**Title page**

## Title:

Integrating gene mutation spectra from tumors and the general population with gene expression topological networks to identify novel cancer driver genes

## Author affiliations

Dan He1,2#, Ling Li1,2#, Zhiya Lu1,2, Shaoying Li1,2, Tianjun Lan3, Feiyi Liu1,2, Huasong Zhang1,2, David N. Cooper4, Huiying Zhao1,2,\*

1Department of Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou 510006, China

2Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangzhou 510006, China

3Department of Oral and Maxillofacial Surgery, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou 510010, China

4Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

# Contributed equally

## Correspondence email

Huiying Zhao, PhD

Department of Medical Research Center, Sun Yat-Sen Memorial Hospital, 107 Yan Jiang West Road, Guangzhou, P.R. China, 500001

Zhaohy8@mail.sysu.edu.cn

**Abstract**

**Background:** Discovering cancer driver genes is critical for improving survival rates. Current methods often overlook the varying functional impacts of mutations. It is necessary to develop a method integrating mutation pathogenicity and gene expression data, enhancing the identification of novel cancer drivers.

**Methods:** To predict cancer drivers, we have developed a framework (DGAT-cancer) that integrates the pathogenicity of somatic mutation in tumors and germline variants in the healthy population, with topological networks of gene expression in tumors, and the gene expressions in tumor and paracancerous tissues. This integration overcomes the limitation of current methods that assume a uniform impact of all mutations by leveraging a comprehensive view of mutation function within its biological context. The framework employs topological networks to harmonize mutation and expression data, allowing for a more discerning evaluation of each gene's role.These features were filtered by an unsupervised approach, Laplacian selection, and combined by Hotelling and Box-Cox transformations to score genes. By using gene scores as weights, Gibbs sampling was performed to identify cancer drivers.

**Results:** DGAT-cancer was applied to 6,643 samples derived from seven types of cancer cohorts, and achieved the best area under the precision-recall curve compared to five commonly used methods. Specifically, 505 genes were identified across seven cancer types, with 398 being unique to a single cancer type and 38 being common to three or more cancer types. Knockdown of the top novel gene, *EEF1A1* indicated a ~ 41-50% decrease in glioma size and improved the temozolomide sensitivity of glioma cells.

**Conclusion:** By combining heterogeneous genomics and transcriptomics data, DGAT-cancer has significantly improved our ability to detect novel cancer drivers, and is an innovative approach revealing novel cancer therapeutic targets, thereby advancing the development of more precise and effective cancer treatments.

**Keywords**: Cancer drivers, Pathogenic status of mutations, Laplacian selection, Hotelling and Box-Cox transformations, Gibbs sampling.

# Background

The identification of cancer driver genes is important for the early diagnosis of cancer, for identifying efficacious anti-cancer therapeutics, and for investigating the underlying mechanisms of tumorigenesis. Traditionally, cancer driver genes have been recognized on the basis of their being recurrently altered in tumors[1](#_ENREF_1),[2](#_ENREF_2). However, the ability of many somatic mutations to alter gene function is often uncertain, making it hard to identify cancer driver genes unambiguously. Clearly, we require information other than somatic mutation data in order to reliably detect cancer driver genes.

Cancer driver genes could in principle be identified by integrating gene mutation data with gene expression data. As an example, a previous study performed enrichment analysis to integrate genomic and transcriptomic alterations from whole-exomes and functional data from protein function predictions with gene interaction networks to reveal breast cancer driver genes[3](#_ENREF_3). Recent approaches have been developed by combining mutation scores with biological network protein-protein interaction (PPI) data to predict cancer driver genes[4](#_ENREF_4),[5](#_ENREF_5). These approaches did not consider functional data for mutations and cancer tissue-specific PPI networks, which may reduce the ability to predicting novel cancer drivers.

To provide valuable insights, methods like MutSigCV[6](#_ENREF_6" \o "Lawrence, 2013 #19) have been developed to identify cancer genes by assessing mutation heterogeneity, highlighting those that deviate from expected random mutational patterns. OncodriveFML[7](#_ENREF_7" \o "Mularoni, 2016 #18) and OncodriveCLUSTL[8](#_ENREF_8) further contribute by detecting drivers based on pan-cancer mutation patterns and employing sequence-based clustering approaches, respectively. While each method provided a unique genomic perspective on cancer, they collectively share limitations in fully accounting for the functional implications of mutations or adapting to the tissue-specificity of PPI networks, which are essential for the comprehensive prediction of novel cancer drivers.

To address these shortcomings, we devised a new model, DGAT-cancer (Distinguish cancer drivers using Genomics and Transcriptome data), which integrates the predicted pathogenicity scores of somatic mutations in cancers and germline mutations in the healthy population, with gene expression in tumors and paracancerous tissues in order to detect cancer driver genes. DGAT-cancer utilized Topological Data Analysis (TDA) to identify tumor clusters with similar gene expression profiles, crucial for understanding cancer heterogeneity[9](#_ENREF_9). TDA has been instrumental in uncovering novel cancer driver genes through its ability to detect meaningful correlations between mutation frequency and gene expression alterations. DGAT-cancer analyzed 6,643 samples from seven cancer types, identifying 505 genes as potential drivers. Notably, 398 of these genes were unique to individual cancers, while 38 were prevalent across at least three cancer types, highlighting DGAT-cancer's ability to detect both cancer-specific and shared genetic determinants.

# Methods

## Data collection

***Somatic mutation data.*** We collected simple somatic mutation (SSM) data from 12 cancer cohorts from the Broad Institute GDAC Firehose Portal, the International Cancer Genome Consortium (ICGC) Data Portal and The Cancer Genome Atlas (TCGA) (Table S1). The coordinates of the data are by reference to the genome assembly version hg19.We collected germline mutation data of 2,557 individual samples from Phase 3 of the 1000 Genomes Project (GRCh38). Somatic mutation entries that were duplicated between multiple databases were removed such that only one non-redundant entry was retained.

***The pathogenicity of mutations.*** There are many methods available with which to predict the pathogenicity of mutations. We employed dbNSFP (version: dbnsfp30a, also called LJB\*)[10](#_ENREF_10) of ANNOVAR[11](#_ENREF_11) to annotate mutations from the 1000 Genomes Project (1000GP) and the somatic mutations (Table S1). The dbNSFP includes 19 predictors which provide scores representing the probability of the non-synonymous variants being pathogenic (Table S2). These predictors take into account both loss-of-function (LoF) and gain-of-function (GoF) mutations in the context of tumor biology[12](#_ENREF_12),[13](#_ENREF_13). The scores given by each predictor were normalized into the range of 0 to1 in this study.

***Known cancer driver genes.*** The accuracy of the predictions made by DGAT-cancer was examined using a set of known cancer driver genes. This set of 1,168 unique cancer driver genes comprised a non-redundant set of 723 genes (including 576 class Tier 1 genes and 147 class Tier 2 genes) from the COSMIC Cancer Gene Census[14](#_ENREF_14" \o "Sondka, 2018 #70) (CGC) and 1,064 genes from OncoKB[15](#_ENREF_15" \o "Chakravarty, 2017 #60). OncoKB contains more cancer driver genes than CGC because it collects cancer genes from various panels, including class Tier 1 genes in CGC (576), the MSK-IMPACT™ panel (505), the MSK-IMPACT™ Heme and HemePACT panels (575), the FoundationOne CDx panel (324), the FoundationOne Heme panel (593), and data from Vogelstein et al.(125)[16](#_ENREF_16" \o "Vogelstein, 2013 #74). We also collated gene sets containing cancer type-specific driver genes, which were collected from IntOGen[17](#_ENREF_17" \o "Martinez-Jimenez, 2020 #9). The number of known driver genes for each cancer type are shown in Additional file: Table S3. These genes were used as gold standard cancer-associated genes to evaluate the accuracy of the predictions made in this study.

