Understanding the Impact of Non-coding Mutations in Colorectal Cancer

PRERNA GOEL
MS16093

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Under the guidance of
Dr. Sabarinathan Radhakrishnan, NCBS Bangalore

Indian Institute of Science Education and Research Mohali
April 2021

Certificate of Examination

This is to certify that the dissertation titled "Understanding the impact of non-coding mutations in Colorectal cancer" submitted by Ms. Prerna Goel (Reg. No. MS16093) for the partial fulfillment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Rajesh Ramachandran Dr. Shashi B. Pandit Dr. Kuljeet Singh Sandhu

(Local Supervisor)

Dated: April 30, 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Sabarinathan Radhakrishnan at the National Centre of Biological Sciences, Bangalore. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Prerna Goel
(Candidate)
April 30, 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Sabarinathan Radhakrishnan

(Supervisor)

Acknowledgements

I'd like to extend my sincere thanks to my advisor, Dr. Sabarinathan Radhakrishnan, for his guidance, support and patience without whose help and supervision, this thesis would have never been possible. The discussions that I had with him have enhanced my capabilities as a researcher. I am grateful to him for giving me this opportunity and allowing me to be a part of a wonderful research environment.

I'd like to thank all the members of the Lab: Bhavya, Anurag, Naveen, Rahul and Faseela for all their help and feedback. I'm incredibly grateful to them for their friendship, support and all the invaluable and fun conversations.

I'm extremely grateful to IISER Mohali for giving me the opportunity to experience the bliss of doing research over the last five years. In this time, I've grown a lot as a person, researcher, and learner. The interdisciplinary nature of the course structure has allowed me to pursue my interests for the love of Biology and data science, which still seem like two worlds apart to many people.

I'd also like to thank my thesis committee members Dr. Shashi Bhushan Pandit, Dr. Rajesh Ramachandran and my local coordinator, Dr. Kuljeet Singh Sandhu for their support and guidance. I'd like to sincerely thank Dr. N. G. Prasad, who was assigned to me as my faculty mentor at IISER. I'll always be grateful to him for the discussions I had with him during my core years at IISER and for all the amazing advice he gave to an utterly confused and introverted kid that I was. I would like to express my gratitude to KVPY and INSPIRE for the monetary support over the course of the last five years.

I'm thankful to my friends, Vishal and Tejendra for all their support and encouragement and to all my friends at IISER for just being there in this journey of learning and growing together.

Lastly, I want to thank my parents and brother, Shubham for all their love and for being incredibly supportive. I am sincerely thankful to them for being my pillars of strength and for always believing in me.

— Prerna Goel

List of Figures

List of Databases and Softwares

1. Bedtools
2. CrossMap for genome coordinates conversion between different assemblies
3. JASPAR
4. HOMER
5. NCBI Gene Expression Omnibus accession no. GSE143653
6. PCAWG
7. Promoter Capture HiC data - Orlando et al.
8. SMuRF
9. TCGA
10. UCSC Genome Browser

List of Abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
CD274	Cluster of differentiation 274
ChIP-seq	Chromatin immunoprecipitation sequencing
CRE	Cis-regulatory elements
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
ETS	Erythroblast Transformation Specific
FGFR2	Fibroblast growth factor receptor 2
HOMER	Hypergeometric Optimization of Motif EnRichment
Indels	Insertions and deletions
MSI	Microsatellite instability or microsatellite unstable
MSS	Microsatellite stable
NCBI	National Center for Biotechnology Information
PCAWG	Pan-Cancer Analysis of Whole Genomes
POLE	DNA polymerase epsilon
PCHi-C	Promoter Capture Hi-C
SMuRF	Significantly Mutated Region Finder
TCGA	The Cancer Genome Atlas
TERT	Telomerase reverse transcriptase
TF	Transcription Factor
TFBS	Transcription factor binding site
UCSC	University of California, Santa Cruz
UTR	Untranslated region
WGS	Whole-genome sequencing

Content

Abstract

The studies characterizing the genomic landscape of cancers have been majorly focused on the identification of driver mutations within the protein-coding gene regions. However, the non-coding region occupies a significantly larger proportion of the genome, and functional mutations have been reported in the regulatory regions (of non-coding regions) which can affect signaling pathways implicated in cancer. Cis-regulatory elements (CREs) are an enriched subset of the non-coding DNA and can regulate the gene expression of neighboring genes. CREs can be highly tissue-specific and hence, it becomes important to study tissue type-specific gene regulation. In this study, we used capture Hi-C data for 19,023 promoter fragments in the Colorectal cancer cell line (HT-29) from Orlando et al. and integrated it with the whole genome somatic mutation and gene expression data from PCAWG. We used the SMuRF tool to identify significantly mutated (qvalue<0.05) CREs. We identified five genes, ALCAM, PRKCH, TSC22D1, NFIB and FGFR2 with significant differential expression (p-value<0.05 and absolute fold change >= 1) in the mutated group (samples having mutations in the CRE interacting with the gene) versus the non-mutated group. Out of these five genes, we focused our analysis on FGFR2 which is a well-known cancer-driver gene, but the impact of non-coding mutations on this gene in colorectal cancer has not been reported before. We identified multiple TFBS and histone modifications in the FGFR2 CRE. We thus report a non-coding CRE interacting with the FGFR2 gene as a potential non-coding cancer driver in colorectal cancer.

Chapter 1

Introduction

1.1 Cancer

Cancer is the term used to denote a set of diseases characterized by unregulated cellular proliferation and the spread of cells from the site of origin to different parts of the body. Cancer results from genomic mutations that provide a growth advantage to the cells. (Cooper, 2000). In 2018, it is estimated that 17.0 million new cancer cases and 9.5 million cancer deaths were reported worldwide which is expected to grow to 27.5 million new cancer cases and 16.3 million cancer deaths by 2040 (Bray *et al.*, 2018). The field of cancer research has developed progressively since it was first suggested by Theodore Boveri that cancer is caused by chromosomal abnormalities a century ago (Holland and Cleveland, 2009). With the advances in genomic technologies, the complexity and heterogeneity of cancer have become much clearer to the extent that we now know that no two patients' cancers are exactly the same and hence, neither should be their treatments. This outlook has led to the development of new treatments specific to the particular subtypes of cancer-based on the studies of multiple types of omics data from cancer patients (Krzyszczyk *et al.*, 2018). However, much work still needs to be done in the domain to identify every possible genetic aberration that can cause cancer and provide what can be called precision medicine or personalized treatment for cancer.

1.1.1 Cancer genome sequencing

It is now a well-established fact in science that cancer is a genetic disease (Weinstein *et al.*, 2013). However, it is in recent years that cancer research has seen unprecedented technological advances, higher emphasis on understanding the genomic landscape of cancer and reduction in the cost of sequencing which has made the large scale cancer genome sequencing projects like The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013) and the International Cancer Genome Consortium (ICGC) (Consortium and The International Cancer Genome Consortium, 2010) feasible. The development of the revolutionary next-generation sequencing (NGS) technology enabled us to parallelly sequence millions of DNA fragments without any prior knowledge of the genomic sequence. One of the most important impacts of genome sequencing is that it has made it possible to compare the cancer genome and the normal genome of the same tissue (Mardis and Wilson, 2009) and identify what it is that makes cancer genomes so intricately complex as opposed to the genome in normal tissues. The majority of the large-scale cancer studies done so far are based on targeted sequencing or whole-exome sequencing and have accumulated loads of mutational data on the protein-coding regions (Nakagawa and Fujita, 2018). Protein-coding exons represent only 1.2% of the human genome (International Human Genome Sequencing Consortium, 2004) and hence, whole-exome sequencing is a low-cost alternative for WGS. However, exome sequencing fails to provide any information on the mutations in the non-coding regions, which have been shown to have functional and clinical implications in cancer (Nakagawa and Fujita, 2018). Whole-genome sequencing has made it possible to identify mutations in the regulatory regions like promoters, enhancers and understand their functional potential in cancer (Sakthikumar *et al.*, 2020). Still, there are many challenges to the analysis of the non-coding genome. Firstly, whole-genome sequencing has been escalated recently and not many WGS datasets are available as compared to whole-exome sequencing (Zhu *et al.*, 2020). Regulatory regions, comprising a large portion of the non-coding genome (ENCODE Project Consortium, 2012) pose another challenge as the genomic regulation in humans is highly tissue-specific.

In this study, we have used gene expression and mutation data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) which is an international collaboration to identify common

patterns of mutation in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium.

1.1.2 Colorectal Cancer

Colorectal cancer is the third most common cancer worldwide and somatic mutations leading to its progression have been widely characterized over the years (Yang *et al.*, 2019). Colon cancer and rectal cancer are together considered as a single tumor entity and called colorectal cancer. This is because colon and rectal tissues are essentially a part of the same larger organ with similar histology, functionality and certain genomic alterations in both the tumors have also been found to be similar (Cancer Genome Atlas Network, 2012).

Of all the colorectal cancers (CRCs), ~ 15% display MSI (microsatellite instability pathway). MSI results when genes that are part of DNA Mismatch Repair don't work correctly. Mismatch Repair Genes (MMR) function by correcting errors in DNA as cells divide. The remaining 80% to 85% of CRCs are microsatellite stable (Chapelle, de la Chapelle and Hampel, 2010). With respect to the MSI status, three distinct phenotypes have been defined: high microsatellite instability (MSI-H) if two or more microsatellite markers are mutated, low microsatellite instability (MSI-L); if one microsatellite marker is mutated, and microsatellite stability (MSS) if none of the examined loci is mutated (Nojadeh, Behrouz Sharif and Sakhinia, 2018; Bonneville *et al.*, 2020). At present, clinical research tends to classify MSS-L and MSS together as MSS.

1.1.3 Driver mutations in cancer

A normal cell becomes a cancer cell through the accumulation of a number of somatic mutations (Luzzatto, 2011). Mutations can be caused due to exposure to chemicals, radiation, hormones or other endogenous and exogenous factors or simply by aging (Parsa, 2012). These mutations whose accumulation can result in abnormal cell proliferation and further lead to tumor formation are known as driver mutations (Morjaria, 2020). The 'driver' mutations conferring a proliferative advantage to the cell get positively selected during the evolution of

cancer. However, any mutation found in a tumorous tissue, in fact, most of the mutations cannot be termed as driver mutations. There may be what are called passenger mutations in the cancer cells that do not undergo selection and thus, have no role in the promotion of growth of cancer cells. (Greenman *et al.*, 2007) It, therefore, becomes an important task to distinguish between these two classes of mutations when studying the genome of a cancer patient.

