Use of Somatic Mutations for Classification of Endometrial Carcinomas with CpG Island Methylator Phenotype

A thesis presented to

the faculty of

the Russ College of Engineering and Technology of Ohio University

In partial fulfillment

of the requirements for the degree

Master of Science

Jonathan R. Feige

April 2022

This thesis titled

Use of Somatic Mutations for Classification of Endometrial Carcinomas with CpG Island Methylator Phenotype

by

JONATHAN R. FEIGE

has been approved for

the School of Electrical Engineering and Computer Science

and the Russ College of Engineering and Technology by

Lonnie R. Welch

Professor of Electrical Engineering and Computer Science

Mei Wei

Dean, Russ College of Engineering and Technology

Abstract

FEIGE, JONATHAN, M.S., April 2022, Computer Science

Use of Somatic Mutations for Classification of Endometrial Carcinomas with CpG Island Methylator Phenotype

Director of Thesis: Lonnie Welch

Endometrial carcinoma begins in the cells within the inner lining of the uterus that, like many other cancers, grows out of control. A subset of these tumors shows genome wide hypermethylation. Hypermethylation results in down-regulation of tumor suppressor genes resulting in CpG Island Methylator Phenotype (CIMP) tumors. Individuals with this hypermethylation are classified as CIMP+ and have increase in cancer reproduction and growth. We have hypothesized that by using CIMP related samples and the mutations associated with them, we can classify with high accuracy, an unknown sample using only mutational data. Using machine learning, we found that it is possible to correctly classify unknown CIMP samples with 90% accuracies just using the somatic mutations within each sample. This breakthrough will be used for diagnostics and treatment of endometrial cancers.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Chapter 1: Introduction and Background

The CpG island methylator phenotype (CIMP) is a phenotype in cancer which is caused by hypermethylation at CpG islands. These CpG islands can function as a driver for silencing of tumor suppressing genes like PTEN [1]. Individuals with the CIMP+ phenotype in endometrial carcinoma are much more resistant to chemotherapies compared to the individuals that are CIMP- [2]. When individuals develop cancer a full methylation profiling is not commonly taken, the mutations within the cells are much more commonly collected and analyzed. By analyzing the mutations that occur within CIMP samples we can better understand the biology of the tumors when the targeted chemotherapies fail, a better understanding of the CIMP phenotype, provide potential mutations as divers for CIMP, and provide potentially new treatments and diagnostics for CIMP+ individuals.

To solve this problem, we developed a pipeline for analyzing the samples and mutations from The Cancer Genome Atlas (TCGA) as they provided data with the samples with the CIMP phenotypes and mutation data. We aim to evaluate the hypothesis that we can use mutations for Classification of Endometrial Carcinomas with CIMP phenotype. First, we will choose statistical ways to select the most prevalent and CIMP+ related samples. Then with these mutation groups we will use four unique supervised machine learning techniques to evaluate the connection between CIMP and the mutations in a sample. Once mutations are selected, the mutations will then be used for correlation analysis and for biological interpretation. The research was broken down into three major aims. Aim one is to discover the most CIMP related mutations. Aim two is to use the

mutations discovered in aim one to preform association and correlation analysis. Aim three is to use the mutations from aim one and interpret the biological significance of the mutations.

This study will increase our understanding of the relationship between CIMP samples and the mutations that occur in them. By using a novel approach to CIMP samples and mutations we will develop a pipeline that uncovers the relationship between CIMP and mutations. This will then help the scientific community to develop new hypothesis for further experimentation and deepen the understanding of CIMP in cancer.

1.1: DNA Methylation

All cancers have shown to be extraordinarily complex diseases. Endometrial carcinoma is no exception. One of these significant changed that can occur in the genome is DNA methylation. DNA methylation occurs when a methyl group (CH3) attaches to an adenine or a cytosine. This can be from via three unique DNA methyltransferases DNMT1, DNMT3a and DNMT3b which are all enzymes for the regulation of methylation in the genome [3]. DNMT1 controls the maintenance of the DNA methylation, maintains the DNA from parent strand to newly generated strand and assists in binding to the hemi-methylated strands [3]. DNMT3a and DNMT3b do not drive methylation and they do not need the hemi-methylated DNA. DNMT3a and DNMT3b do add additional methyl groups to locations across the DNA [3]. When these DNMT's No longer work properly hypermethylation can occur.

The most common methylation in humans is a methyl group -CH3 bonded to a cytosine on the DNA [4]. DNA methylation can regulate gene expression by inhibiting the binding to and from transcription factors (proteins that transcribe DNA) in the DNA [3]. DNA methylation is found in normal tissues along with cancer cells in the control of transcription factors across the DNA. A key difference in cancer cells is that this process can quickly grow out of control in either failing to remove the existing methyl groups or rapidly adding methyl groups to DNA loci that would not need it [4]. In the situation that this occurs in a tumor suppression gene like MLH1, the cancer cell loses its ability to apoptosis, this leading to widespread cancer within an individual. Having these increased

levels of methylation is dangerous to individuals with cancer due to how rapidly the cancer progress.

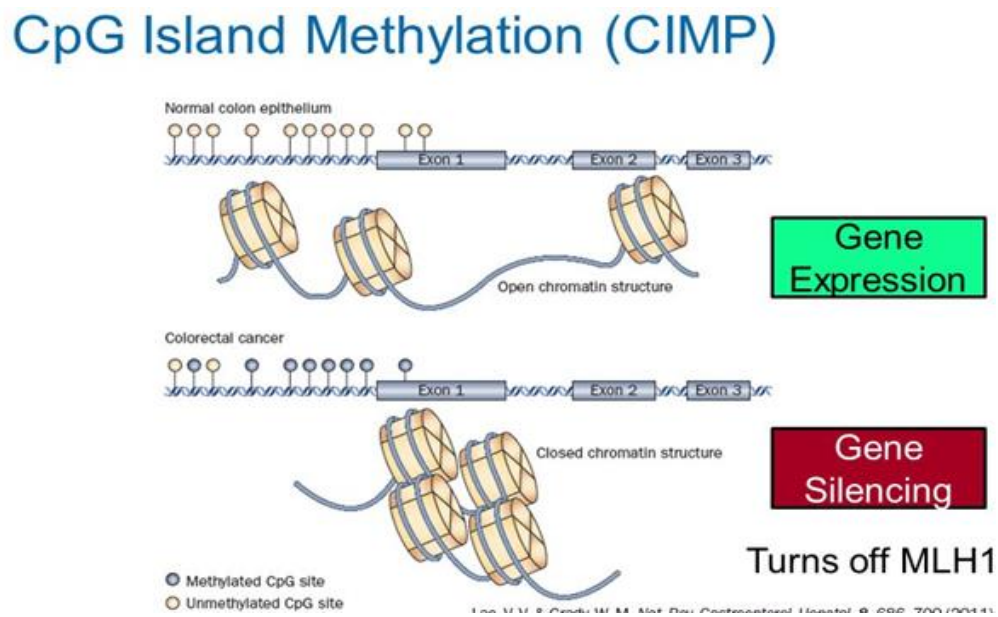
1.2: CpG Island Methylator Phenotype

A cytosine – phosphate – guanine (CpG) is a pattern of base pairs that occur across the genome. In order to qualify as a CpG island the DNA CpG chain must be greater than two hundred base pairs in length and the C and G bases must count for at least 50% of total base pairs [5]. Sixty percent of mammalian promoters have unmethylated CpG islands [5]. CpG islands have three major functions. The first function is cell type expression, suppression of testis specific genes and the control of imprinted genes [5]. These CpG's also tend to be outside the gene promoter due to their rule closely relating to gene suppression and silencing. The CpG islands are a common point for mutation due to methylation through deamination (A removal of an amino group). Methylation is a common occurrence due to the DNMT enzymes targeting the cytosine groupings within the CpG [5].

In the DNA there can be large changes to the 5-methylcytosine (5mC) that can have major alterations in the progression of cancer cells [6]. The majority of the 5mC are located on CG nucleotide pair which are denoted CpG's [6]. If 5mC upregulates or downregulates a gene upstream of a gene promoter in cancer, it is considered a CpG island methylator phenotype (CIMP) location [7]. In normal samples the DNA has an open structure so it can be read, with the added CIMP the DNA becomes tightly coiled and is no longer able to be read.

Figure 1

CpG Island Methylation Phenotype



Note: Healthy DNA versus hypermethylated DNA on the gene MLH1. The diagram shows the effect of CIMP on DNA.

Source: [7]

As shown in the example above, the gene MLH1, which is a DNA repair gene, is silenced due to CIMP. Each yellow and blue dot represents a point on the DNA that there is a CpG site. The graphic demonstrates how the CpG itself is not harmful but the methylation that occurs at a CpG site can cause various problems within the cell. The blue Markers represent a methylated CpG island and, as seen on the bottom half of the graphic, causes exon (a location that encodes an amino acid) silencing. Due to the methylation of the CpG islands, the DNA is tightly coiled and no longer able to be read by the RNA polymerase causing gene silencing. Gene silencing in MLH1 causes the cancer cell to no longer repair mutations, leading to the cell to mutate out of control [7].

CIMP was originally identified in colorectal cancer, but CIMP has since been identified in multiple cancer types including, but not limited to, gastric cancer and endometrial cancers [7]. CIMP has been shown to be a non-random occurrence in pan-cancers with a major relationship between CIMP and cancer growth [8]. CIMP tends to arise early in the tumor development process leading to much more aggressive tumors over the span of the lifetime. Identifying the CIMP phenotype for a sample takes substantial amounts of methylation profiling. It has been shown that using DNA probes across the genome can determine the hypermethylation level across the tumor [7]. Researchers used the HumanMethylation450 bead chip to determine methylation status across each samples genome [1]. In uterine carcinoma there were 1430 probes across the genome that were used to classify samples [1]. Each probe collects a methylation value, which is then converted to a beta value. A beta value is produced by the Illumina methylation assay and shows how high in methylation a probe is between 0 (low) and 1 (high). These beta values are used in a K-Means Clustering where *k* (The number of clusters in the system) is equal to 3 representing positive, negative, and intermediate [7].

K-means clustering is a supervised machine learning technique used to group like samples into *k* groups. In this study the k groups were substantial amounts of methylation across the genome, mild methylation across the genome, and low to no methylation across the genome. Each sample is then grouped into the category it fits best based on the samples already in each beta grouping. The highest valued beta grouping is considered CIMP positive (CIMP+), the smallest beta grouping is considered CIMP negative

(CIMP-) , lastly, the ambiguous grouping in the middle was coined CIMP Intermediate (CIMPi) [9].

1.3: Research Hypothesis and Aims

The goal is to use mutation data to accurately classify CIMP status in endometrial carcinoma samples by using random forest, support vector machines, multilayer perceptron's, and K-nearest neighbor supervised machine learning algorithms.

Understanding how the CIMP subtype is related to mutations, which may lead to new treatment techniques. The results come from three targeted aims. The first is to find the relevant mutations that are strongly related to CIMP via supervised machine learning. We expect prominent levels of accuracy using mutations to classify unknown CIMP samples. Once the mutations are found, aim two looks at the correlation / causation between the mutations and attempts to uncover the relationship between cancer samples, CIMP, and the significant mutations. The goal would be to build a deeper understanding of the interrelationships between mutations and the CIMP phenotype. The last analysis analyses the biological significance of the mutations selected by analyzing signaling pathways and large scale effects on the cell due to mutations. When an individual has cancer the first thing done is not typically a 450 methylation probing to determine CIMP status, but an analysis of the mutations always occurs [1]. Using mutations to classify CIMP would help give new insights into CIMP's effect on cancer cells and how CIMP alters the progression in these cells. This may also lead to new diagnostic techniques for individuals with the CIMP+ phenotype.

