driveR: A Novel Method for Prioritizing Cancer Driver Genes Using Somatic Genomics Data

Ege Ülgen[1*], O. Uğur Sezerman[1]

[1]Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey.

*Correspondence to:
Ege Ulgen
Address: Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey
E-mail: egeulgen@gmail.com

Abstract
Cancer develops due to “driver” genomic alterations. Numerous approaches exist for predicting cancer drivers from cohort-scale genomic data. However, methods for personalized analysis of cancer driver genes are underdeveloped.
In this study, we developed a novel personalized and batch analysis approach for driver gene prediction utilizing somatic genomic data, called driveR. Combining somatic genomic information and prior biological knowledge, driveR accurately prioritizes cancer driver genes via a multi-task learning model.
Testing on 28 different datasets, this study demonstrates that our approach performs adequately, outperforms existing approaches, and is an accurate and easy-to-utilize approach for prioritizing driver genes in cancer genomes.

Keywords
driver gene; prioritization; cancer; somatic

1. Background

Cancer develops due to changes that have occurred in the DNA sequence of the genomes of cancer cells, somatic mutations acquired during the lifetime of an individual [1]. Cancer genomes contain large numbers of somatic mutations, but most are "passengers" that emerge simply as a result of genome instability during cancer progression and do not contribute to cancer development, and a small proportion are "drivers" that are implicated in oncogenesis [2-4]. Identification of molecular cancer driver genes is critical for personalized oncology as accurate identification of personalized driver genes will result in precise diagnosis and will allow the clinicians to possibly define personalized therapeutic targets [5, 6].

Several computational batch analysis approaches, extensively reviewed by Tokheim et al. and Cheng et al. [7, 8], have been developed to identify cancer driver genes. Some notable approaches for batch analysis include MuSiC [9], MutSigCV [10], MutPanning [11], MEMo [12], Hierarchical HotNet [13], TieDie [14], DriverNet [15], CaDrA [16] and OncodriveFML [17]. MuSiC uses the significance of higher than expected rate of mutations, pathway mutation rate, and correlation with clinical features to detect drivers. MutSigCV investigates the mutational significance of genes by identifying genes that were mutated more often than expected by chance given background mutation processes. MutPanning uses deviation of mutational context from characteristic contexts around passenger mutations in addition to traditionally-used features (e.g. higher than expected mutation rate) for driver gene identification. MEMo tries to detect small subnetworks of genes in the same pathway, exhibiting internal mutual exclusivity patterns. Hierarchical HotNet incorporates knowledge from protein-protein interaction networks (PINs) to find a hierarchy of altered subnetworks, containing frequently mutated genes. TieDie incorporates PIN and mRNA expression data to find overlapping subnetworks exhibiting a high degree of mutation and expression values using heat diffusion. DriverNet tries to detect driver genes via their effect on mRNA expression networks, by identifying a set of genes with mutations/copy-number alterations that are linked to genes with deregulated expression in a given PIN. CaDrA uses a step-wise heuristic search approach to identify functionally relevant subsets of genomic features, maximally associated with a specific outcome of interest. OncodriveFML aims to detect driver genes by analyzing the functional impact bias of observed somatic mutations.

Personalized driver gene prioritization is important for numerous reasons: (i) it is important to identify true driver genes for the patient as some patients have alterations in many known driver genes, (ii) there may be a need to identify putative driver genes in patients without any alteration in any known driver gene, (iii) as the number of therapies that can be given at the same time is limited due to toxicity and adverse events [18, 19], there is a need to prioritize driver genes for the patient. Although there are several approaches for identifying cancer driver genes in cohort-scale genomics data, personalized driver gene identification approaches are still

underdeveloped. Some approaches that operate on the data of a single patient to identify and rank patient-specific driver genes have been developed: DawnRank uses a PageRank algorithm to rank mutated genes according to effect on expression deregulation of downstream differentially-expressed genes (DEGs) in a directed PIN [20]. The single-sample controller strategy (SCS) aims to identify a set of mutated genes that are linked to downstream DEGs in a directed PIN [21]. iCAGES utilizes a statistical framework to identify driver variants by integrating contributions from coding, non-coding, and structural variants and identifies driver genes by combining genomic information and prior biological knowledge [22]. PRODIGY analyzes the expression and mutation profiles of the patient along with data on known pathways and PIN to quantify the impact of each mutated gene on every deregulated pathway using the prize-collecting Steiner tree model [23].

As stated above, most approaches for driver gene prioritization are batch-analysis-based. Furthermore, most approaches utilize only genomics and/or transcriptomics data without exploiting prior biological knowledge. To improve on existing approaches, we aimed to develop a novel driver gene prioritization approach that utilizes somatic genomics information incorporating prior biological knowledge, which we called driveR. We developed driveR intending to establish an accurate and reliable approach for driver gene prioritization. The approach allows for personalized or batch analysis of genomics data for driver gene prioritization by combining genomic information and prior biological knowledge. As features, driveR uses coding impact metaprediction scores, non-coding impact scores, somatic copy number alteration scores, hotspot gene/double-hit gene condition, Phenolyzer [24] gene scores, and memberships to cancer-related Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] pathways. It uses these features to estimate cancer-type-specific probabilities of being a driver for each gene using the related task of a multi-task learning model. In this article, we demonstrate that our approach can help increase the accuracy of cancer driver gene detection and prioritization, and therefore, facilitate precise diagnosis, personalized therapy, and overall result in better clinical decision-making.

2. Methods

2.1. Coding variant impact metaprediction

Initially, we fitted coding impact metapredictor models to assign an estimated probability of damaging impact for each somatic variant. These probabilities were later used as a feature in the multi-task learning classification model.

For training the coding impact metapredictor models, a benchmarking dataset from the Martelotto et al. study [26] was obtained. All mutations within the dataset were annotated using ANNOVAR[27] with the 12 impact predictors: SIFT [28], PolyPhen-2 [29] (HumDiv scores), LRT [30], MutationTaster [31], Mutation Assessor [32], FATHMM [33], GERP++ [34], PhyloP [35], CADD [36], VEST [37], SiPhy [38] and

DANN [39]. Variants with any missing predictor scores were excluded. Additionally, variants with the label "uncertain" were excluded, hence, the final dataset consisted of 135 "neutral" and 814 "non-neutral" variants.
Before training the models, pairwise Pearson correlations of all variables (including the outcome variable) were investigated to determine any collinearity issue and to evaluate the predictive strength of the predictors (Figure S1A). The correlogram revealed that each of the individual scores provides a different facet of information for a coding variant. Additionally, violin plots of all 12 predictors by outcome status (i.e., "neutral" or "non-neutral") were visualized for investigating the distributions of the two status labels and detect any outliers (Figure S1B). The violin plots revealed that the distributions of scores in "neutral" and "non-neutral" variants were different. There were no noticeable outliers.
The final dataset was split into training and test datasets by randomly selecting 75% of each "neutral" and "non-neutral" variants as the training set and the remaining as the test set.
In total, 6 different classification models were trained and evaluated: logistic regression, naïve Bayes, support vector machine (SVM) with linear kernel, SVM with radial kernel, random forest, and gradient boosting machine. Each of the 6 classification models was trained using the training dataset using 10-fold 3-times-repeated cross-validation, maximizing the Area Under Curve (AUC) metric, and tested on the test dataset. Figure S2 shows the receiver operating characteristic (ROC) curves and AUC values for each of the 6 classification models in the training, test, and validation datasets.
The model with the best performance in all of the datasets was the random forest classification model and this model was used as the coding variant impact metapredictor.

2.2. Driver gene prioritization

The overview diagram of our approach to prioritize driver genes is presented in Figure 1.
For prioritizing cancer driver genes in a cancer-type-specific context, a multi-task learning (MTL) classification model was trained. For this purpose, initially, all somatic mutation and somatic copy number alteration (SCNA) data from The Cancer Genome Atlas (TCGA) program available on the International Cancer Genome Consortium (ICGC) data portal [40] was obtained. This data consisted of genomics data for 21 different cancer types (Supplementary Table 1). Genomics data from each of these different cancer types were randomly split into training (75% of patients) and test (25% of patients) data. Additional datasets for testing were obtained for 4 different cancer types from cBio Cancer Genomics Portal [41] (namely, BRCA-METABRIC [42], COAD-CPTAC [43] and LUAD-ONCOSG [44]) and 4 different cancer types from ICGC (namely OV-AU, PACA-AU, PACA-CA, and PRAD-CA).

"True positive" driver genes were defined as the experimentally-validated driver genes curated by the Cancer Gene Census (CGC) [45]. Using ANNOVAR annotations, SCNA tables, and Phenolyzer gene scores, 26 features were generated:

The maximum coding variant impact score

The maximal coding variant metaprediction score for each gene was used as a feature in the driver gene classification model.

The maximum non-coding variant impact score

For non-coding variants, the Phred-scaled CADD scores were used. The maximal score for each gene was used as a feature in the driver gene classification model.

Proxy SCNA score

To score gene-level SCNA events, firstly, the gene-level SCNA events were determined using segment-level SCNA data. If a segment overlapped at least 25% (the default threshold value) of the gene, the gene-level SCNA event was called. The $\log_2$ ratio for any gene was determined as the ratio with the maximum $|\log_2 \text{ratio}|$ value over all segments overlapping the gene. Genes on sex chromosomes were excluded. Gene-level SCNA events with $|\log_2 \text{ratio}| < 0.25$ (the default threshold value) were also excluded.

A Minimal Common Region (MCR) is defined as the minimal region of copy number amplifications or deletions representing a common genomic alteration across the examined cancers [46]. We obtained pan-cancer MCR data from Kim et al. [47] who analyzed chromosomal aberrations in 8000 cancer genomes. Genes overlapping an MCR region with the same direction of SCNA event (amplification or deletion) were assigned the SCNA density (SCNA / Mb) of the MCR. This SCNA density was used as a feature in the driver gene classification model.

Hotspot or double-hit gene condition

Genes containing hotspot mutations (used as an indication of oncogenes) were determined using the Catalogue of Somatic Mutations in Cancer (COSMIC) [48] v91 occurrence annotations. A mutation with an occurrence greater than 5 (the default threshold) was defined as a hotspot mutation. Additionally, genes with a non-synonymous mutation and a homozygous copy number loss (defined as $|\log_2 \text{ratio}| < -1$) was defined as a double-hit gene (used as an indication of tumor suppressor genes). Whether or not a given gene was a hotspot gene or a double-hit gene was used as a feature in the driver gene classification model.

Phenotype based gene analyzer score

Phenotype Based Gene Analyzer (Phenolyzer) [24] is a tool for phenotype-based prioritization of candidate genes using prior biological knowledge and phenotype information. All genes from genomics data were used as an input for Phenolyzer and were scored based on previous biological knowledge regarding the specific cancer type. The cancer-specific Phenolyzer gene scores were used as a feature in the driver gene classification model.