For each gene, DGAT-cancer calculated a unidirectional Earth Mover’s Difference score (uEMD) to evaluate the difference between the predicted pathogenic scores (obtained from 19 predictors in dbNSFP[10](#_ENREF_10" \o "Liu, 2016 #47)) of somatic mutational spectra in cancers and that of germline variants in healthy populations. The influence of mutation on gene expression has been evaluated by integrating gene expression data in tumors with somatic mutations through topological data analysis (TDA)[18](#_ENREF_18" \o "Rabadan, 2020 #20). The Jensen–Shannon divergence between the gene expression profile and mutation profile of each gene on the topological network was further computed and used as another feature of the gene. The most effective features were identified by Laplacian selection in an unsupervised way. The selected features were integrated by means of the Hotelling and Box-Cox transformations to score the genes. Finally, by using gene scores as weights, we performed Gibbs sampling to identify cancer drivers. One of the top novel genes was validated to be a cancer driver ofglioma. The work scheme of DGAT-cancer is shown in Fig. 1. To evaluate the performance of DGAT-cancer, we also collected genes lacking missense mutations, genes affecting proliferation and/or viability, drug response genes, and genes associated with prognosis of cancer patients. The detailed methods are in supplementary material.

## Pathogenicity score profiles of mutations in cancers and in the healthy population cohort

Each tumor sample or specific individual from the 1000GP may harbor more than one mutation in each gene. The pathogenic status of these mutations was scored by 19 distinct approaches each representing different detrimental effects of mutations on gene function. The mutation score of a gene for a specific sample determined by a particular predictive approach was defined as the average mutation score across all mutations in that gene. Mutation scores of a gene from all tumor samples or all samples from 1000GP (Table S1) were considered to represent the mutation score profiles of that gene calculated by one particular predictive approach. For each gene, we constructed a density distribution for mutation scores in each cancer cohort and in the healthy population cohort, respectively. The density distribution was divided into 100 evenly spaced bins. A gene was filtered out from the construction of mutation score profiles if all the mutations in the gene were derived from fewer than five samples.

In order to compare the mutation score profiles in tumors with that of the healthy population cohort, we calculated the difference between the two profiles by means of a unidirectional Earth Mover’s Difference score (uEMD, *Equation (1)*). For each gene, we repeated this calculation for all 19 types of functional score and obtained 19 uEMD scores. These uEMD scores were termed uEMD-Mut scores.

(1),

where is the fraction of normalized scores in the -th bin for gene in the density distribution of a cancer cohort, is the fraction of normalized scores in the -th bin for gene in the density distribution of the general population, and is the index of bins in the density distribution. Thus, genes with a significantly different distribution in cancer tissues from the general cohort would be given higher uEMD scores.

## Comparing gene expression profiles of tumors and paracancerous tissues

The RNA-seq data of eight types of cancers were collected by methods described in supplementary material. RNA-seq data from paracancerous tissues or tumors with a sample size fewer than five were not included in the study. The sample sizes of RNA-seq data for each type of cancer are shown in Table S4. For each gene, we calculated the median expression level across samples. Then, we computed the uEMD score for each gene to measure the difference in the density distribution of gene expression in tumor and paracancerous tissues. This uEMD score of the gene was termed uEMD-Ex.

## Topological network constructed from gene expression data and somatic mutations in tumors

By using gene expression data derived from tumor samples, we constructed a topologicalnetwork for each type of cancer (Table S4 and Table S5) using Mapper algorithm, an R package in TDAmapper (the detailed input arguments are listed in Table S6)[19](#_ENREF_19). Briefly, tumor samples with similar expression profiles are clustered into one node. Where two nodes have at least one tumor sample in common, they are connected by an edge. For each gene in the topological network, we evaluated the divergence of the mutation frequency of the gene in samples across similar gene-expression clusters (*Equation (2)*). The strategy used here is similar to that described in previous studies[18](#_ENREF_18),[20](#_ENREF_20). Prior to the calculation, we used a threshold of MAF > 0.001 to filter out mutations occurring at low frequencies in tumors.

(2),

where denotes the set of nodes in the topological network, is the adjacency matrix of the topological network, is the number of nodes in , and is the average frequency of non-synonymous mutations of gene for samples in the node . A lower represents a higher divergence of the mutation frequency between similar clusters.

To evaluate the similarity between the profiles of mutation frequency and mRNA expression, we computed the Jensen–Shannon divergence between the expression and mutation profiles of each gene based on the topological network (*Equation (3)*). and were termed MutExTDA scores.

(3),

where denotes the fraction of tumors with gene somatically mutated in node , denotes the average expression of gene in the tumors associated with node , and denotes the set of nodes in the topological network. Prior to calculation, they were normalized to meet . A lower denotes less difference between the two distributions.

## Imputation of missing data

We obtained 24 features to describe each gene, comprising 19 uEMD-Mut scores, one uEMD-Ex score, two MutExTDA scores and two features representing the median expression levels of the gene in tumor and paracancerous tissues, respectively (Table S2). After filtering genes with over half-missing features (Tables S7), we compared k-nearest neighbors (KNN), mean imputation, and the removal of features with missing values on the remaining genes, using AUPRC to evaluate predictive performance. The result indicated that the imputation method, KNN has provided the best performance in the prediction (Fig. S1a). In detail, by using KNN for imputing the data, the DGAT-cancer has achieved average AUPRC 0.761 (ranging from 0.653 to 0.872) in predicting driver genes for 7 types of cancers, which is higher than using the exclusion of all missing values method (average AUPRC 0.152 ranging from 0.379 to 0.849) and mean imputation method (average AUPRC 0.666 ranging from 0.506 to 0.817). More than 99% of genes for cancer types Pheochromocytoma and paraganglioma (PCPG), Testicular germ cell tumors (TGCT) and Thyroid carcinoma (THCA) were lacking uEMD-Mut scores, leading to their exclusion in subsequent analyses.

## Using Laplacian Score to select features

We used an unsupervised method, the Laplacian Score, to select features that have high power to preserve the local geometric structure of the feature space[21](#_ENREF_21" \o "He, 2005 #2). The detailed steps for the application of the Laplacian Score in feature selection were as follows:

1. A network was constructed to connect all candidate genes (). For each pair of genes, we calculated the Euclidean distance, between their feature vectors ( is the feature vector of gene , and is the feature vector of gene ). Based on the Euclidean distance, if gene is among the top (here we set ) of the nearest genes to gene , or the gene is among the top of the nearest gene to the gene (), we set the connection between *i* and *j* as . Otherwise, .
2. In the network , if , we weigh the edge by (). Otherwise, . The weighted network reflects the local structure of the genes in the feature space.
3. Computing the Laplacian Score of each feature. Let denote the -th feature values for all genes. We redefine by removing the mean from the samples as in *Equation (4)*. The Laplacian Score () is computed by *Equation (5)*. The lower the value, the more important the feature is.

(4)

where , .

(5)

where .

