


Identification of potential driver mutations in glioblastoma using machine learning

Medha Pandey, P. Anoosha, Dhanusha Yesudhas and M. Michael Gromiha 

Corresponding author: M. Michael Gromiha, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, 600036, India. Tel: +91 44 2257 4138; Fax: +91 44 2257 4102; E-mail: gromiha@iitm.ac.in

Abstract

Glioblastoma is a fast and aggressively growing tumor in the brain and spinal cord. Mutation of amino acid residues in targets proteins, which are involved in glioblastoma, alters the structure and function and may lead to disease. In this study, we collected a set of 9386 disease-causing (drivers) mutations based on the recurrence in patient samples and experimentally annotated as pathogenic and 8728 as neutral (passenger) mutations. We observed that **Arg** is highly preferred at the **mutant sites of drivers**, whereas **Met** and **Ile** showed preferences in **passengers**. Inspecting neighboring residues at the mutant sites revealed that the motifs YP, CP and GRH, are preferred in drivers, whereas SI, IQ and TVI are dominant in neutral. In addition, we have computed **other sequence-based features** such as conservation scores, Position Specific Scoring Matrices (PSSM) and physicochemical properties, and developed a machine learning-based method, GBMDriver (Glioblastoma Multiforme Drivers), for distinguishing **between driver and passenger mutations**. Our method showed an accuracy and AUC of 73.59% and 0.82, respectively, on 10-fold cross-validation and 81.99% and 0.87 in a blind set of 1809 mutants. The tool is available at <https://web.iitm.ac.in/bioinfo2/GBMDriver/index.html>. We envisage that the present method is helpful to prioritize driver mutations in glioblastoma and assist in identifying therapeutic targets.

Keywords: machine learning, cancer, variants, motifs, driver mutation, glioblastoma

Introduction

Glioblastoma (GBM) is one of the most common malignant tumors of the central nervous system. Understanding its molecular features and pathogenicity facilitates the progress of disease diagnosis, drug discovery and effective treatment. Several investigations on integrated analysis of multi-omics data revealed potential **drug targets** and shed light on deciphering immune signatures and critical biological pathways altered in GBM [1]. Earlier studies identified significantly mutated genes in GBM that include key alterations in receptor tyrosine kinase genes (EGFR and PDGFRA), tumor suppressors (TP53, NF1, PTEN and RB1) and other potential targets including PIK3CA and IDH1 [2–4]. Han et al. [4] showed that R132H/K/C mutations in IDH1 are associated with GBM, and Arg132 is crucial for the recognition of IDH1. Mutation at this position leads to decreased affinity to isocitrate and obstructs the formation of hydrogen bonds with it. As a result, mutant IDH1 cells establish neomorphic activity leading to cellular malignancies and oncogenesis. Recently, Vuong et al. [5] identified oncogenic mutations in the promoter region of telomerase reverse transcriptase (TERT) gene and suggested the same as a potential

therapeutic target. In addition, mutations in BRAF (V600E), IDH2 (R172G/M/K) [6] and TERT (C228T and C250T) are also identified as potential targets by various studies [7].

Despite recent advances in understanding mutational landscape of GBM, therapeutic options remain limited with median survival of <2 years from initial diagnosis. This is mainly because of the emerging rare oncogenic and resistant mutations during the treatment course of the patient. It is essential to evaluate the functional consequences of these variants to advance the therapeutic options for patients with GBM. However, functional validation to identify driver mutations from all the observed variants would be strenuous work. Hence, it is important to develop computational techniques to evaluate the mutations as drivers or passengers. Towards these directions, several computational methods have been developed to predict the impact of mutations in various diseases including cancer [8–11]. These methods are mainly developed with sequence (genomic and protein) [10], and structure-based [12] features. Recent work by Shi et al. [13] highlighted the importance of developing cancer-type specific methods to predict drivers due to high heterogeneity of mutational

Medha Pandey is currently doing PhD in computational biology at Indian Institute of Technology Madras. She has received her MTech Degree in Bioinformatics from Maulana Azad National Institute of Technology, India. Her research is focused on proteins and mutations associated to different cancer types.

P. Anoosha received her PhD in computational biology from Indian Institute of Technology Madras (IIT-M), India. Currently, she is working as a postdoctoral researcher in a precision cancer medicine lab at The Ohio State University (OSU), Columbus, USA. Her main research interests are to study tumor heterogeneity, evolution and clonal inference in different cancers and develop computational methods to assess heterogeneity and find distinct cancer cell populations in a tumor that drive drug resistance.

Dhanusha Yesudhas received her PhD in Medical science from Ajou University, South Korea and pursued her postdoctoral research fellow at the Indian Institute of Technology, Madras. Currently, she is working at NIH, USA. Her main research interests are scRNA-seq data analysis, protein–ligand and protein–nucleic acid interactions.

M. Michael Gromiha is working as a Professor at Indian Institute of Technology Madras, India. His main research interests are structural analysis, prediction, folding and stability of globular and membrane proteins, protein interactions and development of bioinformatics databases and tools. He has published >250 research articles, 50 reviews, 7 editorials and 2 books entitled 'Protein Bioinformatics: From Sequence to Function' by Elsevier/Academic Press and 'Protein Interactions: Computational methods, analysis and prediction' by World Scientific.

Received: July 7, 2022. **Revised:** September 13, 2022. **Accepted:** September 22, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

effects in different cancer types. Thus, we focused on developing a machine learning method that is specific to GBM, one of the aggressive cancer-types, to predict driver and passenger mutations.

In this work, we have analyzed a set of 9386 driver and 8728 passenger mutations in GBM, and identified preferred amino acid residues, and di- and tripeptide motifs at the binding sites of driver and passenger mutations. Furthermore, we have utilized the sequence-based features such as amino acid properties, preferences of di and tripeptides, conservation scores and Position Specific Scoring Matrices (PSSM), and developed a machine learning tool for discriminating driver and passenger mutations. Our method showed an accuracy and Area under the Receiver Operating Characterstics (ROC) Curve (AUC) of 81.99% and 0.87, respectively, in a test set of 1809 mutations, which is better than other existing methods in the literature. Furthermore, prediction performance in proteins with frequent and rare mutations, different functional classes of proteins as well as analysis on the relationship between change in stability and disease causing mutations will be discussed.

