Widespread redundancy in -omics profiles of cancer mutation states

ISMB 2022: General Computational Biology track

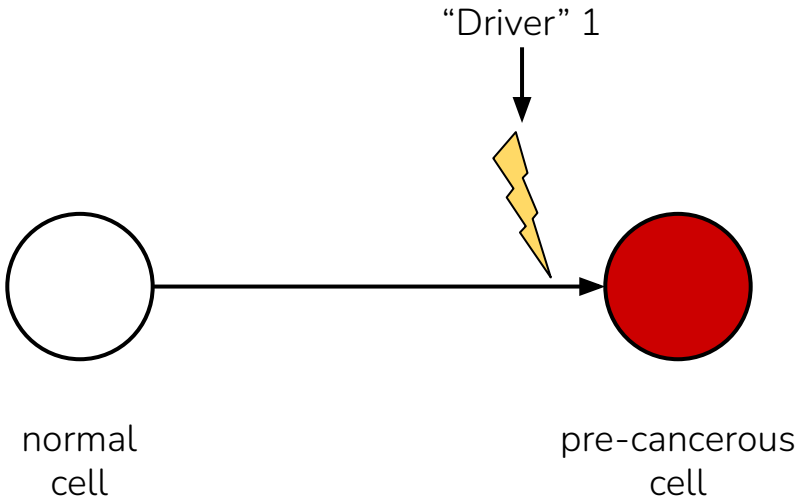
Jake Crawford
PhD Candidate, Greene Lab
University of Pennsylvania

Web: **http://jjc2718.github.io**
Twitter/GitHub/Gmail: **@jjc2718**

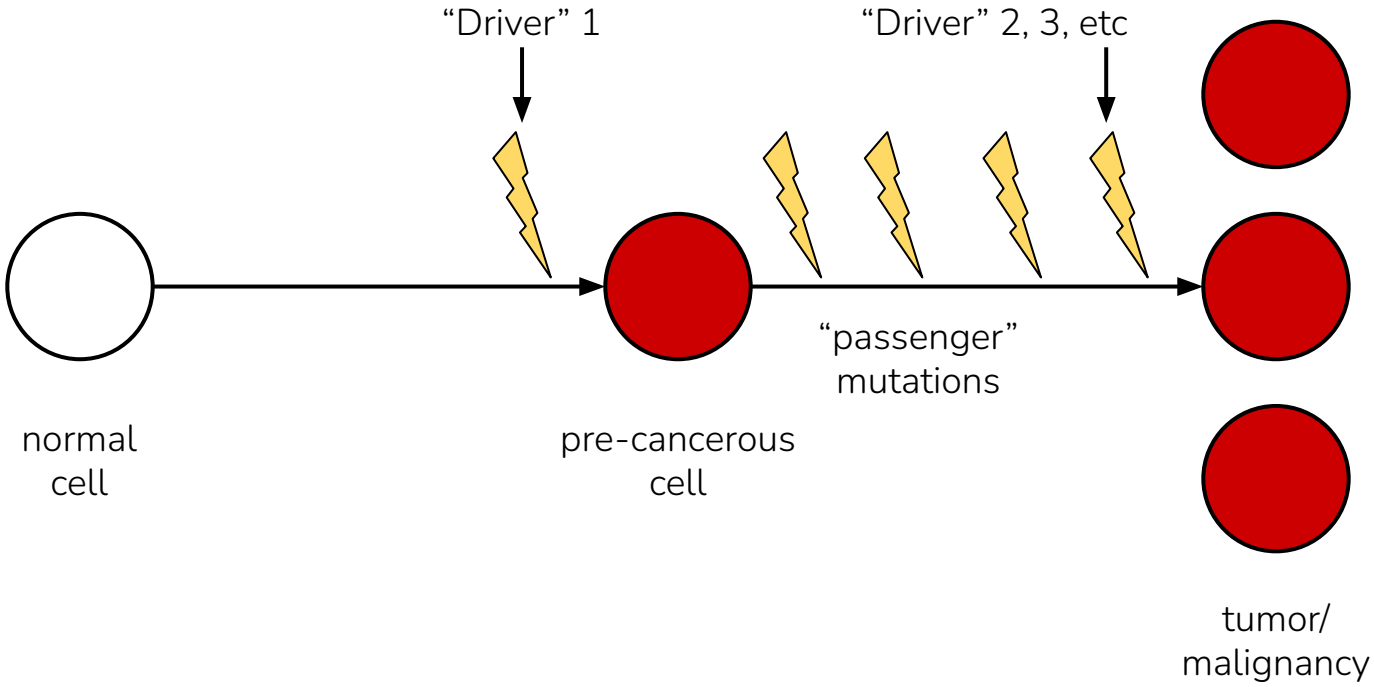
- **Background:**
  - Functional -omics in cancer
  - What are we predicting? Why are we predicting it?

- **Current work:**
  - Which cancer -omics is the best -omics? (for our problem)
  - Survival prediction
  - Multi-omics models

