**Open**

Hello, my name is Jonathan Feige, I am a masters student in computer science and bioinformatics at Ohio university. I worked under the supervision of Dr. Laura Elnitski for research in the performance of somatic mutations in classification of endometrial carcinomas with CpG island methylator phenotypes.

**Abstract**

Endometrial carcinoma is a type of uterine cancer that begins in the endometrium lining within the uterus. A subset of these tumors shows genome hypermethylation, resulting in down-regulation of tumor suppressor genes. This is commonly referred to as CpG island methylator phenotype (CIMP).

Our working hypothesis is DNA mutations in endometrial tumors can be used to predict CIMP. This capability will be useful when DNA methylation data is not available for analysis. Being able to quickly and accurately classify CIMP patients will lead to better diagnostics and therapies.

**Introduction**

Endometrial carcinoma is the 4th most common cancer in women in the US. Approximately 66,000 cases per year with a fatality rate of ~20% (13,000). DNA methylation landscapes are often altered in cancers. In particular we looked at the CpG island methylator phenotype. There are three subtypes of CIMP: CIMP positive, CIMP negative, and CIMP intermediate. In the past it has been shown that tumor suppressor and DNA repair genes are frequently silenced in CIMP+ cases.

**Materials**

The Uterine carcinoma samples were pulled from the cancer genome atlas. The data used consists of 250 unique samples. 108 samples are CIMP+ while the other 142 samples are CIMP-. All CIMPi samples were removed. There were 8085 total mutations. Of the 8085 mutations, 739 demonstrated a strong correlation to either the CIMP+ or the CIMP- grouping all other mutations were removed.

**Methodology**

**Our goal is to correctly classify tumors as CIMP+ or CIMP-** by building a random forest classifier. The diagram below expresses the classification model used to produce the results. Three inner thresholds were used for mutation selection. The first being fishers exact p value, the second being the chi squared significance test and the last only using mutations that exist in CIMP+ samples and no CIMP- samples, we coined these as mutational selectors.

**Results**

In terms of results, using only these mutations we can classify samples with an accuracy between 100 and 70 precent. On average, correctly classify a sample with 87% accuracy.

We found that the CHI Squared mutational selectors had the best overall performance, while the mutational selector that included no CIMP- had perfect specificity in classification.

In the graphic below it expresses the best mutational selectors, we had a primary focus on accuracy, is depicted in blue.

In looking for the most prevalent mutations we found that 159 mutations in the center of the Venn diagram express all mutational selector. We found that these mutations are the driving force behind the results.

**Conclusion**

In conclusion, we found that is possible to correctly classify samples as CIMP with high accuracy using only mutational data. This breakthrough in technology could give us the ability to classify unknown samples which would help a patient with better diagnostics and therapies.

Our next steps will be to Validate the model with an existing dataset, explore the relationships between mutations and to interpret the findings from a biological perspective.

**Thank you for coming**