To avoid overfitting, we did not directly control the number of features in the supervised model. Instead, we selected high-scoring features ranked by the Laplacian Score for inclusion. To assess the impact of feature selection, we compared the AUROC values obtained using 10, 15, 20, and all features across different cancer types. As shown in Fig. S1b, the model using 20 features achieved an average AUROC of 0.673 (ranging from 0.578 to 0.788), which is higher than the average AUROC obtained using 10 features (0.666, ranging from 0.58 to 0.807), 15 features (0.653, ranging from 0.582 to 0.733), and 24 features (0.625, ranging from 0.503 to 0.789). Besides AUROC, model performance was also evaluated using MCC. The model using 20 features had an average MCC of 0.437 (ranging from 0.311 to 0.517), outperforming the models with 10 features (average MCC = 0.432, ranging from 0.309 to 0.525), 15 features (average MCC = 0.418, ranging from 0.296 to 0.503), and 24 features (average MCC = 0.388, ranging from 0.266 to 0.512). Based on the Laplacian Score, we ultimately selected the top 20 features with the lowest scores for each cancer type in the prediction model.

## Data transformation

In order to integrate multiple features of each gene into a risk score (Fig. 1), we combined the features by Hotelling and Box-Cox transformations, which converted the feature values into -values. For a given scaled matrix of genes with features (, here ), the Hotelling transformation is performed as:

where with as the number of chosen principal components of and are eigenvectors for the covariance matrix of corresponding to decreasing eigenvalues with . . Thus, transformed . Then, the Box-Cox transformation is performed as follows,

,

where is the -th row vector of and all elements of vector are forced to be positive before being transformed, and is the parameter for transforming to .

Finally, we standardized each and calculated P-values for elements of that is in a standard Gaussian distribution. The P-values of the elements were combined as by Fisher's method (Fisher's combined probability test). is termed the score of gene .

## Gibbs sampling

In our previous research, we successfully applied Gibbs sampling for to predict genes associated with diseases[22](#_ENREF_22), and indicated that Gibbs sampling offers efficiency in sampling a set of scored genes showing maximus product. This approach ensures a robust convergence towards a stable posterior distribution, accurately assessing gene significance. In our research, Gibbs sampling provided a refined probabilistic approach to identify and prioritize cancer driver genes through the assessment of their posterior probabilitie.

The scores () of genes were used as conditional probabilities in Gibbs sampling to obtain a convergent probability distribution of candidate genes. In the first round of sampling, Gibbs sampling was initiated by randomly selecting () genes from the candidate genes. The genes were assumed to have equal probabilities of being selected. Then, in the second round of sampling, another set of genes was sampled from the remaining genes weighted by their scores . The following rounds were the same as the second round. In each round, the selected frequency () of each gene was updated. All the selected frequencies of candidate genes in the -th round were denoted as a vector (),. When the Euclidean norm of was smaller than ( was set as 0.01), the iteration was stopped. was assigned as the posterior probabilities (PP) of candidate genes. Then, we constructed a null distribution of PP in order to obtain the likelihood of a given gene being a cancer driver gene. The null distribution was generated by giving genes randomly weighted scores that were derived from the uniform distribution with the same range as the true scores. We generated 1,000,000 sets of null distributions of PPs for the genes by running Gibbs sampling. For each gene, we obtained 1,000,000 random PPs. By counting the number of times that a random PP of a gene was larger than the real PP of the gene, an experience -value was estimated. The -values were adjusted by Bonferroni correction and we selected those genes with as being significant.

## Comparison with other cancer driver prediction methods

In our comparative evaluation of DGAT-cancer and other methods for identifying cancer driver genes, we used the area under the precision-recall curve (AUPRC), calculated with the perfMeas package in R[23](#_ENREF_23). Our positive gene set comprised 1,199 known cancer drivers from CGC, OncoKB, and IntOGen (Methods). For the negative gene set, we excluded known drivers and those interacting with them according to the BIOGRID database (version 4.4.214, organism: human sapiens)[24](#_ENREF_24), yielding 8,153 unique human genes. To ensure a fair comparison, we focused on genes identified as potential drivers by at least one method. The numbers of positive and negative genes for each method are detailed in Table S9 and Table S11. The negative gene set in Table S9 was established by excluding known cancer drivers and their interacting genes from a pool of genes evaluated by DGAT-cancer, MutSigCV, OncodriveFML, OncodriveCLUSTL and DiffMut. In contrast, the negative gene set in Table S11 was derived by applying the same exclusion criteria to a broader pool of genes assessed by DGAT-cancer, DGAT-Mut, DGAT-Exp, DGAT-TDA, MutSigCV, OncodriveFML, OncodriveCLUSTL, DiffMut and TDAmut. To evaluate performance, five commonly used metrics in classification models were applied: specificity, F1-score and Matthews correlation coefficient (MCC). MCC is a balanced measure that considers true positives, true negatives, false positives, and false negatives, making it particularly useful for imbalanced datasets[25](#_ENREF_25). We selected the best MCC score and compared the corresponding specificity and F1-score to assess the relative performance of different methods.

DGAT-cancer was compared to five classical methods, MutSigCV[6](#_ENREF_6" \o "Lawrence, 2013 #19) (<https://software.broadinstitute.org/cancer/cga/mutsig>), OncodriveFML[7](#_ENREF_7" \o "Mularoni, 2016 #18) (http://bbglab.irbbarcelona.org/oncodrivefml/home), OncodriveCLUSTL[8](#_ENREF_8) (<http://bbglab.irbbarcelona.org/oncodriveclustl/home>), DiffMut[26](#_ENREF_26" \o "Przytycki, 2017 #1) (<https://github.com/Singh-Lab/Differential-Mutation-Analysis>) and TDAmut[9](#_ENREF_9" \o "Rabadán, 2020 #2) ([https://github.com/CamaraLab/TDA-TCGA/](https://github.com/CamaraLab/TDA-TCGA/" \t "_blank)). The descriptions of these methods are shown in supplementary material, Fig. S2.

## Tissue specimens and patient information

We collected 91 surgical specimens from the Department of Neurosurgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University. These samples include 55 glioblastoma multiforme (GBM) specimens, 31 brain lower grade glioma (LGG) specimens and 5 non-tumor brain tissues, which had been diagnosed between 2012 and 2022. The study was approved by the Ethics Committee of Sun Yat-sen University, and informed consent was obtained from all subjects. The non-tumor brain tissues were obtained from patients with non-tumor diseases and required partial brain excision from patients with traumatic brain injury, or other diseases such as cerebral angiomas or vascular malformations.

## Experimental validation of the role of novel driver genes in cancer

In order to examine the performance of DGAT-cancer, we experimentally validated the roles of the predicted cancer drivers by using surgical specimens, a cell model and a zebrafish model. All experimental methods are provided in supplementary material.

# Results

## Application of DGAT-cancer to multiple types of cancer

DGAT-cancer was applied to 6,643 samples derived from 7 cancer cohorts, which are Bladder urothelial carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Head and neck squamous cell carcinoma (HNSC), Lung adenocarcinoma (LUAD) and Stomach adenocarcinoma (STAD) whose mutation (uEMD-Mut) and gene expression (uEMD-Ex, gene expression level and MutExTDA) features were all available in TCGA database (Methods). The numbers of genes predicted to be cancer drivers by DGAT-cancer are shown in Table S8.

We found significant () overlaps between the predicted cancer drivers and the cancer gene sets, CGC, OncoKB and IntOGen, respectively (Fig. 2a) compared to the background genes 19,350 protein-coding genes form Ensembl version 110).

Since cancer genes are likely to have experienced a slower evolutionary rate and stronger purifying selection than those of non-cancer, Mendelian disease, and orphan disease genes[27](#_ENREF_27), we tested the enrichment of the predicted genes in the genes under selective constraint[28](#_ENREF_28). The predicted cancer drivers of the seven cancer types were significantly () enriched in genes that have been under selective constraint (Fig. 2a). We further evaluated the enrichment of the predicted cancer drivers in the gene sets identified through shRNA and CRISPR screens[29](#_ENREF_29),[30](#_ENREF_30) (Methods). As shown in Fig. 2a, the predicted cancer driver genes in the seven cancer types were significantly () enriched in the shRNA gene set, illustrating the potential roles of the predicted cancer drivers in cancer cell survival. The predicted cancer drivers were also enriched in cancer-related pathways defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[31](#_ENREF_31) database, e.g. Proteoglycans in cancer () and p53 signaling pathway () (Fig. S3). The method for gene-set enrichment analysis is in supplementary material.

