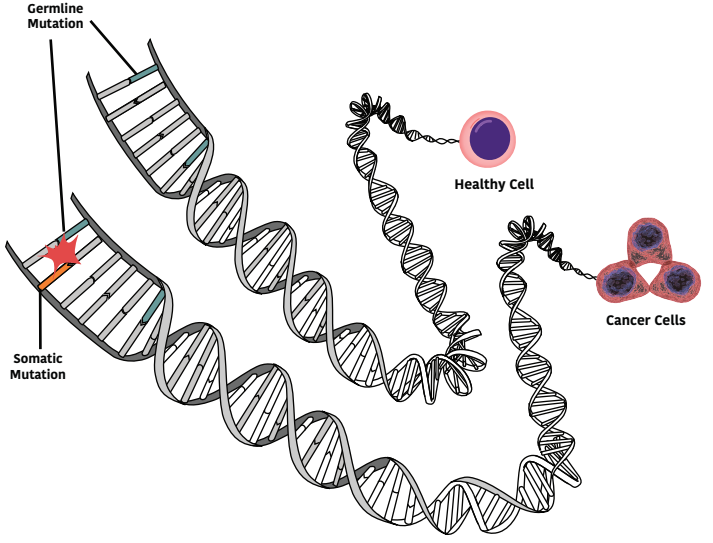
Deep neural networks to identify somatic-germline interactions in cancer

Hussein Mohsen[1,2], Sushant Kumar[1,2], Keniel Yao[3], Joanthan Warrell[1,2], Mark Gerstein[1,2,3,4]

Yale

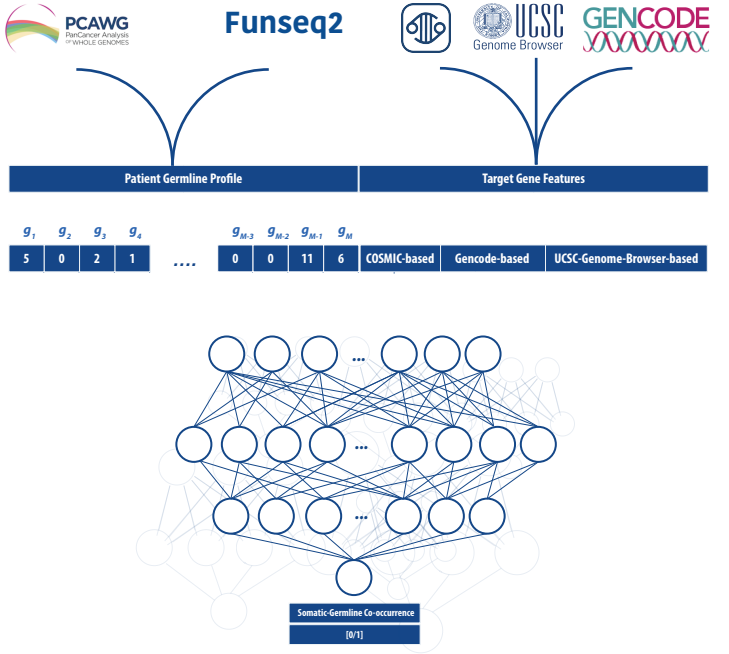
1 Computational Biology & Bioinformatics
2 Molecular Biophysics & Biochemistry
3 Computer Science
4 Statistics & Data Science

In cancer, complex interactions between alterations associated with multiple genes can lead to cooperative, more deleterious effects facilitating progression of the disease. Such epistatic relationships might also vary across cancer types. To this aim, we investigate the co-occurrence of rare somatic and germline variation (SGV) at pan-cancer and cancer-specific scales, a question that has never been investigated at the whole genome level [1, 2]. We train deep neural networks to predict somatic-germline variation co-occurrence at the gene level. Networks are trained on data from the 2800 cases in the PanCancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium (ICGC) consortium, and they are capable of identifying key biological features that would contribute to co-occurring SGV. Ongoing work will further unfold the genetic patterns of SGV co-occurrence and further prioritize genes involved in this type of variation.



