

METHODOLOGY ARTICLE

Open Access



# driveR: a novel method for prioritizing cancer driver genes using somatic genomics data

Ege Ülgen\*  and O. Uğur Sezerman

\*Correspondence:  
egeulgen@gmail.com  
Department of Biostatistics  
and Medical Informatics,  
School of Medicine,  
Acibadem Mehmet Ali  
Aydinlar University, Istanbul,  
Turkey

## Abstract

**Background:** Cancer develops due to “driver” alterations. Numerous approaches exist for predicting cancer drivers from cohort-scale genomics data. However, methods for personalized analysis of driver genes are underdeveloped. In this study, we developed a novel personalized/batch analysis approach for driver gene prioritization utilizing somatic genomics data, called driveR.

**Results:** Combining genomics 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 driveR performs adequately, achieving a median AUC of 0.684 (range 0.651–0.861) on the 28 batch analysis test datasets, and a median AUC of 0.773 (range 0–1) on the 5157 personalized analysis test samples. Moreover, it outperforms existing approaches, achieving a significantly higher median AUC than all of MutSigCV (Wilcoxon rank-sum test  $p < 0.001$ ), DriverNet ( $p < 0.001$ ), OncodriveFML ( $p < 0.001$ ) and MutPanning ( $p < 0.001$ ) on batch analysis test datasets, and a significantly higher median AUC than DawnRank ( $p < 0.001$ ) and PRODIGY ( $p < 0.001$ ) on personalized analysis datasets.

**Conclusions:** This study demonstrates that the proposed method is an accurate and easy-to-utilize approach for prioritizing driver genes in cancer genomes in personalized or batch analyses. driveR is available on CRAN: <https://cran.r-project.org/package=driveR>.

**Keywords:** Driver gene, Prioritization, Cancer, Somatic

## Background

Cancer develops due to changes that have occurred in the DNA sequence of the genomes of cancer cells, somatic mutations acquired during an individual’s lifetime [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]. Identifying molecular cancer driver genes is critical for personalized oncology as accurate identification of personalized driver genes will result



in precise diagnosis and 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], OncodriveFML [17] and LOTUS [18]. MuSiC uses the significance of a 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., a 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 genomics 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. LOTUS is a machine-learning-based method for pan-cancer and cancer-specific driver gene prediction. Combining mutation frequency, functional impact, and pathway-based information, LOTUS provides a single- and multi-task learning algorithm to predict driver genes.

The results of some of the batch analysis approaches mentioned above are available in DriverDBv3 [19]. A recent integrated database and analysis platform, OncoVar, which employed some of the methods discussed above and incorporated known driver events to identify driver mutations, driver genes, and pathogenic pathways with high confidence, enables researchers to assess relationships and cancer drivers [20]. The Integrative OncoGenomics (IntOGen) platform also summarizes somatic mutations, genes, and pathways involved in tumorigenesis [21]. The 2020-02-01 release has identified cancer drivers analyzing 221 tumor cohorts by combining results from seven different methods for cancer driver gene identification.

Although several approaches exist for identifying cancer driver genes in cohort-scale genomics data, and several platforms contain curated drivers for different cancer cohorts, personalized driver gene identification approaches are still underdeveloped. Personalized driver gene prioritization is essential for numerous reasons: (1) it is necessary to identify actual driver genes for the patient as some patients have alterations in many known driver genes, (2) there may be a need to identify putative driver genes in patients without any alteration in any known driver gene, (3) as the number of therapies that can be administered at the same time is limited due to toxicity and adverse events

[22, 23], there is a need to prioritize driver genes for the patient. Some methods that operate on a single patient's data to identify and rank patient-specific driver genes have been developed. DawnRank uses a PageRank algorithm to rank mutated genes according to the effect on expression deregulation of downstream differentially-expressed genes (DEGs) in a directed PIN [24]. The single-sample controller strategy (SCS) aims to identify a set of mutated genes linked to downstream DEGs in a directed PIN [25]. 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 genomics information and prior biological knowledge [26]. 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 [27].

As stated above, most approaches for driver gene prioritization are batch-analysis-based. Furthermore, most methods 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 and incorporates prior biological knowledge, which we call driveR. We developed driveR intending to establish an accurate and reliable method for driver gene prioritization. The approach allows for personalized or batch analysis of genomics data for driver gene prioritization by combining genomics 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 [28] gene scores, and memberships to cancer-related Kyoto Encyclopedia of Genes and Genomes (KEGG) [29] pathways. It uses these features to estimate cancer-type-specific probabilities of being a driver for each gene using 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 with the hopes of facilitating precise diagnosis, personalized therapy, and overall resulting in better clinical decision-making.

