



OPEN

## Novel ratio-metric features enable the identification of new driver genes across cancer types

Malvika Sudhakar<sup>1,2,3</sup>, Raghunathan Rengaswamy<sup>2,3,4</sup>✉ & Karthik Raman<sup>1,2,3</sup>✉

An emergent area of cancer genomics is the identification of driver genes. Driver genes confer a selective growth advantage to the cell. While several driver genes have been discovered, many remain undiscovered, especially those mutated at a low frequency across samples. This study defines new features and builds a pan-cancer model, cTaG, to identify new driver genes. The features capture the functional impact of the mutations as well as their recurrence across samples, which helps build a model unbiased to genes with low frequency. The model classifies genes into the functional categories of driver genes, tumour suppressor genes (TSGs) and oncogenes (OGs), having distinct mutation type profiles. We overcome overfitting and show that certain mutation types, such as nonsense mutations, are more important for classification. Further, cTaG was employed to identify tissue-specific driver genes. Some known cancer driver genes predicted by cTaG as TSGs with high probability are ARID1A, TP53, and RB1. In addition to these known genes, potential driver genes predicted are CD36, ZNF750 and ARHGAP35 as TSGs and TAB3 as an oncogene. Overall, our approach surmounts the issue of low recall and bias towards genes with high mutation rates and predicts potential new driver genes for further experimental screening. cTaG is available at [https://github.com/RamanLab/cTaG\\_](https://github.com/RamanLab/cTaG_).

### Abbreviations

AUROC	Area under receiver operating characteristic
CGC	Cancer gene census
COSMIC	Catalogue of somatic mutations in cancer
FN	False negative
FP	False positive
GO	Gene ontology
HCC	Hepatocellular carcinoma
HiFI	High functional impact
Indels	Insertions and deletions
KS statistic	Kolmogorov–Smirnov statistic
LOF	Loss of function
LoFI	Low functional impact
MiFI	Mid functional impact
OG	Oncogene
TN	True negative
TP	True positive
TSG	Tumour suppressor gene

Cancer is one of the leading causes of morbidity globally, with more than 18.1 million cases reported in the year 2018<sup>1</sup>. A primary focus of cancer research has been the understanding of molecular mechanisms that govern tumorigenesis and the targets that can be used for treatment. Cancer cells are distinct because of their genomes, which give these cells the ability to divide and metastasise to other tissues in the body. It has been observed that mutations in some genes<sup>2,3</sup> confer the ability of oncogenesis to these cells. The term "driver" was coined to refer to mutations in the genome that pushed the cell to oncogenesis<sup>4</sup>. Of all the mutations present in a cancer cell, not all

<sup>1</sup>Department of Biotechnology, Bhupat Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India. <sup>2</sup>Centre for Integrative Biology and Systems mEdicine (IBSE), Indian Institute of Technology Madras, Chennai, India. <sup>3</sup>Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), Indian Institute of Technology Madras, Chennai, India. <sup>4</sup>Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai, India. ✉email: [raghur@iitm.ac.in](mailto:raghur@iitm.ac.in); [kraman@iitm.ac.in](mailto:kraman@iitm.ac.in)

are involved in giving a cellular advantage to the cell to divide uncontrollably. Driver mutations<sup>4,5</sup> are those that were advantageous for tumour development and metastasis during the clonal evolution<sup>6,7</sup>. On the other hand, *passenger* mutations<sup>4,5</sup> are mutations that are accumulated during normal cell division or due to high mutational rates in cancer cells, but their presence or absence does not affect the progression and establishment of tumours.

Driver genes are effectively those genes that harbour mutations that provide them with a selective advantage to divide and grow unchecked. These driver genes not only help the cells bypass the cell cycle checkpoints to divide in an uncontrolled fashion but also give added functionality, such as bypassing the immune system<sup>8,9</sup> and angiogenesis<sup>10,11</sup>, which lead to their persistence in the body. Certain cancers with well-understood mechanisms show that the presence of driver mutations is recurrent in most samples of a cancer type<sup>2</sup>, and driver genes accumulate a higher frequency of mutations. There are others that seem to have mutations that occur at a lower frequency. Note that "recurrence" refers to the repeated occurrence of a given mutation across samples, while "frequency" refers to the number of mutations in a given sample. Driver genes that contain a lower frequency of mutations are challenging to identify<sup>12</sup> because, most likely, these genes work in combination with other genes to confer a selective advantage to the cell, or mutations occur at functionally important locations.

Driver genes can be of two kinds depending on the role of the gene in a normal cell type. A tumour suppressor gene (TSG), as the name suggests, is the cell's defence mechanism from becoming a cancer cell. When such a gene loses its function due to, say, frameshift mutations or nonsense mutations, a selective growth advantage is conferred to the cell. Proto-oncogenes undergo gain of function mutations to become an oncogene (OG). Mutations in both TSGs and OGs tip the balance of a normal cell into becoming a cancer cell. While many TSGs and OGs have been discovered for different cancer types, most of them are highly potent and recurring in different patients. A pan-cancer model will help in identifying patterns that might be lost while studying a cohort or specific cancer type, owing to low sample sizes or mutation recurrence. A key aim of this study is to find low-frequency driver genes by classifying them into TSGs and OGs.

There are broadly two classes of methods for identifying driver genes based on mutational data. The first class of methods<sup>13–15</sup> rely on the rate of mutations in genes for a set of patients to identify driver genes. In these studies, the background mutation rate is estimated, and genes that show statistically different mutation rates are identified as driver genes. The rate of different types of mutations is used to calculate the background mutation rate<sup>14,15</sup>. The methods of identification differ in the statistical method used<sup>14</sup>. The rate of cell division and the length of the gene needs to be taken into account as the mutation rate may change depending on cell type and length and position of the genes<sup>15</sup>.

