Bioinformatics, 34(23), 2018, 3975–3982
doi: 10.1093/bioinformatics/bty467
Advance Access Publication Date: 15 June 2018
Original Paper

OXFORD

Genome analysis

Prioritizing predictive biomarkers for gene essentiality in cancer cells with mRNA expression data and DNA copy number profile

Yuanfang Guan[1,*,†], Tingyang Li[1,†], Hongjiu Zhang[1,†], Fan Zhu[2] and Gilbert S. Omenn[1,3,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, USA, [2]Key Laboratory of Big Data and Intelligent Computing, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China and [3]Departments of Internal Medicine and Human Genetics and School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on July 26, 2017; revised on May 15, 2018; editorial decision on June 8, 2018; accepted on June 12, 2018

Abstract

Motivation: Finding driver genes that are responsible for the aberrant proliferation rate of cancer cells is informative for both cancer research and the development of targeted drugs. The established experimental and computational methods are labor-intensive. To make algorithms feasible in real clinical settings, methods that can predict driver genes using less experimental data are urgently needed.
Results: We designed an effective feature selection method and used Support Vector Machines (SVM) to predict the essentiality of the potential driver genes in cancer cell lines with only 10 genes as features. The accuracy of our predictions was the highest in the Broad-DREAM Gene Essentiality Prediction Challenge. We also found a set of genes whose essentiality could be predicted much more accurately than others, which we called Accurately Predicted (AP) genes. Our method can serve as a new way of assessing the essentiality of genes in cancer cells.
Availability and implementation: The raw data that support the findings of this study are available at Synapse. https://www.synapse.org/#! Synapse: syn2384331/wiki/62825. Source code is available at GitHub. https://github.com/GuanLab/DREAM-Gene-Essentiality-Challenge.
Contact: gyuanfan@umich.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancers result from the mutations accumulated during the lifetime of the patients and are characterized by fast and uncontrollable cell proliferation. These mutations can be sub-classified as either driver mutations or passenger mutations (Stratton *et al.*, 2009; Vogelstein *et al.*, 2013). Finding cancer driver mutations that are responsible for the aberrant proliferation rate of cancer cells is informative for both cancer research and the development of targeted drugs. One of the frequently used approaches is loss-of-function screening (Cowley *et al.*, 2014) where the cancer cells are infected with a large pool of shRNAs. Driver genes can thus be identified according to the changes in the proliferation rate of cancer cells. However, due to the limitation in time and resources, it is impractical to use whole-genome RNAi screening as a routine diagnostic test for patients. More practical ways of finding sample-specific driver genes are needed. So far, various *in silico* driver gene discovery methods have been published, such as Helios (Sanchez-Garcia *et al.*, 2014), CHASM (Carter *et al.*, 2009) and OncoIMPACT (Bertrand *et al.*, 2015). These powerful algorithms share a common overlooked limitation, which is the overwhelming model complexity. Most methods are using a staggering number of experimental observations as features to make predictions, which hinders their application in real clinical settings. An algorithm that is able to use a small set of features to make sample-specific cancer driver gene prediction is urgently needed.

The Broad-DREAM Gene Essentiality Prediction Challenge (https://www.synapse.org/#! Synapse: syn2384331/wiki/) was co-organized by the Broad institute, the DREAM Initiative and SAGE Bio-networks. It aimed to gather researchers in bioinformatics and biomedical informatics to advance the methods of discovering driver genes for cancer cells and provide insights for the development of targeted cancer therapies. This challenge provided a large-scale genome-wide RNAi-mediated cancer cell line screen dataset (Cowley *et al.*, 2014) and asked the participants to design novel algorithms to predict the essentiality of genes in cancer cell lines. Here, gene essentiality (or dependency) refers to the importance of a given gene for the proliferation of a given cancer cell line. The gene expression and copy number profiles of the cell lines were provided by Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012) for feature construction. The sub-challenge 2 (SC2) of the DREAM competition asked the participants to use no >10 features for essentiality prediction. We developed an effective machine learning algorithm that makes cell line-level gene essentiality prediction with only 10 features, and won the first place in SC2 of the DREAM competition. We designed an innovative feature selection method and employed support vector machine (SVM) as the base learner. The feature selection procedure leveraged the correlation between the predictors and the gene essentiality on two complementary scales (global and local scale). Eventually, nine most predictive gene expression features and one copy number feature were selected specifically for each potential cancer driver gene. We found that many previously confirmed biomarkers were re-identified by our method, which supports the biological significance of our feature selection method. The algorithm presented in this manuscript has the potential to serve as a new way of assessing the essentiality of genes in cancers.