The mutant genes that can drive tumor progression can be classified into 1) Proto-oncogenes - Genes that require a gain of function mutation to get activated and transform normal cells to cancerous. 2) Tumor suppressor genes - Genes that normally inhibit cellular proliferation and their inactivation can lead to the development of cancer. Tumor suppressor inactivation can occur through mutations, chromosomal rearrangements or epigenetic silencing, whereas a higher gene expression due to mutations, chromosomal translocations, or abnormal signaling can lead to the activation of proto-oncogenes (Morjaria, 2020). In the past few decades, comprehensive efforts have been done to understand the genomic landscape of cancer and as a result, numerous driver mutations have been identified (Vogelstein *et al.*, 2013). However, from the recent studies, it seems that the number of mutated human cancer genes is much higher than what was predicted before and these functional mutations driving cancer are not only present in the coding genome but also in its larger non-coding counterpart. (Piraino and Furney, 2016)

1.2 The non-coding genome and its role in cancer

The non-coding genome comprises more than 98% of the genome and harbors the vast majority of somatic variants. Research has shown that there are a large number of somatic cancer mutations in regulatory DNA. For instance, Melton *et al*. found that ~40% of all somatic mutations were within regulatory regions across multiple cancer types. For the last two decades, the molecular characterization of cancer was primarily focused on protein-coding genes as described by Volgelstein's model (Ragusa *et al.*, 2015). Much research has been done in this direction, leading to the identification of hundreds of cancer driver genes in the protein-coding region. However, in recent years it has become clear that just a protein-centric approach cannot accurately explain the complexity of cancer and the non-coding genome does harbor

significant functional mutations which can contribute to cancer progression. For example, regulatory drivers have been identified in the *TERT* promoter in many cancer types. Novel non-coding cancer driver candidates are now being found in genes that are previously known to have drivers in the protein-coding region. (Rheinbay *et al.*, 2020) have identified point mutations in the 5′ region of TP53 affecting the transcription start site of the first non-coding exon and reported an important form of *TP53* inactivation by non-coding mutations. Further, mutations occurring within the functional non-coding elements, such as promoters and enhancers, can alter gene expression and affect the epigenetic state (Cuykendall, Rubin and Khurana, 2017). This discovery of a non-coding genome harboring potential cancer driver mutations is thus a huge paradigm shift in the study of cancer genomics.

1.3 Cis-regulatory elements

Cis-regulatory elements (CREs) are an enriched subset of the non-coding DNA and can regulate the gene expression of neighboring genes through the binding of transcription factors. They are termed as *cis* because they are usually located on the same DNA strand as the genes they regulate as opposed to trans, which refers to effects on genes not present on the same strand or farther upstream or downstream, such as transcription factors (TFs) (Davidson, 2010; Cuykendall, Rubin and Khurana, 2017). TFs are DNA-binding proteins that bind to regulatory elements, thereby affecting the rate of transcription of DNA to RNA. Genomic cis-regulatory elements, including promoters, enhancers and insulators, exhibit dynamic activities across different tissue types. Hence, it becomes important to study tissue type-specific gene regulation.

1.3.1 Types of cis-regulatory elements

The key cis-regulatory elements include promoters, enhancers, insulators and silencers. Promoters are regions of DNA responsible for the binding of transcriptional machinery as shown in fig 1.1. Enhancers are DNA elements that also recruit transcription factors, and which physically interact with promoter elements to regulate gene expression. A promoter element is

generally found upstream and in close proximity to the transcription start site of the gene that it regulates, whereas an enhancer element may be located upstream of a gene, within the gene, downstream of a gene, or thousands of nucleotides away. One gene promoter may interact with multiple enhancers, and similarly, one enhancer may alternately bind many different promoter elements (Andersson *et al.*, 2014).

Insulators are DNA elements that can block enhancer-promoter interactions, or function as barriers between heterochromatin and euchromatin (Ghirlando *et al.*, 2012). Silencers are regions that repress transcription either actively by binding 'negative transcription factors' called repressors, or passively by preventing the binding of transcription factors to other cis-regulatory elements (Ogbourne and Antalis, 1998; Ghirlando *et al.*, 2012).

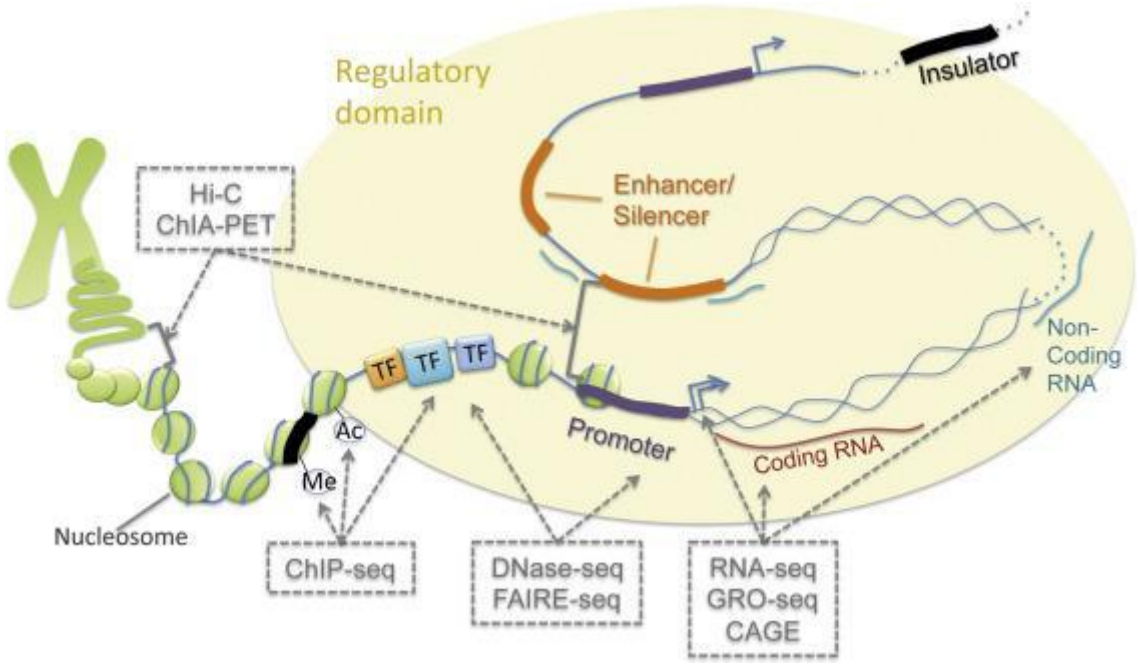


Fig 1.1 Different types of cis-regulatory elements and their identification techniques (Li *et al.*, 2015)

1.3.2 Cis-regulatory somatic driver mutations

Multiple forms of genetic mutations have been reported in the CREs including single nucleotide point mutations, insertion or deletion (indel) mutations and structural rearrangements. Point mutations are mutations affecting only a single nucleotide in a DNA sequence. As mentioned in section 1.2, the TERT promoter is a well-characterized example of a single nucleotide point mutation in the regulatory region. Mutations in the promoter of the TERT gene generate binding sequences for ETS (Erythroblast Transformation Specific) transcription factors, leading to its increased expression and are highly recurrent across multiple tumor types (Piraino and Furney, 2016). This discovery in 2013 was one of the most remarkable discoveries of somatic *cis*-regulatory mutations in cancer. In succeeding years, multiple putative point mutation drivers have been reported in the regulatory region. In breast cancer, a point mutation showing positive selection was identified within the promoter of FOXA1 – a known driver of hormone-receptor positive breast cancer (Rheinbay *et al.*, 2017).

Indels are insertions or deletion of a small number of nucleotides from the DNA sequence whereas structural rearrangements refer to large-scale rearrangement of the genome such as inversions, translocations, duplications or deletions of portions of a chromosome (Vogelstein *et al.*, 2013). For example, in 2014, the relocation of an enhancer for GATA binding protein 2 (GATA2) was demonstrated to cause haploinsufficiency of GATA2 and aberrant activation of ecotropic virus integration site 1 (EVI1). (Gröschel *et al.*, 2014) have thus shown a particular instance where structural variants causing the relocation of a single enhancer can cause deregulation of two unrelated distantly located genes leading to cancer as an outcome.

This evidence shows that many regulatory regions contain mutations subject to positive selection and hence, the regulatory mutations in cancer might have a larger role than was previously estimated.

1.3.3 Identification of cis-regulatory elements

Cis-regulatory networks differ across tissue types, therefore it becomes vital for researchers to ensure that they are analyzing tissue-type specific regulatory regions. The most common techniques to identify CREs have been DNase I hypersensitivity sequencing (DNase-seq) and ChIP-seq. DNase-seq allows the identification of active putative regulatory regions by employing the use of the open chromatin of active regulatory regions and their tendency of cleavage by DNase I enzymes (Boyle *et al.*, 2008). ChIP-seq on the other hand is a method used to identify different regulatory elements through a combination of chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to specifically look for the binding sites of DNA-associated proteins or histone marks. (Fig 1.1) Generally, promoter elements lie in close proximity to the genes that regulate however this is not the case for enhancers. Hence, it may get difficult to associate enhancers with the genes whose expression is regulated by them. Experimental approaches like chromosome conformation capture (3C)-based methods including Hi-C (van Berkum *et al.*, 2010) are now being used to understand interactions between different genomic regions in 3D space. Hi-C can be used to quantify interactions between all possible pairs of DNA fragments at the same time. Capture Hi-C further enhances Hi-C by enriching it for specific regions of interest (van Berkum *et al.*, 2010).

