Published online 12 April 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 12* ***e70***
https://doi.org/10.1093/nar/gkac215

EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants

Saeid Parvandeh[1,*], Lawrence A. Donehower[2,3], Katsonis Panagiotis[1], Teng-Kuei Hsu[4], Jennifer K. Asmussen[1], Kwanghyuk Lee[1] and Olivier Lichtarge[1,4,*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA, [2]Department of Molecular Virology and Microbiology, Houston, TX 77030, USA, [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA and [4]Department of Biochemistry & Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received October 28, 2021; Revised March 17, 2022; Editorial Decision March 19, 2022; Accepted March 21, 2022

ABSTRACT

Discovering rare cancer driver genes is difficult because their mutational frequency is too low for statistical detection by computational methods. EPIMUTESTR is an integrative nearest-neighbor machine learning algorithm that identifies such marginal genes by modeling the fitness of their mutations with the phylogenetic Evolutionary Action (EA) score. Over cohorts of sequenced patients from The Cancer Genome Atlas representing 33 tumor types, EPIMUTESTR detected 214 previously inferred cancer driver genes and 137 new candidates never identified computationally before of which seven genes are supported in the COSMIC Cancer Gene Census. EPIMUTESTR achieved better robustness and specificity than existing methods in a number of benchmark methods and datasets.

INTRODUCTION

Many approaches that link genotype to phenotype were developed to identify cancer driver genes in the hope that experimental screens would then validate some as drug targets for treatment (1–6). Many of the methods are based on genome sequencing of tumors and rely on the frequency to which a gene is mutated to identify genes that are under positive mutational selection in tumors (2,3,7–9). Unfortunately, most tumors have a large fraction of benign passenger mutations (10). Therefore, the methods that rely on mutation frequency to identify driver genes tend to remain insensitive (11–13), regardless of the specific computational and statistical evaluations of whole exome somatic mutations (14–16).

Defining an accurate label for samples is essential for prediction of driver genes and is one of the most challenging tasks for physicians (17). Machine learning is an alternative technique that can discover and identify driver genes in complex diseases (18) and relationships between genotype–phenotype (19–21). Recently, many machine learning methods have been developed to reveal the ambiguity of the genotype–phenotype relationship (3,17–30). In addition, prediction of cancer driver genes that can stratify cancer patient survival rates requires improvements. Given that the most important challenge in machine learning is to define the right label (e.g. a binary classification label like 1's for cases and 0's for controls or quantitative label like continuous values) for samples (31), several studies used pathways as label (23,32–34) or utilized matched normal subjects and cancer patients as a binary classification label (35–38). Despite the strength of these techniques, the low frequency of mutations separating the few functionally relevant ones from the vast majority of random ones with no functional relevance is close to the noise threshold. Both relevant and irrelevant random mutations are likely driven by the same random mutagenic processes in tumor cells (replication errors, DNA damage events from mutagens, etc.). Presumably, sequencing artifacts are mostly eliminated by sophisticated mutation calling methods and germline mutations are easily eliminated by comparing sequences of non-tumor tissues to tumor tissues (11,12,39). Other studies have reported ways to identify mutation drivers through their pattern of mutations (2,40). However, previous studies on the TCGA PanCancer dataset were limited to a sublist of all tumor types (41) and avoided low mutation frequency from low functional impact mutations (13). Additionally, despite experimental and clinical studies on epistasis effects among mutations by considering mutual exclusivity among tumor samples (42–44) PPI network (45–47), there is no previously reported machine learning

[*]To whom correspondence should be addressed. Tel: +1 713 798 7677; Email: saeidparvandeh@gmail.com
Correspondence may also be addressed to Olivier Lichtarge. Tel: +1 713 798 5646; Email: lichtarge@bcm.edu

method to consider the epistasis effects of cancer somatic mutations.

A formal perturbation equation describing the genotype–phenotype relationship was proposed (48) to compute the fitness effect, or Evolutionary Action (EA), of mutations. This equation is a product between the size of a mutation and the functional sensitivity at the position being mutated. EA quantified the likely functional impact of individual mutations on par or better than other methods in blind and objective assessments (49), and also helped in the interpretation of exome data (50), ignoring any interaction effects. Newer methods have begun to integrate EA over entire populations of mutations in cohorts of patients (51–54), Cohort Integral (CI) (55). These attempts sum the separate effects of individual mutational steps to identify genes for specific traits or diseases. None, however, capture interaction epistasis effects, which we now try to do with application to the low mutational frequency of cancer driver genes.

Here, we used whole exome data from The Cancer Genome Atlas (a publicly available MAF file that was compiled by the MC3 working group and is based on the consensus calls from seven software packages) patients through a machine learning procedure based on the nearest-neighbor machine learning feature selection algorithm that uses the advantage of a multigene interaction (epistasis) approach to find driver genes which link to different tumor types. Briefly, EA scores of the functional impact of coding mutations follow nearly exponential action distribution for human coding polymorphisms (48). We build self-defined synthetic controls by random selection from the EA scores of all possible single nucleotide changes in gene perturbation for the human genome sequence. Comparing to EA scores of somatic mutations which tend to have loss of function and gain of function which is not biased to exponential distribution, we will have distributions of cases and controls that are significantly different. Given that, we use the ReliefF feature selection algorithm (56,57) to weigh the genes importance that is based on *k* nearest neighbor (*k*-NN) classification. The output will be a list of genes that could drive cancer across all cancers or in individual cancers. Mapping was performed between specific cancer driver genes and the cell line primary disease annotator from DepMap to assess the recovery of known cancer genes.

MATERIALS AND METHODS

Data preparation

TCGA sequence data was obtained from the MC3 Working Group, which recently compiled a publicly available MAF file (syn7824274, https://gdc.cancer.gov/about-data/publications/mc3-2017) annotated with filter flags to highlight potential artifacts and discrepancies (58). This contains the most uniform attempt to catalogue Cancer Genome Atlas (TCGA) somatic mutations and provides consensus calls from seven software packages for 33 tumors (58). The possible artifact flags include strand-bias, contamination, oxo-guanine artifacts, and low normal read depth, and the 'PASS' label assigned to a mutation that was called by at least two callers. Five of the seven software packages related to Single Nucleotide Variants (SNV) and three of them related to short Insertion Deletion (INDEL) events, where VarScan 2 provides both types of analysis. Seven software packages consist of: VarScan2 (59), MuTect (60), SomaticSniper (61), Indelocator (62), Pindel (63), RADIA (64) and MuSE (65). For our analysis, we used the preprocessed MC3 dataset (22) with the following quality controls: excluding hypermutated samples according to Tukey's outlier condition (more than 1.5 times the interquartile range from the quartiles) if mutation numbers in a sample exceeding 1000, and samples that are flagged by the analysis working group based on pathology. In addition, we filtered out all hypermutated samples exceeding 1,000 mutations. We considered all samples in the dataset and ended up with 9,973 samples from 33 tumor types having a total of 1,087,555 mutations with 1,006,892 missense mutations, 76,386 nonsense mutations, 1,276 nonstop mutations, 1,645 translational start site mutations, 85 splice site mutations, 999 frameshift deletions and 272 frameshift insertions.

Evolutionary action to score functional impact of mutations

Evolutionary Action (EA) (48) is a method that estimates the functional impact of missense mutations using protein homology data. We represent evolutionary fitness as a genotype–phenotype relationship, where the estimation of the fitness effect for each missense mutation from amino acid *X* to any other amino acid *Y* at sequence position *i*, will drive a phenotype change through an evolutionary function:

$$\Delta\varphi \approx \frac{\partial f}{\partial r_i} \cdot \Delta r_{i,\quad X\to Y}.$$

where $\Delta\varphi$ is the action of the mutation Δr on the fitness with inverse amino acid substitution log-odds, · denotes the scalar product, and $\partial f/\partial r$ is the mutated site sensitivity to the genotype changes computed with Evolutionary Trace (ET) (66) ranking of importance. EA accurately predicts the impact of mutations in complex molecular machines. It also outperformed competitive entries in Critical Assessment of Genome Interpretation (CAGI) competitions (49,50). EA scores are available for nonprofit use: http://eaction.lichtargelab.org.

EPIstasis MUTations ESTimator (EPIMUTESTR)

In general, to use machine learning classification methods, we need to have case and control datasets to identify biomarkers that are highly correlated to case/control labels that lead to understanding the cause of the disease (67). Starting with the preprocessed MC3 dataset (22), we annotated it with EA methods to give every mutation an importance score. We designed the EPIMUTESTR algorithm in four parts: (A) take the EA annotated MC3 datasets as input; (B) construct a case matrix (*M*) of samples (*i*) by genes (*j*), where each entity (M_{ij}) is the maximum EA score among all mutations per gene. Construct an empty control matrix (N_{ij}) of the same size of case matrix and fill with the synthetic EA scores by maintaining the same mutational signature rate from the case matrix; (C) from the ReliefF feature selection family (56,57) we used relevant estimation of features that is based on the *k*-nearest neighbor (*k*-NN) algorithm. It takes the case and control matrix of the EA scores of coding variants as training features to identify which genes taken together can best distinguish cases from random controls; (D) the result will be a list of gene weights which, after square root back-transformation into a Gaussian distribution, will let us pick the most important ones in the right tail with q-value < 0.1.

