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.

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, 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

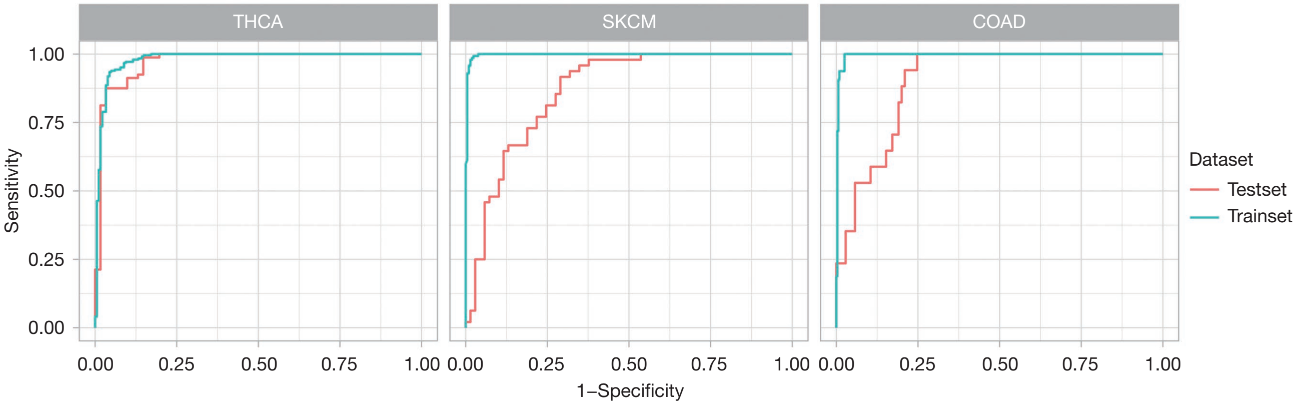

Figure 1 ROC curve of *BRAF* V600E variant prediction. THCA, thyroid carcinoma; SKCM, Cutaneous Melanoma; COAD, Colon adenocarcinoma; ROC, receiver operating characteristic.

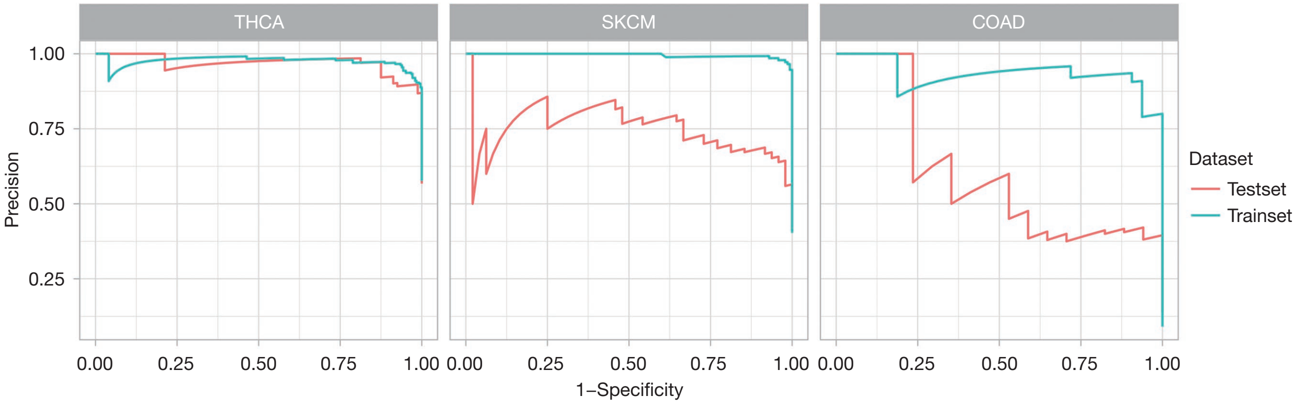

Figure 2 The precision-recall curve of *BRAF* V600E variant prediction. PR, precision-recall.

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

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).

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

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