1.4 Promoter Capture HiC

Promoter Capture Hi-C (PCHi-C) enables the detection of all the distal genomic sequences that interact with gene promoters in a single experiment. In PCHi-C, highly complex Hi-C libraries are enriched for promoter sequences through in-solution hybrid selection with biotinylated RNA baits library targeting the promoter-containing restriction fragments. The objective is to pull down promoter sequences and their interaction partners like enhancers and other potential regulatory elements (Schoenfelder *et al.*, 2018).

Multiple studies utilizing Promoter Capture Hi-C have shown that genomic regions interacting with promoters of genes with increased expression are enriched in marks like the histone marks including H3K27 acetylation, and p300 binding which are known to be linked with enhancer activity (Mifsud et al., 2015). In our analysis, we also report a cis-regulatory region interacting

with the FGFR2 gene which was enriched in different histone marks, predominantly the H3K27ac. It is interesting to note that enhancers represent only 20% of all the promoter interacting regions uncovered by PCHi-C. This suggests that there could be other regulatory elements devoid of any enhancer marks or direct transcription regulation but can still control gene expression possibly through some structural role. From these studies, we can safely say that functional characterization of the promoter interacting regions is yet to be completed as they can comprise a variety of gene regulatory activities many of which are still not known (Schoenfelder et al., 2018).

In this study, we used capture Hi-C data for 19,023 promoter fragments in Colorectal cancer cell lines (HT-29 and Lovo, MSS and MSI, respectively) given by (Orlando *et al.*, 2018) and integrated it with PCAWG (Consortium and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) whole-genome somatic mutations and gene expression data. The analysis was done by (Orlando *et al.*, 2018) on this CRE dataset focused on identifying locus-specific recurrent mutations. (Orlando *et al.*, 2018). We followed a less stringent approach and investigated if a CRE mutation was causing differential target gene expression in colorectal cancer. We also extended our analysis to indels in addition to the single nucleotide variants (SNVs).

Chapter 2
Methodology

2.1 Classification of PCAWG colorectal sample cohort into MSS and MSI

The Pan-Cancer Analysis of Whole Genomes (PCAWG) (Consortium and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) dataset has mutation data for 52 cancer samples (https://dcc.icgc.org/pcawg). We selected only a subset of these samples that were devoid of POLE mutations to avoid any comparison with hyper-mutated tumors and were left with 47 samples. POLE encodes the DNA polymerase ε and mutations of POLE can result in the accumulation of DNA errors (Y. Li *et al.*, 2019). With respect to the MSI status, CRC can be distinguished into three types: high microsatellite instability (MSI-H), low microsatellite instability (MSI-L) and microsatellite stability (MSS) (Bonneville *et al.*, 2020). At present, clinical research tends to classify MSS-L and MSS together as MSS. We followed the same approach to separate the mutation data for MSS and MSI samples. Since the number of samples with MSI-H status was less, the further analysis was focused only on MSS samples (45 samples).

2.2 Mapping non-coding mutations to CREs

We mapped non-coding mutations from the PCAWG dataset to the CREs defined by promoter CHi-C generated on CRC cell line HT29 (MSS, non-POLE samples) using bedtools intersect function.

2.3 Identification of significantly mutated CREs

We used the tool SMuRF(Guilhamon and Lupien, 2018) to identify significantly mutated CREs. The SMuRF tool takes 3 files as input including the variants file, a bed file for the genomic regions of interest (CREs in this case) and GENCODE 1g19 promoter annotation. The significantly mutated regions were selected based on 0.05 qval cutoff given by the software. Further, another filter was set to choose only the CREs with mutations in at least 3 samples so as to have a comparable number in the mutated and non-mutated group.

2.4 Differential Gene Expression analysis

All the significantly mutated CREs were divided into two groups, mutated and non-mutated samples and differential gene expression analysis was done. The Wilcoxon rank-sum test was used to compare the two groups. The significantly expressed genes were selected based on p-value < 0.05 and absolute fold change >= 1.

2.5 Gene copy number analysis

The CREs which showed significant differential expression in the mutated versus non-mutated group were also checked for the gene copy number in all the mutated samples to make sure that the differential expression is an effect of the mutation in the CRE and not because of gene amplification. For the FGFR2 gene, the copy number was also checked for all the colorectal samples used in the analysis for which the data was available (44 samples) to understand the distribution of FGFR2 copy number in the samples used.

2.6 Accessibility and expression correlation in CREs

TCGA ATAC-seq normalized bigwig files for 81 colorectal samples were obtained from Corces & Granja et al (Corces *et al.*, 2018; Guilhamon and Lupien, 2018). The CREs for HT-29 cell line (Microsatellite stable) were obtained from (Orlando *et al.*, 2018). The CRE genomic coordinates obtained were lifted over from hg19 to hg38 using the tool CrossMap (Zhao *et al.*, 2014; Orlando *et al.*, 2018; Poernomo and Kang, 2018) and the chain files for the hg19 to hg38 conversion were obtained from UCSC Genome Browser (Kent, 2002). Each bigwig file was overlapped with the CREs for the colorectal MSS samples. All the CREs had 100 percent overlap with the ATAC seq regions since the bigwig files provide continuous accessibility data for the genome and each 100-bp bin represents the normalized number of insertions that occurred within the corresponding 100 bp. Further, the average accessibility score was calculated for each CRE intersecting with multiple 100bp peaks. TCGA Expression data were available for 25 Colorectal samples. We calculated the Pearson correlation coefficient between gene expression and accessibility for the ~97k unique CREs using the SciPy package in python. We specifically checked the correlation value for the FGFR2 CRE (122666324122667822_FGFR2) and made an accessibility correlation plot for the same.

2.7 Genomic Annotation of CRE peaks

We used the annotatePeaks function from HOMER tool (Heinz *et al.*, 2010) to annotate all the CRE peaks for HT-29 cell line obtained from (Orlando *et al.*, 2018) and identify whether the CRE falls in the TSS (transcription start site), TTS (transcription termination site), Exon (Coding), 5' UTR Exon, 3' UTR Exon, Intronic, or Intergenic. We then specifically checked the annotation of the CREs for which the interacting genes showed significant differential gene expression.

2.8 FGFR2 Specific Analysis

Our analysis yielded five CREs leading to significant differential gene expression in the mutated group of samples versus the non-mutated group, of which FGFR2 is a well known cancer-driver gene. We further focused our analysis on FGFR2 as the impact of non-coding

mutations on this gene in colorectal cancer has not been reported before. We checked for the specific transcriptions factors (TFs) and histone markers overlapping with the FGFR2 CRE (122666324122667822_FGFR2).

2.8.1 Identification of TF motif in the FGFR2 CRE

The FGFR2 CRE was searched for any TFBS using the JASPAR Core Vertebrates collection(2020) (Fornes *et al.*, 2020) for the hg19 genome. The search was performed using the table browser tool on UCSC browser. The mutated regions bed file was uploaded on the browser. The output file obtained contained the chromosome number, genomic coordinates and TF name which were intersecting with the input file. The output TFBS file was overlapped with the mutated regions file using bedtools intersect.

2.8.2 Mapping histone modifications in the FGFR2 CRE

The histone modification marks for the HT-29 cell line(MSS) were obtained from the NCBI Gene Expression Omnibus accession no. GSE143653 ((Kent, 2002; Gopi and Kidder, 2021). The histone peaks were intersected with the FGFR2 CREs whose genes showed significant differential expression in the mutated versus the non-mutated group using bedtools intersect function (Quinlan and Hall, 2010). A graphical representation of the same was also made using the Integrated Genomics Viewer (Robinson *et al.*, 2011).

2.8.3 Pan-cancer level analysis of the FGFR2 CRE

We checked if the FGFR2 CRE was mutated in any cancer types other than colorectal cancer in the PCAWG database. We checked if the expression data were available for all those samples. The Wilcoxon rank-sum statistic test was done between the expression values of the mutated group of samples with all the non-mutated samples from all tissue types.

Chapter 3

Results and Conclusions

3.1 Distribution of mutations in the colorectal MSS samples

We performed the analysis of 45 Colorectal samples (MSS and MSI-L, Non-POLE). The following bar plot (Fig 3.1) represents the number of mutations per sample. The x-axis represents the TCGA Id of all the colorectal samples. The majority of the samples considered have similar mutation counts. However, two samples towards the extreme left seem to have a relatively higher mutation count than the rest of the samples inspite of having excluded the POLE mutated samples. Both of these samples are of MSS type. We have included these two samples in the analysis but we confirmed that the altered CREs with significant differential expression were not solely driven by these hyper-mutated samples as it may represent a bias in the analysis.

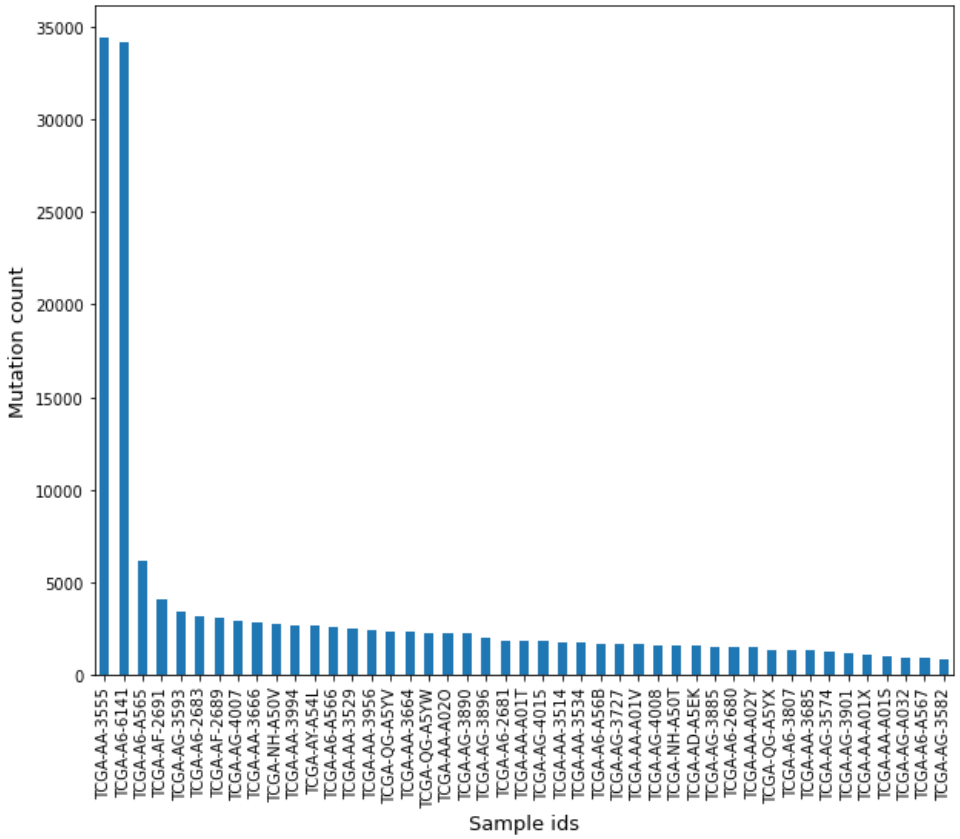


Figure 3.1: Bar plot displaying the number of mutations in each of the 45 samples analyzed

3.2 SMuRF output

We obtained a total of 1991 CRE mutations with a qval cutoff of 0.05 after running the SMuRF tool. We further filtered the CREs with a mutation in at least 3 samples and were left with 514 CREs. The SMuRF output was merged with the original CRE file containing the CRE coordinates along with the interacting gene names. The gene expression data was available for only a subset of the genes which were interacting with a total of unique 422 CREs. A unique CRE id was allocated to each CRE by combining the interacting gene name with the start and end coordinates of the CRE for easy identification. The coordinates used for the id were for hg19 genome assembly.