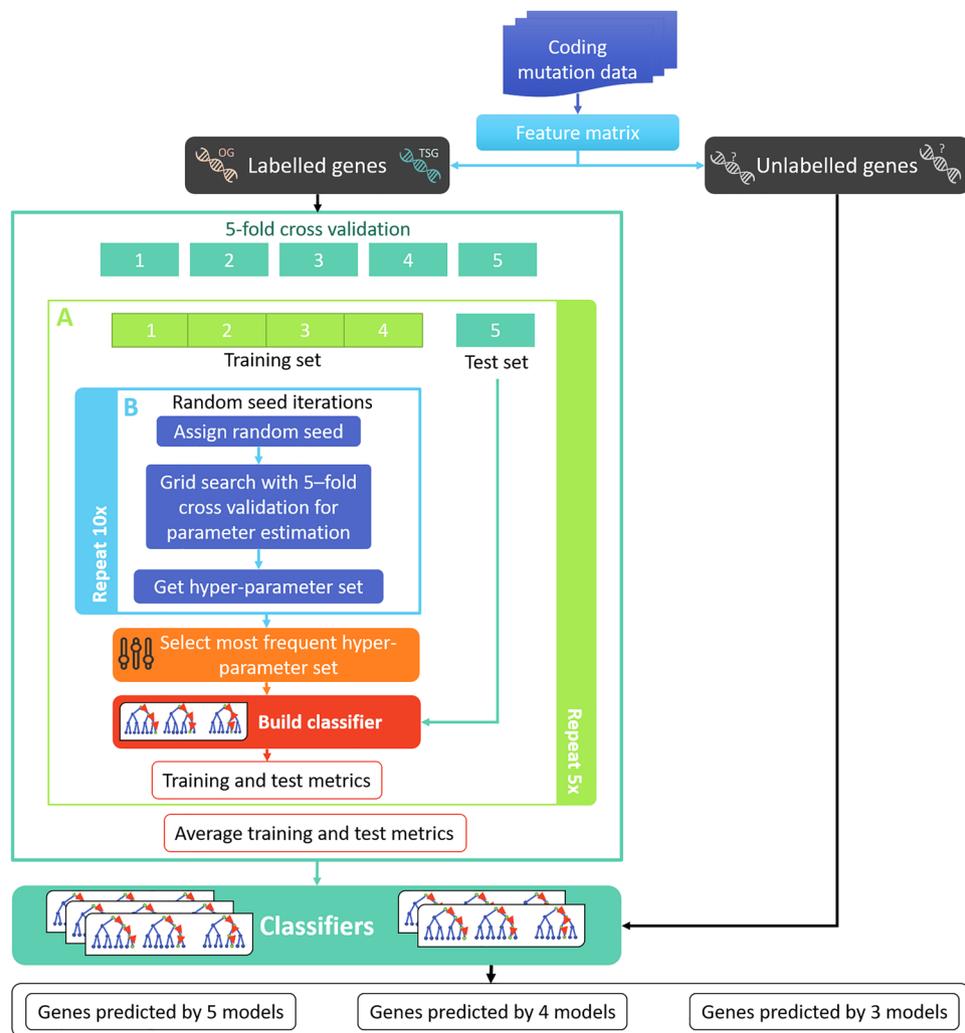
Among the different methods that exist for identifying driver genes, when validated using the Cancer Gene Census (CGC)<sup>16</sup>, it was observed that while the precision of identifying these genes was high, they had a very low recall<sup>12</sup>. Furthermore, genes identified through these approaches have high recurrence across different tumour samples. We now know that the rate of mutation is not sufficient for the identification of driver genes; instead, genes with low mutation rate can be driver genes if a mutation occurs at functionally important positions.

The second class of methods use a ratio-metric approach, where not only the repeated occurrence of mutations is taken into consideration but also the functional impact of the mutations. Ratio-metric algorithms<sup>17–19</sup> capture the proportion at which the different mutation types occur. The type of mutations and their ratios vary and are distinct for TSGs and OGs. For instance, TSGs are more likely to have indels (insertions and deletions), more specifically frameshift mutations, that lead to loss of function of the protein. On the other hand, OGs tend to accumulate missense mutations that confer the protein with a "gain of function"<sup>5,20</sup>. The recurrence of mutations is captured by the entropy features. These features are then used for differentiating between these two types of driver genes.

While these methods do capture some mutation patterns observed across samples, low recall shows that our understanding of the characteristics that define TSGs and OGs is far from complete. In this study, we define new features that calculate entropy and frequency of different mutation types along with other ratio-metric features. Our aim is to identify important features for TSGs and OGs that can help classify a given gene as a TSG or an OG. Since the ratio-metric approach is based on the type of mutations, and these differ for TSGs and OGs, genes were classified into two classes. Further, classification problems are prone to overfitting resulting in high classification scores in the training set, but the model can turn out to be unreliable for predictions using new data. We outline a method for estimating parameters for the given classification algorithm and avoid overfitting. We use the final model, cTaG (classify TSG and OG), to predict new driver genes by classifying a list of unlabelled genes; we validated our predictions by illustrating the presence of known TSGs and OGs among our predictions and through functional analysis of the predicted new genes. We calculated the mutation rates and compared our results with the widely used tool MutSigCV and show that cTaG is able to pick out many driver genes that have very low mutation rates. Further, we used a pan-cancer model to predict driver genes that were tissue-specific.

## Methods

**Outline of cTaG.** We build a model, cTaG, to classify genes as TSG or OG based on the mutation types observed in them. Our entire workflow is summarised in Fig. 1. First genes are labelled as TSGs or OGs for training the model. Next, we use random forest to train and build robust models. Model (i) identifies stable hyper-parameters to avoid overfitting and (ii) trains on features generated from mutation data. We use multiple random iterations (Fig. 1 Block 2) to identify stable hyper-parameters. We conduct fivefold cross-validation to avoid data bias and report average metrics (Fig. 1 Block A). Lastly, the genes not used for training were used for predicting new driver genes. Results from multiple classifiers are used to make the final predictions. The processed data and codes are available from GitHub: <https://github.com/RamanLab/cTaG>.



**Figure 1.** Methodology for identifying new driver genes. The figure presents an overview of the different steps involved in our study as described in “Outline of cTaG” section. The somatic mutation data is used to generate the feature matrix where the rows and columns represent genes and features respectively. The genes labelled as TSG or OG were used for building the models. Block A (light green frame) shows how classifier is built and is repeated five times for each fold (details in “Classification of genes” section). Block B (light blue frame) shows random iterations for estimation of hyper-parameters and is repeated 10 times to identify a set of stable hyper-parameters (details in “Tuning hyperparameters” section). The five models are then used to make predictions on unlabelled genes to identify the new driver genes.

**Mutation data.** We downloaded somatic mutation data from the Catalogue of Somatic Mutations in Cancer (COSMIC) (v79)<sup>21</sup>. These data were pre-processed to exclude hyper-mutated samples (samples containing more than 2000 mutations). Only known SNPs were retained if they were “confirmed somatic mutations”. The final processed data consist of 2,145,044 mutations from 20,667 samples across 37 primary tissues. COSMIC also contains transcript information, where different transcripts of a gene are saved as “gene\_transcript” and are handled as separate genes. Splice site mutations were identified as mutations at 1 or 2 bps after the end of the exon border or 1 or 2 bps before the start of the exon border. We used the popular tool Polyphen2<sup>22</sup> to predict the phenotypic impact of missense mutations. For a few mutation positions, Polyphen2 returned two scores. In such cases, the mean value was considered for the purpose of our analyses. The pre-processed COSMIC data used to build the feature matrix is available from zenodo.com: <https://doi.org/10.5281/zenodo.4153052>.

For tissue-specific analysis, the same data with the primary tissue annotations provided by COSMIC was used to extract mutations for a particular tissue. Tissues with  $\leq 1000$  samples were filtered out, which resulted in 10 primary tissues. The tissues used for the analysis are breast, central nervous system, cervix, endometrium, haematopoietic and lymphoid tissue, kidney, large intestine, liver, pancreas and prostate. The tissue-specific feature matrix was constructed for each of the primary tissues, which was used for the prediction of tissue-specific genes using the pan-cancer model.

TSGs and OGs for training and test were taken from the CGC<sup>16</sup> gene list. Only those genes that were labelled “TSG” or “OG” and not “Fusion” were used for this analysis. A total of 213 driver genes were used, of which 136 were TSGs and 77 were OGs. All genes and transcripts not used for model building were marked as unlabelled

Compound mutations = missense + complex + inframe + nonstop – LoFI
Loss of Function (LOF) = nonsense + frameshift
Damaging = HiFI + MiFI
Benign = silent + LoFI
$\text{ratio}(A/B)_g = \begin{cases} \frac{A_g}{B_g} & \text{if } B_g \neq 0 \\ 2 * \max(A) & \text{if } B_g = 0 \end{cases}$

**Table 1.** Definitions of mutation categories and the ratio of mutation categories. Along with mutation categories annotated by COSMIC, we define additional categories which combine multiple mutation types. These categories together with the 11 mutation categories defined in “[Ratio-metric features](#)” section are used to define ratio-metric features. The ratio of two mutation types A and B is defined for a given gene, where A and B are any two mutation categories.