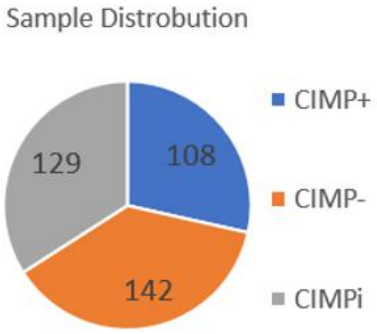
The data was collected from the Cancer Genome Atlas, which is the most comprehensive cancer genomics study to date. The data used consists of 250 unique samples. 108 samples are CIMP+ while the other 142 samples are CIMP-. The CIMPi

group is removed to make each classification problem a clear two group binary classification problem, being that the CIMPi group tends to be a muddy area between the two groups. There are 8085 total mutations. Of the 8085 mutations, 739 demonstrated a strong correlation to either the CIMP+ or the CIMP- grouping.

Figure 2

Data and CIMP distribution

	GOT1_GR(	TEX36_GR	KIAA1217_
TCGA-A5-A	1	1	1
TCGA-A5-A	0	0	0
TCGA-A5-A	0	0	0
TCGA-A5-A	0	0	1



Note: This is an example of the data for mutations GOT1, TEX36, and KIAA1217 and their relationship to the first four samples. On the right is the sample distribution of the CIMP+ vs CIMP- samples.

The **first aim is to find the important CIMP related mutation**. We will select mutations using supervised machine learning classification. Using machine learning we are able to show connections between mutations and CIMP with high confidence

Using the strongly correlated CIMP mutations, the **second aim is to find the relationships between the data features**. This takes the form of association rule mining. The goal will be to build strong correlations between any two mutations. The scope is also extended to finding the relationships between gene-level mutations within a single cancer type.

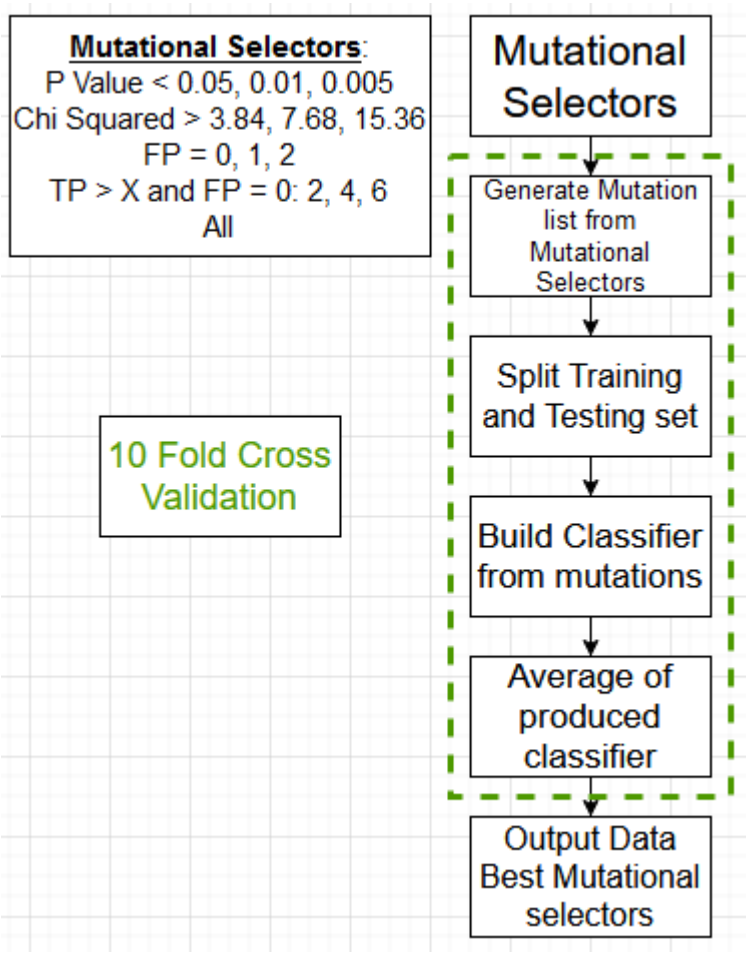
Using the strongly correlated CIMP mutations, the **third aim is to interpret the findings from a biological perspective**. This would take form in various biological analyses to determine a link between biological pathways, cancer, and CIMP.

Chapter 2: Aim 1 – Finding Important CIMP Mutations

The goal is to classify CIMP in cancer using only mutations. Each step expresses small parts of the pipeline in order to reach the classification models used. The goal of the first aim is to use statistical methods in order to find the most influential mutations that exist within the samples for the CIMP+ and CIMP- phenotype. It is important to find these strong mutations in order to have a strong statistical significance in the following two aims.

Figure 3

Pipeline Diagram



Note: This is the broad flowchart of the pipeline used for the classification model.

The approach for this chapter is to identify relevant mutations in relationship to CIMP. The pipeline in the figure above shows the outline of the code used in order to produce results. First the mutational selectors must be established. Then using the mutational selectors, we can choose the mutations we want to use in the classification problem. This will be achieved first by analyzing the raw statistics of a mutation, then evaluating each mutational grouping (mutational selector) in a classification model.

2.1: Mutation Selectors and Classification

The most low-level statistical concept used was the confusion matrix. The confusion matrices used have the standard build of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). As an example, from the data "RPL22_GRCh38_1:6197725-6197725_Frame-Shift-Del_DEL_T-T—" has TP = 27, FP = 10, FN = 81, and TN = 261. From a biological perspective, this actively demonstrates that the mutation RPL22_GRCh38_1:6197725-6197725_Frame-Shift-Del_DEL_T-T— exists in 27 CIMP+ samples (TP) and 10 Non-CIMP+ samples (FP), It does not exist in 81 CIMP+ samples (FN) and does not exist in 261 CIMP- samples (TN). These base level statistics are used in all mutational selector

2.1.1: Mutational Selectors

After analyzing at the base level statistics, we began the development of our mutational selectors. These mutational selectors are metrics that determine if a mutation will or will not be used in the classification problem. The selectors chosen were:

- Fisher's Exact P-Value < 0.05, 0.01, 0.005.
- The Chi-Squared Significance Test > 3.84, 7.68, 15.36.
- FP <= 0, 1, 2
- FP = 0, TP > 2, 3, 4 (TPX_FP0).
- All mutations

Stepping through each of these selectors, Fisher's Exact P-Value is a test that analyzes the significance of a confusion matrix. The main focus of the test is to show how likely

the data is by random chance based on the values of the confusion matrix. The value is most commonly calculated by:

$$p = \frac{(TP + FN)! * (FP + TN)! * (TP + FP)! * (FN + TN)!}{TP! * FN! * \ FP! * TN! * total!}$$

Typically, a significant value is less than 0.05. To ensure this is not too tolerant we also chose mutations that were less than 0.01 and 0.005. Using RPL22 as an example.

$$p = \frac{(27 + 81)! * (10 + 261)! * (27 + 10)! * (81 + 261)!}{10! * 81! * \ 10! * 261! * 379!}$$

We find the p-value for RPL22 to be 3.86 * 10^-9. This falls far under the 0.05 threshold typically used by this statistic.

The Chi-Squared Significance Test is used to determine how unique the data is based on the difference from the expected mean.

$$CHI = \frac{(TP - \overline{TP})^2}{\overline{TP}} + \frac{(FN - \overline{FN})^2}{\overline{FN}} + \frac{(FP - \overline{FP})^2}{\overline{FP}} + \frac{(TN - \overline{TN})^2}{\overline{TN}}$$

TP, FN, FP, and TN represent the values for a single mutation, while $\overline{TP},\ \overline{FN}, \overline{FP},$ $\overline{TN}$ represent the mean of all values for that static. Again, with RPL22 as the example mutations

$$CHI = \frac{(27 - 1.421)^2}{1.421} + \frac{(10 - 1.988)^2}{1.988} + \frac{(81 - 106.579)^2}{106.579} + \frac{(261 - 269.011)^2}{269.011}$$

We find that RPL22 has a CHI squared value of 499.0767, which is well over the 3.84 threshold of significance. CHI is a reliable test that checks if the entire confusion matrix of a mutation is significant. Once a value is calculated, each value is checked against the Chi-squared distribution table. A chosen p-value is on the x-axis and the degrees of freedom are on the y. The p-value is chosen by the users (in this case the

standard 0.05), while the degrees of freedom must be calculated. It can be calculated by the (rows in confusion matrix – 1) * (column in confusion matrix -1). In this case all values will be (2 -1) * (2 – 1) = 1. With a p-value 0f 0.05 and degree of freedom at 1, the table shows that all CHI greater than 3.84 are considered significant. Once again, to have a more thorough exploration of the mutational space, we had two more increased thresholds, each double the previous one. These thresholds are CHI greater than 7.68 and CHI greater than 15.36.

The next selector analysis mutations that exist in little to no no-CIMP+ samples. FP = 0 indicates that the mutation occurs in exactly 0 non-CIMP+ samples making this an extremely strict threshold for classification. The other two options are less strict being FP <= 1, allowing mutations that occur in 1 non-CIMP+ sample to be included and FP <= 2, allowing for mutations that occur in two or less non-CIMP+ samples. Removing the background class has many potential benefits in classification problems. Classifying just the foreground tends to lead to higher accuracies for multiclass classification models.

The strictest selector is the FP = 0, TP > 2. This includes only mutations that exist strictly in the CIMP+ phenotype. This is an important separator because of its pseudo-unary classification system. With only CIMP+ mutations in the dataset, the classification model should have an easier time with classifications. Since in the data preprocessing all mutations that existed in less than two samples were removed. FP = 0, TP > 2 is a redundant statement since TP > 2 would be implied, because of this the results FP = 0, TP > 2 is referred to just as FP = 0 due to the initialization step the removed mutations that

occur in less than 2 CIMP+ samples. There are also more strict options such as FP = 0, TP > 3, and FP = 0, TP > 4

The last mutational selector is labeled “All mutations.” While including all mutations does not seem like a ‘selection,’ it functions as a baseline statistic to compare to other mutational selectors.

2.1.2: Building Classifiers for Mutations

We explored four different classification methods. The four used are Random Forest, Support Vector Machines, K-Nearest Neighbors, and Multilayer Perceptron. Each of these classification models came from the python library SciKitLearn.

After the mutation selection the next step is cross-validation. Cross-validation aims to not over-fit or overgeneralize the classification model for new unknown samples. In this research, k-fold cross-validation was used. Three, five and ten fold cross validation were all evaluated, but ten folds was chosen as the k value due to its ability to have a larger sampling of testing data. With this cross-validation, the dataset is split into two sets, the training set, and the testing set. The sets are split is for each fold of the Random Forest, 9/10 of the data goes into the training set, which is passed to the Random Forest for classification, the other 1/10 is used to evaluate the Random Forest as an accuracy threshold. The samples divided in each group are randomly selected with each fold. Once completed, the 10-fold are averaged to give a single value for the overall performance of the Random Forest. The cross-validation step functions as a for loop around a classifier. Using Sklearn train_test_split the data can be divided into the random training and testing set.

The next step is the tuning of hyperparameters. In any classification model, there are a large number of parameters that can be used to better the results. In hyperparameter tuning, a collection of parameters is evaluated in order to find the best results. We chose the most important hyperparameters and enumerated various combinations of parameters. Each combination of parameters is used and assessed with SciKitLearn GridSearchCV. The classifier with the highest accuracy is chosen from the battery of tests each time. Once finished, the best parameter combination, based on accuracy, is selected as for that fold by using the best_estimator_ function from GridSearchCV.

The best results produced came from the random forest classifier. When given the sample and mutation data, the forest builds a collection of decision trees based on the mutations. With the trees, the samples are classified as CIMP+ or CIMP- based on the decision from the majority of trees. Similar to before, a confusion matrix is built on the correctness of the forest. The random forest was built from SciKitLearn RandomForestClassifier and the list of parameters can be found in appendix B.1. For most future results, the graphics will be based on the random forest statistics due to it having the best performance and the added interpretability due to each mutation being assigned an importance (Gini index) to the classification model. This showing what were the driving mutations that occurred in each iteration of the random forest. The process that was used to create the random forest was repeated three more times on three other classification models.