Membership to cancer-related pathways

The final features for the driver gene classification model were whether or not the gene is in the given selected cancer-related KEGG [25] pathways. The cancer-related KEGG pathways were determined as pathways related to the "Pathways in cancer" pathway: PPAR signaling pathway, MAPK signaling pathway, Calcium signaling pathway, cAMP signaling pathway, Cytokine-cytokine receptor interaction, HIF-1 signaling pathway, Cell cycle, p53 signaling pathway, mTOR signaling pathway, PI3K-Akt signaling pathway, Apoptosis, Wnt signaling pathway, Notch signaling pathway, Hedgehog signaling pathway, TGF-beta signaling pathway, VEGF signaling pathway, Focal adhesion, ECM-receptor interaction, Adherens junction, JAK-STAT signaling pathway, Estrogen signaling pathway.

For training the MTL model, the R package RMTL [49] was used. We trained an MTL model with sparse structure (lasso regularization). The framework for the algorithm used by RMTL is provided in Equation 1:

$$\min_W \sum_{i=1}^{t} L(W_i|X_i, Y_i) + \lambda_1 ||W||_1 + \lambda_2 ||W||_F^2 \quad (1)$$

where *L()* is the logistic loss function. There are *t* tasks. *W* is the coefficient matrix, W_i is the ith column of *W* and refers to the coefficient vector of task *i*. *X* is the predictor matrices and Y is the response vectors of the *t* tasks. $||.||_1$ is the L_1 norm and $||.||_F$ is the Frobenius norm. λ_1 aims to control the effect of cross-task regularization and λ_2 stabilizes the numerical results and is used to improve generalization performance. Features for genes within the training data for all cancer types were further randomly split into training genes (75% of genes) and validation genes (25% of genes) datasets. The optimal λ_2 value was obtained through the assessment of performance on the validation dataset (determined to be 10^{-4}). After determining the optimal λ_1 value using 10-fold cross-validation (determined to be 10^{-4}), the final classification model was built.
For predicting the labels (i.e., "driver", "non-driver") using the probabilities of being a driver gene, cancer-type specific thresholds were determined as probability values maximizing accuracy on the validation datasets.
The performance of each sub-model was assessed by calculating AUC for each test dataset. Additionally, personalized analysis performances of the sub-models were assessed by calculating AUC per each individual in each test dataset.

2.3. Comparison with other batch analysis approaches

The performance of driveR was compared with the performances of the batch analysis approaches MutSigCV, DriverNet, OncodriveFML, and MutPanning by comparing AUC values.
MutSigCV (version 1.3.5) analyses were performed using default settings on the GenePattern platform [50]. DriverNet analyses were performed using DriverNet version 1.28.0 with the BioGRID [51, 52] Homo sapiens PIN (version 4.0.189) using the default settings. OncodriveFML analyses were performed using OncodriveFML

version 2.3.0 with the default settings. MutPanning (version 2.0) analyses were performed using default settings on the GenePattern platform.

2.4. Comparison with other personalized analysis approaches

The personalized analysis performance of driveR was also compared with the performances of the personalized analysis approaches with DawnRank and PRODIGY by comparing AUC values.

As both of these tools required normal tissue expression data, only 16 datasets, for which more than 1 normal tissue expression data were available, were used, namely: BLCA-US, BRCA-US, CESC-US, COAD-US, HNSC-US, KIRC-US, KIRP-US, LIHC-US, LUAD-US, LUSC-US, PAAD-US, PRAD-US, READ-US, STAD-US, THCA-US, and UCEC-US.

DawnRank (version 1.2) analyses were performed with the BioGRID Homo sapiens PIN (version 4.0.189) using the default settings. PRODIGY (version 1.0) analyses were performed with the STRING [53] Homo sapiens PIN (version 11.0) and with KEGG curated pathways using the default settings. PRODIGY gene scores per pathways were aggregated into a single gene score for comparison purposes.

3. Results

The variant impact metapredictor outperforms individual predictors

As described above, we trained a metapredictor model to estimate the pathogenicity of coding variants. Out of 6 different model types, the random forest metapredictor performed the best (Figure S2). As expected, this model also performed better than the individual impact prediction tools that were used as features to train the metapredictor (Figure 2). This model can be used to prioritize coding variants according to their pathogenicity and was used for generating a feature for the MTL classification model.

The cancer-type-specific driver gene prioritization approach performs well both for batch analysis and personalized analysis

Using genomic data and prior biological information, we next trained an MTL classification model for obtaining cancer-type-specific driver gene predictions.

The batch analysis performance of the MTL classification model was assessed on the 28 test datasets (Figure 3). The median AUC on the test datasets was 0.683 (range=0.648 – 0.801, Figure 3A). As expected, the percentages of "True Driver Genes" (as curated by CGC) and "Actionable Genes" (as curated by TARGET [54]) over all predicted driver genes across all datasets increased with increasing threshold values (Figure 3B, range of median percentages of "True Driver Genes" = 66.42-100%, range of median percentages of "Actionable Genes" = 56.12-87.5%). Using the cancer-type specific thresholds, a median of 75.6% predicted driver genes were found to be "True Driver Genes" (Figure 3B left-hand panel), and 66.67% predicted driver genes were found to be "Actionable Genes" (Figure 3B right-hand panel).

Next, the personalized analysis performance of the driveR approach was assessed on the 5091 test patients (Figure 4). The median AUC value over all test patients was 0.756 (Figure 4A, top, range = 0-1). The median AUC values per test dataset were similar (Figure 4A bottom, range of median AUC per dataset = 0.666-1). The median percentages of "True Driver Genes" and "Actionable Genes" over all predicted driver genes in each patient were 100% (Figure 4B). Using the cancer-type specific thresholds, the median percentages were again 100%. This implies that each gene predicted by driveR to be a driver in an individual is a true driver gene and most likely an actionable gene.

The driver gene predictor approach outperforms existing approaches

The ROC curves of the performances of different batch analysis approaches per each dataset are presented in Figure S3. The median AUC of driveR across all datasets (median AUC = 0.683) was significantly higher than all of MutSigCV (0.579, Wilcox rank sum test $p < 0.001$), DriverNet (0.614, $p < 0.001$), OncodriveFML (0.546, $p < 0.001$) and MutPanning (0.591, $p < 0.001$) (Figure 5).

Next, performances of driveR and other personalized analysis tools on all test patients were compared (Figure 6). It was observed that driveR had higher median AUC (0.716) compared to DawnRank (median AUC = 0.694, $p < 0.001$) and PRODIGY (median AUC = 0.679, $p < 0.001$) overall. When the performances were compared for all patients per dataset, driveR again displayed higher performance (Figure S4).

4. Discussion

Alterations in driver genes are the putative underlying causes of oncogenesis and tumor formation [3, 4]. Although numerous driver genes are experimentally validated [55-58], there is great potential clinical benefit in identifying driver genes in individual patients [5, 6].

In this study, we established a simple, model-based approach, driveR, that can accurately prioritize cancer driver genes, sorting through a vast amount of passengers observed in cancer genomes [1]. Using 26 features based on somatic genomics data, we trained a multi-task learning classification model to estimate driver gene probabilities in a cancer-type-specific context for 21 different cancer types. Compared to other approaches, driveR achieved better performance on different test datasets by accurately prioritizing cancer driver genes in analyses on both cohort-scale and personalized data. Below we discuss several unique aspects of driveR.

Driver genes are diverse among different cancer types [55-58]. Different cancers may possess different driver genes. The driveR approach attempts to define cancer-type-specific driver genes based on somatic genomic features, incorporating prior biological knowledge. The multi-task learning model at the core of our approach allows for cancer-type-specific identification of driver genes. The approach is

available for use for 21 different cancer types. Analysis in a cancer-type-specific context allows driveR to more accurately identify driver genes that may be specific to a certain cancer type.

Most approaches for driver gene prioritization are designed for analysis of tumor cohorts. These approaches usually fail to identify low-frequency driver genes, not to mention patient-specific driver genes. Patient-specific driver genes may be rare or not match the disease organ. In this study, we demonstrate that driveR can be utilized to rapidly and accurately analyze both patient-specific and cohort-scale genomic data to prioritize cancer driver genes. This makes driveR a suitable option for studying driver genes for individual patients.

In our approach, we incorporated prior biological knowledge into the MTL model. We used Phenolyzer, a database-mining tool, which integrates 15 different biological knowledge databases to score a gene's prior association with cancer. Additionally, memberships of genes to cancer-related KEGG pathways were also taken into consideration. This integration of extensive biological knowledge improved the accuracy of the MTL model because the final model used for driver gene prioritization was based not only on the genomic features but also guided by expert knowledge from decades of research.

We devised driveR to be less demanding on the data type. Other driver prioritization tools usually require multiple omics data for the patient, including genomic mutation data, tumor expression data, and normal expression data. However, obtaining transcriptomics data is not always feasible due to cost and other practical issues. We devised driveR based only on somatic genomics data (somatic mutation and SCNA) because targeted sequencing or whole exome/genome sequencing is more widely utilized in both the clinical and research settings and technical analysis of genomic data is less complicated compared to transcriptomic data.

We demonstrated that the overall performance of driveR in identifying true drivers is adequate on both batch analyses and personalized analyses. Additionally, a high proportion of driveR predictions were also clinically-actionable genes. We also demonstrate that driveR outperforms other tools in both personalized and batch analyses. These demonstrate that driveR performs well in successfully prioritizing driver genes in a cancer-type-specific context.

5. Conclusions

In this study, we devised a novel approach, driveR, for prioritizing cancer-type-specific driver genes using somatic genomic data. As demonstrated, driveR can be utilized for both analyzing individual cancers or cancer cohorts to accurately prioritize patient-specific or cohort-scale driver genes. We also demonstrated that our approach outperforms existing driver gene prioritization methods. We hope that this approach can provide further insight into cancer driver gene discovery and help progress personalized cancer research.

6. List of abbreviations

PIN: Protein-protein interaction network
SCS: Single-sample controller strategy
DEG: differentially-expressed gene
KEGG: Kyoto Encyclopedia of Genes and Genomes
SVM: Support Vector Machine
ROC: Receiver Operating Characteristic
AUC: Area Under Curve
MTL: Multi-task learning
SCNA: Somatic Copy Number Alteration
TCGA: The Cancer Genome Atlas
ICGC: International Cancer Genome Consortium
CGC: Cancer Gene Census
MCR: Minimal Common Region
COSMIC: Catalogue of Somatic Mutations in Cancer
Phenolyzer: Phenotype Based Gene Analyzer

7. Declarations

Ethics approval and consent to participate
Not applicable
Consent for publication
Not applicable
Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Project name: driveR
Project home page: https://github.com/egeulgen/driveR
Operating system(s): Platform independent
Programming language: R
License: MIT license

Competing interests
The authors declare that they have no competing interests.
Funding
This research received no external funding.
Authors' contributions
Conceptualization, EU and OUS; implementation, analyses, preparing the manuscript draft, EU; study supervision, OUS; review and editing, both authors. Both authors have read and agreed to the published version of the manuscript.

