Original Article

Prediction of *BRAF* V600E variant from cancer gene expression data

Jun Kang[1]^, Jieun Lee[2]^, Ahwon Lee[1,3], Youn Soo Lee[1]

[1]Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea; [2]Division of Medical Oncology, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea; [3]Cancer Research Institute, The Catholic University of Korea, Seoul, Korea

Contributions: (I) Conception and design: J Kang, A Lee, YS Lee; (II) Administrative support: J Kang; (III) Provision of study materials or patients: J Kang; (IV) Collection and assembly of data: J Kang; (V) Data analysis and interpretation: J Kang, J Lee; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Youn Soo Lee, MD, PhD. Department of Hospital Pathology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 222, Banpo-daero, Seocho-gu, Seoul 06591, Korea. Email: lys9908@catholic.ac.kr.

Background: BRAF inhibitors have been approved for the treatment of melanoma, non-small cell lung cancer, and colon cancer. Real-time polymerase chain reaction or next-generation sequencing were clinically used for *BRAF* variant detection to select who responds to BRAF inhibitors. The prediction of *BRAF* variants using gene expression data might be an alternative test when the direct variant sequencing test is not feasible. In this study, we built a prediction model to detect *BRAF* V600 variants with mRNA gene expression data in various cancer types.

Methods: We adopted a penalized logistic regression for the *BRAF* V600E variants prediction model. Ten times bootstrap resampling was done with a combined target variable and cancer type stratification. Data preprocessing included knnimputation for missing value imputation, YeoJohnson transformation for skewness correction, center, and scale for standardization, synthetic minority over-sampling technique for class imbalance. Hyperparameter optimization with a grid search was undertaken for model selection in terms of area under the precision-recall.

Results: The area under the curve of the receiver operating characteristic curve on the test set was 0.98 in thyroid carcinoma, 0.90 in colon adenocarcinoma, and 0.85 in cutaneous melanoma. The area under the precision-recall of the test set was 0.98 in thyroid carcinoma, 0.71 in colon adenocarcinoma, and 0.65 in cutaneous melanoma.

Conclusions: Our penalized logistic regression model can predict *BRAF* V600E variants with good performance in thyroid carcinoma, cutaneous melanoma, and colon adenocarcinoma.

Keywords: *BRAF*; machine learning; The Cancer Genome Atlas (TCGA); BRAF kinase inhibitor

Submitted Mar 31, 2022. Accepted for publication Sep 07, 2022.
doi: 10.21037/tcr-22-883
View this article at: https://dx.doi.org/10.21037/tcr-22-883

Introduction

BRAF gene encodes a serine/threonine kinase and is known to be an oncogene (1,2). BRAF regulates the mitogen-activated protein kinase (MAPK) pathway. The V600E is the most common somatic *BRAF* variant followed by V600K/D/R/M and non-V600 variants (3). Knowing the presence of these *BRAF* variants is important to make a plan for patient treatment, especially in melanoma and colorectal

^ ORCID: Jun Kang, 0000-0002-7967-0917; Jieun Lee, 0000-0002-2656-0650.

© Translational Cancer Research. All rights reserved.

carcinoma.

The presence of *BRAF* variants is a marker to screen Lynch syndrome in microsatellite-unstable (MSI-H) colorectal cancer (4). Lynch syndrome is an autosomal dominant hereditary cancer syndrome associated with mismatch repair gene deficiency. The presence of a BRAF V600E variant suggests that MSI-H colorectal cancer is sporadic tumor rather than a component of Lynch syndrome-associated malignancy (5).

Real-time polymerase chain reaction (PCR) or next-generation sequencing were traditionally used for *BRAF* variant detection to select who will respond to the BRAF inhibitors. Recently immunohistochemistry and digital polymerase chain reaction are used for detecting *BRAF* V600E variant (6,7). BRAF inhibitors have been approved for the treatment of melanoma (8-10), non-small cell lung cancer (11), and colon cancer (12). The prediction of *BRAF* variants using gene expression data might be an alternative test when the direct variant sequencing test is not available or fails.

We have built prediction models to detect *PIK3CA* variants and homologous recombination deficiency with mRNA gene expression data using The Cancer Genome Atlas (TCGA) pan-cancer data (13). TCGA is a large cancer genomic consortium including more than 10,000 specimens from 25 different tumor types with exome sequencing, mRNA gene expression, DNA methylation, and clinical data (14). In this study, we try to develop a prediction model to detect *BRAF* V600E variant with mRNA gene expression data in various cancer types. We present the following article in accordance with the TRIPOD reporting checklist (available at https://tcr.amegroups.com/article/view/10.21037/tcr-22-883/rc).

Methods

Dataset

We used TCGA pan-cancer data. The mRNA gene expression data were downloaded from the National Cancer Institute (NCI)'s Genomic Data Commons (GDC) website (https://gdc.cancer.gov/about-data/publications/pancanatlas). Data of *BRAF* variants were obtained from the cbioportal website (15).

We only included the presence of *BRAF* V600E variants as the target variable because BRAF inhibitors have been approved for cancers with *BRAF* V600E variants but not for other *BRAF* variants. Predictor variables were mRNA gene expression and cancer types. The mRNA gene expression predictor variables were filtered with a median absolute deviation to exclude less informative variables.

