OHIO
UNIVERSITY

Performance of somatic mutations in the classification of endometrial carcinomas with CpG island methylator phenotype

Jonathan Feige

FOREVER OHIO

Introduction

- Endometrial carcinoma is the 4th most common cancer in women in the US, with approximately 66,000 cases per year.
- Individuals with endometrial carcinoma have ~30% fatality rate.
- Individuals that are CpG island Methylator Phenotype positive (CIMP+) tend to tend to be non-responsive to chemotherapies in late-stage tumors compared to those who are CIMP-.

The Cancer Genome Atlas Program

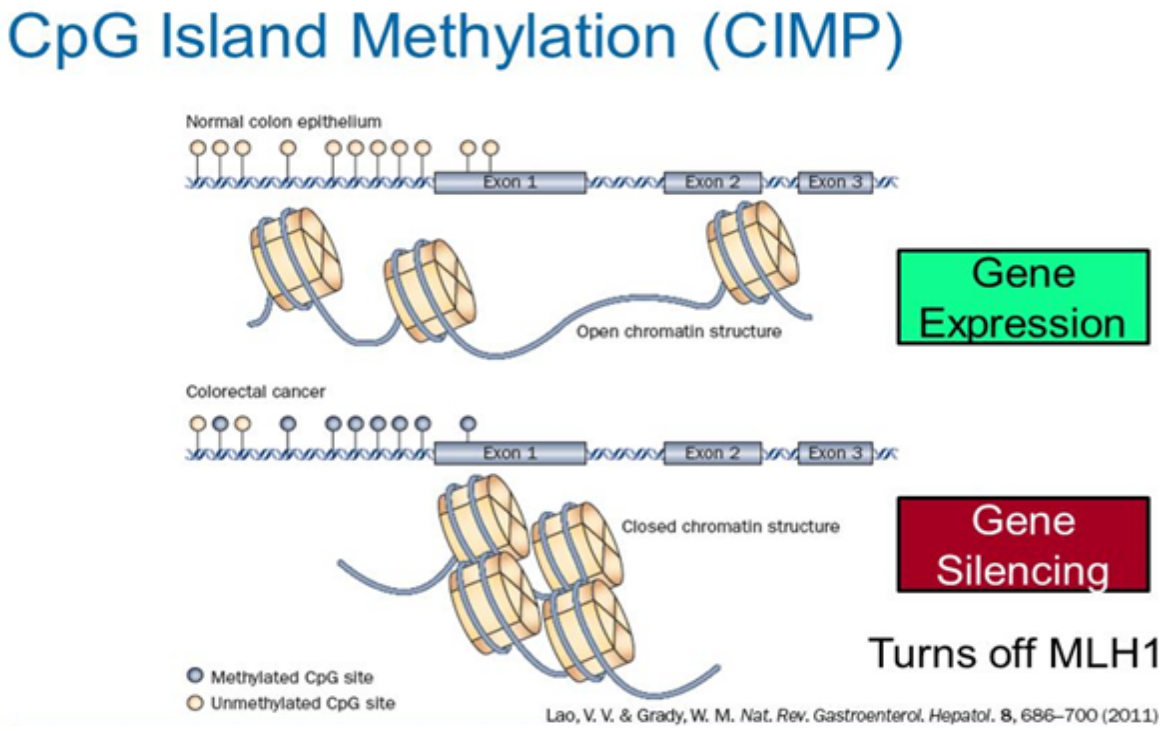
(*The cancer genome atlas program*. National Cancer Institute. (n.d.))

- The Cancer Genome Atlas (TCGA)
- TCGA was a joint effort between National Cancer Institute and the National Human Genome Research Institute that began in 2006.
- Molecularly characterized over 20,000 primary cancer and matched normal samples
- These span across 33 unique cancer types Including:
  - **Endometrial Carcinoma**
  - Gastric Carcinoma
  - Colorectal Carcinoma

What is CIMP?

(Sánchez-Vega et. al, 2015)

- CIMP (CpG Island Methylator Phenotype).
  - CIMP is represented in three classes CIMP+, CIMP-, and CIMPi
- If there is hypermethylation across the genome it is denoted as CIMP+.
- Individuals who are CIMP+ tend to be non-responsive to chemotherapies in late-stage tumors.
- Understanding CIMP gives insights into the biology of cancer.



Hypothesis and Aims

- We hypothesize that by using just mutation data we can accurately classify CIMP status.
- Aim 1 – Find the important mutations relating to CIMP.
- Aim 2 – Find the relationships between the important mutations.
- Aim 3 - Interpret the findings, biologically and medically.
- Significance:
  - Individuals with CIMP+ are non-responsive to chemotherapies in late-stage tumors. Understanding the mutations that co-occur with CIMP can help lead to early diagnostics and treatments.
  - Understanding how the CIMP subtype is related to mutations, which may lead to new treatment techniques.

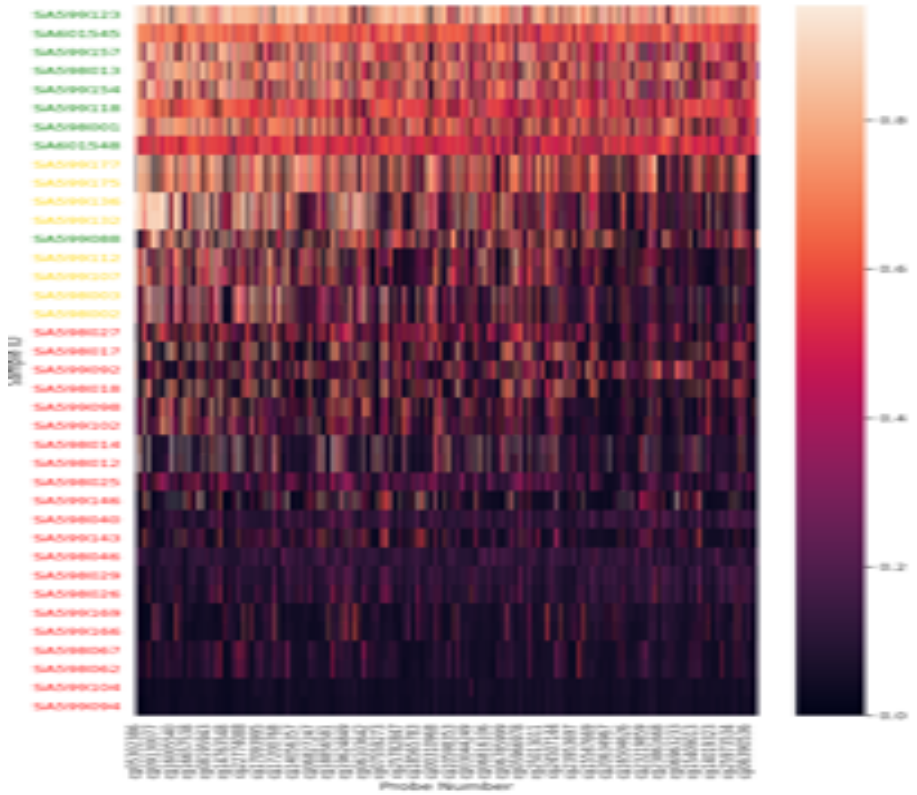
OHIO UNIVERSITY

Classification of CIMP

(Sánchez-Vega et. al, 2015)

- Each sample from TCGA has a collection of methylation probes marked across the genome.
- In endometrial carcinoma there are 1430 probes that measure methylation across each cancer sample.
- Using K-Means Clustering (K=3) samples get grouped by methylation across the genome.
- The samples with the highest methylation are CIMP+
- The samples with the lowest methylation are CIMP-