Multilayer Perceptron is a type of neural network [10]. It consists of three layers: the input layer, output layer, and hidden layer. The input layer receives the input. The

prediction of samples and classification are performed by the output layer. Between the input and output layers, an arbitrary number of hidden layers is placed between the input layer and output layer. This is where all computation is done for the MLP. The major change in the pipeline is hyperparameter tuning. The multilayer perceptron was built from SciKitLearn MLPClassifier and the list of parameters can be found in appendix B.2.

K-Nearest Neighbors classifier is a pseudo-k means clustering approach where samples are grouped by the number of mutations they have in common. The K-Nearest Neighbors algorithm assumes that similar things are near each other. The classifier randomly selects a sample that has already been classified, then takes the K closest (by mutational overlap) and classifies the K samples as the same class as the base. This machine assumes proximity in the data therefor it tends to underperform compared to other classifiers for this dataset due to the sparsity in the data. The K-Nearest Neighbors was built from SciKitLearn KNeighborsClassifier and the only change is the hyperparameter tuning. The list of parameters can be found in appendix B.3.

Support Vector Machines are supervised learning models that uses fit vectors for classification. The objective of the Support Vector Machine algorithm is to find a vector that distinctly classifies the data points [11]. It is done by repetitive vector placements until a best fit is found. In this dataset, the number of features is two, CIMP+ and CIMP-. The support vector machine was built from SciKitLearn svm and the list of parameters can be found in appendix B.4.

2.1.3: Classification Mutation Selection

Once all classifiers have produced results, we selected the feature that had the highest impact on the classifier. This comes in a variety of forms. For the Random Forest classifier, each mutation in the classifier is given a "gini index". The gini index provides a percentage importance of each mutation for each classification. This means the score is relative to the number of mutations in the mutational selector. The K-Nearest Neighbors selects mutations based on the connections made in the neighborhoods within the classification model. The Multilayer Perceptron selects features by selecting the network that was generated which produced the best performance. Once a network is selected, features can be extracted from each of the nodes in the network. Lastly, the Support Vector Machine classifier selects features by choosing the features that are constantly on the same side of the hyperline over numerous iterations, this indicates a feature that is strongly linked to one of the two classes.

2.2: Results

In this data processing and classification stage, there were a large number of mutations found to be incredibly significant in numerous ways for our statistical analysis. Using the classification models, we found accuracies of up to 90% using only the mutations in the TCGA endometrial carcinoma data. The results in this section support the claim that mutations and CIMP are connected and with just mutations we can accurately classify unknown CIMP samples.

2.2.1: Mutational Selectors

The mutational selectors are the basis for the classification models. Without these stricter guidelines all of the following results would fall under the 'All' row, and as seen in the following tables, this tends not to perform well compared to the other selectors (Up to a 20% increase in accuracy). Having these selectors also brings new insight into how the classification models react to different datasets with differing characteristics. Another aspect to consider when looking at the following tables is that the FP=0 classifier always, regardless of accuracy, has a specificity near 100%. All tables for classification have been placed in Appendix C.

Table 1

Mutational Selectors

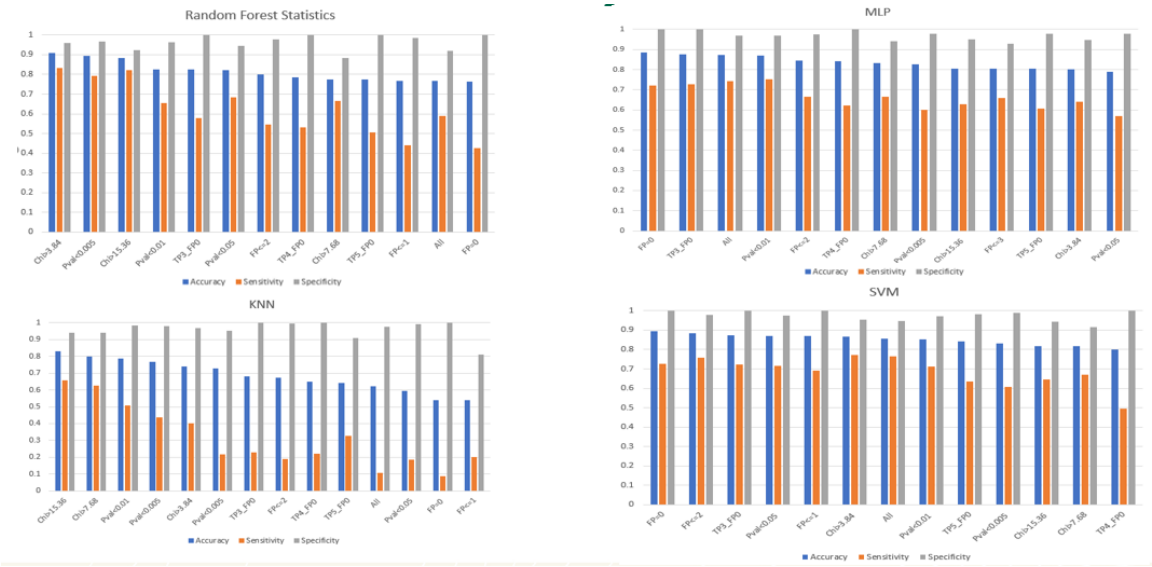
Mutational Separator	**Number of mutations**
Fishers Exact P value > 0.05	845
Fishers Exact P value > 0.01	326
Fishers Exact P value > 0.005	174
Chi Squared > 3.84	739
Chi Squared > 7.68	383
Chi Squared > 15.36	196
FP = 0, TP > 2	556
FP = 0, TP > 3	159
FP = 0, TP > 4	57
TP – 2 * FP > 2	609
TP – 2 * FP > 3	214
TP – 2 * FP > 4	80
All mutations	8085

Note: Table 1 shows the thirteen different mutational selectors described in the previous sections with the number of mutations included within each selector.

The results for each mutational selector and the respective classifier are shown below.

Figure 4

All Results for each mutation selector and classifier



Note: This diagram shows all four classification models, each set of bars represents a mutational selector. The blue bar is accuracy, the orange bar is sensitivity, and the gray bar is specificity.

This graphic shows a strong connection between the CIMP phenotype and the mutations that occur in the cancerous samples. The highest performance from all classifiers is the random forest CHI > 3.84 at just over 90%. The top results from the multilayer perceptron and support vector machine are not far behind with approximately 89% accuracies. Having all classification models produce such high results strengthens the connection between these mutations and the CIMP phenotype due to these high level accuracy classifications.

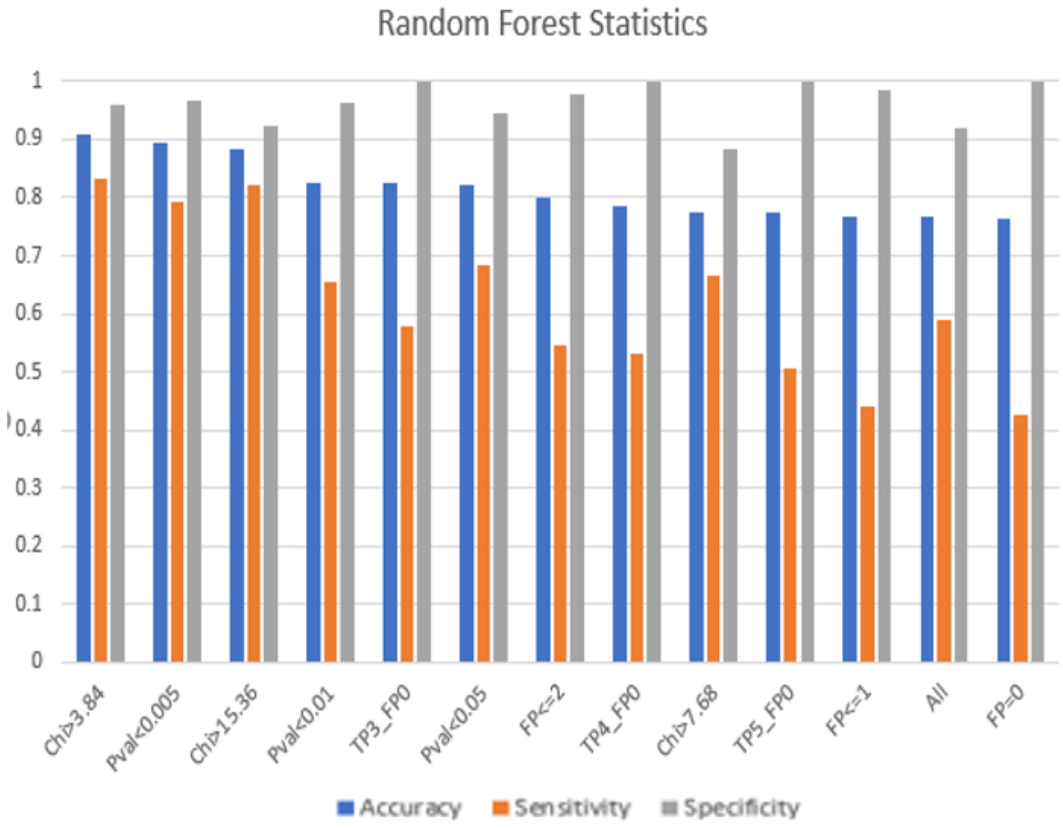
2.2.2 Random Forest

Table C.1 shows how each mutational selector performed. Each selector has a TP, FP, TN, and FN calculated which is the base for the rest of the columns calculations. In

this data, we found that Chi>3.84 is able to correctly identify samples using only mutations 90% of the time, giving testament to how strong some of the mutational selectors are.

Figure 5

Random Forest Statistics



Note: This graphic shows the random forest classifier where each set of bars represents a mutational selector. The blue bar is accuracy, the orange bar is sensitivity, and the gray bar is specificity.

All selectors in the random forest classifier are able to correctly classify samples with at least 73% accuracy. Showing that even picking the worst mutational selector there

is still a reasonable amount of accuracy. On the other hand, these low-performing selectors based on accuracy have achieved 100% specificity. This meaning, which given a mystery sample, the mutational selector can identify a CIMP- sample as negative every single time.

2.2.3: MLP, KNN, SVM results

Table C.2 shows how each mutational selector performed. Each selector has a TP, FP, TN, and FN calculated which is the base for the rest of the columns calculations. In this data, we found that FP=0 is able to correctly identify samples using only mutations 88% of the time.

Table C.3 shows how each mutational selector performed. Each selector has a TP, FP, TN, and FN calculated which is the base for the rest of the columns calculations. In this data, we found that Chi > 5.36 is able to correctly identify samples using only mutations 83% of the time. We have found in the past that the K nearest neighbor algorithm tends to be the worst performing in terms of accuracy. Seeing that the range is from 53% to 83% it is easy to see the drastic falloff in later selectors compared to the other classifiers. The way these classifiers are built does not intuitively work with the way the endometrial data is built. The K Nearest Neighbor algorithm thrives when there are two strongly unique classes. Unfortunately, the line between CIMP+ and CIMP- is not strong enough for this classifier to outperform the others.

Table C.4 shows how each mutational selector performed. Each selector has a TP, FP, TN, and FN calculated which is the basis for the rest of the columns calculations. In this data, we found that FP=0 is able to correctly identify samples using only mutations

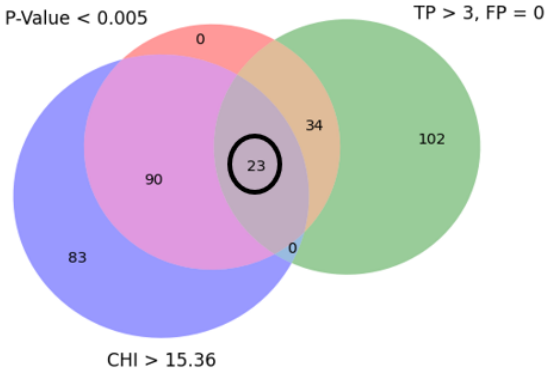
89% of the time. The support vector machine has the largest room to grow. In previous iterations of the data, the top performer has reached an accuracy of up to 94%, but this has not been a consistent occurrence. Being that these high accuracies are not consistent shows that there is potential growth in this classifier.