Additionally, there were 20 genes predicted as cancer drivers by DGAT-cancer in multiple cancer types (Fig. 2b). These genes include *TP53* that was predicted to be a cancer driver in seven types of cancer, with predicted scores ranking between the top 1 to the top 24. The *COL1A2* gene was predicted as a cancer driver in BLCA and HNSC, with predicted scores ranking in the top 10 and the top 4, respectively. Another gene, *PTEN,* was predicted to be a cancer driver in BRCA, CESC, COAD, GBM, LGG and STAD.

Moreover, predicted cancer drivers in BRCA, COAD, GBM and STAD were significantly () enriched in drug response genes (Fig. 2c). We explored the correlation between the expression patterns of the predicted cancer driver genes and drug activities, expressed as 50% growth inhibitory levels (GI50) in the NCI-60 cell line, which were derived from CellMinerCDB[32](#_ENREF_32" \o "Luna, 2021 #69). The analysis was performed by calculating Pearson correlation coefficients between gene expression levels and the z-scores of negative log 10 (GI50) in all NCI-60 cell lines. We found that the expression levels of many predicted cancer drivers were significantly () correlated with drug activities (Fig. 2d), thereby illustrating the potential of these genes for clinical treatment.

## DGAT-cancer outperformed other methods

DGAT-cancer was compared to five other cancer driver identification methods, MutSigCV, OncodriveFML, OncodriveCLUSTL, DiffMut and TDAmut with respect to the area under the precision–recall curve (AUPRC). First, they are compared in predicting cancer drivers from a set of genes (Table S8) that have prediction scores given by at least one of the six methods. Table S9 shows the numbers of positive and negative genes used in evaluating the methods. As shown in Fig. S4a, the AUPRC (in the range of 0.336 to 0.469) of DGAT-cancer in predicting cancer drivers for seven types of cancer (BLCA, BRCA, COAD, HNSC, GBM, LUAD and STAD) were the highest by comparison with the other five methods, MutSigCV (AUPRC ranged in 0.225 to 0.282), OncodriveFML (AUPRC ranged in 0.283 to 0.365) OncodriveCLUSTL (AUPRC ranged in 0.289 to 0.356), DiffMut (AUPRC ranged in 0.212 to 0.248) and TDAmut (AUPRC ranged in 0.274 to 0.334). Then, we assessed the model performance using genes with prediction scores provided by all five methods (Table S8). Due to the unique gene filtering step involved in TDAmut, which results in predictions for only a limited set of genes, TDAmut was not included in the comparison (supplementary material). When assessing the AUPRC for predicting cancer drivers from genes with prediction scores provided by all five methods, the AUPRC (in the range of 0.646 to 0.862) of DGAT-cancer in predicting cancer drivers for seven types of cancer (BLCA, BRCA, COAD, HNSC, GBM, LUAD and STAD) were the highest by comparison with the other five methods, MutSigCV (AUPRC ranged in 0.322 to 0.565), OncodriveFML (AUPRC ranged in 0.417 to 0.629) OncodriveCLUSTL (AUPRC ranged in 0.357 to 0.562) and DiffMut (AUPRC ranged in 0.379 to 0.525) (Fig. 3a). Our method, DGAT-cancer, demonstrated superior performance in terms of MCC across all seven cancer types (Fig. 3b). This indicates that DGAT-cancer effectively balances the detection of true positives and true negatives, making it a robust method for predicting cancer drivers in various contexts. Although DGAT-cancer showed strong F1 score performance across most cancer types (Fig. 3b), yet its lower score in HNSC. Additionally, DGAT-cancer's performance in terms of Balanced Accuracy was consistently the highest across most cancer types, indicating its reliability in providing a balanced assessment of both true positive and true negative rates (Fig. 3b). In conclusion, DGAT-cancer analyzed 6,643 samples from seven cancer types, identifying 505 genes as potential drivers. Notably, 398 of these genes were unique to individual cancers, while 38 were prevalent across at least three cancer types, highlighting DGAT-cancer's ability to detect both cancer-specific and shared genetic determinants (Fig. 3c).

MutSigCV, OncodriveFML and OncodriveCLUSTL yielded a *p*-value to represent the probability of the predicted cancer driver being a false positive. The -value distributions of genes generated by these four methods were then compared with those expected *p-*values from a uniform distribution using quantile-quantile plots. As shown in Fig. S4b, the -values of genes not predicted to be cancer drivers by DGAT-cancer (*p* > 0.05) in BLCA, LUAD and STAD exhibited better agreement with the expected -values than those predicted by other methods. Additionally, the genes predicted (*p* < 0.05) by DGAT-cancer to be cancer drivers in all seven cancer types showed higher inflation from the expected -values than the genes predicted by other methods (Fig. S4c). This suggested that DGAT-cancer has enhanced potential to distinguish novel cancer drivers from random genes.

We next explored the consistency of DGAT-cancer with respect to the other methods. The posterior probabilities (PPs) given by DGAT-cancer for genes predicted to be cancer drivers () by the other methods were compared to the PP scores of the genes not predicted to be cancer drivers () by the other methods, MutSigCV, OncodriveFML, OncodriveCLUSTL, TDAmut and DiffMut, respectively. The differences were evaluated by the Wilcoxon rank-sum test. The total number of cancer drivers with identified by MutSigCV, OncodriveFML, OncodriveCLUSTL, TDAmut and DiffMut are shown in Table S10. As depicted in Fig. 4a, those genes predicted to be cancer drivers in BLCA, BRCA andCOAD, by the four methods were given significantly higher PPs () by DGAT-cancer than the genes not predicted to be cancer drivers by the other methods. These results suggested a high degree of consistency between DGAT-cancer and the other methods.

There were 67, 58,157, 41, 87, 38 and 108 genes predicted to be cancer drivers of seven types of cancer (BLCA, BRCA, COAD, GBM, HNSC,LUAD and STAD) by DGAT-cancer, respectively but not predicted to be cancer drivers by the other five methods (MutSigCV, OncodriveFML, OncodriveCLUSTL, TDAmut and DiffMut). Moreover, 108, 153,191, 22, 204, 153 and 160genes were predicted as cancer drivers in seven types of cancer (BLCA, BRCA, COAD, GBM, HNSC, LUAD and STAD) by at least of one of the five methods, were not predicted to be cancer drivers by DGAT-cancer. We compared these genes in relation to their scores of features used in DGAT-cancer, and found that the gene expression-related features such as uEMD-Ex scores, expression values and MutExTDA scores showed a significant difference in the two groups of genes across multiple cancers (Fig. 4band Fig. S5). Specifically, the genes missed by the other three methods were expressed more highly in both tumors and paracancerous tissues than the genes that were missed by DGAT-cancer in BRCA, COAD, HNSC and LUAD. This illustrated that DGAT-cancer was more likely to detect active genes in cancer and paracancerous tissues than other methods without considering gene expression. DGAT-cancer also used features based on mutation frequency and expression profiles in tumor cohorts, JSD and C score (Methods). For these two scores, genes missed by other methods had relatively lower JSD scores in three cancers (BRCA, COAD and STAD) and lower C scores in four cancers (BRCA, COAD and STAD) than genes missed by DGAT-cancer (Fig. 4b). The C score is an indicator of how mutations are distributed among genes with similar expression patterns. A lower C score suggests that mutations are less uniformly distributed across these genes. Similarly, the JSD score measures the divergence between gene expression and mutation frequency distributions, with a lower score indicating greater similarity. Our findings indicate that DGAT-cancer is particularly effective at identifying genes that exhibit diverse mutation frequencies within similar expression profiles, which other methods might not detect.