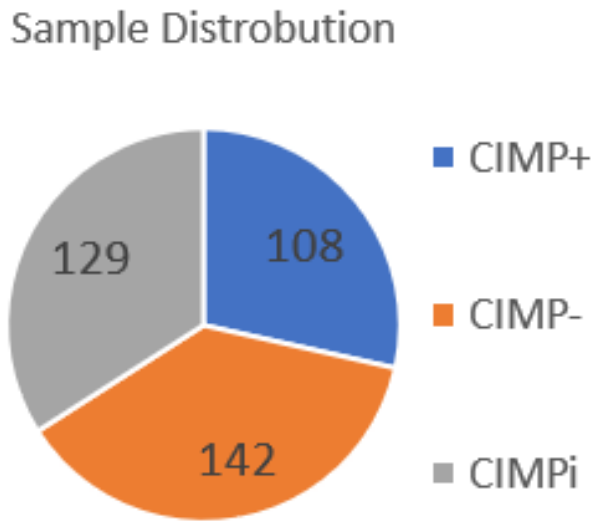
	A	B	C
1		SA599177	SA599143
2	cg0530238	0.912909	0.013218
3	cg1334805	0.024211	0.071551
4	cg2718908	0.888255	0.068849
5	cg1845652	0.824581	0.062639
6	cg1024937	0.855363	0.033493



Data Collection

- The data set used was collected from TCGA and consists of 379 unique samples.
- Each sample in TCGA has been given a CIMP+ / CIMP- / CIMPi classification.
- Across the X axis are a collection of 8085 unique mutations.
- Across the Y axis are the sample ID's.
- Each cell contains a 1 or 0. If 1 then the mutation X occurs in the sample Y.

	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



Mutational Selectors

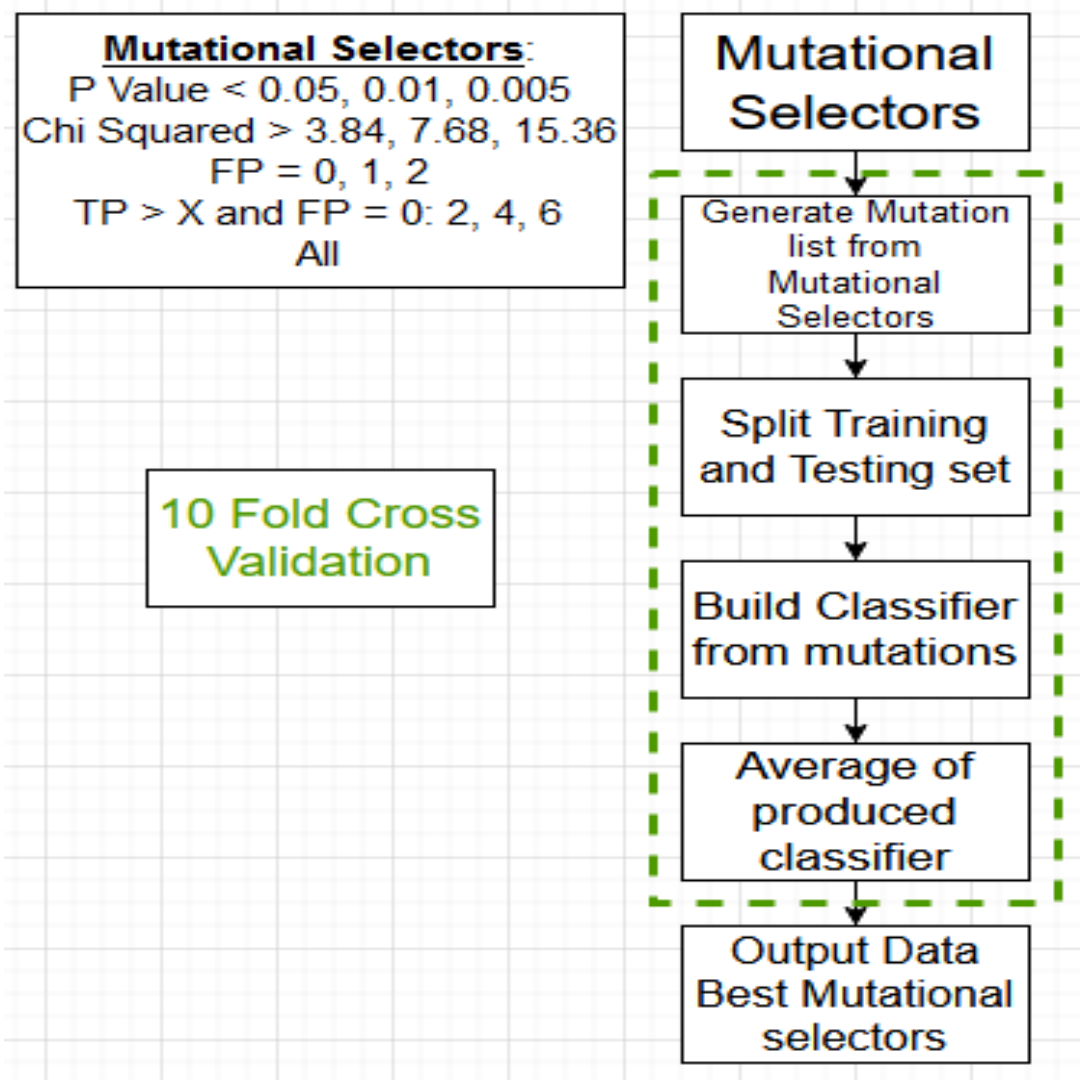
- Mutational selectors characterize a collection of mutations based on each statistical measure.
  - FP <= 0, 1, and 2.
  - TP>X and FP = 0 (TPX_FP0) where X is 3, 4 or 5.
  - Fishers exact p-value < 0.05, 0.01, and 0.005
  - Chi squared > 3.84, 7.68, 15.36
  - Using all mutations (Included as a baseline 8085 mutations)
- TP denotes a mutation occurring in a CIMP+ sample (Foreground).
- FP denotes a mutation occurring in a Non-CIMP+ sample (Background).

	A	B	C	D	E	F	G
1	Feature	TP	FP	TN	FN	p value	CHI Squared
2	RPL22_GR(	27	10	261	81	3.86E-09	499.0766752
3	RNF43_GR	22	13	258	86	1.29E-05	364.2053871
4	KRAS_GRC	22	22	249	86	0.0013	506.4821175

Classification Models

- In order to predict an unknown sample's classification, use supervised machine learning models.
  - Random Forest
  - Support Vector Machine
  - K nearest neighbors
  - Multi-layer perceptron

Pipeline


Mutational Selectors:
P Value < 0.05, 0.01, 0.005
Chi Squared > 3.84, 7.68, 15.36
FP = 0, 1, 2
TP > X and FP = 0: 2, 4, 6
All
Mutational Selectors
Generate Mutation list from Mutational Selectors
Split Training and Testing set
10 Fold Cross Validation
Build Classifier from mutations
Average of produced classifier
Output Data Best Mutational selectors