Materials and methods

Dataset

We collected the data for GBM causing mutations from COSMIC [14] and OncoKB [15] databases and selected the driver mutations based on their recurrence (≥ 2) in the samples and annotated as highly pathogenic, respectively. Neutral mutations were obtained from the dbSNP [16] and dbCPM [17] databases, which are based on the evidence from the functional experiments and high recurrence frequencies in healthy controls. These mutations are non-redundant and the final dataset contains 9386 driver and 8728 neutral mutations.

We have randomly chosen 10% of the data (1809) with appropriate representations in secondary structures (helix, strand and coil) for evaluating the performance of the method as 'test set'. In addition, most of the mutant sites (>95%) in the test set are completely different from the training set.

Calculation of sequence-based features

We have extracted a set of 487 sequence-based features based on six categories: (i) 268 amino acid properties, (ii) 7 groups of di- and tripeptide motifs, (iii) 3 PSSM scores, (iv) 2 predicted secondary structures and solvent accessibility, (v) 188 amino acid mutation matrices from AAindex [18] and (vi) 19 conservation scores [19]. A brief description of each category is discussed below:

Physicochemical properties

Amino acid properties include physicochemical, energetic and conformation parameters for the 20 amino acid residues, which are widely used in the literature for mutational studies [20] and from AAindex database [18].

The change in the property due to mutation is calculated as the difference between the property value in the wild-type and mutant residues using the following equation (Equation (1)):

$$\Delta P = P_{\text{mutant}} - P_{\text{wild}} \quad (1)$$

Where P_{mutant} and P_{wild} are the property of mutant and wild type residues, respectively.

Di- and tri-peptide motifs

We derived di- and tripeptide motifs using the information on neighboring residues towards N and C-termini of the mutant sites with and without gaps [21]. We calculated the odds ratio for all di- and tripeptides (Equation (2)) and used a cut-off of greater than 1.2 and 0.8 for motifs contribute to disease prone and neutral sites, respectively [21]. We encoded the odds ratio into three types: (i) for highly preferred (odds ratio ≥ 1.2), (ii) for not preferred (odds ratio ≤ 0.8) and (c) for moderate ($0.8 < \text{odds ratio} < 1.2$) to build the model.

$$\text{Odds ratio} = (N_{\text{dp}}/N_{\text{d}}) / (N_{\text{np}}/N_{\text{n}}) \quad (2)$$

Where N_{dp} and N_{np} denote the occurrence of a specific di- or tripeptide motif in disease prone and neutral sites, respectively. N_{d} and N_{n} are the respective number of disease prone and neutral sites in the dataset.

PSSM scores

We obtained the evolutionary information of amino acid sequences in terms of PSSM using PsiBLAST [22]. The query sequences were blasted for four iterations against the eukaryotic reference proteome from UniProt with an e-value cut off of 1×10^{-5} . We calculated three different PSSM scores: (a) average score of the mutant position, (b) difference between mutant and wild-type amino acid residue and (c) average over a window-length of 15 residues around the mutation position.

Predicted secondary structures and accessible surface area

We used NetsurfP [23] for predicting the secondary structure of each mutant site (helix, sheet or coil) and relative solvent accessibility (0–100%).

Amino acid mutation matrices

We collected 94 mutation matrices and pair-wise contact potential matrices from AAindex database [18] and substituted the value for each mutation. The difference of amino acid contact potential for a mutation is obtained by subtracting the contact potential value of N-/C-neighbor of mutation position to wild type residue from N-/C-neighbor to mutant residue [24].

Conservation scores

We calculated 19 different amino acid conservation scores using AACon server [19] at the mutation position, which include frequency-based Shenkin score [25], physicochemical property-based Zvelebil score [26], redundancy aware Valdar score [27] and specificity-sensitive SMERFS score [28]. In this study, conservation scores are calculated for each protein and mutant positions in the training and test sets are distinct. Hence, the conservation scores for the test set of mutants are different from the training set, which resulted no leakage of features in the test set.

Inclusion of neighboring residue information

The average property value for a window length of n residues from the disease prone site is computed using the equation:

$$\langle P \rangle = \sum_{k=-n}^{k+n} P(i) / (2n + 1) \quad (3)$$

where $\langle P \rangle$ is the average property, $i = k - n$, n is the window length, which varies from 1 to 10 (3–21 residues) and k is the mutant

position. We optimized the method with a window length of six (13 residues) based on the best performance [24].

Model development

We evaluated seven different machine learning algorithms including Support Vector Machine (SVM) [29], ADaptive Boosting (AdaBoost) [30], random forest (RF) [31], logistic regression [32], XGBoost [33], Multi-Layer Perceptron (MLP) [34] and K-nearest neighbor (KNN) [35] and one deep learning sequential model [36]. The definitions and categories of these methods are presented in [Supplementary Table S1](#). Based on their performances, we have developed three different models to classify driver and passenger mutations in GBM using adaptive boosting (AdaBoost), SVMs and xgBoost classifiers for the mutations in α -helical, β -strand and coil regions, respectively.

AdaBoost is a metaclassifier which utilizes decision stumps and used as an Ensemble method in machine learning [30]. This technique gives higher weightage to misclassified datapoints in the model development and these datapoints were given more importance in the next iteration until the error is reduced. Sample weight is calculated using the following equation:

$$W(x_i, y_i) = 1/N \quad (4)$$

where N is the total number of data points and W is the sample weight.