## Gene expression-based features in tumour and paracancerous tissue served to improve DGAT-cancer

DGAT-cancer integrated both mutation-based features (uEMD-Mut) and gene expression-based features (uEMD-Ex and MutExTDA) to identify cancer drivers. The importance of these two types of features in the prediction was then evaluated. First, we used only uEMD-Mut to predict cancer drivers for each of the nine cancer types (Methods). The comparison between DGAT-Mut with other methods was based on a total of 23,433 genes that have the prediction score from at least one of the methods. Table S11 shows the numbers of positive and negative genes used for evaluating the methods. As shown in Fig. S6, DGAT-Mut yielded higher AUPRC values (ranging from 0.272 to 0.370) for predicting cancer drivers in BLCA, BRCA, COAD, GBM and STAD than OncodriveCLUSTL (ranging from 0.288 to 0.356), OncodriveFML (ranging from 0.283 to 0.365), MutSigCV (ranging from 0.225 to 0.282), DiffMut (ranging from 0.204 to 0.250) and TDAmut (ranging from 0.274 to 0.334). MutExTDA is a method that integrates topological data analysis with mutation and expression profiles, quantifying the correlation between mutation frequency and mRNA levels across cancer. When uEMD-TDA was utilized exclusively as a predictive factor for identifying cancer drivers across the nine cancer types, it demonstrated AUPRC values (ranging from 0.185 to 0.283) that were consistently lower than those achieved by the DGAT-cancer model (ranging from 0.336 to 0.468). The AUPRC values achieved by DGAT-Mut and DGAT-TDA in predicting cancer drivers for ninecancer types were lower than with DGAT-cancer. ​This comparison underscores the enhanced predictive power of the DGAT-cancer model, demonstrating the importance of combining mutation-based and gene expression-based features for the improved identification of cancer driver genes. Although the gene expression-based model, uEMD-Ex, showed lower AUPRC values than DGAT-cancer in two out of seven cancer types, including GBM and STAD, it demonstrated superior predictive performance across all seven types when compared to alternative methods such as MutSigCV, OncodriveFML, OncodriveCLUST, DiffMut, and TDAmut (Fig. S6). DGAT-cancer and DGAT-Exp were not applied to the prediction of cancer drivers for two cancer types (CESC and LGG) because limited paracancer was available from the TCGA for these two types of cancer (Methods). These results confirm the key role of gene expression-based features in the model. Thus, combining both mutation-based features and expression-based features improved the performance of DGAT-cancer.

## DGAT-cancer identifies novel cancer drivers

DGAT-cancer identified many novel cancer driver genes that have not hitherto been reported to be related to cancer in CGC, OncoKB or IntOGen gene sets (Table S12). We have compiled all the gene sets from our study into a comprehensive supplementary spreadsheet (Table S13) for future benchmarking purposes.

We next wondered if these predicted novel cancer drivers might correlate with the prognosis of the cancer patients (Methods). Using pre-analysis data from The Human Protein Atlas (HPA), which utilizes Kaplan-Meier models, we evaluated the link between gene expression and patient survival in multiple cancer types: BRCA, BLCA, COAD, GBM, HNSC, LUAD, and STAD. [33](#_ENREF_33). By comparing the number of genes whose mRNA expression level significantly (log-rank ) correlated with patient survival in at least one cancer type, we observed that the predicted novel cancer drivers contain a significantly higher proportion of genes (total number 481 and proportion 96.83%) (one-sided Fisher’s Exact test, , ) that correlated with patient survival than the genes predicted to be of low probability to be cancer drivers (removing genes contained in CGC, OncoKB, IntOGen and predicted cancer drivers, total number 7,111 and proportion 91.52%). The HPA pathology data revealed significant correlations between *EEF1A1* expression levels and patient survival outcomes in CESC (), COAD (), GBM (), and LGG (). However, the HPA only provides pre-computed survival analysis results but not the original survival data of patients, which has made it impossible to plot the KM curves of the patients. Given that *EEF1A1* was predicted as a potential cancer driver gene across LUAD, GBM, BRCA, BLCA and HNSC in our analysis (Table S12), we proceeded to conduct additional experiments to explore its potential role in glioma.

## *EEF1A1* expression is increased in glioma and correlates with a poor prognosis

We compared the mRNA expression levels of *EEF1A1* in 698 glioma samples from the TCGA GBM/LGG dataset and 1,157 normal brain samples from GTEx[34](#_ENREF_34" \o "Consortium, 2013 #79). The results showed that *EEF1A1* expression was upregulated in tumor tissue relative to normal brain tissue, as well as significantly increased in GBM relative to LGG (Both P<0.001, Fig. 5a). Moreover, the expression level of *EEF1A1* in the glioma samples from TCGA was found to negatively correlate with the overall survival of patients (*p*=0.021, Fig. 5b).

The *EEF1A1* protein level was examined in 91 self-collected glioma specimens by immunohistochemical analysis. Relative to non-tumor brain tissues, *EEF1A1* exhibited a significantly higher level in LGG (, Fig. 5c) and GBM , Fig. 5c). Notably, the expression of *EEF1A1* protein was significantly increased in GBM as compared with LGG (, Fig. 5c). When we cultured GBM cell lines (U251 and U87 cells) and LGG cell lines (Hs683), we found that *EEF1A1* protein was more highly expressed in GBM cells (U251 and U87) than in LGG cells (Hs683) ( for U87 *vs*. Hs683, for U251 *vs*. Hs683, Fig. S7). Thus, we surmised that *EEF1A1* may be involved in glioma tumorigenesis.

## Knockdown of *EEF1A1* inhibited the proliferation and migration of glioma cells

We used shRNA to knock down *EEF1A1* expression in U251 and U87 cells. The real-time quantitative PCR (RT-qPCR) showed that compared with the control group, *EEF1A1* mRNA in the knockdown group decreased by about 77% and 86% in U87 and U251 glioma cells ( for U87 and for U251, Fig. S8a), respectively. The western blot experiments showed that *EEF1A1* protein expression in the knockdown groups decreased by 65% and 45% in U87 and U251 glioma cells ( for U87 and for U251), respectively (Fig. S8b).

To assess the effect of *EEF1A1* knockdown on the proliferation of U251 and U87 cells, we performed a Cell Counting Kit-8 (CCK-8) assay, an EdU flow cytometry assay and a colony formation assay. The results showed that U251 and U87 cells with reduced *EEF1A1* expression exhibited a significant decrease in cell viability within 72h according to the CCK-8 assay (-25% and for U251, -11% and for U87, Fig. S8c), whilst the proportions of EdU-positive cells were decreased by 45% for U251 () and by 38% for U87 () (Fig. 5d), with the numbers of colonies being decreased by 54% for U251 () and by 69% for U87 () (Fig. 5e) compared with the control group. These results demonstrated that *EEF1A1* knockdown significantly inhibited the proliferation of U251 and U87 cells.