Parameter Space

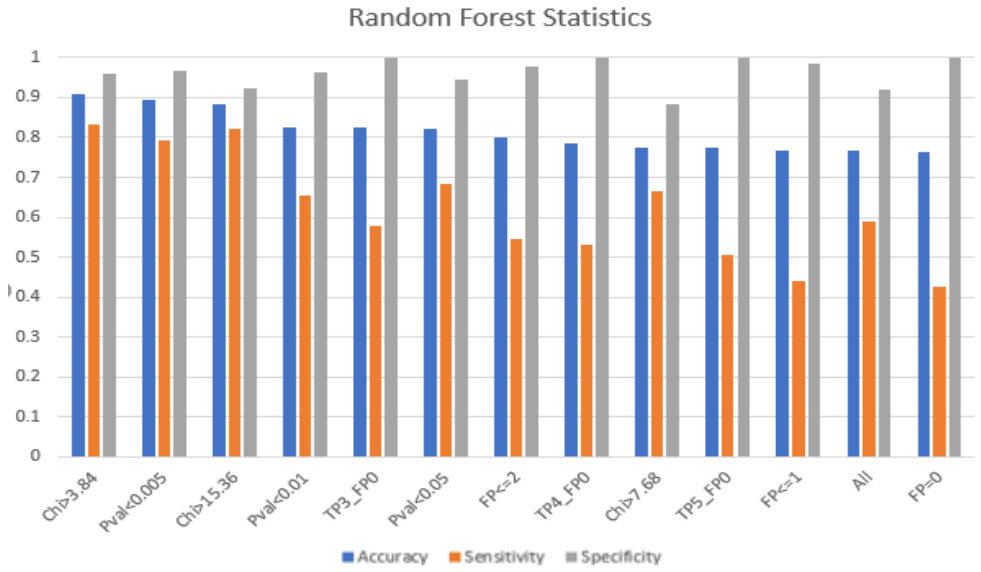
- Using Sklearn's GridsearchCV a dictionary is made of parameters and all setting that we want to explore
- The grid search then produces all possible combinations of parameters
- The best parameters change based on the mutational selector

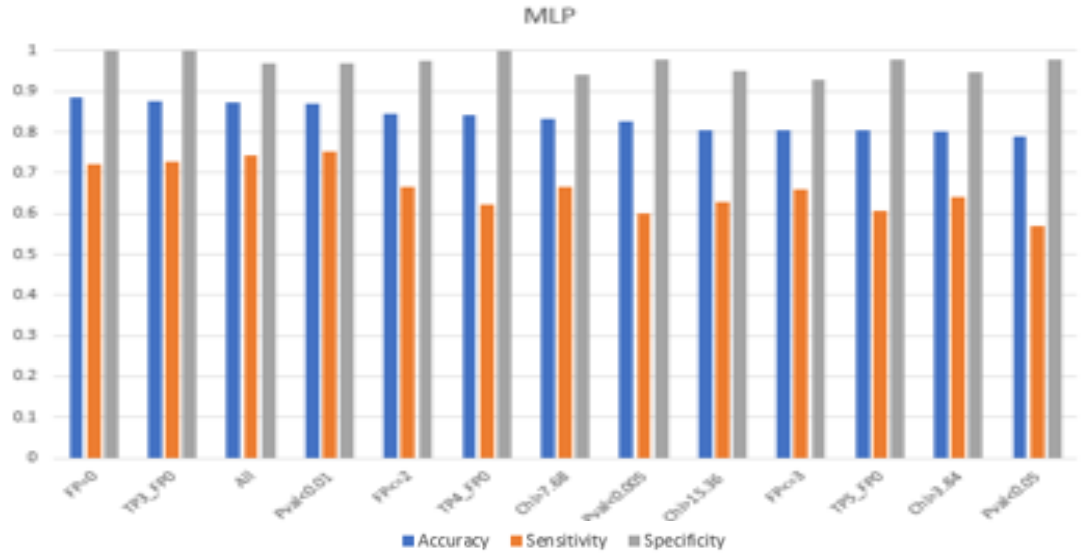
```
space = dict()
space['n_estimators'] = [50, 100, 500]
space['max_features'] = ["auto","sqrt", "log2"]
space['max_samples'] = [0.5, 0.75]
space['criterion'] = ["gini", "entropy"]
```

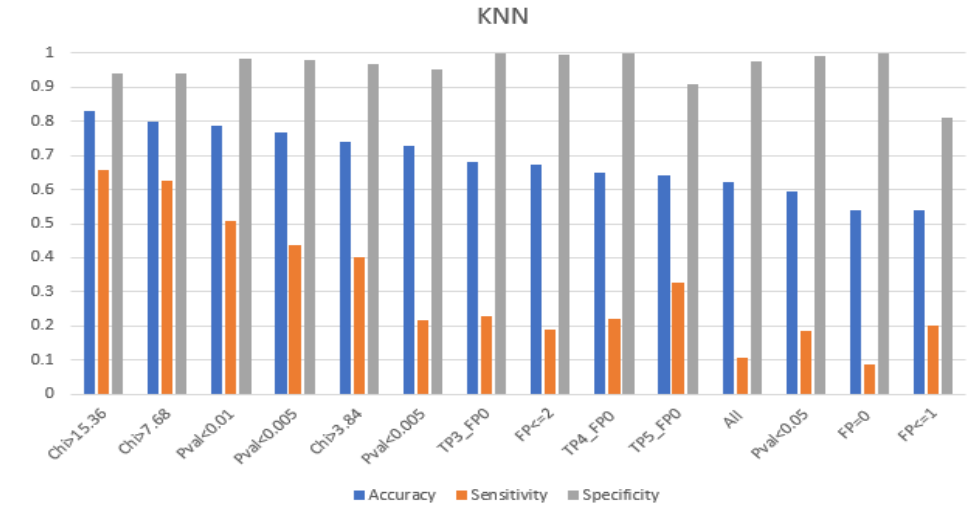
- For example: RF, CHI > 3.86
  - 500, auto, 0.5, gini
  - ~90% accuracy
- RF, p-value <= 0.005
  - 50, auto, 0.5, entropy
  - ~89% accuracy

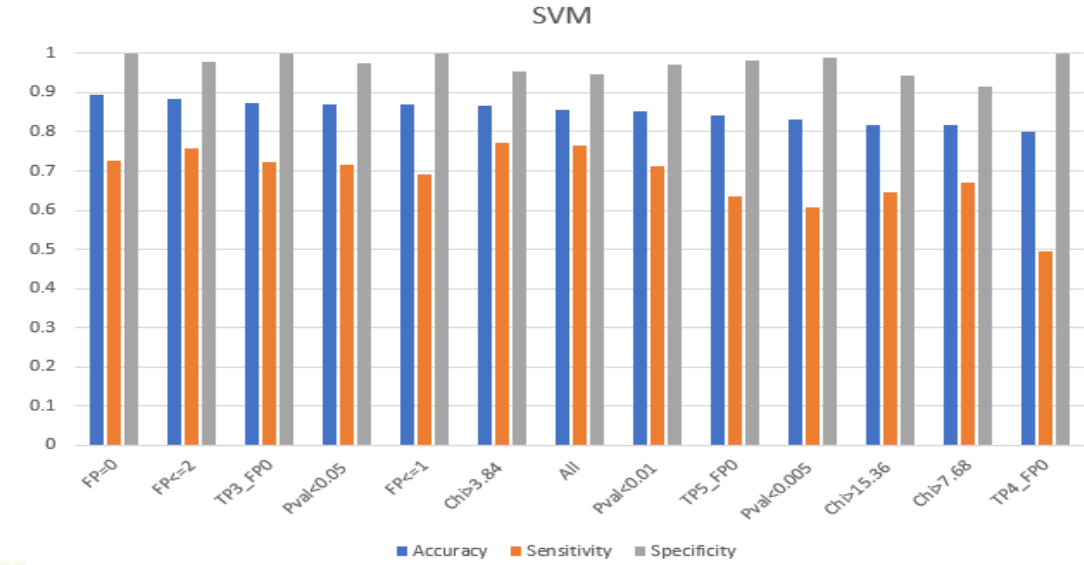
```
space = dict()
space['C'] = [0.5, 1 ,2]
space['gamma'] = ["auto", "scale"]
space['tol'] = [0.001, 0.0001, 0.0000001]
```