Acknowledgements
Not applicable

8. Figure Legends

Figure 1. The overview of the driveR approach for driver gene prioritization.
Figure 2. Metapredictor of coding variant impact outperforms individual predictors. ROC curves displaying the performances of the random forest (RF) metapredictor and individual impact prediction tools.
Figure 3. Batch analysis performance of driveR on 28 test datasets. (A) ROC curves of driveR prioritization results per each test dataset. AUC value per test dataset is indicated in the bottom-right legend. Median AUC and range are indicated on the top-left. (B) Boxplots of percentages of "True Driver Genes" (left) and "Actionable Genes" (right) over all predicted driver genes by using the specified thresholds across all datasets. "cancer-specific" indicates that the cancer-type-specific thresholds were used to predict driver genes.
Figure 4. Personalized analysis performance of driveR on 5091 patients. (A) Boxplot displaying the distribution of AUC values in all patients (top). Boxplots displaying the distributions of AUC values per individual in each of the 28 test datasets. (B) Boxplots of percentages of "True Driver Genes" (left) and "Actionable Genes" (right) over all predicted driver genes by using the specified thresholds across all test patients. "cancer-specific" indicates that the cancer-type-specific thresholds were used to predict driver genes.
Figure 5. Comparison of performance of driveR with batch analysis approaches. Boxplots displaying the distributions of AUC values of each approach across the 28 test datasets. The brackets display p-values per each comparison.
Figure 6. Comparison of performance of driveR with personalized analysis approaches. Boxplots displaying the distributions of AUC values of each approach across all test patients. The brackets display p-values per each comparison.
Figure S1. Overall analysis of individual variant impact predictors. (A) Correlogram displaying correlations between the outcome and individual variant impact predictors. (B) Violin plots displaying the distributions of scores of individual variant impact predictors in driver and non-driver variants.
Figure S2. Performance of different coding variant impact metapredictor models in the training (A) and test (B) datasets
Figure S3. Comparison of performance of driveR with batch analysis approaches per test dataset. ROC curves for assessing the performance of each approach per each test dataset. The bottom-right legends display AUC per each approach.
Figure S4. Comparison of performance of driveR with personalized analysis approaches per test dataset. Boxplots displaying the distributions of AUC values of each approach across all patients per test dataset. The bracket display p values per each comparison.

9. References

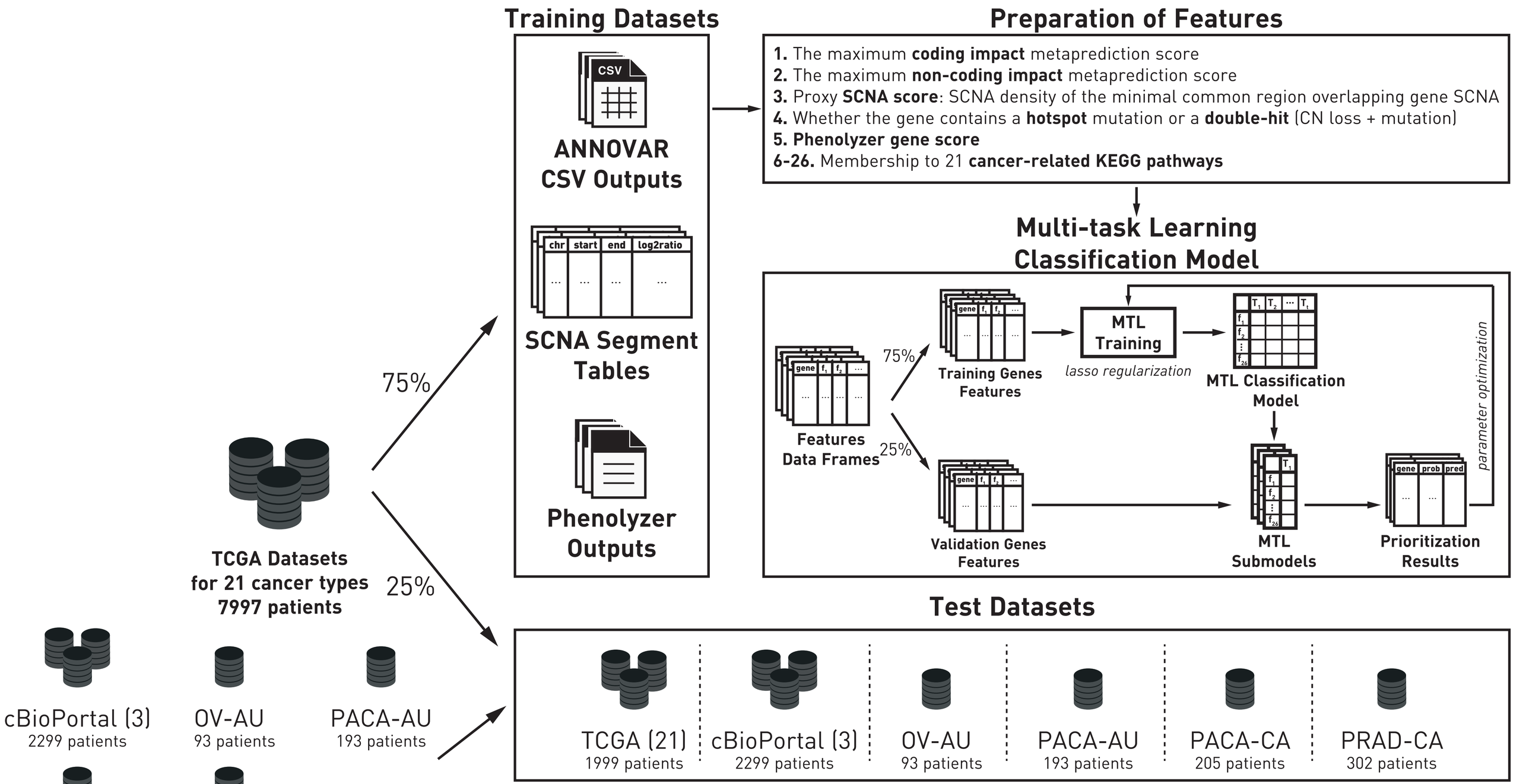
1. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome**. *Nature* 2009, **458**(7239):719-724.
2. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C *et al*: **Patterns of somatic mutation in human cancer genomes**. *Nature* 2007, **446**(7132):153-158.
3. Haber DA, Settleman J: **Cancer: drivers and passengers**. *Nature* 2007, **446**(7132):145-146.
4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: **Cancer Genome Landscapes**. *Science* 2013, **339**(6127):1546-1558.
5. Chin L, Andersen JN, Futreal PA: **Cancer genomics: from discovery science to personalized medicine**. *Nat Med* 2011, **17**(3):297-303.
6. Schilsky RL: **Personalized medicine in oncology: the future is now**. *Nature Reviews Drug Discovery* 2010, **9**(5):363-366.
7. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R: **Evaluating the evaluation of cancer driver genes**. *Proc Natl Acad Sci U S A* 2016, **113**(50):14330-14335.
8. Cheng F, Zhao J, Zhao Z: **Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes**. *Brief Bioinform* 2016, **17**(4):642-656.
9. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER *et al*: **MuSiC: identifying mutational significance in cancer genomes**. *Genome Res* 2012, **22**(8):1589-1598.
10. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA *et al*: **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457):214-218.
11. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR: **Identification of cancer driver genes based on nucleotide context**. *Nature Genetics* 2020, **52**(2):208-218.
12. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules**. *Genome Res* 2012, **22**(2):398-406.
13. Reyna MA, Leiserson MDM, Raphael BJ: **Hierarchical HotNet: identifying hierarchies of altered subnetworks**. *Bioinformatics* 2018, **34**(17):i972-i980.
14. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM: **Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)**. *Bioinformatics* 2013, **29**(21):2757-2764.
15. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP: **DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer**. *Genome Biology* 2012, **13**(12):R124.