The *BRAF* V600E variants were frequently observed in thyroid carcinoma, cutaneous melanoma, and colon adenocarcinoma and very rarely observed in other cancer types. We used three-quarters of the three cancer types with a high prevalence of *BRAF* V600E variants for the training set and the remaining test set. The other cancer types with a low prevalence of *BRAF* V600E variants were regarded as an unseen test set.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Ethical approval is not required because we used public databases according to the TCGA publication guidelines (https://cancergenome.nih.gov/publications/guidelines).

Dataset summary

The number of included cases of the training set, the test set, and the unseen test set was 1,136, 376, and 9,377, respectively. A total of 5,129 mRNA gene expression predictors were selected after filtering with median absolute deviation. The prevalence of *BRAF* V600E variants was 0.57 (326/568 cases) for thyroid carcinoma, 0.33 (190/469 cases) for cutaneous melanoma, and 0.10 (49/475) for colon adenocarcinoma. Cancer type abbreviation of pan TCGA dataset and number of cases of each cancer type are summarized in Table S1.

Prediction modeling

We adopt a penalized logistic regression for the *BRAF* V600E variants prediction model (16). Tidymodels was used for the modeling process. Tidymodels is a framework that is a collection of R packages (R project for Statistical Computing, RRID:SCR_001905) for modeling and machine learning.

Penalized logistic regression has two hyperparameters which are the amount of regularization (λ) and the proportion of lasso penalty (α). Bootstrap resampling was used to determine those hyperparameters. Ten times bootstrap resampling was performed with a combined target variable and cancer type stratification.

Data preprocessing included knnimputation for missing value imputation, and YeoJohnson transformation for skewness correction, center, and scale for standardization,

© Translational Cancer Research. All rights reserved.

with the synthetic minority over-sampling technique (smote) for class imbalance.

Hyperparameter optimization with a grid search was done for model selection in terms of area under the precision-recall (AUPR). AUPR is better than area under the receiver operating characteristic (AUROC) to compare model performance with an imbalanced dataset (17). The hyperparameter grid was set into λ (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}) and α (0.0, 0.25, 0.5, 0.75, 1.0).

Assessing model performance

Model performance was estimated on the test set of the cancer types with a high prevalence of *BRAF* V600E variants and the test set of other cancer types with a low prevalence of *BRAF* V600E variants as an unseen test set in terms of AUPR.

Gene ontology test

The gene ontology test was done with the PANTHER overrepresentation test (18) to determine which pathways are important in predicting the *BRAF* V600E variants. The selected predictor genes after final model fitting with entire training set were evaluated for gene ontology test with following detailed PANTHER parameters (analysis type: PANTHER Overrepresentation Test (Released 20210224), Annotation Version and Release Date: PANTHER version 16.0 Released 2020-12-01, Reference List: Homo sapiens (all genes in database), Test Type: FISHER, Correction: FDR).

Statistical analysis

All statistical analysis was done using R (R Project for Statistical Computing, RRID:SCR_001905).

Results

Model summary

The hyperparameter was chosen as 10^{-5} for λ and 0.25 for α. Those hyperparameter values showed the highest AUPR by 10 times bootstrap resampling. After model fitting with the entire training set and selected hyperparameters, 546 predictors were included in the final model. The cancer types were excluded from the final model. The coefficient values of genes that were included in the final model are summarized in Table S2. A predicted probability was calculated by the final logistic model after pre-determined data preprocessing. Genes with the largest positive coefficient value included *ETS variant transcription factor 1* (*ETV1*), *AKT serine/threonine kinase 2* (*AKT2*), *neurofibromin 1* (*NF1*) and *nuclear factor kappa B subunit 1* (*NFKB1*).

Performance of prediction model

The AUROC of *BRAF* V600E variant prediction on the training set was 0.99 in thyroid carcinoma, and 1.00 in colon adenocarcinoma and cutaneous melanoma. The AUROC on the test set was 0.98 in thyroid carcinoma, 0.90 in colon adenocarcinoma, and 0.85 in cutaneous melanoma. The receiver operating characteristic curve (ROC curve) is illustrated in *Figure 1*.

The AUPR of *BRAF* V600 variant prediction on the training set was 0.99 in thyroid carcinoma, 1.00 in colon adenocarcinoma, and cutaneous melanoma. The AUPR on the test set was 0.98 in thyroid carcinoma, 0.71 in colon adenocarcinoma, and 0.65 in cutaneous melanoma. The precision-recall curve (PR curve) was illustrated in *Figure 2*.

AUROC was 0.52 and AUPR was 0.002 with 0.002 baselines on an unseen test set of other cancer types with a low prevalence of *BRAF* V600E variants.

Gene ontology test

The selected predictor genes were overrepresented in the following pathways: Insulin/IGF pathway-protein kinase B signaling cascade, PI3 kinase pathway, Endothelin signaling pathway, Integrin signaling pathway, Apoptosis signaling pathway, T cell activation, CCKR signaling map, Inflammation mediated by chemokine and cytokine signaling pathway, Gonadotropin-releasing hormone receptor pathway. Detailed gene ontology results are described in the Table S3.

Discussion

Our *BRAF* V600 variant prediction model showed very good performance on the test set of the cancer types including thyroid carcinoma, colon adenocarcinoma, and cutaneous melanoma. Those cancer types have a high prevalence of *BRAF* V600E variants. This result suggests that a *BRAF* V600 variant prediction model can help to select patients for treatment with BRAF inhibitors.