The mutations can predict CIMP status



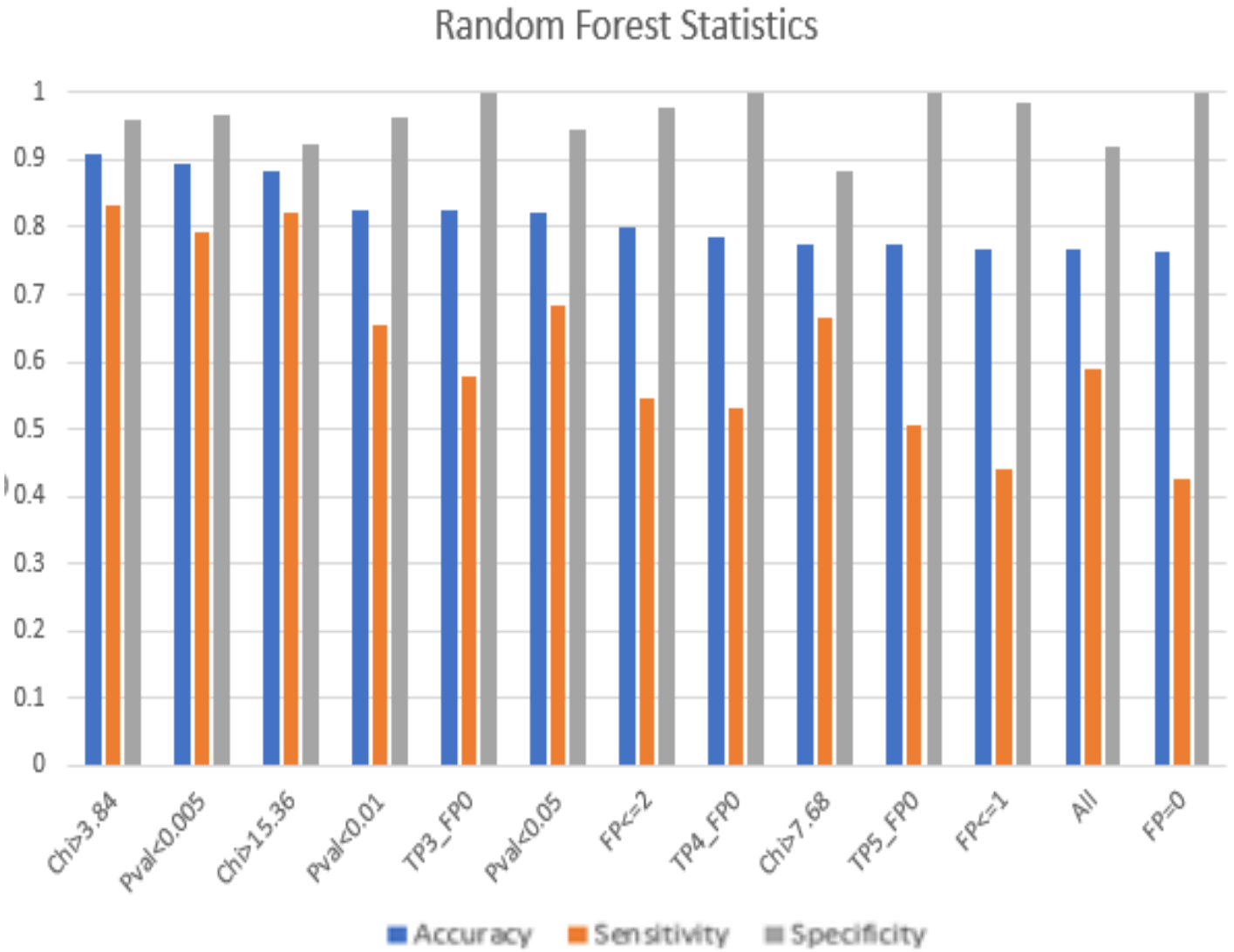






Results

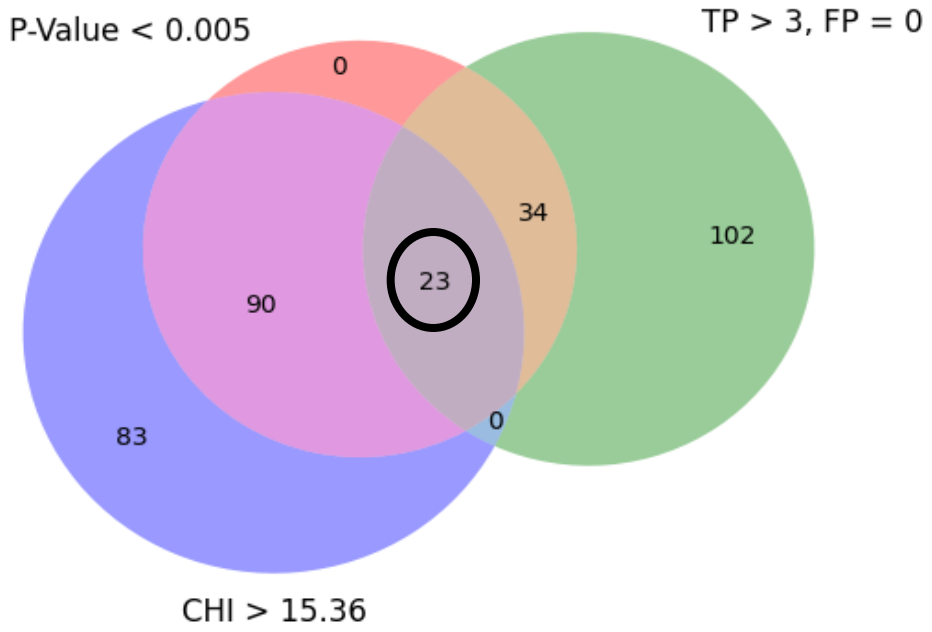
- Random forest is the highest performing supervised machine learning model for this data set, (~90% accuracy)
- This result alone shows a connection between CIMP and the mutations that occur within the cancer cell.
- The random forest is chosen for its interpretability due to the gini index assigned to each mutation per classification