## Results

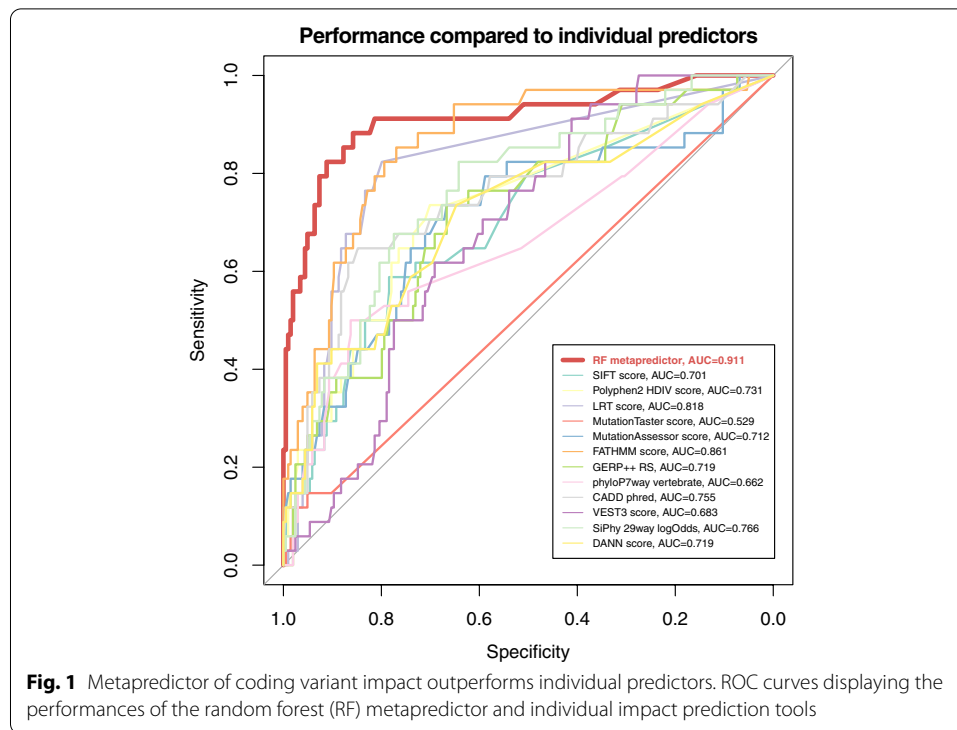
### The variant impact metapredictor outperforms individual predictors

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 (Additional file 2: Figure S2). As expected, this model also performed better than the individual impact prediction tools that were used as features to train the metapredictor, achieving an AUC of 0.911 (Fig. 1).

This metapredictor can be used to prioritize coding variants according to their pathogenicity and was used for generating one of the features of the MTL classification model.

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

Using genomics data and prior biological information, we next trained an MTL classification model for obtaining cancer-type-specific driver gene predictions. The Biclustered coefficient matrix for the MTL model classification model is presented in Fig. 2. Based



on biclustering, three major categories of features were identified: one cluster containing only Phenolyzer score, a cluster of positively weighted features, and another cluster of negatively weighted features. Overall, the Phenolyzer score had relatively higher coefficient estimates, highlighting the importance of incorporating prior biological knowledge into the MTL model.

The MTL classification model's batch analysis performance was assessed on the 28 test datasets (Fig. 3). The median AUC on the test datasets was 0.684 (range 0.651–0.861, Fig. 3A). Using different thresholds, the median number of predicted driver genes per test dataset ranged from 13 to 42.5 and, as expected, decreased with the increasing threshold (Fig. 3B). The median number of predicted driver genes per test dataset was 35 (range 2–104) using cancer-type-specific thresholds. As expected, the percentages of “True Driver Genes” (as curated by CGC) and “Actionable Genes” (as curated by TARGET [30], containing a total of 135 actionable genes) among all predicted driver genes across all datasets increased with increasing threshold values (Fig. 3C, range of median percentages of “True Driver Genes” = 63.84–94.59%, range of median percentages of “Actionable Genes” = 55.29–85.71%). Using the cancer-type-specific thresholds (that maximized the accuracy on the validation datasets), a median of 71.29% predicted driver genes were found to be “True Driver Genes” (Fig. 3C left-hand panel), and 59.46% predicted driver genes were found to be “Actionable Genes” (Fig. 3C right-hand panel).