2.2.4: Strongest Mutation from Classification

From the random forest classification model came a collection of mutations that promoted high accuracy for the random forest. P-value < 0.005 contained 147 mutations and had an accuracy of 89%. CHI > 15.36 contained 196 mutations with 88% accuracy. TP > 3, FP = 0 contained 159 mutations with 83% accuracy. These selectors were chosen due to their high accuracy while having a minimal number of mutations. The P-value < 0.005 and CHI > 15.36 came in second and third respectively by accuracy in the random forest, and the TP > 3, FP = 0 came in second for the multilayer perceptron and third for the support vector machine. The mutations in each selector are clear divers of the classification model and all exist as good candidates for a CIMP and mutation relationship. Looking at the intersection of these statistics provided a list of mutations that are the drivers across all three classifiers.

Figure 6

Mutation Overlap



AFF1	ABHD13
CDC25A	ARF5
DENND1C	BTBD11
HAS2	CNTLN
KCNA4	DDX3X
MAGI1	ELOVL5
MESDC1	GRHL3
PCDH9	MAPK1
RSBN1L	PLXND1
SLC44A1	PXDN
	SIKE1
	TCEAL1
	ZMIZ1

Note: The Venn diagram shows the overlap of the mutational selectors p-value < 0.005, TP > 3 FP = 0, and CHI > 15.36. The chart of the right shows the center most twenty-three mutations. The mutations in red indicate mutations with existing relationships to cancer.

The most central twenty-three mutations occur in all three mutational selectors and function as the driver to all three classifiers. These can function as the base for the significant mutations that relate to CIMP. Combining the inner circles will also lead to insights into CIMP.

Table 2

P-Value < 0.005 ∩ TP >3, FP=0 mutations

AMOT	ELAVL2	LRRC57	SLMAP
ARID1A	FAHD2A	MGAT3	SRRT
ARL2BP	FOXJ3	MMP16	SRRT
B4GALNT4	FOXP1	PLEKHA3	TP63
C11orf84	HMBOX1	RABL2B	TTC3
CD47	HOXD10	RTKN2	UBE2G1
COL19A1	ITGAV	SENP2	VEGFA
CYTH1	LCP2	SEPHS1	
DYNC2LI1	LMAN1	SHC4	

Note: The list depicts the intersection of P-Value < 0.005 and TP >3, FP=0. This intersection results in thirty-four mutations that have a significant relationship to CIMP.

This interesting list is the right inner circle in figure 8. It contains thirty-four genes. These mutations are still especially important, but a deeper analysis would be required in order to fully understand their connections to CIMP and cancer.

Table 3

P-Value < 0.005 ∩ CHI > 15.36 mutations

ACVR2A	BCOR	CSNK1G1	DZIP1	IRX3	NCOA3	PTENP1	ST8SIA4	UPF3A
AK2	BTBD7	CTCF	ESRP1	JAK1	NDST1	RAB1B	SYDE2	ZBTB20
ANKH	C3orf70	CTNNA2	FAM222A	KMT2C	NFIA	RBM12B	SYNE1	ZBTB34
ANKRD13C	CAMSAP2	DDHD1	FAT1	KRAS	NSD1	RC3H1	TCERG1	ZFP91
API5	CHD3	DENND6A	FBXO48	LMAN1	POF1B	RNF2	TFAP2B	ZNF217
ARID1A	CLVS1	DOCK3	FMR1	MAF	POTEG	RNF43	TMC7	ZNF503
ASB7	CNNM4	DONSON	FOXP2	MAP3K2	POU4F2	RPL22	TMEM184B	ZNF609
BCAS3	COBLL1	DPF3	G3BP1	MSH3	PPP2R1A	RYBP	TNRC18	ZNF621
BCL11B	CPEB2	DRD5	GRIA2	MUM1L1	PRKCE	SENP1	TRIM2	ZNRF1
BCL7A	CSNK1A1	DRD5	INPPL1	MYO1A	PTEN	SETD1B	TTK	ZXDB

Note: The list depicts the intersection of P-Value < 0.005 and CHI > 15.36. This intersection results in ninety mutations that have a significant relationship to CIMP.

This interesting list is the left inner circle in figure 8. It contains ninety genes. These mutations are still especially important, but a deeper analysis would be required in order to fully understand their connections to CIMP and cancer. Combining the pink, orange, and gray middle circles we find a list of 147 mutations that have a strong relationship to the CIMP phenotype through classification. This list of mutations will be important to understanding the biology of the CIMP phenotype and the following aims of the research.

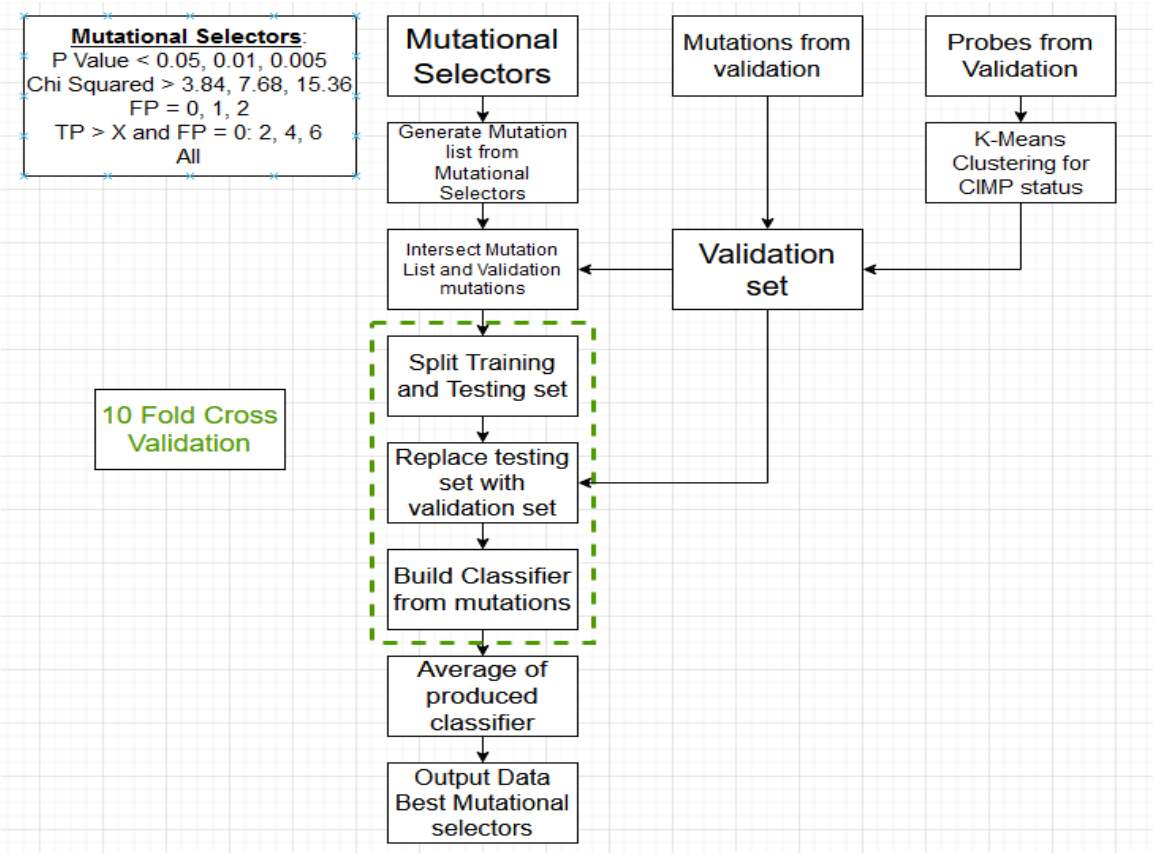
Chapter 3: Using Uterine Carcinoma Mutations to Classify Other Cancer Types

This section focuses on the classification of other cancer types using the TCGA data. The goal is to prove the model built in chapter two is a strong and accurate model using various other datasets as validation. The approach is wildly the same as the previous chapter, but with aim to validate our model and classify CIMP from a pan-cancer perspective.

In this chapter we will evaluate four unique datasets (Colorectal TCGA, Gastric TCGA, Uterine ICGC, and Uterine Cell-line) in classification with the TCGA data.

Figure 7

Validation Flow Chart



Note: This flowchart is similar to the previous chart with the exception of the generation and inclusion of the validation set.

The process for using the Endometrial Classifier to classify other cancers is similar to the previous section on Random Forests with additional preprocessing data steps. First, if the samples do not have a CIMP classification (Cell line and ICGC) they must first go through the K-means clustering process described in the in the first chapter and Sánchez-Vega paper [1]. At this point all four datasets will be at the same stage of preparation. The mutations in endometrial must be the same as the mutations in other

cancers. The validation cancer dataset was prepared in the same way endometrial cancer was prepared. The mutation columns that do not exist in both cancers will be removed. The mutational selectors are then applied to the existing mutations based on the mutational selectors from endometrial cancer. Once the subset is generated the endometrial cancer is split into the training and testing set. The training set builds the chosen classifier, and the testing set is set aside and replaced with the validation cancer data. This produces a similar table as shown in appendix C.

3.1: Results

The classification was performed on all four classification models. These tables tend to be quite large, therefore, will be included as supplementary tables. Each supplementary table is in the same format as the classification model tables in appendix C. Each dataset performed in a remarkably equivalent way giving ~80% max accuracy across most runs.

3.1.1: Validation Mutational Selectors

Almost identical to single cancer classification, mutations must be selected for the classification model. There is a slight alteration because they are two unique cancer types; all mutations used in one, must be included in the other or the classification model will no longer function.

Table 4

Mutational Selectors on validation data

Mutational Selectors	**ColRec**	**Gastric**	**ICGC**	**Elnitski Cell line**
Fishers Exact P value > 0.05	210	66	5	121
Fishers Exact P value > 0.01	85	38	3	63
Fishers Exact P value > 0.005	45	22	3	36
Chi Squared > 3.84	153	33	10	52
Chi Squared > 7.68	75	20	4	26
Chi Squared > 15.36	30	11	3	16
FP = 0, TP > 2	131	22	3	79
FP = 0, TP > 3	42	10	1	36
FP = 0, TP > 4	19	3	1	15
TP – 2 * FP > 2	151	39	1	23
TP – 2 * FP > 3	55	20	1	9
TP – 2 * FP > 4	26	10	3	59
All mutations	1460	215	28	345

Note: Table 2 shows the number of mutations for each mutational selector and the chosen secondary dataset. The process to find the values functions similar to an intersection where the mutations included are the mutations in TCGA that fit the mutational selector and exist in the validation set.

As seen in the table 2 above, there is a drastic falloff in mutations when the two sets are compared. Where previously, there were 8085 total mutations that were spread across the mutational selectors, there are only 1460 mutations to choose from in ColRec. It is important to know that all 1460 mutations came from the original 8085 mutations and that there are no new mutations included from colorectal cancer dataset. This lower

level of mutations occurs across all four validation sets. Regardless of the mutation falloff the classification models were able to still perform up to 80% accuracy.

3.1.2: TCGA Colorectal Cancer

The Random Forest classifier showed Endometrial Carcinoma classifying colorectal cancer at 81.60% accuracy using the p-value < 0.005 mutational Selector. The worst performance was from TP >4 and FP = 0 at 70.63% accuracy. The Multilayer Perceptron classifier showed Endometrial Carcinoma classifying Colorectal Cancer at 76.69% accuracy using the Chi-Squared > 15.36 mutational selector. The worst performance was from TP >4 and FP = 0 at 67.40% accuracy. The K Nearest Neighbors classifier showed Endometrial Carcinoma classifying Colorectal Cancer at 76.42% accuracy using the Fisher's exact p-value < 0.005 mutational selector. The worst performance was from TP-FP > 2 at 65.73% accuracy. The Support Vector Machine classifier showed Endometrial Carcinoma classifying Colorectal Cancer at 79.39 accuracies using the All mutational selector. The worst performance was from TP > 3 and FP = 0 at 65.93% accuracy.

