# Title

Use of Somatic Mutations For classification of Endometrial Carcinomas with CpG Island Methylator Phenotype

# Authors

Jonathan Feige, Catherine Baugher, Lonnie Welch, Laura Elnitski

# Affiliations

# Abstract

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. Then validating the classification model with four unique datasets we found 80% accuracies. Then the mutations were used in order to uncover new discoveries about the CIMP genotype (by association rules) and phenotype. This breakthrough will be used for diagnostics and treatment of endometrial cancers.

Claims:

* It is possible to correctly classify unknown CIMP samples with just mutations (AIM 1)
  + Mutation Selection
  + RF, MLP, KNN, SVM Results
    - 90% accuracy in classification
  + Strongest Mutations For classification
    - 149 Most prevalent mutations
  + Using Uterine Carcinoma Mutations to Classify Other Cancer Types
    - 80% accuracy in validation (2 uterine and 2 pan cancer)
* Use mutations in order to uncover new discoveries about the CIMP (Genotype and Phenotype) (AIM 2/3)
  + Association Rule mining
  + Gene relationships to cancer
  + Pathway analysis

# Author Summary

* + We ask that all authors of research articles include a 150-200 word non-technical summary of the work as part of the manuscript to immediately follow the abstract. This text is subject to editorial change, should be written in the first-person voice, and should be distinct from the scientific abstract.
  + Aim to highlight where your work fits within a broader context; present the significance or possible implications of your work simply and objectively; and avoid the use of acronyms and complex terminology wherever possible. The goal is to make your findings accessible to a wide audience that includes both scientists and non-scientists.
  + Authors may benefit from consulting with a science writer or press officer to ensure they effectively communicate their findings to a general audience.

# Introduction

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.

## **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.

## **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.

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].

## **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.

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. The mutations were analyzed first by mutational selectors, then by classification models, then lastly by association rule mining.

Understanding how the CIMP subtype is related to mutations, which may lead to new treatment techniques. The first step 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 the association rules between the mutations are found this is to uncover the relationship between cancer samples, CIMP, and the significant mutations. 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.

# Results

Using just mutations, we can accurately classify unknow CIMP samples. The classification models gave a list of mutations that have a significant relationship to CIMP. Then the mutations were used in order to uncover new discoveries about the CIMP genotype (by associations) and phenotype. 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.

## **Mutational Selectors**

The mutational selectors are the basis for the classification models. Having these selectors also brings new insight into how the classification models react to different datasets with differing characteristics.

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 1*All Results for each mutation selector and classifier*

A picture containing writing implement, stationary, pencil

Description automatically generated

*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.

## **RF, MLP, KNN, SVM Results**

**Figure 4***All Results for each mutation selector and classifier*

A picture containing writing implement, stationary, pencil

Description automatically generated

*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.

Supplementary table 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.

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.

Supplementary table 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.

Supplementary able 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.

Supplementary table 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.

## **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 3***Mutation Overlap*

**Chart

Description automatically generated**

*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*

Table

Description automatically generated

*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*

Table

Description automatically generated with medium confidence

*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.

## **Using Uterine Carcinoma Mutations to Classify Other Cancer Types**

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.

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 4***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.

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 6***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.

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.

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.

## **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%.

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.

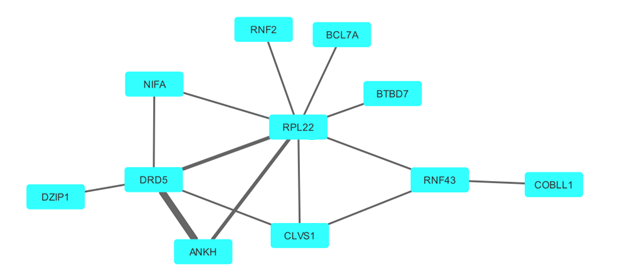
**Table 7***Association Rules*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **A** | **B** | **Support(A)** |  | **Support** | **Confidence** | **Supp \* Conf** |
| ANKH | DRD51 | 11 | 9 (8+, 1-) | 0.023747 | 0.818182 | 0.019429 |
| ANKH, RPL22 | DRD51 | 6 | 6 (5+, 1-) | 0.015831 | 1 | 0.015831 |
| CLVS1, RNF43 | RPL22 | 7 | 6 (4+, 2-) | 0.015831 | 0.857143 | 0.01357 |
| DRD52, 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 |
| DRD52, NFIA | DRD51 | 5 | 5 (4+, 1-) | 0.013193 | 1 | 0.013193 |
| DRD52, DZIP1 | DRD51 | 5 | 5 (4+, 1-) | 0.013193 | 1 | 0.013193 |
| DRD52, CLVS1 | RPL22 | 5 | 5 (4+, 1-) | 0.013193 | 1 | 0.013193 |
| DRD51, JAK1 | DRD52 | 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.

**Figure 4***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.

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.

# Discussion

## **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*

Text

Description automatically generated

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

Source: [15]

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 8***List of mutations without connection to cancer*

Table

Description automatically generated

*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: [15]

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.

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 [14].

**Figure 4***KEGG Pathway for Endometrial Cancer*

*Diagram

Description automatically generated*

*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.

## **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.

# Materials and Methods

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 5***Pipeline Diagram*

A picture containing text, sign, parking, screenshot

Description automatically generated

*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.

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

## **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 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.

## **Classification**

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, ten-fold cross-validation was used.

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.

# Acknowledgments

# 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].

# Supporting information captions