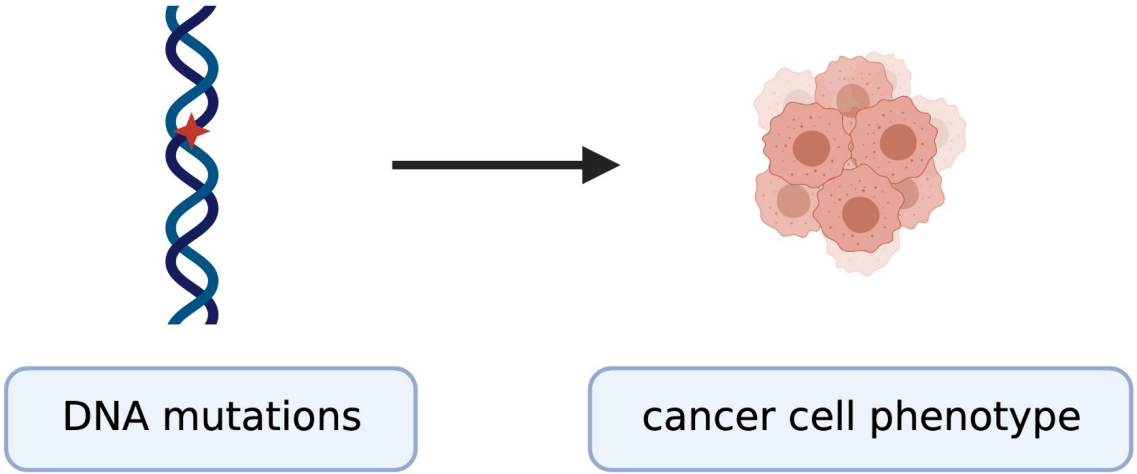
Genetic alterations in cancer

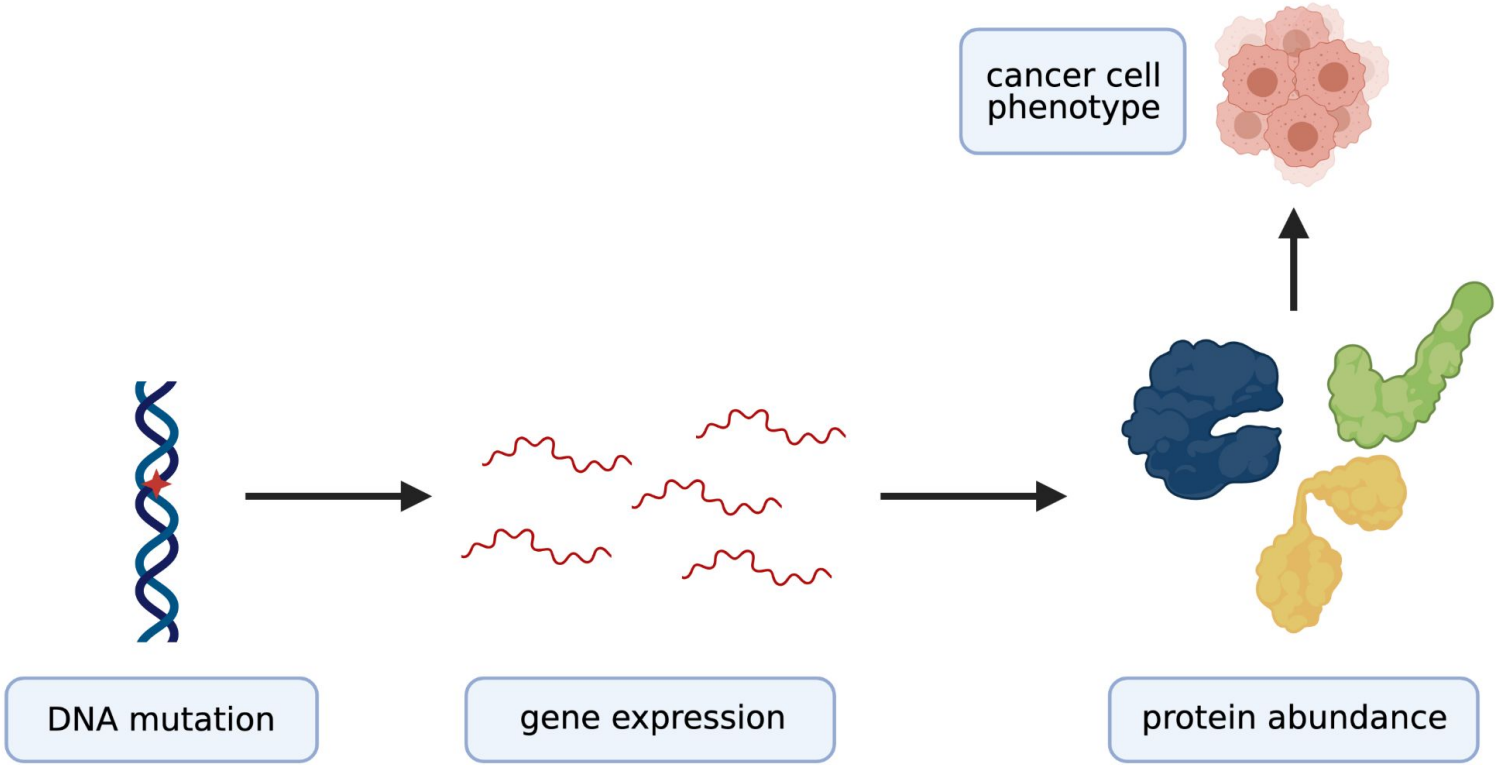


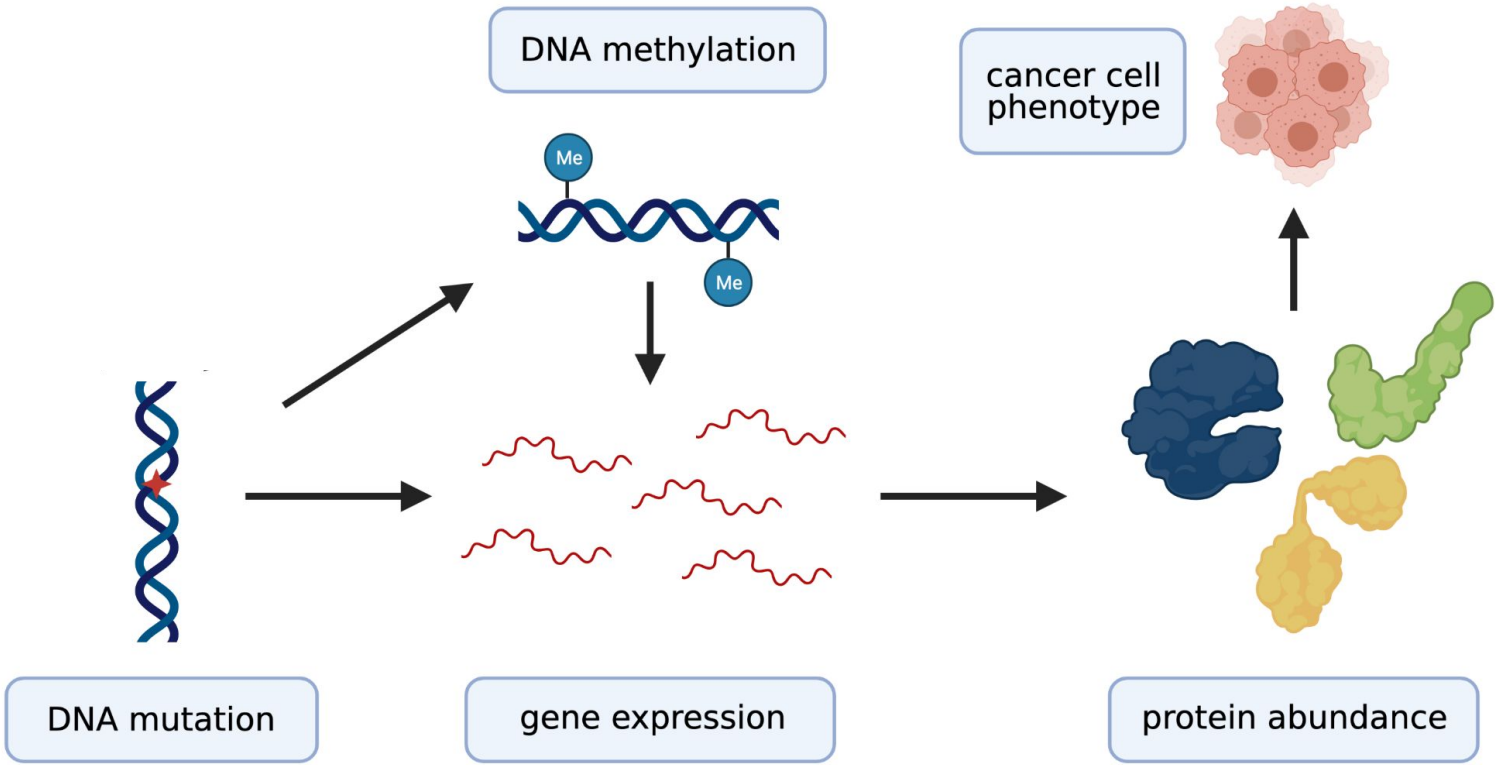
Genetic alterations in cancer

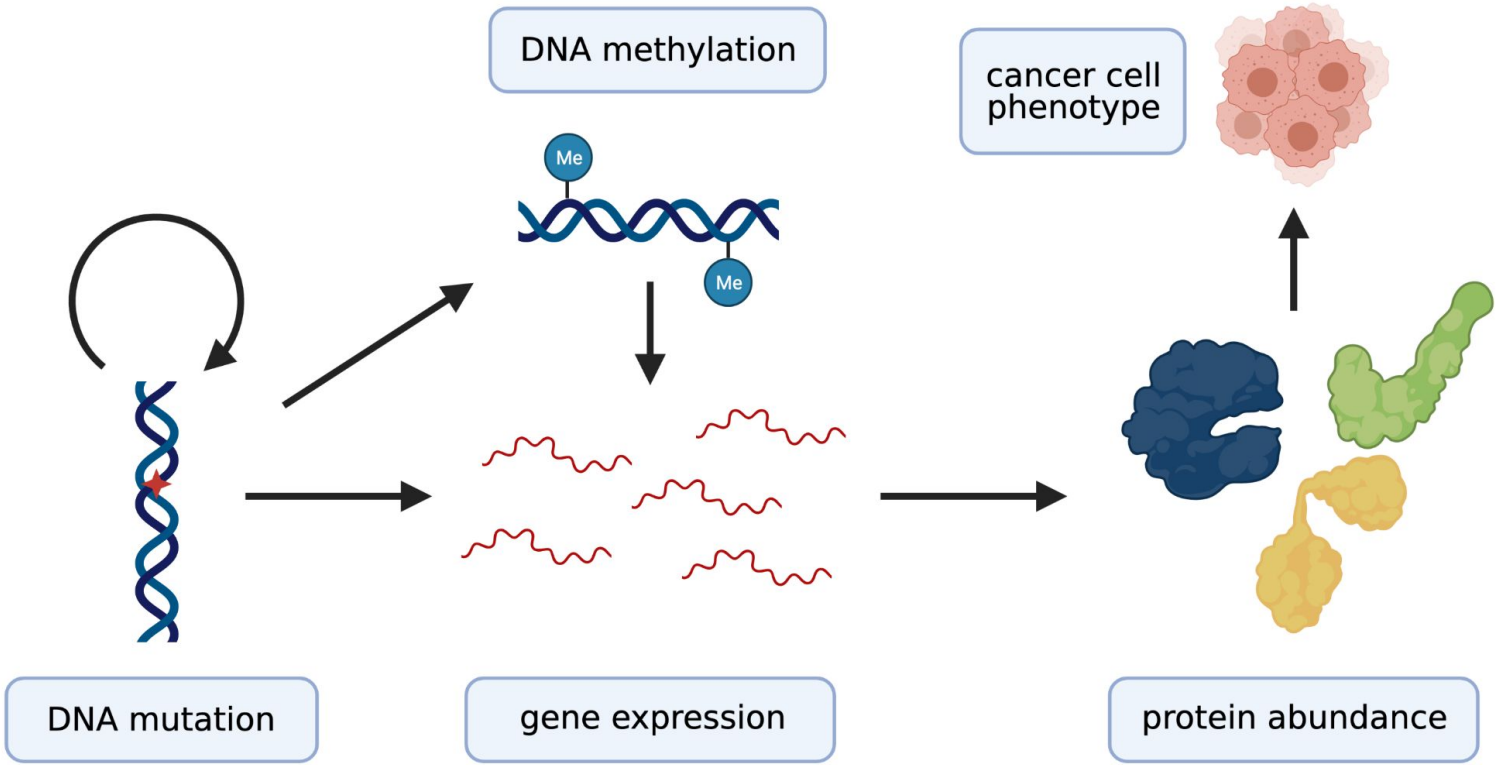


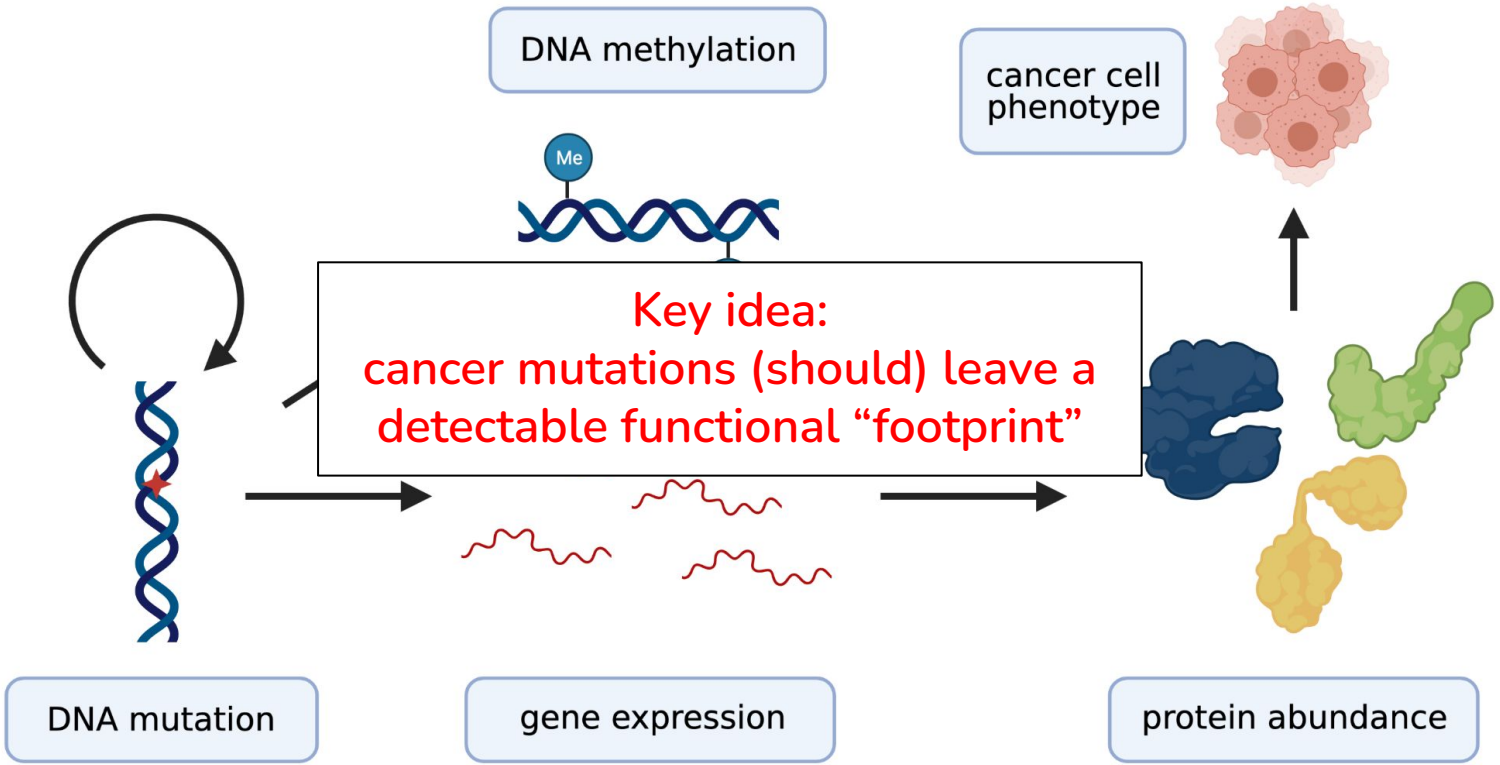
Functional -omics in cancer



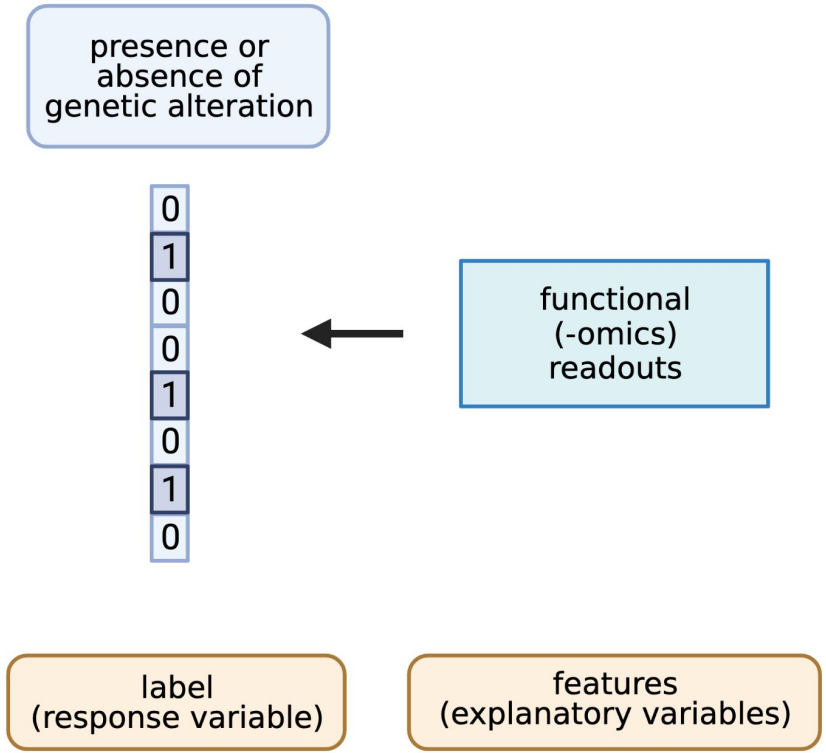

cancer cell phenotype
DNA mutation
gene expression
protein abundance

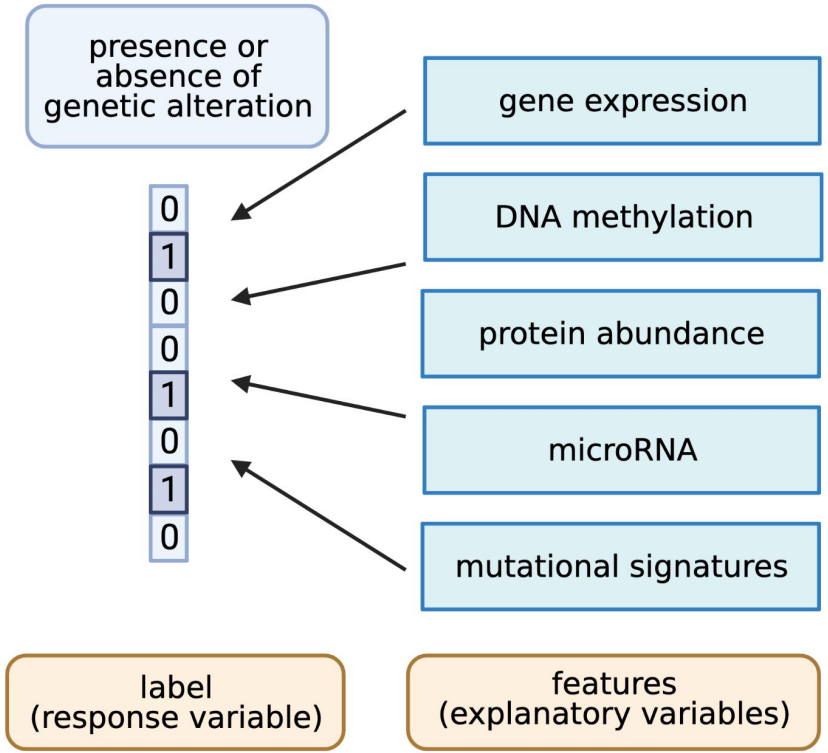

DNA methylation
Me
Me
cancer cell phenotype
DNA mutation
gene expression
protein abundance