The performance of the decision stump (α) is calculated using the log error using Equation (5).

$$\alpha = 1/2 \log(1 - (x_i, y_i) / (x_i, y_i)) \quad (5)$$

Correctly and incorrectly classified instances have negative and positive α , and the weights will be updated using Equation (6).

$$W'(x_i, y_i) = W(x_i, y_i) * e^{\pm \alpha} \quad (6)$$

SVM is a robust classifier for generating optimal results and depends on the selection of extreme points/vectors with the maximum margin to form a hyperplane [30]. Considering a two-dimensional linearly separable data with the line function using the following equation:

$$Y = ax + b \quad (7)$$

If we rename x with x1 and y with x2, Equation (7) can be rewritten as:

$$ax1 - x2 + b = 0 \quad (8)$$

And hence the equation of hyperplane can be written as the following equation:

$$W.x + b = 0 \quad (9)$$

where x and W can be defined as $x(x1, x2)$ and $W(a, -1)$, respectively. Equation (9) represents the equation of hyperplane and the

following hypothesis can be used to make predictions:

$$h(x_i) = \begin{cases} +1, & \text{if } W.x + b \geq 0 \\ -1 & \text{if } W.x + b < 0 \end{cases}$$

XGBoost is an optimized distributed gradient boosting algorithm and highly efficient, where trees are built sequentially with reduction of errors [33]. It provides a parallel tree boosting to solve the problems in a fast and accurate way. This boosting ensemble technique consists of the following three steps:

Step 1: An initial model F_0 is defined to predict the target variable y. This model will be associated with a residual $(y - F_0)$.

Step 2: A new model h_1 is fit to the residuals from the step 1

1) Now, F_0 and h_1 are combined to give F_1 , the boosted version of F_0 . The mean squared error from F_1 will be lower than that from F_0 :

$$F_1(x) = F_0(x) + h_1(x) \quad (10)$$

1) To improve the performance of F_1 , we could model after the residuals of F_1 and create a new model F_2 :

$$F_2(x) = F_1(x) + h_2(x) \quad (11)$$

Step 3: Steps 1 and 2 will be repeated iteratively till the error is as minimum as possible.

The models were developed using Scikit-learn Python library and the hyperparameters were optimized to obtain the best prediction results.

Feature selection

The set of n-best features for discriminating driver and passenger mutations is selected using the package 'feature importance method (feat_importances.nlargest)' in scikit-learn [37]. This method (1) iterates the number of features from 1 to n until there is no improvement in the model performance and (2) selects the combination of features with the best performance based on sensitivity, specificity and accuracy, (3) provides scores for all features and (4) the most important feature has the highest score for discriminating driver and passenger mutations. We used the selected features to evaluate the model performance in the test set.

Model optimization

We trained the classifiers by optimizing their hyper parameters in scikit-learn and predict the output class as driver or passenger. In alpha class, AdaBoost uses real boosting algorithm SAMME to achieve the minimum error, learning rate as 1.0 and number of estimators as 50. In beta class, SVM uses RBC kernel with C as 10, whereas in coil class, XGBoost uses gamma as 2.58 and reg_lambda as 0.462. The details on the optimized hyper parameters of each class are mentioned in [Supplementary Table S2](#).

Model evaluation

We used 10-fold cross validation to evaluate the performance of the model. In this procedure, the dataset was divided into 10 equal parts, nine of which are used for training and the remaining part was used for test. This process was repeated 10 times so that all the data are used in the test at least once in the prediction model.

The following measures assess the performance of the model:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (12)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (13)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (15)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

where TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively and MCC is Mathews Correlation Coefficient. Disease-prone sites are considered as positive and neutral sites as negative class. Furthermore, the performance of the model was assessed with AUC.

Results and discussions

Distribution of driver and passenger mutations in GBM

We have computed the amino acid composition of driver/passenger mutations and normalized with that of respective residues in the whole dataset. The results presented in [Supplementary Figure S1\(a\)](#) showed that Arg is preferred in wild type, whereas Met and Ile showed high preference in passenger mutations (difference of >1). The composition of amino acid residues for the mutated sides revealed that His is preferred in drivers, whereas Val and Ser are preferred in passengers as seen in [Supplementary Figure S1\(b\)](#).

We calculated the odds ratio between driver and passenger mutations, and the results are presented in [Supplementary Table S3](#). We observed that the mutations C→F, E→V, K→M, L→Q, N→Y, R→C, R→I, W→C, W→L and Y→N are preferred in drivers with an odds ratio of ≥4. On the other hand, A→I, A→L, C→Q, D→Q, D→R, F→G, F→V, H→S, I→V, K→V, L→A, L→T, M→V, M→T, Q→D, Q→R, S→D, S→W, V→T and Y→V are preferred in passenger mutations with an odds ratio of ≤0.3. Some mutations such as A→C, A→H, etc. were not reported as drivers (in the dataset used in this study) for which the odds ratio are inconclusive and hence the table contains the odds ratio for 161 mutations.

We classified the amino acids into six groups based on their physicochemical properties: aliphatic (G, A, L, I and V), aromatic (F, Y and W), sulfur-containing (M and C), polar (N, Q, S, T and P), negatively charged (D and E) and positively charged (R, H and K). The preferred changes of driver mutations from one group to another are calculated using the following equation.

$$\Delta P_{D(i \rightarrow j)} = \left[\left(\frac{N_{D(i \rightarrow j)}}{N_{D(i \rightarrow j)}} \right) - \left(\frac{N_{D(j \rightarrow i)}}{N_{D(j \rightarrow i)}} \right) \right] * 100 \quad (17)$$

where $\Delta P_{D(i \rightarrow j)}$, $N_{D(i \rightarrow j)}$, $N_{(i \rightarrow j)}$, $N_{D(j \rightarrow i)}$ and $N_{(j \rightarrow i)}$ represent the percentage difference between driver mutations from group i to j and j to i, number of driver mutations from group i to j, the total number of mutations from group i to j, number of driver mutations from group j to i and the total number of mutations from group j to i, respectively. The results are presented in [Supplementary Table S4](#).

We found that mutations from polar to aromatic, sulfur to aromatic, positive charged to aromatic/sulfur and negative charge to aromatic/polar residues are preferred in driver mutations with a difference of >10%. In passenger mutations, the preferred amino acid groups are sulfur to aliphatic/polar and positive charge to aliphatic residues.

