METASTATIC MUTATION PREDICTION FROM PRIMARY CANCER ANNOTATIONS IN COLORECTAL CANCER

A Thesis

Presented to the Faculty of the Weill Cornell Graduate School of
Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

by

Surya Vishnubhatt

May 2024

ABSTRACT

Metastatic cancer occurs when cancer cells break from the original tumor site, enter the bloodstream or lymphatic system, and spread to other areas of the body; it is considered the final stage of cancer progression. Previous studies have been able to link cancer mutation density to regulatory genomic annotations. This investigation leverages this knowledge, employing a Random Forest model to develop a predictive tool that can serve to explain metastatic cancer mutations, using annotations in primary cancer, which lends key insight towards identifying the cell type of origin of metastasized cancers (i.e. the site of metastasis). Currently, this project focuses on primary annotations of colorectal cancer. These annotations include chromatin accessibility (via ATAC-seq), histone modification (one dimensional H3K27ac), and HCT116 replication timing data. Current results suggest the model's potential utility in predicting the metastatic mutation distribution, however, further refinement and validation of the model are necessary to enhance accuracy, applicability, and generalizability.

BIOGRAPHICAL SKETCH

Surya Vishnubhatt is a Master's student in the Computational Biology Program at Weill Cornell Graduate School of Medical Sciences. He graduated from the University of California, Davis in 2022 with a B.S. in Biomedical Engineering. He is currently conducting his thesis research in the lab of Dr. Ekta Khurana at the Weill Cornell Medicine Sandra and Edward Meyer Cancer Center.

For my grandmother, Mamma.

To her, there was no higher pursuit than research.

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Ekta Khurana for her mentorship throughout the course of my thesis project. I would also like to thank my co-mentor Dr. Doron Betel and program director Dr. Trine Krogh-Madsen. Additionally, I would like to thank Susie Cheng and Dr. Alex Martinez Fundichely for their help in getting this project off the ground.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Metastatic, or Stage IV, cancer occurs when cancer cells break from the original tumor site, enter the bloodstream or lymphatic system, and spread to other areas of the body. The overall goal of this project is to develop a predictive tool that can predict mutational density in cancer metastasis based on primary cancer annotations, lending key insight towards identifying the cell type of origin of metastasized cancers [1]. The eventual results of this project may serve to aid in early detection and treatment of likely metastatic sites of a given cancer. Currently, this project focuses on colorectal cancer which is the third most common cancer worldwide and thus has a large sample size and availability of data for implementation. Additionally, according to a 2016 study published in Nature's Scientific Reports, approximately 56 percent of patients with colorectal cancer die from cancer metastasis with 20 percent of patients already having metastasis at diagnosis, making it an ideal candidate for this investigation [2].

The fundamental basis of this investigation comes from a 2015 Nature paper by Polak et al. which used a Random Forest regression model to provide a link between regulatory annotations in the genome and cancer mutation density wherein the cell type of origin of a cancer could be accurately determined by observing the distribution of mutations that occur along the genome. These researchers found that the best predictors for cancer mutation density were epigenomic features, namely chromatin accessibility, histone modifications, as well as replication timing [1].

For the purposes of this project, the chromatin accessibility data will be extracted through ATAC-seq (Assay for Transposase-Accessible Chromatin with sequencing). In ATAC-seq, genomic DNA is exposed to the transposase Tn5 which acts to fragment the DNA and then inserts into open chromatin sites wherein sequencing primers are then added. The DNA is then sequenced to identify areas where reads align densely (indicating regions where the Tn5 transposase was able to insert into the genome), these regions can then be visualized as peaks in the data, which are indicative of open chromatin [3]. The ATAC-seq data corresponding to primary colorectal cancer for this experiment has already been derived from the The Cancer Genome Atlas (TCGA) by Corces et al. in 2018 [4].

Additionally, one dimensional H3K27ac signal enrichment will be implemented to determine histone modifications. The data used is in HiChIP (Hi-C chromatin immunoprecipitation) form which combines Hi-C, a high throughput chromosome conformation capture technique, with the specificity of chromatin immunoprecipitation-sequencing (ChIP-seq) [5]. This allows for the identification of long-range chromatin interactions and regulatory landscapes associated with the histone modification H3K27ac. The data is then processed into FitHiChIP form which is a computational method to identify chromatin contacts that occur among regulatory regions from HiChIP data. This FitHiChIP data was further processed into one dimensional data by extracting all interacting bins (that also correspond to a ChIP-seq peak) and treating each individual bin as a peak.