3.3 Differentially expressed genes with mutations in CRE

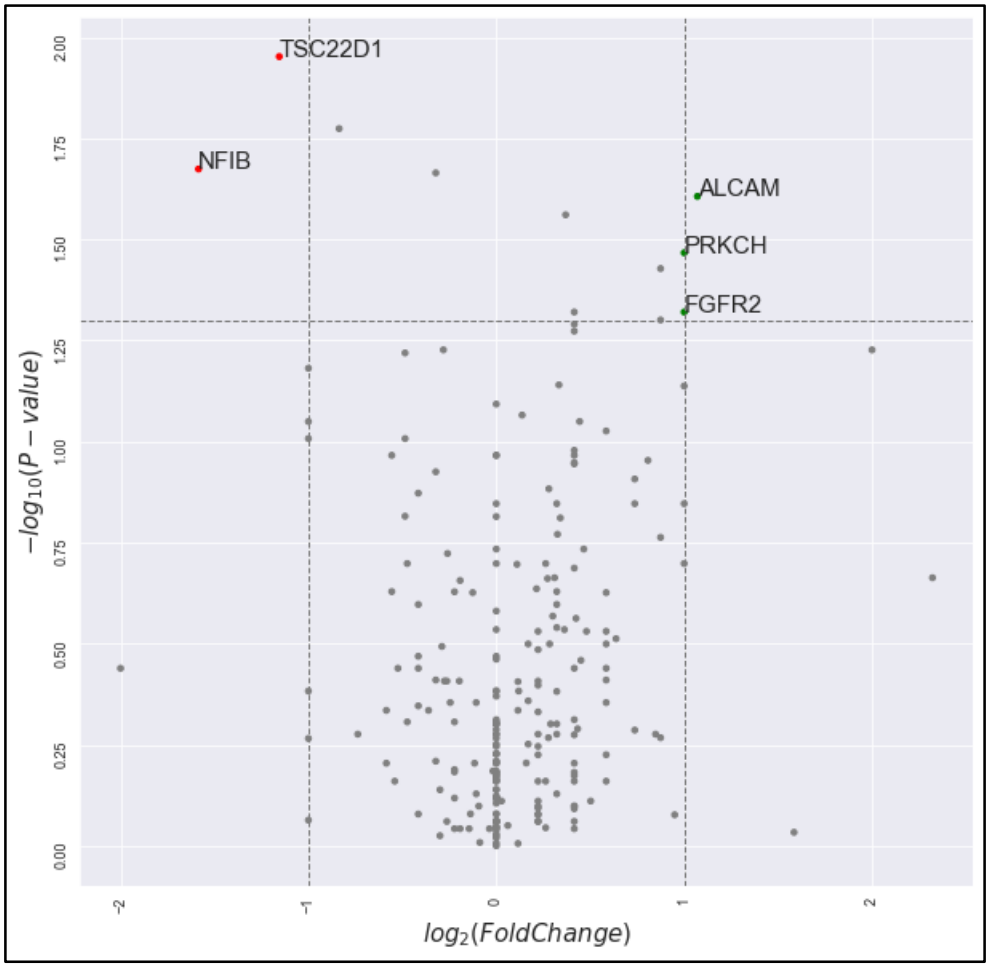


Figure 3.2: Volcano plot for the differentially expressed genes with mutations in the CRE. Each dot represents a gene. The red dots represent genes whose expression is downregulated upon mutation and green dots represent the upregulated ones. The grey dots denote the genes with significantly mutated CRE but not having a significant effect on the gene expression of the interacting gene.

The Volcano plot here has been generated by the Bioinfokit package in python, with default log fold-change thresholds of -1 and 1 and an adjusted P-value threshold of 0.05.

The genes ALCAM, PRKCH, TSC22D1, NFIB and FGFR2 were found to be differentially expressed with a p value<0.05 (Wilcoxon rank-sum statistic) and fold change expression >=

|1|. We see that ALCAM, PRCKCH and FGFR2 are upregulated in the CRE mutated samples whereas TSC22D1 and NFIB are downregulated. ALCAM (Activated leukocyte cell adhesion molecule) codes for a protein that binds to T-cell differentiation antigen CD6, and is thought to be involved in thymic epithelial cells and thymocyte interaction. ALCAM has been shown to have higher expression in colorectal cancer (Weichert *et al.*, 2004). Protein Kinase C (PRKCH) is a member of the serine- and threonine-specific protein kinases family and is known to be involved in diverse cellular signaling pathways. It is known to be involved in cell proliferation and differentiation and hence its overexpression can have a significant contribution in cancer (Weichert *et al.*, 2004; Basu, 2019).

Fibroblast growth factor receptor 2 (FGFR2) is a receptor tyrosine kinase that induces cell proliferation and migration. It is a well-known cancer driver gene and its overexpression has been shown to contribute to tumor progression in multiple cancer types (Szybowska *et al.*, 2019).

While FGFR2 amplification has been reported recurrently in breast and gastric cancer, its activating mutations in the protein-coding regions have also been found in several types of cancer (Carter *et al.*, 2017). However, there have been no reports of mutations in any non-coding region affecting the expression of FGFR2 and hence, contributing to cancer progression. In this study, we identified a cis-regulatory region interacting with the FGFR2 gene which caused a significant differential expression (p-value = 0.047) in the samples having a mutation in that CRE. We further checked the copy number variation in the samples with a mutation in the FGFR2 CRE and found that all three samples had no copy number variation in the FGFR2 gene ruling out the possibility that the increase in gene expression could be because of the amplification of FGFR2. The total sample cohort taken for analysis (45 samples) had only one sample with a reported gene amplification of FGFR2.

TSC22D1 (TSC22 Domain Family Member 1) is known to be a transcriptional repressor that regulates the transcription of multiple genes and is a putative tumor suppressor (Nakamura *et al.*, 2012). We see a clear drop in the gene expression of TSC22D1 upon mutation of its interacting CRE, thereby indicating the importance of the regulatory region in the expression of TSC22D1. It is possible that TSC22D1 was regulating the transcription of some important oncogene and the downregulation of TSC22D1 was now contributing to cancer progression.

Nuclear factor I/B (NFIB) regulates the transcription of a variety of genes and is known to have both tumor-suppressive and oncogenic roles (Becker-Santos *et al.*, 2017). In the current analysis, we see the downregulation of NFIB upon mutation in its interacting regulatory region.

P-value - 0.024

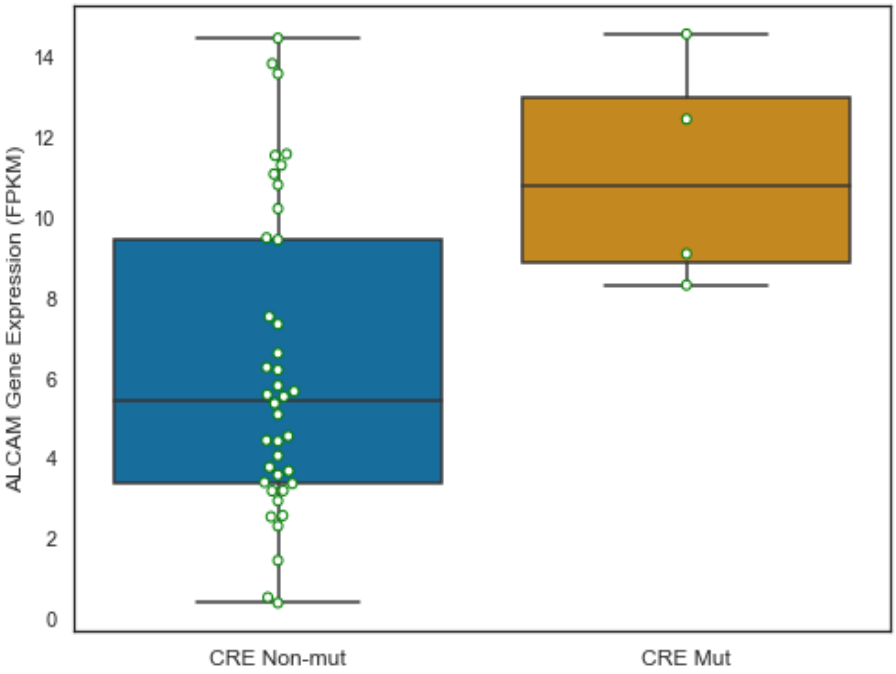


Figure 3.3: ALCAM Gene Expression in CRE mutated versus non-mutated samples

In Fig 3.3, the green bordered dots represent an individual sample. The Wilcoxon rank-sum statistic test was done between the expression values of the mutated group of samples with all the non-mutated samples.

