Widespread redundancy in -omics profiles of cancer mutational states

Jake Crawford[1], Brock C. Christensen[2], Maria Chikina[3], Casey S. Greene[4,5]

[1]Genomics and Computational Biology (GCB) Graduate Group, Perelman School of Medicine, University of Pennsylvania
[2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College
[3]Department of Computational and Systems Biology, University of Pittsburgh School of Medicine
[4]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine
[5]Center for Health AI, University of Colorado School of Medicine

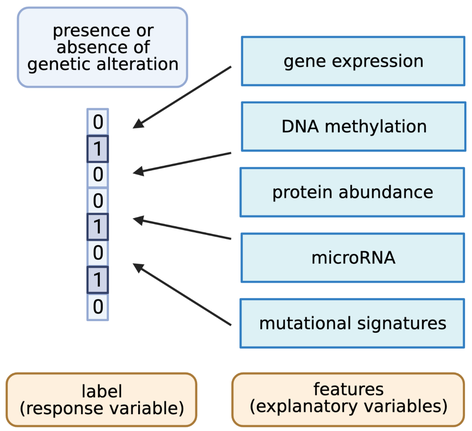
Overview

Research questions:

- Which -omics type captures the functional signatures of cancer mutations most effectively? Is this dependent on the gene(s) that are mutated?
- Does combining multiple -omics types improve detection?

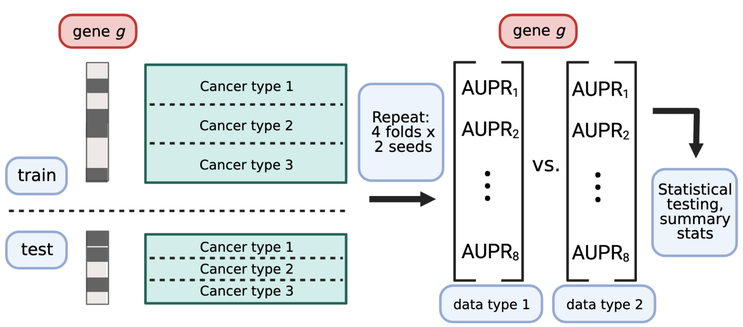
Framing as a prediction problem:

We want to predict cancer mutation presence or absence using -omics data available in the TCGA Pan-Cancer Atlas: gene expression, DNA methylation, reverse phase protein array (RPPA), microRNA, mutational signatures.



Approach

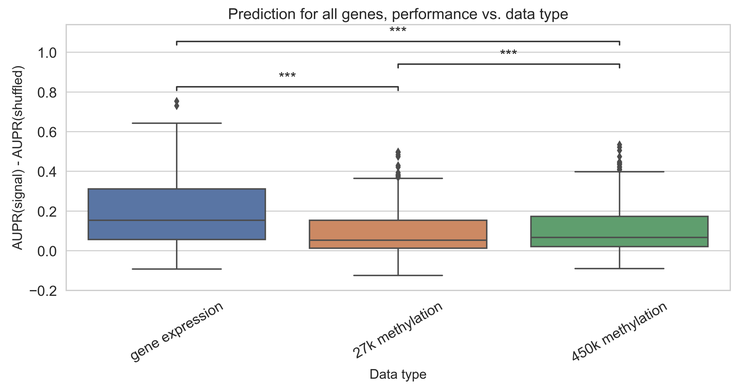
- 268 cancer-related genes collected from existing surveys[1–3]
- Pan-cancer model for each gene
- Elastic net logistic regression (+ 3-layer neural network for multi-omics)
- 2 replicates (random seeds) x 4-fold CV, stratified by cancer type
- Compare classifiers against baseline with permuted labels, and compare directly between data types