Next, replication timing data will also be implemented from Gene Expression Omnibus (GEO) using the colorectal cancer cell line HCT116 [6]. Replication timing was found through high resolution Repli-Seq which maps newly synthesized DNA replication strands during each cell cycle phase throughout the whole genome to obtain genome-wide DNA replication timing. Finally, the metastatic genome information (to obtain somatic mutational density) will be retrieved through the Hartwig Medical Foundation's (HMF) metastatic cancer genome database which contains genomic information for patients with a primary colorectal tumor site whose cancer has since metastasized.

The aim of this project is to also use a Random Forest regression model which will take in primary cancer annotations for colorectal cancer and predict the resulting mutational density in metastasis. Random Forest regression is a supervised, non-parametric, ensemble machine learning method. The method involves the construction of multiple regression trees which are each based on a randomly selected subset of training data (with replacement). The remaining (out-of-bag) data is used as the test set for a given tree, this is implemented to calculate the mean squared prediction error of a tree. The final prediction can then be derived by averaging the resulting predictions from all trees wherein the observation was part of the out-of-bag data [1]. Thus, because Random Forest regression leverages the collective insight of many regression trees, it is more robust against overfitting and sensitive to outliers and noise, making it an ideal candidate for this investigation.

Essentially, the model will have to find associations between the primary cancer annotations (ATAC-seq, 1D H3K27ac, and replication timing data) and the mutational distribution associated with the various metastatic cancer genomes accessed through HMF. For the purposes of this investigation, the mutational density/distribution is a measure of the number of mutations found per one megabase (1-Mb) window along the genome. Figure 1 illustrates the overall workflow. The final, fully trained model should be able to take in a patient's primary cancer annotations and output the associated metastatic mutation distribution.

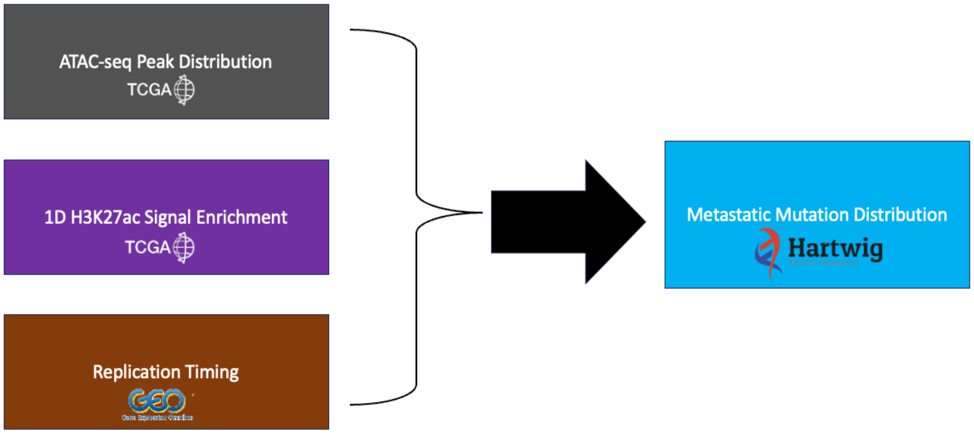

ATAC-seq Peak Distribution
TCGA
1D H3K27ac Signal Enrichment
TCGA
Replication Timing
GEO
Metastatic Mutation Distribution
Hartwig

***Figure 1**: Overall Project Workflow*

The current implementation of the project thus far implements a rudimentary Random Forest model. The datasets inputted into the current model were processed by dividing the genome into 1-Mb regions and calculating, across all respective samples, for each window: the average number of ATAC-seq peaks in primary colorectal samples, the peak intersection across all 1D H3K27ac patient files with primary colorectal cancer, the average HCT116 replication timing, and the average number of somatic metastatic mutations in all HMF samples with a primary colorectal site. The results do show promise given the decent Pearson correlation observed in the resulting Random Forest model, however further steps need to be taken to truly refine the methodologies implemented in model creation and execution.

METHODS

The goal of this project is to develop a predictive tool that can serve to explain metastatic cancer mutations using annotations in primary cancers. As previously mentioned, the primary cancer targeted is colorectal cancer. The aforementioned Polak et al. paper states that the most important epigenomic features needed to input into their model were chromatin accessibility, histone modification data, and replication timing. Additionally, the metastatic mutation distribution of cancers with a primary colorectal site needs to be obtained as well. Further data for analysis includes healthy (undiseased) colorectal ATAC-seq data and the mutational distribution of primary colorectal cancer.