Previously defined in the literature (18 features)	Silent/kb, Total Missense, Total Splicing, Total LOF, Missense/kb, LOF/kb, <b>LOF/Silent</b> , <b>Splicing/Silent</b> , <b>Missense/Silent</b> , <b>LOF/Benign</b> , <b>Splicing/Benign</b> , <b>Missense/Benign</b> , average Polyphen2 score, LOF/Total, Missense/Total, Splicing/Total, <b>LOF/Missense</b> , Missense entropy
Defined in this paper (19 features)	<b>HiFI/LoFI</b> , <b>HiFI/Benign</b> , MiFI/kb, Nonstop/kb, Inframe/kb, Complex/kb, <b>Compound/Benign</b> , <b>Compound/kB</b> , Damaging/kb, <b>Damaging/Benign</b> , <b>Damaging/LoFI</b> , High Missense frequency, Frameshift entropy, High Frameshift frequency, Splicing entropy, High Splicing frequency, Nonsense entropy, High Nonsense frequency, Total MiFI

**Table 2.** The features used in this study for classification. The features in bold are ratio-metric features. The different mutation types and method for calculating the features are defined in “[Mutation data](#)” section.

and used for the prediction of new driver genes. The TSG:OG ratio was maintained during all cross-validation steps and in both training and test sets.

**Ratio-metric features.** Mutations were divided into 11 different categories<sup>17,22</sup>: silent, missense, splicing, High Functional Impact (HiFI), Mid Functional Impact (MiFI), Low Functional Impact (LoFI), nonsense, frameshift, in-frame, nonstop or complex (annotations from COSMIC). Not all missense mutations are equally deleterious—labelling them into HiFI, MiFI and LoFI categories helps differentiate genes that have a large number of mutations with low impact from genes that have relatively fewer mutations but with larger functional impact. We use PolyPhen2 scores to categorise mutations as HiFI ( $\geq 0.85$ ), LoFI ( $\leq 0.15$ ) and MiFI (between 0.15 and 0.85) to differentiate between high confidence pathogenic mutation predictions.

Additionally, other mutation categories were defined, which clubbed multiple mutations into one, such as ‘compound’ and ‘damaging’. Compound mutations are included because mutations types such as in-frame, nonsense and complex occur at a lower rate than single nucleotide missense mutations, which might lead to patterns and the impact of these mutations being masked. Since the functional impact is similar to missense mutations, combining similar mutation types might help capture information of these less frequently observed mutation types. Loss of function (LOF) mutations introduce significant changes in proteins, disrupting function. Damaging mutations are the sum of HiFI and MiFI mutations; these capture impact of multiple MiFI and sparse HiFI mutations. Many features compute a ratio of mutation types, as outlined in Table 1. We defined 37 features in all, with 18 being similar to those defined as Davoli et al.<sup>17</sup> (Table 2).

**Entropy and frequency features.** Entropy and frequency features were defined for four mutation types. A mutation ( $M_i$ ) in a given gene  $i$  is represented by its location. For missense mutations,  $M_i$  is represented as a tuple ( $loc$ ,  $wt$ ,  $mt$ ) where  $loc$  is the location of the mutation,  $wt$  is the wild type nucleotide, and  $mt$  is the mutated nucleotide. If  $k$  unique mutations are present in a gene,  $f_i$  gives the frequency for each of the mutations.

$$f_i = \frac{n_M}{n}$$

where  $n_M$  is the number of occurrences of mutation  $M$  and  $n$  is the number of mutations in gene  $i$ .

$$S = \sum_{i=1}^k f_i \log f_i$$

$$\text{Entropy} = \log k - S$$

**Classification of genes.** Different machine learning algorithms such as random forest, support vector machines and logistic regression were used, among which random forest gave the highest accuracy. Random

forest was used for building a robust model and classifying TSGs and OGs. We used fivefold cross-validation to split data into training to test set ratio of 8:2; where each fold acts as a test set. We used the implementation of Random forest from the Python package Sci-Kit Learn<sup>23</sup>. For each cross-validation set (Fig. 1 Block A), we tuned the parameters using a fivefold cross-validation grid search along with multiple random iterations of random seed (“[Tuning hyperparameters](#)” section). The classification was re-run using the given parameters, and features were ranked. The model was used to predict the classification of test set genes. All models built during cross-validation was used to make predictions. The consensus of the top genes across the models was used to make the final predictions.

**Tuning hyperparameters.** A grid search with fivefold cross-validation was done for multiple different random seeds (Fig. 1 Block B). The parameters tuned are *n\_estimator* (from 5 to 40), *max\_features* ('sqrt' or 'log2'), *max\_depth* (2–4) and *criterion* ('gini' or 'entropy'). The number of maximum features each decision tree considers is given by the parameter *max\_features*, which can be calculated in two ways: either the square root or  $\log_2$  of the total number of features. Optimum parameters were selected by first estimating parameter '*n\_estimator*' and using it to estimate other parameters. Recurrence of '*n\_estimator*' across different random seeds was counted, and the maximum count was considered as the best '*n\_estimator*' to be given to the model. For the random iterations with the given best '*n\_estimator*', the hyper-parameter set giving maximum accuracy was chosen. If multiple estimators were chosen, maximum accuracy during cross-validation was used to select estimator and the corresponding hyper-parameter set.

**Feature comparison and ranking.** All features defined were used for classification and ranked depending on their contribution to the random forest model. The average rank was calculated across the five models, one for each validation set. The list of all features is given in Table 2.

**Estimating the robustness of the classifier.** Our initial results showed variation in classification depending on the random seed that was selected for classifying, even though cross-validation was used while estimating parameters. We used balanced bagging classifier to take into consideration the class imbalance and estimated parameters using cross-validation, which is the standard method. Poor results for this model led us to estimate hyper-parameters differently. To avoid this variation, classification and parameter selection were done for multiple random seeds (Fig. 1 Block B). To estimate the effect of the number of random iterations on parameter estimation, the classifier was built on a varying number of iterations of random seeds (10, 20, 40, 80, 160, 320). The stability of the hyper-parameters selected was analysed based on the variation in the accuracy of the test dataset.

**Identification and functional analysis of new TSGs and OGs.** We used the model built on the combined set of 37 features to classify unlabelled genes and transcripts into TSGs and OGs. In total, 26,866 genes/transcripts were classified as TSGs or OGs and ranked using their probabilities for each class. The feature matrix used for classification is available on GitHub. The genes given for classification contains different transcripts of the same gene symbol as different genes. In all, the gene list contained 18,951 unique gene symbols. Genes were labelled TSG and OG depending on their presence in the top 5 percentile and consensus across models built during cross-validation. Since not all genes are necessarily TSGs or OGs, genes that didn't fulfil these criteria remained unlabelled. New TSG and OG gene list predicted by greater than four models were further used for functional analysis to find the major pathways and gene ontologies these genes are enriched for. Functional analysis was carried out using DAVID<sup>24,25</sup> for both the genes above the threshold as well as training set genes, and the results were compared.