These are two completely unique types of cancer, which do not even share biological systems. The ability to classify at ~80% attests to the capability to use mutations in the classification of CIMP.

3.1.3: TCGA Gastric Cancer

Another completely unique dataset evaluated against was TCGA's gastric cancer. The ICGC dataset contained 256 samples with a 215 mutation overlap. The data was classified at about 80% accuracy. The highest preforming mutational selector was all

mutations selectors. Other selectors came close with 79% accuracies (p value, CHI, TP-FP). The smallest high accuracy set was CHI > 7.68 only including 20 of the 215 mutations.

3.1.4: International Cancer Genome Consortium Dataset

The ICGC dataset is classified as uterine carcinoma. The data used for the previous research focused on the endometrium inside the uterus. From a biological perspective, this dataset compares the entirety of the uterine cancer compared to the endometrium within the uterus. The test was performed again using two more unique datasets. The first was a uterine carcinoma dataset that came from the International Cancer Genome Consortium (ICGC). The ICGC dataset contained thirty-eight samples with a twenty-eight mutation overlap. The data was classified at about 80% accuracy. The highest preforming mutational selector was the p value < 0.05. This included only five of the twenty-eight mutations that were included in the data. It is important to note that while some of the selectors used less mutations they did not perform to the same level as the p value mutational selector.

Table 5

ICGC Mutations

Gene	Location	Mutation
KRAS	12:25398284-25398284	Missense-Mutation_SNP_C-C-A_C-C-T_C-C-G
PTEN	10:89692923-89692923	Missense-Mutation_SNP_G-G-A_G-G-T
BCOR	X:39921444-39921444	Missense-Mutation_SNP_T-T-C
SET	9:131457395-131457395	3'UTR_SNP_C-C-T
VEGFA	6:43753880-43753880	3'Flank_SNP_G-G-A

Note: This shows the most useful mutations in the classification of the ICGC data.

3.1.5: Elnitski Cell Line Dataset

The Elnitski cell line dataset was provided by the Elnitski lab at NIH. The data is classified as uterine carcinoma (12 samples) and ovarian carcinoma (8 samples). The data contained 25 samples and ~400 mutation overlap, this as well classified around the 80% marker. There is a small collection of samples that was the driving factor in the classification, of the ~400 mutations only 9 of them were used in the TP >= 4 and FP = 0 mutational selector. This metric produced equivalent results to those who used more mutations.

The nine mutations are provided in the table below.

Table 6

Elnitski Cell line Mutations

Gene	Location	Mutation
BTBD11	12:107544001-107544001	Frame-Shift-Helideck-C--
DDX3X	X:41347847-41347847	3'UTR_DEL_T-T--
LMAN1	18:59346053-59346053	Splice-Site_DEL_T-T--
ZMIZ1	10:79312689-79312689	Frame-Shift-Del_DEL_C-C--
PLXND1	3:129555548-129555548	3'UTR_DEL_A-A--
ARID1A	1:26779439-26779440	Frame-Shift-Ins_INS_----G
SRRT	100881710-100881711	Frame-Shift-Ins_INS_----G
TTC3	21:37151943-37151943	Frame-Shift-Del_DEL_A-A--
KCNA4	30012016-30012016	Frame-Shift-Del_DEL_A-A--

Note: This shows the most useful mutations in the classification of the Elnitski Cell line data.

Chapter 4: Aim 2 – Finding Groups of Associated Mutations

The second aim of the research is to find the relationships between mutations, samples, and other types of cancers in correspondence to uterine cancer and a selection of other cancer types. The goal of the second aim is to find the associations between mutations in comparison to other mutations that show strong correlation in the data. Finding the association is a crucial step because there is not just one mutation that causes CIMP. CIMP is typically characterized by a collection of mutations. This will be done through a variety of statistical measures. This may give insights that a single mutation or gene cancer cannot be done alone.

Association rule mining is a tool used in data mining that builds an association between two different objects. By using the python library mlxtend, we can take the mutation data and the samples the mutations exist in to perform an association rule mining. First, you choose a K, where K is the minimum number of samples two mutations need to have in common. These mutations are added to an association set together showing that one mutation implies the next. This process is repeated iterating through all mutations and all association sets combining sets that share K samples in common. This continues until there are no more mutations that can be added to sets. The result is a large set of mutations that all share common samples. Large sets, with large samples sizes, show a strong association in comparison to smaller sets, with smaller sample sizes, which show a weaker association.

To build the rules the mutations that are not in the 149 from figure 6 are removed. The data frame is then used as an input for the mlxtend apriori algorithm with the

parameter min support ≥ 4 samples (0.01) which returns all frequent sets that contain at least four samples. The frequent item sets are then used in the mlxtend association rules function along with the minimum confidence threshold of 80%. The output is a list of association rules that have a support > 0.01 and a confidence > 80%.

Each association rule functions like an "if then" statement. The antecedent functions as the "if" statement. The consequent functions as the "then" statement. If these mutations exist, then the consequent exists. The most common way association rules are measured is by support and confidence. Support is calculated by taking the number of samples covered by the rule and dividing it by the total number of samples.

$$Support = P(A \cap \mathrm{B}) = \frac{The\ number\ of\ samples\ in\ A\ and\ B}{Total\ number\ of\ samples}$$

Confidence is calculated by taking the support of the rule and dividing it by the support of the antecedent.

$$Confidence = \frac{support\ (A \cap \mathrm{B})}{Support\ (A)} = \frac{The\ number\ of\ samples\ in\ A\ and\ B}{The\ number\ of\ samples\ in\ A}$$

This shows if there are any situations the new rule does not apply in the existing rule. Together these two metrics provide insight into the power of each association rule. Below only rule that have a support * confidence ≥ 0.013193 are shown. all other rules are in the supplementary files.

4.1: Association Rule Mining

The full table will be included in the supplementary tables. The association rule mining was performed again on the 149 selected mutations from figure 6. As an example, for calculations where the first rule covers 9 samples, the total number of samples is 379 and the antecedent covers 11 samples:

$$support = \frac{9}{379} = 0.023747$$

$$Confidence = \frac{9}{11} = 0.818182$$

Confidence is calculated by taking the support of the rule and dividing it by the support of the antecedent. The association rules in the tables below are filtered by support * confidence > 0.013. For table 7 each association rule has a support greater than 0.01%, confidence greater than 80% and support * confidence being greater than 0.013%.

Table 7

Association Rules

A	**B**	**Support(A)**	$\boldsymbol{support}$ ($\boldsymbol{A} \cap$ **B**)	**Support**	**Confidence**	**Supp * Conf**
ANKH	$DRD5_1$	11	9 (8+, 1-)	0.023747	0.818182	0.019429
ANKH, RPL22	$DRD5_1$	6	6 (5+, 1-)	0.015831	1	0.015831
CLVS1, RNF43	RPL22	7	6 (4+, 2-)	0.015831	0.857143	0.01357
$DRD5_2$, BTBD7	RPL22	7	6 (6+, 0-)	0.015831	0.857143	0.01357
COBLL1	RNF43	7	6 (5+, 1-)	0.015831	0.857143	0.01357
RNF2	RPL22	7	6 (5+, 1-)	0.015831	0.857143	0.01357
NFIA, BCL7A	RPL22	5	5 (5+, 0-)	0.013193	1	0.013193
$DRD5_2$, NFIA	$DRD5_1$	5	5 (4+, 1-)	0.013193	1	0.013193
$DRD5_2$, DZIP1	$DRD5_1$	5	5 (4+, 1-)	0.013193	1	0.013193
$DRD5_2$, CLVS1	RPL22	5	5 (4+, 1-)	0.013193	1	0.013193
$DRD5_1$, JAK1	$DRD5_2$	5	5 (5+, 0-)	0.013193	1	0.013193

Note: This table depicts 12 of 144 association rules present from the 149 mutations. There are two unique DRD5 mutations that exist each DRD5 mutations has been denoted with a subscript to identify each mutation.

The first row shows the highest support from the association rules the mutation ANKH and DRD5 had a support of 0.023 (nine samples in common). This is the first DRD5 (DRD5_GRCh38_4:9783725-9783725_3'UTR_SNP_G-G-C). The second one is indicated by a subscript two (DRD5_GRCh38_4:9783797-9783797_3'UTR_SNP_G-G-T). The following two rows show variations of the longest association rule. The two rules have a common RPL22 mutation between each set. This shows that not one mutation contributes to the CIMP classification, but various collections of mutations can attribute to CIMP status.

Table 8

CIMP Relationship

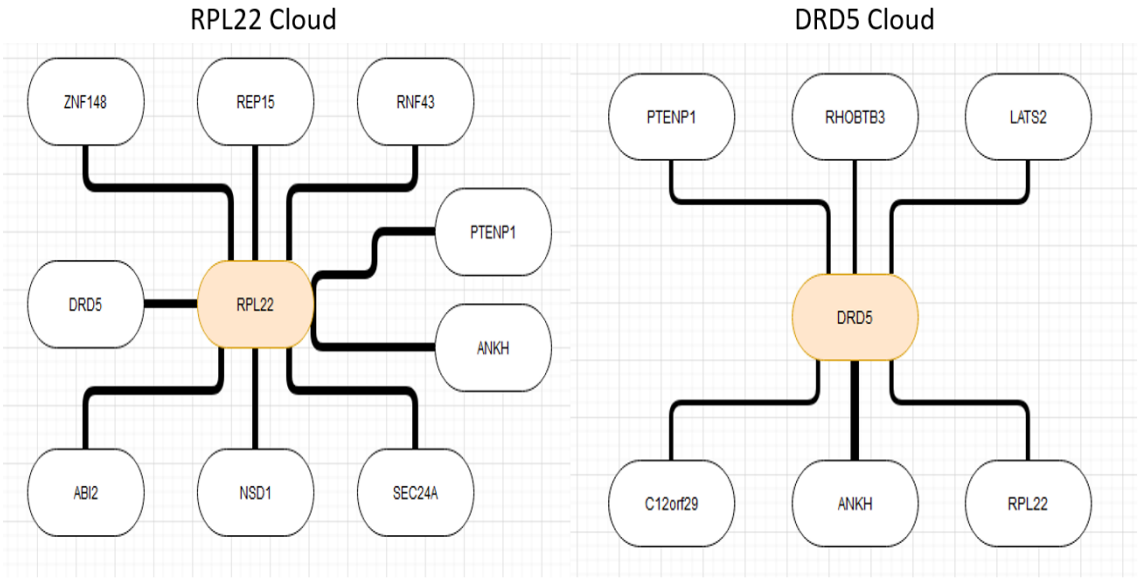
Mutation	TP	FP	FN	TN	P-Value	CHI Squared
ANKH	9	2	269	99	0.000307	43.07737
DRD5	14	5	266	94	3.61E-05	118.6723
RPL22	27	10	261	81	3.86E-09	499.0767
CLVS1	11	3	268	97	0.000109	67.69992
PRKCE	6	1	270	102	0.002592	17.5628
RNF43	22	13	258	86	1.29E-05	364.2054
ZFP91	10	1	270	98	2.16E-05	54.69303
JAK1	11	5	266	97	0.000754	71.97775

Note: The table shows the previous 8 mutations found from association rule mining.

This table shows the relationship between the mutations and CIMP. Each of these mutations have. It shows that all mutations found in the association rule mining for the most prominent mutations are all strongly CIMP+ and have a large sample coverage.

Figure 8

RPL22 / DRD5 Cloud for 149 mutations



Note: The graphic above shows the genes that had association rules with either RPL22 or DRD5. Where the weight of a line represents the support x confidence

The image above shows both a RPL22 cloud and a DRD5 cloud. A mutational cloud is a type of graphic that represents the connection between mutations. The central orange mutation is the center of the cloud with every white mutation being connected to the center by high support x confidence. These show the mutations that co-occurred in the association rule mining for the 149 mutations. At least one of the two central genes above existed in almost every rule in table 7. Showing possible relationships between mutations that could lead to the CIMP+ phenotype.