Gene expression signature has been used as a predictive biomarker in the practice of patient selection. Gene

© Translational Cancer Research. All rights reserved.

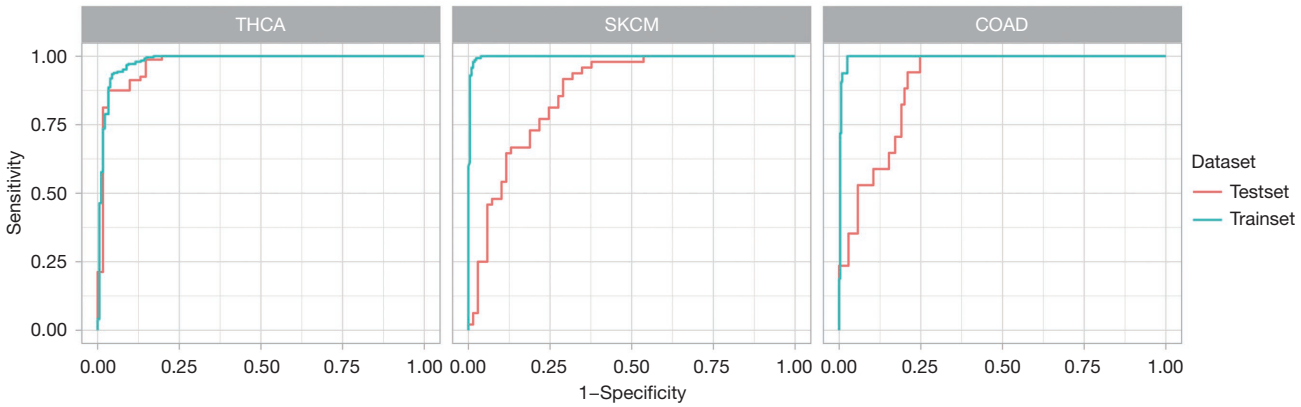

Figure 1 ROC curve of *BRAF* V600E variant prediction. THCA, thyroid carcinoma; SKCM, Cutaneous Melanoma; COAD, Colon adenocarcinoma; ROC, receiver operating characteristic.

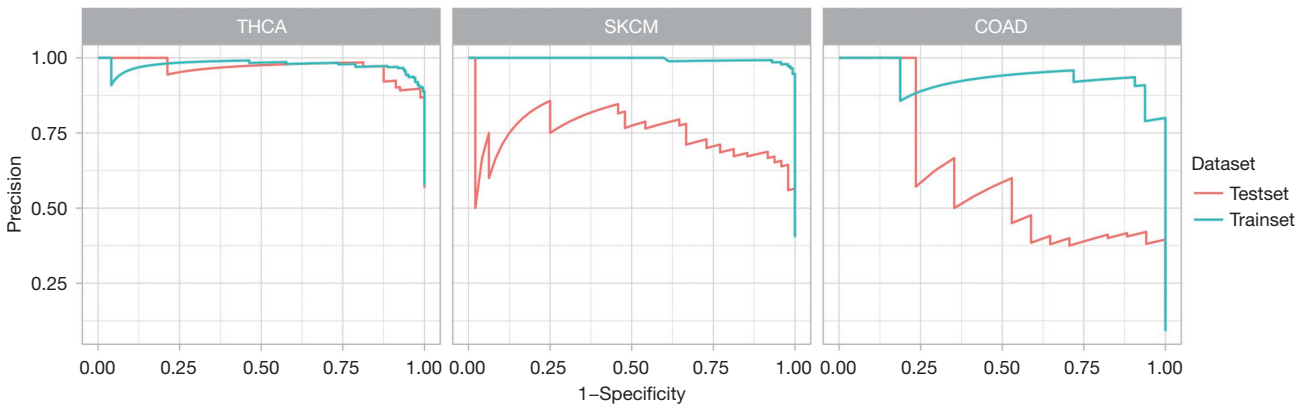

Figure 2 The precision-recall curve of *BRAF* V600E variant prediction. PR, precision-recall.

expression signature assay is recommended to select breast cancer patients who will benefit from receiving chemotherapy (19). These gene signature assay allow many breast cancer patients avoid adjuvant chemotherapy.

Although the purpose of this study is to investigate the possibility of *BRAF* V600E variants predictive model with mRNA gene expression data, we found that our model is biologically relevant because some genes that are biologically related to *BRAF* V600E variants had larger coefficient values. *ETV1* is the predictor with the largest positive coefficient value. *ETV1* is a member of the E twenty-six (ETS) family of transcription factors. ETS family genes make translocations with the *ewing sarcoma breakpoint region 1* (*EWSR1*) gene in Ewing's sarcoma/peripheral neuroectodermal tumor (PNET) spectrum and prostate cancer (20,21). The *BRAF* V600E variant is associated with ETV1 expression and brain metastasis in melanoma (22). ETS factors including *ETV1* are upregulated in papillary thyroid cancer with the *BRAF* V600E variant and showed synergistic effect with *TERT* promoter mutation (23). Nuclear factor κB (NF-κB) is activated by *BRAF* V600E variant and promotes invasiveness in thyroid cancer (24,25). The *BRAF* V600E variant induces NF-κB activation and increases melanoma cell survival in melanoma (26). Genes in the RAF-MEK-ERK signal transduction pathway, including *AKT serine/threonine kinase 2* (*AKT2*) and *NF1*, also showed larger coefficient values.

