Novel Machine Learning Methods for Cancer Research

Colin Campbell

University of Bristol

Collaborators

- Mark Rogers
- Tom Gaunt
- . . . others (see papers) on various aspects of the project:
- Hash Shihab, Madeleine Darbyshire, Michael Ferlaino (indels), Zach du Toit.

1. Predicting the pathogenic impact of sequence variation in the human genome

- Using integrative methods from **machine learning** we have developed a variety of classifiers for predicting if a variant in the human genome is functional (or not) in disease.
- Input Data: discrete, continuous, graph, sequence (ACGT)), up to 30 types of data can be used by the algorithm.
- **One example of input data**: sequence conservation across species (exploit evolution). A variant in a region highly conserved across species has a higher probability of being functional in disease relative to a region with high variability across species (it's important, Nature can't mess with it).

Main tools developed

- generic (FATHMM-MKL, -XF): fathmm.biocompute.org.uk
- cancer (CScape, CScape-somatic): cscape.biocompute.org.uk
- indels (FATHMM-indel): indels.biocompute.org.uk
- visualisation: gtb.biocompute.org.uk
- haploinsufficiency

Main contexts

- Single nucleotide variant (SNV): $AACTA\mathbf{G}GTA \leftrightarrow AACTA\mathbf{A}GTA$
- Indel (insertion or deletion of genetic code): $ACCGTATACG \leftrightarrow ACCGCG$
- Ongoing research programme (*CScape-somatic*, *CScape-indel*, applications projects)

FATHMM-XF

- Positive examples: Human Gene Mutation Database (HGMD)
- Neutral examples: the 1000 Genomes Project Consortium
- We restrict neutral data to SNVs with a global minor allele frequency $\leq 1\%$ and remove any that appear in the pathogenic dataset
- To mitigate potential bias, we filter neutral examples, selecting only those within 1000 positions of a pathogenic mutation.
- Our final training set consists of 156775 coding examples and 25720 non-coding.
- The model uses six feature groups and reaches 88.0% test accuracy.

2. *CScape*: cancer-specific prediction (2017)

- Used cancer data from the COSMIC archive (*FATHMM-MKL* is main variant annotator) (the positives).
- Used data from 1000 Genomes Project (the negatives)
- Associates a confidence measure to the predicted label (disease-driver or neutral): Platt scaling, gives a p-score (a proxy p-value)

Typical feature groups

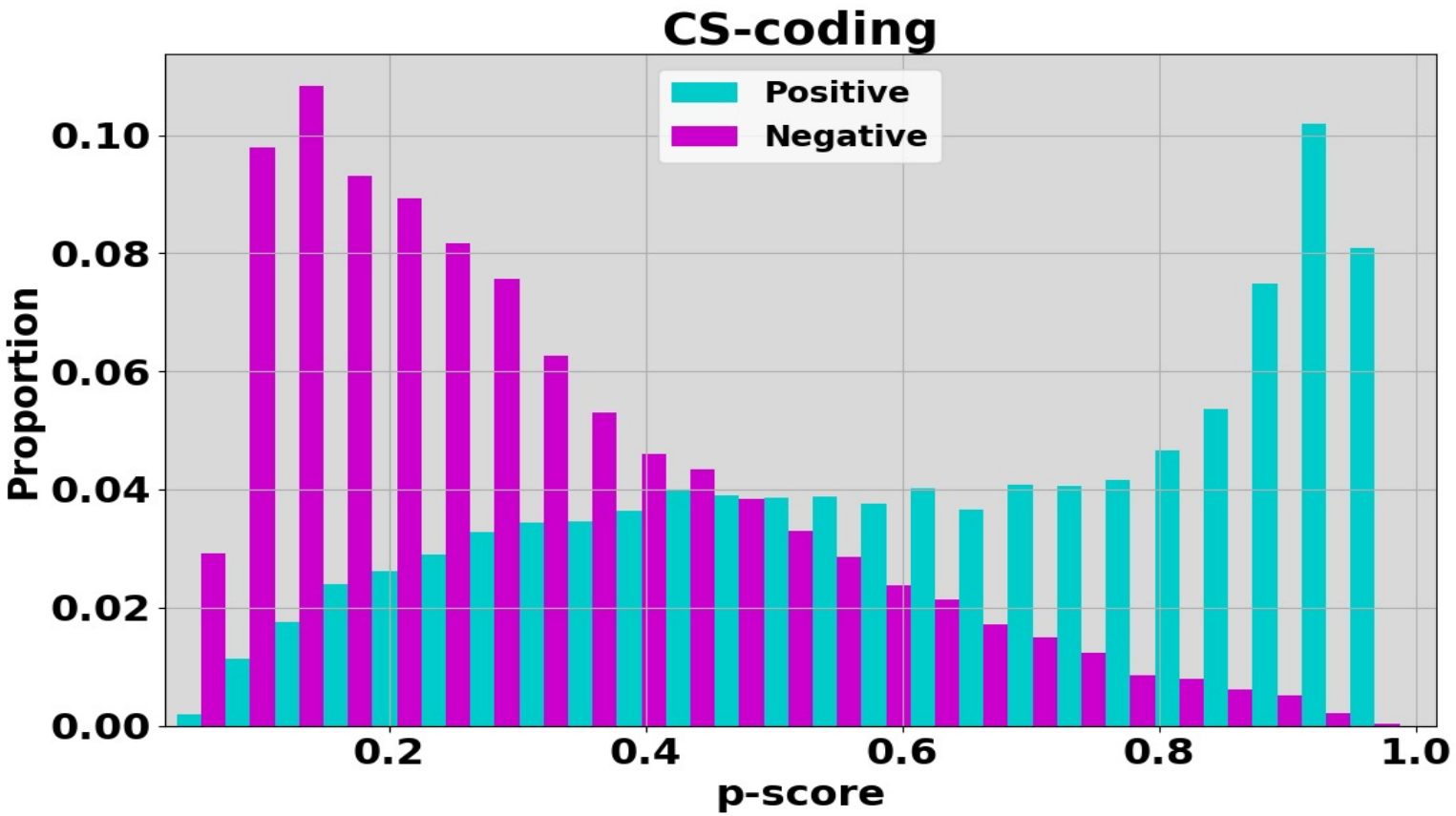
- Genomic and Evolutionary: a comprehensive set of conservation-based measures
- Histone Modifications
- Open Chromatin
- Transcription Factor Binding Sites
- Gene Expression
- Sensitivity to methylation
- Digital Genomic Footprinting Sites
- Network data

for the coding predictor only we also used a variety of protein structure measures.