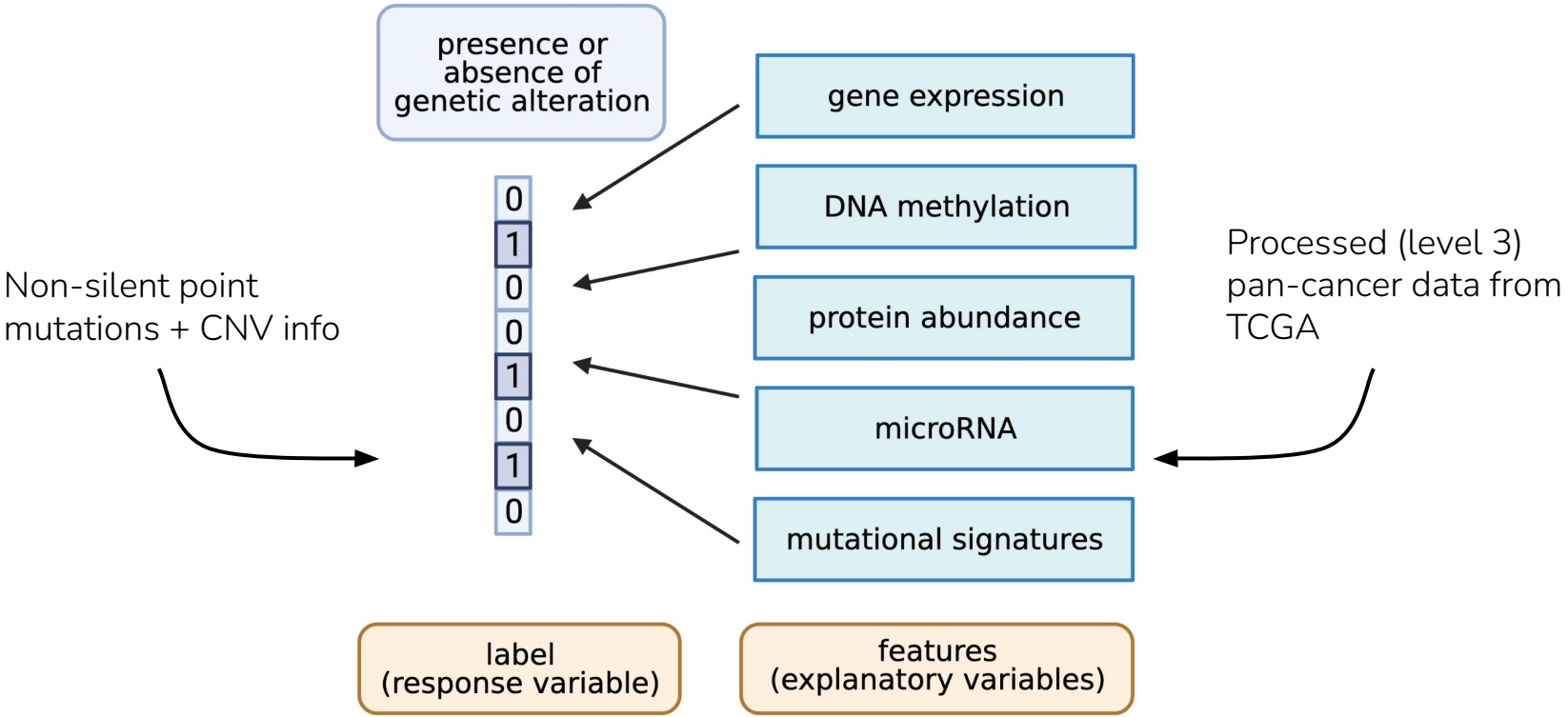

DNA methylation
Me
Me
cancer cell phenotype
DNA mutation
gene expression
protein abundance

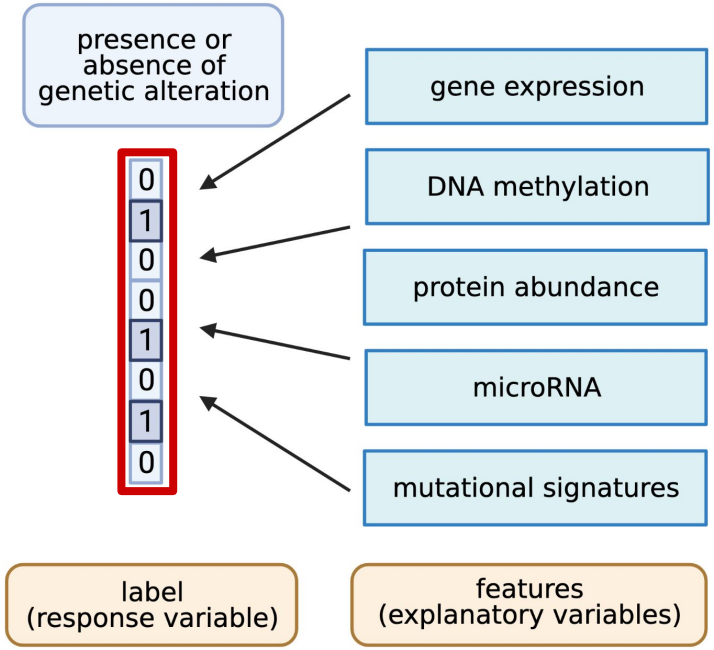

DNA methylation
cancer cell phenotype
Me
Key idea:
cancer mutations (should) leave a detectable functional “footprint”
DNA mutation
gene expression
protein abundance

Predicting mutation status from -omics data



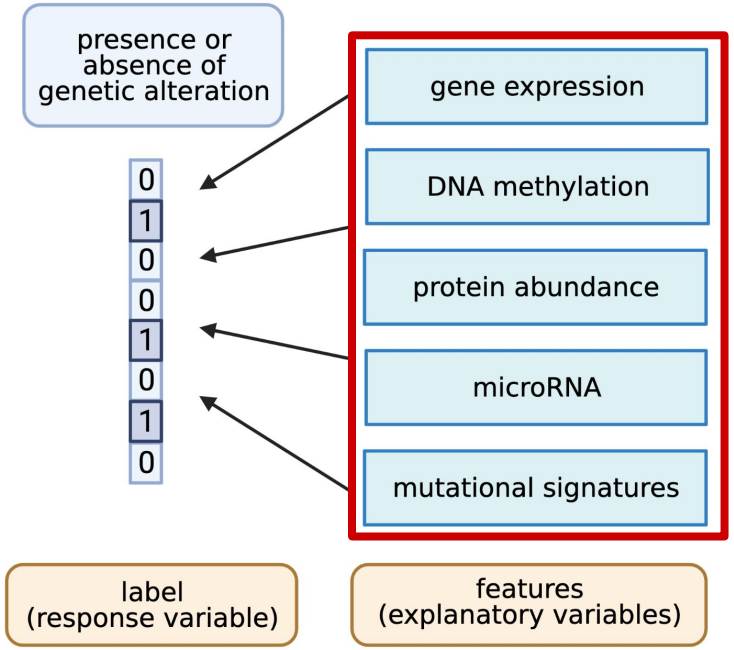

presence or absence of genetic alteration
0
1
0
0
1
0
1
0
gene expression
DNA methylation
protein abundance
microRNA
mutational signatures
label (response variable)
features (explanatory variables)


presence or absence of genetic alteration
0
1
0
0
1
0
1
0
gene expression
DNA methylation
protein abundance
microRNA
mutational signatures
Non-silent point mutations + CNV info
Processed (level 3) pan-cancer data from TCGA
label (response variable)
features (explanatory variables)



Can we detect a functional signature...

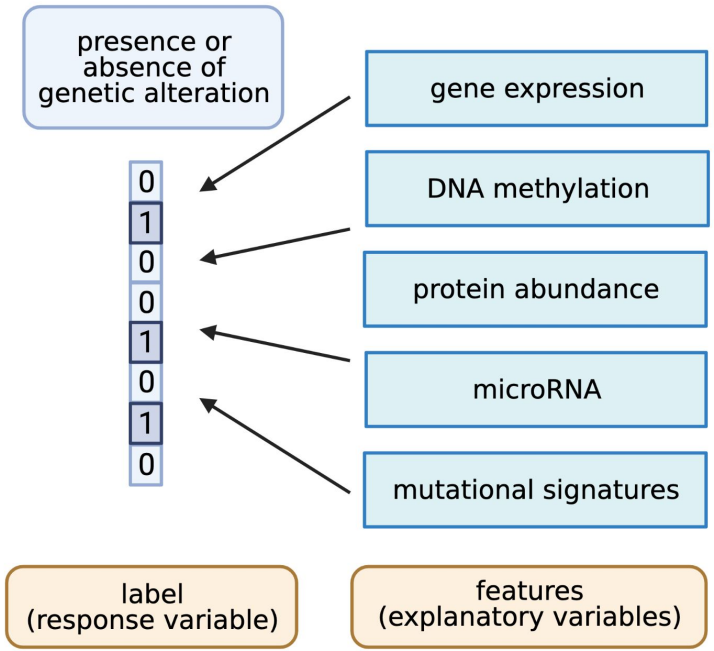
- for this mutation?



Can we detect a functional signature...

- for this mutation?
- for this data type?

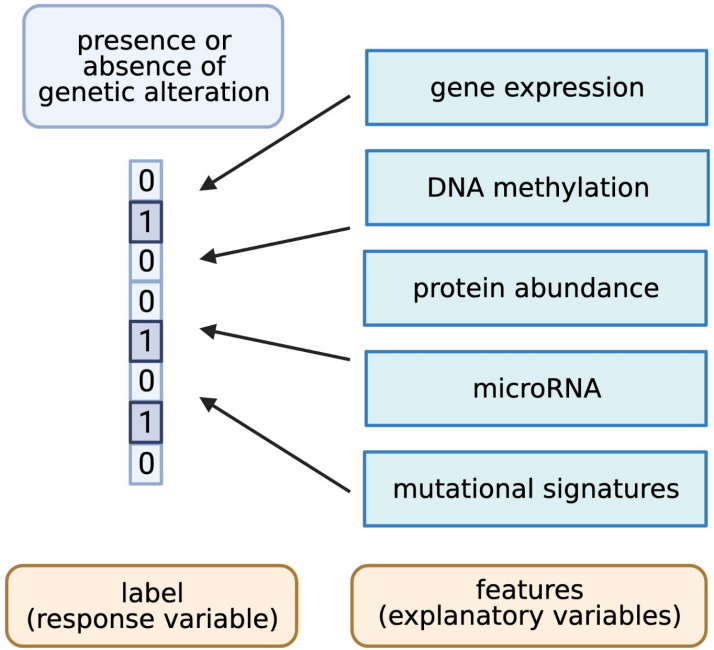
But why?



- Benchmark for cancer-related information content

  Question: Does data source or mutated gene contribute most to predictive signal?

But why?



- Benchmark for cancer-related information content

  Question: Does data source or mutated gene contribute most to predictive signal?

- Study design

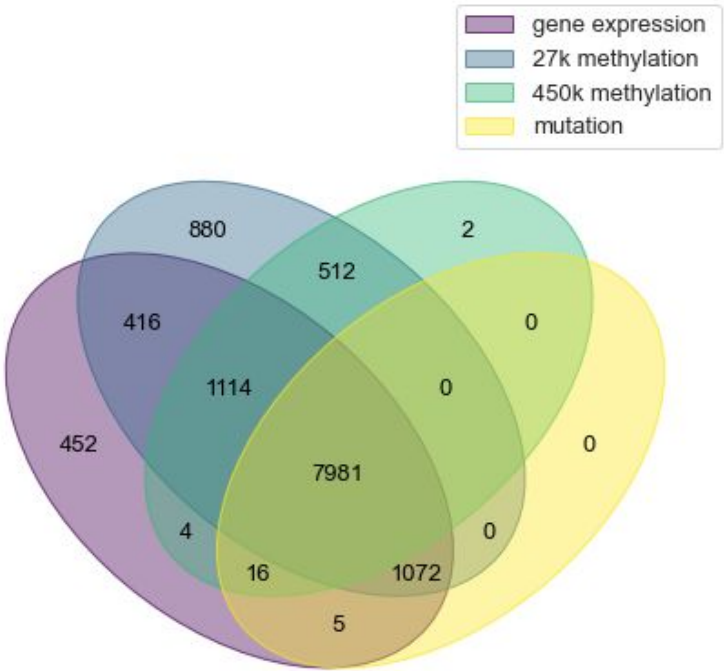
  Question: When are data sources redundant? When might they provide unique information?

Performance comparison

Which data type is best for mutation prediction?