Further, the pan-cancer classifier was used to predict genes in different cancer types based on the primary tissue where the tumour is formed. The data were filtered based on primary tissue, and the feature matrix was generated for tissues with > 1000 samples. The data was then standardised and run using cTaG (pan-cancer model) described earlier.

We compared and calculated mutation rates using MutSigCV. Since the ground truth is not known for these predicted genes, we compared the genes used for training and calculated the recall of these genes. Since MutSigCV does not classify genes as TSG or OG, the classes considered were Driver and Passenger. Further, we were interested in looking at the mutation rate distribution across the genes predicted. Since the distribution of mutation rates is unknown, we compared the similarity of the distribution of the predicted genes with the genes used for training (Kolmogorov–Smirnov statistic). Similarly, the similarity was compared for genes predicted by MutSigCV.

We compared the predicted driver genes from other methods, TUSON, 20/20+ and DriverNet<sup>26</sup>, to ours. Other tools such as DawnRank<sup>27</sup> and Prodigy<sup>28</sup> were not included in the comparison as they are personalized driver gene predictors. We compared the methods based on the precision in predicting curated genes identified in Bailey et al. *undefined*<sup>29</sup> and CGC. The precision of a method is calculated as total number of predicted known driver genes by total number of genes predicted as driver by the method (Precision = TP/(TP + FP)). A total of 165 pan-cancer genes were identified by Bailey et al. using multiple computational methods, which were extended to include 35 manually curated “rescued” genes. We excluded genes used for training the models leaving 92 and 25 genes in the pan-cancer and “rescued” gene list. We considered top-ranking genes reported with *q*-value < 0.05 for TUSON and 2020+ for comparison. The union of all genes reported by DriverNet on different datasets, was used. Genes with a *p*-value of < 0.05 were considered in the final list.

			Accuracy	F1 score	Precision	Recall
cTaG	Training set	OG	0.86 ± 0.04	0.77 ± 0.07	<b>0.93 ± 0.04</b>	0.67 ± 0.09
		TSG		<b>0.90 ± 0.03</b>	0.84 ± 0.04	<b>0.97 ± 0.01</b>
	Test set	OG	0.76 ± 0.03	0.59 ± 0.10	<b>0.79 ± 0.12</b>	0.50 ± 0.19
		TSG		<b>0.83 ± 0.02</b>	0.77 ± 0.07	<b>0.91 ± 0.07</b>
BalancedBagging	Training set	OG	0.93 ± 0.05	0.92 ± 0.06	0.86 ± 0.09	0.99 ± 0.01
		TSG		0.94 ± 0.04	1.00 ± 0.01	0.90 ± 0.07
	Test set	OG	0.69 ± 0.06	0.64 ± 0.06	0.56 ± 0.07	0.75 ± 0.06
		TSG		0.73 ± 0.06	0.82 ± 0.04	0.65 ± 0.09

**Table 3.** Classification metrics for training and test set. Numbers in bold indicate best performances for each metric between TSG and OG. The metrics are standard, and are defined as follows (T stands for True, F for false, P for positives and N for negatives): Accuracy =  $(TP + TN)/(TP + FP + TN + FN)$ ; Precision =  $TP/(TP + FP)$ ; Recall =  $TP/(TP + FN)$ ; F1 score is the harmonic mean of Precision and Recall.

## Results

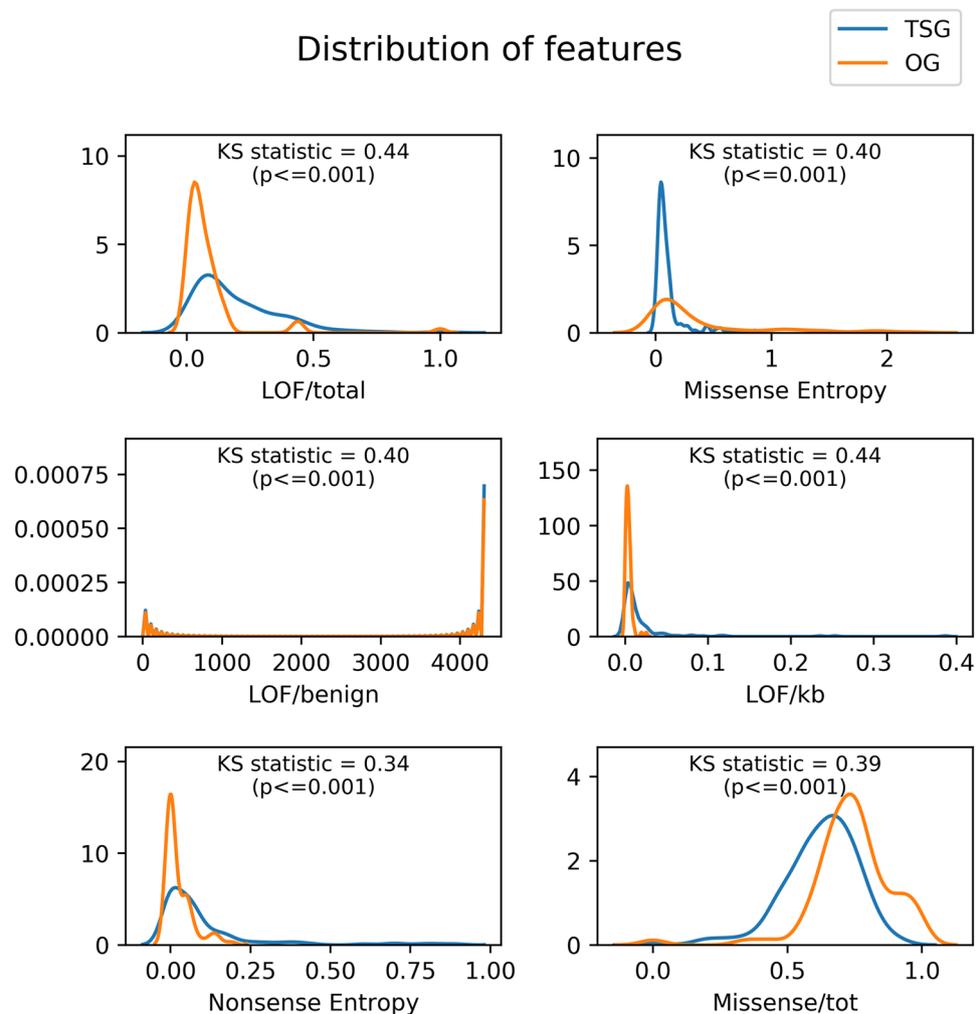
We define novel features and a method to estimate parameters and build a classifier using pan-cancer data to predict TSGs and OGs. The classifier is further used to predict labels for unlabelled genes at pan-cancer and tissue-specific levels, which are analysed for functional enrichment.