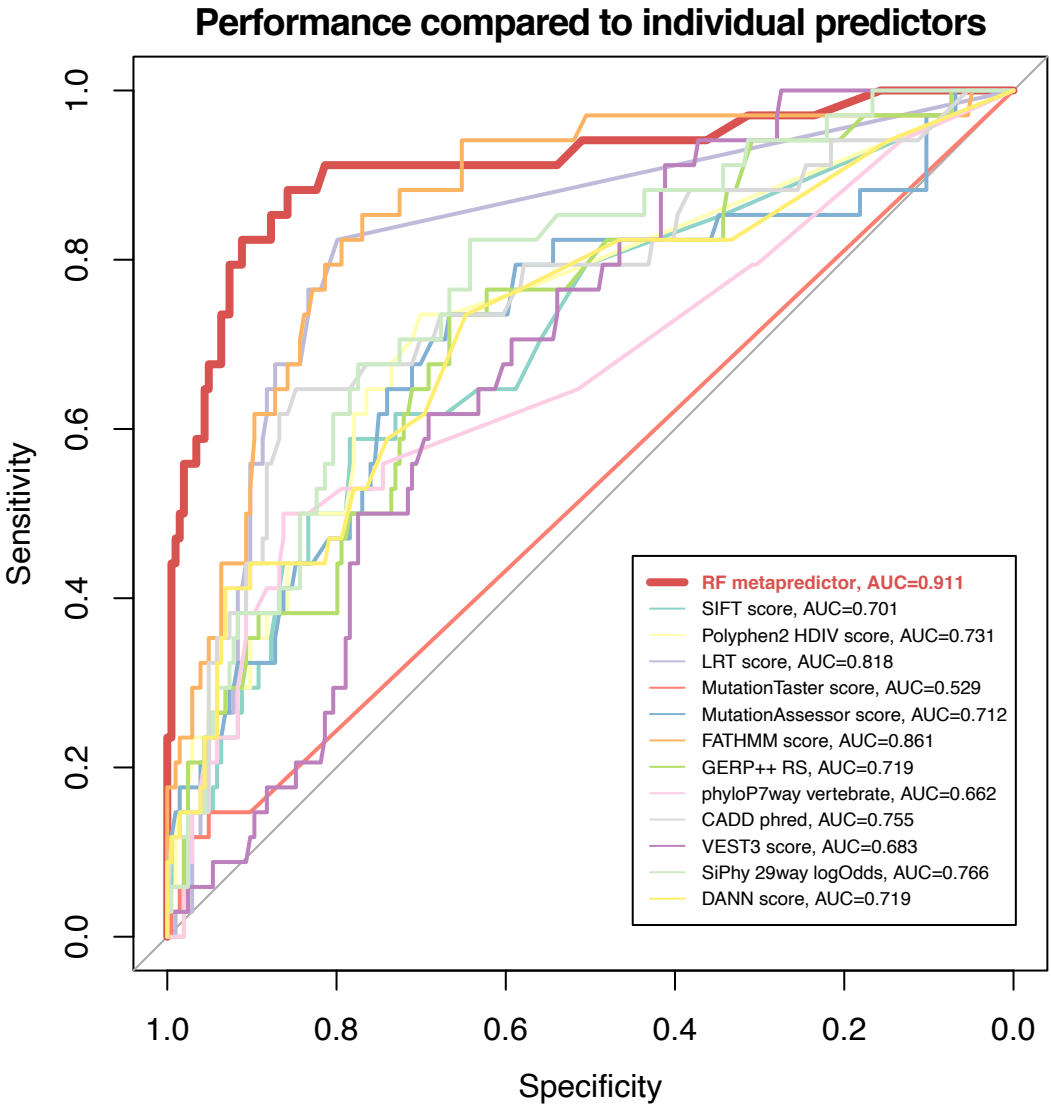
16. Kartha VK, Sebastiani P, Kern JG, Zhang L, Varelas X, Monti S: **CaDrA: A Computational Framework for Performing Candidate Driver Analyses Using Genomic Features**. *Frontiers in Genetics* 2019, **10**(121).
17. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N: **OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations**. *Genome Biology* 2016, **17**(1):128.
18. Kroschinsky F, Stölzel F, von Bonin S, Beutel G, Kochanek M, Kiehl M, Schellongowski P: **New drugs, new toxicities: severe side effects of modern targeted and immunotherapy of cancer and their management**. *Crit Care* 2017, **21**(1):89.
19. Park SR, Davis M, Doroshow JH, Kummar S: **Safety and feasibility of targeted agent combinations in solid tumours**. *Nature Reviews Clinical Oncology* 2013, **10**(3):154-168.
20. Hou JP, Ma J: **DawnRank: discovering personalized driver genes in cancer**. *Genome Med* 2014, **6**(7):56.
21. Guo WF, Zhang SW, Liu LL, Liu F, Shi QQ, Zhang L, Tang Y, Zeng T, Chen L: **Discovering personalized driver mutation profiles of single samples in cancer by network control strategy**. *Bioinformatics* 2018, **34**(11):1893-1903.
22. Dong C, Guo Y, Yang H, He Z, Liu X, Wang K: **iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes**. *Genome Medicine* 2016, **8**(1):135.
23. Dinstag G, Shamir R: **PRODIGY: personalized prioritization of driver genes**. *Bioinformatics* 2019, **36**(6):1831-1839.
24. Yang H, Robinson PN, Wang K: **Phenolyzer: phenotype-based prioritization of candidate genes for human diseases**. *Nat Methods* 2015, **12**(9):841-843.
25. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
26. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, Shen R, Norton L, Reis-Filho JS, Weigelt B: **Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations**. *Genome Biol* 2014, **15**(10):484.
27. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Research* 2010, **38**(16):e164-e164.
28. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**(13):3812-3814.
29. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.
30. Chun S, Fay JC: **Identification of deleterious mutations within three human genomes**. *Genome Res* 2009, **19**(9):1553-1561.
31. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age**. *Nature Methods* 2014, **11**(4):361-362.

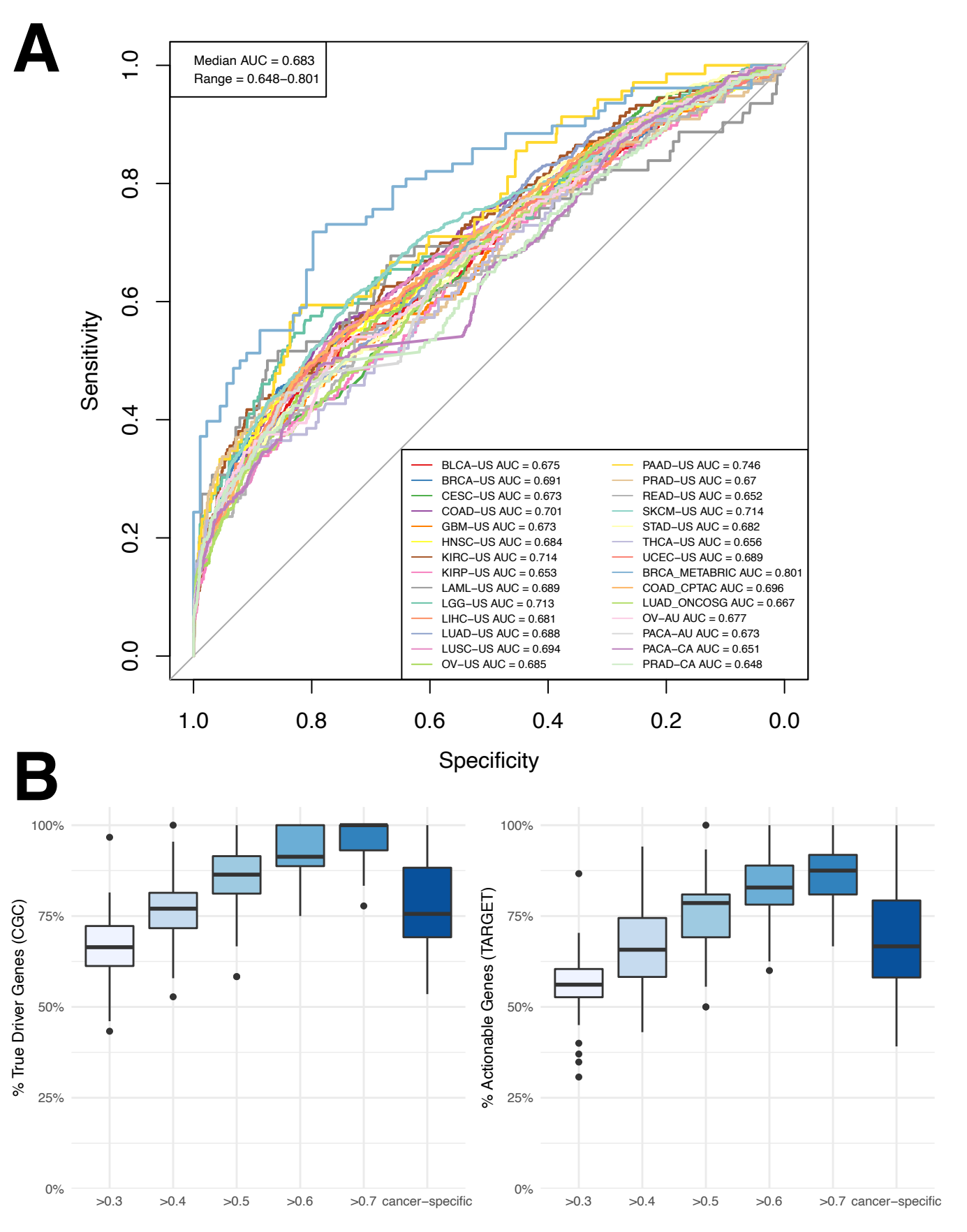
32. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics**. *Nucleic Acids Research* 2011, **39**(17):e118-e118.
33. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR: **Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models**. *Hum Mutat* 2013, **34**(1):57-65.
34. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S: **Identifying a high fraction of the human genome to be under selective constraint using GERP++**. *PLoS Comput Biol* 2010, **6**(12):e1001025.
35. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**(1):110-121.
36. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M: **CADD: predicting the deleteriousness of variants throughout the human genome**. *Nucleic Acids Res* 2019, **47**(D1):D886-d894.
37. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R: **Identifying Mendelian disease genes with the variant effect scoring tool**. *BMC Genomics* 2013, **14 Suppl 3**(Suppl 3):S3.
38. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X: **Identifying novel constrained elements by exploiting biased substitution patterns**. *Bioinformatics* 2009, **25**(12):i54-i62.
39. Quang D, Chen Y, Xie X: **DANN: a deep learning approach for annotating the pathogenicity of genetic variants**. *Bioinformatics* 2015, **31**(5):761-763.
40. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B *et al*: **International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data**. *Database (Oxford)* 2011, **2011**:bar026.
41. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. *Cancer Discov* 2012, **2**(5):401-404.
42. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ *et al*: **The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes**. *Nat Commun* 2016, **7**:11479.
43. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, Dou Y, Zhang Y, Shi Z, Arshad OA *et al*: **Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities**. *Cell* 2019, **177**(4):1035-1049.e1019.
44. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, Alvarez JJS, Lu B, Lim JQ, Takano A *et al*: **Genomic landscape of lung adenocarcinoma in East Asians**. *Nat Genet* 2020, **52**(2):177-186.

45. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA: **The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers**. *Nat Rev Cancer* 2018, **18**(11):696-705.
46. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS: **A census of amplified and overexpressed human cancer genes**. *Nat Rev Cancer* 2010, **10**(1):59-64.
47. Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ: **Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes**. *Genome Res* 2013, **23**(2):217-227.
48. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E *et al*: **COSMIC: the Catalogue Of Somatic Mutations In Cancer**. *Nucleic Acids Research* 2018, **47**(D1):D941-D947.
49. Cao H, Zhou J, Schwarz E: **RMTL: an R library for multi-task learning**. *Bioinformatics* 2018, **35**(10):1797-1798.
50. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0**. *Nat Genet* 2006, **38**(5):500-501.
51. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Res* 2006, **34**(Database issue):D535-539.
52. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R *et al*: **The BioGRID interaction database: 2019 update**. *Nucleic Acids Research* 2018, **47**(D1):D529-D541.
53. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P *et al*: **STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets**. *Nucleic Acids Res* 2019, **47**(D1):D607-d613.
54. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, Jane-Valbuena J, Friedrich DC, Kryukov G, Carter SL *et al*: **Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine**. *Nat Med* 2014, **20**(6):682-688.
55. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B *et al*: **Comprehensive Characterization of Cancer Driver Genes and Mutations**. *Cell* 2018, **173**(2):371-385.e318.
56. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N: **IntOGen-mutations identifies cancer drivers across tumor types**. *Nat Methods* 2013, **10**(11):1081-1082.
57. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L *et al*: **Comprehensive identification of mutational cancer driver genes across 12 tumor types**. *Scientific Reports* 2013, **3**(1):2650.