In order to account for chromatin accessibility, ATAC-seq data is to be used. The ATAC-seq data implemented was derived from TCGA tumor samples by Corces et al. in 2018 [4]. It was done by profiling chromatin accessibility for 23 types of primary human cancers consisting of 410 high quality tumor samples derived from 404 donors. From these samples, 562,709 pan-cancer peaks were consistently observed across multiple samples (technical replicates). The colon adenocarcinoma ATAC-seq data consists of 38 patients and has 61,425 unique overlapping peaks and 488,617 total overlapping peaks across 38 patients. Figure 2 shows the peak overlap across all patients in 1-Mb windows along the genome for chromosome 2. A high overlap can be seen across all patient samples providing a good consensus going forward. The data was then further processed by calculating the average number of

ATAC-seq peaks that exist at a given chromosomal location across all samples, then taking the average of these values in overlapping 1-Mb windows across all samples.

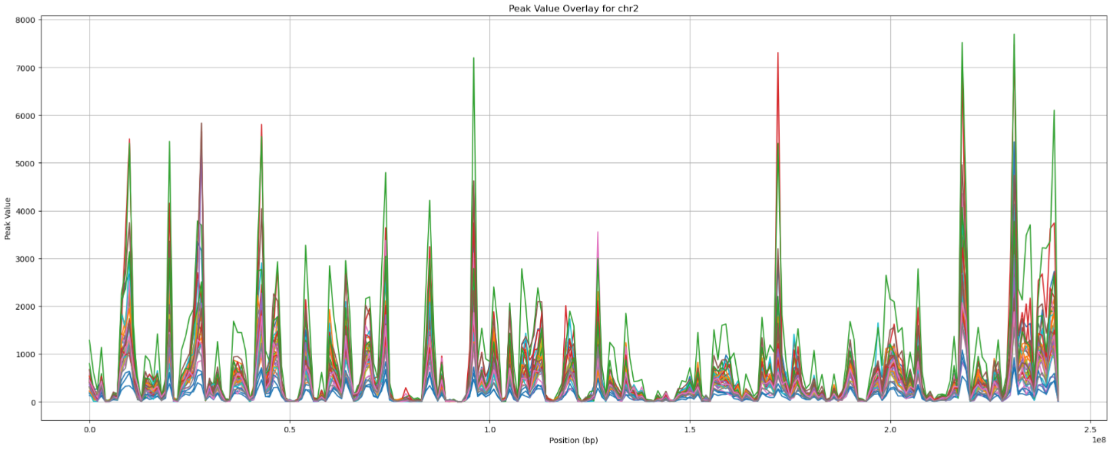


Figure 2*: Overlay of the number of ATAC-seq peaks across all 38 primary colorectal patient samples (each patient represented by a color) for chromosome 2 in 1-Mb windows. On the x-axis, we see the 1-Mb window occupied, on the y-axis, we see the number of peaks noted for a given window.*

Healthy (undiseased) colorectal patient ATAC-seq samples were found through ENCODE [7]. They consist of 4 sigmoid colon samples, 4 transverse colon samples, 2 mucosa of descending colon samples, 1 colonic mucosa sample, and 1 sample from the left colon. Figure 3 shows the number of peaks per 1-Mb window across all patients for chromosome 2. A good peak overlap can be seen across all patient samples. These 12 sample files were then processed by finding the average number of peaks per 1-Mb window across all samples. These peaks were Irreproducible Discovery Rate (IDR) thresholded to ensure good quality and reproducibility across replicates. These healthy colorectal patient samples are important to provide a control group with which to assess our findings.

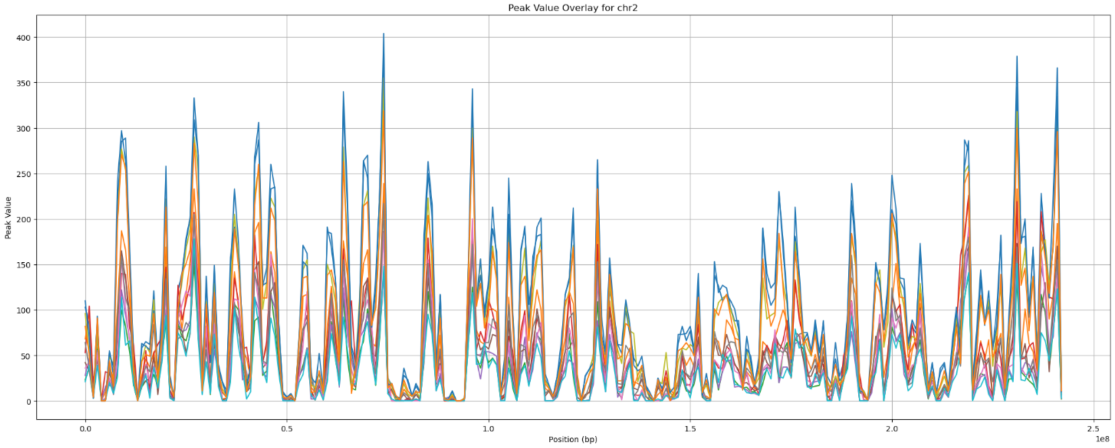


***Figure 3**: Number of ATAC-seq peaks per 1-Mb window across all healthy patients (each patient represented by a color) for chromosome 2. On the x-axis, we see the 1-Mb window occupied, on the y-axis, we see the number of peaks noted for a given window.*