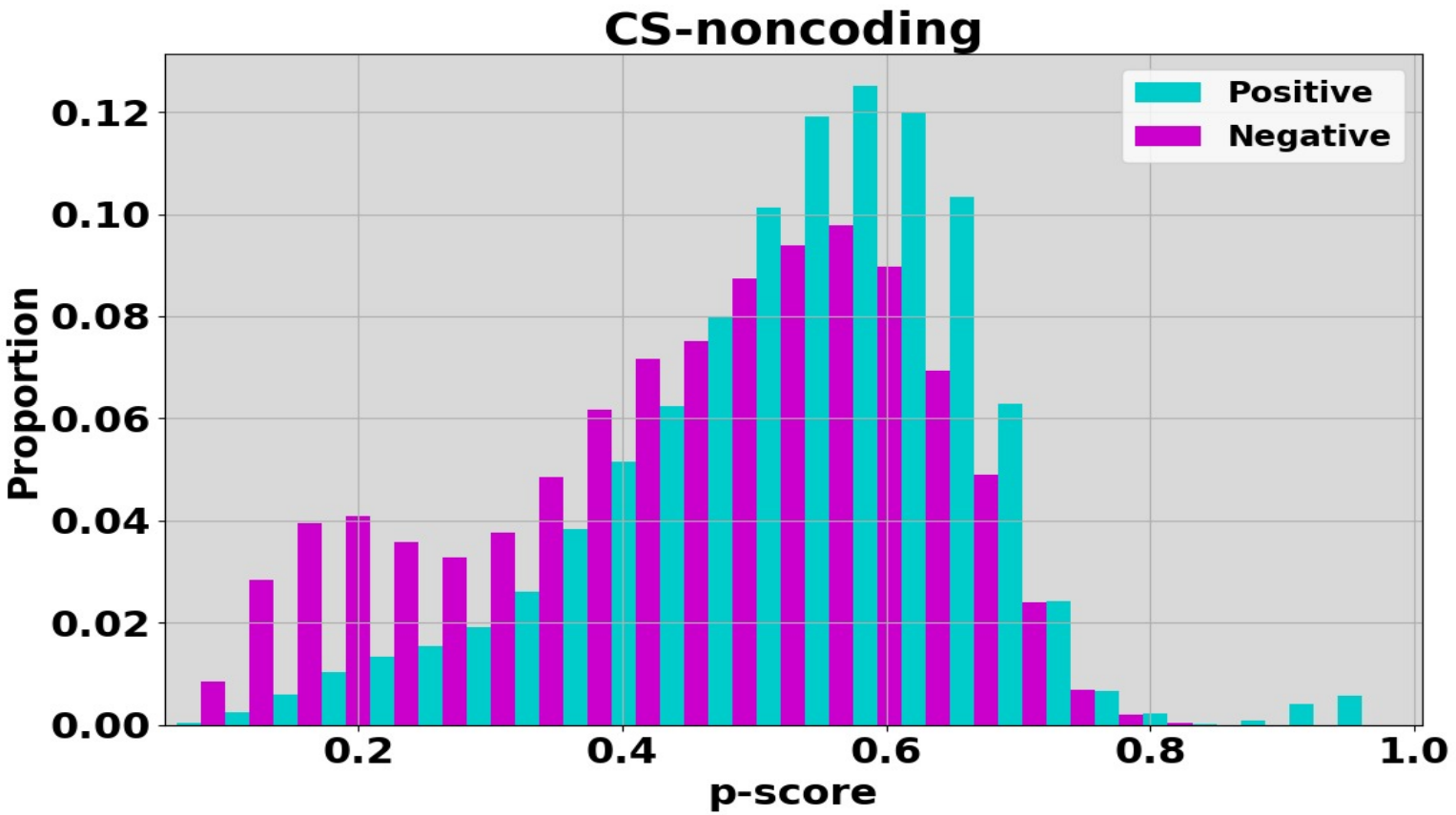
Test Accuracy (*CScape* predictor, 2017)

- Positives: COSMIC data which shows little evidence of bias and provides enough training examples to build a classifier (balanced data). With this criterion we selected a recurrence threshold of $r = 5$ in coding regions and $r = 3$ in non-coding regions for the positives.
- Negatives: 1000 Genomes.
- Using balanced test sets, and LOCO-CV testing (LOCO: leave one chromosome out), the classifier achieves a test accuracy of 72.3% in coding regions and 62.3% in non-coding regions with some higher test accuracies on independent datasets (training and test sets approximately balanced: no bias towards false positives or false negatives).