The process of simulating the synthetic control matrix is as follows: for a given gene, we randomly selected the same number of mutations across all samples from all possible single nucleotide changes in a given canonical transcript of the gene by keeping the mutational signature rate. In step C, k-NN is the core part of EPIMUTESTR. k-NN uses a distance metric (Euclidean or Manhattan) to classify the sample (cases/controls) in which we could get the advantage of distance metric through this algorithm in order to weigh the genes. Basically, samples that are in one class (case or control) have very close EA scores (this is also true in fitness effect of mutations) and those samples that are in different classes (case and control) have very different EA values. We used this technique to weigh the genes using their EA scores for the samples that are in the same class or different classes. Therefore, each gene g_i (the ith gene) starts with an initial weight $W_0(g_i) = 0$. For random sample s, this weight is updated iteratively:

$$W_{(t+1)}(g_i) = W_t(g_i) - diff(g_i,\ s,\ h)^2 + diff(g_i, s,\ m)^2$$

to favor or penalize weights that yield large Manhattan separation distances between s and nearest neighbor of the opposite (miss (m)) or of the same class (hit (h)), respectively. Terms $diff(g_i,\ s,\ h)^2$ and $diff(g_i, s,\ m)^2$ are used for calculating the distances between random sample(s) and the nearest neighbors (hit (h) or miss (m)). We maintain these two weights for all genes and repeat for all the samples and add or subtract the distance value from previously maintained weights. The final gene's weight is simply the sum of distances across all genes. k-NN can efficiently take into consideration all genes at once and is robust with respect to incomplete data. There are two different approaches for neighbor finding: specify fixed k or neighborhood radius that varies for each sample. The radius for each sample can be defined as the average of all Manhattan distances of sample to all other samples subtracted by half of their standard deviation (68). It has been empirically shown that for balanced cases and controls datasets an approximation to the expected number of neighbors within the radius approaches is $(n - 1) \times 0.154$, where n is the total number of samples (69), but the choice can have a large impact and can be optimized for feature estimation (70).

Avana CRISPR screen

We used the Avana CRISPR screen database downloaded from DepMap (https://depmap.org/) that includes the CRISPR (Broad Avana) screen data (71,72), the cell lines, and merged mutation calls (coding region, germline filtered) from Cancer Cell Line Encyclopedia (CCLE) (73,74). We annotated all the missense variants using the EA equation and put them into two categories: moderate EA ($30 \leq$ EA score < 70) and others (low EA variants $0 \leq$ EA < 30, high EA variants $70 \leq$ EA < 100, nonsense variants). We have six gene sets to be compared: (i) random genes consisting of 100 random genes selected from all the cell lines in the Avana set which is 17,632 screened genes, (ii) oncogenes from the Cancer Gene Census (COSMIC) (v79), (iii) CE: core essential genes from the CRISPR screen database, (iv) NE: non-essential genes from the CRISPR screen database, (v) EPIMUTESTR PanCancer: all candidate genes identified as oncogenes in PanCancer and (vi) EPIMUTESTR individual cancers: all candidate genes identified as oncogenes in individual cancer types. The mapping between specific cancers from TCGA and the cell line primary disease annotator from DepMap includes: BLCA (bladder urothelial carcinoma), BRCA (breast invasive carcinoma), CESC (cervical squamous cell carcinoma), COAD (colorectal adenocarcinoma), GBM (glioblastoma multiforme), HNSC (head and neck squamous cell carcinoma), LIHC (liver hepatocellular carcinoma), OV (ovarian serous cystadenocarcinoma), SKCM (skin cutaneous melanoma), STAD (stomach adenocarcinoma).

Guidelines for evaluating state-of-the-art algorithms

We used four assessment guidelines suggested by Tokheim *et al.* (26) to compare state-of-the-art algorithms with EPIMUTESTR toward cancer driver discovery. Given their preprocessed PanCancer MAF dataset derived from 7916 cancer patients with 34 specific tumors, they defined four evaluation criteria as follow: (a) the fraction of overlap between pan-cancer genes and the COSMIC Cancer Gene Census (75); (b) consensus of mutual co-occurrence genes among seven methods including: MutsigCV (3), ActiveDriver (76), MuSiC (77), OncodriveClust (41), OncodriverFM (78), OncodriverFML (5), Tumor Suppressor and Oncogenes (TUSON) (79); (c) the consistency of top-ranked genes of two random splits of samples and (d) mean absolute $\log_2$ fold changes (MLFC) metric to find the discrepancy deviation between a uniform p-value distribution and p-value distribution reported by a method.

Network of cancer genes Enrichment

Network of Cancer Genes (NCG6.0) (80) database is a manually curated repository of 2,372 known cancer genes associated with cancer. They are collected from 275 publications and 273 cancer sequencing screens of 34,905 donors and multiple primary sites from 100 cancer types. Disease Ontology Semantic and Enrichment (DOSE) analysis (81) provides semantic similarity computations between terms and genes to find the similarities of diseases and gene functions including hypergeometric tests. We used DOSE (R package) to query candidate genes through NCG and perform enrichment analysis.

PubMed literature search

PubMed literature search is a way to query PubMed for each gene, cross-indexing the keywords 'cancer' and 'protein/gene'. We calculated the hypergeometric test of genes that appeared in the PubMed literature to find the gene enrichment with cancer association. We used Entrez package from Bio library (Python package) to query the candidate genes and determine the number of publications associated with the key words.

Classification of tumor suppressor genes and oncogenes

A useful classification tool for candidate driver genes is subcategorization as either tumor suppressor genes (loss of function) or oncogenes (gain of function). We subclassified each candidate driver gene as a likely oncogene or tumor suppressor gene through an empirically derived formula. Using the TCGA PanCancer classification of 299 identified cancer driver genes reported by Bailey *et al*. (22) we determined total numbers of non-synonymous missense and truncating mutations for each of the driver genes labeled as PanCancer oncogenes or tumor suppressor genes. We then determined the ratio of missense mutations to truncating mutations for each of these genes. The oncogene drivers had a mean ratio of 30.0 missense/truncating mutations (ratios ranged from 6.6 to 217.8). The tumor suppressor drivers had a mean ratio of 2.21 (ratios ranged from 0.53 to 4.18). The missense/truncating mutation ratio of all exomic genes (>99% non-driver controls) sequenced by the TCGA PanCancer Project was 7.26.

EPIMUTESTR candidate driver genes were also categorized as likely tumor suppressors (missense/truncating mutation ratios < 4.0) or likely oncogenes (missense/truncating mutation rations > 12.0). Candidate drivers with ratios between 4.0 and 12.0 were classified as indeterminate. Also, candidates with total non-synonymous mutation numbers <15 were classified as indeterminate.

RESULTS

EPIMUTESTR predicted 418 candidate cancer driver genes

The main steps of EPIMUTESTR (Figure 1) are: (A) Input preparation by annotating EA scores to all missense and truncating (nonsense, nonstop, translational start site, splice site, frameshift deletions and frameshift insertions) mutations of each TCGA sample (10,265 samples and 2,013,635 missense and nonsense mutations from the MC3 preprocessed dataset (see Materials and Methods)), and by generating a matrix of samples by genes, where the maximum EA score of each gene will be the training feature. (B) For controls, we also created for each patient a size-matched random matrix. Specifically, we randomly selected for each gene as many mutations as there are for that gene in the patient matrix from all possible nucleotide changes for that gene (incorporating mutational signature) and used maximum EA score as training feature. (C) To train the *k*-NN classifier, each gene was assigned a weight based on a *k* nearest neighbor algorithm in n-dimensional (n = total number of genes) Manhattan space (56,57) with optimized *k* (69). For this, we combined a patient matrix with a synthetic control matrix row wise and assigned case label (1's for patient samples) and control label (0's for synthetic control samples). Then, starting with 0 weight for all the genes, the genes' weights were updated based on the *k*-NN classification (see Materials and Methods). The rationale is that an important feature (or gene) can segregate samples between different classes. Furthermore, all the genes were taken into account for classification, so our method captures interaction (epistasis) effects among genes and local dependencies (82,83). (D) The last step identifies putative cancer driver genes in the form of gene weights in which each gene has a valence, and since there is no standard threshold to select the top genes, we square-transformed the weight distribution of all genes into a Gaussian distribution and calculated *p*-value for each gene using 1-(cdf). We then calculated the adjusted *p*-value using the Benjamini–Hochberg method in order to use the standard *q*-value threshold (0.1) for two reasons: first, to get the most significant genes; and second, to get fewer gene numbers (because the standard *p*-value threshold (0.05) gives us more genes that increase the risk of false positives). In order to minimize false positive genes, we replicated EPIMUTESTR 10 times with different synthetic controls. We sorted the genes based on the lowest *q*-value in each replication and considered the genes with at least five times co-occurrence (70). This procedure identified 407 candidate genes from all cancer types separately, and 11 additional PanCancer genes (Supplementary Table S1). In total, our final gene list is 418 genes (Supplementary Table S1), which falls within the range of number of genes identified by other methods (40–429, (2,3,7,22,26,42,77,79)).

Comparison to other predicted and known cancer driver genes

In order to evaluate our cancer driver gene list, we first compared to previously published cancer driver gene lists. We collected known cancer genes from 10 sources (namely, COSMIC Cancer Gene Census (75), TUSON (79), MuSic (77), MutSigCV (2), MutSig2CV (3), 2020 (40), 2020+ (26), Bailey *et al*. (22), Ding *et al*. (42) and dNdScv (7)) and considered only those genes present in at least two to obtain 428 imputed 'gold standard' putative cancer driver genes (Supplementary Table S2). EPIMUTESTR recovered 168 (39.3%) of these genes (Fisher exact test, *p*-value = 6.7e–147).