P-value - 0.021

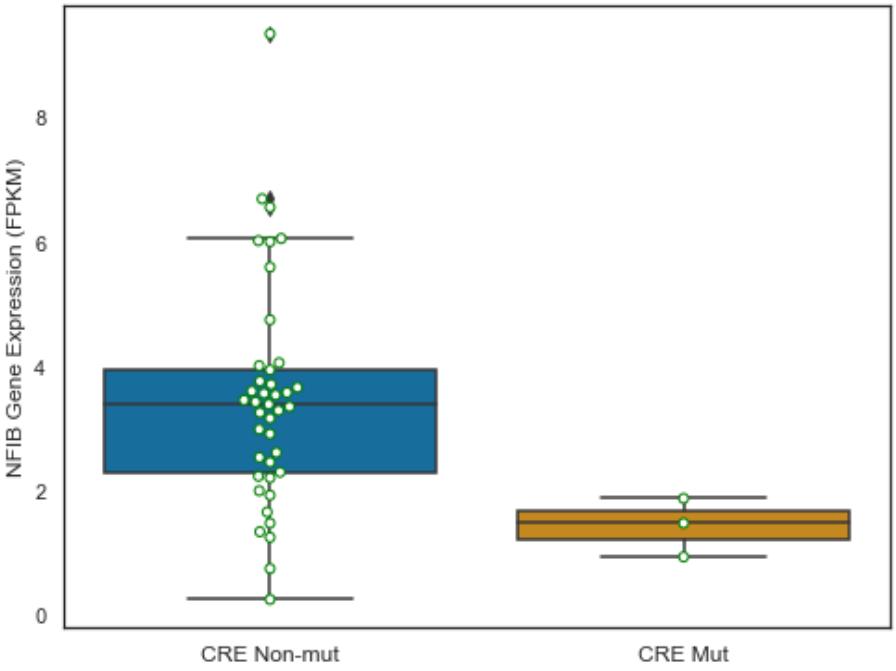


Figure 3.4: NFIB Gene Expression in CRE mutated versus non-mutated samples

P-value - 0.011

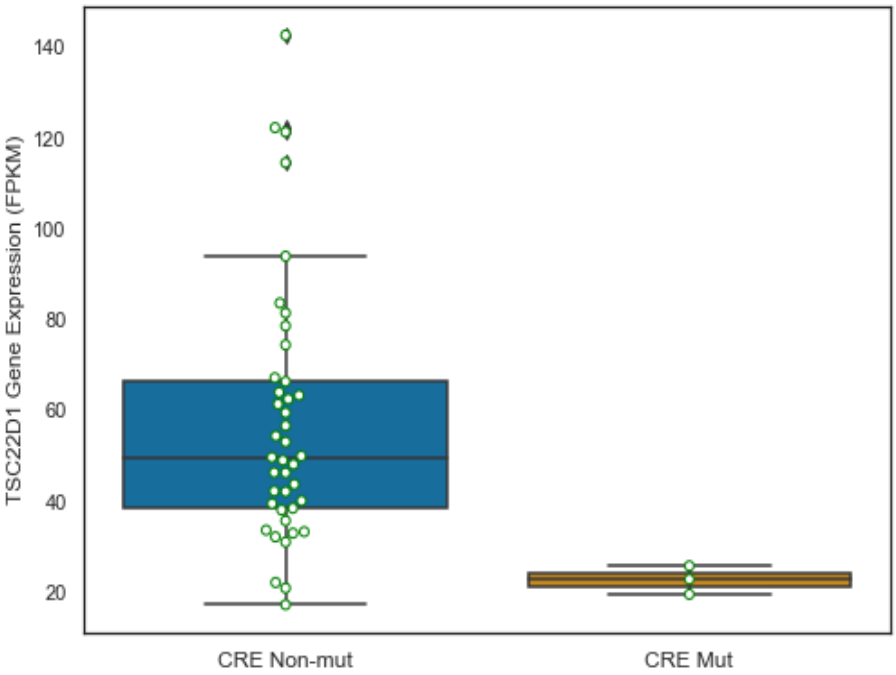


Figure 3.5: TSC22D1 Gene Expression in CRE mutated versus non-mutated samples

P-value - 0.034

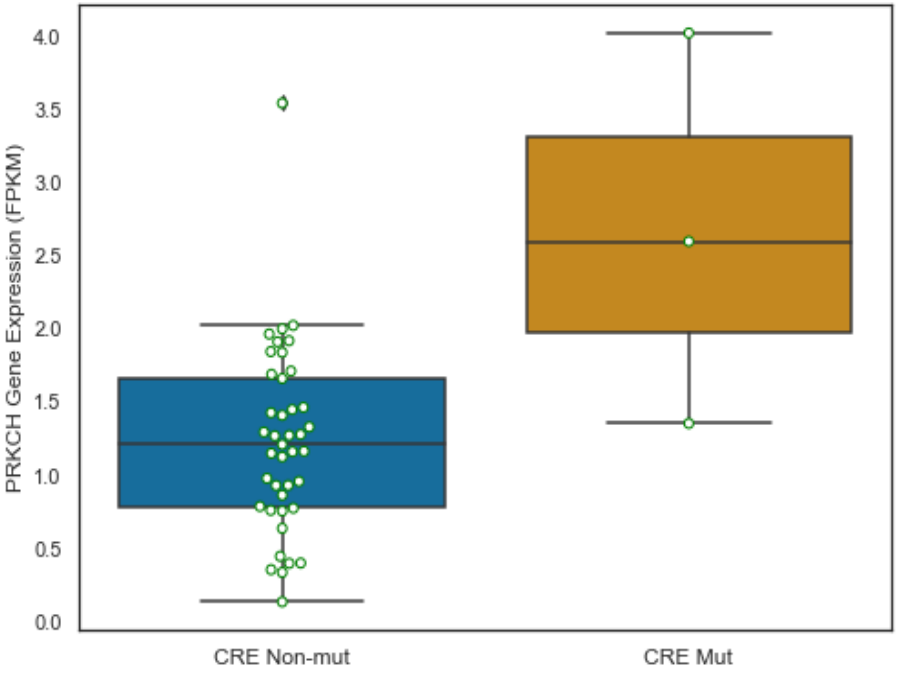


Figure 3.6: PRKCH Gene Expression in CRE mutated versus non-mutated samples

P-value - 0.047

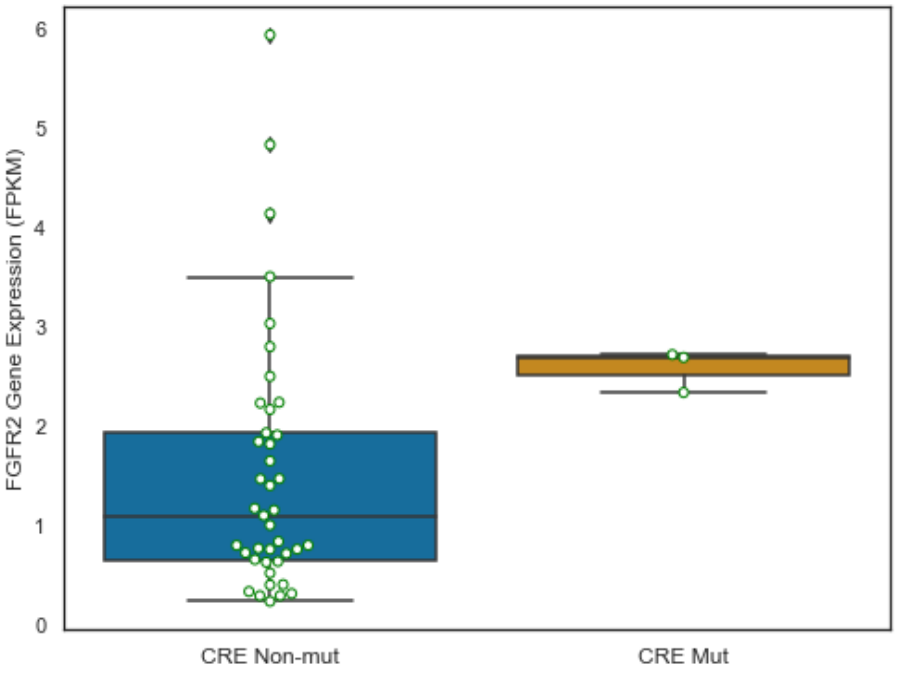


Figure 3.7: FGFR2 Gene Expression in CRE mutated versus non-mutated samples

3.4 CD274 gene expression in samples with and without FGFR2 CRE mutation

FGFR2 gene is known to promote the expression of CD274 gene through JAK/STAT3 signaling pathway in colorectal cancer. CD274 gene has been shown to be a negative immune regulatory protein (P. Li *et al.*, 2019). Thus, FGFR2 overexpression in cancer cells leads to increased cellular proliferation and also induces CD274 overexpression which further suppresses the immune response. We checked if the non-coding mutation in FGFR2 CRE causing its overexpression would also induce the CD274 overexpression in the mutated samples. The CD274 expression in the samples with mutated FGFR2 CRE was higher than the non-mutated counterpart. However, the difference between the two groups was not statistically significant. This could be due to the less number of samples available in the CRE mutated group than the non-mutated group.