We further evaluated the importance of conservation scores for the preferred driver mutations by comparing the mutant sites for drivers and their respective passenger mutations. Our analysis shows that the driver and passenger mutations have the average conservation scores of 0.91 and 0.85, respectively.

Common and unique mutations of GBM with other cancer-types

We have compiled a set of driver mutations in 20 different cancer types based on the recurrence of the mutations across samples (mutation ≥2) and compared with GBM. The results presented in [Supplementary Table S5](#) showed that V→E, Y→H, R→Q, R→C, R→H, T→M, A→V and E→K mutations are common in driver mutations across cancer types. Among these mutations, V600E (BRAF), A289V, R108K, S768I (EGFR), R330H (ACADS) are present with high occurrence in most of the cancer types. The highly preferred mutations across different cancer types are influenced with the mutation of positive charged residues to positively charged (18 cancer types)/ sulfur containing (10 cancer types)/ polar residues (10 cancer types) as well as from aliphatic residues to negatively charged (15 cancer types)/ aliphatic (10 cancer types) residues. On the other hand, driver mutations including R132H (IDH1), K28M (H3F3A), Y390C (CHEK2) and A289T (EGFR) are present with high occurrence specifically in GBM patients.

Important motifs in driver and passenger mutations in GBM

We calculated the odds ratio for the preferred di- and tripeptides motifs around the mutation site using Equation (2) and the topmost five preferred motifs in each class are presented in [Table 1](#). We have considered six different dipeptides with and without gaps (X*M, M*X, X**M, M**X, XM and MX; X: any residue; *: gaps; M: mutation site) and a tripeptide with N- and C-terminal residues around the mutation site (XMX). We observed that Arg containing tripeptide motifs such as GRH, GRV, LRK and KRK as well as Cys containing dipeptide motifs as G*C, C**C, Y*C and CP are preferred in drivers with an odds ratio > 9. Interestingly, GRH, G*C, C*C, etc. are present only in driver mutations. We compared the preferred motifs in GBM with other cancer types [21] and observed that Arg is frequently present in the mutation and proximal sites of drivers. On the other hand, Ala, Ile, Met containing dipeptide and tripeptide motifs as I*V, Q*H, M**P, ME, TVI and KIL are preferred in passenger mutations with an odds ratio of <0.3. It is noteworthy that TVI, Q*H, etc. are present only in passenger mutations. We have further identified the important motifs using machine learning techniques and compared the results with the preferred motifs obtained with odds ratio. The data are presented in [Supplementary Table S6](#). Interestingly, 77.14 and 74.14% of motifs in driver and passengers, respectively, are commonly identified using machine learning methods and odds ratio.

Development of a prediction model for discriminating driver and passenger mutations

We have developed a machine learning model using a training dataset of 8449 drivers and 7856 passengers. It has been reported in the literature that the classification of mutations based on

Table 1. Preferred motifs in driver and passenger mutations

	XXM	X*M	M*X	X**M	M**X	XM	MX
Driver	GRH	G*C	C*C	E**C	G**M	YP	CP
	(11, 0)	(12, 0)	(9, 0)	(8, 0)	(8, 0)	(17, 1)	(15, 0)
	Inf	inf	inf	Inf	Inf	16.61	Inf
	GRV	Y*C	C*I	C**C	C**C	TC	WL
	(10, 0)	(14, 0)	(9, 0)	(10, 1)	(13, 1)	(14, 1)	(10, 1)
	Inf	inf	inf	9.64	12.57	14.65	9.59
	LRL	A*C	G*N	D**C	C**N	VC	PY
	(13, 1)	(9, 1)	(30, 5)	(9, 1)	(11, 1)	(12, 1)	(20, 3)
	12.71	8.59	5.82	8.68	10.64	11.72	6.39
	KRK	V*Y	C*V	V**C	C**P	FD	GV
Passenger	(12, 1)	(13, 2)	(11, 2)	(15, 2)	(13, 2)	(16, 2)	(48, 10)
	11.73	6.20	5.33	7.23	6.28	7.81	4.60
	TRL	M*R	R*W	V**R	P**Y	MA	RH
	(13, 2)	(40, 8)	(29, 6)	(114, 30)	(14, 3)	(22, 5)	(59, 13)
	6.35	4.77	4.68	3.66	4.51	4.29	4.35
	TVI	C*I	Q*H	F**K	M**D	SI	IQ
	(0, 5)	(1, 8)	(0, 12)	(1, 10)	(3, 16)	(9, 55)	(2, 21)
	0	0.11	0	0.09	0.18	0.15	0.09
	KIL	I*I	I*E	E**M	I**V	VM	NE
	(1, 7)	(4, 25)	(9, 41)	(4, 29)	(11, 43)	(2, 12)	(6, 34)
	0.13	0.15	0.21	0.13	0.24	0.17	0.16
	SAV	N*M	M*K	K**I	Q**E	PI	NR
	(2, 10)	(4, 19)	(5, 20)	(7, 35)	(10, 33)	(2, 29)	(4, 21)
	0.19	0.20	0.24	0.13	0.29	0.17	0.18
	EAA	S*I	I*I	N**I	M**S	AM	ID
	(2, 10)	(13, 41)	(8, 30)	(8, 30)	(10, 35)	(4, 20)	(6, 19)
	0.19	0.30	0.25	0.25	0.27	0.19	0.19
	GAP	T*K	I*V	S**I	M**P	TI	ME
	(3, 11)	(8, 23)	(10, 35)	(13, 42)	(3, 10)	(10, 33)	(6, 24)
	0.26	0.33	0.27	0.29	0.29	0.29	0.23

M: mutation site; *: gap and X represents any residue present in N- or C-termini. The number of drivers and passengers in each motif is shown in parenthesis followed by the odds ratio. Mutant position is shown in bold.

secondary structure enhanced the prediction performance [24]. Hence, we classified the mutations into helix, strand and coil and developed separate models. We used the feature selection technique, feature importance method (feat_importances.nlargest) using scikit-learn Python package [33] for all the three secondary structure classes and the selected features with the description are listed in [Supplementary Table S7](#). We found that some of the features such as tripeptide motif, PSSM scoring schemes and conservation scores are commonly selected in all the three models. Interestingly, these features are shown to be important for identifying disease causing mutations in membrane proteins [38, 39], binding sites in protein–protein [27] and protein–nucleic acid complexes [40], predicting the binding affinity upon mutations in protein–carbohydrate complexes [41] and so on.

