Potential problematic statements in the driver paper

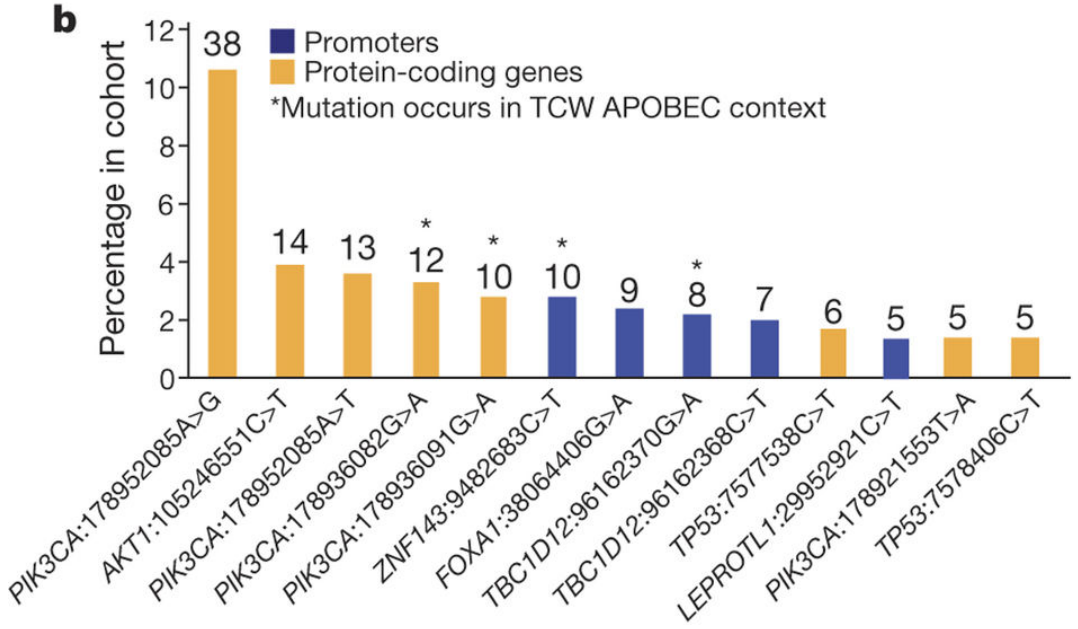
Power analyses indicate that the relative paucity of non-coding drivers is only partially due to small cohorts and smaller functional territory, suggesting that the vast majority of common genetic cancer drivers are protein-coding.

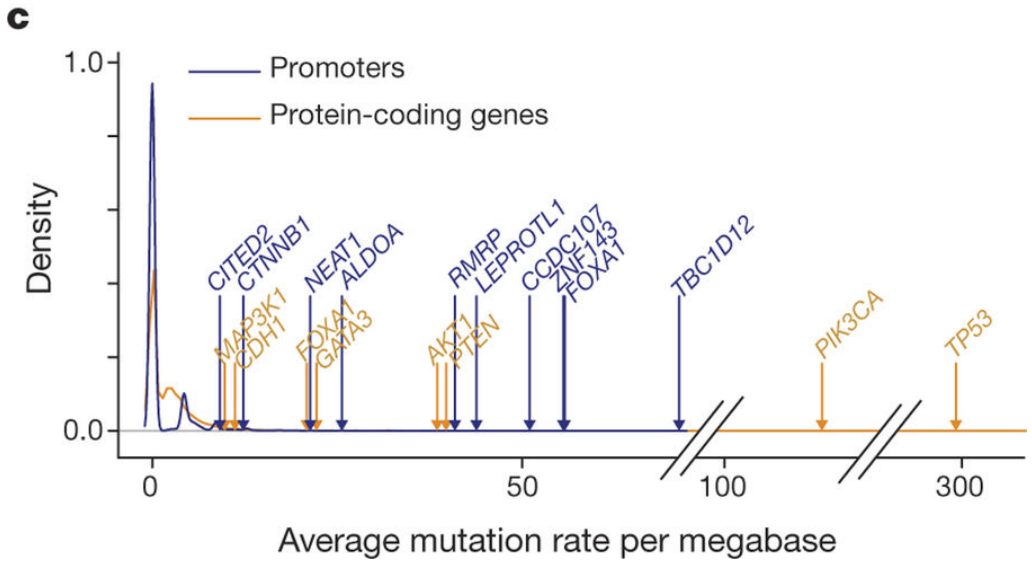
The use of all non-CGC genes as putative passengers for the purpose of estimating the background number of mutations makes this approach conservative. Nevertheless, the results strongly suggest that driver mutations in the promoters of known cancer genes are extremely rare.

"Our study shows that promoter regions harbour recurrent mutations in cancer with functional consequences and that the mutations occur at similar frequencies as in coding regions. Power analyses indicate that more such regions remain to be discovered through deep sequencing of adequately sized cohorts of patients."