After *EEF1A1* knockdown, the percentage of cells migrating through the transwell plate significantly decreased (-69% and for U251, -71% and for U87, Fig. 5f and Fig. S8d). A scratch-wound healing assay yielded a similar result. A significantly shorter migration distance was observed in U251 cells with *EEF1A1* knockdown (-44%, , Fig. S8e) and U87 cells (-30%, , Fig. S8f). To assess the proliferation of glioma cells *in vivo*, U251 and U87 cells transfected with *EEF1A1* shRNA or control shRNA were injected into zebrafish and the areas of GFP fluorescent foci were measured. As shown in Fig. 5g, U251 and U87 cells with *EEF1A1* knockdown exhibited a significant decrease in the areas of fluorescent foci (-50% and for U251, -41% and for U87, Fig. S8g). Taken together, these results suggest that *EEF1A1* plays a role in regulating the proliferation and migration of glioma cells.

## Knockdown of *EEF1A1* increased temozolomide (TMZ) sensitivity in glioma cells

Temozolomide (TMZ) is a chemotherapy drug that treats gliomas by methylating DNA and inducing apoptosis in cancer cells[35](#_ENREF_35). In order to explore the role of *EEF1A1* in the sensitivity of glioma cells to TMZ, U251 and U87cells with *EEF1A1* knockdown, and control glioma cells were cultured in different concentrations of TMZ. The results showed that knockdown of *EEF1A1* significantly decreased cell viability at different concentrations of TMZ ( for U251 cells and for U87 cells, Fig. 6a), as well as at the half maximal inhibitory concentration (IC50) of TMZ in glioma cells (-41% and for U251, -44% and  for U87, Fig. 6b). Next, a colony formation assay was performed in cells with treatment of TMZ (50μM). In the presence of TMZ, knockdown of *EEF1A1* significantly reduced colony formation (-64% and for U251 cells, -58% and for U87 cells, Fig. 6c and d).

Culturing U251 and U87 cells in each group with 200μM TMZ, we observed that the cell viability of the *EEF1A1*-knockdown group was significantly lower () than that of the control group according to CCK-8 assays (Fig. S9a). These results suggest that the knockdown of *EEF1A1* is able to inhibit the proliferation of glioma cell lines in the presence of TMZ.

To investigate the effect of *EEF1A1* on the apoptosis of glioma cells treated by TMZ, we cultured U251 and U87 cells with 200μM TMZ for 24h, and assessed cellular apoptosis by means of a flow cytometer. The results showed that the proportion of apoptotic cells in the *EEF1A1*-knockdown group was significantly higher than that in the control group (+127% and for U251, +129% and for U87, Fig. 6e and f). Moreover, apoptosis-related proteins also showed significant changes after the knockdown of *EEF1A1*. Thus, cleaved caspase-3, the effector of apoptotic activity, was significantly increased, whilst bcl-2, an inhibitor of apoptosis, was significantly decreased in *EEF1A1*-knockdown glioma cells (Both ) (Fig. 6g and h). These findings indicated that the decreased expression of *EEF1A1* can promote apoptosis in glioma cells, thereby increasing the TMZ sensitivity of glioma cells.

# Discussion

The computational identification of cancer driver genes is key to improving our understanding of the underlying mechanisms of tumorigenesis. Most of the current methods for the identification of cancer drivers rely on the use of a single type of genomic data, such as mutations or gene expression[36](#_ENREF_36). Here, we have developed a method, DGAT-cancer, which integrates genomics and transcriptomics data in tumors and the healthy population for predicting cancer drivers. It takes advantage of the huge amount of mutation data from individual samples generated by the 1000 Genomes Project, the TCGA project and the ICGC project to obtain the distribution differences of the predicted pathogenic scores of mutations in the healthy population and in tumor tissues. DGAT-cancer is capable of detecting cancer drivers for any type of cancer by inputting mutation and/or gene expression data.

Our analysis indicated that employing only mutational information in cancer driver prediction was insufficient to achieve enhanced performance. As indicated in Fig. S6, DGAT-Mut only performed better than the method, DGAT-Exp that only uses the gene expression data in its prediction of cancer drivers in BRCA, HNSC and LUAD. This result may reflect the important roles of gene expression data in tumor tissue in predicting cancer drivers. Compared to other methods, DGAT-cancer has the ability to recognize genes with a higher expression level in tumors and paracancerous tissues (Fig. 4b), suggesting the importance of introducing tissue-specific gene expression as a feature in predicting cancer driver genes.

DGAT-cancer predicted many novel candidate genes that require further experimental validation. For example, *AHNAK* was predicted as a cancer driver in BRCA, COAD, LUAD and STAD. It encodes a large structural scaffold protein and has been reported as a tumour suppressor in the proliferation and invasion of triple-negative BRCA[37](#_ENREF_37) and LUAD[38](#_ENREF_38). The DGAT-cancer predicted *EEF1A1* as a cancer driver gene across five cancer types: BRCA, LUAD, BLCA, HNSC, and GBM. The elevated expression of *EEF1A1* in BRCA[39](#_ENREF_39), LUAD[40](#_ENREF_40) and GBM[41](#_ENREF_41) is consistent with prior experimental validations, underscoring its significance and reinforcing the rationale for its further study as a potential therapeutic target. Furthermore, recent bioinformatics studies on GBM and LGG have confirmed high *EEF1A1* expression, bolstering its role as a significant factor in glial tumors[42](#_ENREF_42),[43](#_ENREF_43). To validate its role in GBM, we have initiated supplementary experimental analyses.

*EEF1A1* protein has been reported to interact with key components of the pathway that regulates the synthesis of apoptosis-related proteins[44-46](#_ENREF_44). To explore its clinical relevance, we assessed the impact of EEF1A1 knockdown on U251 and U87 glioma cells treated with TMZ. TMZ is a key chemotherapy agent for gliomas, including GBM, operating through DNA methylation that leads to replication disruption and apoptosis in tumor cells[35](#_ENREF_35). However, we found that EEF1A1 knockdown inhibits apoptosis in glioma cells when treated with TMZ. In contrast, when EEF1A1 was knocked down in glioma cells without TMZ treatment, no change in apoptosis was observed (Fig. S9b). The underlying mechanisms require further investigation. Additionally, our study indicates that overexpression of *EEF1A1* correlates with a poorer prognosis in glioma (Fig. 5b), a finding that contrasts with the results reported by Hassan et al.[43](#_ENREF_43) and Petkovic et al[42](#_ENREF_42). ​ These discrepancies may arise from differences in the datasets or patient cohorts utilized across studies, as well as variations in tumor staging of the samples.The underline reason may due to the complexity of tumor genetic and phenotypic characteristics, which can influence the role of *EEF1A1* in tumor progression and response to therapy*.* The prognosis of gliomas is known to be context-dependent, and the role of *EEF1A1* may further modulate this by potentially influencing the immune response or interacting with other molecular pathways, which could in turn affect treatment response and patient outcomes[47](#_ENREF_47). It is worth noting that, although *EEF1A1* may promote glioma progression, *EEF1A1* overexpression may also potentially improve therapeutic outcomes and patient prognosis. Our study provides valuable insights into the role of *EEF1A1* in glioma resistance to TMZ. However, further research with more extensive experimental data and clinical samples is needed to fully elucidate the precise mechanisms of *EEF1A1* in gliomas.

There are many ways to further improve the accuracy of DGAT-cancer. First, DGAT-cancer is dependent on mutation data from the general population and cancer population. When the sample size is increased, more mutational information will become available which will help to improve the accuracy of predictions. Moreover, DGAT-cancer only takes account of mutation and gene expression data in its predictions. Other features, such as the protein or RNA structures around the mutations could also contribute important information that might be of use in discriminating cancer drivers. Finally, since the performance of DGAT-cancer is influenced by pathogenicity predictions for mutations, more accurate predictors for determining the pathogenicity of mutations could help to improve DGAT-cancer.