Furthermore, we have compared the six embedded features in the final predicted models and the results showed that PSSM is uniformly selected in all the methods (helix, strand and coil). In addition, motifs, conservation scores and amino acid properties are identified to be more important than mutation matrices and neighboring residue information. Specifically, among the dipeptide and tripeptide motifs, XXM is selected in all the three models, whereas M**X is selected in two models, and X*M, M*X and XM in only one model.

We have evaluated the performance of the models using 10-fold cross validation and the results are presented in [Table 2](#). We noticed that the present method is capable of discriminating driver and passenger mutations with an overall accuracy, AUC and MCC of 73.59%, 0.82 and 0.47, respectively.

[Supplementary Figure S2](#) shows the area under the curve for each fold as well as the average in each secondary structure. Furthermore, the performance of the method has been examined with a test set of 1809 mutants, and our method could distinguish between driver and passenger mutations with a sensitivity, specificity, accuracy and AUC of 84.30%, 79.41%, 81.99% and 0.87, respectively. [Figure 1](#) shows the ROC curve for α -helical, β -strand and coil mutations with the test set.

Performance of the model with unique mutations in test dataset

We have examined the performance of the model using independent mutations in the test set by removing all the mutations, which are present at the same position in a given protein, and the results are presented in [Supplementary Table S8](#). We obtained an accuracy of 72.65% with a sensitivity, specificity and AUC of 76.62%, 68.07% and 0.80, respectively. The ROC curve for alpha, beta and coil mutations with the test set of unique mutations is presented in [Supplementary Figure S3](#).

Furthermore, we obtained the dataset used in the CHASM [42] for GBM-specific studies and identified 56 unique mutations from 16 proteins including TP53, EGFR, NF1, IDH1, PIK3CA and PTEN. We tested these mutations with our method and found that 51 mutations were predicted as driver mutations with a sensitivity of 91.07%. Similarly, we used the coding region-based training data obtained from CScape [43]. We mapped the coding region variants with the amino acid changes and obtained 249 mutations

Table 2. Model performance on training, 10-fold cross-validation and test set

Method		Mutation		Sensitivity	Specificity	Accuracy	MCC	AUC
		N _D	N _P					
Alpha (Adaboost)	Training	2452	2286	78.10	75.28	76.74	0.53	0.84
	10-fold CV			75.79	73.58	74.76	0.49	0.83
	Test	272	254	87.13	76.38	81.94	0.64	0.89
Beta (SVM)	Training	1158	919	88.61	84.55	86.81	0.73	0.94
	10-fold CV			75.71	64.23	70.36	0.40	0.82
	Test	128	102	88.61	84.55	86.81	0.73	0.87
Coil (XGBoost)	Training	4839	4651	77.15	77.30	77.22	0.54	0.85
	10-fold CV			76.69	74.55	75.65	0.51	0.82
	Test	537	516	77.15	77.30	77.22	0.54	0.85
Overall	Training	8449	7856	81.29	79.04	80.26	0.6	0.89
	10-fold CV			76.06	70.79	73.59	0.47	0.82
	Test	937	872	84.20	79.41	81.99	0.64	0.87

N_D and N_P represent the number of drivers and passengers, respectively. Results obtained with test set of data are shown in bold.

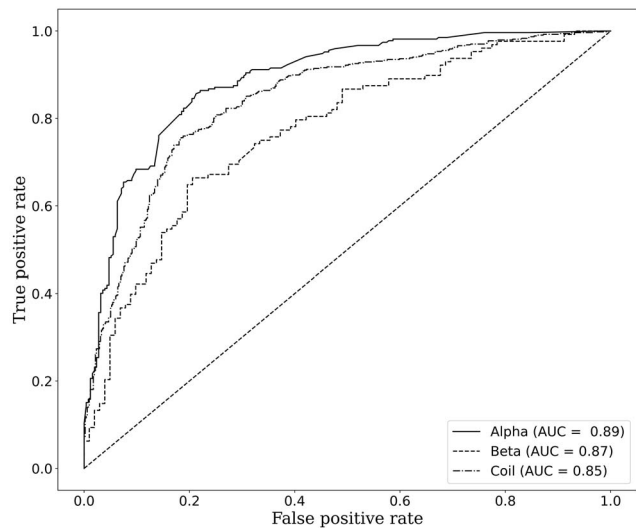


Figure 1. The receiver operator characteristics (ROC) curve for mutations in α -helical, β -strand and coil regions for the test dataset.

with 211 drivers and 38 neutral mutations associated to the 94 census genes involved in the GBM. We evaluated the performance of our method and found that out of 249 mutations, 228 were predicted correctly with an accuracy of 91.57% and sensitivity and specificity of 90.99% (TP: 192, FN: 19) and 94.73% (TN: 36, FP: 2), respectively.

Comparison with other algorithms based on network architectures

We compared the performance of the present method with other algorithms and the results are presented in Table 3. We optimized algorithms such as XGBoost, AdaBoost, Logistic regression (LR), SVM and RF and observed that they were able to achieve the accuracy in the range between 72 and 79%. We also checked the deep neural network architecture which was capable of classifying driver mutations with an accuracy of 78.35%. On the other hand, the present method, GBMDriver (Glioblastoma Multiforme Drivers) based on different classifiers is able to distinguish between driver and passenger mutations at an accuracy of 81.99% with 84.30% sensitivity and 79.41% specificity.