Experimental results to validate EPIMUTESTR oncogenes

The CRISPER-Cas9 system is a powerful tool for multiplexed screening to systematically identify genes relevant to cancer cell division and survival. It has been used to build a Cancer Dependency Map (DepMap) database (84) and catalogs cell-line-specific genetic and chemical vulnerabilities of CRISPR-Cas9 loss-of-function screens in 342 cancer cell lines. The Broad Institute developed CERES (71,72) to estimate the degree of essentiality of individual gene expression for survival and division of cancer cell lines following CRISPR/Cas9-engineered gene ablation. In order to find out whether EPIMUTESTR genes were cancer essential we used the CERES method to compare their ability to match DepMap data compared to common core essential genes, COSMIC genes, EPIMUTESTR genes from each individual cancer, EPIMUTESTR genes from PanCancer, non-essential genes and random genes. We retained genes in the moderate category (30 < EA score < 70) that are considered as likely oncogenes and we observed that EPIMUTESTR performs similar to common core essential genes and outperforms COSMIC genes (*t*-test *p*-value < 0.07) (Figure 2). These data suggest that EPIMUTESTR effectively identified genes that are relevant to cancer-specific vulnerabilities (Kruskal-Wallis, *p*-value < 2.2e–16). We also compared the distribution characteristics of the EA scores for an essen-

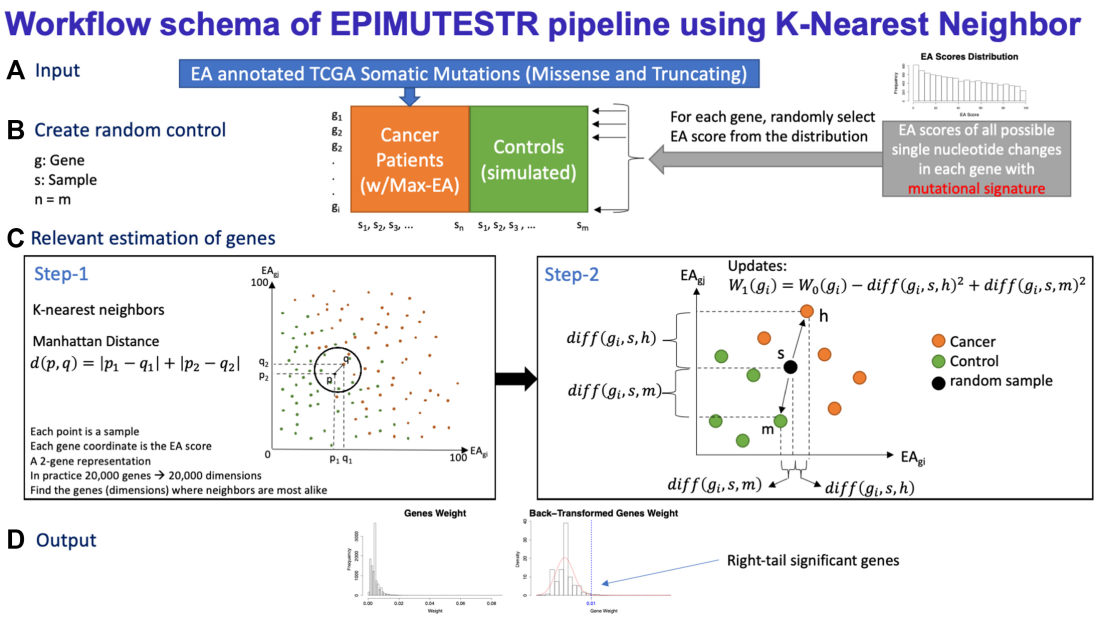
Workflow schema of EPIMUTESTR pipeline using K-Nearest Neighbor
A Input
EA annotated TCGA Somatic Mutations (Missense and Truncating)
EA Scores Distribution
B Create random control
g: Gene
s: Sample
n = m
g_1
g_2
g_2
g_i
Cancer Patients (w/Max-EA)
Controls (simulated)
$s_1, s_2, s_3, \ldots$ s_n $s_1, s_2, s_3, \ldots$ s_m
For each gene, randomly select EA score from the distribution
EA scores of all possible single nucleotide changes in each gene with mutational signature
C Relevant estimation of genes
Step-1
K-nearest neighbors
Manhattan Distance
$d(p,q) = |p_1 - q_1| + |p_2 - q_2|$
Each point is a sample
Each gene coordinate is the EA score
A 2-gene representation
In practice 20,000 genes → 20,000 dimensions
Find the genes (dimensions) where neighbors are most alike
EA_{gj}
100
q_2
p_2
0
p_1 q_1
100
EA_{gi}
Step-2
Updates:
$W_1(g_i) = W_0(g_i) - diff(g_i,s,h)^2 + diff(g_i,s,m)^2$
EA_{gj}
$diff(g_i,s,h)$
$diff(g_i,s,m)$
h
s
m
● Cancer
● Control
● random sample
EA_{gi}
$diff(g_i,s,m)$
$diff(g_i,s,h)$
D Output
Genes Weight
Back-Transformed Genes Weight
Right-tail significant genes

Figure 1. Workflow schema of EPIMUTESTR that is based on k-NN method and developed in four parts. (**A**) Input: missense and truncating mutations from standard MC3 dataset annotated by EA method. (**B**) Create random control: considering maximum EA score among all mutations in a gene as a training feature and creating synthetic control by random selection of EA scores from the background mutations that consists of all possible single nucleotide changes for each transcript (distribution plot at the top right). The orange squares indicate the cases samples, and the green squares indicate the controls samples. (**C**) Relevant estimation of genes: using k-NN algorithm on the training data, where in step-1 a random sample with k nearest neighbors around it are selected and in step-2, starting zero weight for each gene, depends on whether the random sample (black circle) is in the case class (orange circles) or control class (green circles) the difference of normalized EA score will be penalized by $-diff(g_i,\ s,\ h)$ or awarded by $diff(g_i,\ s,\ m)$ to weigh the genes. (**D**) Final genes weights are a non-gaussian distribution that using square root transformation, we back-transform to the Gaussian distribution and select the significant genes at the right tail using cumulative distribution function.

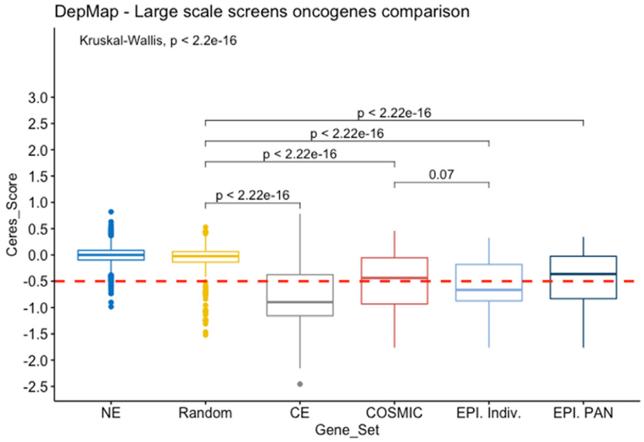

Figure 2. Comparison of DepMap Ceres Score for EPIMUTESTR candidates, COSMIC genes, core essential (CE) genes, random and non-essential (NE) genes. EPIMUTESTR oncogenes splitted to oncogenes in all individual cancers (EPI. Indiv.) and oncogenes in pan-cancer (EPI. PAN). The y-axis shows the Ceres Score from non-essential to essential (–3, +1). And the horizontal dash line indicates the critical cut-off point (–0.5) for oncogenes in DepMap. The significance p-values (t-test) indicate the difference between each pair. And the overall Kruskal–Wallis p-value indicates the significance of comparison.

tial gene, non-essential gene, and a random gene in Pan-Cancer (Supplementary Figure S1). The random gene and non-essential gene distributions have a similar characteristic with bias to low EA scores (exponential distribution), whereas the essential genes bias to high EA scores (reverse exponential distribution). Since the EPIMUTESTR algorithm weighs the essential genes higher than normal genes, they tend to be in the significant genes list. To show this, we calculated hypergeometric test between gold standard genes and essential/non-essential genes to compare the overlap genes, and we noticed a significant enrichment between essential genes and gold standard genes (p-value: 2.3e–19).

Robustness in downsampling

In order to evaluate the robustness of the EPIMUTESTR pipeline, we next performed a downsampling analysis for several tumor types (LUAD, BRCA, BLCA, HNSC, UCEC and OV). We compared EPIMUTESTR with dNdScv (7) which is a well-known method to detect cancer driver genes under positive selection. For this, we iteratively removed 5% of random samples and each time ran EPIMUTESTR and dNdScv to investigate the significance of the number of driver genes. In each step, we replicated the 5% random removal ten times to calculate the error bar (Figure 3) and we observed that both methods can detect the core genes throughout the downsampling flow (Supplementary Table S3) but EPIMUTESTR is able to detect more cancer driver genes. This analysis suggests that EPIMUTESTR is very robust to downsampling and does not need very large patient cohorts to work.