**Novel features used for classification of TSGs and OGs.** We trained multiple random forest models using a subset (80%) of 136 TSGs and 76 OGs for each fold of the cross-validation. We performed fivefold cross-validation while estimating hyper-parameters for the model, followed by multiple random iterations to estimate stable hyper-parameters and avoid overfitting (as defined in “Methods”). The final model was built using the hyper-parameters so identified (Supplementary Table S1). It is important to carefully consider overfitting as the initial training set is not very large. The accuracy of the test set reduces compared to the training set, but this difference is not substantial. We note that TSGs can be predicted with higher accuracy than OGs; it is probable that the features are biased at capturing information regarding TSGs better than OGs. Across the multiple models, an average accuracy of  $0.76 \pm 0.03$  was achieved (Table 3). These models were further used for the identification of new genes as well as tissue-specific analyses. Our model (cTaG) presents a significant improvement in recall for TSGs. For OGs, the recall is similar to those observed in other tools. Nevertheless, an average recall of driver genes (comprising both classes) shows an improvement over the tools reported earlier<sup>12</sup>.

To identify features important for the classification of TSGs and OGs, we calculated the average rank of each feature across all models. We observe that the top-ranking features contain LOF and missense mutations (Supplementary Table S2). The new features that replace old features in the top 18 ranks are Nonsense entropy, High missense frequency, Compound/benign, High Frameshift Frequency, Damaging/kb, Compound/kB, Damaging/LoFI and HiFI/benign. Further, we used the training set genes to compare the distribution of feature values in TSG and OGs, and observed that our top-ranking features show the highest differences between the two distributions (Fig. 2). The model built on only the old features performs marginally lower with an accuracy of 0.75 than the model using all features, but the difference is not statistically significant. While it is common knowledge that LOF mutations accumulate in TSG and recurrent missense mutations in OGs, we formally show that the feature distribution is different for these two functional classes.

**Iterative hyper-parameter estimation avoids overfitting.** Initial analysis using support vector machines (SVM), logistic regression, and random forest showed high accuracy for random forests (Supplementary Table S3). For many trees, random forest (95.3%) gave a higher accuracy score for training sets comparable to 91.9% achieved by Davoli et al.<sup>17</sup>. However, these showed very low accuracy for the test set (Supplementary Table S3), indicating overfitting. Additionally, we observed that changing the random seed showed substantial variation in results. This variation is unexpected and could perhaps stem from non-optimum parameters used for classification or the small size of the data. To avoid this variation, we selected random forest for its best performance and re-estimated the parameters, *n\_estimator*, *max\_features*, *max\_depth* and *criterion*. Changing the *n\_estimator* had a major effect on classification, and a simple grid search with cross-validation did not help in removing overfitting, as seen in our results for balanced bagging (Table 3). Comparison of metrics of our final model with balanced bagging, a similar algorithm that uses decision trees and handles unbalanced data, showed our procedure helps avoid overfitting.

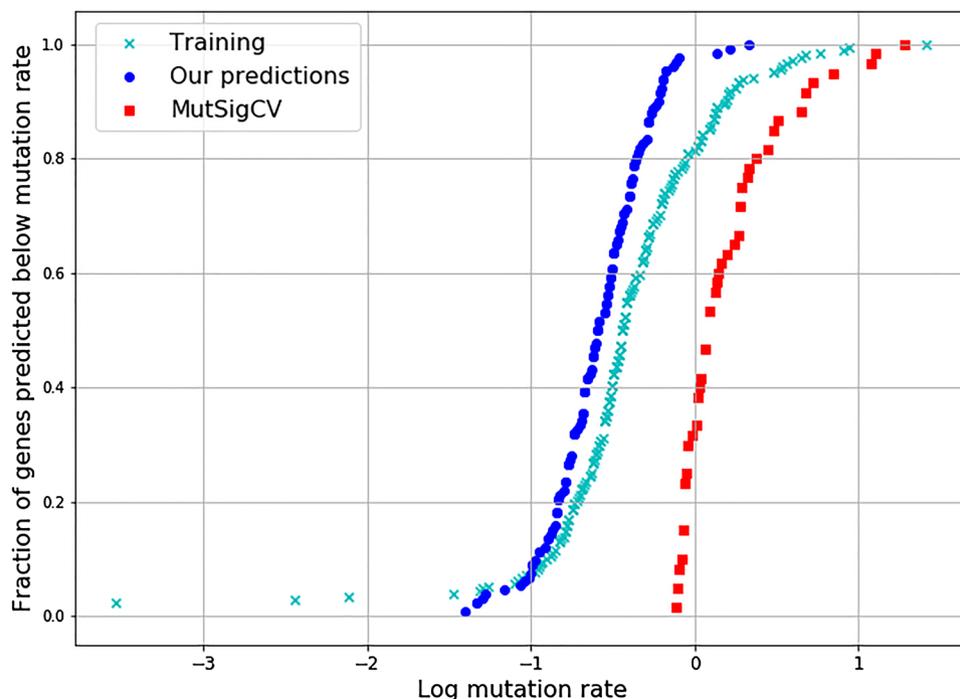
We overcame this by multiple iterations of hyper-parameter estimation by changing the random seed, which helps us identify more stable hyper-parameters. This gave lower accuracy for training sets but improved the accuracy of the test set considerably. When varying sets of random seeds (10, 20, 40, 80, 160, 320) were used, the results were consistent across all cross-validation folds (test set accuracy 0.76 and standard deviation 0.03), implying the increasing number of random seed iterations do not decrease or improve accuracy (Supplementary Table S4). We observe that for a given data fold, the hyper-parameters selected are more stable for varying sets of random seeds. While different parameter sets dominate as the data is changed, the overall results on the test set do not vary.



**Figure 2.** Distribution of top features identified by the classifier for TSG and OG. Training genes were used to study the differences between the distributions of features (kernel density) in TSG and OG. Kolmogorov–Smirnov statistic and the p-value is given for each feature. Higher value of KS statistic shows magnitude of difference of the two distributions.

**cTaG identified new TSGs and OGs along with known driver genes.** All genes that were not used for training the models were classified into TSGs and OGs. This list also contained genes that are known driver genes present in CGC but not used for training. The labels were predicted for the unlabelled genes, of which 126 genes or transcripts showed consensus across all models (Supplementary Table S5). CGC known driver genes contributed to 40.5% of these predictions, which included genes such as ARID1A, ATRX, NF1, TP53, RB1, and STAG1 and their transcripts. Some new genes predicted consistently are SIN3A, ZNF750, IWS1, CD36, ARHGAP35, MGA, and RASA1 as TSGs. The model tends to be biased towards TSGs; out of the 699 genes with consistent predictions across three or more models, only nine are predicted as OGs. The top OGs predicted are U2AF1, BCL2L10, KRAS, MAP1LC3B, C11orf68, TAB3, MED12, MAX, and BRAF. Further, we show not all transcripts of a gene behave like a driver gene, e.g. ATRX transcript ENST00000373344 is labelled as TSG but not ENST00000400866, ENST00000373341. The presence of known driver genes among top TSG and OG shows the validity of cTaG, and those other genes in the list are potential driver genes.

Enrichment analysis of genes for various KEGG and BIOCARTEA pathways revealed genes involved in different cancer pathways such as myeloid leukaemia and pancreatic cancer. Genes are also enriched for various signalling pathways associated with cell growth, such as EGF and PDGF signalling pathways. Further, to validate, a similar analysis was conducted using genes used for training the model. We find GO terms related to cell cycle, regulation of transcription, signalling and cell cycle arrest to be common for both results. These keywords were further clustered with top clusters associated with genes involved in zinc-finger proteins, helicases, ATP-binding, ARID binding and cancer pathways. The analysis shows known driver genes and predicted driver genes enrich for similar pathways.