Comparison with other existing methods

We have compared the performance of our method with other generic algorithms such as SIFT [44], PROVEAN [45], Mutation-Taster [46], LRT [47], DEOGEN2 [48] and cancer-specific methods, VEST4 [42], FATHMM-MKL [49] and SusPect [45], CADD [46], MetaLR [47], CScape [43]. For a fair evaluation of the method, we compared the prediction performance of each of the mentioned methods (obtained using dbNSFP [50] database) with our method using the test dataset. The test dataset contains 937 drivers and 872 passengers. Among them, we could not get the results for 188 mutations by other methods and performed the comparison using the remaining set of 1621 (89.60%) mutations. The results are presented in Table 4. The performance of the method in each secondary structure is shown in Supplementary Table S9. We observed that all the methods have high specificity to capture neutral mutations, whereas the sensitivity was poor. Our method accurately predicted the driver and passenger mutations with a sensitivity, specificity and accuracy of 80.29, 75.85 and 78.11%, respectively.

Prediction performance in proteins with frequent and rare mutations

We evaluated the performance of the present method in proteins with frequently occurring mutations (top 50 proteins with 4030 mutations: 2587 drivers, 1443 passengers) associated with GBM. The major proteins include TP53, BRCA2, SYNE1, BRCA1, USH2A and ATM. We observed that our method predicted driver mutations with an accuracy, sensitivity and specificity of 79.82, 80.74, and 79.51%, respectively.

Furthermore, we examined the performance in proteins with rare mutations (mutations ≤ 5 in each protein; 6886 mutations in 3618 proteins; 5512 drivers and 1374 passengers) and the results are provided in Supplementary Table S10. Our method, GBMDriver showed a sensitivity, specificity and accuracy of 77.52, 72.60 and 76.43%, respectively.

Identification of driver mutations at a large scale

We have selected the topmost 10 proteins, which have the highest occurrence of driver mutations. We mutated each residue in these proteins into all other 19 amino acids, and classified them as drivers/passengers using GBMDriver. The mutations are ranked based on probability scores and a list of 25 mutations are presented in Supplementary Table S11. Specific driver mutations

Table 3. Comparison of different machine learning algorithms on the test set

Method	Sec Structure class	Sensitivity	Specificity	Accuracy	MCC	AUC
XGBoost	Alpha	87.50	75.20	81.56	0.63	0.88
	Beta	69.53	89.22	78.26	0.59	0.87
	Coil	77.15	77.30	77.22	0.54	0.85
	Overall	78.06	80.57	79.01	0.59	0.86
CHASM (VEST4)	Alpha	56.02	79.31	67.44	0.36	-
	Beta	54.46	82.98	67.48	0.38	-
	Coil	41.68	83.27	65.06	0.32	-
	Overall	47.58	83.14	65.57	0.33	-
AdaBoost	Alpha	87.13	76.38	81.94	0.64	0.89
	Beta	74.22	64.71	70	0.39	0.76
	Coil	80.63	70.35	75.59	0.51	0.83
	Overall	80.66	70.48	75.84	0.51	0.83
SVM	Alpha	81.99	74.80	78.52	0.57	0.87
	Beta	88.61	84.55	86.81	0.73	0.87
	Coil	74.12	69.57	71.89	0.44	0.81
	Overall	81.57	76.31	79.07	0.58	0.85
LR	Alpha	81.25	71.26	76.43	0.53	0.85
	Beta	74.22	58.82	67.39	0.33	0.69
	Coil	72.63	77.91	75.21	0.51	0.77
	Overall	76.03	69.33	73.01	0.46	0.77
MLP	Alpha	85.66	72.05	79.09	0.58	0.88
	Beta	71.09	52.94	63.04	0.24	0.7
	Coil	74.86	73.84	74.36	0.49	0.83
	Overall	77.20	66.28	72.16	0.44	0.80
RF	Alpha	81.99	73.23	77.76	0.55	0.87
	Beta	82.03	47.06	66.52	0.31	0.77
	Coil	71.69	61.24	66.57	0.33	0.75
	Overall	78.57	60.51	70.28	0.40	0.80
KNN	Alpha	51.1	85.83	67.87	0.39	0.76
	Beta	49.22	84.31	64.78	0.35	0.7
	Coil	50.84	77.71	64.01	0.3	0.68
	Overall	50.39	82.62	65.55	0.35	0.71
Deep neural network	Alpha	81.99	74.41	78.33	0.57	0.86
	Beta	74.22	87.25	80.00	0.61	0.87
	Coil	77.09	76.36	76.73	0.53	0.83
	Overall	77.77	79.34	78.35	0.57	0.85
GBMDriver (Present work)	Alpha	87.13	76.38	81.94	0.64	0.89
	Beta	88.61	84.55	86.81	0.73	0.87
	Coil	77.15	77.3	77.22	0.54	0.85
	Overall	84.30	79.41	81.99	0.64	0.87

SVM: support vector machine; XGB: XGBoost; MLP: multi-layered perceptron; KNN: K-nearest neighbor; DNN: deep neural network. Results obtained with overall data are shown in bold.

include A129D and K24N in TP53 (UniProt ID: P04637), N131K, W1490L and V4785A in TITIN (UniProt ID: Q8WZ42), V403I and Q396N in PTEN (UniProt ID: P60484), etc.

Downstream analysis of mutations in GBM based on functional influence

We segregated our test dataset based on the functional classes of proteins. We classified them into membrane proteins, enzymes and nucleic acid binding proteins. We evaluated the performance in the test data belonging to these three functional classes (1567 mutations: 766 drivers, 801 passengers) and our method could discriminate driver and passenger mutations with an accuracy of 80.87, 81.09 and 77.86% in membrane proteins, enzymes and nucleic acid binding proteins, respectively (Table 5). For the cross-validation data (13 878 mutations: 6749 drivers, 7129 passengers), our model distinguished driver and passenger mutations with an accuracy of 80.39, 74.87 and 76.25%, respectively, in membrane proteins, enzymes and nucleic acid binding proteins (Supplementary Table S12).

We further tested our method on the frequently mutated genes in GBM [42], which include EGFR, IDH1, NF1, TP53, etc. and most of the proteins are involved in the regulation of cell signaling required in cell proliferation and survival. Important driver mutations are R132C/H in IDH1, E542K/V, E545K/A and Q546K in PIK3CA, R445 in RB1. We evaluated the performance and the detailed results are reported in Table 6. We observed that the accuracy lies in the range of 67–100%.