Coding regions



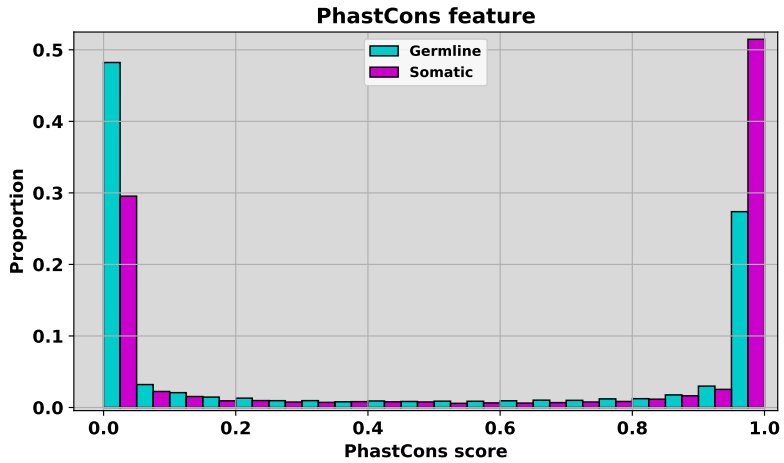
Non-Coding regions

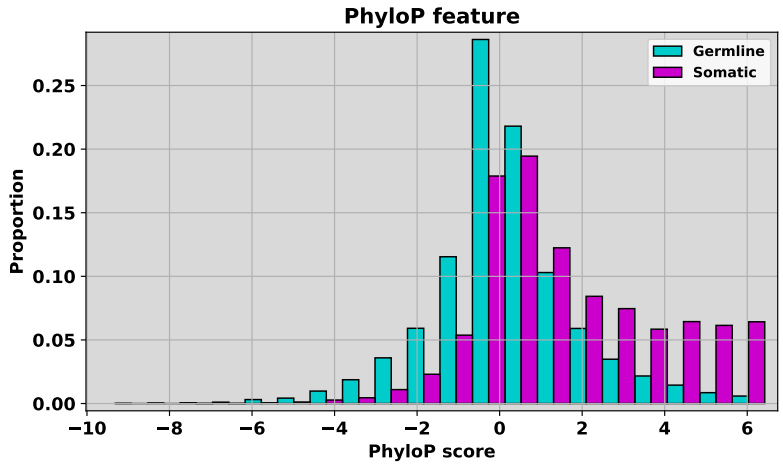


Evaluation

- *Coding* regions: the peak test accuracy is 91.7% (LOCO-testing) which is achieved for a cutoff on the confidence measure at 0.89. For test data taken across the genome, 17.7% of test examples had a high enough confidence for prediction at this level.
- *Non-coding* regions: the peak test accuracy is 76.1%, which is achieved at a cutoff on the confidence of 0.70. Taken across the entire genome, 14.8% of locations in non-coding regions had a predicted label at this accuracy.
- Compared *CScape* against other methods, on unseen data from the International Cancer Genome Consortium, The Cancer Genome Atlas, The Database of Curated Mutations and ClinVar.

Germline vs $r = 1$ somatic: there is a difference in the distributions (here: sequence conservation scores)





CScape-Somatic and more recent predictors, 2019

- *CScape-somatic*: uses purely cancer data ($r = 1$, versus recurrent)
- More accurate: 74% in coding, 69% in non-coding.
- Using more recent data and experimenting with new feature groups: more than 80% test accuracy in coding regions looks tractable (incomplete study).

. . . but what do these methods tell you about cancer?

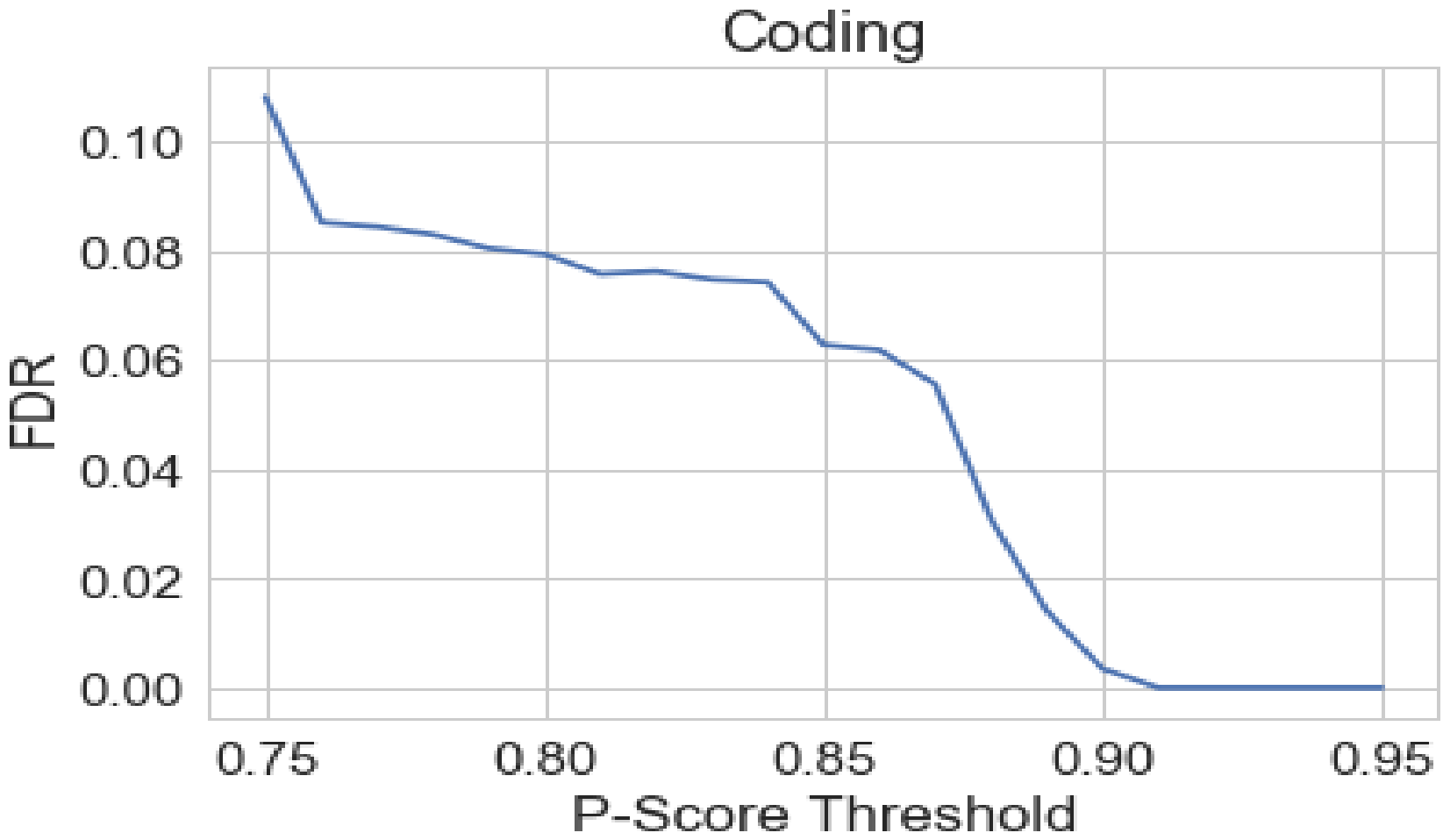
The following discussion is based on *CScape* (2017):