To account for histone modification data, H3K27ac HiChIP data will be implemented. The data is processed to obtain a 1D H3K27ac signal enrichment by isolating peaks in processed primary colorectal adenocarcinoma TCGA FitHiChIP. There are four TCGA patient FitHiChIP files that are under investigation. The 1D H3K27ac signal enrichment is based on overlaps with ChIP-seq peaks for each patient and the peak intersection across all four files is used. Figure 4 shows a breakdown of the 1D H3K27ac data where, although patient TCGA-QL-A97D has a significantly higher number of peaks, Figure 5 shows that we can still see good peak overlap across all patient files. The 1D H3K27ac data is then processed by finding the number of peaks overlapping 1-Mb windows along the genome.

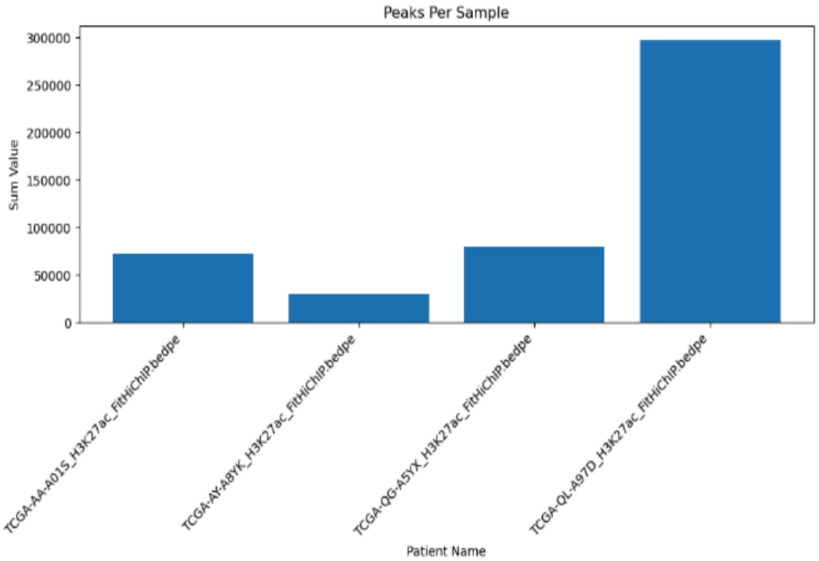


Figure 4*: The total number of 1D H3K27ac peaks found per primary colorectal cancer patient.*