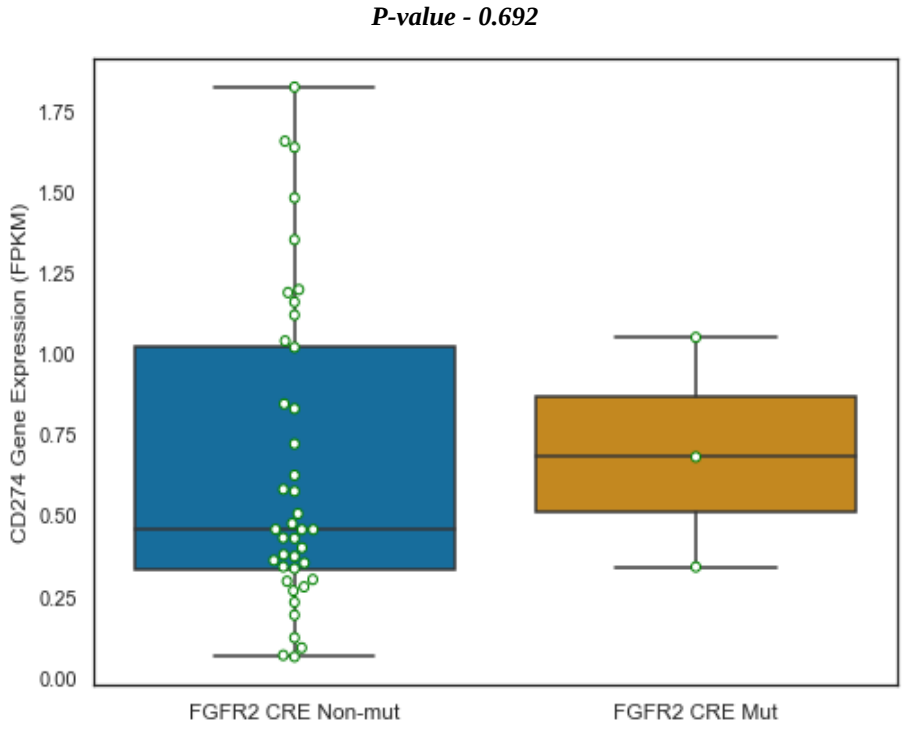


Figure 3.8: CD274 Gene Expression in FGFR2 CRE mutated versus non-mutated samples

3.5 HOMER genomic annotation of the CRE peaks

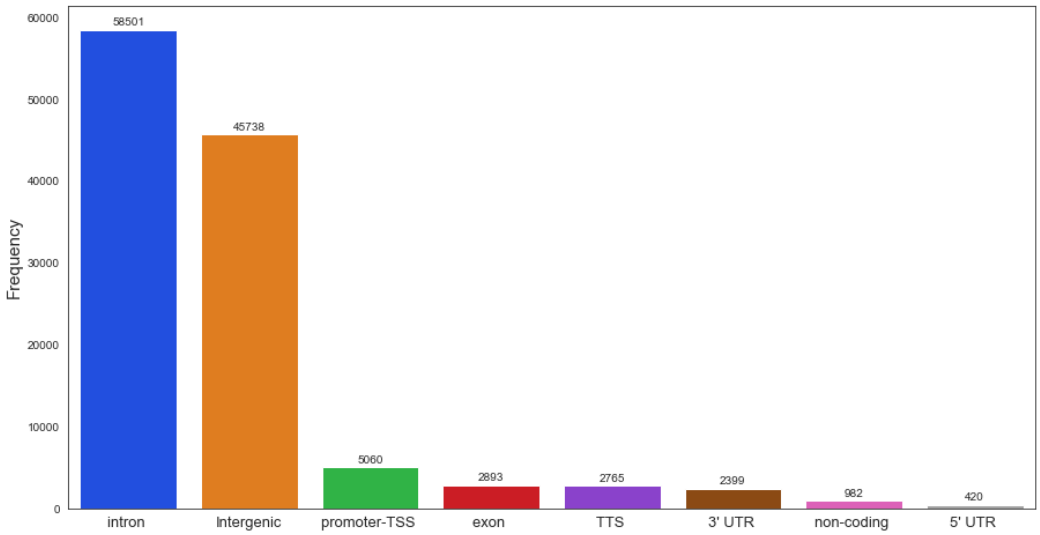


Figure 3.9: Distribution of CRE peaks in functional genomic regions according to HOMER genomic annotation

We see that most of the promoter Hi-C capture peaks were prevalent in the intronic and intergenic regions. Exonic peaks contribute only 2.44% of the total peaks suggesting that the majority of the regions captured by promoter Hi-C are non-coding regions. Based on the genomic annotation, we identified that the FGFR2 CRE falls in the intronic region of the WDR11 gene. It could be seen from the UCSC Browser genome track that the FGFR2 gene lies ~5Mb downstream of WDR11. WDR11 gene is a tumor suppressor gene disrupted in glial tumors (17). The following are the coordinates of the CRE and two genes in the hg19 genome assembly.

FGFR2 CRE - chr10:122,666,324-122,667,822

WDR11 gene - chr10:122,610,687-122,669,038

FGFR2 gene - chr10:123,237,844-123,357,972

We further checked if there was any difference in the gene expression of WDR11 in the

samples with FGFR2 CRE mutated as compared to the non-mutated samples. We found that there is no significant difference in the expression of WDR11 gene in both groups. Thus, the WDR11 gene probably did not contribute to the increased expression of FGFR2 in CRE mutated samples.

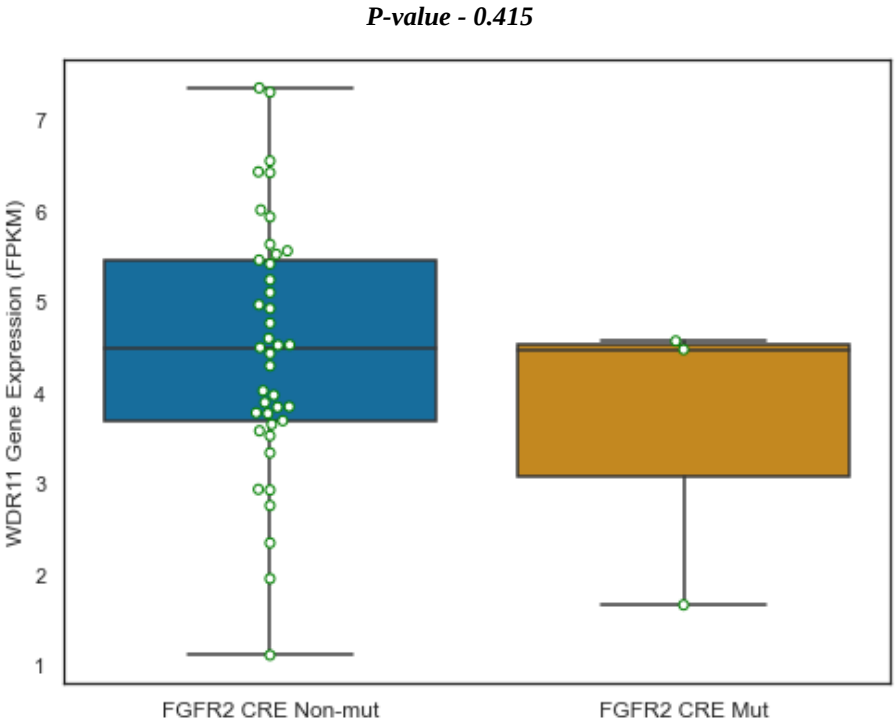


Figure 3.10: WDR11 Gene Expression in CRE mutated versus non-mutated samples

3.6 TFBS in CREs

We found that the FGFR2 CRE mutated region intersects with various transcription factor binding sites including, Rhox11, PRDM4, ZNF384, PHOX2B, MEF2A, FOXC1, FOXC2, Foxj3, STAT1::STAT2 and ONECUT3. This implicates that it is possible that the mutation in the FGFR2 CRE is changing the binding affinity of one or many of the transcription factors to that region and thereby, causing a significant change in the gene expression.

3.7 Histone modifications in the FGFR2 CRE

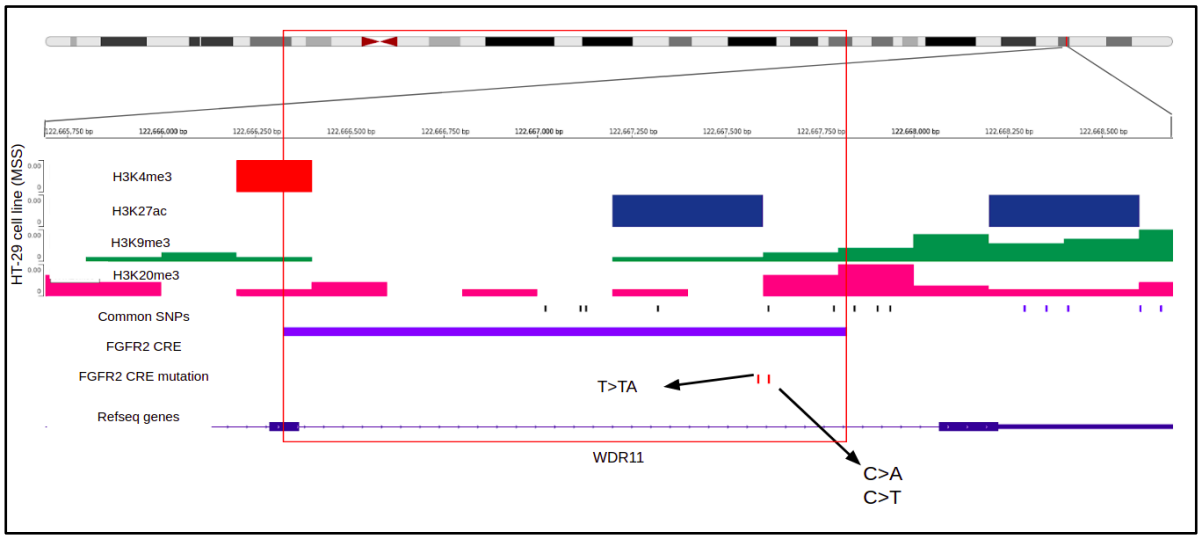


Figure 3.11: Graphical representation of histone modifications on the FGFR2 CRE genomic track.

In Fig 3.11, the red box represents the histone modifications in the FGFR2 CRE. A total of 3 mutations were found in the given CRE including two point mutations and one indel where both the point mutations were identified at the same genomic location in two different samples. The FGFR2 CRE harbors multiple histone modifications including H3K4me3, H3K27ac, H3K9me3 and H4K20me3. In general, the H3K4me3 modification is enriched at active promoters near transcription start sites and regarded as a transcription activation biomarker (Howe *et al.*, 2017). Similarly, the H3K27ac marker denotes active gene transcription (Roth, Denu and Allis, 2001) while H3K9me3 is a heterochromatin-associated histone mark specific for facultative heterochromatin (Saksouk, Simboeck and Déjardin, 2015) and hence, indicates repressed transcriptional activity in neighboring genome regions. H4K20me3 is known to be associated with the formation of pericentric hetereochromatin. Thus, from the above-mentioned four epigenetic marks, H3K4me3 and H3K27ac are associated with transcriptional activation and H3K9me3 and H4K20me3 are associated with transcriptional suppression. All the histone marks were obtained for the HT-29 cell line.