OHIO UNIVERSITY

Results

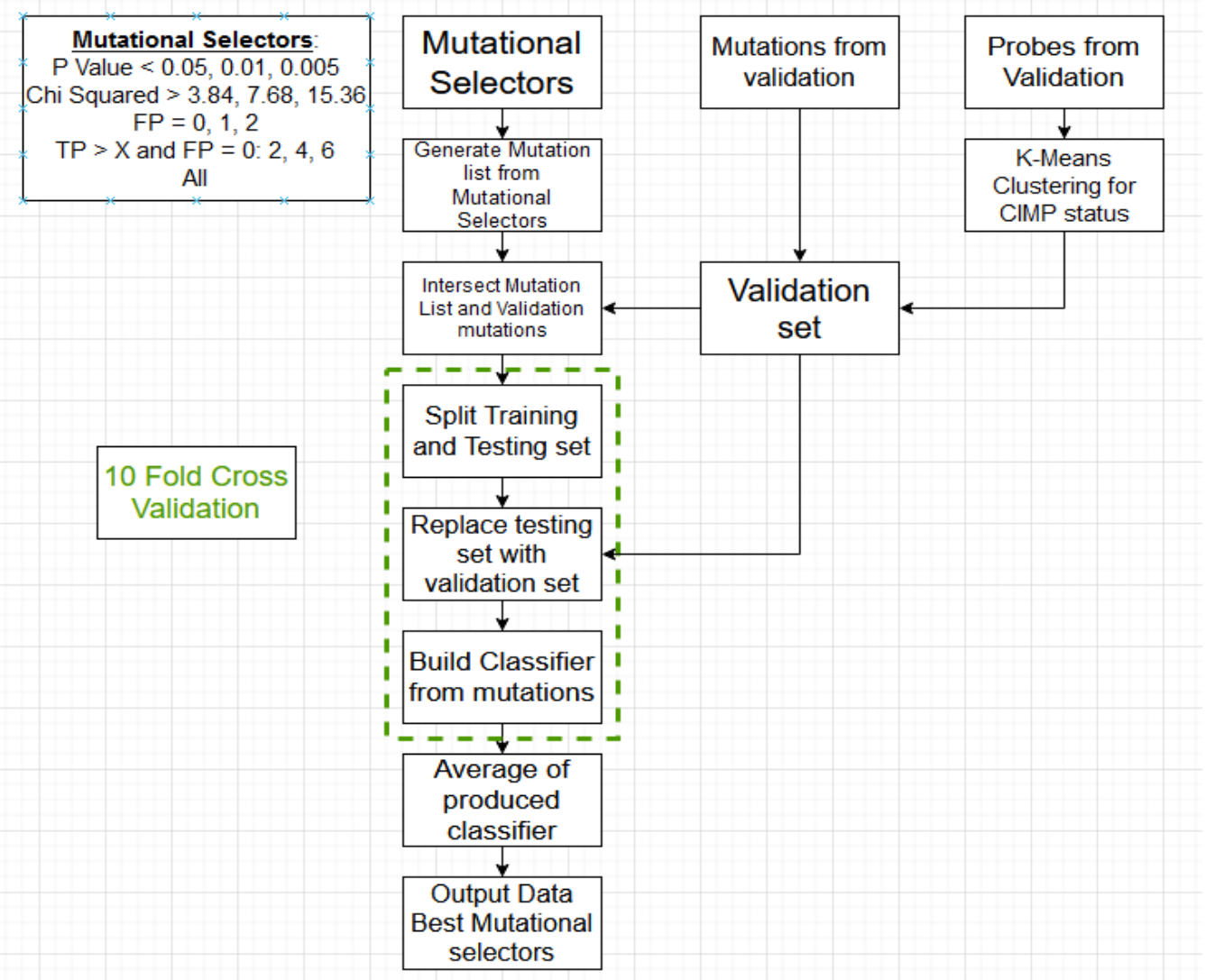
Separator Type	TP	FP	FN	TN	Accuracy	Sensitivity	Specificity	Included Muts
Chi>3.84	8.5	0.6	1.7	14.2	0.908	0.833333	0.959459	739
Pval<0.005	**8.1**	**0.5**	**2.1**	**14.3**	**0.896**	**0.794118**	**0.966216**	**147**
Chi>15.36	**7.9**	**1.2**	**1.7**	**14.2**	**0.884**	**0.822917**	**0.922078**	**196**
Pval<0.01	7.4	0.5	3.9	13.2	0.824	0.654867	0.963504	326
TP3_FP0	**6**	**0**	**4.4**	**14.6**	**0.824**	**0.576923**	**1**	**159**
Pval<0.05	8.2	0.7	3.8	12.3	0.82	0.683333	0.946154	845
FP<=2	5.6	0.3	4.7	14.4	0.8	0.543689	0.979592	5941
TP4_FP0	6.1	0	5.4	13.5	0.784	0.530435	1	57
Chi>7.68	8.1	1.5	4.1	11.3	0.776	0.663934	0.882813	383
TP5_FP0	5.7	0	5.6	13.7	0.776	0.504425	1	23
FP<=1	4.4	0.2	5.6	14.8	0.768	0.44	0.986667	2582
All	6.7	1.1	4.7	12.5	0.768	0.587719	0.919118	8085
FP=0	4.4	0	5.9	14.7	0.764	0.427184	1	556



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

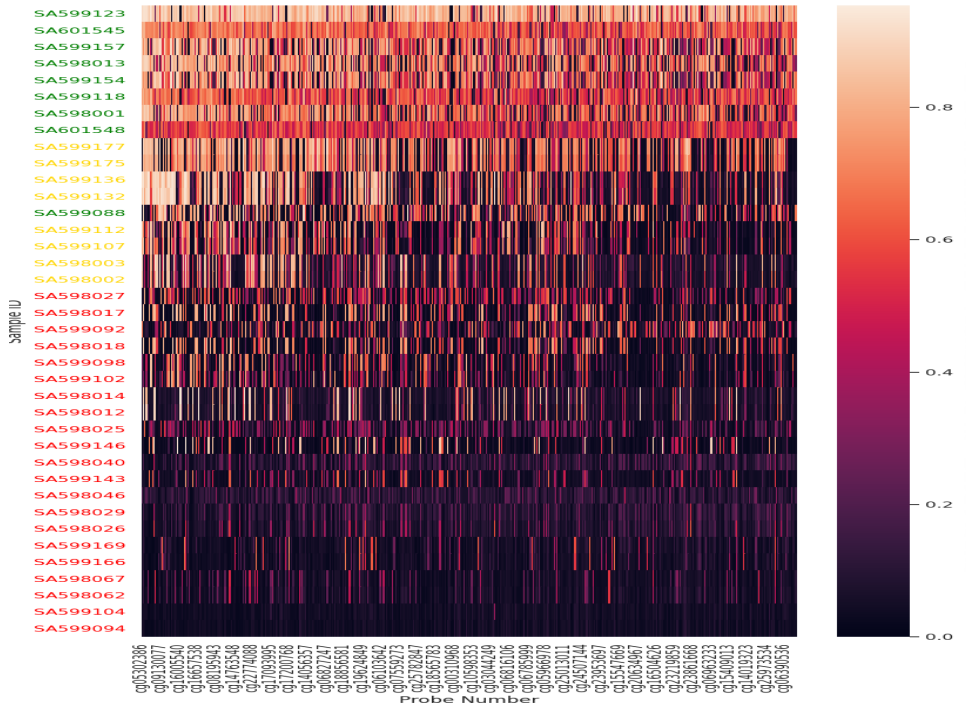
Validation

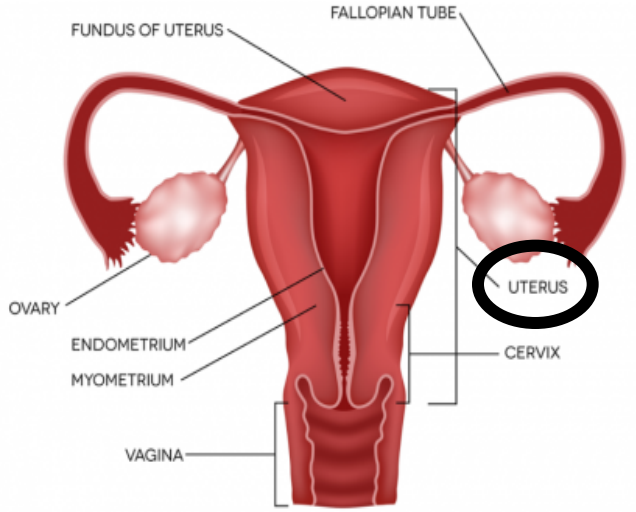
- How do we prove that the classification model is not overfit to the TCGA data?
- 4 validation sets were used:
  - Uterine (Validation):
    - ICGC
    - Elnitski Cell line
  - Non-uterine (Pan-Cancer):
    - TCGA Colorectal Cancer
    - TCGA Gastric Cancer
- If the TCGA classification model can accurately classify other datasets, this will help prove the model's functionality.