Mark Rogers, Hashem Shihab, Tom Gaunt, and Colin Campbell. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Scientific Reports* (Nature) 7, article number: 11597, (2017)

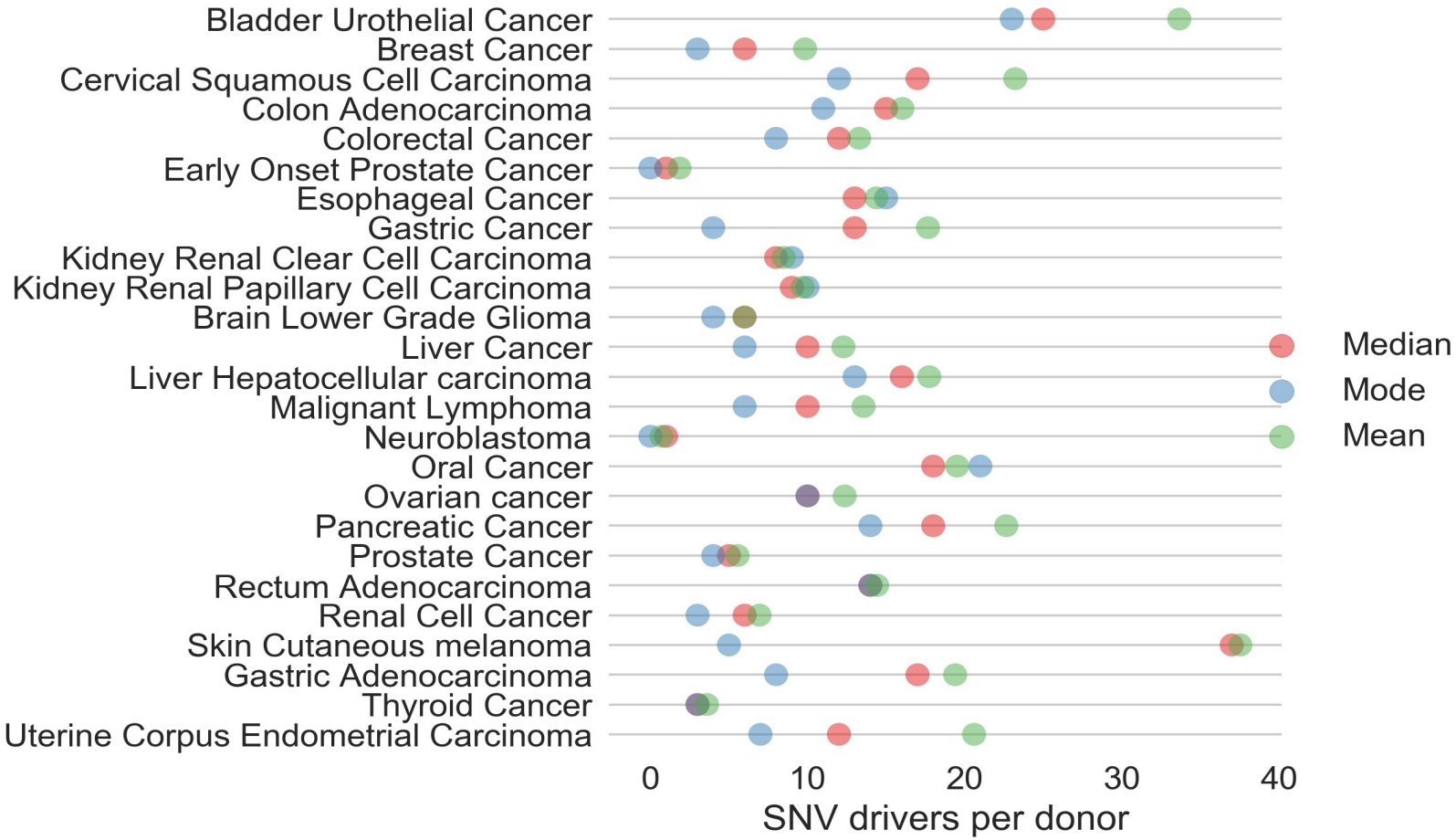
and it is based on this paper:

Madeleine Darbyshire, Zachary du Toit, Mark F. Rogers, Tom R. Gaunt and Colin Campbell. Estimating the Frequency of Single Point Driver Mutations across Common Solid Tumours. *Scientific Reports* (Nature) 9, article number: 13452, (2019).