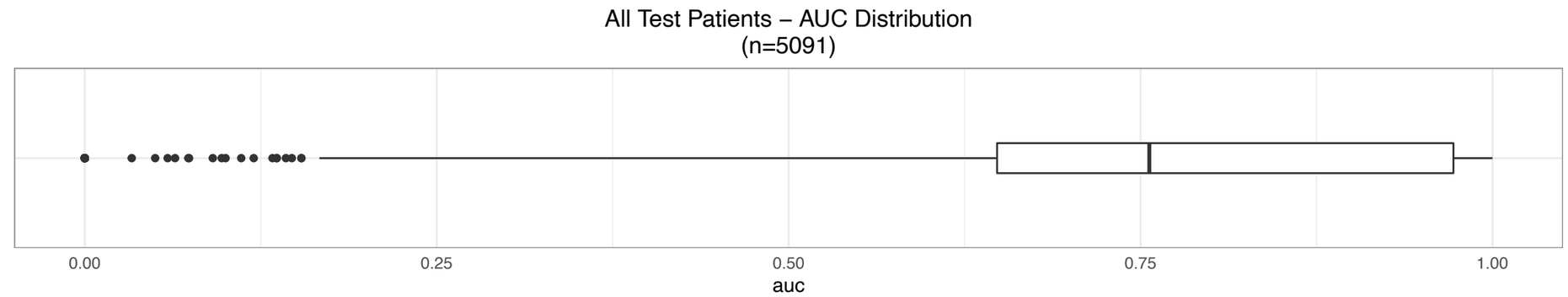
58. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H *et al*: **A compendium of mutational cancer driver genes**. *Nature Reviews Cancer* 2020, **20**(10):555-572.

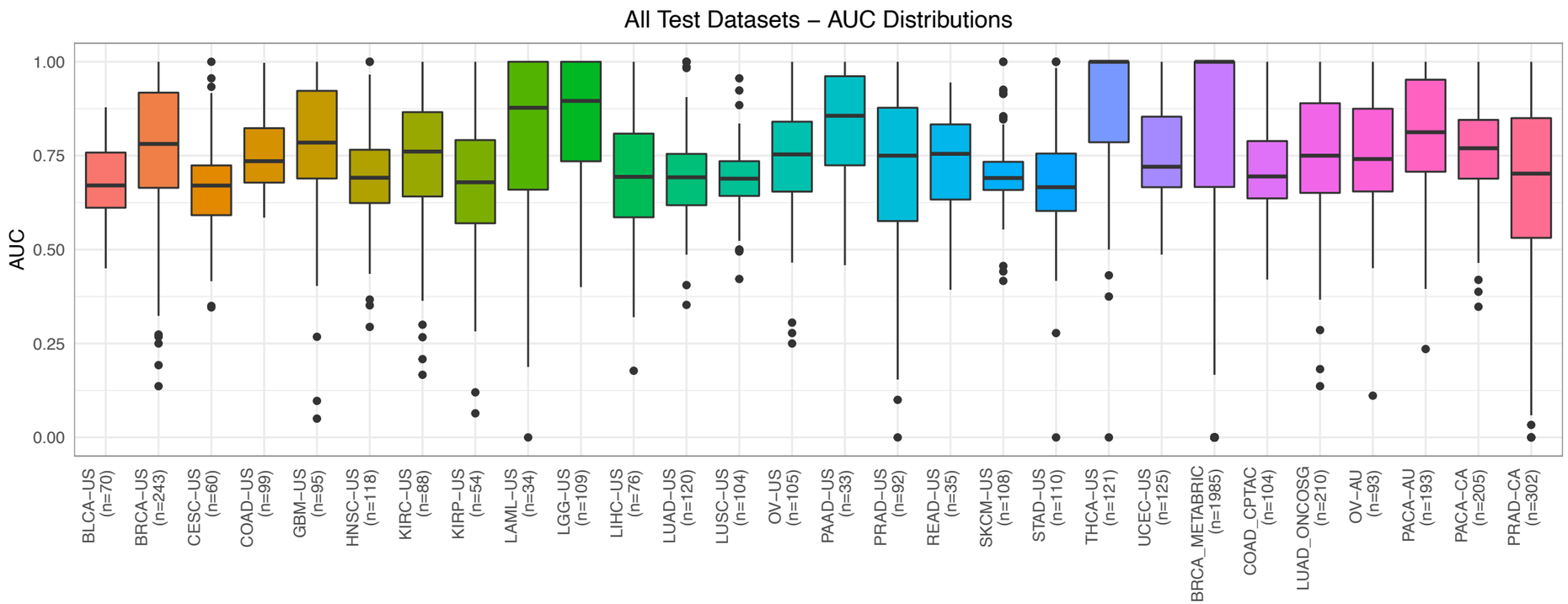

Training Datasets
CSV
ANNOVAR CSV Outputs
chr start end log2ratio
SCNA Segment Tables
Phenolyzer Outputs
Preparation of Features
1. The maximum coding impact metaprediction score
2. The maximum non-coding impact metaprediction score
3. Proxy SCNA score: SCNA density of the minimal common region overlapping gene SCNA
4. Whether the gene contains a hotspot mutation or a double-hit (CN loss + mutation)
5. Phenolyzer gene score
6-26. Membership to 21 cancer-related KEGG pathways
Multi-task Learning Classification Model
Features Data Frames
75%
25%
Training Genes Features
MTL Training
lasso regularization
MTL Classification Model
Validation Genes Features
MTL Submodels
Prioritization Results
parameter optimization
TCGA Datasets for 21 cancer types 7997 patients
75%
25%
Test Datasets
cBioPortal (3) 2299 patients
OV-AU 93 patients
PACA-AU 193 patients
PACA-CA 205 patients
PRAD-CA 302 patients
TCGA (21) 1999 patients
cBioPortal (3) 2299 patients
OV-AU 93 patients
PACA-AU 193 patients
PACA-CA 205 patients
PRAD-CA 302 patients


Performance compared to individual predictors
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
RF metapredictor, AUC=0.911
SIFT score, AUC=0.701
Polyphen2 HDIV score, AUC=0.731
LRT score, AUC=0.818
MutationTaster score, AUC=0.529
MutationAssessor score, AUC=0.712
FATHMM score, AUC=0.861
GERP++ RS, AUC=0.719
phyloP7way vertebrate, AUC=0.662
CADD phred, AUC=0.755
VEST3 score, AUC=0.683
SiPhy 29way logOdds, AUC=0.766
DANN score, AUC=0.719

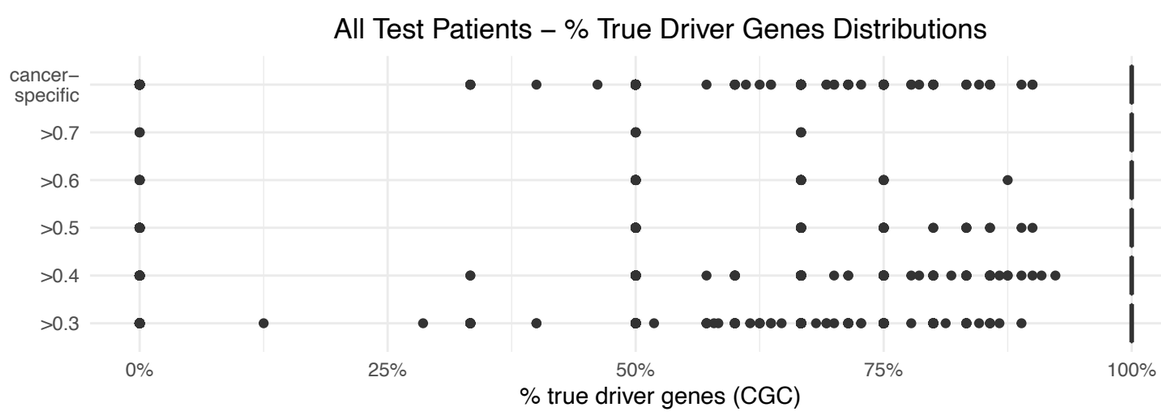

A
Median AUC = 0.683
Range = 0.648–0.801
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
BLCA–US AUC = 0.675
BRCA–US AUC = 0.691
CESC–US AUC = 0.673
COAD–US AUC = 0.701
GBM–US AUC = 0.673
HNSC–US AUC = 0.684
KIRC–US AUC = 0.714
KIRP–US AUC = 0.653
LAML–US AUC = 0.689
LGG–US AUC = 0.713
LIHC–US AUC = 0.681
LUAD–US AUC = 0.688
LUSC–US AUC = 0.694
OV–US AUC = 0.685
PAAD–US AUC = 0.746
PRAD–US AUC = 0.67
READ–US AUC = 0.652
SKCM–US AUC = 0.714
STAD–US AUC = 0.682
THCA–US AUC = 0.656
UCEC–US AUC = 0.689
BRCA_METABRIC AUC = 0.801
COAD_CPTAC AUC = 0.696
LUAD_ONCOSG AUC = 0.667
OV–AU AUC = 0.677
PACA–AU AUC = 0.673
PACA–CA AUC = 0.651
PRAD–CA AUC = 0.648
B
% True Driver Genes (CGC)
% Actionable Genes (TARGET)
100%
75%
50%
25%
0%
>0.3
>0.4
>0.5
>0.6
>0.7
cancer–specific