CIMP Classification for Unclassified Data

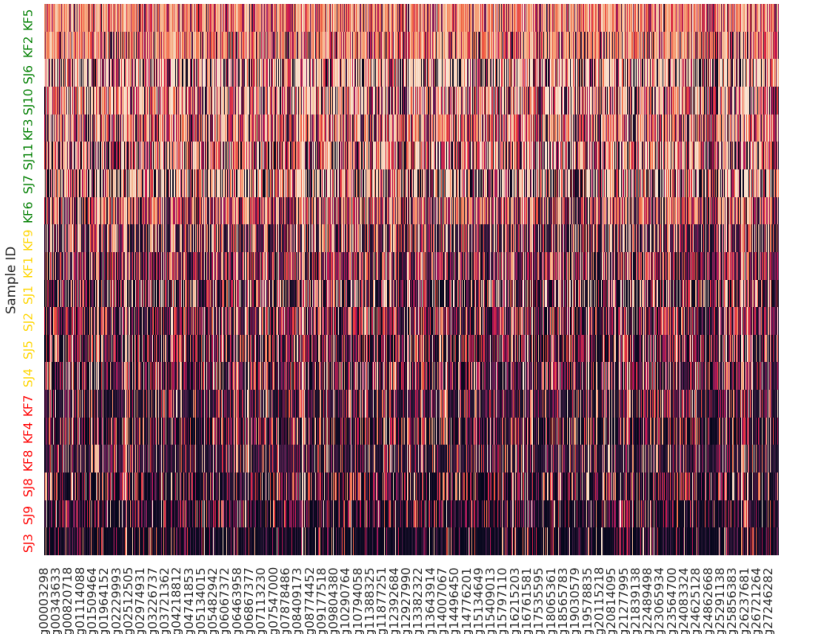
- ICGC statistics
  - Samples: 38
    - CIMP+: 9
    - CIMP-: 21
    - CIMPi: 8
  - Mutation Overlap:28
  - Peak accuracy: 80.1%

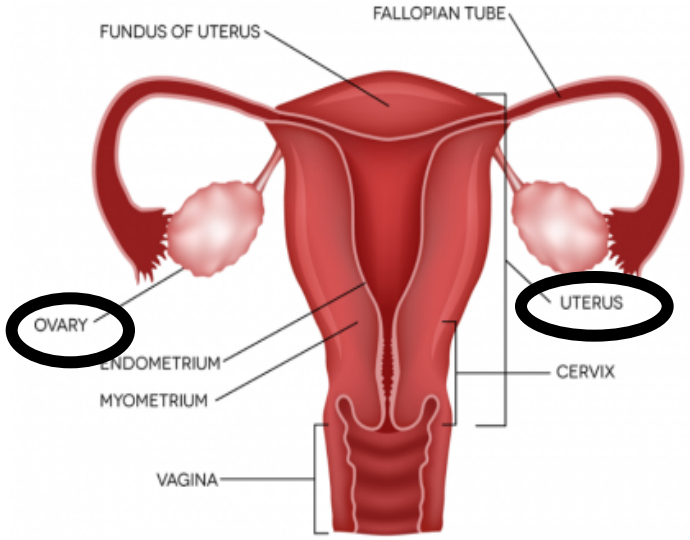




CIMP Classification for Unclassified Data

- Elnitski Lab cell line statistics
  - Samples: 20
    - CIMP+: 8
    - CIMP-: 6
    - CIMPi: 6
  - Mutation Overlap: 345
  - Peak accuracy: 78.9%

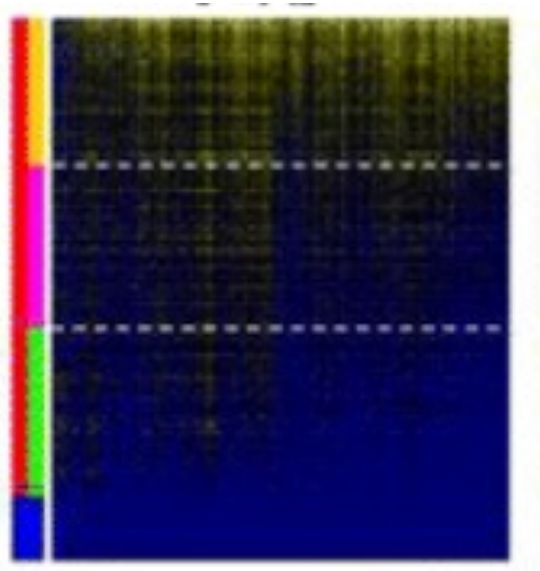


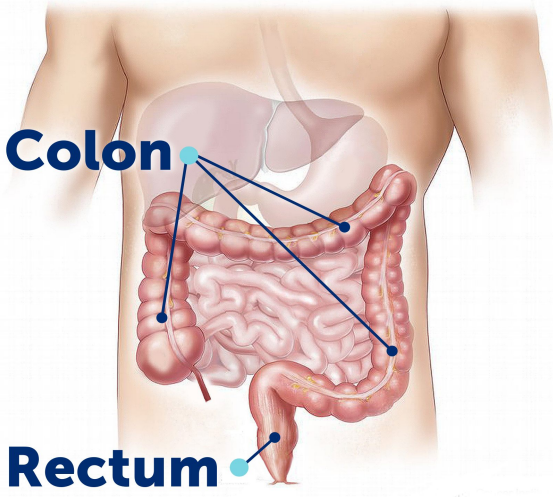


Validation TCGA Data

(Sánchez-Vega et. al, 2015)

- Two validation Sets were used form TCGA
  - Colorectal Cancer
  - Gastric Cancer
- Colorectal Cancer Statistics
  - Samples: 360
    - CIMP+: 110
    - CIMP-: 120
    - CIMPi: 130
  - Mutation Overlap: 540
  - Peak accuracy: 81.0%