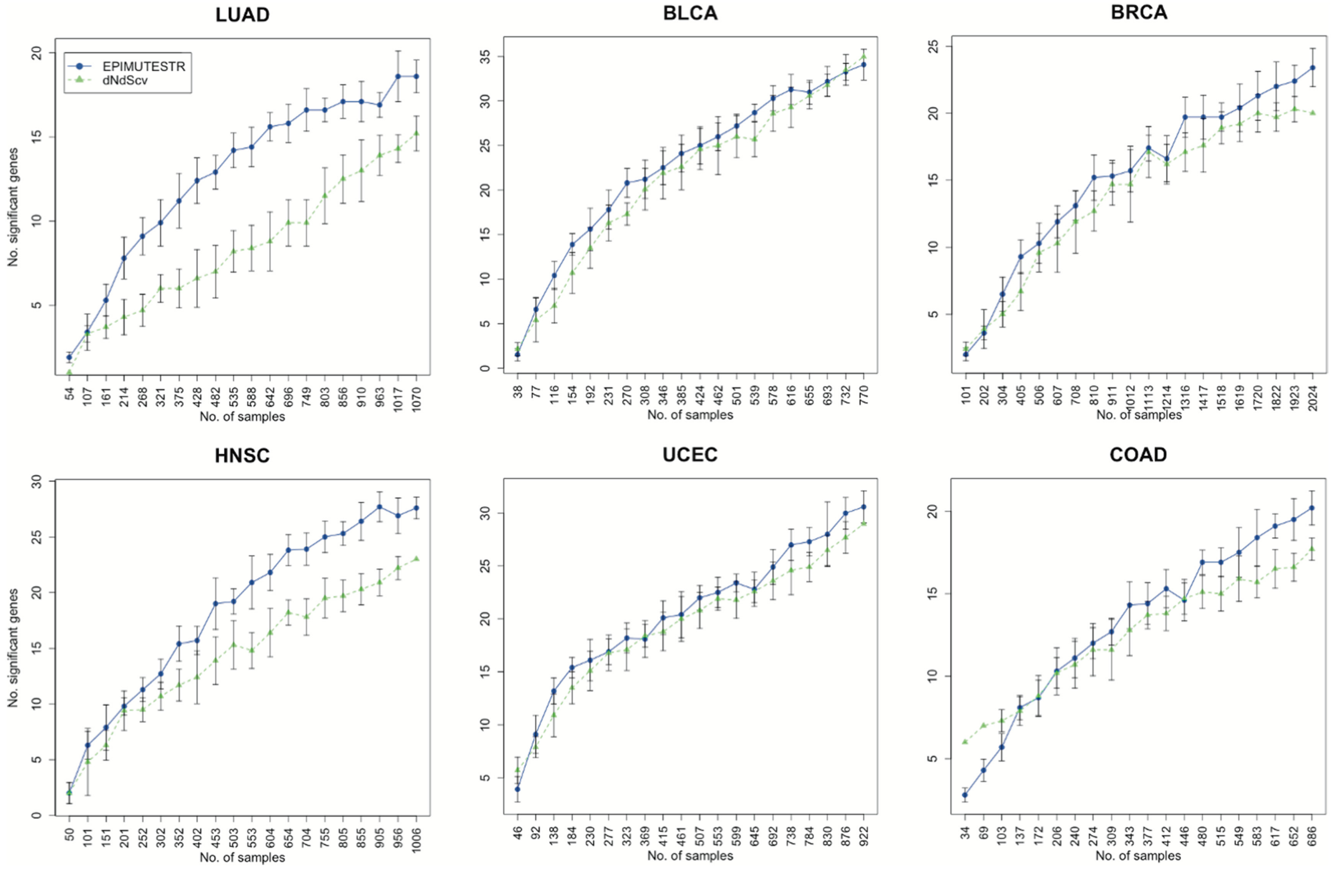

Figure 3. Comparison of robustness between EPIMUTESTR (blue curve) and dNdScv (green curve) by downsampling. The x-axis indicates the number of samples and y-axis indicates the number of overlap genes between candidate genes and manually created list of known cancer driver genes from ten resources. The vertical lines in each step indicate the error bars.

Well-connected to human curated network of cancer genes and PubMed literature searches

We further confirmed an association between EPIMUTESTR genes and cancer with two other types of evidence: using DOSE (81) to enrich for known cancer genes from NCG (80) and occurring in at least 10 PubMed cancer publications (Figure 4A). For the remaining 250 unidentified genes, we found 56 genes supported by both types of evidence, 91 genes supported by one type of evidence, and 103 genes remained unidentified by any of the evidence sources (Figure 4B). We compared the enrichment analysis with all genes from PanCancer (18,696) for NCG (Fisher exact test p-value < 3.8e–116) and PubMed (Fisher exact test p-value < 1.9e–43). Overall, these results suggest that the most EPIMUTESTR candidate genes are connected to the human curated network of cancer genes and the PubMed literature.

Methodological control: EPIMUTESTR outperformed state-of-the-art cancer driver prediction methods and is robust

There are many lists of candidate cancer driver genes generated by various cancer driver prediction methods. In order to compare the performance of EPIMUTESTR to these state-of-the-art methods, we used the assessment criteria suggested by Tokheim *et al.* (26). We used their prepared dataset consisting of 7916 cancer patients affected by 34 different types of specific cancers, and accompanying software to evaluate through standardized tests set by Tokheim *et al.* (26) the performance of EPIMUTESTR against nine state-of-the-art methods in (MutsigCV (2), ActiveDriver (76), MuSiC (77), OncodriveClust (41), OncodriverFM (78), OncodriverFML (5), Tumor suppressor and Oncogenes (TUSON) (79). These criteria were the fraction of overlap with CGC (Figure 5A), agreement between methods (Figure 5B), p-values deviation from standard uniform distribution (Figure 5C), and consistency of top genes over the two random splits of datasets by preserving the same proportion of specific cancer patients in each split (Figure 5D). The EPIMUTESTR overall performance was the best among other methods in Cancer Gene Census (CGC) from COSMIC, consensus genes, and the second best in consistency of subsampling in all the methods; and p-value deviation (Figure 5E). These evaluations and comparisons demonstrate that the EPIMUTESTR outperformed the methods in (26) evaluation guidelines, including the machine learning methods.

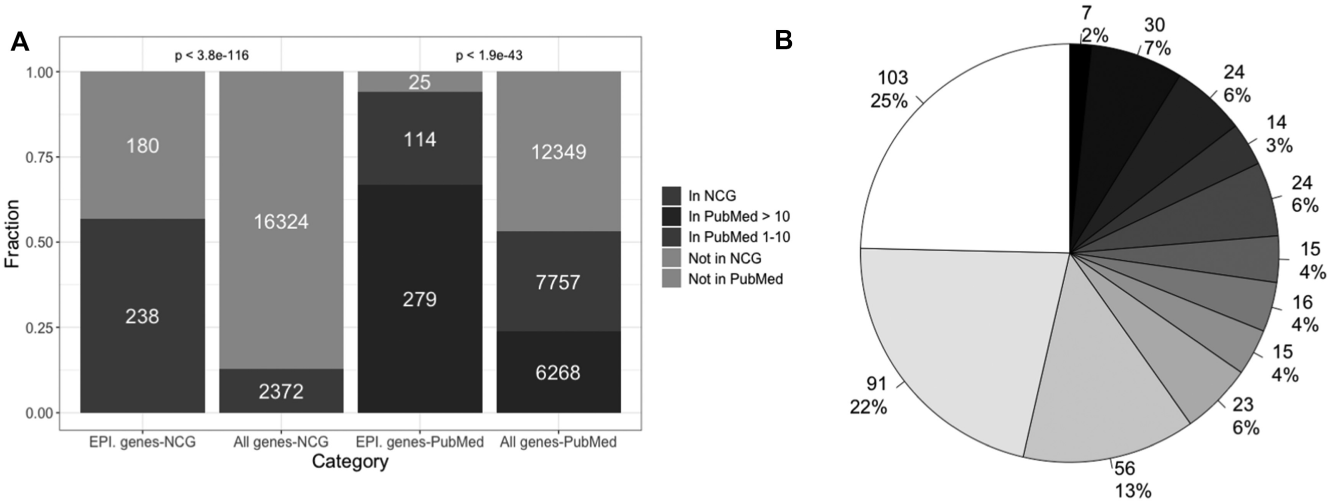

Figure 4. Novel genes well-connected to human curated network of cancer genes and PubMed literature searches. (**A**) Comparing EPIMUTESTR candidate genes and entire genes in Network of Cancer Genes (NCG) and PubMed literature search. The dark gray color indicates the number of genes that found in NCG and PubMed. The light gray color indicates the number of genes that are not in NCG and PubMed. For PubMed, we also looked at the number of genes in 1–10 publications. The significant *p*-value (fisher exact test) compares the two bar plots. (**B**) Confidence of cancer association for 418 EPIMUTESTR candidate genes.