**Figure 3.** Fraction of genes predicted plotted against log transformed mutation rates. Genes predicted by a given method were sorted based on their mutation rate and plotted against the fraction of genes predicted below the given mutation rate.

**Our approach identifies genes with low mutation frequency.** We analysed the mutation frequencies of the predicted genes. Mutation rates were calculated using MutSigCV, a well-known driver gene predictor, which calculates mutation rates to identify driver genes. MutSigCV ranks all genes of which a total of 602 driver genes were identified above the threshold ( $p \leq 0.005$ ,  $q \leq 0.01$ ). Training data labels were used to compare the two methods. MutSigCV identified 40% for our training gene set with 85 genes predicted as a driver, while cTaG did better by predicting 85% of genes. The mutation rates of the genes predicted by the two models were compared. Since MutSigCV ranks all genes, we picked top genes equal in size to cTaG predictions ( $\geq 5$  model consensus) and calculated KS statistic against the training set and plotted the fraction of genes below the mutation rate of each gene. We observe that the distribution of mutation rates is similar to genes used for model building for our predicted genes, while MutSigCV tends to be biased towards genes with higher mutation rates (Fig. 3). The minimum mutation rate predicted for cTaG was 0.35, while for MutSigCV was 0.90. The KS (Kolmogorov–Smirnov) statistic for both models, when compared to the training set, shows the difference is far lesser for cTaG (KS statistic = 0.193,  $p = 0.054$ ) when compared to MutSigCV (KS statistic = 0.774,  $p = 0.0$ ), which shows that the distribution of mutation rates is similar to what is expected.

Further, we compared the precision of predicted driver genes from cTaG, TUSON, 20/20+ and DriverNet to the pan-cancer genes listed by Bailey et al. *undefined*, and the CGC driver gene list (Supplementary Table S6). We compared with feature-based methods, TUSON and 20/20+ as well as with network-based method DriverNet. For each method, we considered the top-ranking genes and compared the overlap with the pan-cancer gene list. Based on driver genes listed by Bailey et al., cTaG performs best followed by 20/20+, TUSON and DriverNet. For driver genes listed in CGC 20/20+ performs best followed by TUSON, cTaG and DriverNet (Supplementary Fig. 2). Some pan-cancer genes are identified as “rescued” as they were excluded as outliers from the initial list before being included in the final list. None of the rescued genes were identified by cTaG, while the three methods identified 4 (TUSON), 3 (20/20+), and 5 (DriverNet) genes. We do not expect a large overlap with rescued genes as they are manually curated and included by experts. We also observe an overlap between cTaG and the methods with maximum overlap with TUSON with 43 genes, followed by 20/20+ (31 genes) and DriverNet (9 genes). Since the number of genes predicted by methods vary, DriverNet (473), TUSON (269), 20/20+ (137) and cTaG (94), precision was used to normalize for the number of predicted genes.

**Driver genes are tissue-specific.** Cohort studies tend to be specific to a cancer type. The usefulness of a pan-cancer model is further elucidated when it can be used to identify tissue-specific driver genes (Supplementary Table S5). The objective of predicting genes using a subset of data specific to tumour primary tissue source was to identify genes specific to a cancer type. This helped in identifying genes that might otherwise be lost in biological noise (Table 4). We observe TP53 predicted as TSG across the different tissues. Other known driver genes that weren’t identified by the pan-cancer analysis were identified, such as CBF1, CDH1, PTEN in breast cancer and APOB in the liver. Genes such FAM182A, SOX9, AHNK2, ENSG00000121031, FLT3LG, PMP1, ZFP36L2 in the large intestine, ALB, KRTAP19-1, APOB, CD200, CRYGD, KRTAP24-1, OR6N2 in the liver are

Primary tissue	Genes
Breast cancer	TP53, <b>CBFB</b> , RUNX1, CDH1, GATA3, PTEN, TBX3
Central nervous system	TP53, <b>HCN1</b>
Endometrium	<b>KRAS</b> , PIK3R1, PTEN
Hematopoietic	TP53, B2M, CCND3, HLA-A
Kidney	PBRM1, <b>VHL</b> , TP53
Large intestine	TP53, FBXW7, <b>FAM182A</b> , SOX9, <b>AHNAK2</b> , TCF7L2, ENSG00000121031, <b>FLT3LG</b> , <b>PMEPA1</b> , <b>ZFP36L2</b>
Liver	TP53, <b>ALB</b> , <b>KRTAP19-1</b> , <b>APOB</b> , <b>CD200</b> , <b>CRYGD</b> , <b>KRTAP24-1</b> , <b>OR6N2</b>

**Table 4.** Driver genes predicted for each of the cancer types. The genes reported showed consensus for > 4 CV models. Genes in bold did not find similar consensus in the pan-cancer predictions. New genes are underlined.

novel predictions, and their functions in these cancers can further be studied. We used the pan-cancer model (cTaG) to predict tissue-specific driver genes and identified new genes not reported by the pan-cancer analysis.

Genes identified for breast cancer was validated by supporting literature. CBFB<sup>30</sup> and PTEN<sup>31,32</sup> is a known TSG in breast cancer. PTEN is found to be under-expressed in breast cancer<sup>33,34</sup>. While CDH1 mutations are found mostly in stomach cancer, they are also shown to be frequently occurring in lobular breast cancer<sup>35,36</sup>. Pathway analysis of breast cancer genes shows enrichment of pathways involved in gene expression regulation governed by TP53, RUNX1 and PTEN, which includes pathways that regulate estrogen-mediated transcription. CBFB deletion leads to expression loss of RUNX1<sup>30</sup>, which can no longer regulate NOTCH signalling by repression, which is confirmed by pathway analysis. Some apoptosis pathways are enriched that include CDH1 and TP53 genes. The genes identified by cTaG (pan-cancer model) for breast cancer samples predict genes functionally important in breast tumour cells.

Predictions made for liver cancer mainly were novel, which made literature validation difficult. RNA expression levels of genes APOB, ALB and CD200 were higher compared to all other tissues (as reported by The Human Protein Atlas). Higher albumin levels are known to decrease the risk of HCC (Hepatocellular carcinoma)<sup>37</sup>. APOB mutational signatures are shown computationally to be significant to predict prognosis by loss of regulation of genes such as TP53, PTEN, HGF<sup>38</sup>. While the role of other genes is difficult to elucidate, our method helps identify research gaps that can be filled by studying these potential driver genes.

## Discussion

Identification of driver genes has been an important focus area of cancer research because these genes are potential targets for therapy and biomarkers. Different approaches have been used for identification using mutational information<sup>17,18,39</sup>, gene expression levels<sup>40</sup>, protein structural information<sup>41</sup>, network analysis<sup>42,43</sup> or using multiple data sources<sup>40</sup>. Advances in sequencing technologies have made mutational information easily available, and different tools have been developed to identify driver genes. Driver genes are further classified into TSGs and OGs based on the functional impact of the mutations they harbour.

We adopt a classification approach that is able to predict TSGs and OGs by leveraging a set of ratio-metric and other new features. Traditional methods identify genes based on the mutation rate. Compared to previous approaches, we ascribe a higher significance to functional impact along with the position of the mutations, as the genes might contain mutations in functionally important regions even though the mutation rate may not be very different from the background mutation rate. Features like nonsense entropy, frameshift frequency captures the recurrence of a mutation when multiple samples are considered, thus taking into account the position at which the mutation occurs.