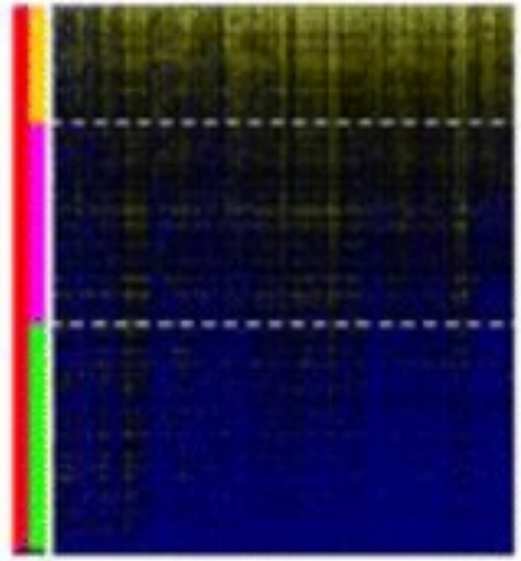


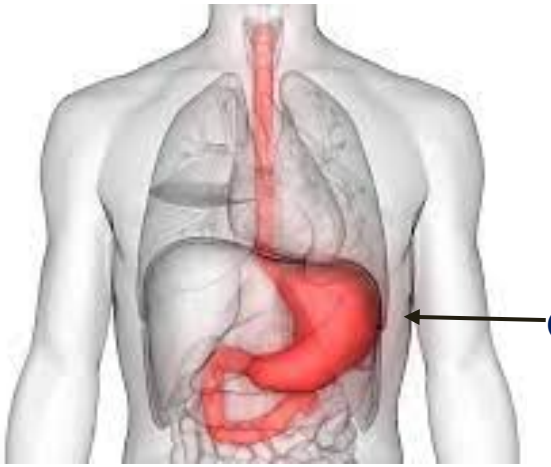


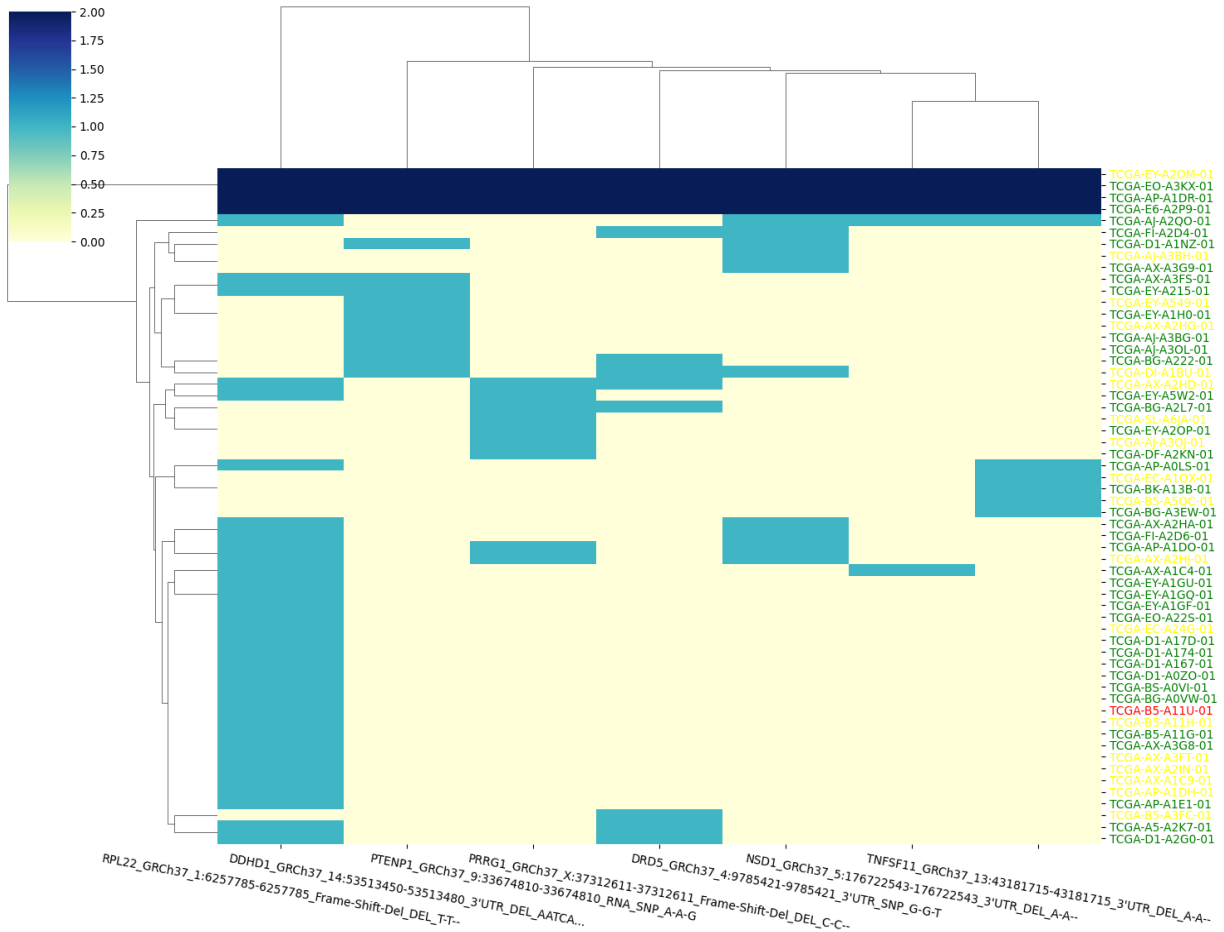
Validation TCGA Data

(Sánchez-Vega et. al, 2015)

- Gastric Cancer Statistics
  - Samples: 256
    - CIMP+: 56
    - CIMP-: 106
    - CIMPi: 94
  - Mutation Overlap: 215
  - Peak accuracy: 80.6%
- Each cancer classifying at a high accuracy shows that the mutations have a strong linkage to the CIMP phenotype.



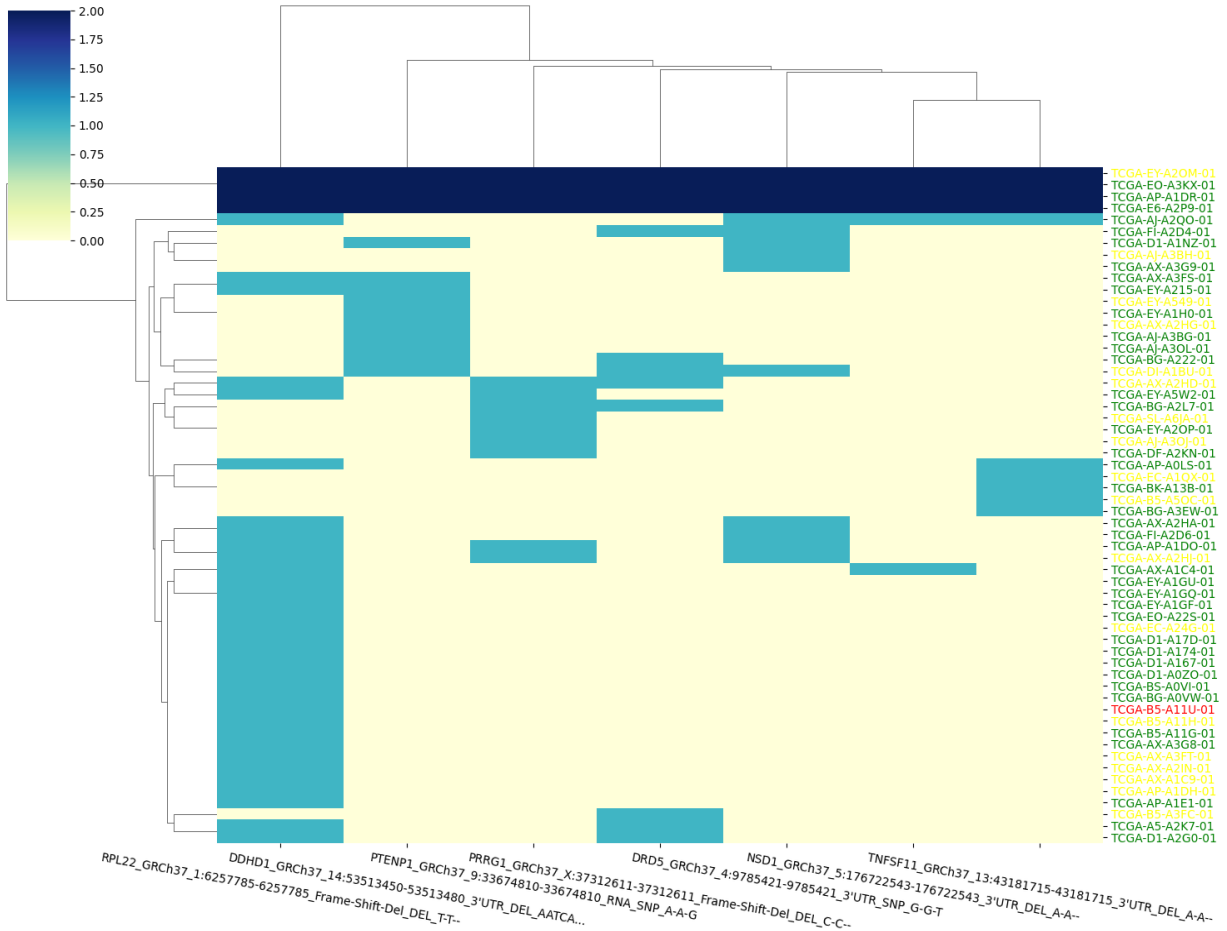




Mutations	P-Value
DRD5_GRCh37_4:9785421-9785421_3'UTR_SNP_G-G-T	3.61E-05
PRRG1_GRCh37_X:37312611-37312611_Frame-Shift-Del_DEL_C-C--	0.005
PTENP1_GRCh37_9:33674810-33674810_RNA_SNP_A-A-G	0.01
DDHD1_GRCh37_14:53513450-53513480_3'UTR_DEL_AATCAGTTT...	8.59E-04
NSD1_GRCh37_5:176722543-176722543_3'UTR_DEL_A-A--	0.002
TNFSF11_GRCh37_13:43181715-43181715_3'UTR_DEL_A-A--	0.007
RPL22_GRCh37_1:6257785-6257785_Frame-Shift-Del_DEL_T-T-	3.86E-09

Aim 2 – Find the relationships between the important mutations.