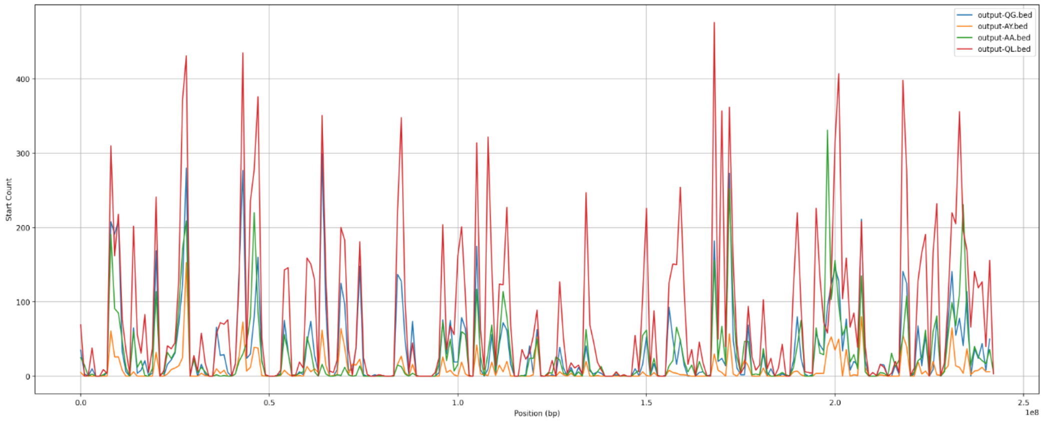
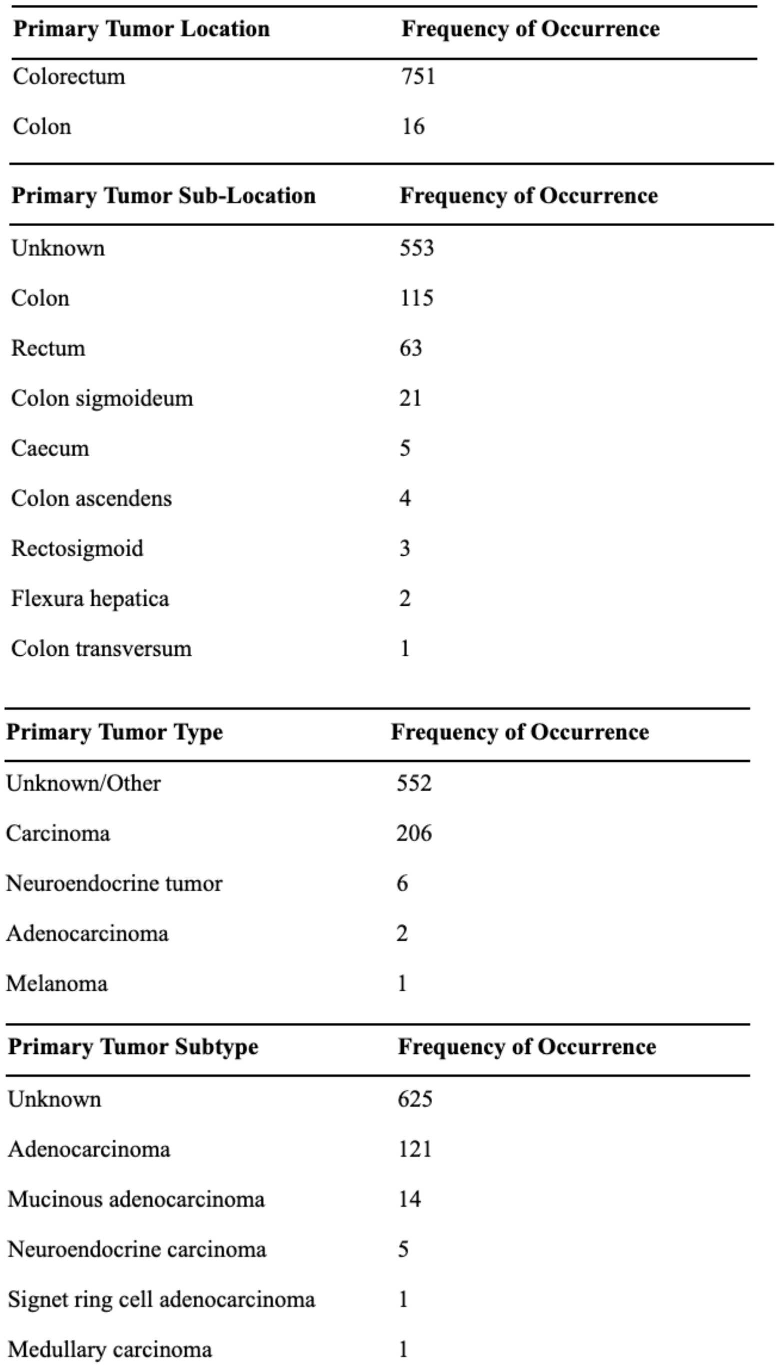


Figure 5*: Overlap of the number of 1D H3K27ac peaks across all 4 patients, indicating patterns of histone modification in 1-Mb windows along the genome. On the x-axis, we see the 1-Mb window occupied, on the y-axis, we see the number of peaks noted for a given window.*

The replication timing data has been retrieved from the HCT116 cell line through GEO from a Florida State University study which measured log2-transformed replication timing enrichment levels [6]. It was processed by also dividing the data into 1-Mb windows and calculating the mean replication timing for each window along the genome.

