Machine Learning approaches with interpretability to improve clinical applications for colorectal cancer classification

Inbar Leaf [1,2,3], **Ahmad Aghaebrahimian** [1,2], **Maria Anisimova** [1,2]

[1] FS Bioinformatics, ZHAW Zurich University of Applied Sciences, Switzerland;
[2] SIB Swiss Institute of Bioinformatics, Switzerland; [3] University of Zurich, Switzerland

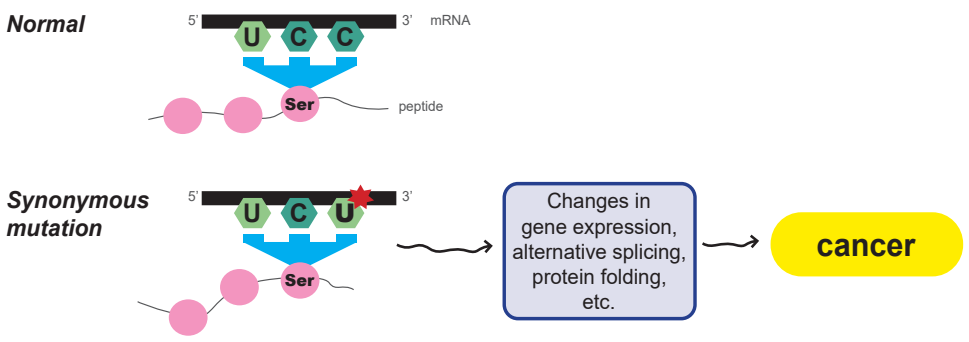
Why?

◊ The consensus molecular subtype classification (CMS) is used to classify colorectal cancer (CRC) into 4 subtypes, to provide better personalised treatment [1]. However, it is insufficient based on gene expression data alone.

◊ Synonymous mutations can alter cellular processes, and some are linked to cancer [2,3].

◊ We aim to improve the classification of CRC and to explore the role of synonymous mutations in CRC subtypes.

Synonymous mutations could be linked to colorectal cancer subtypes



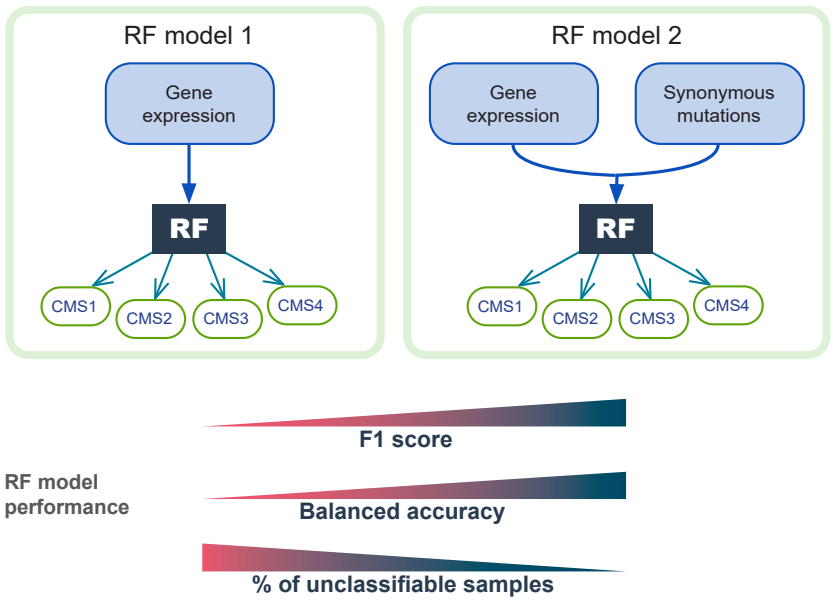
How?

◊ Aggregate synonymous mutation numbers: codon change, amino acid, genes

◊ Train two Random Forest (RF) models on gene expression with and without synonymous mutation data from CRC samples (n=351, TCGA).

◊ Validate by repeated testing* and applied t-test on the scores. Repeated testing: sampled from the test set 10K times with replacement, and calculated scores.

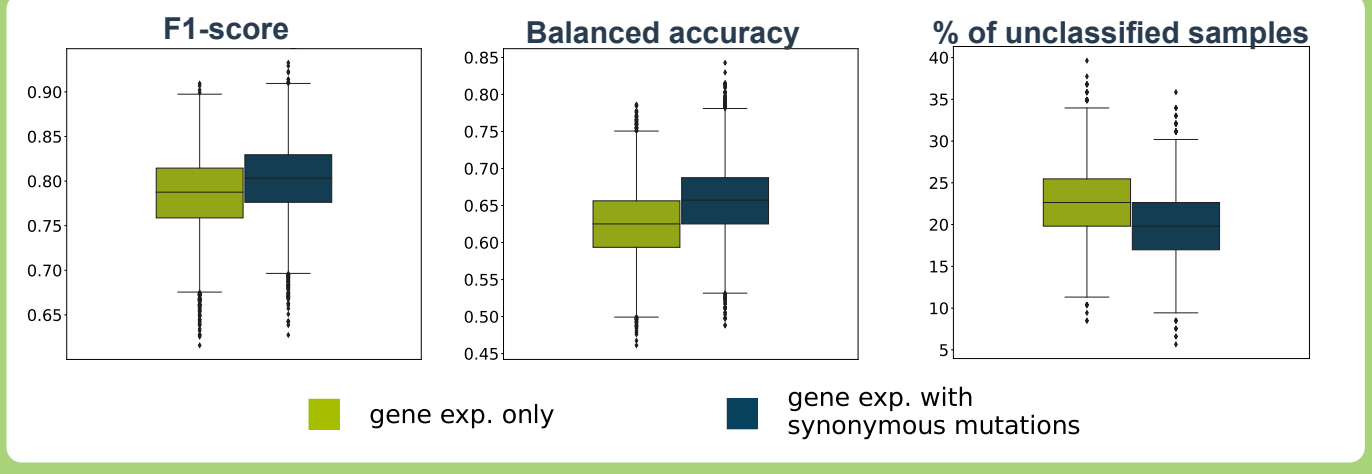
We found that synonymous mutation data improves the consensus molecular subtype classification of colorectal cancer



Adding synonymous mutation data improved the balanced accuracy, F1 score of the CMS classification and reduced the percentage of unclassified samples relative to gene expression alone.

What?

The RF model with synonymous mutation data resulted in increased balanced accuracy, F1 score and reduced percentage of unclassified samples relative to the model trained with gene expression data alone. These differences were significant (p-value < 0.001).



Conclusions

◊ Synonymous mutations showed signal in association with colorectal cancer subtypes

◊ Further research is needed: next we will train a deep learning model, to allow identification of complex patterns in the synonymous mutation data, and add interpretability.

References:
1. Guinney, J. et. al. (2015). The consensus molecular subtypes of colorectal cancer. Nat Med(21), 1350–1356.
2. Sharma, Y. et al. (2019). A pan-cancer analysis of synonymous mutations. Nat Commun(10), 2569.
3. Dimitrieva S, Anisimova M (2014) Unraveling Patterns of Site-to-Site Synonymous Rates Variation and Associated Gene Properties of Protein Domains and Families. PLoS ONE 9(6): e95034.