability of the tumor sample. Similarly, we approximate the regulatory neighborhood of each gene in a tumor cell to be the same as that in a normal cell. Therefore, tumor-induced formation and silencing of regulatory interactions have not been accounted for. We also considered a cell type-agnostic approach toward characterizing replication timing primarily due to a limitation in the number of Repli-Seq profiles available.

We envision a number of potential avenues of expansion for this work. In addition to more rigorously incorporating reference epigenomes and interactomes that are representative of tumor-specific chromatin organization, overlaying three dimensional interaction data from HiC or ChiA-PET experiments could be insightful. Additionally, the functional impact of individual non-coding variants and TF-binding affinity can be rigorously characterized through sequence motif-based methods such as IGR. We believe that the current framework can be strengthened significantly with the availability of larger tissue cohorts allowing better capture of regional variation in the mutational background. Accounting for the inherent instability of the cancer genome from copy number variations, gene fusions, or chromosomal translocations and integrating these distinct modalities would enable more accurate modelling of the regulatory interactome. Additionally, as non-coding variants are not immunogenic, their accumulation in the tumor cell would be masked from immune surveillance. Our model can be expanded to characterize this period of divergent evolution thereby providing functional insight on the distinct clonal populations of tumor cells that emerge. Finally, empowering our framework to incorporate clinical characteristics for each tumor sample coupled with longitudinal profiling over the course of treatment can potentially bring about translational discoveries. Ultimately, this could drive progress towards improved patient stratification and accelerate the development of personalized therapeutics.

Appendix A

Data download and analysis

Whole genome variant calls for the TCGA cohort for 9 tumor types based on the study by Fredriksson et al. are downloaded from Synapse (SYN2882201). Colon and rectal carcinomas were considered together as one tumor type. SAMtools and VarScan were used to call somatic mutations. A minimum variant frequency of 0.2 is imposed. The whole genome variant calls for the Alexandrov cohort are obtained from `ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl`. We download the raw mutations text file for each of 8 tumor types. Somatic mutational prevalence across the genome are calculated based on all identified mutations with the assumption that the average whole genome has sufficient coverage over 2.8 gigabases. Both the TCGA and the Alexandrov cohort contain samples of breast cancer and lung adenocarcinoma.

Reference epigenomes for each of the tumor types are obtained from the Roadmap Epigenomics dataset. We download the 11 reference epigenomes (E031, E033, E059, E066, E081, E086, E096, E098, E101, E119, E125) from the Washington University web portal - `http://egg2.wustl.edu/roadmap/web_portal/`. For each cell or tissue type, we obtain the BigWig signal (p-value) files as well as the peak calling results as NarrowPeak BED files for all histone modifications and DNase hypersensitivity tracks available. Additionally, we also download the 15-state chromatin state annotation for each of the cell and tissue types.

BigWig enrichment files based on the Repli-Seq assay are downloaded for four cell-cycle fractions (G1B, S1, S4, G2) from ENCODE (`https://www.encodeproject.`

org/). We utilize the bigWigToBedGraph tools from the UCSC binary utilities directory (http://hgdownload.soe.ucsc.edu/admin/exe/) to convert the BigWig signal files to the BedGraph format. This is followed by peak calling using the Model-based analysis of ChiP-seq (MACS) algorithm (https://github.com/taoliu/MACS) to identify NarrowPeaks. We then aggregate the number of NarrowPeaks over the genomic segment of interest for each cell cycle fraction. Replication timing is then represented by an early-to-late ratio given by (G1B+S1)/(S4+G2).

For the Random Forest regression framework, the predictor variables are the NarrowPeak counts over 1Mb bins across the entire genome (giving rise to 3,113 bins). The target variable is the 1Mb binned count of total mutations over all samples of each tumor type. To compute the epigenomically-adjusted mutation density, each unique mutation for a tumor type is annotated with its local epigenomic context. For this analysis, the average of the BigWig waveform for every available histone mark, DNase hypersensitivity track and Repli-Seq cell cycle profile is considered around a 1kb window of each mutation. Replication timing in the local context of the mutation is calculated as described above.

Gene expression profiles for 20,530 genes across the 9 tumor types from TCGA are obtained from the Broad Institute Firehose project (http://gdac.broadinstitute.org/). Samples from colon adenocarcinoma and rectum adenocarcinoma are collapsed into the CRC cohort. All of the analyses described in this study are performed in the cluster computing environment at the Broad Institute and in local Linux workstations (Intel i7-6900K @ 3.2GHz) of the Computational Biology group at MIT.