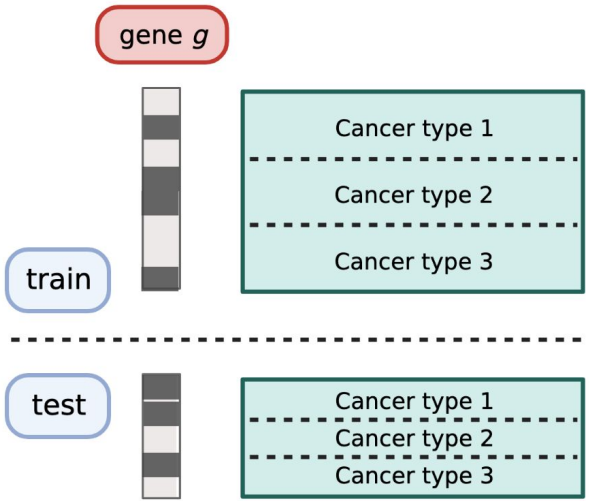
Gene expression vs. DNA methylation

- Gene set collected from [1], [2], [3] (~280 cancer-related genes)
- RNA-seq, 27K and 450K methylation array data from TCGA
- Only samples with all three data types measured -> **7,981** total samples

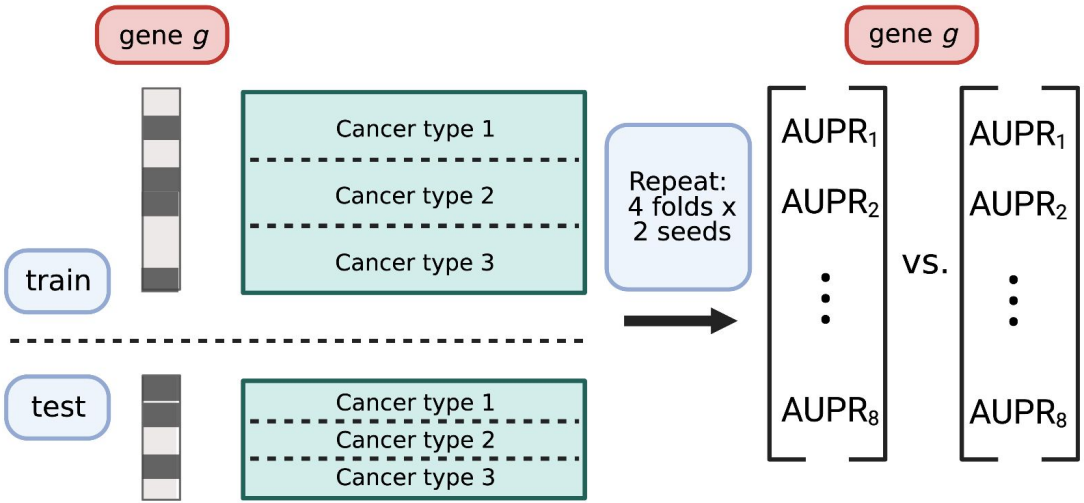


[1] Vogelstein et al. *Science* 2013, [2] Bailey et al. *Cell* 2018, [3] COSMIC Cancer Gene Census

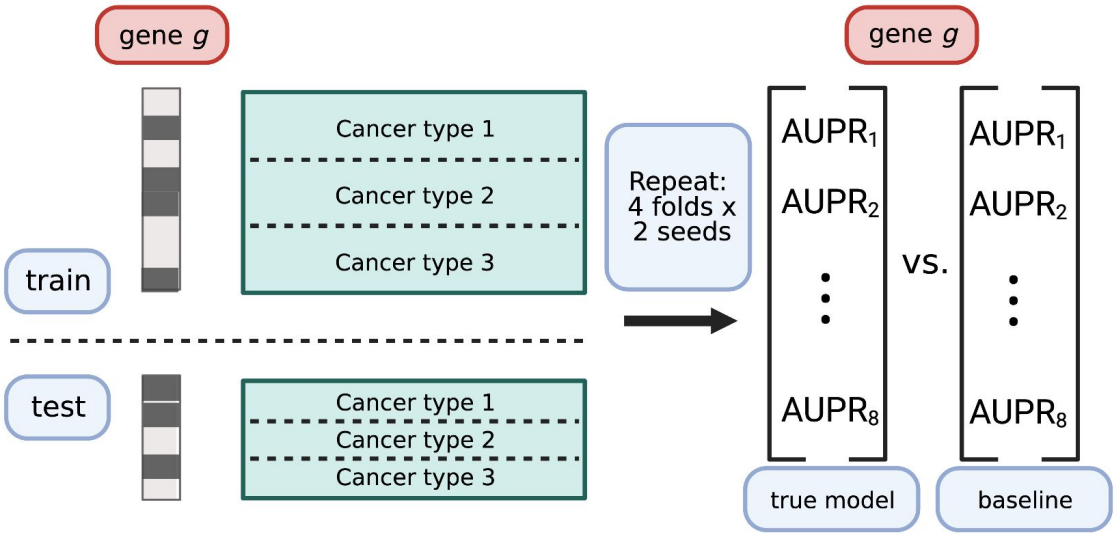
Performance evaluation



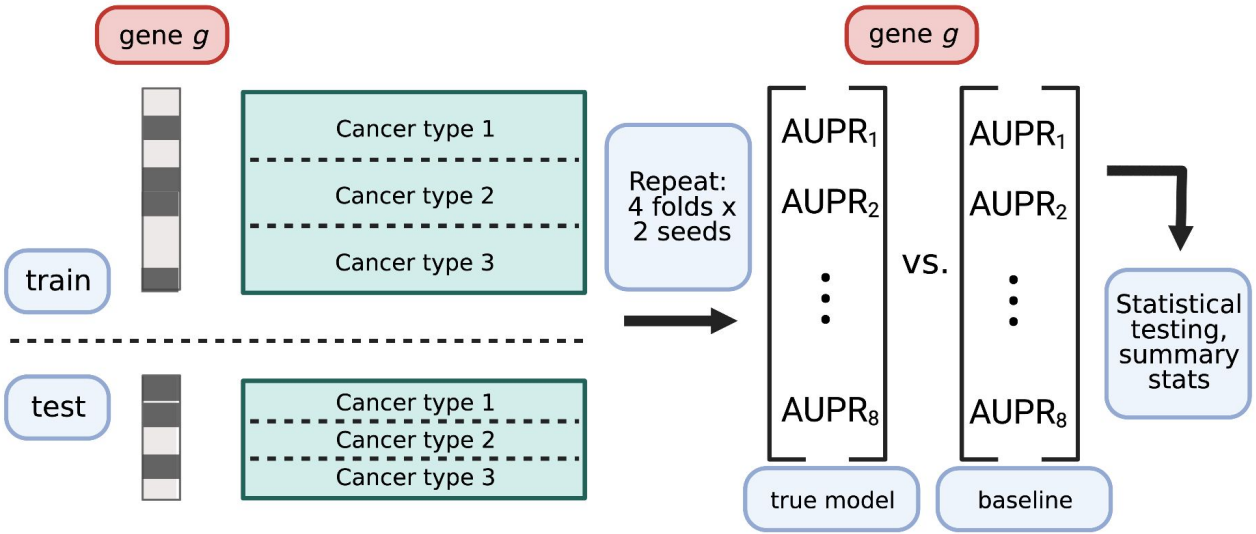
4 folds x 2 random seeds cross-validation,
stratified by cancer type



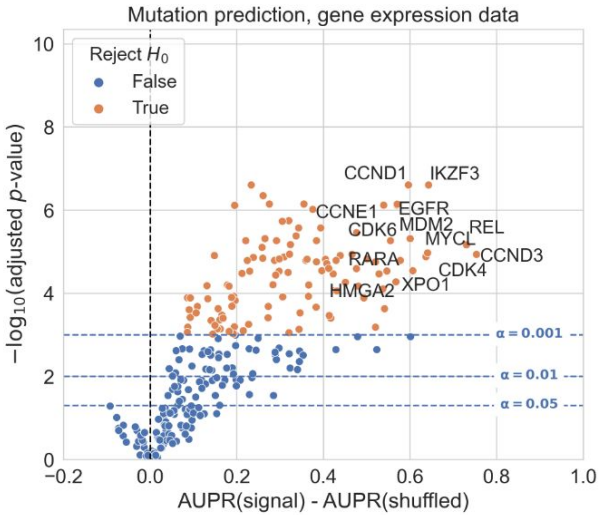
Metric: AUPR
(handles continuous output,
works well for imbalanced labels)

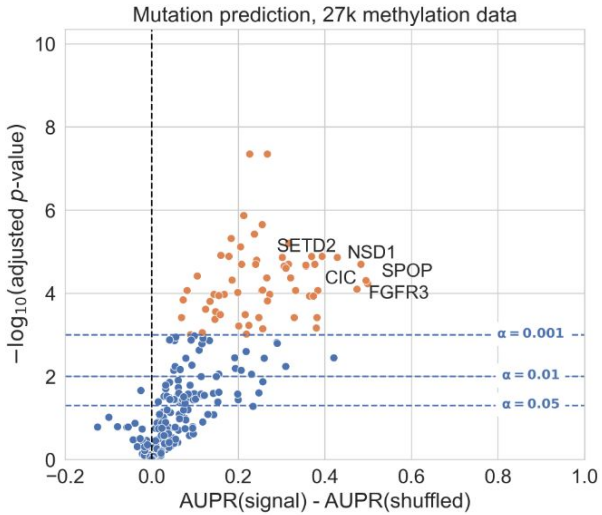


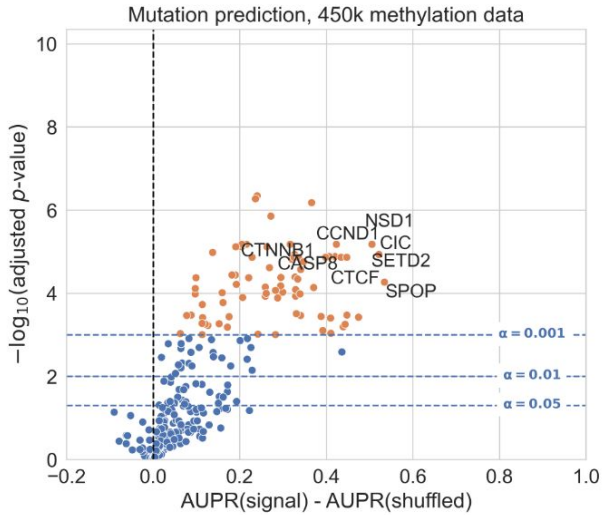
Baseline: model with permuted mutation labels (controls for cancer type)


gene g
train
test
Cancer type 1
Cancer type 2
Cancer type 3
Cancer type 1
Cancer type 2
Cancer type 3
Repeat:
4 folds x
2 seeds
gene g
AUPR1
AUPR2
AUPR8
vs.
AUPR1
AUPR2
AUPR8
true model
baseline
Statistical
testing,
summary
stats

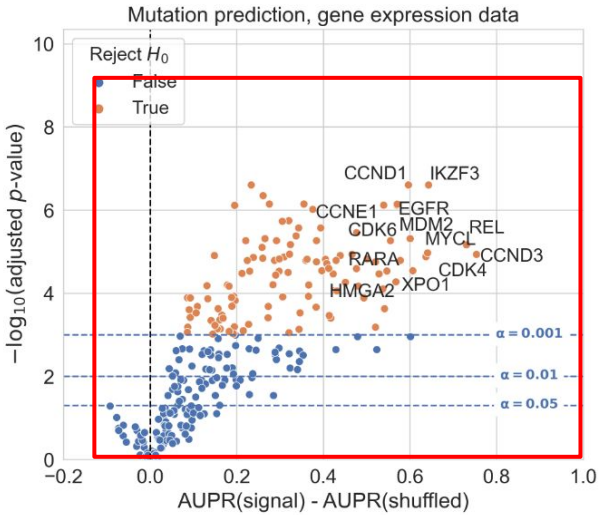
Which classifiers beat the baseline?

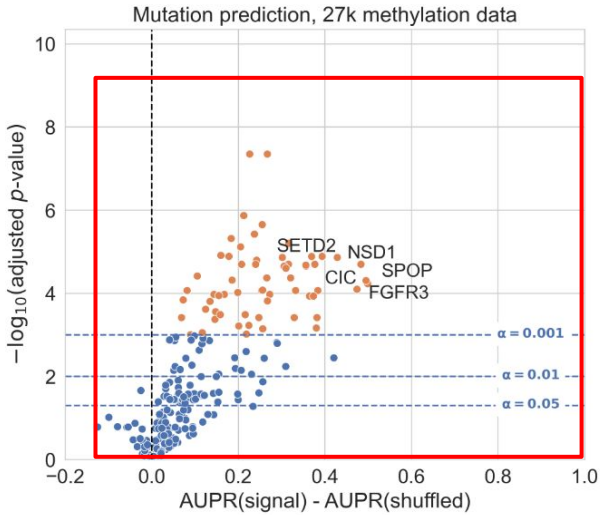


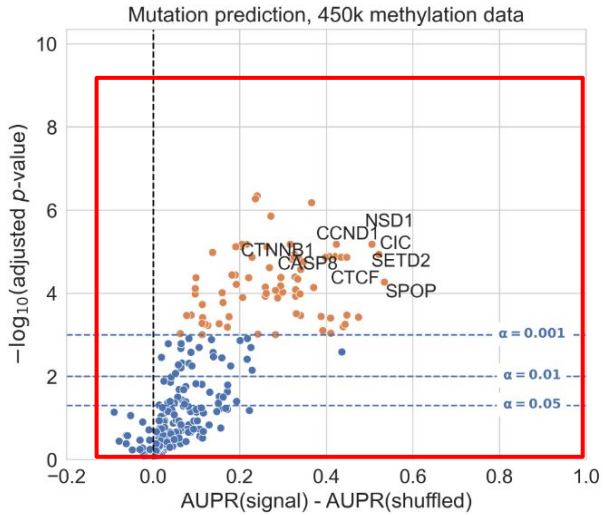


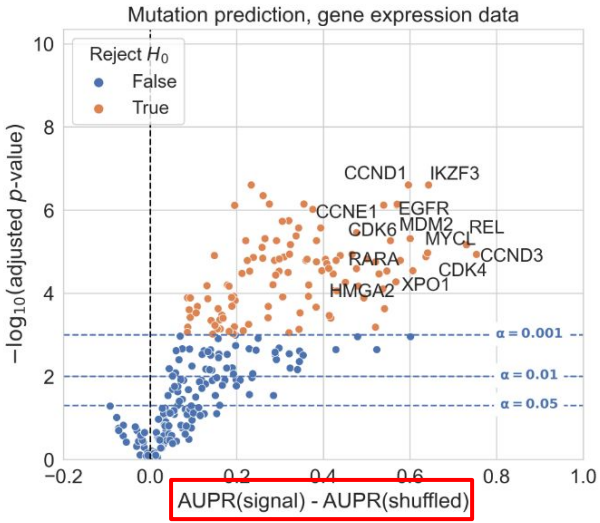


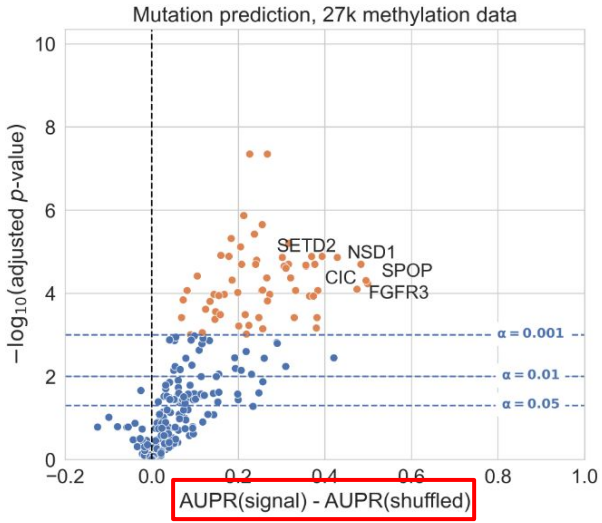
Each point = results for one gene in one data type