Next, the personalized analysis performance of the driveR approach was assessed on the 5157 test patients (Fig. 4). The median AUC value among all test patients was 0.773 (Fig. 4A, top, range 0–1). The median AUC values of patients per test dataset were similar (Fig. 4A bottom, range of median AUC per dataset = 0.655–1). For each different threshold, the median number of predicted driver genes per patient was



1 (Fig. 4B). The median number of predicted driver genes per patient was 1 (range 0–26) using cancer-type-specific thresholds. The median percentages of “True Driver Genes” and “Actionable Genes” among all predicted driver genes in each patient per all different thresholds were 100% (Fig. 4C). Using the cancer-type specific thresholds, the median percentages were again 100%. This implies that each gene (if any) predicted by driveR to be a driver in an individual is most likely a true driver gene and an actionable gene.

(See figure on next page.)

**Fig. 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 numbers of predicted driver genes using the specified thresholds across all test datasets. “cancer-specific” indicates that the cancer-type-specific thresholds were used to predict driver genes. **c** Boxplots of percentages of “True Driver Genes” (left) and “Actionable Genes” (right) among all predicted driver genes by using the specified thresholds across all test datasets. “cancer-specific” indicates that the cancer-type-specific thresholds were used to predict driver genes

### 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 Additional file 3: Figure S3. The median AUC of driveR across all test datasets (median AUC = 0.684) was significantly higher than all of MutSigCV (0.579, Wilcoxon rank-sum test  $p < 0.001$ ), DriverNet (0.614,  $p < 0.001$ ), OncodriveFML (0.546,  $p < 0.001$ ) and MutPanning (0.591,  $p < 0.001$ ) (Fig. 5).

Next, performances of driveR and other personalized analysis tools on test patients from 16 datasets, for which analyses with all approaches could be performed, were compared (Fig. 6). It was observed that driveR had higher median AUC (0.728) compared to DawnRank (median AUC = 0.693,  $p < 0.001$ ) 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 (Additional file 4: Figure S4).