Relationship between change in stability and driver mutations

We have mapped the predicted driver mutations in top nine frequently mutated genes in GBM with their structures and predicted the free energy change upon mutations using three different methods, mCSM [51], FoldX [52] and CUPSAT [53]. The results are presented in Supplementary Table S13. We observed that 93.42, 89.15 and 85.09% of predicted driver mutations are destabilized using mCSM, FoldX and CUPSAT, respectively.

Table 4. Comparison of the GBMDriver performance with existing methods

Classification method	Sensitivity	Specificity	Accuracy	MCC
General				
SIFT [44]	57.19	95.85	76.14	0.62
PROVEAN [45]	51.03	97.74	73.92	0.58
MutationTaster [46]	67.96	50.06	59.19	0.24
LRT [47]	58.33	69.69	64.06	0.22
DEOGEN2 [48]	27.27	92.76	59.69	0.24
MetaSVM [47]	25.89	88.54	56.73	0.18
M-CAP [49]	59.35	44.79	52.90	0.04
Cancer specific				
VEST4 (CHASM) [42]	48.37	83.14	65.41	0.33
FATHMM-MKL [54]	70.70	44.46	57.84	0.16
SuSPect [55]	21.77	93.84	57.09	0.22
CADD [56]	82.10	34.09	58.57	0.46
MetaLR [47]	26.86	83.38	54.68	0.12
CScape [43]	52.27	73.40	62.62	0.26
DANN [57]	94.92	13.71	55.12	0.15
LIST-S2 [58]	59.04	72.58	65.78	0.32
GBMDriver (Present work)	80.29	75.85	78.11	0.63

Table 5. Performance of the method for discriminating driver and passenger mutations belonging to three functional classes of proteins using test set of data

Functional class	Sec Str	Number of proteins	Mutations		Sensitivity (%)	Specificity (%)	Accuracy (%)
			N _D	N _P			
Membrane proteins	Helix	150	117	105	85.47	80.95	83.33
	Strand	83	55	51	89.0	86.27	87.73
	Coil	256	187	191	77.54	77.49	77.51
	Overall	412	359	347	81.89	79.83	80.87
Enzymes	Helix	50	32	51	90.63	78.43	83.13
	Strand	39	25	25	88.00	84.00	86.00
	Coil	109	82	113	87.80	72.57	78.97
	Overall	154	139	189	88.49	75.66	81.09
Nucleic acid binding	Helix	99	67	76	80.60	80.26	80.42
	Strand	45	34	22	88.23	86.00	87.50
	Coil	194	167	167	79.04	71.26	75.15
	Overall	282	268	265	80.59	75.09	77.86

N_D and N_P represent the number of drivers and passengers, respectively.

Table 6. Performance of the model with frequently mutated genes in GBM

Protein	Mutations		Sensitivity	Specificity	Accuracy
	Driver	Passenger			
TP53	219	4	98.17	75	97.76
EGFR	75	8	81.33	62.5	79.52
PTEN	65	10	92.31	100	93.33
PIK3CA	46	4	71.74	75	72
NF1	23	21	86.96	66.67	77.27
PIK3R1	19	3	78.95	100	81.82
PTPRD	11	1	63.64	100	66.67
IDH1	6	1	100	100	100
RB1	5	7	80	71.43	75

Development of a web-server

We have developed a web-server, GBMDriver for identifying driver mutations in GBM. It takes UniProt ID and mutation as input, and automatically identifies the secondary structure of the

mutant position using NetSurfP. Based on the secondary structure (helix/strand/coil) it selects the classifier (helix: AdaBoost; strand: SVM; coil: xgboost) to discriminate driver and passenger mutations. The output includes the secondary structure of the

mutation and the predicted effect of mutation as driver or passenger. It also has a feasibility to calculate the features for the mutation of interest in a given protein sequence. The web-server is accessible at <https://web.iitm.ac.in/bioinfo2/GBMDriver/>.

Conclusion

In this study, we have systematically analyzed the missense mutations associated with GBM and observed that GRH, G*C, C*C, YP and CP motifs are preferred in driver mutations, whereas TVI, C*I, I*V, M**P and M**D are dominant in passenger mutations. Utilizing various sequence-based features such as amino acid properties, mutation matrices, motifs, conservation scores and predicted secondary structure and solvent accessibility, we have developed a model for classifying driver and passenger mutations. Our method showed an average accuracy of 81.99% with a sensitivity, specificity and MCC of 84.20 and 79.41% and 0.64, respectively on a test set of 1809 mutants. Furthermore, it showed consistently a similar accuracy in both proteins with frequent and rare mutations. We also compared the performance with different machine learning algorithms, and existing generic and cancer-specific methods, and observed that the present method is better than other methods in the literature. We suggest that the present method is a valuable resource for identifying potential driver mutations in GBM and to develop drug design strategies.

Key Points

- Tripeptides GRH, GRV, LRK are preferred in driver mutations, whereas TVI, KIL and SAV are preferred in passengers
- Developed a classification model to distinguish between driver and passenger mutations in GBM
- Conservation scores, PSSM and tripeptide motifs are important for discrimination
- Identified the driver mutations with an accuracy and AUC of 81.99% and 0.87, respectively
- Developed a web server GBMDriver for identifying driver mutations associated to GBM

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgement

M.P. thanks MHRD and Department of Biotechnology, IIT Madras for computational facilities.

Funding

Department of Science and Technology, Government of India (No. DST/INT/SWD/P-05/2016).