Appendix B

Supplementary Tables and Figures

Table B.1: Description of tumor cohorts, individual tumor types, corresponding matched reference epigenomes, and TCGA tissue ID.

Tumor Abbr.	Dataset	Cancer Type	Matched Cell-of-origin	TCGA
Breast	Alexandrov	Breast cancer	E119 – HMEC	BRCA
Pilocytic_Astrocytoma	Alexandrov	Pilocytic astrocytoma	E125 – NH-A Astrocyte	N/A
Medulloblastoma	Alexandrov	Medulloblastoma	E081 – Fetal Brain	N/A
BRCA	Fredriksson	Breast cancer	E119 – HMEC	BRCA
Liver	Alexandrov	Hepatocellular carcinoma	E066 – Liver	LIHC
LUAD	Fredriksson	Lung adenocarcinoma	E096 – Lung	LUAD
LUSC	Fredriksson	Lung squamous cell carcinoma	E096 – Lung	LUSC
CRC	Fredriksson	Colorectal carcinoma	E101 – Rectal Mucosa	COAD+READ
SKCM	Fredriksson	Melanoma	E059 – Foreskin Melanocyte	SKCM
KIRC	Fredriksson	Kidney clear cell renal cell carcinoma	E086 – Fetal Kidney	KIRC
CLL	Alexandrov	Chronic lymphocytic leukemia	E033 – T cells cord blood	N/A
GBM	Fredriksson	Glioblastoma	E125 – NH-A Astrocyte	GBM
Lung_Adeno	Alexandrov	Lung adenocarcinoma	E096 – Lung	LUAD
Lymphoma _B-cell	Alexandrov	B-cell Lymphoma	E031 - Primary B-cells	DBLC
LGG	Fredriksson	Low-grade glioma	E125 – NH-A Astrocyte	LGG
KICH	Fredriksson	Kidney chromophobe	E086 – Fetal Kidney	KICH
Pancreas	Alexandrov	Pancreatic cancer	E098 – Pancreas	PAAD

Table B.2: Number of samples with whole-genome mutation calls, mean and standard deviation of overall mutational count per sample, and number of unique variants for each tumor type.

Tumor Abbr.	# WGS	Mean Mut. Count	Standard Deviation Mut. Count	Unique SNV Count
Breast	119	5679.81	7269.79	659093
Pilocytic_Astrocytoma	101	105.21	113.85	10556
Medulloblastoma	100	1253.33	1190.04	124865
BRCA	96	6651.33	5337.09	600155
Liver	88	10012.97	5491.45	870110
LUAD	46	28114.17	31570.63	1202789
LUSC	45	36402.53	15941.42	1578870
CRC	42	117896.33	219933.91	4152409
SKCM	38	94655.24	117199.79	3529151
KIRC	29	4848.41	1612.62	116240
CLL	28	2080.57	874.87	51749
GBM	27	7056.26	1847.3	174164
Lung_Adeno	24	66494.83	61637.33	1531734
Lymphoma _B-cell	24	5329	7254.75	127407
LGG	18	2090.11	1242.65	33315
KICH	15	1919.53	557.68	21954
Pancreas	15	7471.2	3993.39	112018

Table B.3: Chromatin state abbreviations, definition and active/inactive potential from the Roadmap Epigenomics project.

Abbreviation	Chromatin State	Active / Inactive
1_TssA	Active TSS	Active
2_TssAFlnk	Flanking active TSS	Active
3_TxFlnk	Transcr. at gene 5' and 3'	Active
4_Tx	Strong Transcription	Active
5_TxWk	Weak Transcription	Active
6_EnhG	Genic enhancers	Active
7_Enh	Enhancers	Active
8_ZNF/Rpts	ZNF genes + repeats	Active
9_Het	Heterochromatin	Inactive
10_TssBiv	Bivalent/poised TSS	Inactive
11_BivFlnk	Flanking bivalent TSS/Enh	Inactive
12_EnhBiv	Bivalent enhancer	Inactive
13_ReprPC	Repressed Polycomb	Inactive
14_ReprPCWk	Weak repressed Polycomb	Inactive
15_Quies	Quiescent	Inactive