***Table 1**: Description of primary tumors being investigated for metastatic mutation distribution from the Hartwig Metastatic Mutation Database. For each feature, the frequency of occurrence corresponds to the number of files for which it is present across all 767 samples.*



Primary Tumor Location	Frequency of Occurrence
Colorectum	751
Colon	16

Primary Tumor Sub-Location	Frequency of Occurrence
Unknown	553
Colon	115
Rectum	63
Colon sigmoideum	21
Caecum	5
Colon ascendens	4
Rectosigmoid	3
Flexura hepatica	2
Colon transversum	1

Primary Tumor Type	Frequency of Occurrence
Unknown/Other	552
Carcinoma	206
Neuroendocrine tumor	6
Adenocarcinoma	2
Melanoma	1

Primary Tumor Subtype	Frequency of Occurrence
Unknown	625
Adenocarcinoma	121
Mucinous adenocarcinoma	14
Neuroendocrine carcinoma	5
Signet ring cell adenocarcinoma	1
Medullary carcinoma	1

To obtain a metastatic mutation distribution, the Hartwig metastatic mutation database was used to identify all patients with primary colorectal cancer. This resulted in a total of 728 patients with 767 samples. Table 1 provides a description of the characteristics of the primary tumor sites being investigated for metastatic mutation data across all samples. The primary mutation type of interest are somatic single nucleotide variants (SNVs) given that they are an established characteristic of cancer genomes [8]. These were processed by first finding the average number of mutations for 1-Mb windows along the genome across all samples for a given patient. After repeating this for all patients, another average is then computed across patients, finding the average number of mutations per 1-Mb window per patient along the genome. This provides a way to obtain an overall landscape of the metastatic mutation distribution across patients.

Finally, the primary mutational SNV information for colorectal cancer was also implemented. This was done by using data from the Pan-Cancer Analysis of Whole Genomes (PCAWG). The results are stored as a list of SNVs by chromosomal position and were processed by simply counting the number of mutations in 1-Mb windows along the genome. Primary cancer mutations are of importance in order to assess our results, providing a baseline with which to perform comparison.

RESULTS

Spearman's Rank Correlation Coefficient Analyses

After all files are processed into 1-Mb windows, we can begin analysis of the datasets across matching 1-Mb windows along the genome. First, to assess the relationship between the primary colorectal cancer chromatin accessibility (ATAC-seq data) and average metastatic mutation distribution, the Spearman's rank correlation coefficient can be calculated. For this relationship, the correlation was found to be -0.64. This is consistent with the findings in Polak et al. which found that features that indicate regions of active chromatin were associated with low mutation density, and vice versa, those with repressed chromatin features were more closely related with regions of high mutation density [1]. The correlation coefficient found for the relationship between primary chromatin accessibility and the primary cancer mutation distribution was found to be -0.61. This correlation is slightly worse than the comparison to the metastatic mutation distribution but is still quite comparable.

Comparing primary colorectal chromatin accessibility and 1D H3K27ac data resulted in a correlation coefficient of 0.73. This serves as a good indication of the quality of both data sets as we expected to see a high, positive correlation between regions of modified histones and open chromatin regions as accessible chromatin regions are likely bound by regulatory proteins to regulate gene expression. Going further, we can also compare the correlation between the primary chromatin accessibility and replication timing. This was found to be a value of 0.77, this high, positive correlation has a biological basis, indicating the relationship between less condensed, accessible chromatin regions and early replication timing [9].

Looking at the healthy colorectal ATAC-seq data, similar analyses can be implemented. The correlation coefficient between the healthy chromatin accessibility and metastatic mutations was found to be -0.60. The correlation between the same healthy chromatin accessibility and the primary mutations were found to be quite similar at -0.54. Investigating further, we can take the correlation between the primary cancer mutations and the metastatic cancer mutations. This correlation was found to be very high at 0.90.

Preliminary Random Forest Model

Using the primary cancer annotations and metastatic mutation distribution, a Random Forest regression model can be made. The current implementation is bare-bones, just taking in the previously calculated primary colorectal ATAC-seq data, replication timing, and 1D H3K27ac. The data is kept in the same format (1-Mb windows) and inputted into a combined dataframe. The current model uses 1000 trees and splits the data into 90% training and 10% testing. It was also trained using a linear regression model as well, for baseline comparison.