3.8 Correlation between the accessibility of all the CREs and the expression of their linked genes in TCGA colorectal samples

We checked the correlation between the accessibility of all the CREs and the expression of their linked genes for all the TCGA colorectal cancer samples irrespective of their MSI status. The FGFR2 CRE showed a positive correlation of 0.46 between accessibility and expression. The distribution of the correlation values, ranging from -0.818 to 0.897 for all the CREs cataloged have been plotted below. Therefore, the correlation value of 0.46 for the FGFR2 is towards the higher end of the distribution (Fig 3.13).

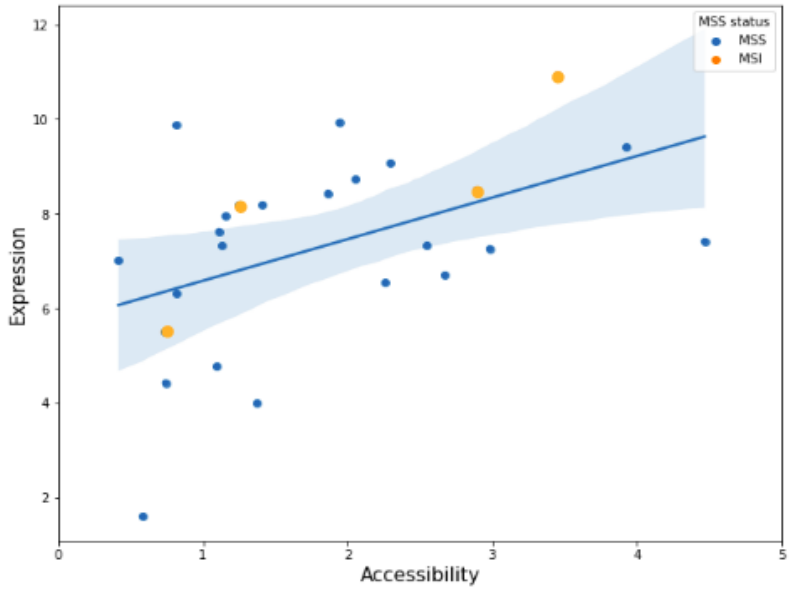


Figure 3.12: Expression versus accessibility correlation plot for FGFR2 CRE in all TCGA Colorectal samples. Each dot represents one sample.

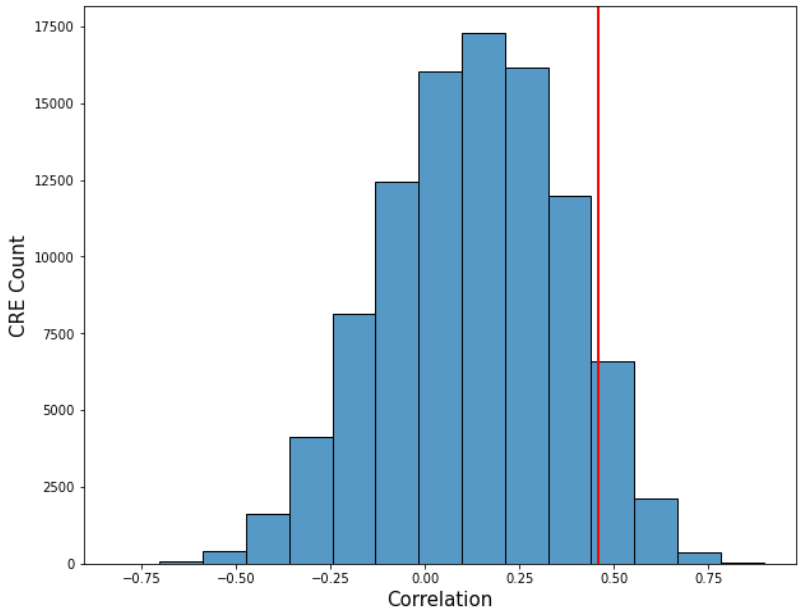
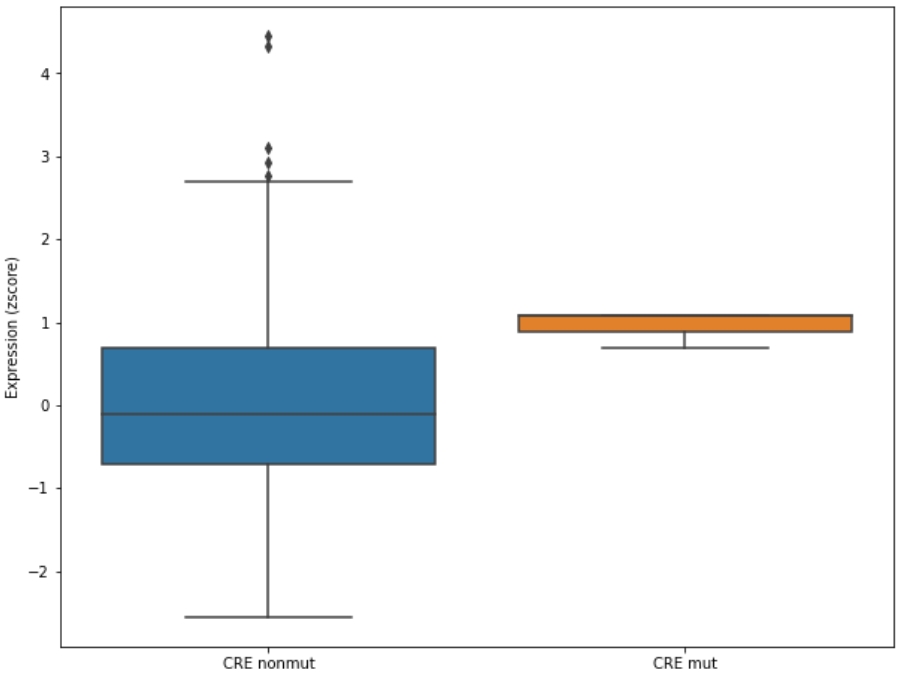


Figure 3.13: Distribution of correlation between the accessibility of all the CREs and the expression of their linked genes in TCGA colorectal samples

3.9 FGFR2 differential gene expression in mutated versus non-mutated CRE group at Pan-cancer level

The FGFR2 CRE was found to be mutated in 8 cancer tissue types other than colorectal in the PCAWG database including Biliary-AdenoCA, Lung-SCC, Panc-AdenoCA, Skin-Melanoma, Stomach-AdenoCA, Eso-AdenoCa, Liver-HCC and Ovary-AdenoCA. The PCAWG gene expression data was available for only 3 of these 8 tumor tissue type samples. Colorectal Cancer was not considered for this particular analysis. The FGFR2 gene expression was significantly higher (pvalue-0.05) in the samples having mutations in the FGFR2 CREs as compared to all the non-mutated samples at the Pan-Cancer level. This indicates that this CRE might be important on a Pan-Cancer level and any mutation in the region could be causing some dysregulation of the gene.

P-value - 0.05



Pan cancer

Lung-SCC

Stomach-AdenoCA

Ovary-AdenoCA

Figure 3.14 FGFR2 gene expression in mutated versus non-mutated CRE samples at Pan-Cancer level

Chapter 4

Discussion

The identification of driver mutations in cancer has been shifted from protein to non-coding in recent years. Hence, we focused on the non-coding regions, specifically the cis-regulatory regions to understand their role in cancer development. The previous study done on the same dataset of cis-regulatory regions in Colorectal cancer (Orlando *et al.*, 2018) has focused on identifying locus-specific recurrent mutations in the CREs affecting gene expression. We performed a similar analysis but used a less stringent approach as the enrichment of cis-regulatory regions is not in very narrow regions.

Our analysis is complementary to the identification of Cis-eQTL. Expression quantitative trait loci (eQTLs) are genomic loci that influence variation in mRNAs expression levels and eQTLs when located near the gene of origin are called as cis-eQTLs. It has been shown that naturally occurring eQTLs are enriched in promoter-interacting regions that are connected to the same genes whose expression is affected by the eQTLs (Javierre et al., 2016).

In our analysis, we identified five CREs whose mutation caused the altered expression of their interacting genes. Out of these five genes, FGFR2, a well-known cancer-driver gene plays an important role in cellular proliferation, migration, anti-apoptosis, angiogenesis, wound healing, and tissue regeneration (Sun *et al.*, 2020). It is known to be amplified prominently in breast and gastric cancer and activating mutations in the coding region have also been reported in many cancer types including gastric cancer, breast cancer, and endometrial carcinoma. (Babina

and Turner, 2017). However, the regulation through regulatory regions in the non-coding genome is still an unexplored domain for the FGFR2 gene. We also confirmed that there is no copy number variation in the samples with a mutation in the FGFR2 CRE ruling out the possibility that the altered gene expression could be because of the amplification of FGFR2. Our analysis explores the domain of studying tissue-specific regulatory regions and provides evidence for the presence of cancer drivers in the non-coding genome.

The limitations of our study are that we have used the CRE data from the HT-29 cell line whereas the mutation and expression data is from colorectal primary tumors. The genes that are mutated in the two groups may not be the same. Hence, it would be better if both the CREs and expression data is from the same source. Also, we need data from the normal cells to see how well the CRE is influencing the expression upon mutation. Another limitation of our analysis is that the sample size is very low and we need more samples to strongly establish the statistical significance of the CRE mutation causing an altered gene expression.

Further analysis can be done to study the effect of structural variants in the cis-regulatory regions on the gene expression of their linked gene. Structural variants are known to cause tumor development by disrupting genes or by causing gene copy number variation and thus, it can be interesting to see if structural variants could cause dysregulation of genes by disrupting the cis-regulatory regions. Experimental perturbation of the analyzed CREs (with significant differential expression) or disrupting the genome with the selected CRE mutations can also be done to observe the effect on gene expression.

Bibliography

Andersson, R. *et al.* (2014) ‘An atlas of active enhancers across human cell types and tissues’, *Nature*, 507(7493), pp. 455–461.

Babina, I. S. and Turner, N. C. (2017) ‘Advances and challenges in targeting FGFR signalling in cancer’, *Nature reviews. Cancer*, 17(5), pp. 318–332.