The result for using all mutations showed a group of several mutations that co-occurred in CIMP samples, using all mutations, in the dataset. For each association rule

in table 9 the support is greater than 0.01%, confidence is greater than 80% and support * confidence is greater than 0.015%. These thresholds are in place in order to only select the most prevalent mutations.

Table 9

Mutations selected from association rule mining from all mutations

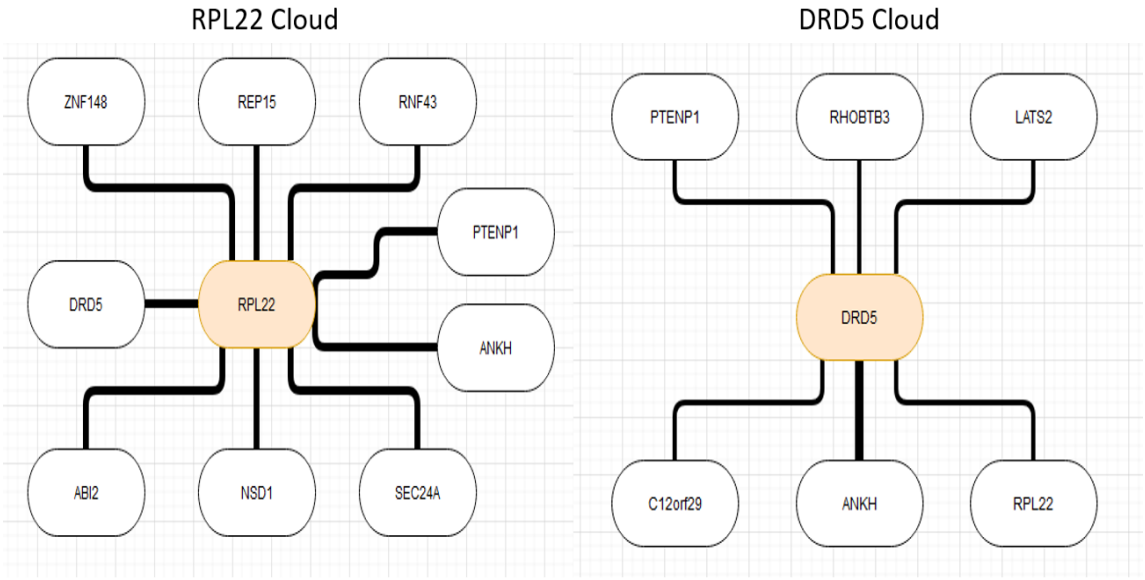
A	**B**	**Support(A)**	***support*** **(*A* ∩ B)**	**Support**	**Confidence**	**Supp * Conf**
ANKH	$DRD5_1$	11	9 (8+, 1-)	0.024	0.818	0.019
ZNF148	RPL22	8	7 (4+,3-)	0.018	0.875	0.016
REP15, RNF43	RPL22	6	6 (4+,2-)	0.016	1	0.016
C12orf29, $DRD5_2$	$DRD5_1$	6	6 (5+, 1-)	0.016	1	0.016
C12orf29, $DRD5_1$	$DRD5_2$	6	6 (5+, 1-)	0.016	1	0.016
LATS2, $DRD5_2$	ANKH	6	6 (5+, 1-)	0.016	1	0.016
LATS2, ANKH	$DRD5_2$	6	6 (5+, 1-)	0.016	1	0.016
LATS2	$DRD5_2$, ANKH	6	6 (5+, 1-)	0.016	1	0.016
PTENP1, $DRD5_1$	RPL22	6	6 (4+, 2-)	0.016	1	0.016
RPL22, ANKH	$DRD5_2$	6	6 (5+, 1-)	0.016	1	0.016
LATS2	ANKH	6	6 (5+, 1-)	0.016	1	0.016
LATS2	$DRD5_2$	6	6 (5+, 1-)	0.016	1	0.016
RHOBTB3	$DRD5_2$	6	6 (5+, 1-)	0.016	1	0.016
ABI2	RPL22	6	6 (5+, 1-)	0.016	1	0.016
NSD1	RPL22	6	6 (5+, 1-)	0.016	1	0.016
SEC24A	RPL22	6	6 (4+, 2-)	0.016	1	0.016

Note: The list of mutations above are the highest order association rule (7) to be produced while using all mutations.

Above are the seven mutations found from association rule mining. These mutations all have at least four samples in common. Each mutation was also given a support value, which shows the percentage of samples covered by the mutation. These mutations are formatted in the heatmap below to show the sample coverage.

Figure 9

RPL22 / DRD5 Cloud for all mutations



Note: The graphic above shows the genes that had association rules with either RPL22 or DRD5. Where the weight of a line represents the support x confidence.

The image above shows both a RPL22 cloud and a DRD5 cloud. These show the mutations that co-occurred in the association rule mining for all mutations. These one of

the two hub genes above existed in almost every rule. These two mutations were also the driving factor for the smaller, more significant set of mutations. This strengthens the idea of relationships between mutations that could lead to the CIMP+ phenotype

As an example, PTENP1 is in the PTEN family which is responsible for a tumor suppressor protein [13]. PTENP1 is in an association rule with both RPL22 and DRD5 showing six samples that have a relationship with cancer, RPL22, and DRD5. This chaining effect could be applied across all of the mutation clouds shown showing how closely related the top mutations in association are to CIMP samples.

Each cancer will provide its own unique set of mutations that are categorized across the cancer type. While this process will need refinement in future research this shows the beginnings of mutation and CIMP relationships between samples. Finding the associations between mutations within a single cancer then expanding to multiple cancers could build a network of mutations to categorize CIMP from a Pan-Cancer perspective.

Chapter 5: Aim 3 – Biological Interpretation of Mutations

The last aim is to interpret the findings from the previous two aims and interpret the biological significance of the mutations. This aim is highly theoretical and is the focus of the next steps, but in this section, we explored ways to move forward with this aim. The first interpretation was to analyze mutations that already have a connection to cancer.

Table 10

List of mutations with connection to cancer

Mutation	GeneCard	TP	FP	P-Value
HAS2	HAS2 (Hyaluronan Synthase 2) is a Protein Coding gene. bladder, lung, ovarian and breast cancers	9	0	9.68E-06
AFF1	This gene encodes a member of the AF4/ lymphoid nuclear protein B-Lymphoblastic Leukemia and acute Leukemia.	8	0	3.59E-05
RSBN1L	SBN1L (Round Spermatid Basic Protein 1 Like) is a Protein Coding gene Pilocytic Astrocytoma (Central Nervous system)	6	0	0.000484
KCNA4	KCNA4 (Potassium Voltage-Gated Channel Subfamily A Member 4) is a Protein Coding gene. Lung and prostate cancer	7	0	0.000132
SLC44A1	SLC44A1 (Solute Carrier Family 44 Member 1) is a Protein Coding gene Eye and Brain cancer	6	0	0.000484
MAGI1	This gene is a member of the membrane-associated guanylate kinase homologue (MAGUK) family Cervical Large cell Neuroendocrine carcinoma	7	0	0.000132
MESDC1	This gene encodes a protein that is regulated by micro-RNA MiR-574-3 Bladder cancer	9	0	9.68E-06
PCDH9	PCDH9 (Protocadherin 9) is a Protein Coding gene. Tumor suppressor Gene	9	0	9.68E-06
DENND1C	The protein encoded by this gene functions as a guanine nucleotide exchange factor Renal Cancer	6	0	0.000484
CDC25A	CDC25A is required for progression from G1 to the S phase of the cell cycle. frequently found in many cancers, and are often associated with high-grade tumors and poor prognosis	6	0	0.000484

Note: The list depicts the mutations that have an existing connection to cancer and the effect the mutation has in the genome.

Source: [14]

The table shows ten mutations that existed in the most central twenty-three mutations from the Venn diagram from figure 8. These red mutations all have connections to different cancer cells. To highlight the most important ones; HAS2 is found to commonly occur in Ovarian cancers, MAGI1 is common in cervical cancers, and CDC25A is a tumor suppressor gene that we found to be very prevalent in classifying CIMP samples.

Table 11

List of mutations without connection to cancer

Mutation	GeneCard	TP	FP	P-Value
TCEAL1	This gene encodes a member of the transcription elongation factor A (SII)-like (TCEAL) gene family.	6	0	0.000484
ELOVL5	This gene belongs to the ELO family. It is highly expressed in the adrenal gland and testis	6	0	0.000484
ABHD13	ABHD13 (Abhydrolase Domain Containing 13) is a Protein Coding gene.	6	0	0.000484
BTBD11	BTBD11 (BTB Domain Containing 11) is a Protein Coding gene.	6	0	0.000484
ZMIZ1	This gene encodes a member of the PIAS (protein inhibitor of activated STAT) family of proteins.	6	0	0.000484
CNTLN	CNTLN (Centlein) is a Protein Coding gene.	7	0	0.000132
MAPK1	This gene encodes a member of the MAP kinase family. MAP kinases, also known as extracellular signal-regulated kinases (ERKs).	6	0	0.000484
PLXND1	PLXND1 (Plexin D1) is a Protein Coding gene.	6	0	0.000484
GRHL3	This gene encodes a member of the grainyhead family of transcription factors.	6	0	0.000484
SIKE1	SIKE interacts with IKK-epsilon (IKBKE; MIM 605048) and TBK1 (MIM 604834) and acts as a suppressor of TLR3 (MIM 603029) and virus-triggered interferon activation pathways	7	0	0.000132
ARF5	This gene is a member of the human ADP-ribosylation factor (ARF) gene family.	6	0	0.000484
PXDN	This gene encodes a heme-containing peroxidase that is secreted into the extracellular matrix.	6	0	0.000484
DDX3X	The protein encoded by this gene is a member of the large DEAD-box protein family,	6	0	0.000484

Note: The list depicts the mutations that do not have an existing connection to cancer and the effect the mutation has in the genome.

Source: [14]

There are also genes that do not have an apparent connection to cancer, but we found to be highly correlated with CIMP. These mutations are remarkably interesting because we found them in endometrial carcinoma, indicating they may have connections to other cancers, but the mutation is not commonly documented.

Table 12

Association Rule mining mutations and Cancer

Mutation	GeneCard	TP	FP	P-Value
DRD5	This gene encodes the D5 subtype of the dopamine receptor. The D5 subtype is a G-protein coupled receptor which stimulates adenylyl cyclase.	14	5	3.61E-05
PRRG1	This gene encodes a vitamin K-dependent, gamma-carboxy-glutamic acid (Gla)-containing, single-pass transmembrane protein.	8	4	0.005964
PTENP1	PTENP1 represents a highly homologous processed pseudogene of PTEN (phosphatase and tensin homolog). Endometrial Cancer	8	5	0.01182
DDHD1	The protein encoded by this gene preferentially hydrolyzes phosphatidic acid. colon cancer	10	4	0.000859
NSD1	This gene encodes a protein containing a SET domain, 2 LXXLL motifs, 3 nuclear translocation signals (NLSs), 4 plant homeodomain (PHD) finger regions, and a proline-rich region.	7	2	0.002771
TNFSF11	This gene encodes a member of the tumor necrosis factor (TNF) cytokine family which is a ligand for osteoprotegerin and functions as a key factor for osteoclast differentiation and activation. Expressed in a large amount of cancer	7	3	0.007054
RPL22	RPL22 (Ribosomal Protein L22) is a Protein Coding gene. Colon and Uterine carcinoma	27	10	3.86E-09

Note: The list depicts the mutations that occurred in association rule mining. A red gene represents a mutation that has a connection to cancer, a black gene indicates one that does not.