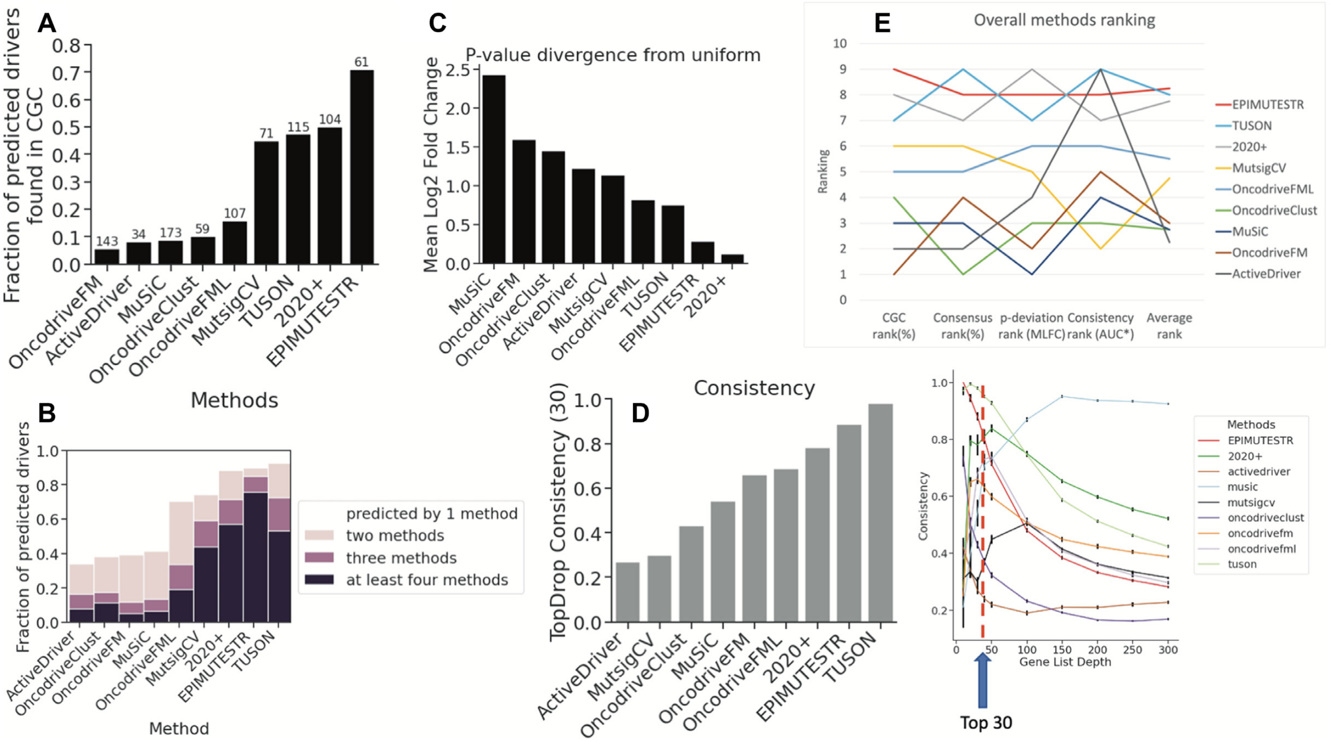

Figure 5. Panel of comparison for state-of-the-art cancer driver prediction methods (Tokheim *et al.*, 2016). (**A**) Cancer Gene Consensus (CGC) overlap: the bar plot compares the number of overlap genes between each method and CGC genes (downloaded 29 March 2021). The y-axis indicates the fraction of driver genes ($q < 0.1$) with CGC. Number of overlapped driver genes indicated at the top of each bar. (**B**) Method consensus: the number of agreements between methods. The black color indicates the number of agreements that is at least in four methods, purple color indicates the number of agreements in three methods, yellow color indicates the number of agreements in two methods, and the white color indicates the predicted by one method. The y-axis of bar plot is calculated based on the fraction of predicated genes over all genes. (**C**) The *p*-value deviation: this bar plot implicated the divergence from uniform *p*-values, measured as a mean log fold change (MLFC) between the method observed and theoretical *p*-values, to show how far the top significant genes are from each other. (**D**) Consistency bar plot to show the top drop genes consistency. We have filtered the top 30 genes across 10 random splits of PanCancer by preserving the proportion of each cancer type in each split. We repeated this process by increasing the cut-off point for selecting the top 20, 30, 40, 50, 100, 150, 200, 250 and 300 genes. EPIMUTESTR started to perform worse after we increase the top genes. (**E**) Overall ranking of methods. The x-axis indicates the four guidelines test, and the y-axis indicates the ranking. The higher the ranking the better the performance.

Table 1. Confidence of cancer association

	EPIMUTESTR (418)		
Method agreements	No. overlap	*P* value	Unidentified
All methods (7)	7	2.1e-23	381
Nine methods (37)	37	2.0e-94	
Eight methods (65)	61	4.5e-135	357
Seven methods (89)	75	4.8e-148	343
Six methods (121)	99	1.6e-179	319
Five methods (161)	114	5.7e-181	304
Four methods (208)	130	4.4e-182	288
Three methods (279)	145	2.5e-170	273
Two methods (428)	168	2.8e-144	250
One method (1358)	214	3.8e-24	204

Potential discovery/functional roles: pathway enrichment

Among all EPIMUTESTR candidate genes, there were 250 genes that were not identified in the manually created list of known cancer driver genes resources (Table 1). We reviewed the EA score distribution of these genes and systematically narrowed the focus to genes with 80% of mutations having an EA score greater than 30. We ended up with 137 novel candidate cancer driver genes (Figure 6). We used pathway enrichment assessment to characterize the biological function of these genes. We queried 137 candidate genes through Molecular Signature Data Base (MSigDB) (85) to enrich for known pathways and calculated the hypergeometric test and false discovery rate (FDR) using the Benjamini–Hochberg method to correct for multiple testing. We found 10 Reactome pathways that were enriched with FDR < 0.05 including pathways directly related to cancer such as Integrin cell surface interactions ($q < 4.0$–e02) (86), RNA polymerase II transcription ($q < 4.1$–e02) (87), and NCAM signaling for neurite out-growth (88) is potential cancer related ($q < 1.8$–e02) (Table 2).

Novel cancer driver candidates implicated by EPIMUTESTR

After EA-based filtering EPIMUTESTR identified 137 novel cancer driver gene candidates not identified by the other driver gene identification methods described in Figure 6. To demonstrate that these novel driver candidates were enriched for bona fide cancer drivers, we performed a comparison of our 137 novel candidates with a random set of 137 genes not identified by our methods or any of the previous computational methods. First, we screened the EPIMUTESTR novel candidates in the COSMIC Cancer Gene Census, an expert panel curated catalogue of over 700 genes exhibiting cancer mutational patterns and functional impact data sufficient to be labeled as cancer drivers (89). We identified 7 of our 137 EPIMUTESTR novel candidates in the Cancer Gene Census, whereas none of the random 137 genes were found in the Cancer Gene Census. The seven genes were TP63 (TP53 family member), BCL6 (lymphoid cell transcriptional repressor), ZEB1 (transcription factor mediating epithelial mesenchymal transition), ACSL3 (lipid biosynthesis enzyme), CNTNAP2 (neural cell adhesion molecule), NTRK3 (neurotropic tyrosine kinase receptor), and IKZF1 (hematopoietic transcription factor associated with chromatin remodeling).

We also determined the number of cancer-associated publications in PubMed for each of the 137 EPIMUTESTR novel candidates and the 137 random gene controls. The EPIMUTESTR candidates averaged 131.2 cancer-associated publications per gene versus 43.6 such publications per random control gene. The seven novel candidate drivers discussed above that were in the Cancer Gene Census were also characterized by high numbers of cancer-associated publications in PubMed, with three of these genes discussed in over 1000 cancer-associated publications. However, the most frequent number of cancer citations found among the novel cancer candidates was for TOP2A. TOP2A was not listed in the Cancer Gene Census but has been linked to cancer in over 3500 publications in the cancer literature. TOP2A encodes DNA topoisomerase II alpha, an enzyme that catalyzes transient breaking and rejoining of double stranded DNA, thus facilitating DNA helix winding and unwinding (90,91). TOP2A is involved in critical cellular processes such as chromosome condensation and separation, as well as DNA transcription and replication. Importantly, it is often overexpressed in multiple cancer types and is targeted by 10 approved chemotherapeutic agents (92,93). The Cancer Dependency Map (DepMap) Project lists TOP2A as a common essential gene for proliferation of many cancer cell types (94).

Thus, among our 137 novel candidates we have identified seven novel expert curated cancer genes and one highly cited oncogene not identified by previous computational methods, indicating a further enrichment for previously undiscovered cancer driver genes. Future functional and cancer genomics studies should provide additional light on whether these novel candidates are bona fide cancer drivers.

DISCUSSION

Many clinical and experimental approaches have attempted to find driver genes in cancer by simple mathematical and statistical models (1,2,4,15,25,78), but they have often been limited by mutation frequency (11–13), and epistasis effects of mutations (42–44). In addition, advanced statistical studies used machine learning approaches to optimize the somatic mutation detection in human cancer. However, recent studies revealed machine learning methods can accurately detect candidate driver genes in TCGA data, but the key point to success in these approaches is to use the right cohort as training data (3,17–27). Our approach relies on the Evolutionary Action (EA) method, to score the fitness effect of coding variants. Other methods compute related scores (95–99). However, not all such scores are the same and several reasons led us to the choice of using EA. First, the Evolutionary Action method estimates the fitness effect of mutations. Other methods, such as Polyphen2 rather offer the probability of each mutation to be deleterious or neutral, ignoring the fact that gain of function variants (as those seen in oncogenes) are neither deleterious nor neutral. Second, the overall EA scores of fitness effect in the human genome follows an exponential distribution (Supplementary Figure S1), where the outcome of other methods may be mostly bimodal or very different than the exponential distribution expected for a distance approximation (as proposed by Fisher in his geometric model of fit-

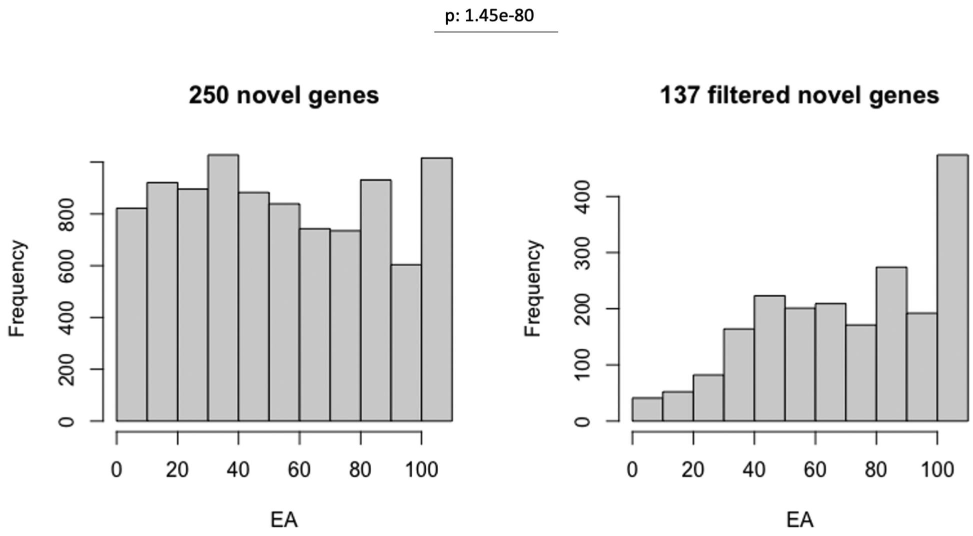

Figure 6. Filtering novel cancer driver candidates implicated by EPIMUTESTR. The left figure illustrates the EA distribution of 250 unidentified genes, and the left figure illustrates the filtered genes with EA > 30 in more than 80% of mutations. The x-axis indicates the EA scores, and y-axis indicates the frequency in both plots. The bar over 100 EA scores indicated the truncating mutations. KS test *p*-value indicates that the two distribu