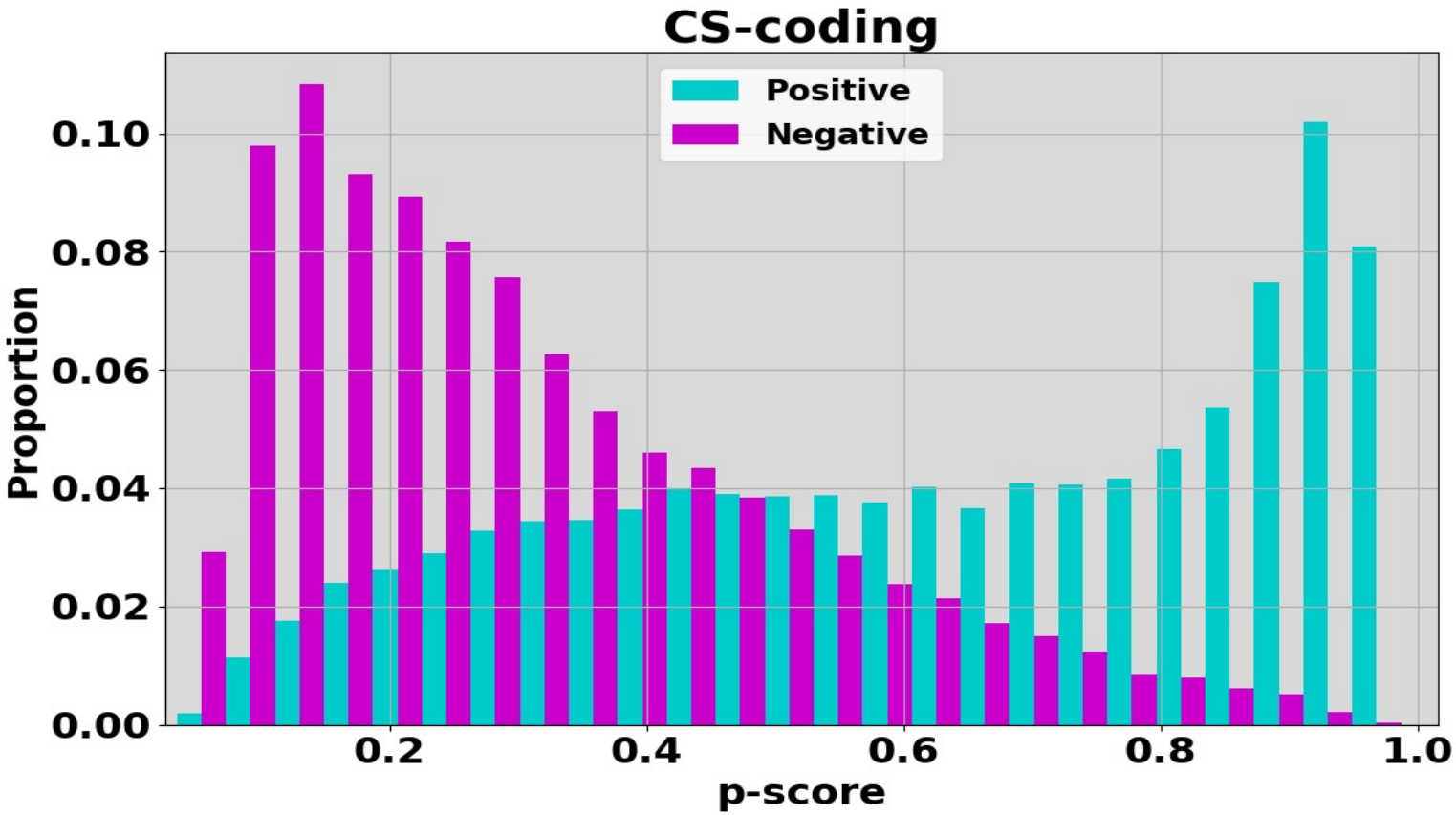
Coding regions: false discovery rate and p-score



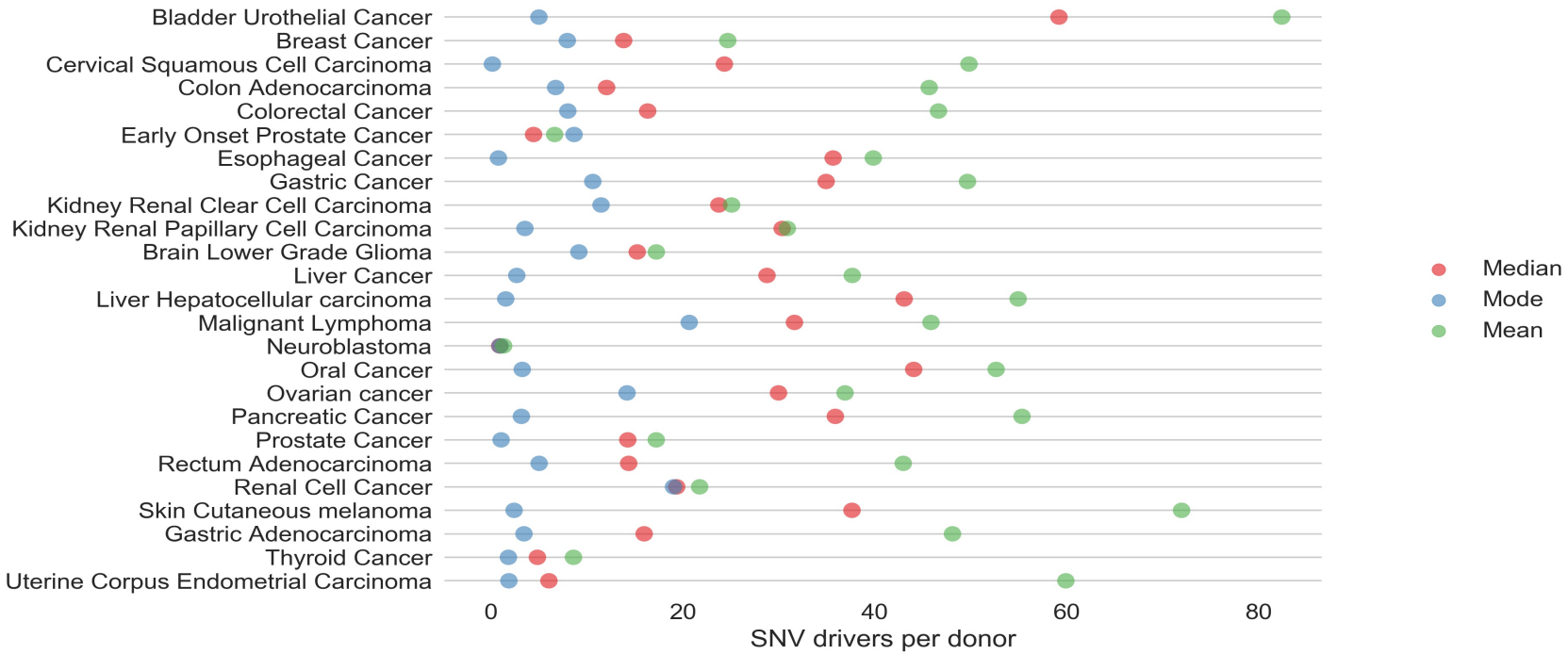
Coding regions (FDR of 5%)