The unexpected paucity of regulatory non-coding mutations suggests that SNVs and small indels do not easily alter the function of non-coding regulatory elements and that directly mutating protein-coding sequences or altering expression levels by larger structural events, such as copy number changes, are much more likely to have large fitness effects in cancer cells.

Relevant result from the breast cancer paper





"promoter hotspots were among the most frequent single-site recurrent events across all sequenced territory, including both coding and non-coding."

"values of μ_f for the promoters were similar to or exceeded those of several well-known coding drivers (Fig. 4c), supporting the view that the low observed frequency of promoter mutations may be due, at least in part, to their smaller functional genomic footprint."

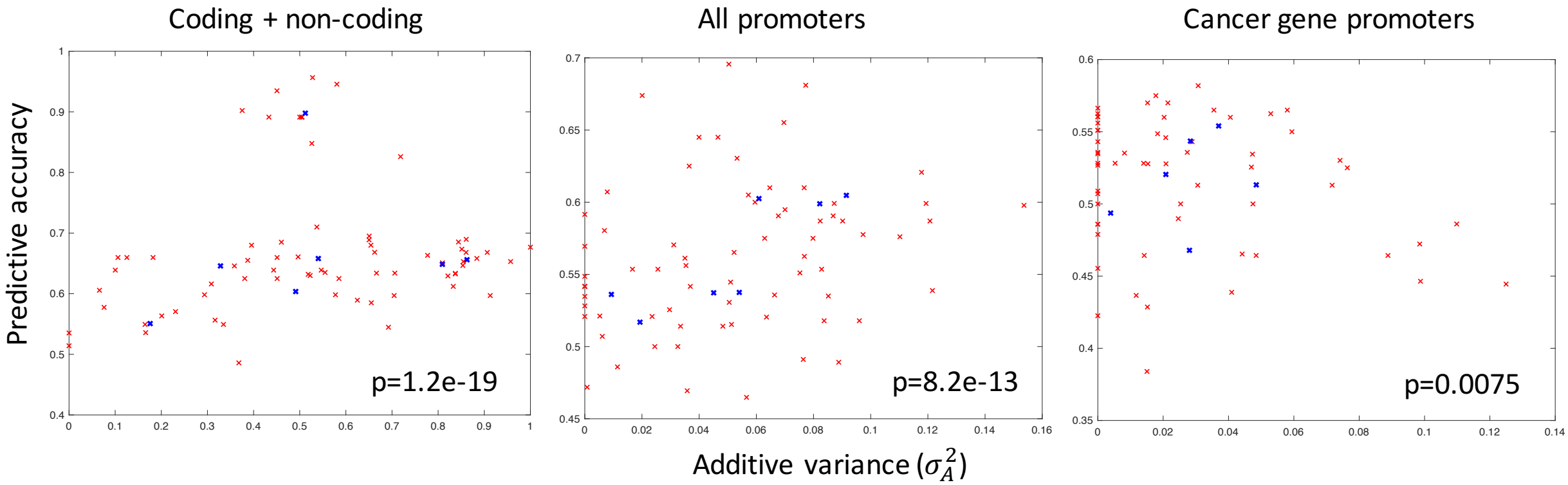
Driver paper aggregate analysis

- General issues:
  1. Analysis assumes there will be an enrichment of drivers among the cancer gene promoters if non-neutral effects are present
  2. Negative binomial regression model (NBR) uses an expected mutation rate per promoter at the cohort level; for the aggregate analysis, a single ‘pan-cancer cohort’ is used, excluding melanoma and lymphoma cohorts (hence collapsing variation across remaining types)
  3. Analysis assumes that the 603 CGC promoters will have the largest non-neutral effects if they are present, hence non-cancer promoters may be used to model the background

Our analysis

- We investigate these assumptions/issues using a version of the additive variance model recast as a predictive model (using the Best Linear Unbiased Predictor, BLUP); we make 10 splits of the data into equal-sized test and training partitions for each cohort, and train separate models on each
- We find:
  - Both all and cancer-only promoters have significant predictive power for cancer phenotype
  - There is not a significant enrichment of positive effects in the cancer gene promoters
  - The CGC promoters are not significantly more predictive at the pan-cancer level, although they are for some tumors

Promoters have predictive power for cancer phenotype



Blue: Cohort means; Red: 10 X test/train splits for each cohort;
p-value: Binomial-test against null hypothesis that predictive_accuracy=0.5

NB: Half of the data only used for training, which is expected to lower predictive accuracy and increase variance on the σ_A^2 estimate

There is a mixture of positive and negative effects in cancer gene promoters

- Proportion of positive coefficients assigned to CGC promoters in BLUP model (10 test/train splits each):

Cohort	Breast	CNS	Kidney	Ovary	Pancreas	Prostate	Skin
Proportion	0.48	0.43	0.49	0.51	0.46	0.42	0.32

CGC promoters are more predictive than others only in certain tumor types