2 Materials and methods

2.1 Data acquisition

Gene essentiality scores of 14 738 genes in 105 human cancer cell lines were provided by Project Achilles, and were downloaded from the DREAM challenge website. The essentiality scores were obtained with a genome-wide loss-of-function shRNA screening assay (Cowley *et al.*, 2014). First, ~98 000 shRNAs targeting ~17 000 genes were lentivirally delivered to cancer cells at an MOI (Multiplicity of Infection) of 0.3. Each cell would receive either zero or one shRNA. Cells were harvested after 16 population doublings or 40 days in culture, whichever came first. The essentiality of genes was measured by the relative abundance of shRNA in these cells with regard to the initial shRNA pool using next generation sequencing (NGS) technology. The raw read counts obtained by NGS were first normalized with the following function:

$$\mathrm{N} = \frac{\mathrm{R}}{\mathrm{T} \times 10^6} \tag{1}$$

where R is the raw read count for each shRNA and T is the total raw read counts.

Then, the normalized read counts were log2 transformed:

$$Nt = \log[N+1] \tag{2}$$

Quality control was performed on the normalized and transformed read counts using GenePattern module (Reich *et al.*, 2006), which enabled the removal of overlapped shRNAs and shRNAs that had low abundance in DNA reference. GenePattern module was then used to calculate an shRNA-level score for each cell line, and map shRNAs to genes. DEMETER (Tsherniak *et al.*, 2017) was used to calculate a gene-level essentiality score (ES) for each cell line. DEMETER decomposes the effect of an shRNA into a linear combination of the gene effect (on-target effect) and seed effect (off-target effect). The essentiality score, which is a unit-free coefficient in the model, represents the relative strength of the on-target effect. For a given gene, ES= −k(negative) means the essentiality is k standard deviations more dependent than the average essentiality in all cell lines, while positive ES scores indicate lower-than-average essentiality. In sum, lower ES indicate higher essentiality (Supplementary Fig. S1A).

Gene expression levels of 18 960 genes (CCLE-EXP genes) and copy number profiles of 23 288 genes (CCLE-CN genes) in 149 cell lines (105 training, 44 testing) were provided by CCLE and downloaded from the DREAM challenge website. The DNA copy numbers were measured by a genome-wide human Affymetrix SNP array 6.0. The mRNA expression levels were measured using Affymetrix Human Genome U133 Plus 2.0 arrays. The raw data were processed as described in a previously published paper (Barretina *et al.*, 2012). A gene list containing 2647 prioritized genes (PR genes) was downloaded from the DREAM challenge website. The prioritized gene list overlapped with CCLE-EXP genes and CCLE-CN genes but was not a subset of them. In other words, not all PR genes had available expression data or copy number data. The PR gene list was a subset of Achilles genes, which in other words means every PR gene had available Achilles gene essentiality scores in the 105 training cell lines. The essentiality scores of PR genes in the 44 testing cell lines are hidden at the model development stage. A detailed description of the DREAM challenge and the generation of the aforementioned datasets is available online at https://www.synapse.org/#! Synapse: syn2384331/wiki/62826. The community-authored summary has been published (Cell Systems, Nov 2017).

2.2 Selection of ten predictors

The expression features were selected as shown in Figure 1. First, the Pearson correlation coefficients between the Achilles essentiality scores of every PR gene and expression levels of every CCLE-EXP gene were calculated using the canonical function:

$$\text{Pearson correlation score} = \frac{\text{covariance}(\mathrm{x},\mathrm{y})}{\text{deviation}(\mathrm{x}) \times \text{deviation}(\mathrm{y})} \tag{3}$$

where vector x is the expression levels of a CCLE-EXP gene in all training cell lines, and vector y is the essentiality scores of a certain PR gene in all training cell lines. Thus, the total number of Pearson correlation scores is:

$$\begin{aligned} &18\,960 \ (\text{CCLE EXP genes}) \times 2647 \ (\text{PR genes}) \\ &= 50\,187\,120 \ (\text{Pearson correlation scores}) \end{aligned} \tag{4}$$