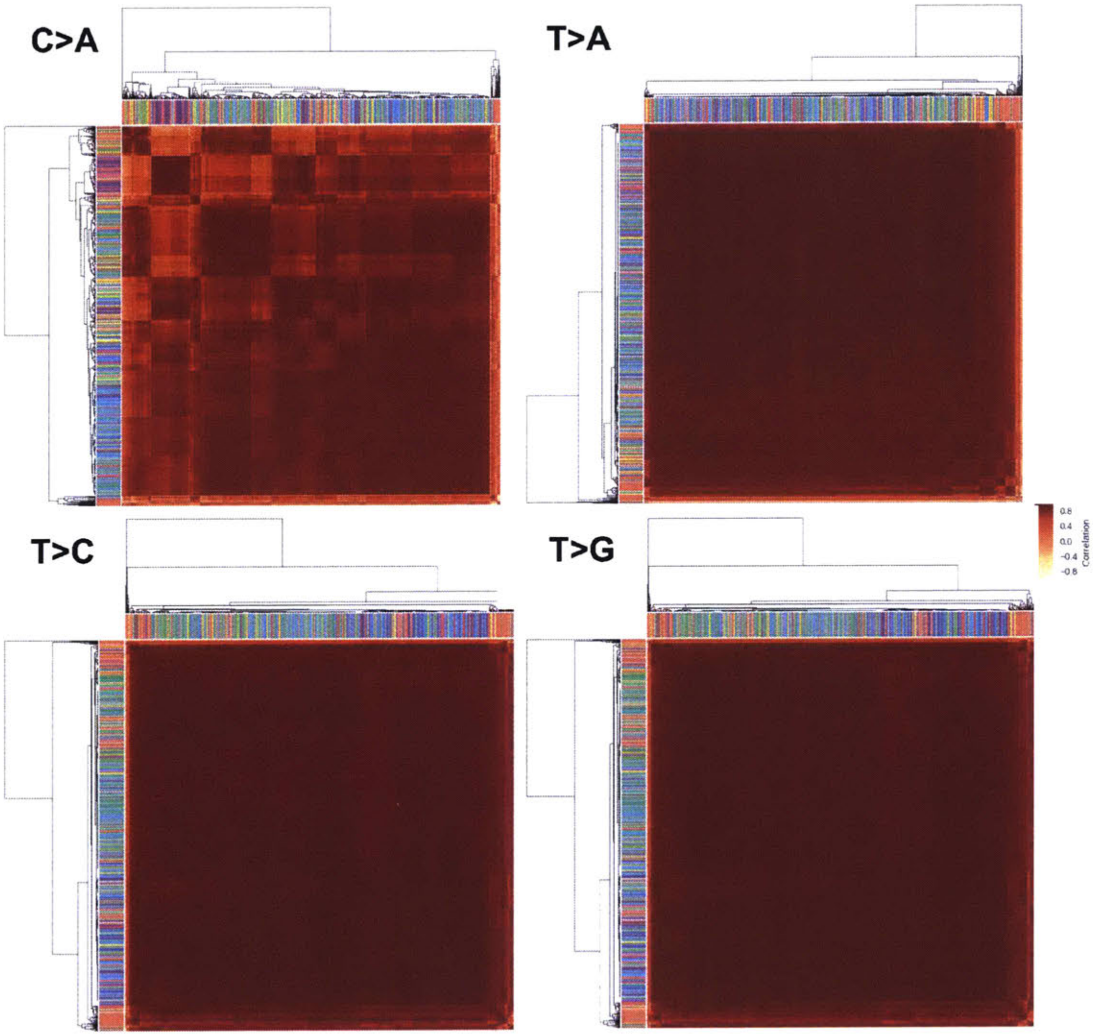


Figure B-1: Sequence context analysis for C>A, T>A, T>C, and T>G mutational categories.

Bibliography

[1] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[2] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[3] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108, 2016.

[4] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[5] Terrence S Furey. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

[6] Lingyun Song and Gregory E Crawford. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb–prot5384, 2010.

[7] Zachary D Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.

[8] Christoph Bock. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

[9] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587, 2013.

[10] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[11] Franklin W Huang, Eran Hodis, Mary Jue Xu, Gregory V Kryukov, Lynda Chin, and Levi A Garraway. Highly recurrent tert promoter mutations in human melanoma. *Science*, 339(6122):957–959, 2013.