Table 2. Pathway enrichment

Gene set name	No. genes in gene Set (K)	Description	No. genes in overlap (*k*)	*k*/*K*	*p*-value	FDR *q*-value
REACTOME COLLAGEN FORMATION	90	Collagen formation	5	0.0556	1.47E-05	1.25E-02
REACTOME COLLAGEN CHAIN TRIMERIZATION	44	Collagen chain trimerization	4	0.0909	1.56E-05	1.25E-02
REACTOME NERVOUS SYSTEM DEVELOPMENT	580	Nervous system development	10	0.0172	3.15E-05	1.68E-02
REACTOME MUSCLE CONTRACTION	195	Muscle contraction	6	0.0308	5.76E-05	1.85E-02
REACTOME NCAM SIGNALING FOR NEURITE OUT GROWTH	63	NCAM signaling for neurite out-growth	4	0.0635	6.51E-05	1.85E-02
REACTOME COLLAGEN DEGRADATION	64	Collagen degradation	4	0.0625	6.92E-05	1.85E-02
REACTOME COLLAGEN BIOSYNTHESIS AND MODIFYING ENZYMES	67	Collagen biosynthesis and modifying enzymes	4	0.0597	8.29E-05	1.90E-02
REACTOME INTEGRIN CELL SURFACE INTERACTIONS	85	Integrin cell surface interactions	4	0.0471	2.09E-04	4.07E-02
REACTOME PROTEIN PROTEIN INTERACTIONS AT SYNAPSES	87	Protein-protein interactions at synapses	4	0.046	2.28E-04	4.07E-02
REACTOME RNA POLYMERASE II TRANSCRIPTION	1374	RNA Polymerase II Transcription	14	0.0102	2.56E-04	4.11E-02

ness effects (100,101)). As such, we cannot use these methods without proper transformation. Third, the EA method had good performance in blind, objective contests of the CAGI community for predicting the impact of mutations (49,50). Fourth, it has been useful in molecular and clinical applications (52,102–104). Fifth, it is untrained rather than dependent on prior human clinical data (105). Last but not least, it describes well the overall functional impact and the organismal fitness outcome (106,107). We exploited the advantages of EA, a method to score mutations in protein coding sequences of the human genome, to better annotate the somatic mutations in cancer patients. Essentially, we used EA scores from human genome background mutations that produce a nearly exponential distribution to create simulated controls by random selection of the EA scores from the distribution of all possible single nucleotide changes in canonical transcripts of each gene. And we combined the cancer patients with synthetic controls and labelled them as 1's for cases and 0's for the controls. Here, we developed EPIMUTESTR, a machine learning pipeline, that takes the combined cases/controls dataset and identifies the driver genes in human cancer. We found that EPIMUTESTR accurately identifies driver genes for all 33 individual cancer types and across all 33 cancers assessed by the TCGA Pan-Cancer project, EPIMUTESTR minimizes the false positive calls and may improve specificity for the purpose of patient targeted therapeutic drugs. We chose the term epistasis because EPIMUTESTR is based on nearest-neighbor algorithm and all genes participate in samples classification. Technically, when some genes have different EA scores (lower or higher) than others, this will give additional award or penalty to the gene's weight, but the effect size depends on the distance between genes. Therefore, we believe that genes interactively affect each other and consequently gene weight as we have shown in Supplementary Figures S2 and S4.

An important aspect of our approach is the use of synthetic controls to detect infrequently mutated yet important cancer driver genes. On the other hand, simulating synthetic mutations from all possible nucleotide mutations may carry some risk of false positives. Recent studies revealed nested cross-validation chooses a more parsimonious set of features with fewer false positives (70,108,109). Therefore, we replicated the analysis with 10 different synthetic controls and considered the consensus set of genes that co-occurred in more than half of the replications. However, we note that the relatively tight error bars in the robustness curves for six tumor types in Figure 3 suggest that the risk of false positives with the synthetic control is low and that replicating the analysis with more synthetic controls may not be always helpful. Thus, we compared the discovered cancer genes over the various thresholds for each frequency. Supplementary Figure S3 indicates the overall performance of the number of gene frequencies from 10 replications of synthetic controls in which as the number of gene frequencies go up the number of overlap genes drop. On the other hand, the low and high thresholds tend to have the low and high specificity and sensitivity. Therefore, we chose threshold 0.1 based on three reasons: first, a q-value < 0.1 is a standard threshold for statistical significance; second, a tradeoff between the smaller number of unidentified cancer genes and the higher number of discovered cancer genes meet at 0.1; and third, the number of discovered cancer genes falls within the range of the number of genes identified by other methods (40–42).

Fundamentally, the k nearest neighbor algorithm has two important steps, first, generating a distance matrix across all samples, second, classify samples randomly based on the k nearest neighbor. In step two, all the genes are considered for classification, thus, we can conclude that k nearest neighbor considers the polygenic aspect. To show this, we generated a heatmap of samples by significant genes for the LUAD cancer type, where each entity refers to EA score and highlights the significance of the genes and the reason for selection in Supplementary Figure S3. Thus, given all missense and truncating somatic mutations for cancer patients and simulated controls, we were able to consider all the genes in n-dimensions (n = total number of genes) of Manhattan space to capture epistasis effect of genes (68,83). We also compared the EA scores distributions for TP53 as a tumor suppressor and BRAF as an oncogene in Supplementary Figure S4 to show different cancer drivers may impact cancer in different ways (loss of function or gain of function). We also have observed the similar impact to the gene weights as we have seen in Supplementary Figure S1. And, we noticed that p-values of these two genes in LUAD and SKCM cancers are subjective, where TP53 is the leading driver in LUAD (TP53 p-value: 3.06E–137, BRAF p-value: 1.57E–21) and BRAF is the leading driver in SKCM (BRAF p-value: 3.23E–168, TP53 p-value: 1.07E–12). Consequently, EPIMUTESTR picks up oncogenes or tumor suppressor genes subjectively and it highly depends on EA scores distribution. In addition, in order to test the robustness of the EPIMUTESTR, we reduced the number of samples by random subsampling iteratively. We iteratively removed 5% of samples by preserving the ratio of cancer patients and simulated controls in each iteration. Finally, we used our predefined gold-standard gene list and considered the overlap genes that were identified in each iteration. This analysis is illustrated in Figure 3 and shows that EPIMUTESTR is robust even at smaller sample sizes.

Despite the advantages of our approach, this study may have some limitations. First, machine learning methods are highly dependent on the input data and because our input data are somatic mutations with synthetic controls (i.e. incomplete data), the consistency assessments will fail as we reduce the power by subsampling (Figure 5D). Second, since we accept all mutations in a gene, it may reduce our sensitivity due to different pathogenic mutations. Therefore, in future directions we plan to the use UK Biobank where there is control data to calculate odds ratio of a gene. Third, gene weights calculated based on the classification of cancer patients against synthetic controls may not order top genes correctly as we rely on synthetic controls for classification, and therefore we replicated the analysis with ten different random controls. Although our approach is designed to work with somatic coding mutations and there are many factors contribute orthogonal information to the search for cancer driver genes and the success will arise, eventually, when each is properly use and then integrated with the others. Therefore, besides coding mutation scores, such as from EA, PolyPhen, or SIFT, copy number variation (30), gene expression (42), methylation (110), etc. can be used to build a multi-variate model. In the future we do plan to study performance improvements when integrating these other sources of data in our model (111).

In summary, EPIMUTESTR improves cancer driver gene identification where it finds 418 genes with high specificity (Figure 4B) across 33 different cancer types from the TCGA project. 168 candidate driver genes are well-known according to ten recent sources of reported cancer genes. Most have been identified in other sources of evidence such as cancer publications in PubMed searches and the Network of Cancer Genes. The remaining genes that are less studied are enriched in eight pathways related to cancer. Additionally, our remaining genes included many genes that are unidentified by previously reported genetic studies that are important genes in cancer pathways.

DATA AVAILABILITY

EPIMUTESTR is under MIT license and publicly available in the GitHub repository (https://github.com/LichtargeLab/EPIMUTESTR)

MC3 MAF file is publicly available at (https://gdc.cancer.gov/about-data/publications/mc3-2017)

An online version of EA scores is available for nonprofit use at (http://eaction.lichtargelab.org)

DepMap is publicly available at (https://depmap.org/)

20/20+ is publicly available at (https://github.com/KarchinLab/2020plus)

dNdScv is publicly available at (https://github.com/im3sanger/dndscv)

DOSE is publicly available at (http://dx.doi.org/10.1093/bioinformatics/btu684)

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Brigitta D. Wastuwidyaningtyas for the valuable feedback toward editing. We thank Daniel M. Konecki for helping on DepMap analysis.

FUNDING

OL gratefully acknowledges support from NIH [GM066099, AG061105, AG068214, AG074009]; S.P. Supported by a Computational Cancer Biology Training Program fellowship CPRIT [RP170593]; J.A. was supported by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science [T15LM007093]. Funding for open access charge: Research fund.

Conflict of interest statement. None declared.