A previous study predicts *BRAF* variants using Affymetrix mRNA gene expression data with a support vector machine model from a panel of 63 melanoma cell lines with 0.794 ROCAUC (27). *BRAF* prediction studies using image data have been published. Ultrasound images with radiomics data were used for *BRAF* variant prediction with 0.651

© Translational Cancer Research. All rights reserved.

ROCAUC (28). A deep learning model from the histologic image was also used for *BRAF* variant prediction in melanoma with 0.83 ROCAUC (29).

Our prediction model has some limitations. Our model showed poor performance on the test set of other cancer types with a low prevalence of *BRAF* V600E variants. BRAF inhibitors have been approved in patients with lung non-small cell carcinoma and *BRAF* V600E variants. The lung non-small cell carcinoma shows a low prevalence of *BRAF* V600E variants. Therefore, our prediction model cannot be applied to lung non-small cell carcinoma patients or other cancer types with a low prevalence of *BRAF* V600E variants. Gene expression data are expensive and still complex for clinical use.

In conclusion, our penalized logistic regression model can predict *BRAF* V600E variant with good performance in thyroid carcinoma, cutaneous melanoma and colon adenocarcinoma.

Acknowledgments

Funding: This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2021R1I1A1A01043754).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tcr.amegroups.com/article/view/10.21037/tcr-22-883/rc

Peer Review File: Available at https://tcr.amegroups.com/article/view/10.21037/tcr-22-883/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tcr.amegroups.com/article/view/10.21037/tcr-22-883/coif). The Catholic University of Korea, Industry-Academic Cooperation Foundation has been filed a patent for "Modeling method for BRAF variant prediction model" (Application No. 10-2022-0014717). All authors are listed as inventors of the patent.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. Nature 2002;417:949-54.
2. Wan PT, Garnett MJ, Roe SM, et al. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell 2004;116:855-67.
3. Yao Z, Yaeger R, Rodrik-Outmezguine VS, et al. Tumours with class 3 BRAF mutants are sensitive to the inhibition of activated RAS. Nature 2017;548:234-8.
4. Chen W, Frankel WL. A practical guide to biomarkers for the evaluation of colorectal cancer. Mod Pathol 2019;32:1-15.
5. Thiel A, Heinonen M, Kantonen J, et al. BRAF mutation in sporadic colorectal cancer and Lynch syndrome. Virchows Arch 2013;463:613-21.
6. Ilie M, Long E, Hofman V, et al. Diagnostic value of immunohistochemistry for the detection of the BRAFV600E mutation in primary lung adenocarcinoma Caucasian patients. Ann Oncol 2013;24:742-8.
7. Fu G, Chazen RS, MacMillan C, et al. Development of a Molecular Assay for Detection and Quantification of the BRAF Variation in Residual Tissue From Thyroid Nodule Fine-Needle Aspiration Biopsy Specimens. JAMA Netw Open 2021;4:e2127243.
8. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 2011;364:2507-16.
9. Dummer R, Hauschild A, Santinami M, et al. Five-Year Analysis of Adjuvant Dabrafenib plus Trametinib in Stage III Melanoma. N Engl J Med 2020;383:1139-48.
10. Robert C, Karaszewska B, Schachter J, et al. Improved overall survival in melanoma with combined dabrafenib and trametinib. N Engl J Med 2015;372:30-9.
11. Planchard D, Smit EF, Groen HJM, et al. Dabrafenib

© Translational Cancer Research. All rights reserved.

plus trametinib in patients with previously untreated BRAFV600E-mutant metastatic non-small-cell lung cancer: an open-label, phase 2 trial. Lancet Oncol 2017;18:1307-16.
12. Sharma V, Vanidassane I. Encorafenib, Binimetinib, and Cetuximab in BRAF V600E-Mutated Colorectal Cancer. N Engl J Med 2020;382:876.
13. Kang J, Lee A, Lee YS. Prediction of PIK3CA mutations from cancer gene expression data. PLoS One 2020;15:e0241514.
14. Cancer Genome Atlas Research Network; Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113-20.
15. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401-4.
16. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1-22.
17. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
18. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2013;41:D377-86.
19. Giorgi Rossi P, Lebeau A, Canelo-Aybar C, et al. Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative. Br J Cancer 2021;124:1503-12.
20. Delattre O, Zucman J, Plougastel B, et al. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. Nature 1992;359:162-5.
21. Kedage V, Selvaraj N, Nicholas TR, et al. An Interaction with Ewing's Sarcoma Breakpoint Protein EWS Defines a Specific Oncogenic Mechanism of ETS Factors Rearranged in Prostate Cancer. Cell Rep 2016;17:1289-301.
22. Birner P, Berghoff AS, Dinhof C, et al. MAP kinase activity supported by BRAF (V600E) mutation rather than gene amplification is associated with ETV1 expression in melanoma brain metastases. Arch Dermatol Res 2014;306:873-84.
23. Song YS, Yoo SK, Kim HH, et al. Interaction of BRAF-induced ETS factors with mutant TERT promoter in papillary thyroid cancer. Endocr Relat Cancer 2019;26:629-41.
24. Bommarito A, Richiusa P, Carissimi E, et al. BRAFV600E mutation, TIMP-1 upregulation, and NF-κB activation: closing the loop on the papillary thyroid cancer trilogy. Endocr Relat Cancer 2011;18:669-85.
25. Palona I, Namba H, Mitsutake N, et al. BRAFV600E promotes invasiveness of thyroid cancer cells through nuclear factor kappaB activation. Endocrinology 2006;147:5699-707.
26. Liu J, Suresh Kumar KG, et al. Oncogenic BRAF regulates beta-Trcp expression and NF-kappaB activity in human melanoma cells. Oncogene 2007;26:1954-8.
27. Johansson P, Pavey S, Hayward N. Confirmation of a BRAF mutation-associated gene expression signature in melanoma. Pigment Cell Res 2007;20:216-21.
28. Kwon MR, Shin JH, Park H, et al. Radiomics Study of Thyroid Ultrasound for Predicting BRAF Mutation in Papillary Thyroid Carcinoma: Preliminary Results. AJNR Am J Neuroradiol 2020;41:700-5.
29. Kim RH, Nomikou S, Dawood Z, et al. A Deep Learning Approach for Rapid Mutational Screening in Melanoma. bioRxiv 2019:610311.