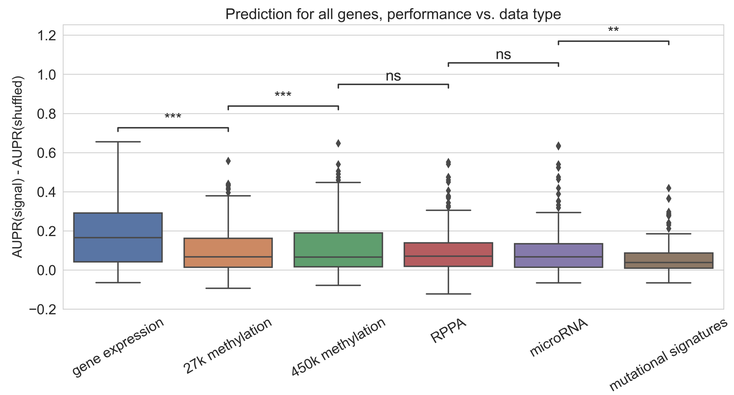
Figures above created with BioRender.com

Results

On aggregate over the cancer-associated gene set, gene expression is a slightly more effective predictor than the methylation arrays (Illumina 27K/450K merged and Illumina 450K).

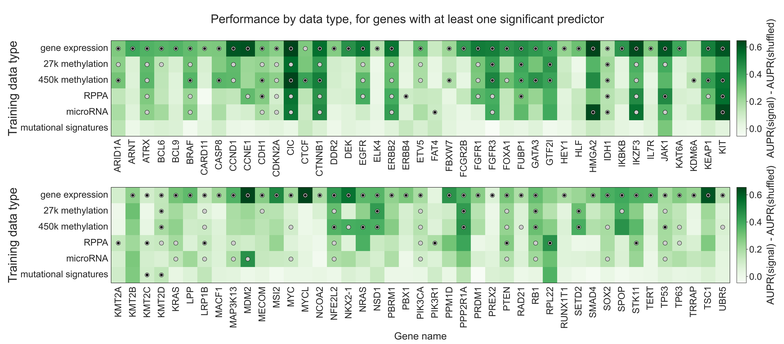


When we compare all data types using all cancer genes, the gene expression dataset significantly outperforms the remaining data types.



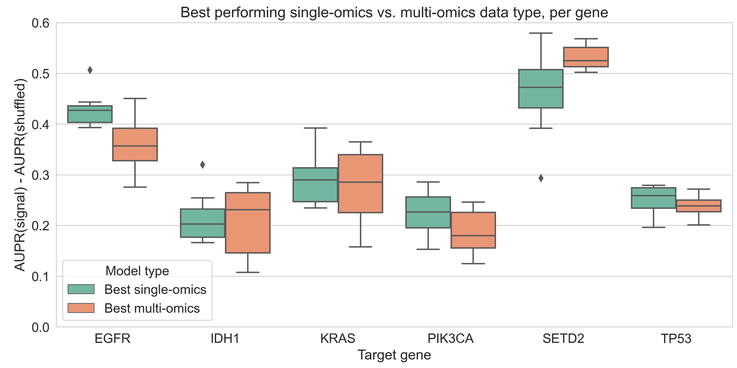
Of 86 genes that are "well-predicted" using $\geq$ one data type, 52/86 (60.5%) are well-predicted by multiple data types (vertical "stripes" in heatmap).

28/34 (82.4%) of the remaining genes are best predicted by models using gene expression.



We also built multi-omics models by concatenating combinations of the expression and methylation datasets. For each data type, we used the top 5000 principal components as predictive features.

Using six pan-cancer driver genes as targets, none of the multi-omics models significantly outperformed the best-performing single-omics model.



Conclusions

Main takeaways:

- On average, gene expression is the most effective functional readout.
- However, strong cancer drivers tend to perturb ***most or all data types***, resulting in detectable functional signatures.
- Multi-omics models do not tend to outperform their single-omics counterparts, suggesting the existence of redundant information across data types.

Why is this significant?

We anticipate that these results will be useful in ***study design:*** for most genes, multiple readouts produce similarly effective models.

In our paper (link below) we provide a table which can be used by cancer researchers to identify effective readouts for genes of interest.

Where can I learn more?

Data and code:

https://github.com/greenelab/mpmp

Paper (published in *Genome Biology*):

https://doi.org/10.1186/s13059-022-02705-y

Link to this poster:

http://jjc2718.github.io/ismb_2022_poster.pdf

References

[1] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[2] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, *et al.*, "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, no. 2, pp. 371–385, 2018.

[3] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, no. 11, pp. 696–705, 2018.