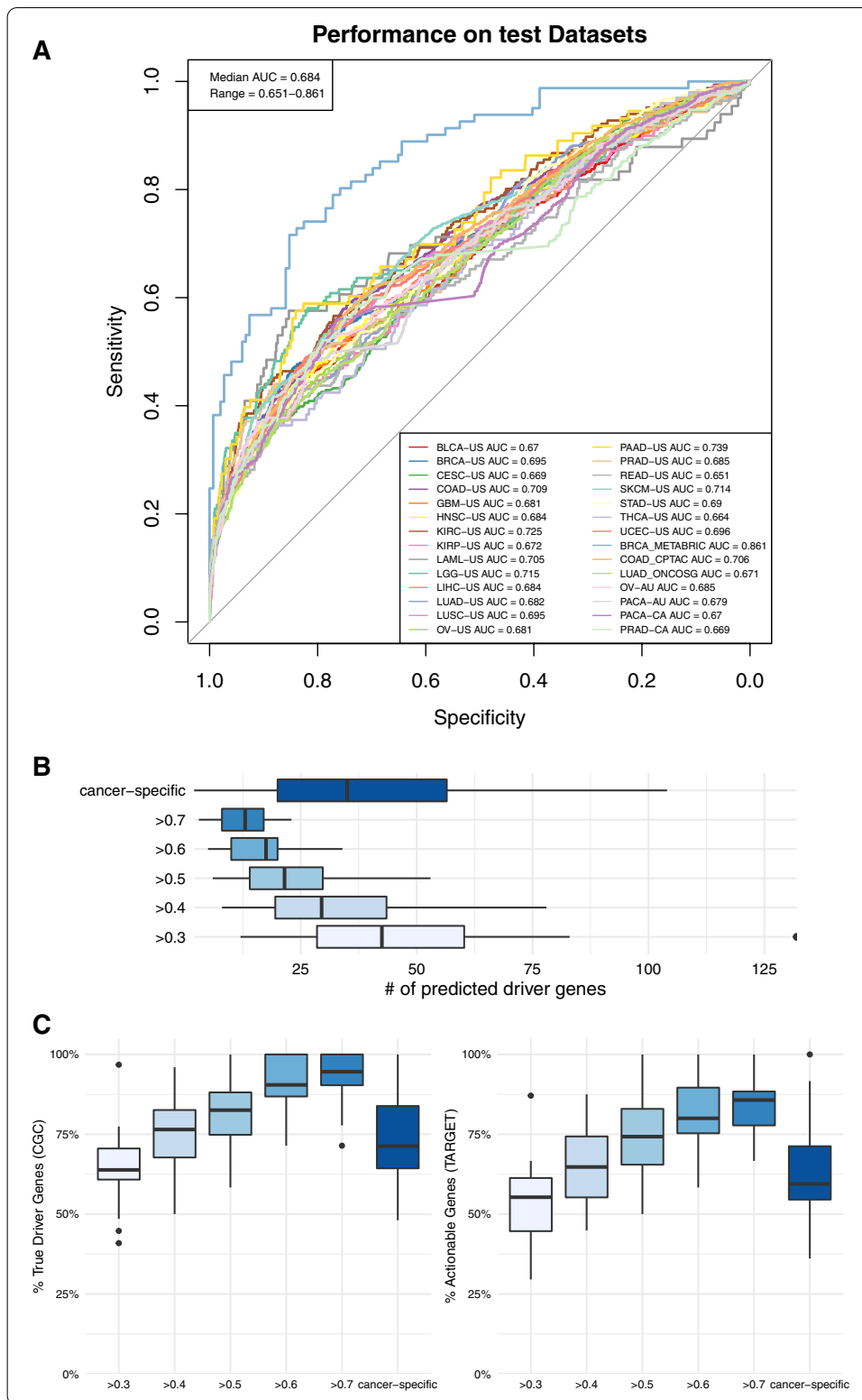
### Discussion

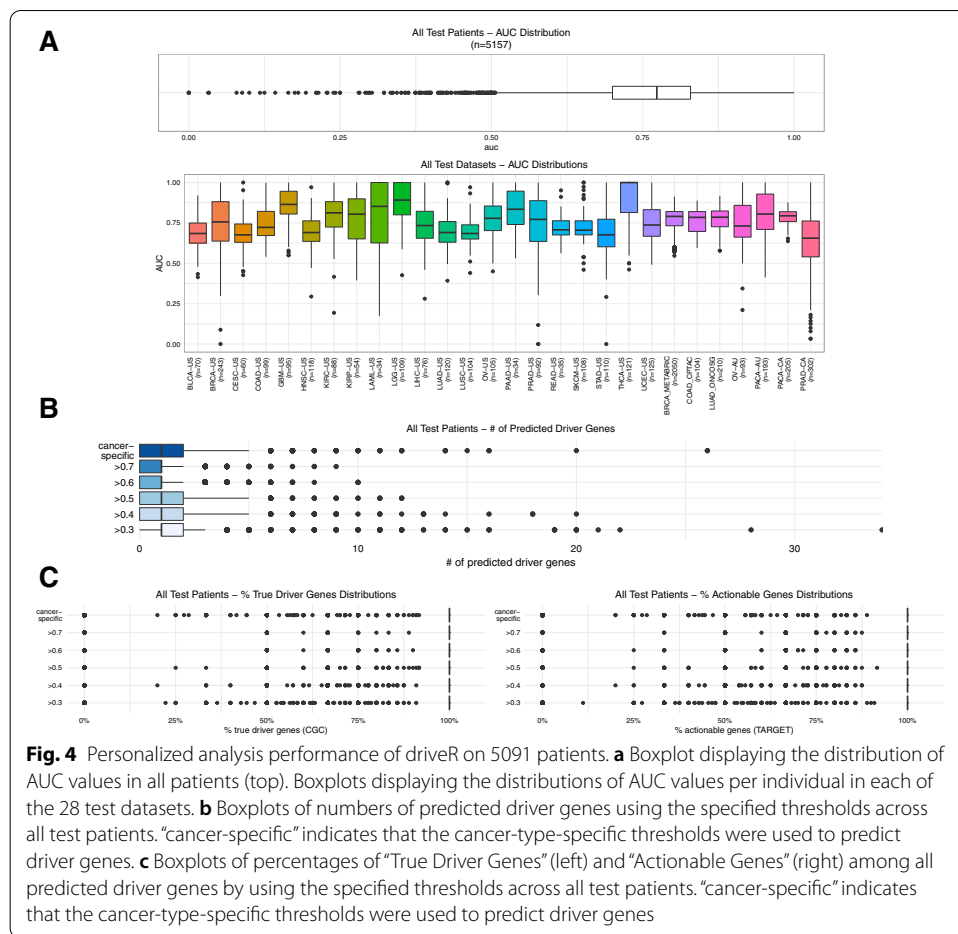
Alterations in driver genes are the putative underlying causes of oncogenesis and tumor formation [3, 4]. Although numerous driver genes are experimentally validated [21, 31–33], there is great potential clinical benefit in identifying individual patients’ driver genes [5, 6].