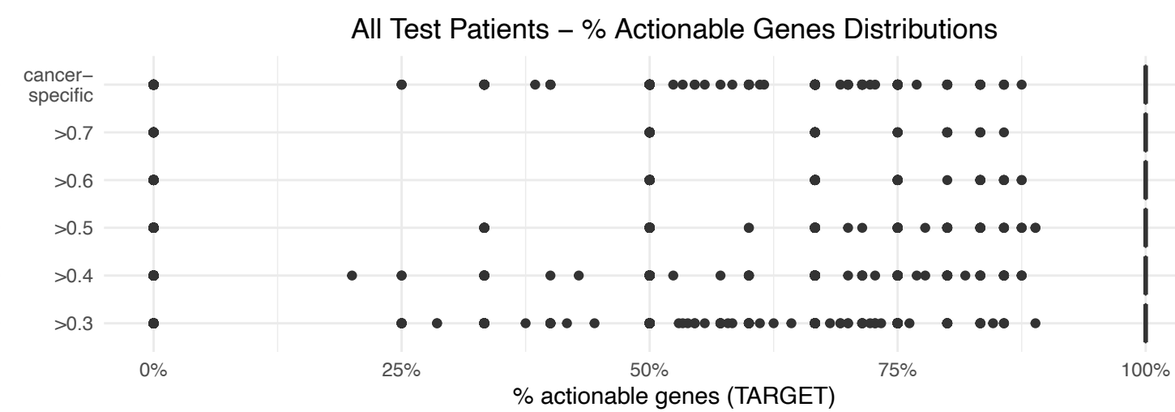
A

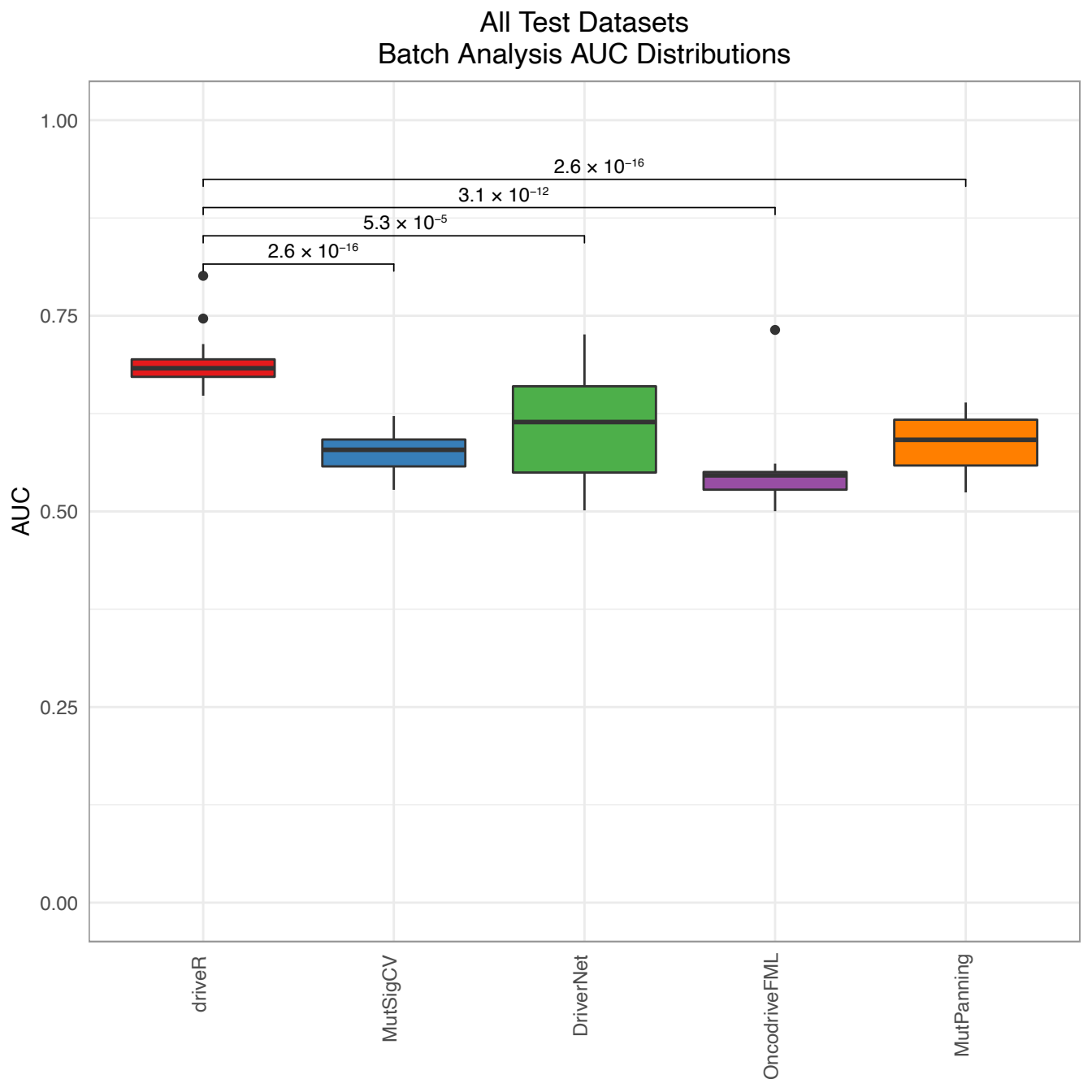

All Test Patients – AUC Distribution
(n=5091)
0.00
0.25
0.50
0.75
1.00
auc

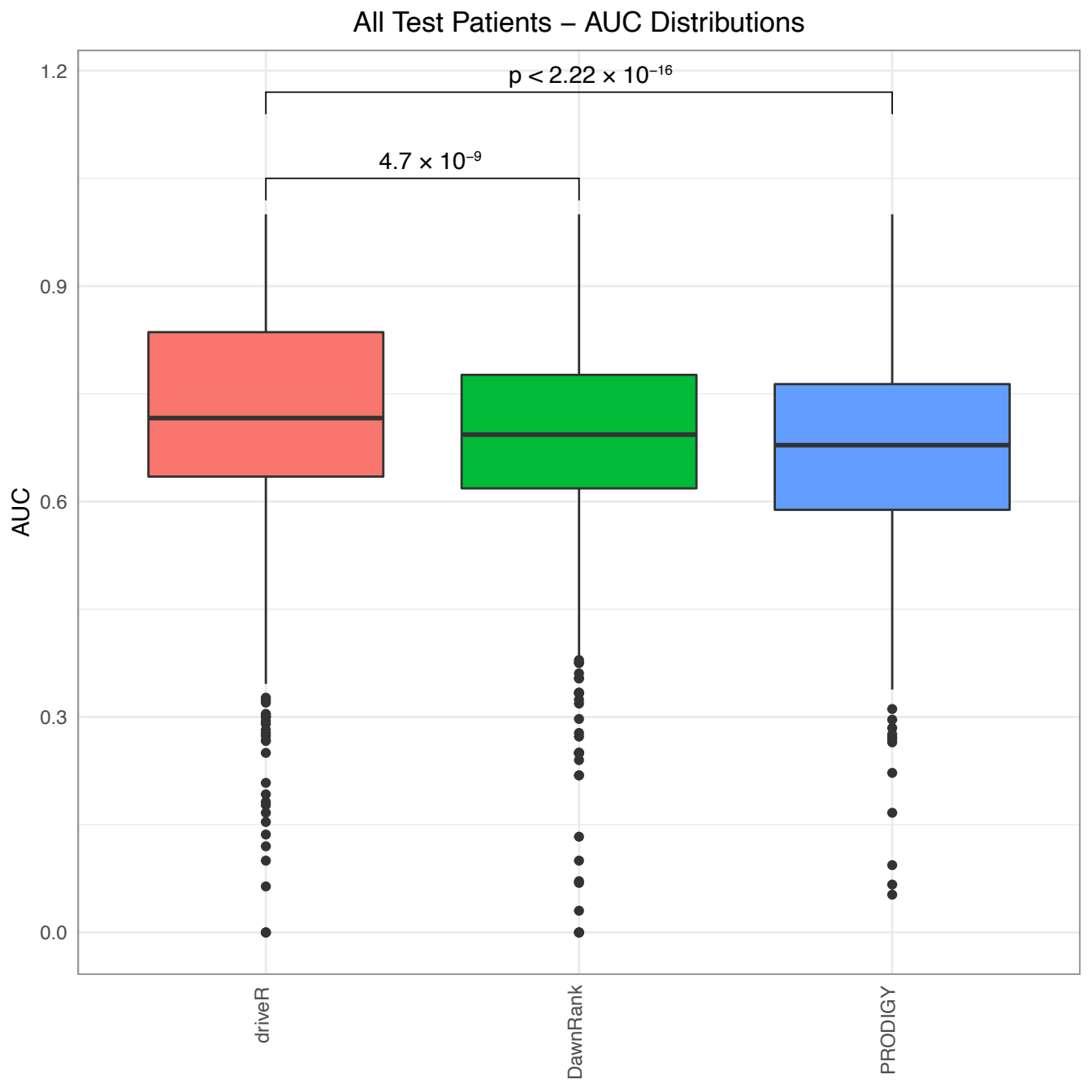

All Test Datasets – AUC Distributions
AUC
1.00
0.75
0.50
0.25
0.00
BLCA–US (n=70)
BRCA–US (n=243)
CESC–US (n=60)
COAD–US (n=99)
GBM–US (n=95)
HNSC–US (n=118)
KIRC–US (n=88)
KIRP–US (n=54)
LAML–US (n=34)
LGG–US (n=109)
LIHC–US (n=76)
LUAD–US (n=120)
LUSC–US (n=104)
OV–US (n=105)
PAAD–US (n=33)
PRAD–US (n=92)
READ–US (n=35)
SKCM–US (n=108)
STAD–US (n=110)
THCA–US (n=121)
UCEC–US (n=125)
BRCA_METABRIC (n=1985)
COAD_CPTAC (n=104)
LUAD_ONCOSG (n=210)
OV–AU (n=93)
PACA–AU (n=193)
PACA–CA (n=205)
PRAD–CA (n=302)