These scores measured the significance of the correlation between the expression levels of CCLE-EXP genes and the essentiality scores of PR genes. We named them local scores as they represented 1 versus 1 association between each feature and target pair. Then we calculated global scores as a complement to the local scores to represent the correlation between a certain CCLE-EXP gene and all PR genes (1 versus all association). To obtain the global scores, first

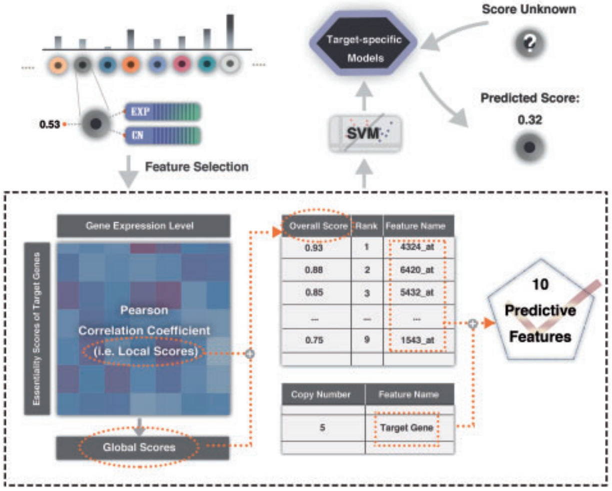

Fig. 1. Method pipeline. In training datasets, the known information included gene expression levels, copy number profile and Achilles essentiality scores of the prioritized genes in the training cell lines. After feature selection, 10 features, including nine expression features and one copy number feature, were selected for each prioritized gene. Then SVM was performed to build gene-specific models using only 10 features. These gene-specific models were then used to predict the essentiality of genes in the training cancer cell lines

we ranked CCLE-EXP genes according to their local scores for each PR gene. Then we counted how many times a CCLE-EXP gene appears in top 10 of the ranked local scores. The cutoff '10' was decided empirically. This yielded 18 960 non-negative integers which were $\leq$2647. These numbers were used to calculate the global scores using the following formula.

$$\text{Global scores} = \frac{\text{count of occurrence in top 10}}{\max\{\text{count of occurrence in top 10}\}} \tag{5}$$

Here the denominator is a constant and can be viewed as a scaling factor (Supplementary Fig. S1B). The global scores are thus all on the scale of 0 to 1. The local scores range from -1 to 1. Local scores are different for each CCLE-EXP & PR gene pair. Global score is a length of 18 960 vector corresponding to the CCLE-EXP genes, or, in other words, a rank-one matrix corresponding to the CCLE-EXP & PR gene pairs.

To combine the global and local correlation information, we calculated the weighted average of the global and local scores using the following formula, and optimized the weight (α) via cross-validation. This naive formula is chosen to avoid heavy calculation process. It is also a common practice in machine learning to use weighted averages to mix information coming from different resources.

$$\begin{aligned}\text{Combined correlation score} = (1-\alpha) \times \text{local score}(x, y) + \alpha \\ \times \text{global score}(x)\end{aligned} \tag{6}$$

i.e. for each PR gene y, we give a local score of local (x, y) for the correlation with the PR gene and a global score for x, for the global frequency of this feature over all PR genes. The alpha score is determined through a grid search (Supplementary Fig. S3) in cross-validation, and reflects the relative importance of the local model and the global model. The CCLE-EXP genes that have the nine largest combined correlation scores were then included as predictive features. If the copy number of the PR gene itself is available, it will be selected as the 10th feature; if the copy number was not available, we use expression data of the top 10 CCLE-EXP genes as 10 predictive features.

2.3 Build the gene-specific prediction model

We built gene-specific models so that each model predicts essentiality scores for one specific PR gene in all testing cell lines (Supplementary Fig. S2). The target values are Achilles gene essentiality scores of all PR genes in the training cell lines. The essentiality scores were first normalized for each PR gene using the following formula:

$$\text{Scaled Score} = \frac{\text{Original Score} - \min}{\max - \min} \tag{7}$$

where min is the minimum Achilles score for a PR gene, and max is the maximum Achilles score for a PR gene. After scaling, all essentiality scores fall in the range 0–1, and the maximum essentiality scores for all PR genes are shifted to 1.