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 [21, 31–33]. Different cancers may possess different driver genes. The driveR approach attempts to define cancer-type-specific driver genes based on somatic genomics 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 particular cancer type.

Most approaches for driver gene prioritization are designed for analyzing 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 organ-of-origin. In this study, we demonstrate that driveR can be utilized to





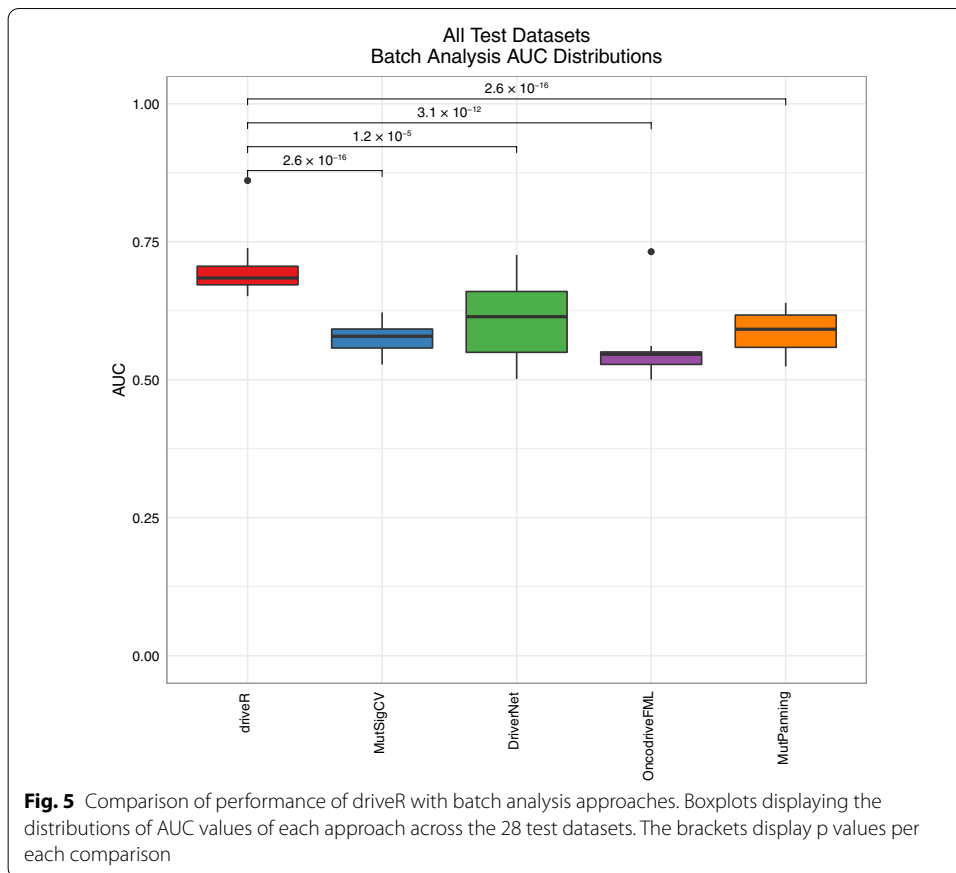
**Fig. 4** Personalized analysis performance of driveR on 5091 patients. **a** Boxplot displaying the distribution of AUC values in all patients (top). Boxplots displaying the distributions of AUC values per individual in each of the 28 test datasets. **b** Boxplots of numbers of predicted driver genes using the specified thresholds across all test patients. “cancer-specific” indicates that the cancer-type-specific thresholds were used to predict driver genes. **c** Boxplots of percentages of “True Driver Genes” (left) and “Actionable Genes” (right) among 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

rapidly and accurately analyze both patient-specific and cohort-scale genomics 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 genomics 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 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 genomics data is less complicated compared to transcriptomic data.