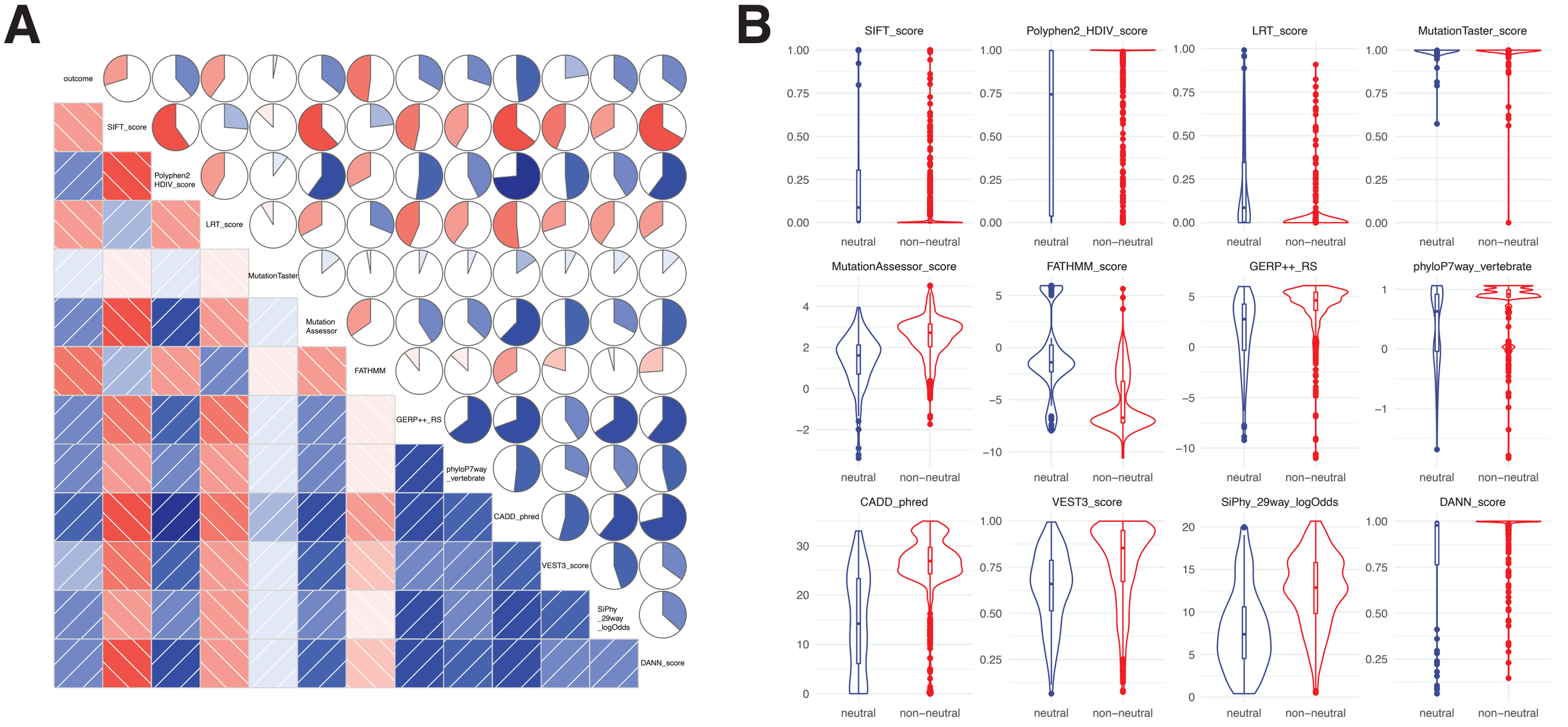
B

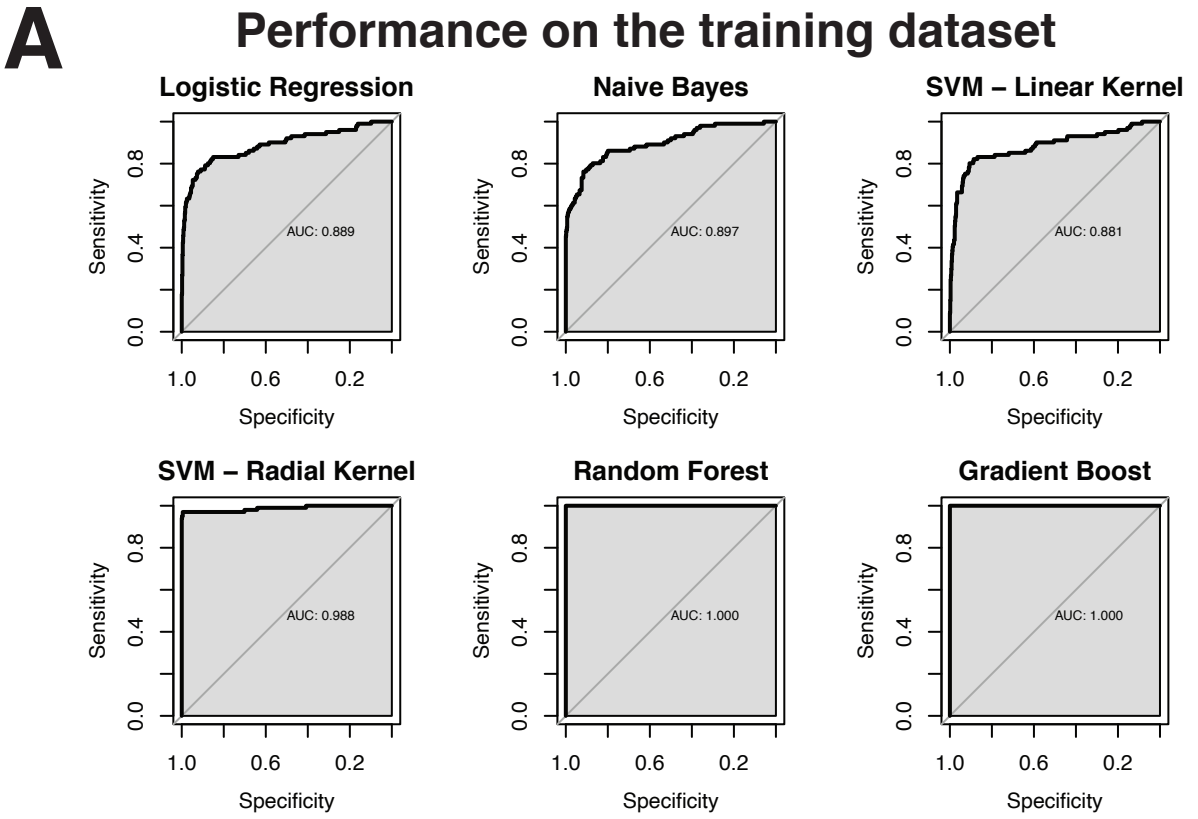

All Test Patients – % True Driver Genes Distributions
cancer–specific
>0.7
>0.6
>0.5
>0.4
>0.3
0%
25%
50%
75%
100%
% true driver genes (CGC)

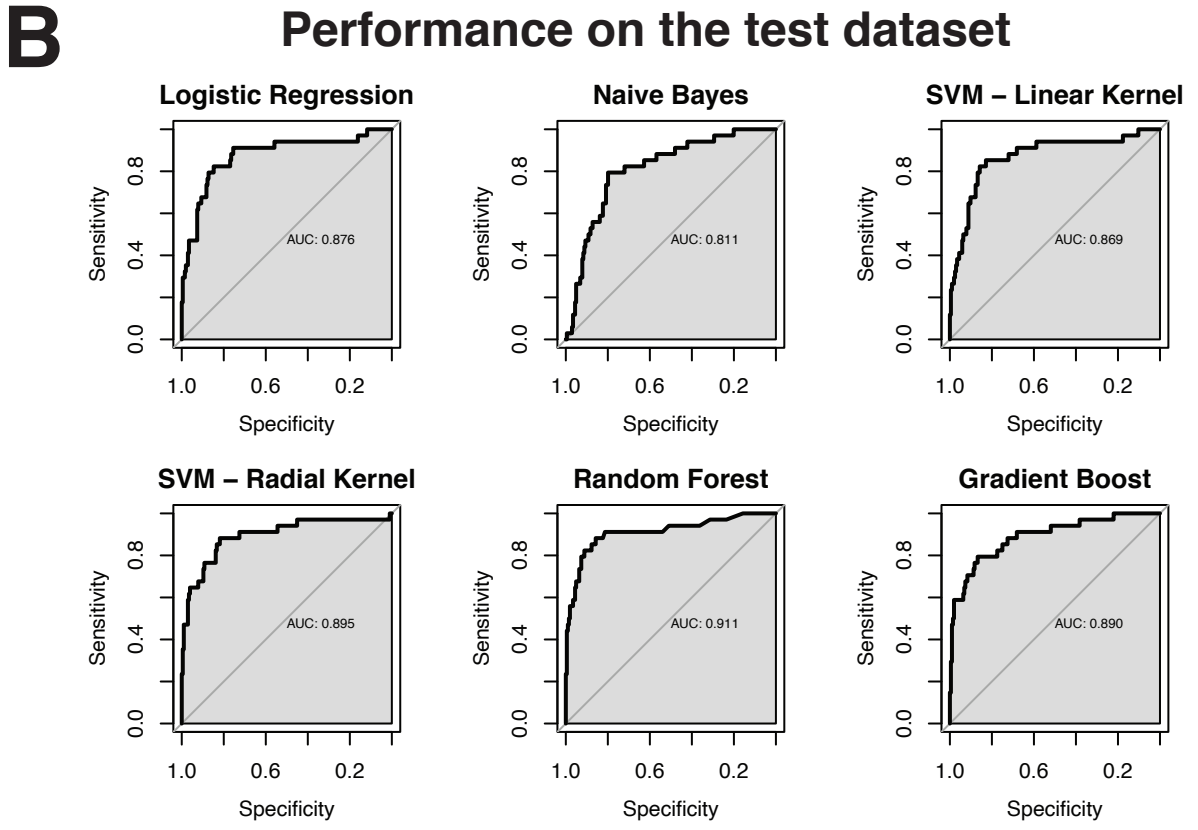

All Test Patients – % Actionable Genes Distributions
cancer–specific
>0.7
>0.6
>0.5
>0.4
>0.3
0%
25%
50%
75%
100%
% actionable genes (TARGET)

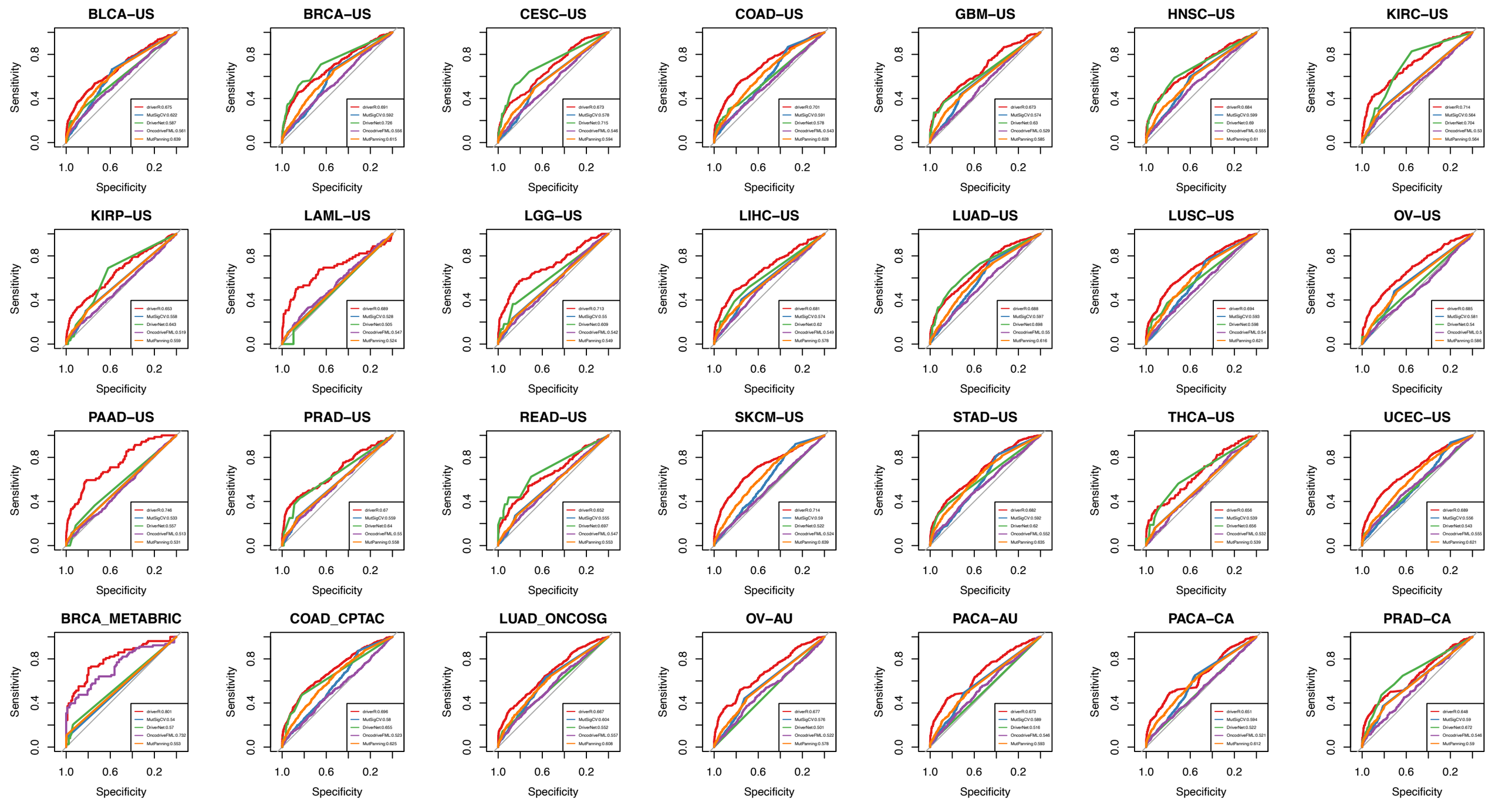

All Test Datasets
Batch Analysis AUC Distributions
AUC
1.00
0.75
0.50
0.25
0.00
2.6 × 10⁻¹⁶
3.1 × 10⁻¹²
5.3 × 10⁻⁵
2.6 × 10⁻¹⁶
driveR
MutSigCV
DriverNet
OncodriveFML
MutPanning