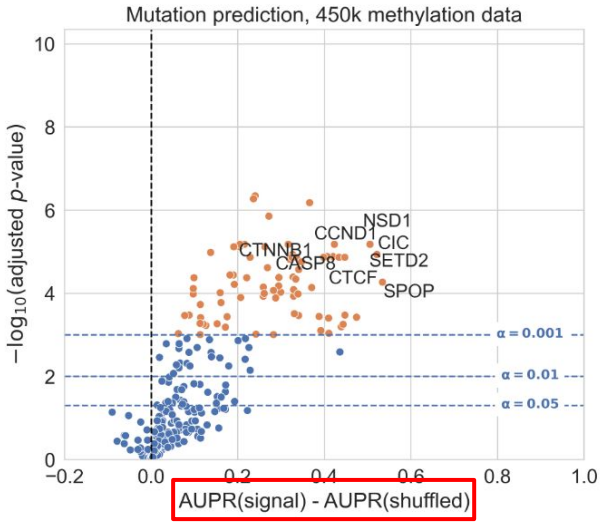




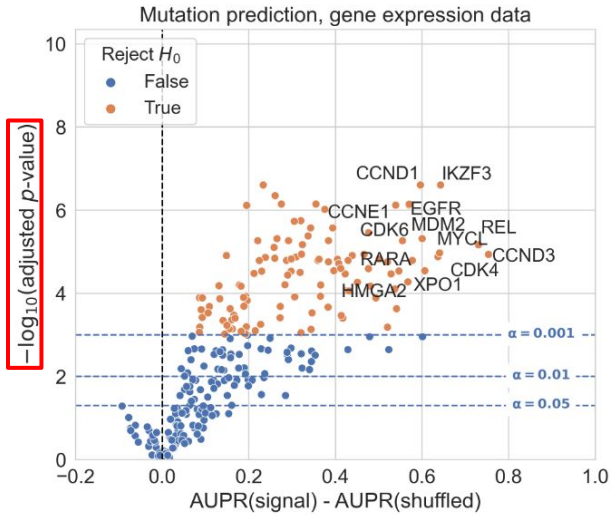


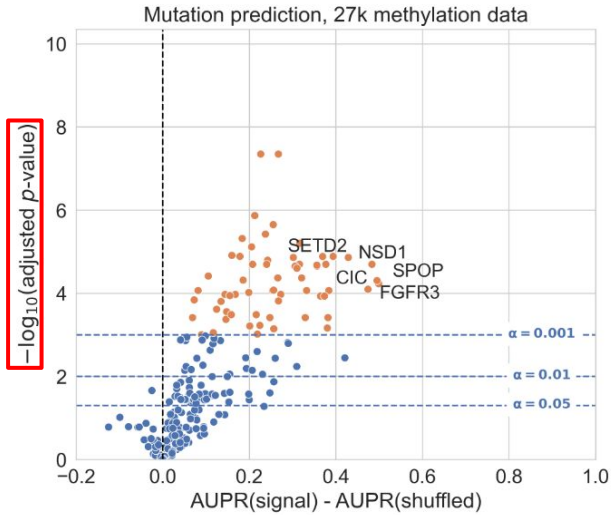


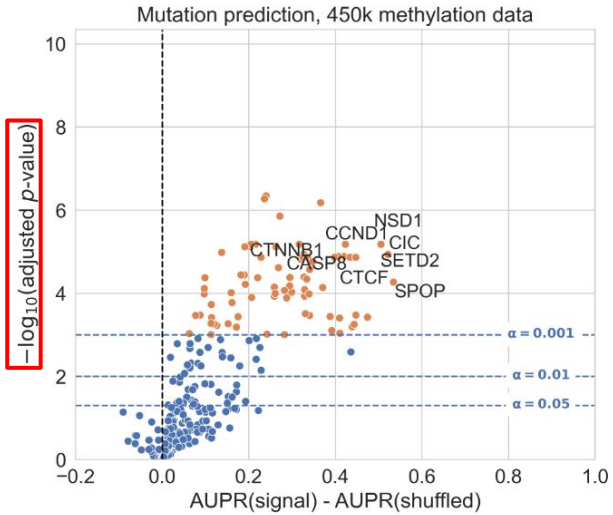




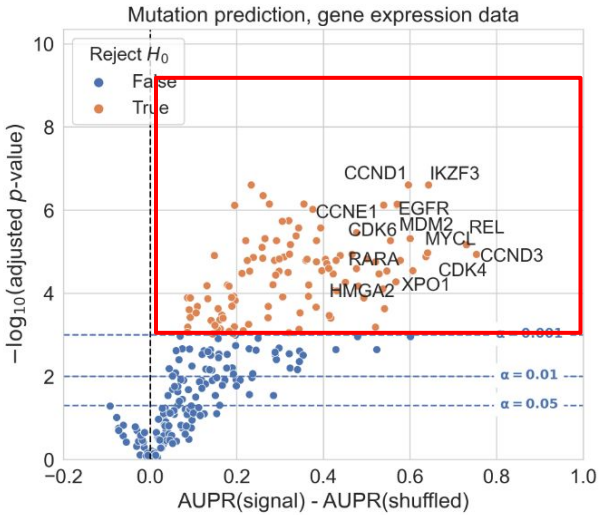
mean difference between true model and baseline with permuted labels



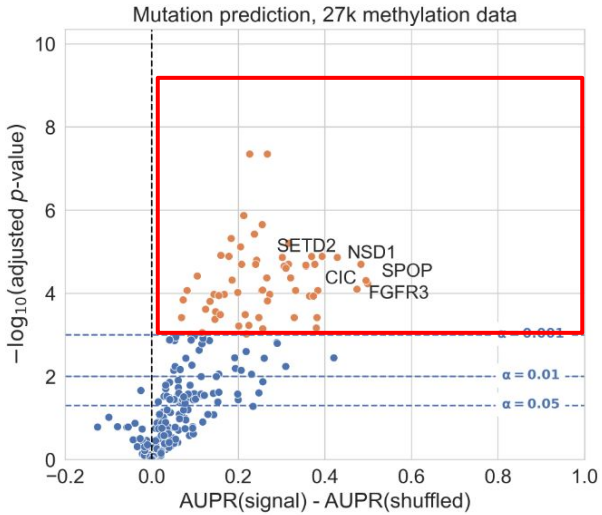




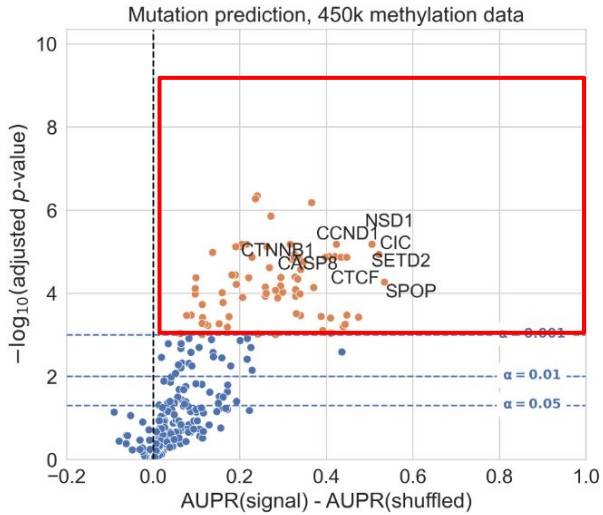
***t*-test *p*-value, across 4 cross-validation folds x 2 random seeds**