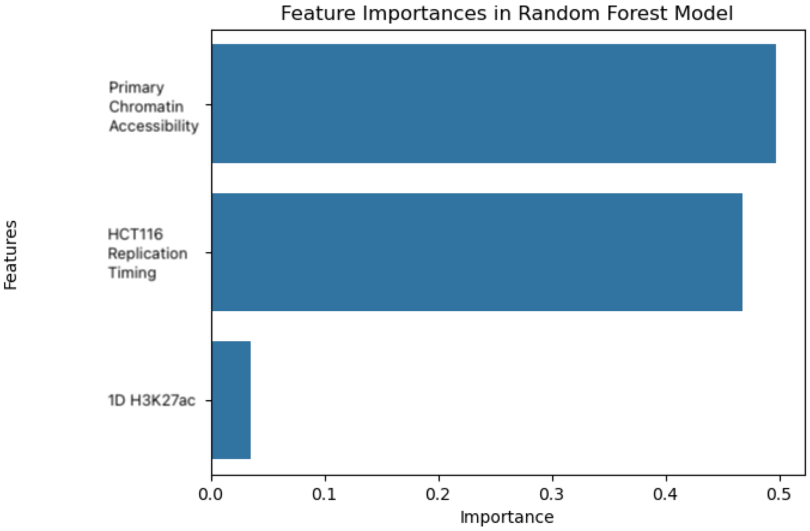


Figure 6*: Random Forest regression model feature importances where we can see that primary chromatin accessibility is ranked the highest at 0.497 with HCT116 replication timing comprising 0.467 and the 1D H3K27ac is ranked the lowest at 0.035.*

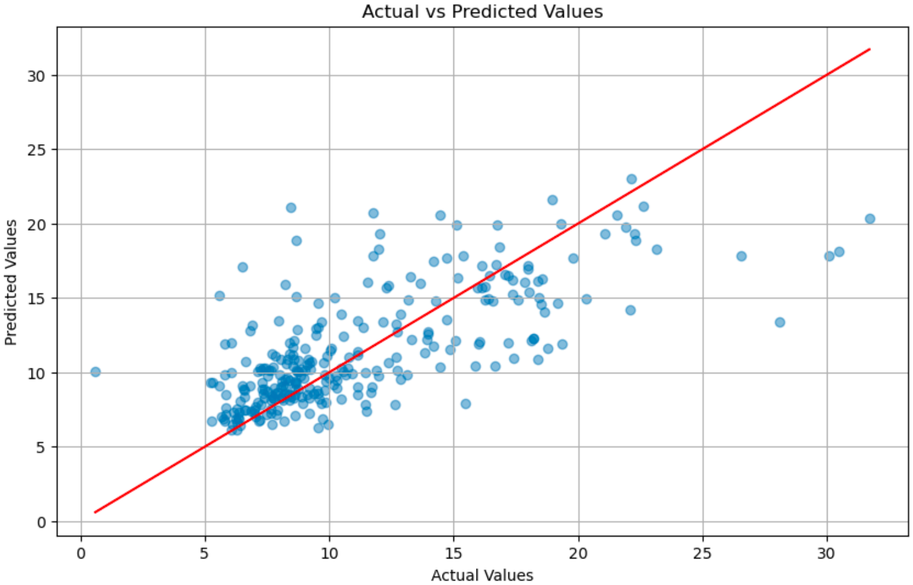


Figure 7*: Model results showing actual versus predicted metastatic mutation density (i.e. the number of mutations per 1-Mb window for a given region along the genome)*

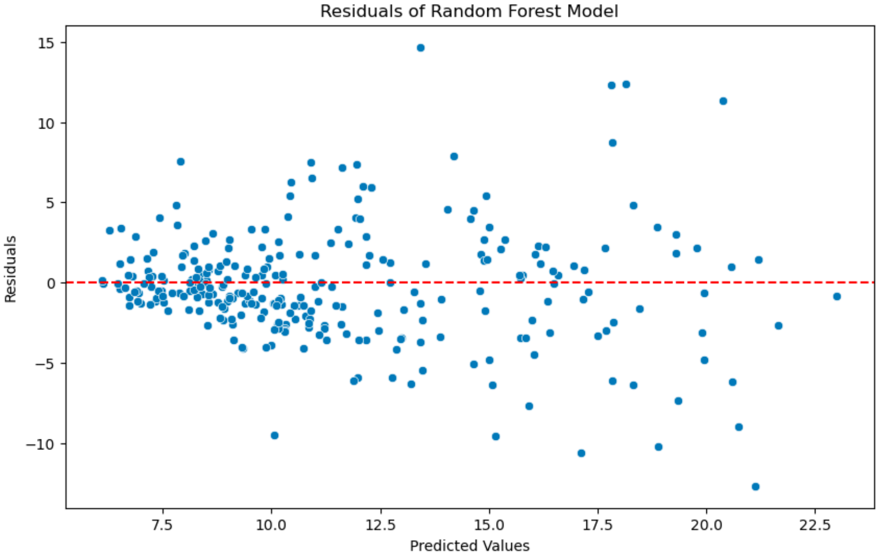


***Figure 8**: Plot of residuals from the model showing the difference between the predicted and actual metastatic mutation density.*

The results for the Random Forest model show a Pearson correlation of 0.71 which indicates a strong positive correlation between the actual and predicted mutational values. Figure 6 shows the feature importances where chromatin accessibility and replication timing are the dominant features. Figure 7 shows a plot of the actual versus predicted mutational values and Figure 8 shows the residuals. The MSE of the model was found to be 12.34 and the R^2 was 0.495, which is quite low for explaining variance. The random forest, as expected, outperformed the linear regression model which had a lower Pearson correlation of 0.56, a lower R^2 (0.32), and higher MSE (16.54).