REFERENCES

1. Dietlein,F., Weghorn,D., Taylor-Weiner,A., Richters,A., Reardon,B., Liu,D., Lander,E.S., van Allen,E.M. and Sunyaev,S.R. (2020) Identification of cancer driver genes based on nucleotide context. *Nat. Genet.*, **52**, 208–218.
2. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
3. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
4. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
5. Mularoni,L., Sabarinathan,R., Deu-Pons,J., Gonzalez-Perez,A. and López-Bigas,N. (2016) OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.*, **17**, 128.
6. Porta-Pardo,E. and Godzik,A. (2014) e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics*, **30**, 3109–3114.
7. Martincorena,I., Raine,K.M., Gerstung,M., Dawson,K.J., Haase,K., van Loo,P., Davies,H., Stratton,M.R. and Campbell,P.J. (2017) Universal patterns of selection in cancer and somatic tissues. *Cell*, **171**, 1029–1041.
8. Weghorn,D. and Sunyaev,S. (2017) Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.*, **49**, 1785–1788.
9. Zhao,S., Liu,J., Nanga,P., Liu,Y., Cicek,A.E., Knoblauch,N., He,C., Stephens,M. and He,X. (2019) Detailed modeling of positive selection improves detection of cancer driver genes. *Nat. Commun.*, **10**, 3399.
10. Tomasetti,C. and Vogelstein,B. (2015) Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (New York, N.Y.)*, **347**, 78–81.
11. Ahn,E.H. and Lee,S.H. (2019) Detection of low-frequency mutations and identification of heat-induced artifactual mutations using duplex sequencing. *Int. J. Mol. Sci.*, **20**, 199.
12. Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, **19**, 269–285.
13. Zhao,X., Little,P., Hoyle,A.P., Pegna,G.J., Hayward,M.C., Ivanova,A., Parker,J.S., Marron,D.L., Soloway,M.G., Jo,H. *et al.* (2019) The prognostic significance of low-frequency somatic mutations in metastatic cutaneous melanoma. *Front. Oncol.*, **8**, 584.
14. Hardy,J. and Singleton,A. (2009) Genomewide association studies and human disease. *N. Engl. J. Med.*, **360**, 1759–1768.
15. Hirschhorn,J.N. and Daly,M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
16. He,Q., He,Q., Liu,X., Wei,Y., Shen,S., Hu,X., Li,Q., Peng,X., Wang,L. and Yu,L. (2014) Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data. *Am. J. Cancer Res.*, **4**, 394–410.
17. Kourou,K., Exarchos,T.P., Exarchos,K.P., Karamouzis,M.V. and Fotiadis,D.I. (2014) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, **13**, 8–17.
18. Han,Y., Yang,J., Qian,X., Cheng,W.-C., Liu,S.-H., Hua,X., Zhou,L., Yang,Y., Wu,Q., Liu,P. *et al.* (2019) DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res.*, **47**, e45.
19. Basile,A.O. and Ritchie,M.D. (2018) Informatics and machine learning to define the phenotype. *Expert Rev. Mol. Diagn.*, **18**, 219–226.
20. Drouin,A., Letarte,G., Raymond,F., Marchand,M., Corbeil,J. and Laviolette,F. (2019) Interpretable genotype-to-phenotype classifiers with performance guarantees. *Sci. Rep.*, **9**, 4071.
21. Grinberg,N.F., Orhobor,O.I. and King,R.D. (2020) An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Mach. Learn.*, **109**, 251–277.
22. Bailey,M.H., Tokheim,C., Porta-Pardo,E., Sengupta,S., Bertrand,D., Weerasinghe,A., Colaprico,A., Wendl,M.C., Kim,J., Reardon,B. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **173**, 371–385.
23. Colaprico,A., Olsen,C., Bailey,M.H., Odom,G.J., Terkelsen,T., Silva,T.C., Olsen,A.V, Cantini,L., Zinovyev,A., Barillot,E. *et al.* (2020) Interpreting pathways to discover cancer driver genes with moonlight. *Nat. Commun.*, **11**, 69.
24. Collier,O., Stoven,V. and Vert,J.-P. (2019) LOTUS: a single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput. Biol.*, **15**, e1007381.
25. Kumar,R.D., Searleman,A.C., Swamidass,S.J., Griffith,O.L. and Bose,R. (2015) Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics*, **31**, 3561–3568.
26. Tokheim,C.J., Papadopoulos,N., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2016) Evaluating the evaluation of cancer driver genes. *PNAS*, **113**, 14330–14335.
27. Tokheim,C. and Karchin,R. (2019) CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst.*, **9**, 9–23.
28. Malebary,S.J. and Khan,Y.D. (2021) Evaluating machine learning methodologies for identification of cancer driver genes. *Sci. Rep.*, **11**, 12281.
29. Luo,P., Ding,Y., Lei,X. and Wu,F.-X. (2019) deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Frontiers in Genetics*, **10**, 13.
30. Chen,Y., Hao,J., Jiang,W., He,T., Zhang,X., Jiang,T. and Jiang,R. (2013) Identifying potential cancer driver genes by genomic data integration. *Sci. Rep.*, **3**, 3538.
31. Zhou,Y.-H. and Gallins,P. (2019) A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.*, **10**, 579.
32. Cuperlovic-Culf,M. (2018) Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites*, **8**, 4.
33. Sanchez-Vega,F., Mina,M., Armenia,J., Chatila,W.K., Luna,A., La,K.C., Dimitriadoy,S., Liu,D.L., Kantheti,H.S., Saghafinia,S.

et al. (2018) Oncogenic signaling pathways in the cancer genome atlas. *Cell*, **173**, 321–337.
34. Way,G.P., Sanchez-Vega,F., La,K., Armenia,J., Chatila,W.K., Luna,A., Sander,C., Cherniack,A.D., Mina,M., Ciriello,G. *et al.* (2018) Machine learning detects Pan-cancer ras pathway activation in the cancer genome atlas. *Cell Rep.*, **23**, 172–180.
35. Jones,S., Anagnostou,V., Lytle,K., Parpart-Li,S., Nesselbush,M., Riley,D.R., Shukla,M., Chesnick,B., Kadan,M., Papp,E. *et al.* (2015) Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.*, **7**, 283ra53.
36. Mandelker,D., Zhang,L., Kemel,Y., Stadler,Z.K., Joseph,V., Zehir,A., Pradhan,N., Arnold,A., Walsh,M.F., Li,Y. *et al.* (2017) Mutation detection in patients with advanced cancer by universal sequencing of cancer-related genes in tumor and normal DNA vs guideline-based germline testing. *JAMA*, **318**, 825–835.
37. Schrader,K.A., Cheng,D.T., Joseph,V., Prasad,M., Walsh,M., Zehir,A., Ni,A., Thomas,T., Benayed,R., Ashraf,A. *et al.* (2016) Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol.*, **2**, 104–111.
38. Wood,D.E., White,J.R., Georgiadis,A., van Emburgh,B., Parpart-Li,S., Mitchell,J., Anagnostou,V., Niknafs,N., Karchin,R., Papp,E. *et al.* (2018) A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.*, **10**, eaar7939.
39. Arbeithuber,B., Makova,K.D. and Tiemann-Boege,I. (2016) Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res.*, **23**, 547–559.
40. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer genome landscapes. *Science (New York, N.Y.)*, **339**, 1546–1558.
41. Tamborero,D., Gonzalez-Perez,A. and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
42. Ding,L., Bailey,M.H., Porta-Pardo,E., Thorsson,V., Colaprico,A., Bertrand,D., Gibbs,D.L., Weerasinghe,A., Huang,K.-L., Tokheim,C. *et al.* (2018) Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell*, **173**, 305–320.
43. van de Haar,J., Canisius,S., Yu,M.K., Voest,E.E., Wessels,L.F.A. and Ideker,T. (2019) Identifying epistasis in cancer genomes: a delicate affair. *Cell*, **177**, 1375–1383.
44. Wang,X., Fu,A.Q., McNerney,M.E. and White,K.P. (2014) Widespread genetic epistasis among cancer genes. *Nat. Commun.*, **5**, 4828.
45. Gumpinger,A.C., Lage,K., Horn,H. and Borgwardt,K. (2020) Prediction of cancer driver genes through network-based moment propagation of mutation scores. *Bioinformatics*, **36**, i508–i515.
46. Shi,K., Gao,L. and Wang,B. (2016) Discovering potential cancer driver genes by an integrated network-based approach. *Mol. BioSyst.*, **12**, 2921–2931.
47. Kobren,S.N., Chazelle,B. and Singh,M. (2020) PertInInt: an integrative, analytical approach to rapidly uncover cancer driver genes with perturbed interactions and functionalities. *Cell Syst.*, **11**, 63–74.
48. Katsonis,P. and Lichtarge,O. (2014) A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res.*, **24**, 2050–2058.
49. Katsonis,P. and Lichtarge,O. (2017) Objective assessment of the evolutionary action equation for the fitness effect of missense mutations across CAGI-blinded contests. *Hum. Mutat.*, **38**, 1072–1084.
50. Katsonis,P. and Lichtarge,O. (2019) CAGI5: objective performance assessments of predictions based on the evolutionary action equation. *Hum. Mutat.*, **40**, 1436–1454.
51. Clarke,C.N., Katsonis,P., Hsu,T.-K., Koire,A.M., Silva-Figueroa,A., Christakis,I., Williams,M.D., Kutahyalioglu,M., Kwatampora,L., Xi,Y. *et al.* (2018) Comprehensive genomic characterization of parathyroid cancer identifies novel candidate driver mutations and core pathways. *J. Endocr. Soc.*, **3**, 544–559.
52. Neskey,D.M., Osman,A.A., Ow,T.J., Katsonis,P., McDonald,T., Hicks,S.C., Hsu,T.-K., Pickering,C.R., Ward,A., Patel,A. *et al.* (2015) Evolutionary action score of TP53 identifies high-risk mutations associated with decreased survival and increased distant metastases in head and neck cancer. *Cancer Res.*, **75**, 1527–1536.
53. Osman,A.A., Neskey,D.M., Katsonis,P., Patel,A.A., Ward,A.M., Hsu,T.-K., Hicks,S.C., McDonald,T.O., Ow,T.J., Alves,M.O. *et al.* (2015) Evolutionary action score of TP53 coding variants is predictive of platinum response in head and neck cancer patients. *Cancer Res.*, **75**, 1205–1215.
54. wheeler@bcm.edu,C.G.A.R.Network.E. address: and Network,C.G.A.R. (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, **169**, 1327–1341.
55. Hsu,T.-K., Asmussen,J.K., Koire,A.M., Choi,B.-K., Gadhikar,M.A., Huh,E., Lin,C.-H., Konecki,D.M., Kim,Y.W., Pickering,C. *et al.* (2022) A general calculus of fitness landscapes finds genes under selection in cancers. *Genome Res.*, https://doi.org/10.1101/gr.275811.121.
56. Kononenko,I. (1994) Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano,F. and de Raedt,L. (eds). *Machine Learning: ECML-94*. Springer, Berlin, Heidelberg, pp. 171–182.
57. Robnik-Šikonja,M. and Kononenko,I. (2003) Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.*, **53**, 23–69.
58. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., Hess,J., Ma,S., Chiotti,K.E., McLellan,M. *et al.* (2018) Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.*, **6**, 271–281.
59. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
60. Cibulskis,K., Lawrence,M.S., Carter,S.L., Sivachenko,A., Jaffe,D., Sougnez,C., Gabriel,S., Meyerson,M., Lander,E.S. and Getz,G. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.
61. Larson,D.E., Harris,C.C., Chen,K., Koboldt,D.C., Abbott,T.E., Dooling,D.J., Ley,T.J., Mardis,E.R., Wilson,R.K. and Ding,L. (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, **28**, 311–317.
62. Chapman,M.A., Lawrence,M.S., Keats,J.J., Cibulskis,K., Sougnez,C., Schinzel,A.C., Harview,C.L., Brunet,J.-P., Ahmann,G.J., Adli,M. *et al.* (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature*, **471**, 467–472.
63. Ye,K., Wang,J., Jayasinghe,R., Lameijer,E.-W., McMichael,J.F., Ning,J., McLellan,M.D., Xie,M., Cao,S., Yellapantula,V. *et al.* (2016) Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.*, **22**, 97–104.
64. Radenbaugh,A.J., Ma,S., Ewing,A., Stuart,J.M., Collisson,E.A., Zhu,J. and Haussler,D. (2014) RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One*, **9**, e111516.
65. Fan,Y., Xi,L., Hughes,D.S.T., Zhang,J., Zhang,J., Futreal,P.A., Wheeler,D.A. and Wang,W. (2016) MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, **17**, 178.
66. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
67. Stanfill,B., Reehl,S., Bramer,L., Nakayasu,E.S., Rich,S.S., Metz,T.O., Rewers,M., Webb-Robertson,B.-J. and Group,T.S. (2019) Extending classification algorithms to case-control studies. *Biomed. Eng. Comput. Biol.*, **10**, 1179597219858954.
68. Urbanowicz,R.J., Olson,R.S., Schmitt,P., Meeker,M. and Moore,J.H. (2018) Benchmarking relief-based feature selection methods for bioinformatics data mining. *J. Biomed. Inform.*, **85**, 168–188.
69. Le,T.T., Urbanowicz,R.J., Moore,J.H. and McKinney,B.A. (2019) STatistical inference relief (STIR) feature selection. *Bioinformatics*, **35**, 1358–1365.
70. Parvandeh,S., Yeh,H.-W., Paulus,M.P. and McKinney,B.A. (2020) Consensus features nested cross-validation. *Bioinformatics*, **36**, 3093–3098.
71. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
72. Meyers,R.M., Bryan,J.G., McFarland,J.M., Weir,B.A., Sizemore,A.E., Xu,H., Dharia,N.v, Montgomery,P.G., Cowley,G.S., Pantel,S. *et al.* (2017) Computational correction of copy number

effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.

73. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G., Sonkin,D. *et al.* (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
74. Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, **528**, 84–87.
75. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
76. Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.
77. Dees,N.D., Zhang,Q., Kandoth,C., Wendl,M.C., Schierding,W., Koboldt,D.C., Mooney,T.B., Callaway,M.B., Dooling,D., Mardis,E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
78. Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.
79. Davoli,T., Xu,A.W., Mengwasser,K.E., Sack,L.M., Yoon,J.C., Park,P.J. and Elledge,S.J. (2013) Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, **155**, 948–962.
80. Repana,D., Nulsen,J., Dressler,L., Bortolomeazzi,M., Venkata,S.K., Tourna,A., Yakovleva,A., Palmieri,T. and Ciccarelli,F.D. (2019) The network of cancer genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.*, **20**, 1.
81. Yu,G., Wang,L.-G., Yan,G.-R. and He,Q.-Y. (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.
82. Bolón-Canedo,V., Sánchez-Maroño,N. and Alonso-Betanzos,A. (2013) A review of feature selection methods on synthetic data. *Knowl. Inform. Syst.*, **34**, 483–519.
83. Urbanowicz,R.J., Meeker,M., la Cava,W., Olson,R.S. and Moore,J.H. (2018) Relief-based feature selection: introduction and review. *J. Biomed. Inform.*, **85**, 189–203.
84. Tsherniak,A., Vazquez,F., Montgomery,P.G., Weir,B.A., Kryukov,G., Cowley,G.S., Gill,S., Harrington,W.F., Pantel,S., Krill-Burger,J.M. *et al.* (2017) Defining a cancer dependency map. *Cell*, **170**, 564–576.
85. Liberzon,A., Birger,C., Thorvaldsdóttir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
86. Cooper,J. and Giancotti,F.G. (2019) Integrin signaling in cancer: mechanotransduction, stemness, epithelial plasticity, and therapeutic resistance. *Cancer Cell*, **35**, 347–367.
87. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
88. Seidenfaden,R., Krauter,A., Schertzinger,F., Gerardy-Schahn,R. and Hildebrandt,H. (2003) Polysialic acid directs tumor cell growth by controlling heterophilic neural cell adhesion molecule interactions. *Mol. Cell. Biol.*, **23**, 5908–5918.
89. Sondka,Z., Bamford,S., Cole,C.G., Ward,S.A., Dunham,I. and Forbes,S.A. (2018) The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer*, **18**, 696–705.
90. Lee,J.H. and Berger,J.M. (2019) Cell cycle-dependent control and roles of DNA topoisomerase II. *Genes*, **10**, 859.
91. Nitiss,J.L. (2009) DNA topoisomerase II and its growing repertoire of biological functions. *Nat. Rev. Cancer*, **9**, 327–337.
92. Ali,Y. and Abd Hamid,S. (2016) Human topoisomerase II alpha as a prognostic biomarker in cancer chemotherapy. *Tumor Biol.*, **37**, 47–55.
93. Nitiss,J.L. (2009) Targeting DNA topoisomerase II in cancer chemotherapy. *Nat. Rev. Cancer*, **9**, 338–350.
94. McFarland,J.M., Ho,Z.V., Kugener,G., Dempster,J.M., Montgomery,P.G., Bryan,J.G., Krill-Burger,J.M., Green,T.M., Vazquez,F., Boehm,J.S. *et al.* (2018) Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.*, **9**, 4610.
95. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
96. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
97. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
98. Ioannidis,N.M., Rothstein,J.H., Pejaver,V., Middha,S., McDonnell,S.K., Baheti,S., Musolf,A., Li,Q., Holzinger,E., Karyadi,D. *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.*, **99**, 877–885.
99. Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.
100. Orr,H.A. (2005) The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.*, **6**, 119–127.
101. Edwards,A.W. (2000) The genetical theory of natural selection. *Genetics*, **154**, 1419–1426.
102. Chun,Y.S., Passot,G., Yamashita,S., Nusrat,M., Katsonis,P., Loree,J.M., Conrad,C., Tzeng,C.-W.D., Xiao,L., Aloia,T.A. *et al.* (2019) Deleterious effect of RAS and evolutionary High-risk TP53 double mutation in colorectal liver metastases. *Ann. Surg.*, **269**, 917–923.
103. Kanagal-Shamanna,R., Montalban-Bravo,G., Katsonis,P., Sasaki,K., Class,C.A., Jabbour,E., Sallman,D., Hunter,A.M., Benton,C., Chien,K.S. *et al.* (2021) Evolutionary action score identifies a subset of TP53 mutated myelodysplastic syndrome with favorable prognosis. *Blood Cancer J.*, **11**, 52.
104. Cea-Rama,I., Coscolín,C., Katsonis,P., Bargiela,R., Golyshin,P.N., Lichtarge,O., Ferrer,M. and Sanz-Aparicio,J. (2021) Structure and evolutionary trace-assisted screening of a residue swapping the substrate ambiguity and chiral specificity in an esterase. *Comput. Struct. Biotechnol. J.*, **19**, 2307–2317.
105. Grimm,D.G., Azencott,C.-A., Aicheler,F., Gieraths,U., MacArthur,D.G., Samocha,K.E., Cooper,D.N., Stenson,P.D., Daly,M.J., Smoller,J.W. *et al.* (2015) The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, **36**, 513–523.
106. Kim,Y.W., Al-Ramahi,I., Koire,A., Wilson,S.J., Konecki,D.M., Mota,S., Soleimani,S., Botas,J. and Lichtarge,O. (2021) Harnessing the paradoxical phenotypes of APOE ε2 and APOE ε4 to identify genetic modifiers in Alzheimer's disease. *Alzheimers Dementia*, **17**, 831–846.
107. Amanda,K., Panagiotis,K., Won,K.Y., Christie,B., J,W.S. and Olivier,L. (2021) A method to delineate de novo missense variants across pathways prioritizes genes linked to autism. *Sci. Transl. Med.*, **13**, eabc1739.
108. Parvandeh,S. and McKinney,B.A. (2019) EpistasisRank and epistasiskatz: interaction network centrality methods that integrate prior knowledge networks. *Bioinformatics*, **35**, 2329–2331.
109. Parvandeh,S., Poland,G.A., Kennedy,R.B. and McKinney,B.A. (2019) Multi-Level model to predict antibody response to influenza vaccine using gene expression interaction network feature selection. *Microorganisms*, **7**, 79.
110. Pan,H., Renaud,L., Chaligne,R., Bloehdorn,J., Tausch,E., Mertens,D., Fink,A.M., Fischer,K., Zhang,C., Betel,D. *et al.* (2021) Discovery of candidate DNA methylation cancer driver genes. *Cancer Discov.*, **11**, 2266.
111. Althubaiti,S., Karwath,A., Dallol,A., Noor,A., Alkhayyat,S.S., Alwassia,R., Mineta,K., Gojobori,T., Beggs,A.D., Schofield,P.N. *et al.* (2019) Ontology-based prediction of cancer driver genes. *Sci. Rep.*, **9**, 17405.