[12] João Vinagre, Ana Almeida, Helena Pópulo, Rui Batista, Joana Lyra, Vasco Pinto, Ricardo Coelho, Ricardo Celestino, Hugo Prazeres, Luis Lima, et al. Frequency of tert promoter mutations in human cancers. *Nature communications*, 4, 2013.

[13] Victor I Sementchenko and Dennis K Watson. Ets target genes: past, present and future. *Oncogene*, 19(55):6533, 2000.

[14] Yongfa Jiao, Zhongming Jiang, Yuxia Wu, Xiaochong Chen, Xing Xiao, and Haiying Yu. A functional polymorphism (rs937283) in the mdm2 promoter region is associated with poor prognosis of retinoblastoma in chinese han population. *Scientific Reports*, 6, 2016.

[15] Jonas A Nilsson and John L Cleveland. Myc pathways provoking cell suicide and cancer. *Oncogene*, 22(56):9007–9021, 2003.

[16] Li Ding, Michael C Wendl, Joshua F McMichael, and Benjamin J Raphael. Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics*, 15(8):556–570, 2014.

[17] Kwoneel Kim, Kiwon Jang, Woojin Yang, Eun-Young Choi, Seong-Min Park, Mingyun Bae, Youn-Jae Kim, and Jung Kyoon Choi. Chromatin structure-based prediction of recurrent noncoding mutations in cancer. *Nature Genetics*, 48(11):1321–1326, 2016.

[18] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[19] Ian R Watson, Koichi Takahashi, P Andrew Futreal, and Lynda Chin. Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics*, 14(10):703–718, 2013.

[20] Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, 2015.

[21] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

[22] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[23] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg,

Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[24] Douglas E Brash. Uv signature mutations. *Photochemistry and photobiology*, 91(1):15–26, 2015.

[25] Song Ling Poon, See-Tong Pang, John R McPherson, Willie Yu, Kie Kyon Huang, Peiyong Guan, Wen-Hui Weng, Ee Yan Siew, Yujing Liu, Hong Lee Heng, et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Science translational medicine*, 5(197):197ra101–197ra101, 2013.

[26] Steven A Roberts and Dmitry A Gordenin. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer*, 14(12):786–800, 2014.

[27] Kenneth H Wolfe, Paul M Sharp, and Wen-Hsiung Li. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285, 1989.

[28] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766, 2011.

[29] Paz Polak, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S Lawrence, Alex Reynolds, Eric Rynes, Kristian Vlahoviček, John A Stamatoyannopoulos, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364, 2015.

[30] Benjamin Schuster-Böckler and Ben Lehner. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature*, 488(7412):504–507, 2012.

[31] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11):1160–1165, 2014.

[32] Johannes Schödel, Chiara Bardella, Lina K Sciesielski, Jill M Brown, Chris W Pugh, Veronica Buckle, Ian P Tomlinson, Peter J Ratcliffe, and David R Mole. Common genetic variants at the 11q13. 3 renal cancer susceptibility locus influence binding of hif to an enhancer of cyclin d1 expression. *Nature genetics*, 44(4):420–425, 2012.

[33] Collin Melton, Jason A Reuter, Damek V Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature genetics*, 47(7):710–716, 2015.

[34] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216, 2012.

[35] Andrew P Feinberg, Michael A Koldobskiy, and Anita Göndör. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nature Reviews Genetics*, 2016.

[36] Nathan Donley and Mathew J Thayer. Dna replication timing, genome stability and cancer: late and/or delayed dna replication timing is associated with increased genomic instability. In *Seminars in cancer biology*, volume 23, pages 80–89. Elsevier, 2013.

[37] Fran Supek and Ben Lehner. Differential dna mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84, 2015.

[38] John A Stamatoyannopoulos, Ivan Adzhubei, Robert E Thurman, Gregory V Kryukov, Sergei M Mirkin, and Shamil R Sunyaev. Human mutation rate associated with dna replication timing. *Nature genetics*, 41(4):393–395, 2009.

[39] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[40] Wouter G Touw, Jumamurat R Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha AFT van Hijum. Data mining in the life

sciences with random forest: a walk in the park or lost in the jungle? *Briefings in bioinformatics*, page bbs034, 2012.