CONCLUSIONS

Discussion

From the results, we can observe that primary colorectal chromatin accessibility has a stronger correlation to the distribution of metastatic mutations when compared to the healthy colorectal chromatin accessibility, however the values differ by only -0.04. Although, when inputting the healthy colorectal chromatin accessibility, instead of the primary colorectal chromatin accessibility, to the Random Forest model seen above, a significant drop in the feature importance of chromatin accessibility was noted (dropping to 0.30) and HCT116 replication timing had a larger feature importance to compensate (0.66), and the resulting Pearson Correlation was dropped to 0.66. This shows that primary colorectal chromatin accessibility is a better indicator of metastatic mutational distribution than the healthy patient chromatin accessibility.

Another observation from the results can be seen in the large overlap between the primary and metastatic mutation distributions. This means that these mutational distributions are quite similar and that metastatic mutations retain a lot of mutational characteristics seen at the primary level. This serves as an explanation for the similarities seen in comparing both primary colorectal and healthy colorectal chromatin accessibility to the primary and metastatic mutational distributions, respectively.

Overall, the current results indicate the model's potential in predicting the metastatic mutation distribution. However, further refinement and validation of the model is essential in overcoming its current shortcomings.

Future Direction

Observing the results of the above Random Forest, we can see that the highest feature importance is chromatin accessibility followed by replication timing and lastly the 1D H3K27ac. The very low feature importance of 1D H3K27ac is likely due to the way it was implemented. Going forward, this data needs to be more closely analyzed and, instead of using the number of peaks per 1-Mb window, the signal intensity should be used. Additionally, ChIP-Seq should also be incorporated to validate the data as well.

The data preparation procedure for the current investigation does not incorporate any normalization procedure across different samples, nor does it filter any of the datasets for outliers. Going forward, both normalization and quality assessments of the samples should be implemented to obtain clearer and more representative results. Additionally, the current data preparation and model development only implements 3 features, however, many more features will need to be incorporated in the final result, in addition to more features for chromatin accessibility, histone modification, and replication timing, there will also need to be features to account for sequence and expression.

Finally, the Hartwig metastatic database has information regarding patient/sample treatment responses as well as pre- and post-biopsy drugs. Going forward, these should be implemented to better understand the role that treatment plays within the context of metastasis.

REFERENCES

1. P. Polak *et al.*, “Cell-of-origin chromatin organization shapes the mutational landscape of cancer,” *Nature*, vol. 518, no. 7539, pp. 360–364, Feb. 2015. doi:10.1038/nature14221
2. M. Riihimäki, A. Hemminki, J. Sundquist, and K. Hemminki, “Patterns of metastasis in colon and rectal cancer,” *Scientific Reports*, vol. 6, no. 1, Jul. 2016. doi:10.1038/srep29765
3. J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, “ATAC-Seq: A method for assaying chromatin accessibility genome-wide,” *Current Protocols in Molecular Biology*, vol. 109, no. 1, Jan. 2015. doi:10.1002/0471142727.mb2129s109
4. M. R. Corces *et al.*, “The chromatin accessibility landscape of primary human cancers,” *Science*, vol. 362, no. 6413, Oct. 2018. doi:10.1126/science.aav1898
5. S. Bhattacharyya, V. Chandra, P. Vijayanand, and F. Ay, “Identification of significant chromatin contacts from hichip data by FitHiChIP,” *Nature Communications*, vol. 10, no. 1, Sep. 2019. doi:10.1038/s41467-019-11950-y
6. “GEO Accession Viewer,” National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4875400 (accessed Feb. 16, 2024).
7. “Experiment matrix,” ENCODE, https://www.encodeproject.org/matrix/?type=Experiment&control_type%21=

%2A&status=released&perturbed=false&assay_title=ATAC-seq&biosample_ontology.organ_slims=colon (accessed Feb. 17, 2024).

8. L. Liu, S. De, and F. Michor, “DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes,” *Nature Communications*, vol. 4, no. 1, Feb. 2013. doi:10.1038/ncomms2502

9. H. Fu, A. Baris, and M. I. Aladjem, “Replication timing and nuclear structure,” *Current Opinion in Cell Biology*, vol. 52, pp. 43–50, Jun. 2018. doi:10.1016/j.ceb.2018.01.004