The features in the input data were expression levels of nine CCLE-EXP genes, plus the copy number of the interested PR gene in each of 105 cell lines. The input file for each PR gene was a 105 (training cell lines) $\times$ (10 features) matrix targeting to predict the gene essentiality score. Similarly, the testing data for each PR gene was a 44 (testing cell lines) $\times$ (10 features) matrix. The models were then used to make predictions for the essentiality scores of PR genes in the testing cell lines.

To avoid overfitting, we used repeated 5-fold cross-validation (CV) to optimize our model. In each round of CV, the whole dataset was evenly divided into five parts (i.e. 21 cell lines each). We iteratively withheld one part of the dataset, used the remaining four parts to train our regression model and made predictions on the withheld data. The test was repeated five times to reduce the variance of our performance estimation. After we obtained the predicted gene essentiality scores, Spearman coefficients between the predicted scores and the Achilles essentiality scores of the 2647 PR genes in 21 cell lines were calculated. The mean of 2647 Spearman scores was calculated, which would be the performance of a single round of cross-validation. As described before, we performed five rounds of 5-fold validation; thus, in total we got $5 \times 5 = 25$ scores. The performance of the current model was represented by the mean of these 25 scores.

3 Results

3.1 Searching predictive features by leveraging the global and local prediction power

In the gene essentiality DREAM challenge, we developed a model to predict the essentiality scores of 2647 prioritized genes (PR genes) in 44 testing cell lines. The training data comprise the Achilles gene essentiality scores of the 2647 PR genes in 105 cell lines (Fig. 1). The predictors in the model were constructed from the gene expression profiles and copy number profiles of the 105 cell lines.

To simplify feature space while maintaining the prediction power, we selected 10 most predictive features for each PR gene. The top nine CCLE-EXP genes whose expression levels were in strong correlation with the Achilles essentiality scores of PR genes were selected as predictive features. The strength of the correlations was measured by the combined correlation scores which were the weighted summation of global scores and local scores. The weights assigned to global and local scores were controlled by the parameter α. After cross validation, 0.7 is selected as the optimal value for α (CV results shown in Supplementary Fig. S3A). The CV results confirmed that a combination of the information from two scales (global/local) has more predictive power than any individual scale. The global scores represent the relative frequency of having a high local score and can be viewed as a correction parameter for local scores that introduces generalizability. On the other hand, local scores add specificity to global scores.

Apart from expression features, we used the copy number of the PR gene itself as the 10th feature. However, not all PR genes had available copy number profile. In these cases, we used top 10 instead of top nine CCLE-EXP genes as predictive features. These 10 features were then used to build models for each PR gene. We compared the performance of five alternative machine learning algorithms using 5-fold cross-validation (CV results shown in Supplementary Fig. S3B), and chose support vector machine (SVM), which performed best as the base learner.

3.2 Accurately predicted (AP) genes are top 50 most accurately predicted PR genes

Figure 2A shows the distribution of the Spearman correlation coefficients between the predicted and the observed Achilles essentiality scores of 2647 PR genes in 44 testing cell lines. The mean Spearman correlation was ∼0.201 (see Supplementary Table S1 for all scores). Figure 2B shows the performance of our methods in individual cancer cell lines (see Supplementary Table S2 for all scores). The cell lines that had the best prediction performance were SUDHL4 (diffuse B-cell lymphoma cell line) and MELHO (melanoma cell lines). On the other hand, some cell lines were predicted with suboptimal accuracy, which indicates that these cell lines might need further exploration.

In view of the complexity of biological processes and the diverse characters and functions of genes, we are not expecting to obtain equal prediction accuracy for all potential driver genes. Indeed, some PR genes were predicted with extraordinary accuracy. The top three most accurately predicted genes are *PSMD2* (mean correlations = 0.771), *SF3B3* (mean correlations = 0.756) and *SF3A1* (mean correlations = 0.771) (see Fig. 2C and Supplementary Fig. S4). We define genes with top 50 Spearman correlations as Accurately Predicted (AP) genes. We investigated the cause of the varied prediction performance of AP genes compared with non-AP genes (genes not among the top 50) from several aspects.