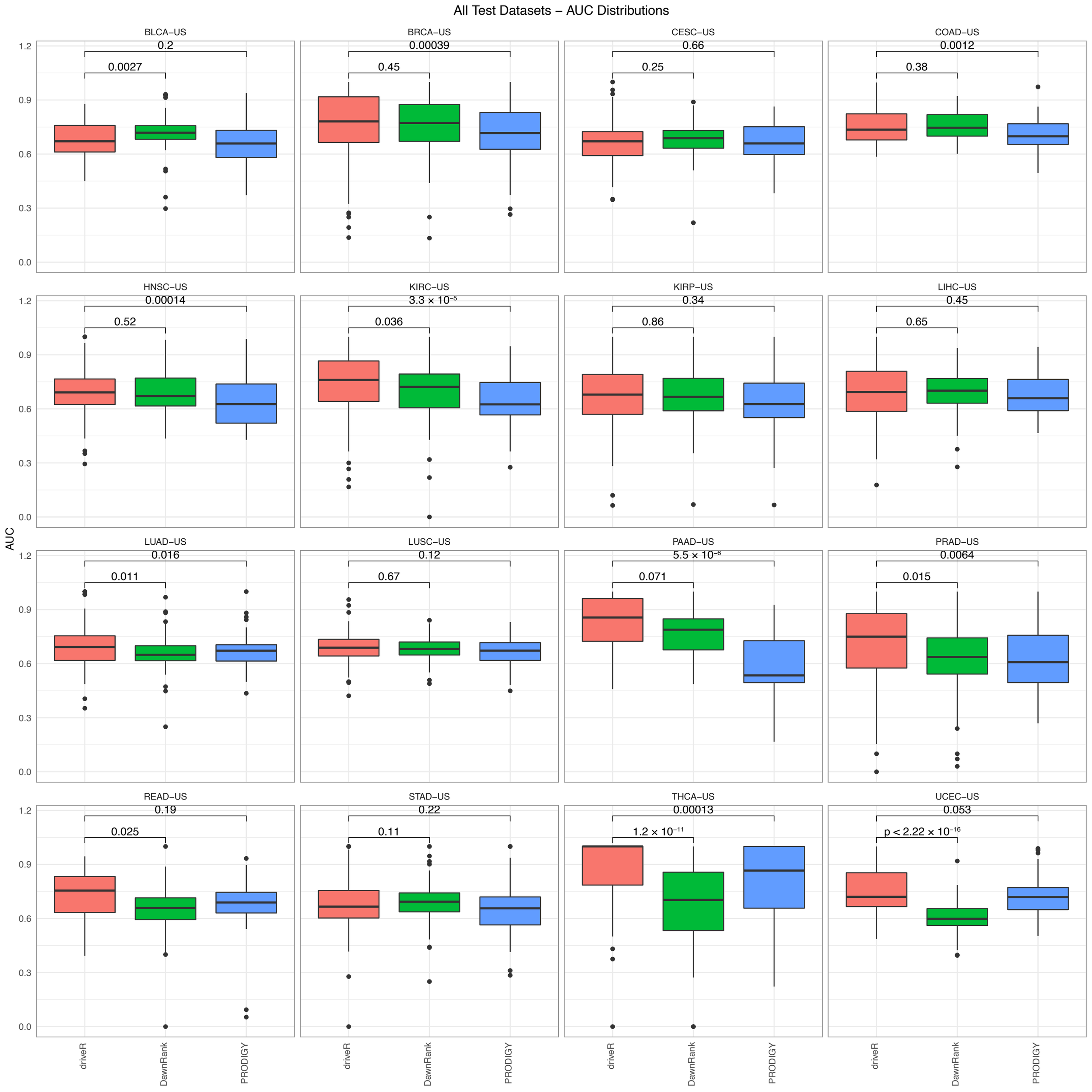

All Test Patients – AUC Distributions
p < 2.22 × 10⁻¹⁶
4.7 × 10⁻⁹
AUC
1.2
0.9
0.6
0.3
0.0
driveR
DawnRank
PRODIGY


A
outcome
SIFT_score
Polyphen2 HDIV_score
LRT_score
MutationTaster
Mutation Assessor
FATHMM
GERP++_RS
phyloP7way _vertebrate
CADD_phred
VEST3_score
SiPhy _29way _logOdds
DANN_score
B
SIFT_score
Polyphen2_HDIV_score
LRT_score
MutationTaster_score
MutationAssessor_score
FATHMM_score
GERP++_RS
phyloP7way_vertebrate
CADD_phred
VEST3_score
SiPhy_29way_logOdds
DANN_score
neutral
non−neutral


A
Performance on the training dataset
Logistic Regression
AUC: 0.889
Naive Bayes
AUC: 0.897
SVM – Linear Kernel
AUC: 0.881
SVM – Radial Kernel
AUC: 0.988
Random Forest
AUC: 1.000
Gradient Boost
AUC: 1.000
Sensitivity
Specificity
0.0
0.4
0.8
1.0
0.6
0.2


B
Performance on the test dataset
Logistic Regression
AUC: 0.876
Naive Bayes
AUC: 0.811
SVM – Linear Kernel
AUC: 0.869
SVM – Radial Kernel
AUC: 0.895
Random Forest
AUC: 0.911
Gradient Boost
AUC: 0.890
Sensitivity
Specificity
0.0
0.4
0.8
1.0
0.6
0.2


BLCA–US
driverR:0.675
MutSigCV:0.622
DriverNet:0.587
OncodriveFML:0.561
MutPanning:0.639
BRCA–US
driverR:0.691
MutSigCV:0.592
DriverNet:0.726
OncodriveFML:0.556
MutPanning:0.615
CESC–US
driverR:0.673
MutSigCV:0.578
DriverNet:0.715
OncodriveFML:0.546
MutPanning:0.594
COAD–US
driverR:0.701
MutSigCV:0.591
DriverNet:0.578
OncodriveFML:0.543
MutPanning:0.628
GBM–US
driverR:0.673
MutSigCV:0.574
DriverNet:0.63
OncodriveFML:0.529
MutPanning:0.585
HNSC–US
driverR:0.684
MutSigCV:0.599
DriverNet:0.69
OncodriveFML:0.555
MutPanning:0.61
KIRC–US
driverR:0.714
MutSigCV:0.564
DriverNet:0.704
OncodriveFML:0.53
MutPanning:0.564
KIRP–US
driverR:0.653
MutSigCV:0.558
DriverNet:0.643
OncodriveFML:0.519
MutPanning:0.559
LAML–US
driverR:0.689
MutSigCV:0.528
DriverNet:0.505
OncodriveFML:0.547
MutPanning:0.524
LGG–US
driverR:0.713
MutSigCV:0.55
DriverNet:0.609
OncodriveFML:0.542
MutPanning:0.549
LIHC–US
driverR:0.681
MutSigCV:0.574
DriverNet:0.62
OncodriveFML:0.549
MutPanning:0.578
LUAD–US
driverR:0.688
MutSigCV:0.597
DriverNet:0.698
OncodriveFML:0.55
MutPanning:0.616
LUSC–US
driverR:0.694
MutSigCV:0.593
DriverNet:0.598
OncodriveFML:0.54
MutPanning:0.621
OV–US
driverR:0.685
MutSigCV:0.581
DriverNet:0.54
OncodriveFML:0.5
MutPanning:0.586
PAAD–US
driverR:0.746
MutSigCV:0.533
DriverNet:0.557
OncodriveFML:0.513
MutPanning:0.531
PRAD–US
driverR:0.67
MutSigCV:0.559
DriverNet:0.64
OncodriveFML:0.55
MutPanning:0.556
READ–US
driverR:0.652
MutSigCV:0.555
DriverNet:0.697
OncodriveFML:0.547
MutPanning:0.553
SKCM–US
driverR:0.714
MutSigCV:0.59
DriverNet:0.522
OncodriveFML:0.524
MutPanning:0.639
STAD–US
driverR:0.682
MutSigCV:0.592
DriverNet:0.62
OncodriveFML:0.552
MutPanning:0.635
THCA–US
driverR:0.656
MutSigCV:0.539
DriverNet:0.656
OncodriveFML:0.532
MutPanning:0.539
UCEC–US
driverR:0.689
MutSigCV:0.556
DriverNet:0.543
OncodriveFML:0.555
MutPanning:0.621
BRCA_METABRIC
driverR:0.801
MutSigCV:0.54
DriverNet:0.57
OncodriveFML:0.732
MutPanning:0.553
COAD_CPTAC
driverR:0.696
MutSigCV:0.58
DriverNet:0.655
OncodriveFML:0.523
MutPanning:0.625
LUAD_ONCOSG
driverR:0.667
MutSigCV:0.604
DriverNet:0.552
OncodriveFML:0.557
MutPanning:0.608
OV–AU
driverR:0.677
MutSigCV:0.576
DriverNet:0.501
OncodriveFML:0.522
MutPanning:0.578
PACA–AU
driverR:0.673
MutSigCV:0.589
DriverNet:0.516
OncodriveFML:0.546
MutPanning:0.593
PACA–CA
driverR:0.651
MutSigCV:0.594
DriverNet:0.522
OncodriveFML:0.521
MutPanning:0.612
PRAD–CA
driverR:0.648
MutSigCV:0.59
DriverNet:0.672
OncodriveFML:0.546
MutPanning:0.59
Sensitivity
Specificity
0.0
0.4
0.8
1.0
0.6
0.2


All Test Datasets – AUC Distributions
AUC
BLCA–US
0.2
0.0027
BRCA–US
0.00039
0.45
CESC–US
0.66
0.25
COAD–US
0.0012
0.38
HNSC–US
0.00014
0.52
KIRC–US
3.3 × 10⁻⁵
0.036
KIRP–US
0.34
0.86
LIHC–US
0.45
0.65
LUAD–US
0.016
0.011
LUSC–US
0.12
0.67
PAAD–US
5.5 × 10⁻⁶
0.071
PRAD–US
0.0064
0.015
READ–US
0.19
0.025
STAD–US
0.22
0.11
THCA–US
0.00013
1.2 × 10⁻¹¹
UCEC–US
0.053
p < 2.22 × 10⁻¹⁶
1.2
0.9
0.6
0.3
0.0
driveR
DawnRank
PRODIGY