113/272 genes

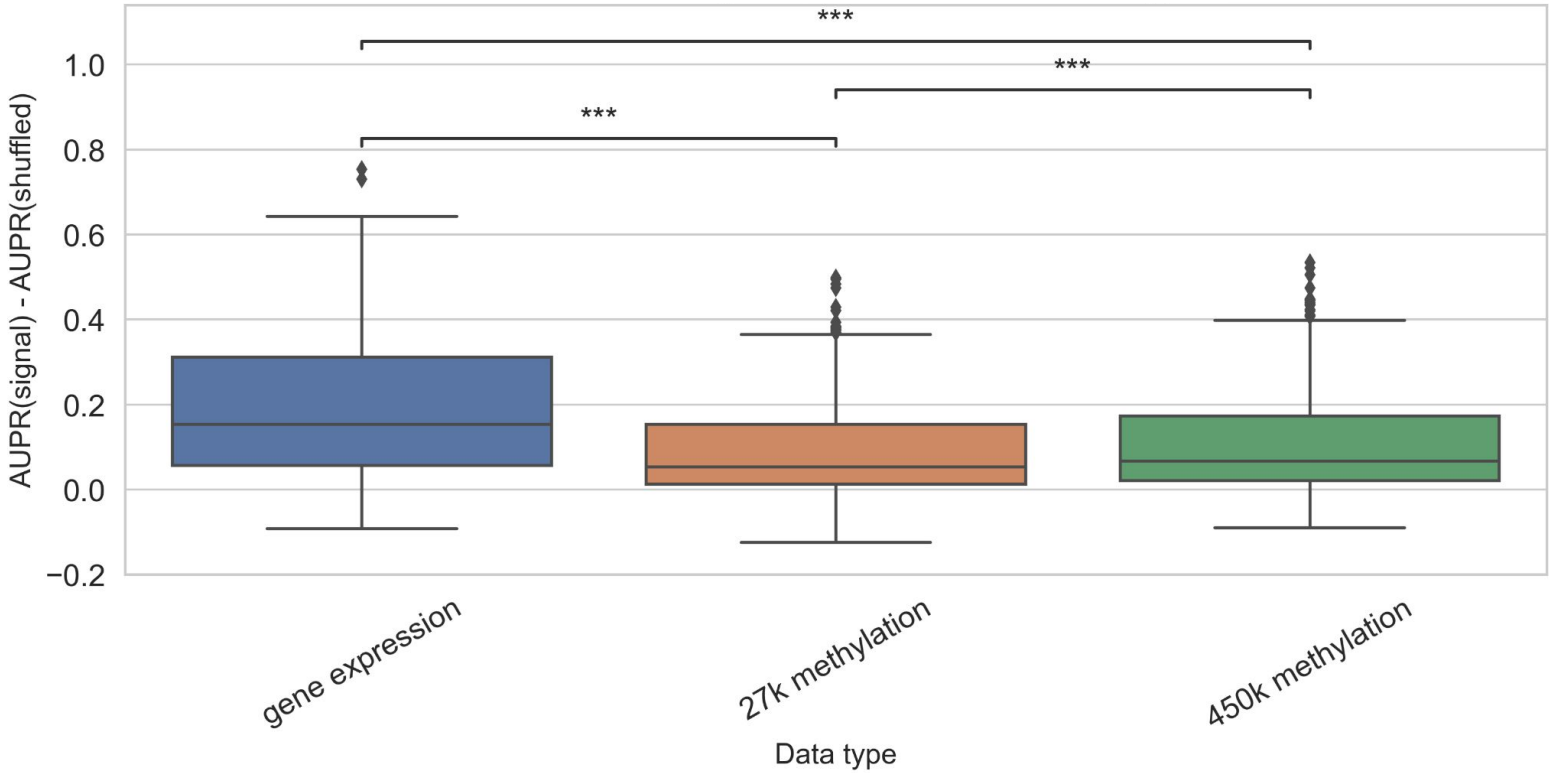


62/272 genes



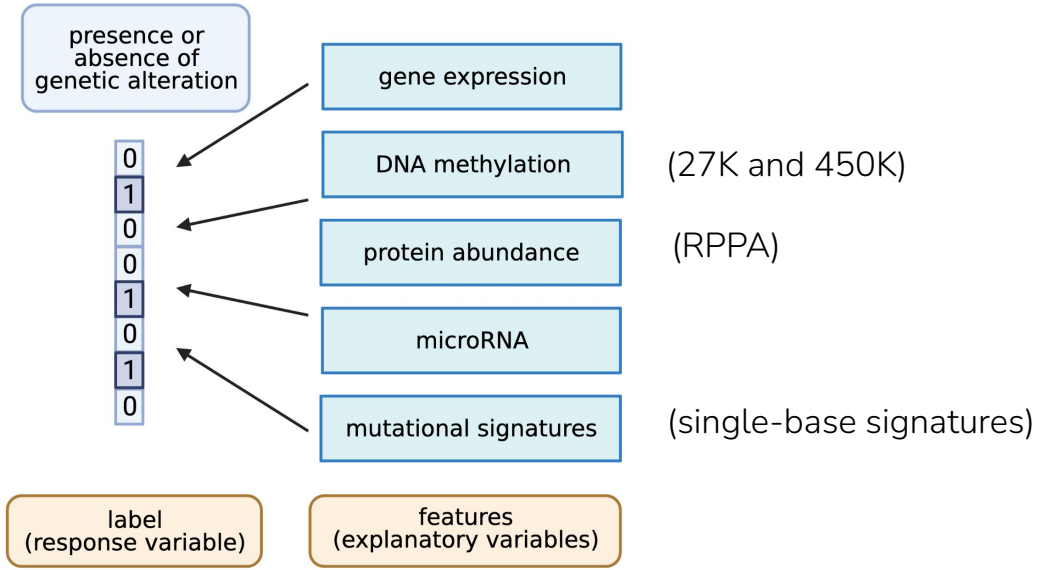
77/272 genes

Expression/methylation summary across genes



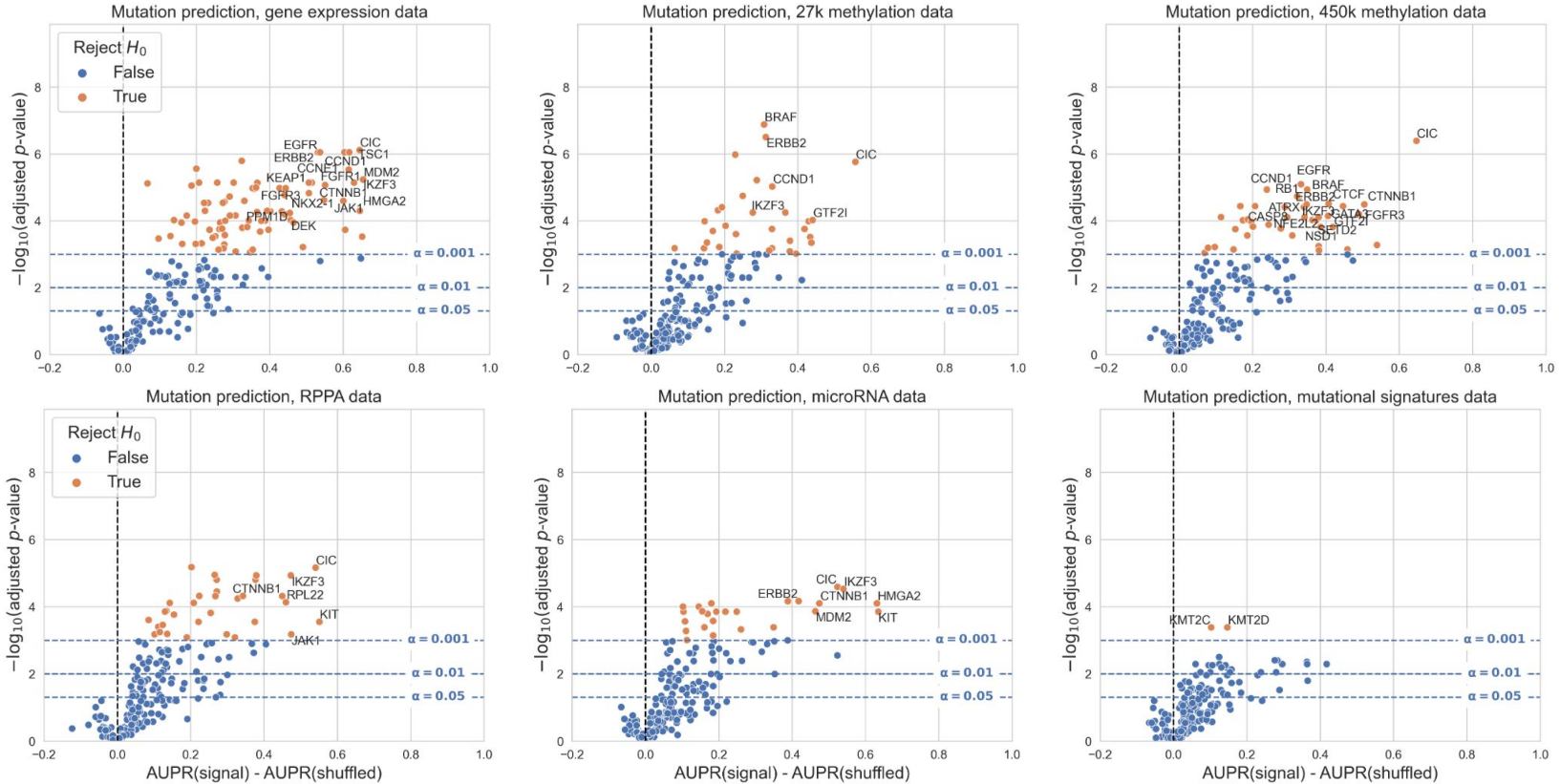
Comparing more -omics data types

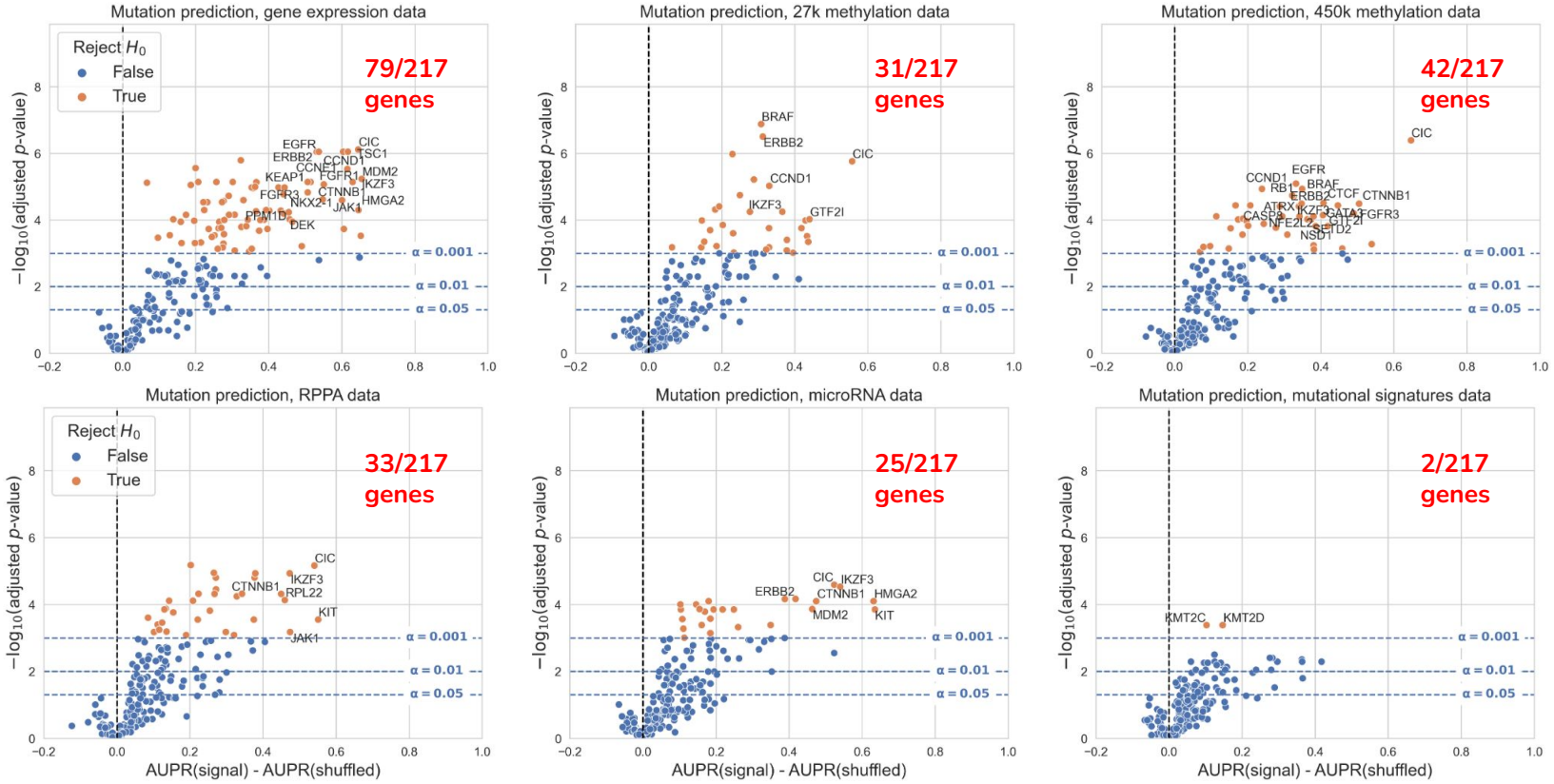
- Data types from TCGA:



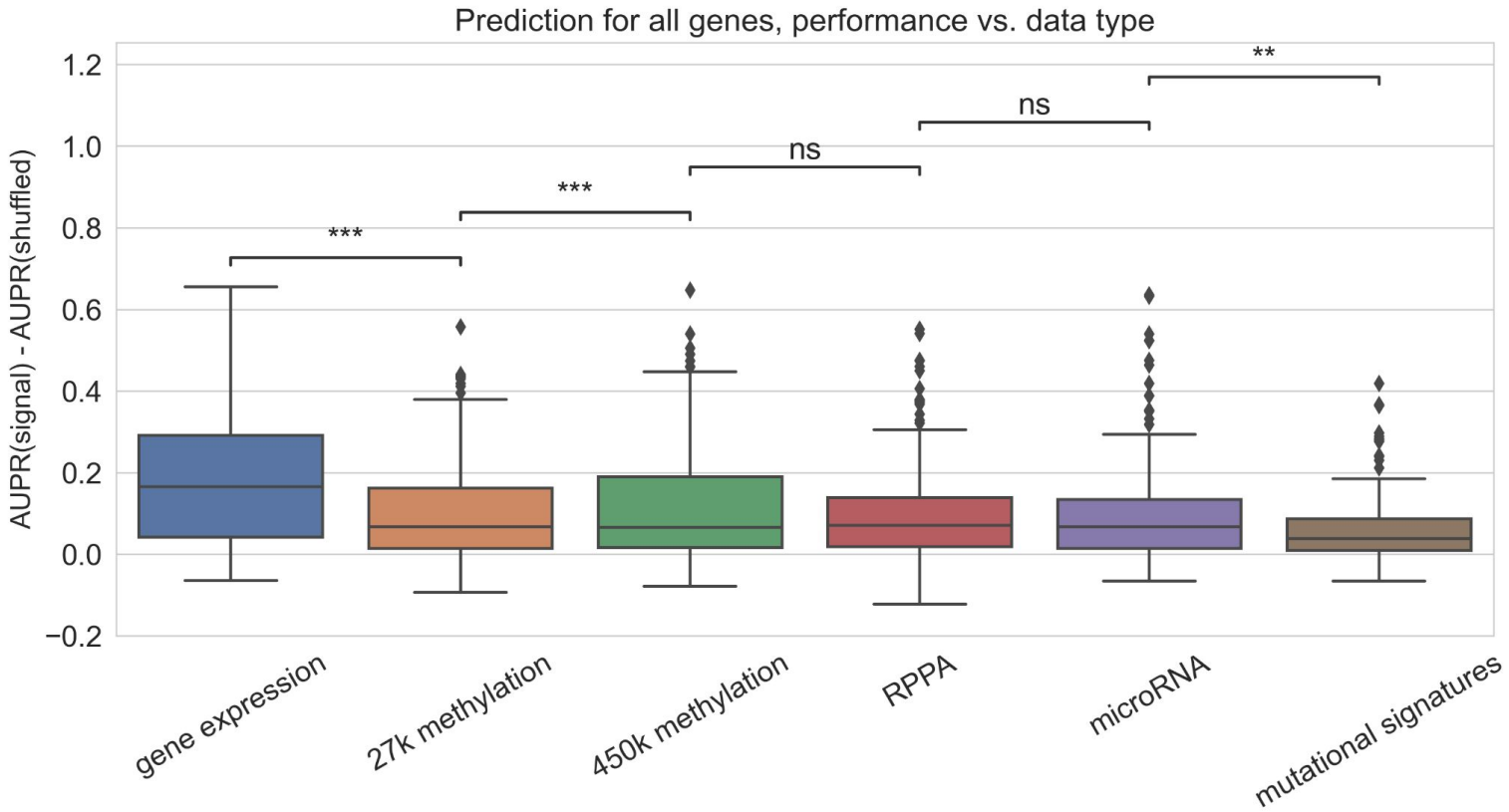
- Use samples where *all data types* are measured -> **5,226** total samples