Basu, A. (2019) ‘The Enigmatic Protein Kinase C-eta’, *Cancers*, 11(2). doi: 10.3390/cancers11020214.

Becker-Santos, D. D. *et al.* (2017) ‘Nuclear Factor I/B: A Master Regulator of Cell Differentiation with Paradoxical Roles in Cancer’, *EBioMedicine*, pp. 2–9. doi: 10.1016/j.ebiom.2017.05.027.

van Berkum, N. L. *et al.* (2010) ‘Hi-C: A Method to Study the Three-dimensional Architecture of Genomes’, *Journal of Visualized Experiments*. doi: 10.3791/1869.

Bonneville, R. *et al.* (2020) ‘Detection of Microsatellite Instability Biomarkers via Next-Generation Sequencing’, *Methods in molecular biology* , 2055, pp. 119–132.

Boyle, A. P. *et al.* (2008) ‘High-resolution mapping and characterization of open chromatin across the genome’, *Cell*, 132(2), pp. 311–322.

Bray, F. *et al.* (2018) ‘Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries’, *CA: a cancer journal for clinicians*, 68(6), pp. 394–424.

Cancer Genome Atlas Network (2012) ‘Comprehensive molecular characterization of human colon and rectal cancer’, *Nature*, 487(7407), pp. 330–337.

Carter, J. H. *et al.* (2017) ‘amplification in colorectal adenocarcinoma’, *Cold Spring Harbor molecular case studies*, 3(6). doi: 10.1101/mcs.a001495.

Chapelle, A. de la, de la Chapelle, A. and Hampel, H. (2010) ‘Clinical Relevance of Microsatellite Instability in Colorectal Cancer’, *Journal of Clinical Oncology*, pp. 3380–3387. doi: 10.1200/jco.2009.27.0652.

Consortium, T. I. C. G. and The International Cancer Genome Consortium (2010) 'International network of cancer genome projects', *Nature*, pp. 993–998. doi: 10.1038/nature08987.

Consortium, T. I. P.-C. A. of W. G. and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) 'Pan-cancer analysis of whole genomes', *Nature*, pp. 82–93. doi: 10.1038/s41586-020-1969-6.

Cooper, G. M. (2000) *The Cell: A Molecular Approach*. Sinauer Associates.

Corces, M. R. *et al.* (2018) 'The chromatin accessibility landscape of primary human cancers', *Science*, 362(6413). doi: 10.1126/science.aav1898.

Cuykendall, T. N., Rubin, M. A. and Khurana, E. (2017) 'Non-coding genetic variation in cancer', *Current Opinion in Systems Biology*, pp. 9–15. doi: 10.1016/j.coisb.2016.12.017.

Davidson, E. H. (2010) *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Elsevier.

ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74.

Fornes, O. *et al.* (2020) 'JASPAR 2020: update of the open-access database of transcription factor binding profiles', *Nucleic acids research*, 48(D1), pp. D87–D92.

Ghirlando, R. *et al.* (2012) 'Chromatin domains, insulators, and the regulation of gene expression', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, pp. 644–651. doi: 10.1016/j.bbagrm.2012.01.016.

Gopi, L. K. and Kidder, B. L. (2021) 'Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains', *Nature communications*, 12(1), p. 1419.

Greenman, C. *et al.* (2007) 'Patterns of somatic mutation in human cancer genomes', *Nature*, 446(7132), pp. 153–158.

Gröschel, S. *et al.* (2014) 'A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia', *Cell*, 157(2), pp. 369–381.

Guilhamon, P. and Lupien, M. (2018) 'SMuRF: a novel tool to identify regulatory elements enriched for somatic point mutations', *BMC bioinformatics*, 19(1), p. 454.

Heinz, S. *et al.* (2010) 'Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities', *Molecular Cell*, pp. 576–589. doi: 10.1016/j.molcel.2010.05.004.

Holland, A. J. and Cleveland, D. W. (2009) 'Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis', *Nature reviews. Molecular cell biology*, 10(7), pp. 478–487.

Howe, F. S. *et al.* (2017) 'Is H3K4me3 instructive for transcription activation?', *BioEssays: news and reviews in molecular, cellular and developmental biology*, 39(1), pp. 1–12.

International Human Genome Sequencing Consortium (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, 431(7011), pp. 931–945.

Javierre, B. M. et al. (2016) 'Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters', Cell, 167(5), pp. 1369–1384.e19.

Kent, W. J. (2002) 'The Human Genome Browser at UCSC', *Genome Research*, pp. 996–1006. doi: 10.1101/gr.229102.

Krzyszczyk, P. *et al.* (2018) 'The growing role of precision and personalized medicine for cancer treatment', *Technology*, 6(3-4), pp. 79–100.

Li, P. *et al.* (2019) 'FGFR2 Promotes Expression of PD-L1 in Colorectal Cancer via the JAK/STAT3 Signaling Pathway', *Journal of immunology*, 202(10), pp. 3065–3075.

Li, Y. *et al.* (2015) 'The identification of cis-regulatory elements: A review from a machine learning perspective', *Bio Systems*, 138, pp. 6–17.

Li, Y. *et al.* (2019) 'POLE mutations improve the prognosis of endometrial cancer via regulating cellular metabolism through AMF/AMFR signal transduction', *BMC medical genetics*, 20(1), p. 202.

Luzzatto, L. (2011) 'Somatic mutations in cancer development', *Environmental health: a global access science source*, 10 Suppl 1, p. S12.

Mardis, E. R. and Wilson, R. K. (2009) 'Cancer genome sequencing: a review', *Human Molecular Genetics*, pp. R163–R168. doi: 10.1093/hmg/ddp396.

Morjaria, S. (2020) 'Driver mutations in oncogenesis', *International Journal of Molecular & Immuno Oncology*, pp. 1–3. doi: 10.25259/ijmio_26_2020.

Mifsud, B. et al. (2015) 'Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C', Nature genetics, 47(6), pp. 598–606.

Nakagawa, H. and Fujita, M. (2018) 'Whole genome sequencing analysis for cancer genomics and precision medicine', *Cancer Science*, pp. 513–522. doi: 10.1111/cas.13505.

Nakamura, M. *et al.* (2012) 'Transforming growth factor-β-stimulated clone-22 is a negative-feedback regulator of Ras / Raf signaling: Implications for tumorigenesis', *Cancer science*, 103(1), pp. 26–33.

Nojadeh, J. N., Behrouz Sharif, S. and Sakhinia, E. (2018) 'Microsatellite instability in colorectal cancer', *EXCLI journal*, 17, pp. 159–168.

Ogbourne, S. and Antalis, T. M. (1998) 'Transcriptional control and the role of silencers in

transcriptional regulation in eukaryotes', *Biochemical Journal*, 331 (Pt 1), pp. 1–14.

Orlando, G. *et al.* (2018) 'Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer', *Nature genetics*, 50(10), pp. 1375–1380.

Parsa, N. (2012) 'Environmental factors inducing human cancers', *Iranian journal of public health*, 41(11), pp. 1–9.

Piraino, S. W. and Furney, S. J. (2016) 'Beyond the exome: the role of non-coding somatic mutations in cancer', *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 27(2), pp. 240–248.

Poernomo, A. and Kang, D.-K. (2018) 'Biased Dropout and Crossmap Dropout: Learning towards effective Dropout regularization in convolutional neural network', *Neural networks: the official journal of the International Neural Network Society*, 104, pp. 60–67.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, pp. 841–842. doi: 10.1093/bioinformatics/btq033.

Ragusa, M. *et al.* (2015) 'Non-coding landscapes of colorectal cancer', *World journal of gastroenterology: WJG*, 21(41), pp. 11709–11739.

Rheinbay, E. *et al.* (2017) 'Recurrent and functional regulatory mutations in breast cancer', *Nature*, 547(7661), pp. 55–60.

Rheinbay, E. *et al.* (2020) 'Analyses of non-coding somatic drivers in 2,658 cancer whole genomes', *Nature*, 578(7793), pp. 102–111.

Robinson, J. T. *et al.* (2011) 'Integrative genomics viewer', *Nature Biotechnology*, pp. 24–26. doi: 10.1038/nbt.1754.

Roth, S. Y., Denu, J. M. and Allis, C. D. (2001) 'Histone acetyltransferases', *Annual review of biochemistry*, 70, pp. 81–120.

Saksouk, N., Simboeck, E. and Déjardin, J. (2015) 'Constitutive heterochromatin formation and transcription in mammals', *Epigenetics & chromatin*, 8, p. 3.

Sakthikumar, S. *et al.* (2020) 'Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes', *Genome biology*, 21(1), p. 127.

Schoenfelder, S. *et al.* (2018) 'Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions', *Journal of visualized experiments: JoVE*, (136). doi: 10.3791/57320.

Sun, Y. *et al.* (2020) 'A comprehensive pan-cancer study of fibroblast growth factor receptor aberrations in Chinese cancer patients', *Annals of translational medicine*, 8(20), p. 1290.

Szybowska, P. *et al.* (2019) 'Cancer Mutations in FGFR2 Prevent a Negative Feedback Loop Mediated by the ERK1/2 Pathway', *Cells* , 8(6). doi: 10.3390/cells8060518.

Vogelstein, B. *et al.* (2013) 'Cancer genome landscapes', *Science*, 339(6127), pp. 1546–1558.

Weichert, W. *et al.* (2004) 'ALCAM/CD166 is overexpressed in colorectal carcinoma and correlates with shortened patient survival', *Journal of clinical pathology*, 57(11), pp. 1160–1164.

Weinstein, J. N. *et al.* (2013) 'The Cancer Genome Atlas Pan-Cancer analysis project', *Nature Genetics*, pp. 1113–1120. doi: 10.1038/ng.2764.

Yang, L. *et al.* (2019) 'An enhanced genetic model of colorectal cancer progression history', *Genome biology*, 20(1), p. 168.

Zhao, H. *et al.* (2014) 'CrossMap: a versatile tool for coordinate conversion between genome assemblies', *Bioinformatics*, 30(7), pp. 1006–1007.

Zhu, H. *et al.* (2020) 'Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks', *Molecular cell*, 77(6), pp. 1307–1321.e10.

Linda Vidarsdottir, *thesis*, Karolinska Institutet, Stockholm, Sweden (2017)