Cite this article as: Kang J, Lee J, Lee A, Lee YS. Prediction of *BRAF* V600E variant from cancer gene expression data. Transl Cancer Res 2022;11(11):4051-4056. doi: 10.21037/tcr-22-883

© Translational Cancer Research. All rights reserved.

Table S1 Study abbreviation and number of cases

Study abbreviation	Study name	Number of cases
LAML	Acute myeloid leukemia	173
ACC	Adrenocortical carcinoma	78
BLCA	Bladder urothelial carcinoma	426
LGG	Brain lower grade glioma	527
BRCA	Breast invasive carcinoma	1201
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	299
CHOL	Cholangiocarcinoma	45
COAD	Colon adenocarcinoma	475
ESCA	Esophageal carcinoma	192
GBM	Glioblastoma multiforme	201
HNSC	Head and neck squamous cell carcinoma	561
KICH	Kidney chromophobe	89
KIRC	Kidney renal clear cell carcinoma	581
KIRP	Kidney renal papillary cell carcinoma	316
LIHC	Liver hepatocellular carcinoma	418
LUAD	Lung adenocarcinoma	569
LUSC	Lung squamous cell carcinoma	534
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	48
MESO	Mesothelioma	87
OV	Ovarian serous cystadenocarcinoma	303
PAAD	Pancreatic adenocarcinoma	182
PCPG	Pheochromocytoma and paraganglioma	186
PRAD	Prostate adenocarcinoma	546
READ	Rectum adenocarcinoma	163
SARC	Sarcoma	259
SKCM	Cutaneous melanoma	469
STAD	Stomach adenocarcinoma	447
TGCT	Testicular germ cell tumors	138
THYM	Thymoma	121
THCA	Thyroid carcinoma	568
UCS	Uterine carcinosarcoma	57
UCEC	Uterine corpus endometrial carcinoma	550
UVM	Uveal melanoma	80

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 The coefficient values of genes are included in the final model

Predictor	Coefficient value
C2orf18	0.300852407
PTK2	0.242215773
FKBP5	0.231662872
DCN	0.223877955
FOSB	0.221057365
PTPRE	0.210010009
CAV1	0.207495453
CAMK2N1	0.196625042
ETV1	0.194502905
CHD3	0.190458013
EXT2	0.188940626
CD200	0.185461899
MMP9	0.18499858
RARG	0.184466599
SOX13	0.18263483
PDLIM4	0.182004634
HNRNPH2	0.169171491
NEK7	0.159721077
SULF2	0.158557026
JAZF1	0.155588028
DDRGK1	0.154423318
SORL1	0.153924494
LIMA1	0.149032802
UNC45A	0.14470599
ITIH5	0.143173781
ETNK1	0.141935617
GBAS	0.141174046
ALDH1A1	0.138736872
C1S	0.138443724
PIGQ	0.131092706
HSPA8	0.126238315
RTKN	0.125980941
VAPB	0.123207175
DHCR24	0.123195818
TTC3	0.118698045
OXCT1	0.118598478
TBC1D1	0.116858115
PCMT1	0.116309156
PPP2R5A	0.114435214
PSMD1	0.113489528
CRISPLD2	0.113425698
MRPS24	0.113278028
SPRED2	0.113043484
WIPI2	0.110292944
LY6E	0.109968565
NRIP1	0.10895298
CCT4	0.108176808
TXNIP	0.107679718
PPP1R9A	0.106847612
AKT2	0.105905525
ITGA2	0.10515576
MLPH	0.100495855
NFKB1	0.100488573
GDE1	0.099253853
NF1	0.09861373
C1QTNF1	0.098279261
MARCH6	0.09818392
RETSAT	0.097431727
FBXO34	0.094880043
NFATC4	0.094651548
C5orf62	0.092537745
ATF5	0.092266395
SAMD4A	0.091851926
C10orf58	0.091821872
LPHN1	0.091206168
ADCY6	0.089464002
CD55	0.089235938
CCND1	0.088506258

Table S2 (*continued*)

Table S2 (*continued*)