References

- Jayaram S, Gupta MK, Raju R, et al. Multi-omics data integration and mapping of altered kinases to pathways reveal gonadotropin hormone signaling in glioblastoma. *Omics* 2016;**20**(12):736–46.
- Verhaak RG, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 2010;**17**(1):98–110.
- Chen Z, Hambardzumyan D. Immune microenvironment in glioblastoma subtypes. *Front Immunol* 2018;**9**:1004.
- Han S, Liu Y, Cai SJ, et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *Br J Cancer* 2020;**122**(11):1580–9.
- Vuong HG, Nguyen TQ, Ngo TN, et al. The interaction between TERT promoter mutation and MGMT promoter methylation on overall survival of glioma patients: a meta-analysis. *BMC Cancer* 2020;**20**(1):1–9.
- Yan H, Parsons DW, Jin G, et al. IDH1 and IDH2 mutations in gliomas. *N Engl J Med* 2009;**360**(8):765–73.
- Brennan CW, Verhaak RG, McKenna A, et al. The somatic genomic landscape of glioblastoma. *Cell* 2014;**157**(3):753.
- Wang D, Li J, Wang Y, et al. A comparison on predicting functional impact of genomic variants. *NAR Genom Bioinform* 2022;**4**(1):lqab122.
- Rogers MF, Gaunt TR, Campbell C. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics* 2020;**36**(12):3637–44.
- Rogers MF, Gaunt TR, Campbell C. Prediction of driver variants in the cancer genome via machine learning methodologies. *Brief Bioinform* 2021;**22**(4):bbaa250.
- Gnad F, Baucom A, Mukhyala K, et al. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 2013;**14**(S3):1–13.
- Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248–9.
- Shi X, Teng H, Shi L, et al. Comprehensive evaluation of computational methods for predicting cancer driver genes. *Brief Bioinform* 2022;**23**(2):bbab548.
- Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 2019;**47**(D1):D941–7.
- Chakravarty D, Gao J, Phillips S, et al. OncoKB: a precision oncology knowledge base. *Precis Oncol* 2017;**1**:1–16.
- Smigielski EM, Sirotkin K, Ward M, et al. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000;**28**(1):352–5.
- Yue Z, Zhao L, Xia J. dbCPM: a manually curated database for exploring the cancer passenger mutations. *Brief Bioinform* 2020;**21**(1):309–17.
- Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 2000;**28**(1):374–4.
- Valdar WS. Scoring residue conservation. *Proteins*, 2002;**48**:227–41.
- Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem* 1999;**82**(1):51–67.
- Pandey M, Gromiha MM. Predicting potential residues associated with lung cancer using deep neural network. *Mutat Res* 2021;**822**:111737.
- Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**(17):3389–402.
- Petersen B, Petersen TN, Andersen P, et al. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct Biol* 2009;**9**(1):1–10.

24. Anoshua P, Huang LT, Sakthivel R, et al. Discrimination of driver and passenger mutations in epidermal growth factor receptor in cancer. *Mutat Res* 2015;**780**:24–34.
25. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. *Proteins* 1991;**11**(4): 297–313.
26. Zvelebil MJ, Barton GJ, Taylor WR, et al. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* 1987;**195**(4):957–61.
27. Valdar WS, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 2001;**42**(1): 108–24.
28. Manning JR, Jefferson ER, Barton GJ. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinform* 2008;**9**(1):1–16.
29. Osisanwo FY, Akinsola JET, Awodele O, et al. Supervised machine learning algorithms: classification and comparison. *Int J Comput Trends Technol* 2017;**48**(3):128–38.
30. Collins M, Schapire RE, Singer Y. Logistic regression, AdaBoost and Bregman distances. *Mach Learn* 2002;**48**(1/3):253–85.
31. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
32. Hosmer DW, Jr DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd Edition, Wiley & Sons. John Hoboken, NJ, 2013.
33. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California (pp. 785–94). New York, USA: Association for Computing Machinery (ACM), 2016.
34. Taud H, Mas JF. Multilayer perceptron (MLP). In: *Geomatic Approaches for Modeling Land Change Scenarios*. Olmedo MTC, Paegelow M, Mas J-F, Escobar F (Eds.), Cham: Springer, 2018, 451–5.
35. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 1967;**13**(1):21–7.
36. Chollet F. *Deep Learning with Python* 2nd Edition. New York: Manning Publications, 2021.
37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikitlearn: machine learning in python. *J Machine Learning Res* 2011;**12**:2825–30.
38. Guo L, Wang S, Li M, et al. Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning. *BMC bioinformatics* 2019;**20**(S25):1–17.
39. Kulandaisamy A, Zaucha J, Frishman D, et al. MPTherm-pred: analysis and prediction of thermal stability changes upon mutations in transmembrane proteins. *J Mol Biol* 2021;**433**(11):166646.
40. Li G, Du X, Li X, et al. Prediction of DNA binding proteins using local features and long-term dependencies with primary sequences based on deep learning. *PeerJ* 2021;**9**:e11262.
41. Siva Shanmugam NR, Jino Blessy J, Veluraja K, et al. Prediction of protein–carbohydrate complex binding affinity using structural features. *Brief Bioinform* 2021;**22**(4):bbaa319.
42. Carter H, Chen S, Isik L, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;**69**(16):6660–7.
43. Rogers MF, Shihab HA, Gaunt TR, et al. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci Rep* 2017;**7**(1):1–10.
44. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**(13):3812–4.
45. Choi Y, Sims GE, Murphy S, et al. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012;**7**(10):e46688.
46. Schwarz JM, Cooper DN, Schuelke M, et al. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014;**11**(4):361–2.
47. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009;**19**(9):1553–61.
48. Raimondi D, Tanyalcin I, Ferte J, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res* 2017;**45**(W1): W201–6.
49. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**(12):1581–6.
50. Liu X, Wu C, Li C, et al. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat* 2016;**37**(3):235–41.
51. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2015;**30**:335–42.
52. Schymkowitz J, Borg J, Stricher F, et al. The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**(Web Server): W382–8.
53. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res* 2006;**34**:W239–42.
54. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**(10):1536–43.
55. Yates CM, Filippis I, Kelley LA, et al. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol* 2014;**426**(14):2692–701.
56. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310–5.
57. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**(5):761–3.
58. Malhis N, Jacobson M, Jones SJ, et al. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. *Nucleic Acids Res* 2020;**48**(W1):W154–61.