All data types prediction: comparisons vs. baseline




Mutation prediction, gene expression data
Reject H_0
False
True
79/217 genes
EGFR
CIC
ERBB2
TSC1
CCND1
CCNE1
FGFR1
MDM2
KEAP1
IKZF3
FGFR3
CTNNB1
NKX2-1
JAK1
HMGA2
PPM1D
DEK
α = 0.001
α = 0.01
α = 0.05
$-\log_{10}$(adjusted p-value)
Mutation prediction, 27k methylation data
31/217 genes
BRAF
ERBB2
CIC
CCND1
IKZF3
GTF2I
Mutation prediction, 450k methylation data
42/217 genes
CIC
CCND1
EGFR
RB1
BRAF
ERBB2
CTCF
CTNNB1
ATRX
IKZF1
CASP8
GATA3
FGFR3
NFE2L2
GTF2I
SETD2
NSD1
Mutation prediction, RPPA data
33/217 genes
CIC
IKZF3
CTNNB1
RPL22
KIT
JAK1
AUPR(signal) - AUPR(shuffled)
Mutation prediction, microRNA data
25/217 genes
CIC
IKZF3
ERBB2
CTNNB1
HMGA2
MDM2
KIT
Mutation prediction, mutational signatures data
2/217 genes
KMT2C
KMT2D

All data types summary across genes



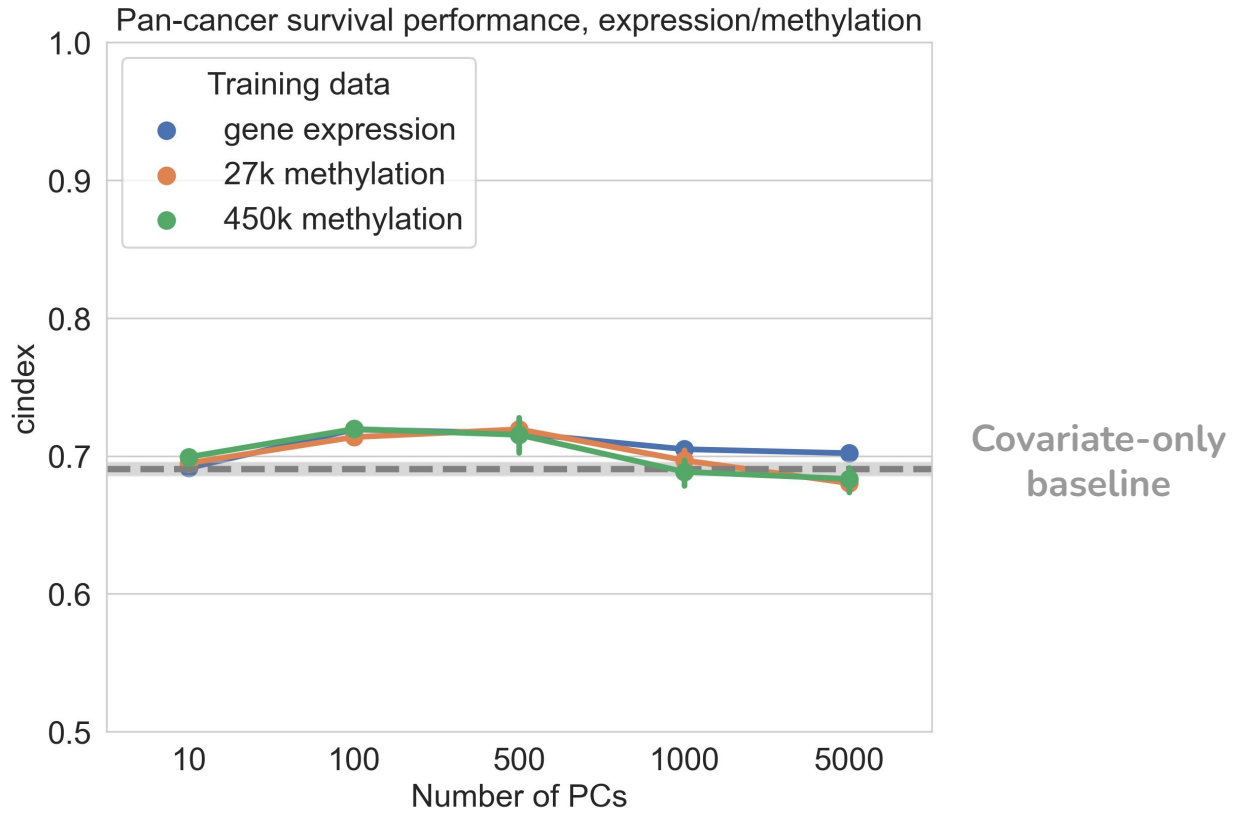
Survival analysis

Do similar results hold for prognosis prediction?

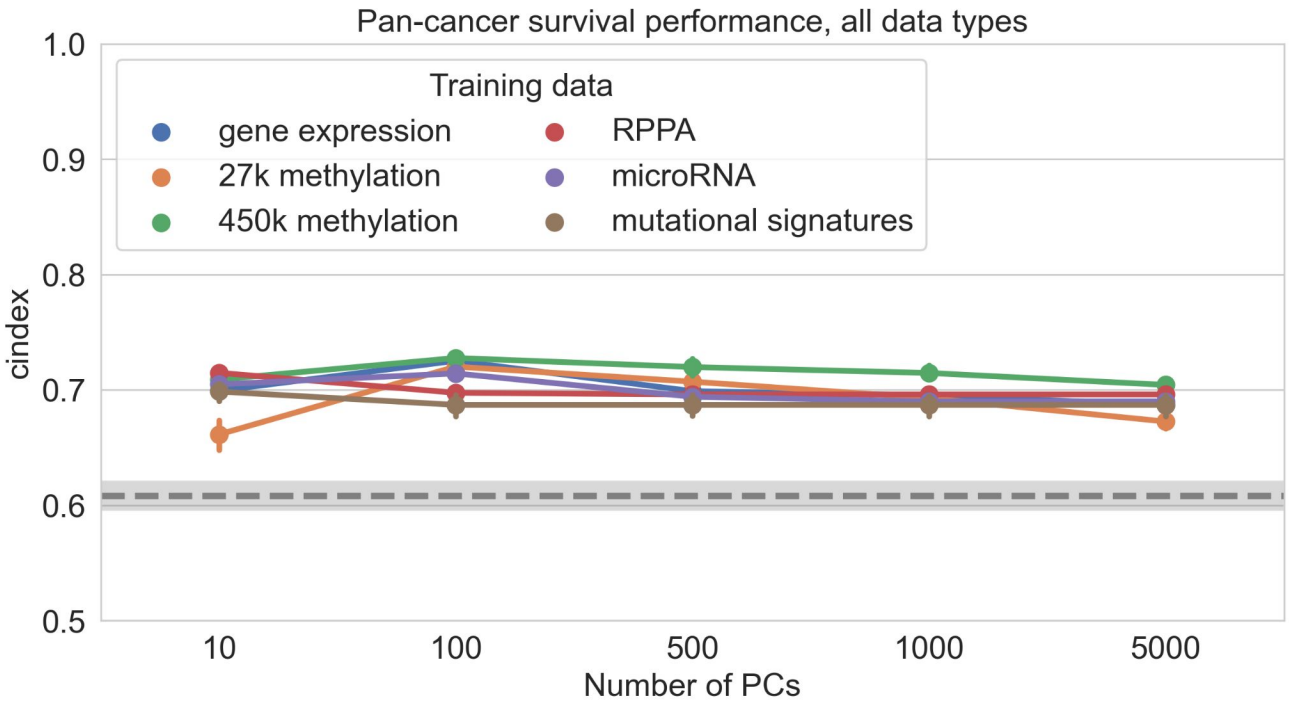
Survival analysis

- Prediction of TCGA clinical endpoints from -omics data
    - PCA preprocessing for all -omics types
    - Elastic net Cox regression
    - Covariates for age, cancer type, mutation burden
    - Baseline predictor uses only non-omics covariates
- Metric: concordance index (higher = better)

Gene expression vs. methylation for survival



All data types for survival



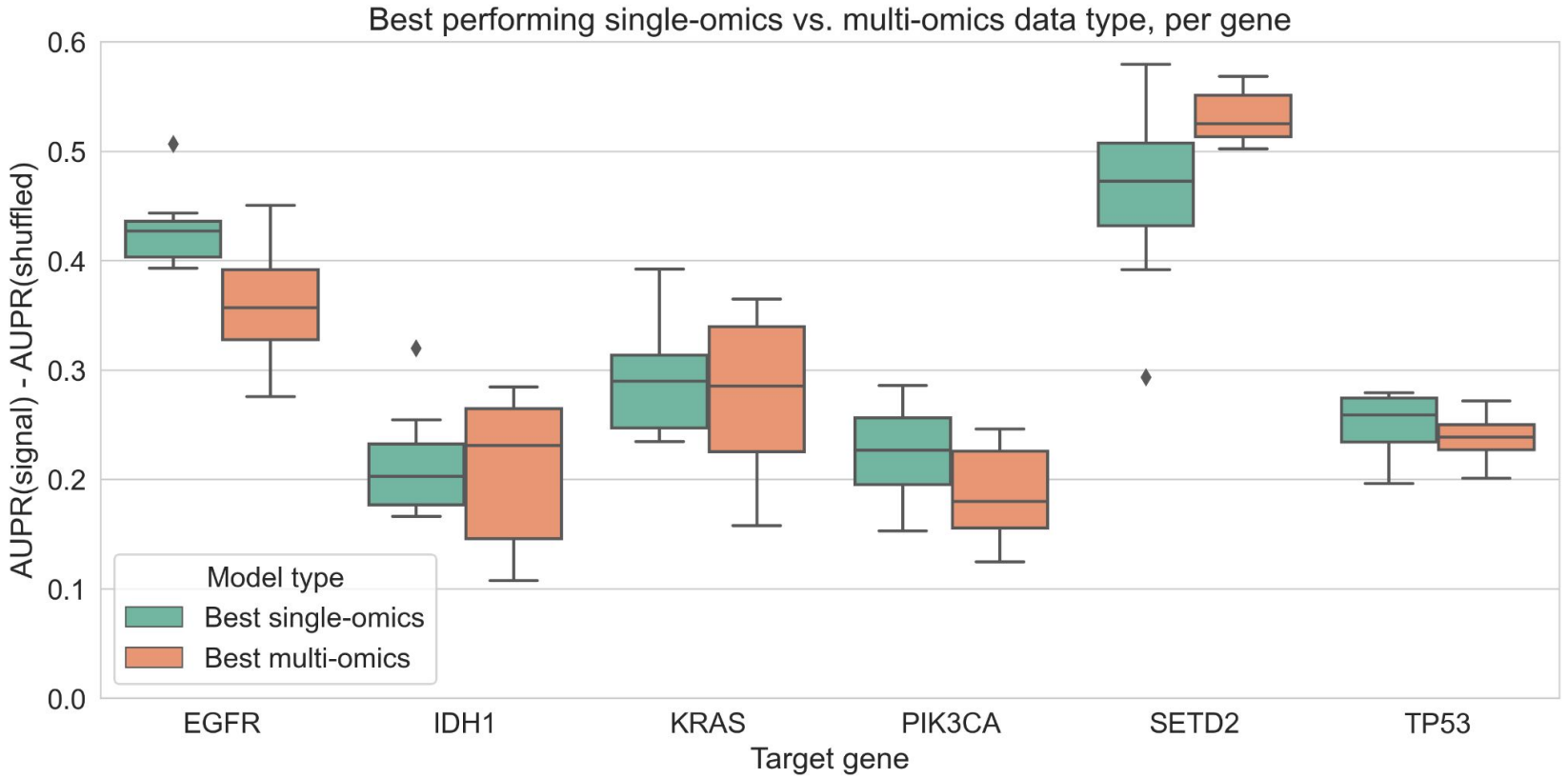
Multi-omics models

Does combining data types improve performance?