# Conclusions

We demonstrate that DGAT-cancer is powerful in predicting cancer drivers using mutation and/or gene expression data, and has a superior performance compared to three commonly used methods. The importance of gene expression data and mutation information in predicting cancer drivers was evidenced. DGAT-cancer has broadened our path to detect cancer driver genes and shed a light on cancer therapy.

# Declarations

## Ethics approval and consent to participate

The study was approved by the Ethics Committee of Sun Yat-sen University, and informed consent was obtained from all subjects.

## Consent for publication

Not applicable.

## Availability of data and materials

The DGAT-cancer method is available as an open-source software package on the GitHub repository (https://github.com/Dan-He/DGAT-cancer). Simple somatic mutation (SSM) data from 12 cancer cohorts were downloaded from the Broad Institute GDAC Firehose Portal (http://gdac.broadinstitute.org/), the International Cancer Genome Consortium (ICGC) Data Portal (https://dcc.icgc.org/releases/current/Projects/) and The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). The germline mutation data of a healthy population were collected from Phase 3 of the 1000 Genomes Project (https://www.internationalgenome.org/data, GRCh38). RNA-seq data of tumors from 12 cancer types in TCGA were downloaded from the UCSC Xena platform (http://xena.ucsc.edu/).

Additional information is available at the website.

## Competing interests

All authors declared that they have no competing interests.

## Funding

This work was supported by the National Key Research and Development Program of China (2020YFB0204803), the Natural Science Foundation of China (81801132, 81971190, 61772566), Guangdong Key Field Research and Development Plan (2019B020228001, 2018B010109006, and 2021A1515010256), Introducing Innovative, Guangzhou Science and Technology Research Plan (202007030010).

## Author’s contributions

HY.Z designed the study. D.H performed the model construction and analyses, with assistance from Z.L and T.L. L.L performed the experiments, with assistance from HS.Z, F.L, S.L and B.L. D.H, L.L and Z.L wrote the manuscript. HY.Z and D.N.C supervised the study. All authors discussed the results and interpretation and contributed to the final version of the paper.

## Acknowledgments

The authors thank many resources for making data available.

# References

1 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).

2 Cheng, F., Zhao, J. & Zhao, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief Bioinform* **17**, 642-656, doi:10.1093/bib/bbv068 (2016).

3 Suo, C. *et al.* Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics* **31**, 2607-2613, doi:10.1093/bioinformatics/btv164 (2015).

4 Gumpinger, A. C., Lage, K., Horn, H. & Borgwardt, K. Prediction of cancer driver genes through network-based moment propagation of mutation scores. *Bioinformatics* **36**, i508-i515, doi:10.1093/bioinformatics/btaa452 (2020).

5 Petrov, I. & Alexeyenko, A. Individualized discovery of rare cancer drivers in global network context. *Elife* **11**, doi:10.7554/eLife.74010 (2022).

6 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

7 Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**, 128, doi:10.1186/s13059-016-0994-0 (2016).

8 Arnedo-Pac, C., Mularoni, L., Muinos, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788-4790, doi:10.1093/bioinformatics/btz501 (2019).

9 Rabadán, R., Mohamedi, Y., Rubin, U., Chu, T. & Alghalith, A. N. Identification of relevant genetic alterations in cancer using topological data analysis. **11**, 3808, doi:10.1038/s41467-020-17659-7 (2020).

10 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. *Human mutation* **37**, 235-241, doi:10.1002/humu.22932 (2016).

11 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

12 Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics (Oxford, England)* **25**, 2744-2750, doi:10.1093/bioinformatics/btp528 (2009).

13 Flanagan, S. E., Patch, A. M. & Ellard, S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers* **14**, 533-537, doi:10.1089/gtmb.2010.0036 (2010).

14 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

15 Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, doi:10.1200/PO.17.00011 (2017).

16 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

17 Martinez-Jimenez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555-572, doi:10.1038/s41568-020-0290-x (2020).

18 Rabadan, R. *et al.* Identification of relevant genetic alterations in cancer using topological data analysis. *Nat Commun* **11**, 3808, doi:10.1038/s41467-020-17659-7 (2020).

19 Singh, G., Mémoli, F. & Carlsson, G. E. in *PBG@Eurographics.*

20 Shuai, S. *et al.* Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nat Commun* **11**, 734, doi:10.1038/s41467-019-13929-1 (2020).

21 He, X., Cai, D. & Niyogi, P. in *Proceedings of the 18th International Conference on Neural Information Processing Systems* 507–514 (MIT Press, Vancouver, British Columbia, Canada, 2005).

22 He, D., Fan, C., Qi, M., Yang, Y. & Cooper, D. N. Prioritization of schizophrenia risk genes from GWAS results by integrating multi-omics data. **11**, 175, doi:10.1038/s41398-021-01294-x (2021).

23 Valentini., G. & Re., M. PerfMeas: Performance Measures for ranking and classification tasks. (2014).

24 Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187-200, doi:10.1002/pro.3978 (2021).

25 Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. **21**, 6, doi:10.1186/s12864-019-6413-7 (2020).

26 Przytycki, P. F. & Singh, M. Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. *Genome medicine* **9**, 79, doi:10.1186/s13073-017-0465-6 (2017).

27 Cheng, F. *et al.* Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol Biol Evol* **31**, 2156-2169, doi:10.1093/molbev/msu167 (2014).

28 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950, doi:10.1038/ng.3050 (2014).

29 Cowley, G. S. *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci Data* **1**, 140035, doi:10.1038/sdata.2014.35 (2014).

30 Pacini, C. *et al.* A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer cell* **42**, 301-316.e309, doi:10.1016/j.ccell.2023.12.016 (2024).

31 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).

32 Luna, A. *et al.* CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics. *Nucleic Acids Res* **49**, D1083-D1093, doi:10.1093/nar/gkaa968 (2021).

33 Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, doi:10.1126/science.aan2507 (2017).

34 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).

35 Stupp, R. *et al.* Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England journal of medicine* **352**, 987-996, doi:10.1056/NEJMoa043330 (2005).

36 Pham, V. V. H. *et al.* Computational methods for cancer driver discovery: A survey. *Theranostics* **11**, 5553-5568, doi:10.7150/thno.52670 (2021).

37 Chen, B. *et al.* AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. *J Exp Clin Cancer Res* **36**, 65, doi:10.1186/s13046-017-0522-4 (2017).

38 Park, J. W. *et al.* AHNAK Loss in Mice Promotes Type II Pneumocyte Hyperplasia and Lung Tumor Development. *Mol Cancer Res* **16**, 1287-1298, doi:10.1158/1541-7786.MCR-17-0726 (2018).

39 Tomlinson, V. A. *et al.* Translation elongation factor eEF1A2 is a potential oncoprotein that is overexpressed in two-thirds of breast tumours. *BMC cancer* **5**, 113, doi:10.1186/1471-2407-5-113 (2005).

40 Wu, Y. *et al.* [Role of Eukaryotic Translation Elongation Factor 1 Family Members in the Tumorigenesis and Progression of Lung Adenocarcinoma]. *Zhongguo yi xue ke xue yuan xue bao. Acta Academiae Medicinae Sinicae* **45**, 867-885, doi:10.3881/j.issn.1000-503X.15580 (2023).

41 Scrideli, C. A. *et al.* Gene expression profile analysis of primary glioblastomas and non-neoplastic brain tissue: identification of potential target genes by oligonucleotide microarray and real-time quantitative PCR. *Journal of neuro-oncology* **88**, 281-291, doi:10.1007/s11060-008-9579-4 (2008).