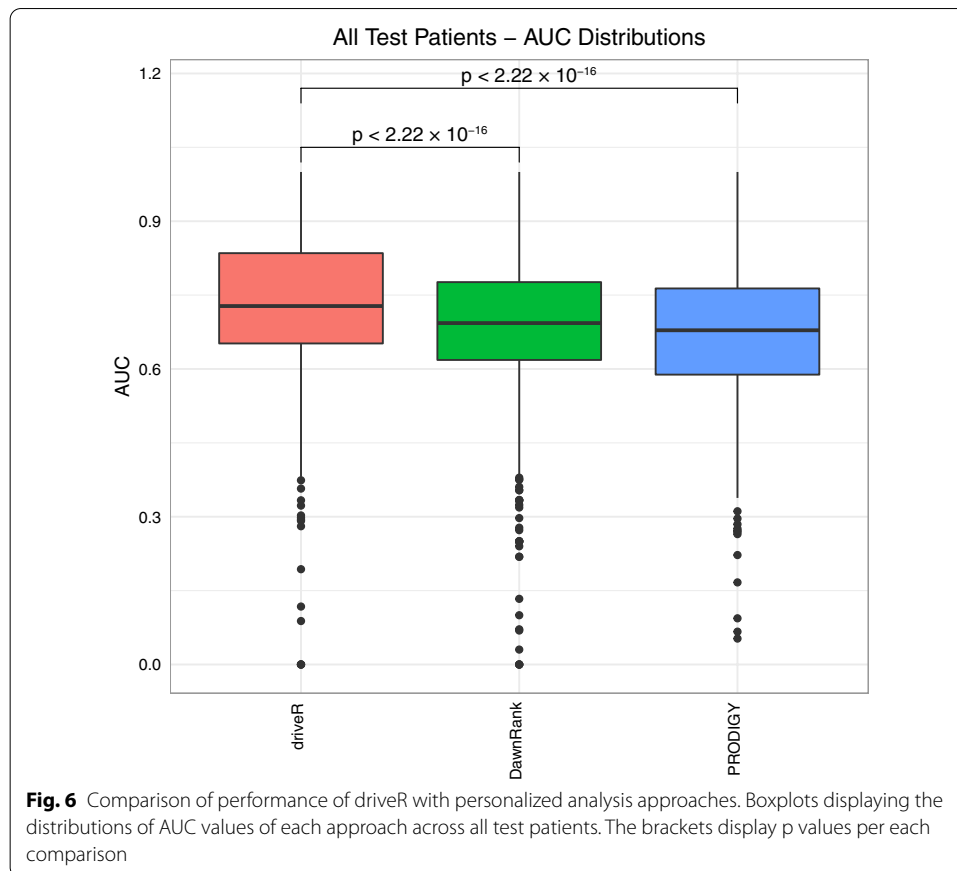




We demonstrated that the overall performance of driveR in identifying true drivers is adequate on both batch analyses and personalized analyses. Additionally, a high proportion of driveR predictions were also clinically-actionable genes. We also demonstrate that driveR outperforms other tools in both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

### Conclusions

In this study, we devised a novel approach, driveR, for prioritizing cancer-type-specific driver genes using somatic genomics data. As demonstrated, driveR can be utilized for both analyzing individual cancers and 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.



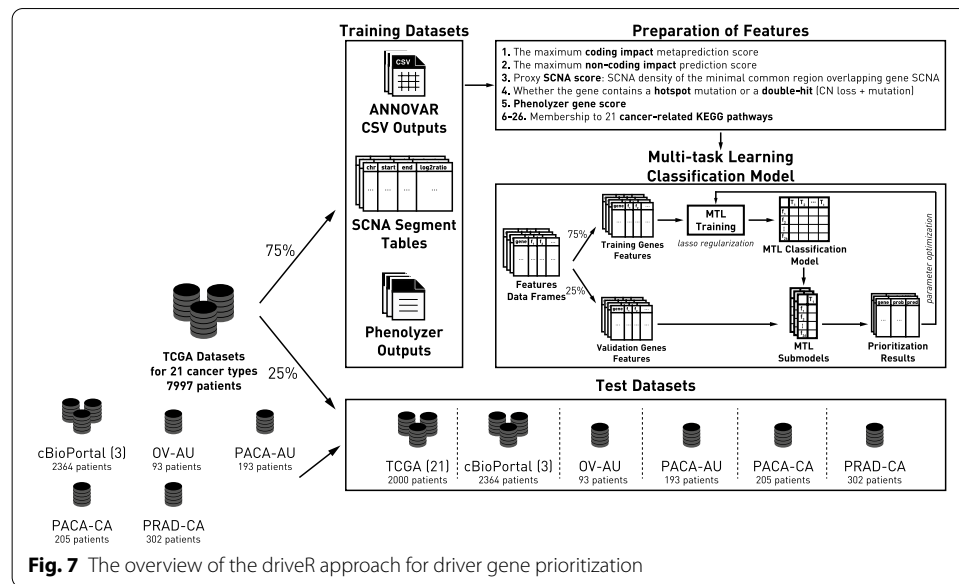
## Methods

### Coding variant impact metaprediction

Initially, we fitted coding impact metapredictor models to assign an estimated probability of damaging each somatic variant's impact. We later used this model to generate a feature for each gene in the multi-task learning classification model. We ensured that the genomic coordinates used in this study are hg19.