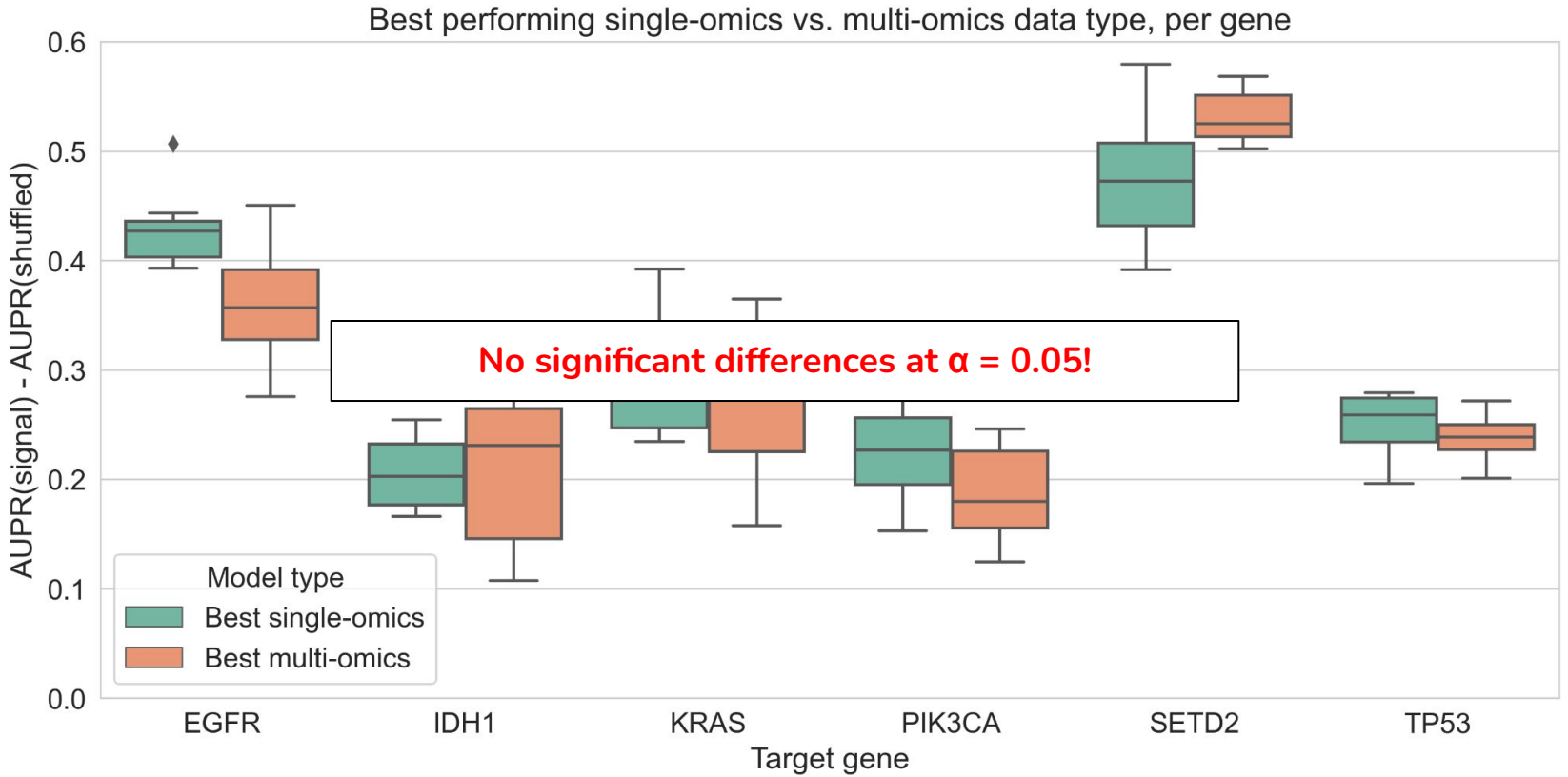
Experimental design

- Can we improve performance by combining more than one data type?
- Gene expression, 27K methylation, 450K methylation
    - Top 5000 PCs for each (results with raw features were similar)
    - Concatenate datasets to form “multi-omics” model
    - All pairs + combination of all 3 data types
- Focused on six “well-predicted” driver genes
    - *EGFR*, *IDH1*, *KRAS*, *PIK3CA*, *SETD2*, *TP53*

Single-omics or multi-omics?



Single-omics or multi-omics?

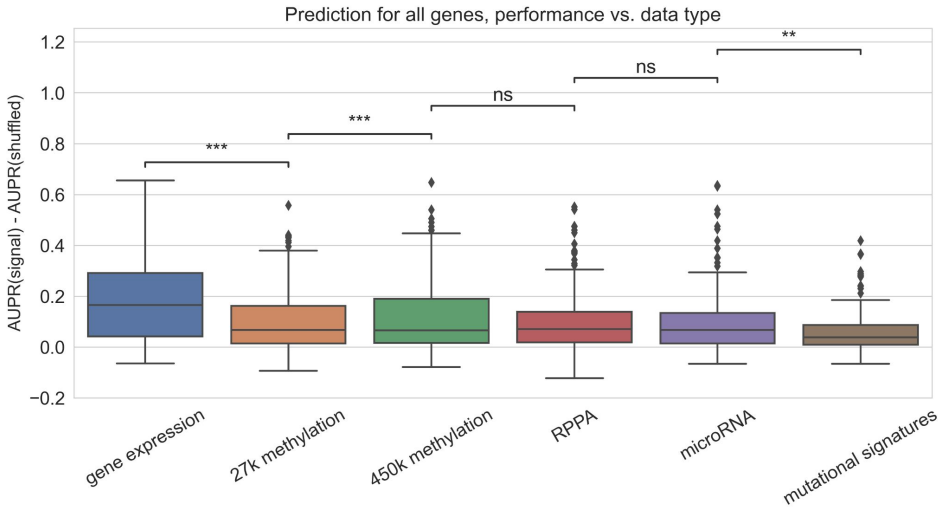


Takeaways

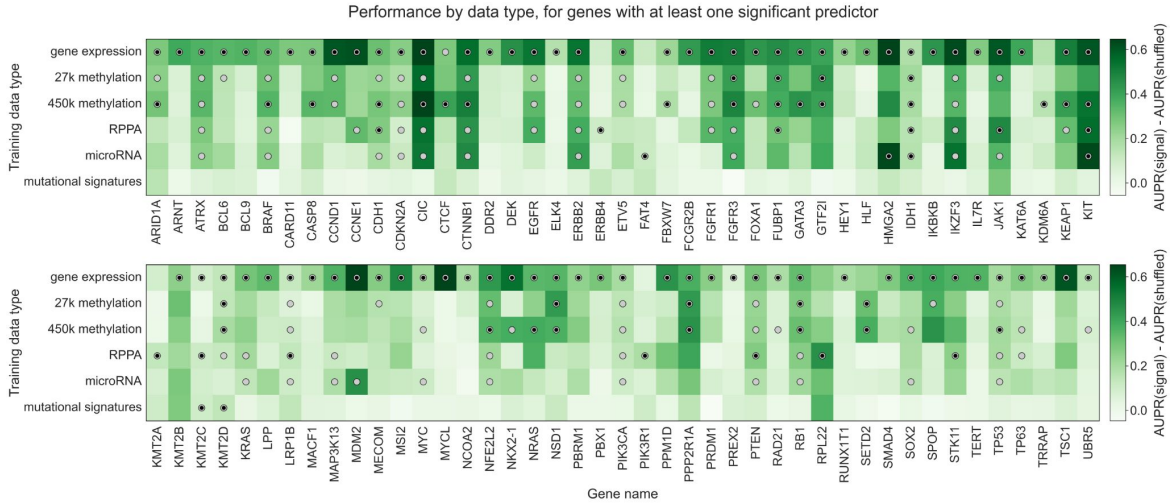
What have we learned?

Gene, not data type, is the primary source of variability

- Gene expression tends to capture the most signal
- Many genes predictable using one -ome are predictable using multiple -omes
- Mutations with strong functional signatures tend to perturb *all* the -omes, to some degree



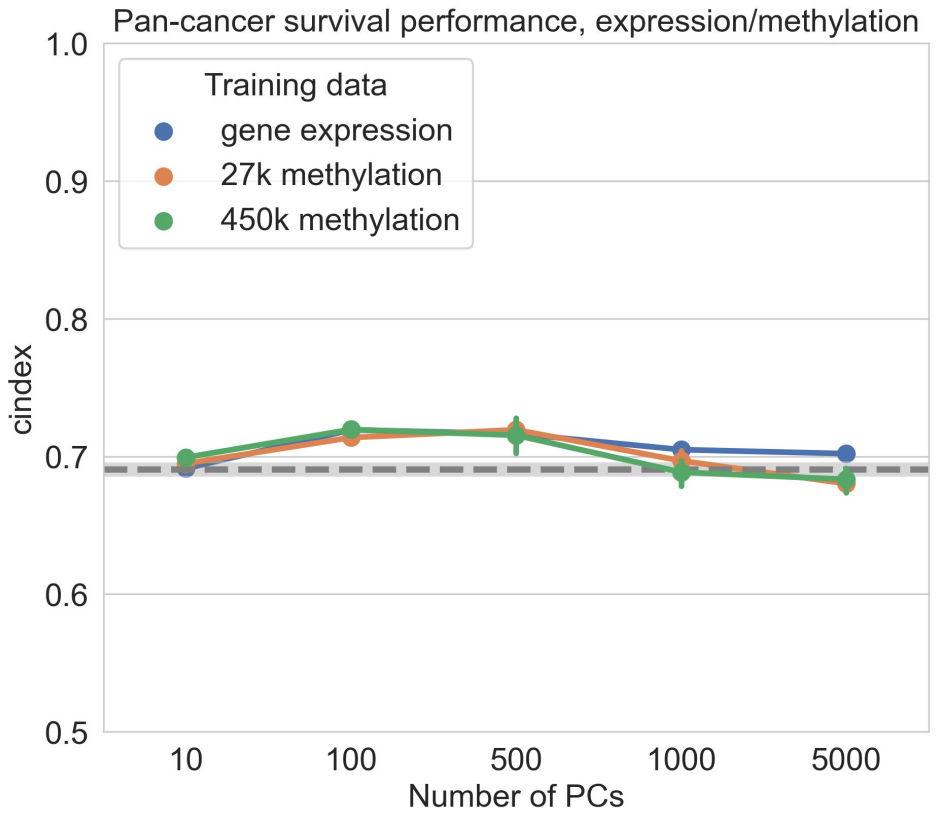
Gene, not data type, is the primary source of variability



- Grey dot = significantly better than baseline
- Black dot = same, and statistically equivalent to “best” predictor

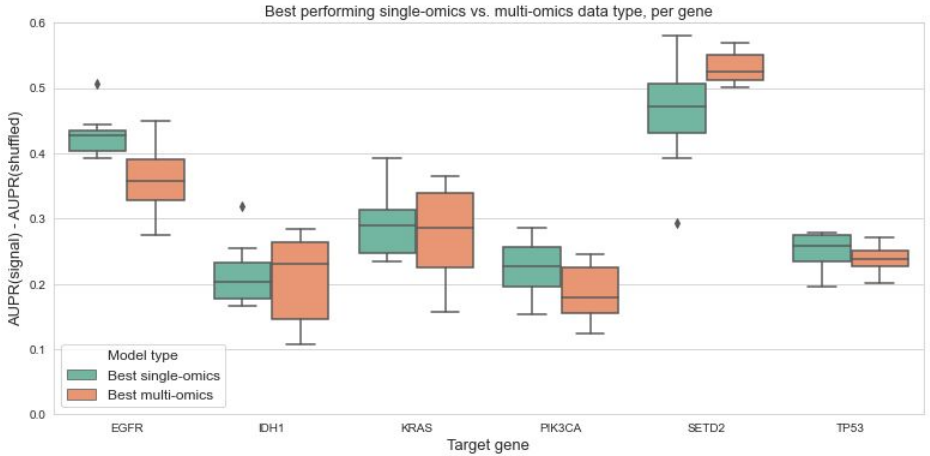
Survival prediction is similar between -omes

- Not much difference between -omics types
- Generally fairly difficult to beat the covariate-only baseline

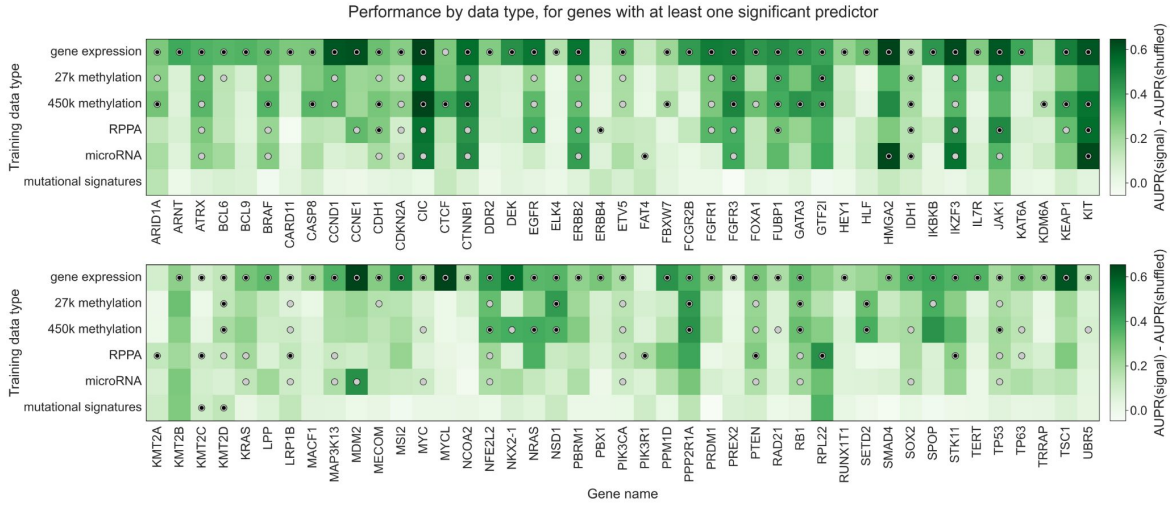


Multi-omics integration doesn’t seem to help

- No discernable difference between best single-omics and multi-omics model
- Most “useful” -omics type varies with target gene
- Relatively simple method, maybe room for improvement



Resource for experimental design



https://greenelab.github.io/mpmp-manuscript

(Figure 6; Supp. Figure 9)

Acknowledgements

Greene Lab:

- **Casey Greene**
- Natalie Davidson
- Ben Heil
- Ariel Hippen
- Alex Lee
- David Nicholson
- Milton Pividori
- Halie Rando
- Taylor Reiter
- Vince Rubinetti

Collaborators:

- **Brock Christensen** (Dartmouth)
- **Maria Chikina** (U of Pittsburgh)

Thesis Committee:

- Marylyn Ritchie
- Mingyao Li
- Kai Tan
- Pablo Cámara
- Donna Slonim (Tufts)

Code and analyses:
https://github.com/greenelab/mpmp

Paper in *Genome Biology*:
https://doi.org/10.1186/s13059-022-02705-y

Greene Lab