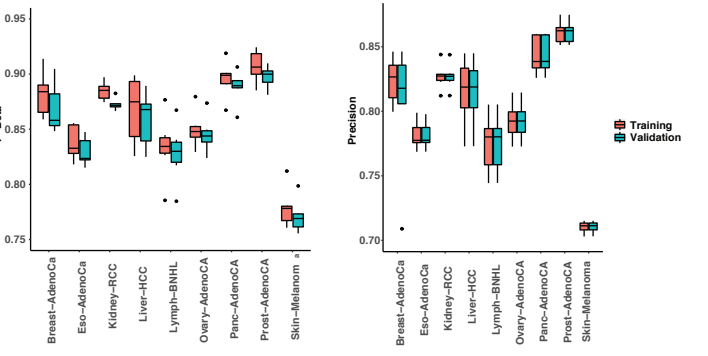
Data. PCAWG genomic variant data are used to train networks for our analysis. Somatic and germline rare variants (VAF < 0.05%) are singled out for 9 cancer types: Liver hepatocellular carcinoma, Skin Melanoma, B-cell Non-Hodgkins Lymphoma, and Pancreatic, Prostate, Breast, Ovary and Esophageal Adenocarcinomas, with SGV co-occurrence calculated at the gene level using Funseq2 annotation.

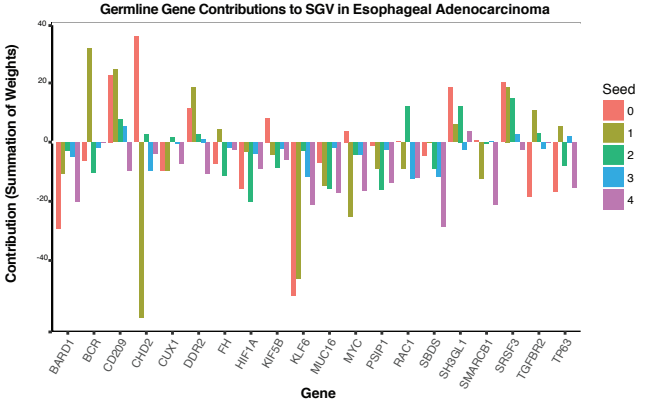
Training. The predictive task according to which the neural networks have been trained is the detection of somatic and germline variation co-occurrence at the gene level for the 718 COSMIC census genes fetched on May 08, 2018. Input data include 43 features ranging from germline variant signatures of known cancer genes alongside a set of biological features extracted from multiple data and annotation repositories, namely COSMIC, UCSC Genome Browser and Gencode v27, generating more than 850,000 data points. Automated hyperparameter optimization and rigorous data balancing are performed using HyperOpt [3] and Imbalanced-learn [4], respectively.



Cancer Type	Samples	COSMIC genes	Dataset Size	0-1 Label Imbalance (%)
Liver-HCC	309	679	209,811	98.53-1.57%
Panc-AdenoCA	230	619	142,370	98.86-1.14%
Prost-AdenoCA	197	463	91,211	99.30-0.70%
Breast-AdenoCa	189	636	120,204	98.72-1.28%
Kidney-RCC	140	578	80,920	98.68-1.32%
Ovary-AdenoCA	109	575	62,675	98.36-1.64%
Lymph-BNHL	104	561	58,344	98.00-2.00%
Eso-AdenoCa	94	605	56,870	97.37-2.63%
Skin-Melanoma	90	685	61,650	92.36-7.64%

Results. Trained networks across cancer types achieved precision values in [0.7, 0.86] and weighted harmonic mean of precision and recall, F-beta score, in [0.74-0.89] over 5 runs with random selection at the gene level for each cancer type. Preliminary feature prioritization led to cancer specific ranked lists of biological features and germline genes' contribution to SGV.





[1] Carter, H., Marty, R., and Hofree, M. et al. (2017), Cancer Discovery 7(4).
[2] Kanchi, K.L., Johnson, K.J., and McLellan, M.D. et al. (2014), Nat. Comm. 5:3156.
[3] Bergstra, J., Yamins, D., and Cox, D.D, SCIPY'2013.
[4] Lemaitre, G., Nogueira, F., and Aridas, C.K. (2017), JMLR.