- The seven mutations have at least 4 samples in common.
- Using association rule mining, a collection of common and relevant mutations are discovered.

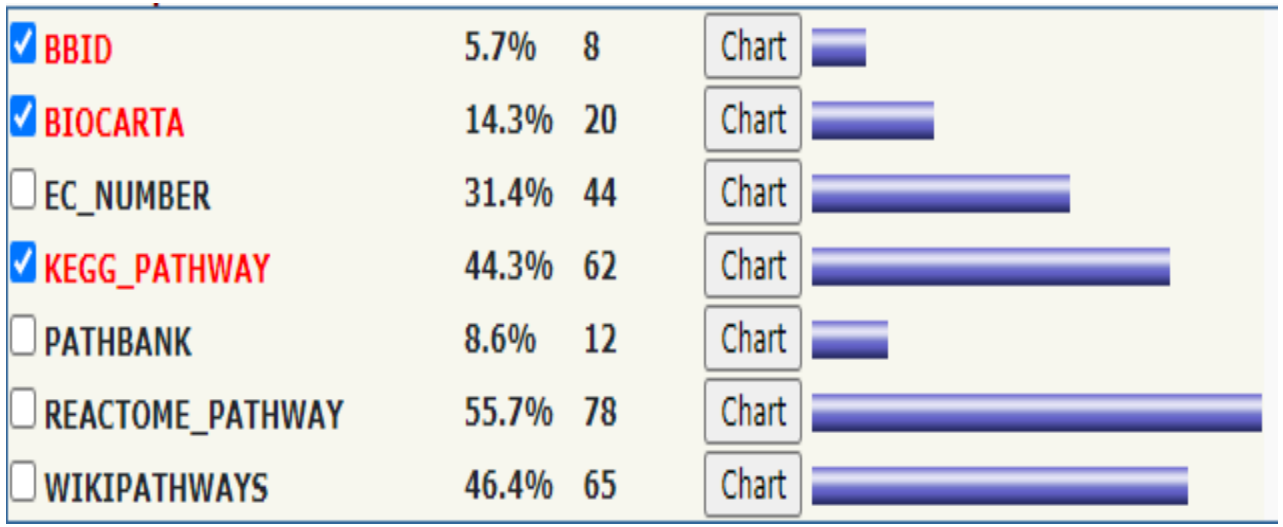


OHIO UNIVERSITY

Aim 3 - Interpret the findings, biologically and medically.

This aim would consist of:

- Identifying a collection of mutations and their effects in cancer.
- Further research into the relationship between CIMP related mutations and their corresponding signaling pathways.

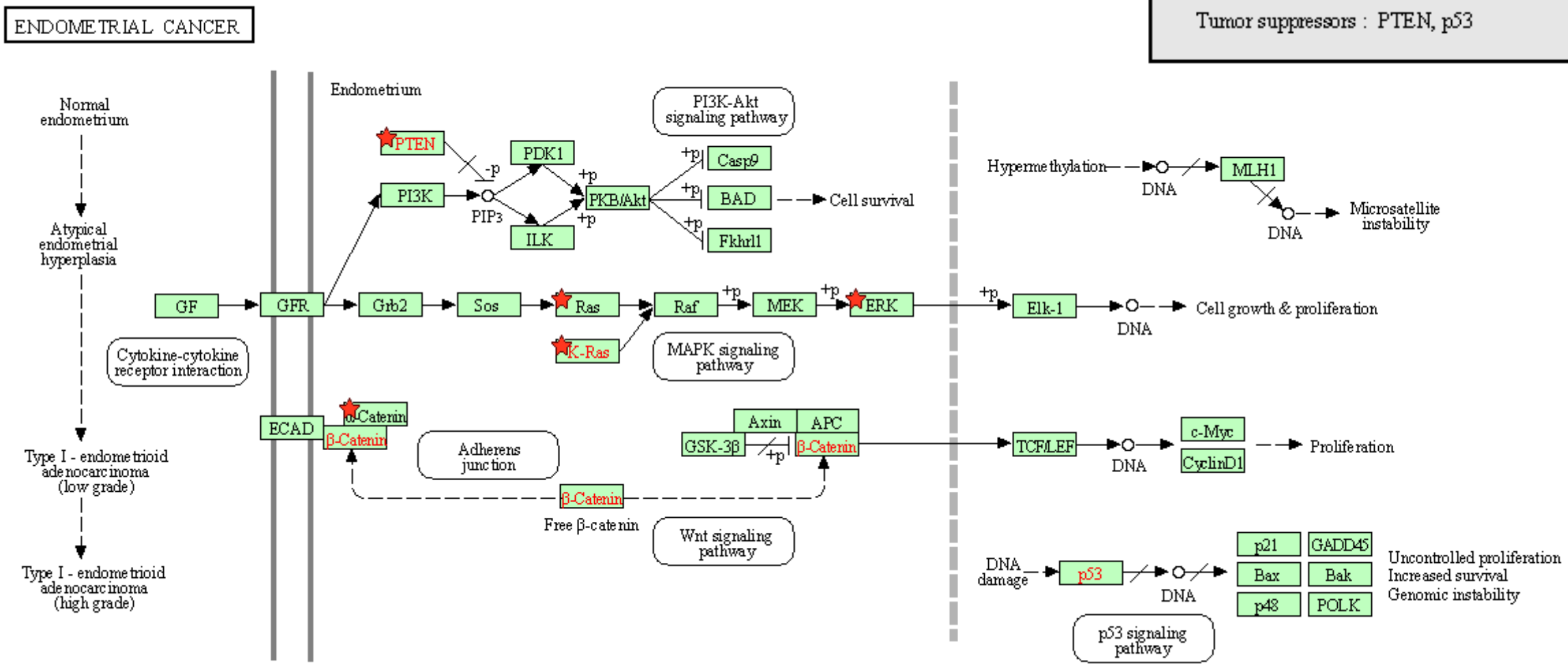


KEGG Pathways in Endometrial Carcinoma

Genetic alterations

Oncogenes : K-Ras, β-Catenin, Her2/neu

Tumor suppressors : PTEN, p53



23 Classification Mutations (Cancer)

(*Genecards: The Human Gene Database*)

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

23 Classification Mutations (Non-Cancer)

(*Genecards: The Human Gene Database*)

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

Association Rule Mining Mutations

(*Genecards: The Human Gene Database*)

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

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	

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

- In conclusion:
  - The results from the previous three aims show a strong connection between CIMP 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.
  - We have found a distinct selection of mutations that can accurately separate 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.
- Next Steps:
  - Building classification models that would be able to study CIMP on a pan-cancer scale.
  - Analyze the most important mutations identified by the classification model by research into the relationships between CIMP related mutations and biological pathways.
  - Generate hypotheses for future experiments.

Reference List

- Database, G. C. H. G. (n.d.). *Genecards®: The Human Gene Database*. GeneCards. Retrieved March 15, 2022, from https://www.genecards.org/
- *The cancer genome atlas program*. National Cancer Institute. (n.d.). Retrieved March 15, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
- 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/
- 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/
- 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/

Acknowledgments

- NIH Summer Internship Program
- Dr. Elnitski, National Human Genome Institute
- Dr. Welch
- Catherine Baugher
- Derek Petrosian
- Julia Cygan
- Trent Davis
- Yingnan Zhang