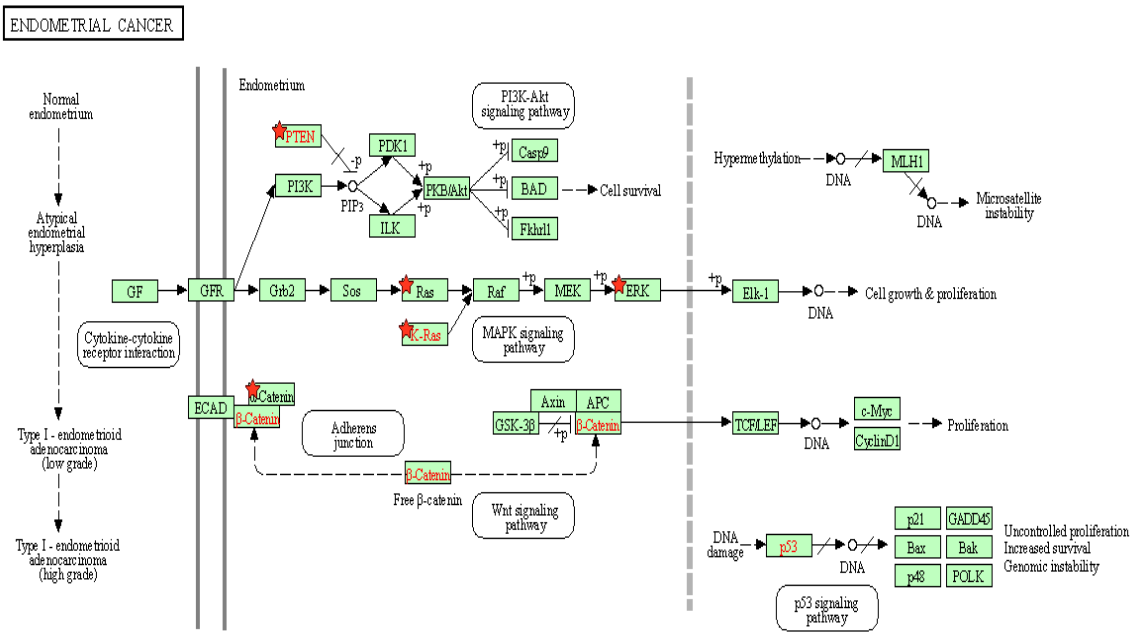
Source: [14]

The last set of mutations for a deeper look came from the association rule mining. These mutations have a support of four samples, and they are all very strongly CIMP+. Four of the seven mutations found in association rule mining were found to be related to cancer. RPL22, PTENP1, DDHD1 are both commonly found in endometrial and colon cancer as seen by TCGA data. TNFSF11 is also found, and just like KRAS, is a tumor suppressor gene that is commonly found mutated.

The last way we chose to interpret the data was to take all previously discovered mutations and check them against a DAVID pathway analysis. This analysis takes a list of genes and shows what mutations in those genes would affect across the cell. In the case below we looked at the Endometrial KEGG pathway. The KEGG pathway is one of the major pathways in cell cycle and cell life regulation [15].

Figure 10

KEGG Pathway for Endometrial Cancer



Note: The graphic shows the Endometrial KEGG pathway where each green box is a gene, a gene with a red name is a tumor suppressor and if a gene has a red star the gene is mutated in the TCGA data.

As seen in the graphic above there are a large number of mutated genes present in this pathway. The cell cycle line, represented by the top section, has a single mutation on

the PTEN gene. PTEN in this pathway functions as a tumor suppressor gene therefor a mutation may lead to cancer within the cell. This is found again in the K-Ras gene on the middle path. Even with a small look at the biological significance of the mutations we are already finding key results between CIMP, mutations, and cancer.

Chapter 6: Next Steps and Conclusions

The third aim of the research which was the analysis of the CIMP samples from a biological perspective was not able to be achieved in great detail in the timeframe allowed. Steps that would relate to this would be building classification models that would be able to classify CIMP on a pan-cancer scale, similar to the analysis presented in the third chapter when comparing the endometrial carcinoma to the colorectal carcinoma. Analyzing the most important mutations identified by the classification model by research into the relationships between CIMP related mutations and biological pathways. Another potential step would analyze the most important mutations identified by the classification model. This could be done by a knock down expression approach using RNA interference, then analyzing the change in methylation patterns that occur across the cell. A fully explored system to collect mutations that are relevant to CIMP, and all cancers could be the steppingstone needed to unlock the mystery of CIMP in Cancer.

The results show a strong indication that the mutations in cancer can predict the CIMP status of cancer. The classifiers all scored between 83% and 90% accuracy. The results also show that the highest performing accuracy was from two Chi-squared and two FP = 0. Throughout all classifiers, the sensitivity (the likelihood of a CIMP+ sample being classified as CIMP+) tends to be low, falling as low as 49%. This is the largest problem faced in the data and classifiers. We believe this is due to the sparse amount of CIMP+ grouping within the data. The most CIMP+ mutation (RPL22) only covers 27 of the 108 total CIMP+ mutations. This number only goes down from RPL22.

Being that the Support Vector Machine and the Random Forest Classifier share the highest place in terms of accuracy, either could be used in the future for data prediction or classification of endometrial cancers. Although random forest has the added benefit of each mutation having a Gini Importance, giving more insight to each classification. The results show a strong correlation between the mutations that occur in endometrial cancer and the CIMP phenotype. Proving that mutations can be used to predict CIMP status in unknown samples.

All of the results from the validation show a consistent area of accuracy when classifying validation data with uterine carcinoma, all of which have about an 80% maximum accuracy while only including as little as 6.67% of the mutations in the original classification problem. The validation of the pipeline, showed by using four unique datasets, that there are mutations present across the pan-cancer perspective that can accurately classify CIMP at approximately 80% accuracies. This process is then repeated again for pan-cancer classification and produces ~80% results showing that the mutations in CIMP are not only an endometrial carcinoma occurrence, but a pan-cancer one as well.

The results from the previous chapters show a strong connection between CIMP, uterine cancer, and mutations. It was found that it is possible to correctly classify samples as CIMP with high accuracy of up to 90% on classification using only mutational data. The classification models were then verified with unique datasets at ~80% accuracies over four unique datasets. We have found a distinct selection of mutations that can accurately sperate the CIMP+ cancer samples from the CIMP- samples. This

breakthrough in technology could give us the ability to classify unknown samples, which could lead to improved diagnostics and therapeutics and give a deeper understanding to the CIMP phenotype.

References

1. Sánchez-Vega, Francisco, et al. “Pan-Cancer Stratification of Solid HUMAN Epithelial Tumors and Cancer Cell Lines REVEALS Commonalities and Tissue-Specific Features of the CPG ISLAND METHYLATOR PHENOTYPE.” Epigenetics & Chromatin, vol. 8, no. 1, 2015, doi:10.1186/s13072-015-0007-7., https://pubmed.ncbi.nlm.nih.gov/25960768/
2. S.-W. Jiang, J. Li, K. Podratz, and S. Dowdy, “Application of DNA methylation biomarkers for endometrial cancer management,” *Expert Review of Molecular Diagnostics*, vol. 8, no. 5, pp. 607–616, 2008.
3. L. D. Moore, T. Le, and G. Fan, “DNA methylation and its basic function,” *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2012.
4. Kunitomi, Haruko, et al. “New Use of Microsatellite Instability Analysis in Endometrial Cancer.” Oncology Letters, vol. 14, no. 3, 20 July 2017, 10.3892/ol.2017.6640. Accessed 24 Nov. 2020., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5587995/
5. “CPG-Rich Island (CPG island, CG Island, CGI, HTF island, methylation-free island, Mfi, ‘Bird's island’),” *The Dictionary of Genomics, Transcriptomics and Proteomics*, pp. 1–1, 2015.
6. M. Ehrlich, “DNA hypermethylation in disease: Mechanisms and clinical relevance,” *Epigenetics*, vol. 14, no. 12, pp. 1141–1163, 2019.
7. Trimarchi, Michael P., et al. (2017) “Identification of Endometrial Cancer Methylation Features Using Combined Methylation Analysis Methods.” PLOS ONE, vol. 12, no. 3, 2017, doi:10.1371/journal.pone.0173242., https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5344376/
8. Miller, Brendan, et al. (2016) "The Emergence of Pan-Cancer CIMP and Its Elusive Interpretation." Biomolecules, vol. 6, no. 4, 2016, p. 45., doi:10.3390/biom6040045., https://pubmed.ncbi.nlm.nih.gov/27879658/

9. J. M. Teodoridis, C. Hardie, and R. Brown, “CPG island methylator phenotype (CIMP) in cancer: Causes and implications,” *Cancer Letters*, vol. 268, no. 2, pp. 177–186, 2008.
10. F. Rosenblatt, *Recent work on Theoretical Models of Biological Memory*. Cornell University, 1967.
11. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
12. The cancer genome atlas program. National Cancer Institute. (n.d.). Retrieved September 20, 2021, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
13. GeneCards. (n.d.). Retrieved March 3, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTENP1&keywords=PTENP1
14. J. Yang, L. et al. “Analysis of tumor suppressor genes based on gene ontology and the kegg pathway,” *PLoS ONE*, vol. 9, no. 9, 2014.
15. G. C. H. G. Database, “Genecards®: The Human Gene Database,” *GeneCards*. [Online]. Available: https://www.genecards.org/. [Accessed: 24-Mar-2022].
16. sklearn.ensemble.randomforestclassifier. scikit. (n.d.). Retrieved September 22, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
17. sklearn.neural_network.mlpclassifier. scikit. (n.d.). Retrieved September 22, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
18. sklearn.neighbors.kneighborsclassifier. scikit. (n.d.). Retrieved September 22, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.
19. sklearn.svm.svc. scikit. (n.d.). Retrieved September 22, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

Appendix A: Materials, Preprocessing, and Data Exploration

This appendix covers the original raw data, how CIMP is classified, the data preprocessing and the structure of the data for the exact purpose of the research. The creation of the CIMP classes combined with the processed raw data builds the dataset used for all statistical models in future chapters. In each section, a small part of the overall pipeline will be explained.

This chapter covers the original raw data, how CIMP is classified, the data preprocessing and the structure of the data for the exact purpose of the research. The creation of the CIMP classes combined with the processed raw data builds the dataset used for all statistical models in future chapters. In each section, a small part of the overall pipeline will be explained.

A.1: TCGA Data

The dataset used for this experiment was collected by The Cancer Genome Atlas program. The Cancer Genome Atlas (TCGA) is a database with over 20,000 primary cancers, spanning thirty-three cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006 [12]. The data from TCGA has been used in hundreds of projects around the world. There is a combination of two datasets used for the exact purpose of this research. Both datasets were collected from cBioPortal. cBioPortal contains a diverse collection of biologic samples for the purpose of open source research.

The First dataset selected was Uterine Carcinosarcoma (TCGA, Pan-Cancer Atlas)

(*https://www.cbioportal.org/study/summary?id=ucs_tcga_pan_can_atlas_2018)*

And the second dataset selected was Uterine Corpus Endometrial Carcinoma

(*https://www.cbioportal.org/study/summary?id=ucec_tcga_pan_can_atlas_2018*)

.

When combined, these two datasets form the full file.

A.2: Raw Data Files

When extracted to a folder, there are a variety of files, all containing different information about the cancer, patients, patient information, clinical record, RNA sequences, and other raw metadata. The file needed for data generation is the data_mutations_extended.txt file. The file is easily read if opened in excel. Looking at the most important columns; Column A is the gene name for the mutation, Column D is the build (GrCh37 - human genome 37), Column E is the chromosome number, Columns F and G are the start and end positions in the genome respectively, Column I is the type of mutation, and columns L, M, and N are the change in nucleotide for the mutation. When combining these columns, we can produce a pinpoint mutation that looks like "RPL22_GRCh37_1:6197725-6197725_Frame-Shift-Del_DEL_T-T—". The last two columns of importance are Q and R. This is the tumor sample barcode, and the matched normalized tumor sample, respectively.