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
SFRP2	0.087677229
ZNF83	0.087546468
TPD52L1	0.085712481
LPIN2	0.084126614
GTF2IP1	0.083390487
S100A4	0.08288983
PXN	0.082238383
MYO1E	0.081908354
STK17B	0.080614539
SMG6	0.080131022
MYO5B	0.078605001
TRIB1	0.078588893
SERPING1	0.077370566
EIF3I	0.077357538
SQSTM1	0.077295597
GLOD4	0.07718981
NDUFA10	0.07701059
CCDC6	0.076946383
RNF144A	0.076929418
FYN	0.076595277
CCNG2	0.07645841
ELP2	0.076066824
MFSD1	0.074244674
STC1	0.072967766
PCGF2	0.072520212
PYGO2	0.071519324
SGK223	0.070781512
SNX9	0.068813508
RARRES2	0.068414047
PTPRU	0.068397909
TXNDC12	0.06819452
YWHAE	0.067861886
ATF7IP	0.067669038
NCS1	0.067328748

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
SUMO1	0.066229742
ATP1B1	0.065661254
TCEAL8	0.065435389
AGPAT3	0.065368532
EP400	0.064906137
NBL1	0.063939314
RTN3	0.063369972
NAP1L1	0.062870275
SMC5	0.061187484
PAFAH1B1	0.060463032
TCIRG1	0.060099389
FMOD	0.059363374
SFRS6	0.059319976
TRAK1	0.059221807
DNAJA4	0.058812443
TTC19	0.058609681
UPF1	0.058099781
BBX	0.057647272
SERPINF1	0.057243117
ADAM15	0.057133469
ITPRIPL2	0.056725428
UBB	0.056092814
PSIP1	0.055793463
AKAP2	0.055220868
ATP6V0E2	0.054865422
GGCT	0.05395249
LRRC8A	0.052489697
SMARCC2	0.052266561
RRM1	0.052092938
TMEM9	0.051126731
CD59	0.05071526
SEC11C	0.050690072
ECHDC1	0.050320647
FOSL2	0.050100302

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
TBL1XR1	0.050044447
POLE3	0.050011027
PRDX6	0.049746269
SCYL1	0.04970137
KIAA0284	0.04952663
TMEM115	0.049371505
ASAP2	0.048787608
EWSR1	0.048755607
RAP1GAP2	0.04844853
SPTBN1	0.048277591
TMED3	0.048243962
RELL1	0.047672887
C2orf28	0.046125098
NR2F2	0.045858868
SLC29A1	0.045743246
NUP50	0.045634565
MSN	0.045223306
NAP1L4	0.045084993
GTPBP2	0.044673267
PDGFRA	0.044646466
EFR3B	0.044415726
MCM5	0.044285624
C20orf3	0.043861156
NDUFB9	0.04328223
SKAP2	0.043165334
ETV4	0.042880029
SPOCK2	0.042529934
SEC61G	0.042423094
USP48	0.042373017
CNIH	0.041370304
ERRFI1	0.041155552
SH3GLB2	0.0411296
C1orf116	0.040877909
CTAGE5	0.04087093

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
HIPK2	0.040434807
NXN	0.040413567
UBA1	0.040376974
AK3	0.040244489
TIMP1	0.039665115
VPS53	0.038876352
GBP4	0.038021839
CHD2	0.037810724
NEAT1	0.037157293
MAP3K5	0.037037163
PIGY	0.036502849
RPS17	0.036204355
LSM4	0.03558709
TESC	0.035548457
PRDM1	0.035180784
FAM111A	0.034755251
CBR1	0.0346988
PACSIN2	0.034450275
RRP7A	0.033567084
HCP5	0.033013885
IFNGR2	0.032854811
AFAP1L2	0.032535422
GLG1	0.032162567
DLST	0.032072897
ZNF385A	0.03205089
CDK14	0.032021909
SPTBN2	0.031855716
PLXNB2	0.031240957
TM7SF3	0.031234756
PTP4A2	0.03115106
KIAA1797	0.030899719
NFKBIA	0.030786842
R3HDM2	0.030557894
MBP	0.029986966

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
SUDS3	0.029841016
SCD5	0.029156601
ANGPTL2	0.029023953
RBM15B	0.028549076
MAT2B	0.028383461
BTG1	0.028292035
SIPA1L1	0.027565644
NFIC	0.026843709
HBP1	0.026437415
TGFBR1	0.02628822
HES6	0.026094211
TGFB1	0.026006626
GOLGA2	0.025942691
GHDC	0.025767253
WTAP	0.025103408
GSDMD	0.02419907
GUSB	0.024128359
CYB5R3	0.023177287
MAGEF1	0.023152326
SYAP1	0.022540284
SLC38A5	0.022433217
MCM2	0.022196407
RTN4	0.021069441
MAFF	0.020695304
FNBP4	0.020363354
PSAT1	0.019667722
ARHGAP29	0.01859355
CHMP5	0.018396141
CD109	0.018370657
SAE1	0.018175123
PUM2	0.017946844
ANO6	0.017709732
IPO9	0.017702399
TNIP1	0.01745877

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
POLR2J3	0.017379979
USP53	0.0171379
GNG12	0.017052882
ARAP2	0.016630591
NPM1	0.016008692
ENO2	0.015019148
DCAKD	0.014914673
LOC729678	0.014413783
ENPP2	0.013956892
FBLN2	0.013531858
MAP3K11	0.013529472
RB1CC1	0.013283522
PFKFB2	0.013252072
EIF4B	0.013235598
UXS1	0.013038237
ATP6V1H	0.012893538
MAN2C1	0.012775558
RSL24D1	0.011258685
PLDN	0.011236057
SEPN1	0.010736932
SPTLC1	0.01025044
TNS1	0.010233013
ERGIC1	0.010142683
BTN3A2	0.009788192
VTI1B	0.009341172
GPBP1	0.009219948
LLGL1	0.008821434
BASP1	0.008584885
LYN	0.008461498
CHPF	0.008158752
ZNF655	0.008016446
TBC1D9B	0.007769478
GPRC5B	0.007600647
GCNT1	0.007451532