[41] Bernhard Lehnertz, Yoshihide Ueda, Alwin AHA Derijck, Ulrich Braunschweig, Laura Perez-Burgos, Stefan Kubicek, Taiping Chen, En Li, Thomas Jenuwein, and Antoine HFM Peters. Suv39h-mediated histone h3 lysine 9 methylation directs dna methylation to major satellite repeats at pericentric heterochromatin. *Current Biology*, 13(14):1192–1200, 2003.

[42] Hae Min Jeong, Mi Jeong Kwon, and Young Kee Shin. Overexpression of cancer-associated genes via epigenetic derepression mechanisms in gynecologic cancer. *Frontiers in oncology*, 4:12, 2014.

[43] Jill C Mwenifumbo and Marco A Marra. Cancer genome-sequencing study design. *Nature Reviews Genetics*, 14(5):321–332, 2013.

[44] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3:2650, 2013.

[45] Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, page 201616440, 2016.

[46] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, Matthew B Callaway, David Dooling, Elaine R Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

[47] Jüri Reimand and Gary D Bader. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology*, 9(1):637, 2013.

[48] Teresa Davoli, Andrew Wei Xu, Kristen E Mengwasser, Laura M Sack, John C Yoon, Peter J Park, and Stephen J Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, 2013.

[49] Shamil Sunyaev, Vasily Ramensky, Ina Koch, Warren Lathe III, Alexey S Kondrashov, and Peer Bork. Prediction of deleterious human alleles. *Human molecular genetics*, 10(6):591–597, 2001.

[50] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, page gkr407, 2011.

[51] Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667, 2009.

[52] Sergio Pulido-Tamayo, Bram Weytjens, Dries De Maeyer, and Kathleen Marchal. Ssa-me detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. *Scientific reports*, 6, 2016.

[53] Runjun D Kumar, S Joshua Swamidass, and Ron Bose. Unsupervised detection of cancer driver mutations with parsimony-guided learning. Technical report, Nature Research, 2016.

[54] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein. Funseq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology*, 15(10):480, 2014.

[55] HA Meijer, YW Kong, WT Lu, A Wilczynska, RV Spriggs, SW Robinson, JD Godfrey, AE Willis, and M Bushell. Translational repression and eif4a2 activity are critical for microrna-mediated gene regulation. *Science*, 340(6128):82–85, 2013.

[56] Anan Yu, Christopher R McMaster, David M Byers, Neale D Ridgway, and Harold W Cook. Stimulation of phosphatidylserine biosynthesis and facilitation of uv-induced apoptosis in chinese hamster ovary cells overexpressing phospholipid scramblase 1. *Journal of Biological Chemistry*, 278(11):9706–9714, 2003.

[57] Alessia Errico. Colorectal cancer: Polr2a deletion with tp53 opens a window of opportunity for therapy. *Nature Reviews Clinical Oncology*, 12(7):374–374, 2015.

[58] Richard Cowper-Sal, Xiaoyang Zhang, Jason B Wright, Swneke D Bailey, Michael D Cole, Jerome Eeckhoute, Jason H Moore, Mathieu Lupien, et al. Breast cancer risk-associated snps modulate the affinity of chromatin for foxa1 and alter gene expression. *Nature genetics*, 44(11):1191–1198, 2012.

[59] Richard C Sallari, Nicholas A Sinnott-Armstrong, Juliet D French, Ken J Kron, Jason Ho, Jason H Moore, Vuk Stambolic, Stacey L Edwards, Mathieu Lupien, and Manolis Kellis. Convergence of dispersed regulatory mutations predicts driver genes in prostate cancer. *bioRxiv*, 2016.

[60] Ju-Pi Li, Chia-Yu Yang, Huai-Chia Chuang, Joung-Liang Lan, Der-Yuan Chen, Yi-Ming Chen, Xiaohong Wang, Alice J Chen, John W Belmont, and Tse-Hua Tan. The phosphatase jkap/dusp22 inhibits t-cell receptor signalling and autoimmunity by inactivating lck. *Nature communications*, 5, 2014.

[61] Sandra Bajaña, Sean Turner, Jinny Paul, Erola Ainsua-Enrich, and Susan Kovats. Irf4 and irf8 act in cd11c+ cells to regulate terminal differentiation of lung tissue dendritic cells. *The Journal of Immunology*, 196(4):1666–1677, 2016.