For training the coding impact metapredictor models, a benchmarking dataset from the Martelotto et al. study [34] was obtained. All mutations within the dataset were annotated using ANNOVAR [35] with 12 impact predictors' scores using dbNSFP v3.0 [36]: SIFT [37], PolyPhen-2 [38] (HumDiv scores), LRT [39], MutationTaster [40], Mutation Assessor [41], FATHMM [42], GERP++ [43], PhyloP [44], CADD [45], VEST [46], SiPhy [47] and DANN [48]. Variants with any missing predictor scores were excluded. Additionally, we excluded variants with the label "uncertain"; 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 evaluate the predictors' predictive strength (Additional file 1: 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 (Additional file 1: 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, six 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 six classification models was trained using the training dataset using tenfold 3-times-repeated cross-validation, maximizing the area under curve (AUC) metric, and tested on the test dataset. Additional file 2: Figure S2 shows the receiver operating characteristic (ROC) curves and AUC values for each of the six 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.

### Driver gene prioritization

The overview diagram of our approach to prioritize driver genes is presented in Fig. 7. We again ensured that the genomic coordinates used in this study are hg19.

We trained a multi-task learning (MTL) classification model to prioritize cancer driver genes in a cancer-type-specific context. 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 [49] was obtained. This data consisted of genomics data for 21 different cancer types (Additional file 5: Table S1). Genomics data from each of these different cancer types were randomly split into training (75% of patients) and test (25% of patients) data. Additional data for testing were obtained for four datasets of four different cancer

types from cBio Cancer Genomics Portal [50] (namely, BRCA-METABRIC [51], COAD-CPTAC [52] and LUAD-ONCOSG [53]) and four datasets of three different cancer types from ICGC (namely OV-AU, PACA-AU, PACA-CA, and PRAD-CA).

“True positive” driver genes were defined as the 723 experimentally-validated driver genes curated by the Cancer Gene Census (CGC, v92) [54]. 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 among all segments overlapping it. 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 the minimal region of copy number amplifications or deletions representing a common genomic alteration across the examined cancers [55]. We obtained pan-cancer MCR data from Kim et al. [56] 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 MCR's SCNA density (SCNA / Mb). 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) [57] v92 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 binary feature in the driver gene classification model.

#### **Phenotype based gene analyzer score**

Phenotype Based Gene Analyzer (Phenolyzer) [28] 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-type-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 [29] 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. Memberships to these cancer-related pathways were used as binary features in the driver gene classification model.

For training the MTL model, the R package RMTL [58] was used. We trained an MTL model with sparse structure (lasso regularization). The framework for the algorithm used by RMTL is provided in Eq. 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  $i$ th 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.  $\|\cdot\|_1$  is the  $L_1$  norm and  $\|\cdot\|_F$  is the Frobenius norm.  $\lambda_1$  aims to control the effect of cross-task regularization and  $\lambda_2$  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  $\lambda_2$  value was obtained by assessing the performance on the validation dataset (determined to be  $10^{-4}$ ). After determining the optimal  $\lambda_1$  value using tenfold cross-validation (determined to be  $10^{-5}$ ), 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, the sub-models’ personalized analysis performances were assessed by calculating AUC per each patient in each test dataset.

#### 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 on the test datasets.

MutSigCV (version 1.3.5) analyses were performed using default settings on the GenePattern platform [59]. DriverNet analyses were performed using DriverNet version 1.28.0 with the BioGRID [60, 61] 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.

### Comparison with other personalized analysis approaches

The personalized analysis performance of driveR was also compared with 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 one 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 [62] Homo sapiens PIN (version 11.0) and with KEGG curated pathways using the default settings. PRODIGY gene scores per pathway were aggregated into a single gene score for comparison purposes.

### 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; IntOGen: Integrative OncoGenomics.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04203-7>.

**Additional file 1: Figure S1.** The 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.

**Additional file 2: Figure S2.** Performance of different coding variant impact metapredictor models in the training (A) and test (B) datasets.

**Additional file 3: 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.

**Additional file 4: 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.

**Additional file 5: Table S1.** Datasets used for training and testing the multi-task learning classification model. "Type" indicates whether the dataset was used for training or testing. "Dataset ID" indicates the name of the dataset. "Cancer type" indicates the cancer type of the dataset. "Source" indicates the source from which the dataset was obtained. "Original Source" indicates the original source of the data.