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
DEF8	0.006858562
ELOVL5	0.006762684
NFIX	0.006750099
MAP1S	0.006531209
GLTP	0.005668953
HSPA1B	0.005262617
SURF4	0.005211659
TMEM106B	0.005012405
HR	0.004917522
ATOH8	0.004234544
RXRA	0.003809947
BCL9	0.003490803
MTMR3	0.003296681
GOLPH3	0.002587358
RHEB	0.002064214
CYFIP1	0.001686961
WBSCR22	0.0012452
NDN	0.000731356
CBX1	0.000396352
TSKU	0.000355837
OLA1	9.93E−05
SEC11A	3.87E−05
LMBR1	−0.000257917
RALB	−0.000790209
CPNE1	−0.000876725
EIF1	−0.001085181
HNRNPA3	−0.00119446
SNX30	−0.001462602
MXD4	−0.001550677
DDX17	−0.001758997
MDC1	−0.001947092
SNN	−0.003317511
GPR116	−0.003493555
SMG7	−0.003980306

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
LAPTM4B	−0.004413844
PRDX2	−0.004693262
SLC22A18	−0.00470786
CYB5B	−0.00500726
PRKCZ	−0.005328095
IGSF3	−0.005849043
ERP29	−0.005982709
RPS18	−0.006011407
KIF5B	−0.006309243
AP2B1	−0.006360396
SPIRE1	−0.006815129
IRAK1	−0.00695366
LPCAT3	−0.00799526
CCNB1IP1	−0.00806399
RPS28	−0.008142391
CLIC4	−0.008265078
UNC13B	−0.009548794
ARHGAP21	−0.010145058
NRAS	−0.010344558
MYL12A	−0.010957177
CABC1	−0.011429478
CKB	−0.011443354
TFG	−0.0117612
BTF3	−0.011766326
ATHL1	−0.011953832
SMARCD2	−0.012226032
BAT2	−0.012854295
RPL9	−0.012874824
ATF6B	−0.013189555
COBLL1	−0.013254689
TGFBR2	−0.013435888
CCDC47	−0.013680079
GBP3	−0.013936651
VWA1	−0.014001641

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
PTP4A1	–0.01433804
TMEM8B	–0.014942868
HEBP1	–0.015682821
SOD3	–0.015693624
NR1D1	–0.015990961
FOXO1	–0.016082349
RAB32	–0.016466656
STXBP2	–0.016960306
TOP2B	–0.017124383
WFDC2	–0.017529991
AUTS2	–0.018320643
CDCA7L	–0.019269484
TRAF4	–0.019329704
EMD	–0.020966912
NT5E	–0.021106556
KIAA0114	–0.02136284
BRP44	–0.02167878
VWF	–0.02247171
TUBB6	–0.023098405
KIF13A	–0.023113696
ZNF185	–0.023218593
DBNDD2	–0.023256556
FXYD6	–0.023258567
RDBP	–0.023456228
VAMP3	–0.023808528
CEBPG	–0.024574337
C13orf23	–0.024595323
ST6GALNAC2	–0.024782139
PDIA3P	–0.025981466
GALNT2	–0.026150768
RPL22	–0.026237191
ANXA6	–0.026695156
UACA	–0.026806383
SH3GLB1	–0.02773001

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
RAF1	–0.027736061
WIPI1	–0.027825138
CLMN	–0.027856652
CRELD1	–0.0279438
GNB4	–0.028291925
CERK	–0.028322649
PLEKHG4	–0.028333032
NAGLU	–0.028679814
CHD7	–0.028934513
LASP1	–0.029179397
KAT2B	–0.03025785
PHF10	–0.03044976
TRIM26	–0.031257552
BAIAP2L1	–0.031640993
SLC6A6	–0.0318446
LAP3	–0.032119672
C14orf147	–0.032932233
RGS3	–0.033292326
MSL1	–0.033626781
CDH3	–0.033715721
ZC3H15	–0.033864524
NID1	–0.033926048
SELM	–0.034511258
OAT	–0.03490598
MXRA7	–0.034935448
MICALL1	–0.036654621
TGFBI	–0.036719332
CDC42EP4	–0.03689641
PPIC	–0.036911223
RRAS	–0.038003268
COASY	–0.038211104
WDR46	–0.038339238
FADS2	–0.038479982
XPO6	–0.040258807

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
CBS	–0.040436944
P4HA2	–0.040516737
TP53I11	–0.040963114
CPNE2	–0.041063973
PER3	–0.042268341
FLOT2	–0.04245193
MOGS	–0.043317534
MEPCE	–0.043440585
IK	–0.043984375
UBP1	–0.044100347
POLDIP2	–0.045119453
RAB20	–0.04546407
SYNPO	–0.046583199
SREBF1	–0.046636772
DDB1	–0.047282074
VASP	–0.047463586
ProSAPiP1	–0.048404754
CTR9	–0.048547135
KCTD10	–0.05021229
SH3BP4	–0.050842515
FERMT3	–0.051149121
SLC2A4RG	–0.051497112
LIMD1	–0.051648348
SEPT11	–0.051733494
TMEM87A	–0.051892977
RHOB	–0.053096316
HLA-F	–0.054014581
GALC	–0.054115181
IDH3G	–0.054639863
SLFN11	–0.054945433
CDC16	–0.056489981
GRAMD1A	–0.056871659
TAPBPL	–0.056949825
ACADM	–0.057122018