Coding regions: alternative estimation



Coding regions: alternative estimation



Observation 1: the number of SNV disease-drivers in coding regions is small in size, SNV-drivers are partially identifiable

- Coding: small number of SNV disease-drivers. *There may be many sites where single point mutations can act as drivers, but an individual clone only has a small subset of these*.
- Very variable by type of cancer. For example:
- Thyroid cancer: average number of SNV-drivers in coding regions is 3.8 (492 samples), average across all cancers is 15.9 (5424 samples). Hypothesis testing: the probability that these two distributions are the same is upper-bounded by 10^{-100} (well, the set sizes are large and the means very different).
- Non-coding: not clear at present.

CScape results

- Coding regions: driver sets sizes in single figures or low double figures depending on cancer type.
- Hypermututions is excluded (alterations in proofreading domains of POLE, POLD1)
- Some differentiation within cancer types. Late stage prostate cancer (typecode:PRAD) has twice as many coding SNV-drivers as early stage prostate cancer (typecode:EOPC).
- Even among those cancers with larger coding SNV-driver sets, there are sub-populations (we call *neo-modal*) with smaller driver sets).
- Aligns with earlier arguments based on mutation rates which suggest driver sets are small in size.

Other results along these lines

- Aligns with analysis by Martincorena *et al*, *Cell* **171**, 1029-1041 (2017).
- Martincorena *et al*: use a statistical argument based on the ratio of non-synonymous to synonymous mutations (dN/dS), synonymous mutations give a null base distribution of neutral variants.
- Martincorena *et al*: use data from the International Cancer Genome Consortium (ICGC).
- Ourselves: machine learning argument, use COSMIC (cancer) and 1000 Genomes (neutrals) datasets for training and to derive above plots.
- Both approaches (coding regions): thyroid cancer has one of the smallest driver sets, bladder cancer one of the largest.
- We estimate more SNV-drivers in cancer, but the picture presented is broadly in alignment between the two approaches.

Historical

- Nordling (1953) and Armitage and Doll (1954): age/cancer incidence models, suggested 6 to 7 sequential mutational events.
- More recent (Tomasetti et al, 2015, lung and and colorectal cancer): statistical argument, single digits.

Observation 2: there is limited accumulation of extra coding SNV drivers with stage of disease (for most cancers)

- Amalgamating data across different stages of disease is maybe unwise? Tumour mutational burden may increase with stage of disease and biopsies may be taken at different stages. Unequal sampling rates: successful intervention may deplete samples at later stages of disease.
- *Find*: increasing numbers of SNV drivers with stage of disease is an exception as a phenomenon, not the rule.
- *Early onset prostate cancer (typecode: EOPC, means)*: 2.5 (I), 3.8 (IIA), 6.5 (IIIB). Low start and increases.
- *Prostate cancer (PRAD), means*: 4.4 (IIB) and 7.6 (IIC), 16.4 (IIIB), 21.8 (IVA). Low start but increases.

... but the norm

- *Renal cell (RECA)*: 3.4 (I), 3.3 (II), 3.2 (III), 3.0 (IV). Low and stays low.
- *Esophageal (ESCA)*: 13.5 (I), 9.9 (II), 10.0 (III), 12.2 (IVA). Higher but constant.
- Same conclusion as Martincorena *et al* (in *Cell*) (cf. their Figure S4C, dN/dS) who argue (stage I versus stage IV) there is little evidence for significant increases in the number of drivers as disease progresses (the tumour mutational burden, the number of non-synonymous SNVs *could* increase irrespective of the number of drivers due to loss of genomic repair mechanisms).

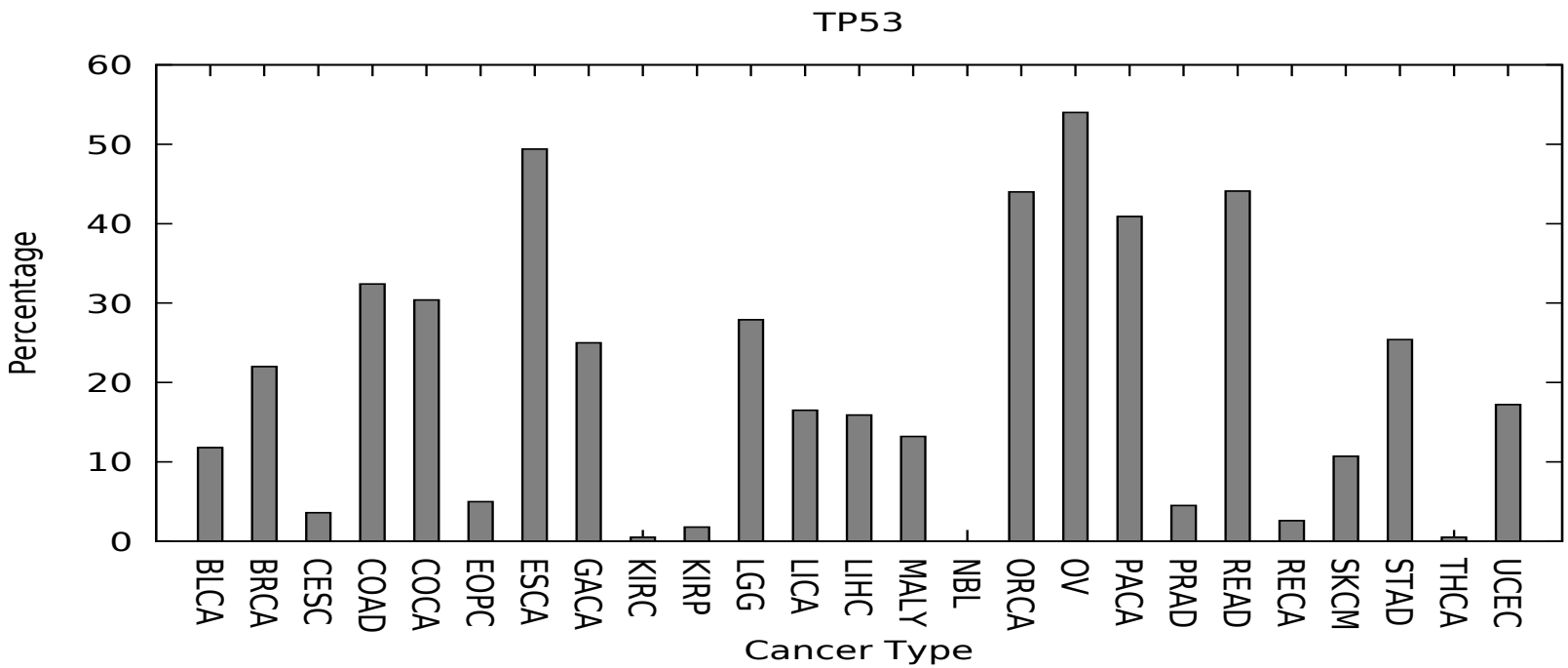
Observation 3: certain genes are common as drivers