[62] Sheng Liu, Thao Nheu, Rodney Luwor, Sandra E Nicholson, and Hong-Jian Zhu. Spsb1, a novel negative regulator of the transforming growth factor-β signaling pathway targeting the type ii receptor. *Journal of Biological Chemistry*, 290(29):17894–17908, 2015.

[63] Anne-Sophie Jannot, Roubila Meziani, Guylene Bertrand, Benedicte Gérard, Vincent Descamps, Alain Archimbaud, Catherine Picard, Laurence Ollivaud,

Nicole Basset-Seguin, Delphine Kerob, et al. Allele variations in the oca2 gene (pink-eyed-dilution locus) are associated with genetic susceptibility to melanoma. *European Journal of Human Genetics*, 13(8):913–920, 2005.

[64] Aristidis Moustakas, Serhiy Souchelnytskyi, and Carl-Henrik Heldin. Smad regulation in tgf-β signal transduction. *Journal of cell science*, 114(24):4359–4369, 2001.

[65] Sebastien Jauliac, Cristina López-Rodriguez, Leslie M Shaw, Lawrence F Brown, Anjana Rao, and Alex Toker. The role of nfat transcription factors in integrin-mediated carcinoma invasion. *Nature cell biology*, 4(7):540–544, 2002.

[66] Li Wang, Jie Xing, Rui Cheng, Ying Shao, Peng Li, Shengtao Zhu, and Shutian Zhang. Abnormal localization and tumor suppressor function of epithelial tissue-specific transcription factor ese3 in esophageal squamous cell carcinoma. *PloS one*, 10(5):e0126319, 2015.

[67] Sai Zhang, Amale Laouar, Lisa K Denzin, and Derek B Sant'Angelo. Zbtb16 (plzf) is stably suppressed and not inducible in non-innate t cells via t cell receptor-mediated signaling. *Scientific reports*, 5, 2015.

[68] Hua-yu Zhu, Wen-dong Bai, Jia-qi Liu, Zhao Zheng, Hao Guan, Qin Zhou, Lin-lin Su, Song-tao Xie, Yun-chuan Wang, Jun Li, et al. Up-regulation of fgfbp1 signaling contributes to mir-146a-induced angiogenesis in human umbilical vein endothelial cells. *Scientific reports*, 6, 2016.

[69] Ronan Quéré, Laetitia Saint-Paul, Virginie Carmignac, Romain Z Martin, Marie-Lorraine Chrétien, Anne Largeot, Arlette Hammann, Jean-Paul Pais de Barros, Jean-Noël Bastie, and Laurent Delva. Tif1γ regulates the tgf-β1 receptor and promotes physiological aging of hematopoietic stem cells. *Proceedings of the National Academy of Sciences*, 111(29):10592–10597, 2014.

[70] Holger Nückel, Ulrich H Frey, Ludger Sellmann, Crista H Collins, Ulrich Dührsen, and Winfried Siffert. The ikzf3 (aiolos) transcription factor is highly upregu-

lated and inversely correlated with clinical progression in chronic lymphocytic leukaemia. *British journal of haematology*, 144(2):268–270, 2009.

[71] S Kasar, J Kim, R Improgo, G Tiao, P Polak, N Haradhvala, MS Lawrence, A Kiezun, SM Fernandes, S Bahl, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature communications*, 6, 2015.

[72] Ari Hashimoto, Tsukasa Oikawa, Shigeru Hashimoto, Hirokazu Sugino, Ayumu Yoshikawa, Yutaro Otsuka, Haruka Handa, Yasuhito Onodera, Jin-Min Nam, Chitose Oneyama, et al. P53-and mevalonate pathway–driven malignancies require arf6 for metastasis and drug resistance. *J Cell Biol*, pages jcb–201510002, 2016.

[73] Daniel Schramek, Ataman Sendoel, Jeremy P Segal, Slobodan Beronja, Evan Heller, Daniel Oristian, Boris Reva, and Elaine Fuchs. Direct in vivo rnai screen unveils myosin iia as a tumor suppressor of squamous cell carcinomas. *Science*, 343(6168):309–313, 2014.

[74] Mahdi Mojallal, Yujuan Zheng, Sara Hultin, Stéphane Audebert, Tanja Van Harn, Per Johnsson, Claes Lenander, Nicolas Fritz, Christin Mieth, Martin Corcoran, et al. Amotl2 disrupts apical–basal cell polarity and promotes tumour invasion. *Nature communications*, 5, 2014.