For classification, many different algorithms are available, but the performance of the algorithm is dependent on the data and estimation of parameters. It is especially important while solving biological problems, where the training data might be small, to build robust models. High performance on a given data might also be due to overfitting. We sought to avoid overfitting by performing a standard fivefold cross-validation while estimating random forest parameters as well as multiple iterations for estimation of stable parameters. We developed a procedure to verify that the predictions are reasonably stable. An ensemble of models is used to make final predictions.

It is important that the estimated parameters are robust to changes in data. For random forest, we estimated four parameters out of which *n\_estimator* seemed to have a large effect on the classification. For large values of *n\_estimator*, we were able to show high accuracy scores similar to Davoli et al.<sup>17</sup>, but the accuracy scores for the test set were much lower. We were not able to compare our performance on the test set with that of Davoli et al.<sup>17</sup>, as their test set results have not been published. To build a better model that is not biased to data, we needed a more robust classifier that is sufficiently generalised and not dependent on the training data.

The models generated were used to find which of the new features are important for classification. To evaluate the cTaG, we used fivefold cross-validation with 20% test dataset while maintaining the ratio between TSGs and OGs and calculated metrics such as accuracy and F1 score. Instead of AUROC (Area under Receiver Operating Characteristic), we chose to show accuracy and F1 score, as AUROC only helps in estimating if the model can separate the given classes but tells us very little about the classification power for each of these classes. The F1 score is calculated for each of the given classes and helps understand if the model is biased towards any one of the classes. The accuracy score on the test set shows that mere accuracy is not sufficient for judging a model. The models perform slightly better for TSGs, though it is far poorer at classifying OGs.

While assessing the model, it is important to use metrics such as the F1 score, as it scores predictions for each of the classes. Studies reporting only AUROC statistics present an incomplete picture and are not effective in estimating the performance of the model, especially in datasets having a class imbalance<sup>44</sup>. This is evident when we compare AUROC of balanced bagging model ( $0.76 \pm 0.07$ ) with cTaG ( $0.54 \pm 0.07$ ). AUROC gives measures the models ability to separate the classes and not the prediction power. By reporting both accuracy as well as F1 score, we show that the cTaG does not perform equally for both classes but tends to be better at classifying TSG than OG. This indicates that the chosen features are not sufficient to classify oncogenes.

Feature ranking shows that features containing information about LOF, nonsense, frameshift and missense mutations are important. Nonsense and frameshift mutations are frequently seen in TSGs, while recurrent missense mutations are characteristic of OGs as they lead to "gain of function".

The list of genes classified contained known driver genes and other transcript data for genes present in training and test set. We found that TSGs such as ATRX, PTCH1, and STAG2 were classified as TSGs with high probability. KDM6A gene and its transcripts (ENST00000377967, ENST00000382899) feature among the top, which shows that cTaG can also help classify a particular transcript of a gene. Similarly, TP53 and its six transcripts were all classified as TSGs. Genes U2AF1, KRAS, BRAF, MED12 and MAX were classified as OGs among the top genes identified as OGs. As the probability scores for OGs tend to be lesser than TSGs, relatively fewer OGs make the cut-off for the top 5 percentile.

Among the top TSGs identified, CD36 (previously known as FAT) is a receptor protein for fatty acids. CD36 is also a prognostic marker for different cancer types<sup>45,46</sup> and found in metastatic cells<sup>45,47</sup>. While the expression of a gene is markedly different from normal cells, the molecular mechanism that enables metastasis is not well understood. Another gene, ARHGAP35, is a glucocorticoid receptor DNA binding factor, which has also been previously identified as a potential driver gene by other methods<sup>48,49</sup>. ZNF750, zinc finger protein 750, has been established as a tumour suppressor in oesophageal squamous cell carcinoma<sup>50–52</sup> though it is absent from the CGC driver gene list. Some other potential TSGs not present in the CGC list are MBD6 and RASA1. In the human protein atlas, MAP1LC3B is labelled as a prognostic marker for renal and stomach cancer among the three shortlisted OGs.

Our model, cTaG, does have some limitations. We have used binary classification to identify TSGs and OGs, which classifies all genes as either TSG or OG. All genes containing mutations are not driver genes, and thus, a majority of genes are neutral. We overcome this by taking consensus across the five models built. It may be possible to improve on this classification by solving a multi-class problem where each gene is identified as TSG, an OG or neutral gene. The difficulty in this problem stems from the huge class imbalance in the data as well as the definition of neutral genes. While there are studies showing the importance of a gene in tumour evolution, it is challenging to define genes that are not involved in cancer progression. Most methods use a list of genes that do not contain cancer driver genes and genes involved in cancer pathways, but this does not exclude potential driver genes.

We compare our method with one of the most commonly used driver identification tools, MutSigCV, but other methods exist that use similar features (TUSON and 20/20+). We used data from Tokheim et al. *undefined*<sup>19</sup> to compare with other methods, but missense and frameshift entropy and frequency could not be calculated. Due to the lack of feature information for one of the high ranked features in our model, we find the fractional overlap of 0.07 with 188 CGC genes compared in the study. We instead compare the genes reported by TUSON, 20/20+, DriverNet and cTaG to the validated pan-cancer driver gene list by Bailey et al.<sup>29</sup>. The overlap with the driver gene list corresponds to number of predicted genes, with DriverNet being the exception. DriverNet employs network and gene expression data but a comparison between the two methods show that cTaG can predict a larger fraction of known driver genes. Additionally, it has been seen that mutations are not always the reason for the change in functionality, and regulation might also lead to a change in expression at transcriptomic and proteomic levels. Other than adding new features to the analysis, including transcriptomic and proteomic data along with genomic mutation data might further improve the classification of genes.

## Conclusion

In summary, we see two main contributions of our paper. First, we developed a classifier, which enabled an improved recall of TSGs and OGs compared to previously proposed methods in the literature. We carefully avoided overfitting for achieving consistent and high confidence results. Second, we predicted many potential TSGs and OGs at both the pan-cancer and tissue-specific level, which form a ready shortlist for further experimental investigation. Some of the top predictions were already well-known cancer drivers, while others are reported in multiple cancer studies though their role in tumorigenesis is not yet well understood. Our approach is also readily amenable to the integration of other omic datasets as they become available.

## Data availability

Data for this analysis was downloaded from COSMIC (v79). The pre-processed COSMIC data used to build the feature matrix is available from zenodo.com: <https://doi.org/10.5281/zenodo.4153052>. The processed data and codes are available from GitHub: <https://github.com/RamanLab/cTaG>.

Received: 26 March 2021; Accepted: 13 December 2021

Published online: 07 January 2022

## References

1. Ferlay, J. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **144**, 1941–1953. <https://doi.org/10.1002/ijc.31937> (2019).

2. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).
3. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
4. Stratton, M., Campbell, P. & Futreal, P. The cancer genome. *Nature* **458**, 719–724 (2009).
5. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
6. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
7. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
8. Beishline, K. & Azizkhan-Clifford, J. Sp1 and the ‘hallmarks of cancer’. *FEBS J.* **282**, 224–258 (2015).
9. Cavallo, F., De Giovanni, C., Nanni, P., Forni, G. & Lollini, P. L. The immune hallmarks of cancer. *Cancer Immunol. Immunother.* **60**, 319–326 (2011).
10. Shahmarvand, N., Nagy, A., Shahryari, J. & Ohgami, R. S. Mutations in the signal transducer and activator of transcription family of genes in cancer. *Cancer Sci.* **109**, 926–933. <https://doi.org/10.1111/cas.13525> (2018).
11. Zhang, E. *et al.* Roles of PI3K/Akt and c-Jun signaling pathways in human papillomavirus type 16 oncoprotein-induced HIF-1 $\alpha$ , VEGF, and IL-8 expression and in vitro angiogenesis in non-small cell lung cancer cells. *PLoS ONE* **9**, e103440 (2014).
12. Hofree, M. *et al.* Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat. Commun.* **7**, 12096 (2016).
13. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
14. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
15. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
16. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
17. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
18. Melloni, G. E. *et al.* DOTS-Finder: A comprehensive tool for assessing driver genes in cancer genomes. *Genome Med.* **6**, 44 (2014).
19. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci.* **113**, 14330–14335 (2016).
20. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
21. Forbes, S. A. *et al.* COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
22. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
23. Pedregosa, F. & Varoquaux, G. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
24. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
25. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**(1), 1–13. <https://doi.org/10.1093/nar/gkn923> (2009).
26. Bashashati, A. *et al.* DriverNet: Uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **13**, R124 (2012).
27. Hou, J. P. & Ma, J. DawnRank: Discovering personalized driver genes in cancer. *Genome Med.* **6**, 56 (2014).
28. Dinstag, G. & Shamir, R. PRODIGY: Personalized prioritization of driver genes. *Bioinformatics* **36**, 1831–1839 (2020).
29. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
30. Malik, N. *et al.* The transcription factor CBF $\beta$  suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* **10**(1), 2071. <https://doi.org/10.1038/s41467-019-10102-6> (2019).
31. Lu, Y. *et al.* The PTEN/MMAC1/TEP tumor suppressor gene decreases cell growth and induces apoptosis and anoikis in breast cancer cells. *Oncogene* **18**(50), 7034–7045. <https://doi.org/10.1038/sj.onc.1203183> (1999).
32. Weng, L.-P. PTEN coordinates G1 arrest by down-regulating cyclin D1 via its protein phosphatase activity and up-regulating p27 via its lipid phosphatase activity in a breast cancer model. *Hum. Mol. Genet.* **10**(6), 599–604. <https://doi.org/10.1093/hmg/10.6.599> (2001).
33. Li, S. *et al.* Loss of PTEN expression in breast cancer: Association with clinicopathological characteristics and prognosis. *Oncotarget* **8**(19), 32043–32054. <https://doi.org/10.18632/oncotarget.16761> (2017).
34. Zhang, H. Y., Liang, F., Jia, Z. L., Song, S. T. & Jiang, Z. F. PTEN mutation, methylation and expression in breast cancer patients. *Oncol. Lett.* **6**(1), 161–168. <https://doi.org/10.3892/ol.2013.1331> (2013).
35. Hansford, S. *et al.* Hereditary diffuse gastric cancer syndrome: CDH1 mutations and beyond. *JAMA Oncol.* **1**(1), 23. <https://doi.org/10.1001/jamaoncol.2014.168> (2015).
36. Schrader, K. A. *et al.* Hereditary diffuse gastric cancer: Association with lobular breast cancer. *Fam. Cancer* **7**(1), 73–82. <https://doi.org/10.1007/s10689-007-9172-6> (2008).
37. Nojiri, S. & Joh, T. Albumin suppresses human hepatocellular carcinoma proliferation and the cell cycle. *Int. J. Mol. Sci.* **15**(3), 5163–5174. <https://doi.org/10.3390/ijms15035163> (2014).
38. Lee, G. *et al.* Clinical significance of APOB inactivation in hepatocellular carcinoma. *Exp. Mol. Med.* **50**, 147 (2018).
39. Kumar, R. D., Searleman, A. C., Swamidass, S. J., Griffith, O. L. & Bose, R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics* **31**, 3561–3568 (2015).
40. Sanchez-Garcia, F. *et al.* Integration of genomic data enables selective discovery of breast cancer drivers. *Cell* **159**, 1461–1475 (2014).
41. Fujimoto, A. *et al.* Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *Sci. Rep.* **6**, 26483 (2016).
42. Ramsahai, E., Walkins, K., Tripathi, V. & John, M. The use of gene interaction networks to improve the identification of cancer driver genes. *PeerJ* **5**, e2568 (2017).
43. Chen, Y. *et al.* Identifying potential cancer driver genes by genomic data integration. *Sci. Rep.* **3**, 3538 (2013).
44. Jeni, L. A., Cohn, J. F. & De La Torre, F. Facing imbalanced data—recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* 245–251 (2013). <https://doi.org/10.1109/ACII.2013.47>
45. Ladanyi, A. *et al.* Adipocyte-induced CD36 expression drives ovarian cancer progression and metastasis. *Oncogene* **37**(17), 2285–2301. <https://doi.org/10.1038/s41388-017-0093-z> (2018).
46. Hale, J. S. *et al.* Cancer stem cell-specific scavenger receptor 36 drives glioblastoma progression. *Stem Cells* **32**, 1746–1758. <https://doi.org/10.1002/stem.1716> (2014).
47. Pascual, G. *et al.* Targeting metastasis-initiating cells through the fatty acid receptor CD36. *Nature* **541**(7635), 41–45. <https://doi.org/10.1038/nature20791> (2017).
48. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
49. Zhang, Y. *et al.* Genetic variations in cancer-related significantly mutated genes and lung cancer susceptibility. *Ann. Oncol.* **28**, 1625–1630 (2017).
50. Lin, D. C. *et al.* Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* **46**(5), 467–473. <https://doi.org/10.1038/ng.2935> (2014).

51. Otsuka, R. *et al.* ZNF750 expression is a potential prognostic biomarker in esophageal squamous cell carcinoma. *Oncology* **94**, 142–148 (2018).
52. Hazawa, M. *et al.* ZNF750 is a lineage-specific tumour suppressor in squamous cell carcinoma. *Oncogene* **36**, 2243–2254 (2017).
53. Sudhakar, M., Rengaswamy, R. & Raman, K. Novel ratio-metric features enable the identification of new driver genes across cancer types. *bioRxiv* <https://doi.org/10.1101/2020.01.17.910075> (2020).

## Acknowledgements

This manuscript has been released as a pre-print at bioRxiv<sup>53</sup>.

## Author contributions

M.S., R.R. and K.R. conceived and designed the study. M.S., R.R., and K.R. were involved in the analysis and interpretation of data. M.S., R.R. and K.R. drafted the manuscript. The study was supervised by R.R. and K.R. All authors read and approved the final manuscript.

## Funding

This work was supported by Department of Biotechnology, Government of India (DBT) (BT/PR16710/BID/7/680/2016), IIT Madras, Centre for Integrative Biology and Systems mEdicine (IBSE) and Robert Bosch Center for Data Science and Artificial Intelligence (RBCDSAI).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-04015-y>.

**Correspondence** and requests for materials should be addressed to R.R. or K.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022