### Acknowledgements

The authors would like to thank Assoc. Prof. Burcu Güngör, Ph.D., Prof. Koray Özdoğan, M.D. and Ozan Özişik, Ph.D. for their critical review and useful suggestions.

**Authors' contributions**

Conceptualization, EU and OUS; implementation, analyses, preparing the manuscript draft, EU; study supervision, OUS; review and editing, both authors. Both authors read and approved the final manuscript.

**Funding**

E.U. receives scholarship from TUBITAK under the Grant 118C039. The TUBITAK Grant 118C039 is a project on personalized oncology and provides E.U. a monthly scholarship.

**Availability of data and materials**

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

The approach presented in this study is available as an R package:uu

**Project name:** driveR.

**Project home page:** <https://github.com/egeulgen/driveR>

**Operating system(s):** Platform independent.

**Programming language:** R

**License:** MIT license.

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 28 December 2020 Accepted: 17 May 2021

Published online: 24 May 2021

**References**

- Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719–24.
- Greenman C, Stephens P, Smith R, Dalgleish 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):153–8.
- Haber DA, Settleman J. Cancer: drivers and passengers. *Nature*. 2007;446(7132):145–6.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science*. 2013;339(6127):1546–58.
- Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med*. 2011;17(3):297–303.
- Schilsky RL. Personalized medicine in oncology: the future is now. *Nat Rev Drug Discovery*. 2010;9(5):363–6.
- 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):14330–5.
- 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):642–56.
- 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):1589–98.
- 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):214–8.
- 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. *Nat Genet*. 2020;52(2):208–18.
- Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*. 2012;22(2):398–406.
- Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics*. 2018;34(17):i972–80.
- 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):2757–64.
- 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 Biol*. 2012;13(12):R124.
- Kartha VK, Sebastiani P, Kern JG, Zhang L, Varelas X, Monti S. CaDrA: a computational framework for performing candidate driver analyses using genomic features. *Front Genet*. 2019. <https://doi.org/10.3389/fgene.2019.00121>.
- 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 Biol*. 2016;17(1):128.
- Collier O, Stoven V, Vert JP. LOTUS: A single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput Biol*. 2019;15(9):e1007381.

19. Liu S-H, Shen P-C, Chen C-Y, Hsu A-N, Cho Y-C, Lai Y-L, Chen F-H, Li C-Y, Wang S-C, Chen M, et al. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* 2019;48(D1):D863–70.
20. Wang T, Ruan S, Zhao X, Shi X, Teng H, Zhong J, You M, Xia K, Sun Z, Mao F. OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res.* 2020;49:D1289–301.
21. 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. *Nat Rev Cancer.* 2020;20(10):555–72.
22. 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.
23. Park SR, Davis M, Doroshow JH, Kummar S. Safety and feasibility of targeted agent combinations in solid tumours. *Nat Rev Clin Oncol.* 2013;10(3):154–68.
24. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. *Genome Med.* 2014;6(7):56.
25. 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):1893–903.
26. Dong C, Guo Y, Yang H, He Z, Liu X, Wang K. iCAGES: integrated CAnceR GEName Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Medicine.* 2016;8(1):135.
27. Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. *Bioinformatics.* 2019;36(6):1831–9.
28. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods.* 2015;12(9):841–3.
29. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
30. 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):682–8.
31. 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.
32. 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):1081–2.
33. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandath C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep.* 2013;3(1):2650.
34. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Pisuoglio 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.
35. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164–e164.
36. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat.* 2016;37(3):235–41.
37. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003;31(13):3812–4.
38. 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):248–9.
39. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res.* 2009;19(9):1553–61.
40. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods.* 2014;11(4):361–2.
41. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39(17):e118–e118.
42. 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.
43. 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.
44. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 2010;20(1):110–21.
45. Rentsch 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.
46. 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):S3.
47. 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):i54–62.
48. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics.* 2015;31(5):761–3.
49. 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.
50. 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):401–4.
51. 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.



52. 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):1035–49.
53. 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):177–86.
54. 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.
55. 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.
56. 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):217–27.
57. 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 Res*. 2018;47(D1):D941–7.
58. Cao H, Zhou J, Schwarz E. RMTL: an R library for multi-task learning. *Bioinformatics*. 2018;35(10):1797–8.
59. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Genet*. 2006;38(5):500–1.
60. 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.
61. 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 Res*. 2018;47(D1):D529–41.
62. 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):D607–13.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