A.3: CIMP Classifications

Alongside the cancer data, there is one other file needed to build the initial dataset. This file contains the CIMP classes for every sample in the TCGA database. In this CIMP classes file, there are three columns. The first column is the sample name, second is the cancer type (for the purposes of the research the cancer type is UCEC). The last column counts the CIMP typing. This can be Positive (+), Negative (-), or intermediate (i). There are also samples in the data marked as "control," these will not be used. The patterns for producing CIMP phenotypes stems from 2015 research on pan-cancer stratification of solid human epithelial tumors and cancer cell lines which revealed commonalities between cancers and tissue-specific features of the CpG island methylator phenotype [1]. The researchers analyzed DNA methylation data from the *Illumina HumanMethylation450K* platform. 5253 solid tumors were used in this study. Being that it was k-means clustering, the approximate lowest third by total levels of methylation was classified as CIMP-, the middle third was classified as CIMPi, and the upper third was classified as CIMP+. The CIMP classification file was the result of this research and the samples within are the classified samples in the study. By matching the samples in the CIMP classification file with the mutation and sample file, each sample will receive a CIMP classing.

A.4: Cimpshare Tools

There is already an existing tool to combine the CIMP classifications with the cancer mutation and methylation data. “Cimpshare” (*https://github.com/CatherineBaugher/cimpshare*), produced by Catherine Baugher, was the tool used to combine the two datasets. This tool requires that Python3 and pip3 be installed. Additional python packages are required. These packages can all be pip installed via terminal (NumPy, pandas, CSV, and SciPy). Before running the tool, some file management needs to be done. First, it is best to create a "mutations" file where the *data_mutations_extended.txt* files can be placed. There is only one per cancer type, but by combining cancer types (uterine carcinosarcoma and uterine corpus endometrial carcinoma) it is possible to produce a single data file. Also needed is the path to the CIMP_Classification file referred to earlier. The data processing part of the tool requires two command-line arguments: a path to a directory where the data_mutations_extended.txt files may be found and a path to the CSV file with the CIMP_Classification data. To run the tool, the command,

python3 analyzecimp.py -p mutations/dir/path class/CSV/path -s --outputdir output –verbose

is used. From left to right, python 3 initializes the python code architecture, analyzecimp.py is the main code base for the tool, -p processes the raw data and builds the 1/0 matrix architecture used for future calculations, mutations/dir/path class/CSV/path is the directory to the data_mutations_extended.txt files, -s takes the 0/1 matrix from -p and builds a statistics file based on the matrix, -outputdir output declares an output directory named output, and lastly -verbose gives a detailed log of all processes done to the raw data.

Once the program is finished running, there will be a new folder in the working directory called "output" with three files: a 0/1 matrix called *mutfeats.csv*, the table *statrankedmuts.csv*, and a txt file of details log.

A.5: Mutations and Samples

The data is in the format of samples on the y axis and mutations on the x-axis. This forms a 1/0 grid in the data. If there is a one for any (x,y), it indicates that the mutation y occurs in sample x. A zero indicates the opposite where the mutation y did not occur in the sample x. There are a total of 8085 mutations that exist in two or more samples. There are 379 unique samples in the dataset. Each sample in the data has a respective class indicated by the "class" column. These values can be 1 – CIMP+, -1 – CIMP- and 2 – CIMPi. The breakdown of samples is CIMP+: 108, CIMP-:142, and CIMPi: 129.

Appendix B: Classification Parameters

This appendix covers the dictionary of each parameter used for the classification models. Each following section shows the combination of parameters and setting via the specified classification model. This was done using the Sklearn library GridsearchCV.

B.1: Random Forrest

Table 13

Random Forest parameters

n_estimators	*10*	*50*	*100*	*1000*	*5000*
max_features	*Auto*	*Sqrt*	*Log2*		
Max_samples	*0.5*	*0.65*	*0.75*		
criterion	*Gini*	*entropy*			

Note: Table B.1 shows the parameters for RF

n_estimators is the number of trees in the forest.

max_features are the number of features to consider when looking for the best split.

max_samples are the number of samples to draw from X to train each base estimator.

criterion is the function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. [16]

B.2: Multilayer Perceptron

Table 14

Multilayer Perceptron Parameters

solver	Adam	sgd		
hidden_layer_sizes	*(500, 400, 300, 200, 100),*	*(400, 400, 400, 400, 400)*	*(300, 300, 300, 300, 300),*	*(200, 200, 200, 200, 200)*
activation	Logistic	Tanh	relu	
alpha	*0.0001*	*0.001*	*0.005*	
early_stopping	*True*	*False*		

Note: Table B.2 shows the parameters for MLP

Solver is the solver for weight optimization. 'sgd' refers to stochastic gradient descent. 'adam' refers to a stochastic gradient-based optimizer.

hidden_layer_sizes are the ith element represents the number of neurons in the ith hidden layer.

Activation is the function for the hidden layer. 'Logistic,' the logistic sigmoid function, returns f(x) = 1 / (1 + exp(-x)). 'Tanh,' the hyperbolic tan function, returns f(x) = tanh(x). 'Relu,' the rectified linear unit function, returns f(x) = max(0, x).

Alpha is the L2 penalty parameter.

early_stopping is whether to use early stopping to terminate training when validation score is not improving. [17]

B.3: K-Nearest Neighbors

Table 15

K-Nearest Neighbor parameters

n_neighbors	*2*	*3*	*4*	*5*	*10*
algorithm	*Auto*	*ball_tree*	*Kd_tree*	*brute*	
weights	*uniform*	*distance*			

Note: Table B.3 shows the parameters for KNN

n_neighbors is Number of neighbors to use by default for kneighbors queries.

algorithm is used to compute the nearest neighbors.

weights are a function used in prediction. [18]

B.4: Support Vector Machine

Table 16

Support Vector Machine Parameters

C	0.5	1	2
gamma	Auto	scale	
tol	0.001	0.0001	0.0000001

Note: Table B.4 shows the parameters for SVM

C is the regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive.

gamma is the kernel coefficient.

tol is Tolerance and for stopping criterion. [19]

Appendix C: Classifcation Tables

Below are the results for each individual classifier given in a table format. The tables express the separator, True Positive value, False Positive value, False Negative vale, True Negative vale, Accuracy, Sensitivity , Specificity, Precision, ROC_AUC, and Precision Recall Logistics Curve.

C.1: Random Forest

Table 17

Random Forest Results

Separator Type	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity
Chi>3.84	8.5	0.6	1.7	14.2	0.908	0.833333	0.959459
Pval<0.005	8.1	0.5	2.1	14.3	0.896	0.794118	0.966216
Chi>15.36	7.9	1.2	1.7	14.2	0.884	0.822917	0.922078
Pval<0.01	7.4	0.5	3.9	13.2	0.824	0.654867	0.963504
TP3_FP0	6	0	4.4	14.6	0.824	0.576923	1
Pval<0.05	8.2	0.7	3.8	12.3	0.82	0.683333	0.946154
FP<=2	5.6	0.3	4.7	14.4	0.8	0.543689	0.979592
TP4_FP0	6.1	0	5.4	13.5	0.784	0.530435	1
Chi>7.68	8.1	1.5	4.1	11.3	0.776	0.663934	0.882813
TP5_FP0	5.7	0	5.6	13.7	0.776	0.504425	1
FP<=1	4.4	0.2	5.6	14.8	0.768	0.44	0.986667
All	6.7	1.1	4.7	12.5	0.768	0.587719	0.919118
FP=0	4.4	0	5.9	14.7	0.764	0.427184	1

Note: Table C.1 shows the results of 10-fold cross-validation using the random forest classifier as the base classification metric. Each row represents a mutational selector, each column is a different statistical measure.

C.2: Multilayer Perceptron

Table 18

Multilayer perceptron Results

Separator Type	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity
FP=0	6.72727	0	2.63636	13.6363	0.88537	0.720094	1
TP3_FP0	7.72727	0	2.81818	12.4545	0.87747	0.726432	1
All	6.8181	0.454	2.4545	13.272	0.87351	0.744205	0.968254
Pval<0.01	7.6363	0.4545	2.5454	12.363	0.86956	0.753777	0.967657
FP <= 1	6.7272	0.3636	3.1818	12.727	0.84585	0.666286	0.974392
TP4_FP0	6	0	3.6363	13.363	0.84189	0.622084	1
Chi>7.68	6.4545	0.727	3.0909	12.727	0.83399	0.665586	0.940422
Pval<0.005	5.6363	0.272	3.7272	13.363	0.82608	0.599357	0.979181
Chi>15.36	6.4545	0.636	3.8181	12.090	0.80632	0.628591	0.949369
FP <= 2	7	0.909	3.5454	11.545	0.80632	0.660535	0.927454
TP5_FP0	6.1818	0.272	4.1818	12.363	0.80632	0.605323	0.979181
Chi>3.84	7.3636	0.636	3.9090	11.090	0.80237	0.640598	0.946052
Pval<0.05	6.2727	0.272	4.5454	11.909	0.79051	0.568576	0.978682

Note: Table C.2 shows the results of 10-fold cross-validation using the Multilayer Perceptron as the base classification metric. Each row represents a mutational selector, each column is a different statistical measure.

C.3 K-Nearest Neighbors

Table 19

K-Nearest Neighbor Results

Separator Type	TP	FP	FN	TN	Accuracy	Sensitivity
Chi>15.36	5.545455	0.909091	3	13.54545	0.83004	0.657556
Chi>7.68	6.636364	0.727273	3.909091	11.72727	0.798419	0.625187
Pval<0.01	5.181818	0.181818	4.727273	12.90909	0.786561	0.507327
Pval<0.005	4.090909	0.272727	5.090909	13.54545	0.766798	0.438292
Chi>3.84	3.636364	0.454545	5.545455	13.36364	0.73913	0.400551
Pval<0.005	23.81818	12.09091	86.18182	237.9091	0.72702	0.216529
TP3_FP0	2.181818	0	7.363636	13.45455	0.679842	0.228343
TP5_FP0	1.727273	0.090909	7.454545	13.72727	0.671937	0.18824
TP4_FP0	2.545455	0	8.090909	12.36364	0.648221	0.21883
FP <= 2	3.272727	1.272727	7	11.45455	0.640316	0.326699
All	1	0.363636	8.363636	13.27273	0.620553	0.107369
Pval<0.05	2.181818	0.090909	9.272727	11.45455	0.592885	0.186167
FP <= 1	1.909091	2.454545	8.181818	10.45455	0.537549	0.200503
FP=0	1	0	10.63636	11.36364	0.537549	0.087619

Note: Table C.3 shows the results of 10-fold cross-validation using the K-Nearest Neighbor classifier as the base classification metric. Each row represents a mutational selector, each column is a different statistical measure.

C.4: Support Vector Machine

Table 20

Support Vector Machine Results

Separator Type	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity
FP=0	6.54545	0	2.45454	14	0.89328	0.726358	1
FP <= 1	7.18181	0.27272	2.36363	13.1818	0.88537	0.758953	0.978749
TP3_FP0	7.09090	0	2.90909	13	0.87351	0.723753	1
Pval<0.05	6.18181	0.36363	2.63636	13.8181	0.86956	0.714948	0.975917
FP <= 2	6.90909	0	3	13.0909	0.86956	0.690674	1
Chi>3.84	7.81818	0.63636	2.45454	12.0909	0.86561	0.772591	0.954317
All	8.09090	0.72727	2.54545	11.6363	0.85770	0.766217	0.946065
Pval<0.01	7.36363	0.27272	3.09090	12.2727	0.85375	0.71247	0.972174
TP5_FP0	5.63636	0.27272	3.36363	13.7272	0.84189	0.636652	0.980583
Pval<0.005	5.54545	0.18181	3.72727	13.5454	0.83004	0.606061	0.988456
Chi>7.68	5.90909	1.18181	3	12.9090	0.81818	0.669998	0.915438
Chi>15.36	6.54545	0.636364	3.545455	12.27273	0.818182	0.645362	0.943939
TP4_FP0	4.636364	0	4.636364	13.72727	0.798419	0.494933	1

Note: Table C.4 shows the results of 10-fold cross-validation using the Support Vector Machine as the base classification metric. Each row represents a mutational selector, each column is a different statistical measure.

Thesis and Dissertation Services