Table S2 (*continued*)

Table S2 (*continued*)

Predictor	Coefficient value
CISH	–0.057345796
CAPN2	–0.05738308
ADAMTS1	–0.057389454
ATP5G3	–0.057841547
STOML2	–0.057904155
GM2A	–0.058118052
C1orf9	–0.058571428
SEMA5A	–0.05870784
UBE2D2	–0.059462263
CALR	–0.06083565
SLC39A7	–0.060907386
CD97	–0.06165926
SLC11A2	–0.062257361
TIMM17B	–0.062804166
APOLD1	–0.063116755
FAM198B	–0.063272058
NME4	–0.063512399
GIT1	–0.064421909
ELF1	–0.065827339
PTK7	–0.066277901
NOMO1	–0.066736764
TRIM47	–0.068284061
PURB	–0.068406362
BAK1	–0.068698233
SCRN2	–0.069026122
CORO7	–0.069467822
PTEN	–0.069512183
SLC39A10	–0.070002984
PIK3R1	–0.070371231
THBS1	–0.070565654
JHDM1D	–0.07084167
PIGS	–0.071583087
PTP4A3	–0.07162628
PLCE1	–0.071852887

Table S2 (*continued*)

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S2 (*continued*)

Predictor	Coefficient value
COL4A1	–0.072707057
RAP2A	–0.072733308
PCIF1	–0.073744716
TP53INP1	–0.074066594
RAPH1	–0.074366176
DUS1L	–0.074762256
BMI1	–0.07703215
PDIA3	–0.078121799
COL17A1	–0.07873817
AEN	–0.079981193
SLC39A8	–0.081255312
NRBP1	–0.081584261
INTS10	–0.081838736
RMND5A	–0.082888045
HES1	–0.084381521
PPP1R3B	–0.086304276
PRDX1	–0.087152088
NNT	–0.089377209
CAPNS1	–0.090180193
PLIN2	–0.091672262
C11orf95	–0.093546732
EIF2AK4	–0.093723433
FLOT1	–0.09394406
IRS1	–0.094171528
DDX27	–0.094357701
PPT1	–0.094385778
EXTL3	–0.096233924
DAZAP2	–0.096820772
PRDX5	–0.098132458
PDZK1IP1	–0.099555814
THNSL2	–0.10147727
YARS	–0.10168223
LAMA1	–0.102704687
IGF2R	–0.102987157
CHST3	–0.104077894
GLYR1	–0.107735232
ALKBH7	–0.109885778
TMEM64	–0.110595607
PHC2	–0.111057907
GLCE	–0.111719993
PBRM1	–0.112828642
RGMB	–0.120761461
GPCPD1	–0.125270358
PPM1H	–0.125579953
DGAT2	–0.126890155
INPP5D	–0.127119869
INPPL1	–0.127660696
OXR1	–0.132848307
SLC39A11	–0.135392652
SPOCK1	–0.137435566
NKIRAS2	–0.137814176
SPRY4	–0.148713899
TMEM132A	–0.150682888
ERBB2IP	–0.156661288
ECM1	–0.157230177
EPDR1	–0.157640704
NEO1	–0.161370461
GSPT1	–0.16385327
RHOA	–0.165882881
EIF4EBP2	–0.174006524
SH3BGRL	–0.178800121
GALNT10	–0.178853545
CASK	–0.184428293
IER2	–0.191224485
TBC1D5	–0.21945529
PLCB4	–0.232494231
VAV3	–0.235682487
EGR1	–0.266768911
CRYZ	–0.27413205
IQSEC1	–0.280122156

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883

Table S3 Gene ontology test result

PANTHER Pathways	Homo sapiens - REFLIST (20,595)	Predictor genes (544)	Predictor genes (expected)	Predictor genes (over/under)	Predictor genes (fold enrichment)	Predictor genes (raw P value)	Predictor genes (FDR)
Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	39	7	1.03	+	6.8	1.61E-04	4.49E-03
PI3 kinase pathway (P00048)	57	9	1.51	+	5.98	4.65E-05	1.94E-03
Endothelin signaling pathway (P00019)	85	8	2.25	+	3.56	2.74E-03	5.09E-02
Integrin signalling pathway (P00034)	193	18	5.1	+	3.53	9.57E-06	7.99E-04
Apoptosis signaling pathway (P00006)	118	11	3.12	+	3.53	5.12E-04	1.22E-02
T cell activation (P00053)	86	8	2.27	+	3.52	2.93E-03	4.90E-02
CCKR signaling map (P06959)	172	15	4.54	+	3.3	1.05E-04	3.51E-03
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	255	20	6.74	+	2.97	3.26E-05	1.81E-03
Gonadotropin-releasing hormone receptor pathway (P06664)	231	15	6.1	+	2.46	1.87E-03	3.89E-02
Unclassified (UNCLASSIFIED)	17,977	435	474.85	–	0.92	2.64E-06	4.40E-04

FDR, false discovery rate; CCKR, cholecystokinin receptors; IGF, insulin-like growth factor.

© Translational Cancer Research. All rights reserved.
https://dx.doi.org/10.21037/tcr-22-883