42 Petkovic, M., Yalçin, M., Heese, O. & Relógio, A. Differential expression of the circadian clock network correlates with tumour progression in gliomas. *BMC medical genomics* **16**, 154, doi:10.1186/s12920-023-01585-w (2023).

43 Hassan, M. K., Kumar, D., Naik, M. & Dixit, M. The expression profile and prognostic significance of eukaryotic translation elongation factors in different cancers. **13**, e0191377, doi:10.1371/journal.pone.0191377 (2018).

44 Cancer Genome Atlas Research Network. Electronic address, w. b. e. & Cancer Genome Atlas Research, N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341 e1323, doi:10.1016/j.cell.2017.05.046 (2017).

45 Kobayashi, Y. & Yonehara, S. Novel cell death by downregulation of eEF1A1 expression in tetraploids. *Cell Death Differ* **16**, 139-150, doi:10.1038/cdd.2008.136 (2009).

46 Vera, M. *et al.* The translation elongation factor eEF1A1 couples transcription to translation during heat shock response. *Elife* **3**, e03164, doi:10.7554/eLife.03164 (2014).

47 Kut, C. & Kleinberg, L. Radiotherapy, lymphopenia and improving the outcome for glioblastoma: a narrative review. *Chinese clinical oncology* **12**, 4, doi:10.21037/cco-22-94 (2023).

# Figure legends

**Fig. 1. Overview of DGAT-cancer.** First, somatic mutations were collected from two databases, ICGC and TCGA, as well as germline mutations from the 1000 Genomes Project. The differences between the pathogenicity scores of mutated genes in cancer (somatic) and those in the general population (germline) were evaluated by DGAT-cancer calculating the uEMD scores. The RNA-seq data from tumors were used to construct a topological network. The gene expression topological network was then used to evaluate the frequencies of mutation spectra occurring in different sample clusters. For the mutations occurring in adjacent sample clusters, we calculated the divergence of their frequencies across clusters, which was used as one feature of the genes harboring the mutations. We also calculated the Jensen–Shannon divergence between the mutation frequency and the mRNA expression of the gene across clusters of the network as one feature. All collected features of genes were filtered using Laplacian scores. We transformed these selected features using Hotelling and Box-Cox transformations in order to integrate them into one composite risk score. By using gene scores as weights, we performed Gibbs sampling to sample genes in order to identify cancer drivers. Finally, we generated null distributions of PP for each gene to compute an experience P-value. The genes with were selected as candidate cancer driver genes.

**Fig. 2. Evaluation of the cancer driver genes predicted by DGAT-cancer.** **a** Enrichment of predicted cancer drivers in known cancer gene sets (CGC, OncoKB, IntOGen) and gene sets related to cancer (constraint: selective constraint genes, and genes with expression associated with functions of cancer cells). **b** Top 20 genes most frequently predicted as cancer drivers. The color depth denotes the rank order of PP (decreasing) for genes generated by DGAT-cancer in that cancer type. **c** Enrichment of predicted cancer drivers in drug-targeted genes. **d** The proportions of predicted cancer drivers whose expression levels were significantly correlated with drug activity (measured in terms of 50% growth inhibitory levels) compared to the background genes. The P-values were obtained by using Fisher’s Exact test to compare the predicted cancer drivers with predicted non-cancer drivers. \*; \*\*; \*\*\*.

**Fig. 3. Comparison of DGAT-cancer with five methods with respect to their prediction of cancer drivers. a** Comparison of DGAT-cancer with other methods in terms of their performance as measured by AUPRC (area under the precision–recall curve). **b** Comparative analysis of the DGAT-cancer method's performance across multiple cancer types, with a focus on MCC, F1 score, and specificity metrics. **c** The UpSet plot visualizes the intersection of predicted cancer driver genes across seven cancer types as identified by DGAT-cancer.

**Fig. 4. Assess the Consistency of DGAT-cancer with Other Methods. a** Comparing posterior probabilities (PPs) given by DGAT-cancer of genes predicted as cancer drivers by MutSigCV, OncodriveCLUSL and OncodriveFML to the genes predicted as non-cancer drivers by those methods. The P-values were obtained by means of the Wilcoxon rank-sum test. In the boxplot, the center lines represent the median, whilst the boxes represent the first and third quartiles. **b** Comparison between genes missed by DGAT-cancer and genes missed by other methods in terms of their feature scores given by DGAT-cancer. Those scores are uEMD-Ex scores, gene expression level in tumor (tumor-med), gene expression level in paracancerous tissues (normal-med), and MutExTDA scores (JSD and C score). Some boxes are blank due to the missing features in that cancer type. The difference was evaluated by Wilcoxon rank-sum test, \* ; \*\* ; \*\*\* .

**Fig. 5. *EEF1A1* is highly expressed in glioma and plays a role in regulating the proliferation and migration of glioma cells. a** Profile of *EEF1A1* mRNA expression in normal, LGG (WHO grade II–III glioma), or GBM (WHO grade IV glioma) patients in the GTEx and TCGA datasets. *EEF1A1* expression was significantly upregulated in glioma relative to normal brain tissue (left) (Mann-Whitney U test), as well as significantly increased in GBM relative to LGG (right) (Kruskal-Wallis test, adjusted by Bonferroni correction). **b** Kaplan–Meier overall survival plot showing that the survival rate of glioma patients with high *EEF1A1* expression (red) in the TCGA data set was significantly lower than those with low *EEF1A1* expression (blue) (two-sided log-rank test). **c** *EEF1A1* protein expression in non-tumor brain tissue, LGG and GBM samples using immunohistochemical analysis. Non-tumor brain tissues were obtained from patients with non-tumor brain diseases who had undergone surgical resection. Representative images of immunohistochemical analysis. Scale bar=10μm. The summary of *EEF1A1* protein expression profile in C (Mann-Whitney U test) was on the right. **d** EdU flow cytometry assay showed that the cell proliferation rate was significantly decreased in U251 and U87 cells with the knockdown of *EEF1A1*. **e** *In vitro* colony formation of U251 and U87 cells was decreased after *EEF1A1* knockdown compared to the control. **f** Representative images of the transwell migration assay for migrated cells stained with crystal violet in control cells and shEEF1A1 knockdown U251 and U87 cells. Scale bar = 1mm. **g** Representative images of the zebrafish xenograft model used to analyze the proliferation of U251 cells after *EEF1A1* knockdown by measuring GFP fluorescent foci. Scale bar = 0.5mm. Unpaired two-sided t test. \*; \*\*; \*\*\*; \*\*\*\*.

**Fig. 6. Knockdown of *EEF1A1* improved TMZ sensitivity in glioma cells. a** Knockdown of *EEF1A1* significantly decreased the viability of U251 and U87 cells under different concentrations of TMZ (24h for U251 cells and 48h for U87 cells) (*n*=6, paired two-sided t test). **b** Knockdown of *EEF1A1* significantly decreased the IC50 of TMZ in U251 and U87 cells (*n*=6, unpaired two-sided t test). **c** In the presence of TMZ (50 ug/mL), the knockdown of *EEF1A1* significantly reduced colony formation. **d** Quantitation of colony formation in (**c**) (*n*=3, unpaired two-sided t test). **e** Flow cytometry results showing the proportions of apoptotic cells (the Annexin V V450-positive cells) in the EEF1A1-knockdown and control groups in U251 and U87 cells**. f** The summary of six independent experiments of (**e**) (unpaired two-sided t test). **g** Western-blot showing the cleaved caspase-3 and bcl-2 protein expression in the *EEF1A1*-knockdown and control groups in U251 and U87 cells. **h** Quantitation of four independent experiments of (**g**). (Mann-Whitney U test) \*; \*\*; \*\*\*; \*\*\*\*.