- Criterion: minimum of *one* high confidence SNV-driver in gene.
- Find: *TP53* in top five driver genes in 17 of the 25 cancer types studied.
- Three more broad-based driver-genes:
  - *PIK3CA* (6 of 25)
  - *KRAS* (5 of 25)
  - *CTC-297N7.11* (4 of 25)

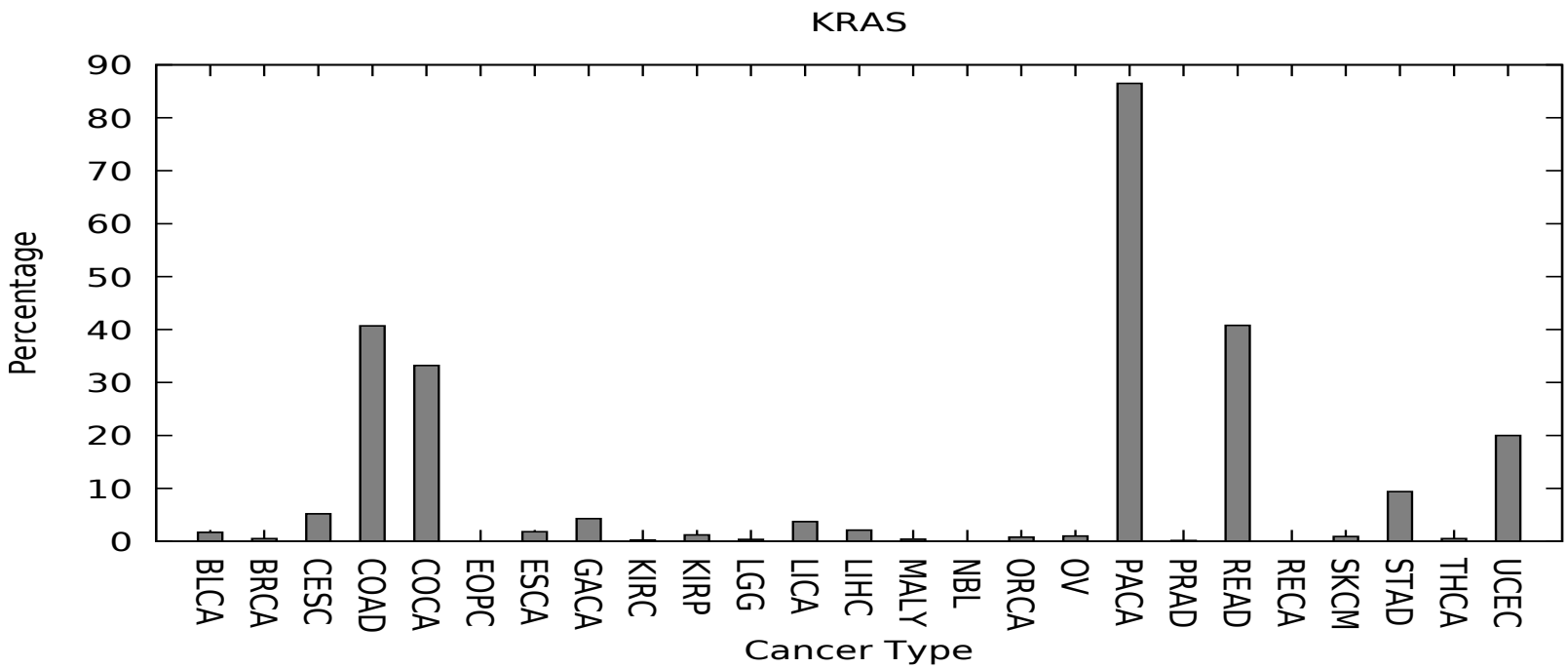
Certain genes are fairly specific to a context

- *APC* is the top ranked driver-gene for colon adenocarcinoma (COAD) and colorectal (COCA) (top five driver-gene ranking the same despite different sample sets).
- *KRAS*: incidence of 86.5% in pancreatic cancer.
- *BRAF* is in the top five driver-genes for skin cutaneous melanoma and thyroid cancer.
- Thyroid: high confidence SNV-drivers are present in *BRAF* in 55.8% of cases, next highest qualifying gene is *NRAS* at 1.3%.
- A given common driver-gene can have varying influence in different cancers:

TP53


TP53
Percentage
60
50
40
30
20
10
0
BLCA
BRCA
CESC
COAD
COCA
EOPC
ESCA
GACA
KIRC
KIRP
LGG
LICA
LIHC
MALY
NBL
ORCA
OV
PACA
PRAD
READ
RECA
SKCM
STAD
THCA
UCEC
Cancer Type

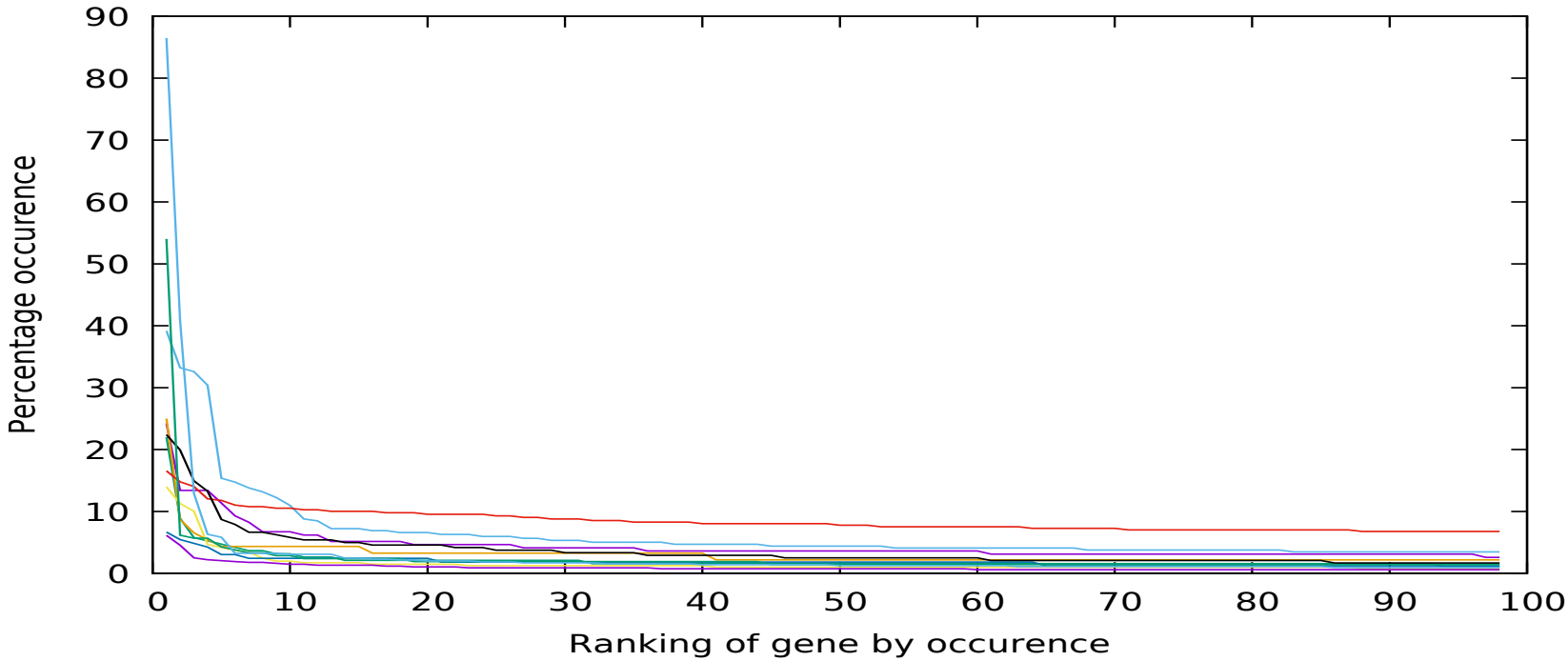
KRAS


KRAS
Percentage
90
80
70
60
50
40
30
20
10
0
BLCA
BRCA
CESC
COAD
COCA
EOPC
ESCA
GACA
KIRC
KIRP
LGG
LICA
LIHC
MALY
NBL
ORCA
OV
PACA
PRAD
READ
RECA
SKCM
STAD
THCA
UCEC
Cancer Type

Observation 4: there are long tails of infrequent driver-genes

- The above driver-genes are well known.
- However, the machine learning methods are partially successful in actually identifying the driver alterations.
- We have been ignoring other types of drivers (e.g. indels, non-coding, etc: neuroblastoma, SNV-drivers barely play a role, other drivers such as indels, copy number variants, etc, must be significant).
- The above genes (*TP53*, *BRAF*, *KRAS*, etc) are *common drivers*, but they are accompanied by long tails of *infrequent driver-genes*.
- Consequence: **the driver-gene set is individual to a patient or tumour** (as expected)

Driver-gene tails



The top common gene-driver is *KRAS* in pancreatic cancer (left side), liver cancer has a high heterogeneity, thyroid cancer the lowest heterogeneity (top and bottom curves on the right).

Observation 5: machine learning classifiers can *generalize* (they are an AI method, we generalize too)

- Example: A recurrent point mutation at chromosome 17, position 64738741 $G \rightarrow C$, introduces a D463H amino acid substitution and this has been described as a hallmark of chordoid glioma (Goode et al. *Nature Commun.* 9, p. 810, (2018)).
- Not in COSMIC and cBioPortal databases but *CScape* (GRCh37) predicts this point mutation as oncogenic with high confidence (0.964).
- Consequence: can predict beyond its training data and would be able, with partial accuracy, to label the driver-status of single point mutations of infrequent genes in the driver tails.

4. Conclusion

- Would be very interesting to look at other types of drivers: indels, copy number variation, methylation, etc.
- *Indels*: a more substantive alteration so higher test accuracies can be achieved relative to single nucleotides variants (SNVs). Have proposed indel predictors (FATHMM-indel: *indels.biocompute.org.uk*)
- The above analysis has highlighted the importance of devising accurate predictors covering **non-coding** regions of the cancer genome (big, missing part of the picture).
- The data rich world of the biomedical sciences is an excellent area for deploying methods from machine learning.