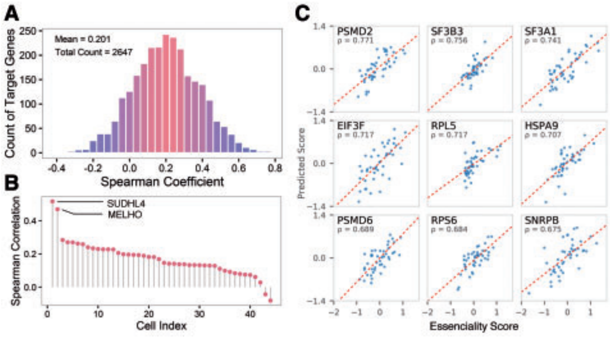

Fig. 2. Overview of prediction results. (**A**) The distribution of the evaluation metrics (i.e. the Spearman coefficients between the Achilles gene essentiality scores and the predicted scores) (**B**) The distribution of the performance of our method in different cells. The Spearman correlations were calculated between the predicted scores and the Achilles essentiality scores of 2647 PR genes in each of the 44 testing cell lines. (**C**) The prediction results of the top nine most accurately predicted PR genes. The predicted scores in this figure were not the original scores we submitted to the DREAM challenge, but were scaled so that they had the same mean and standard deviation as the Achilles essentiality scores. Scaling was performed for visualization and did not change the Spearman coefficients. See also Supplementary Figure S4

3.3 AP genes were connected with more genes in the functional network, and were associated with protein and RNA processing GO terms

We used the human genome-scale gene functional network published by Li *et al.* to investigate the functional connection between AP genes and their related genes (H.-D. Li *et al.*, 2015). This is an isoform-level network built using multiple instance learning (MIL). This gene network integrated isoform-level features and gene-level annotations and is able to show functional associations between pairs of genes. Specifically, this network captures the information of how likely two genes function in the same biological pathway, by integrating large-scale co-expression datasets, protein-protein physical interactions, shared motifs and domains through a Bayesian integration framework, and trained on shared pathways or GO terms. The end result of this network is a connected graph between two genes in the system, where a connection indicates that two genes co-function. As shown in Figure 3A, genes with higher Spearman correlation coefficients tended to have more neighbors in the network. On average, each AP gene had 402.96 neighbors in the network, while the average number of neighbors for all PR genes was 186.19. The reason accounting for the difficulty in predicting genes with fewer neighbors might be that the functions of these genes could not be revealed by the expression level of other genes. Figure 3B is the functional network of the top five PR genes. This network showed that more accurately predicted genes had more neighbors and the top genes had many neighbors in common. These genes could be functionally related to each other. Figure 4A shows the Pearson correlation scores between AP genes and a union of their features. Some genes share similar patterns, i.e. they were highly correlated with the same group of features. This means those genes could share certain pathways.

After performing GO enrichment on AP genes using DAVID (Huang *et al.*, 2009a, b), we found these genes were associated with RNA processing and protein processing GO terms (Fig. 4B). Examples for RNA processing were 'mRNA processing', 'mRNA splicing', 'Spliceosome'. Examples for protein processing were 'proteasome', 'proteasome accessory complex'. We then compared the AP genes and a list of human housekeeping genes (Supplementary Table S1). These housekeeping genes are inferred by a Bayesian classifier based on 68 microarray screens in human cell lines (Hart *et al.*, 2014). We found that 37/50 (74%) AP genes are housekeeping genes;

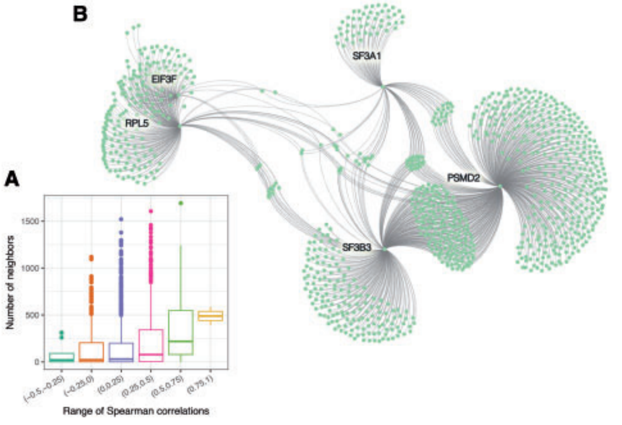

Fig. 3. Networks of top genes. (**A**) The boxplot of the number of neighbors of PR genes in the network. The *x*-axis is the corresponding range of Spearman correlation coefficients of the PR genes. (**B**) The network of top five PR genes. Only the nearest neighbors in the functional network are shown

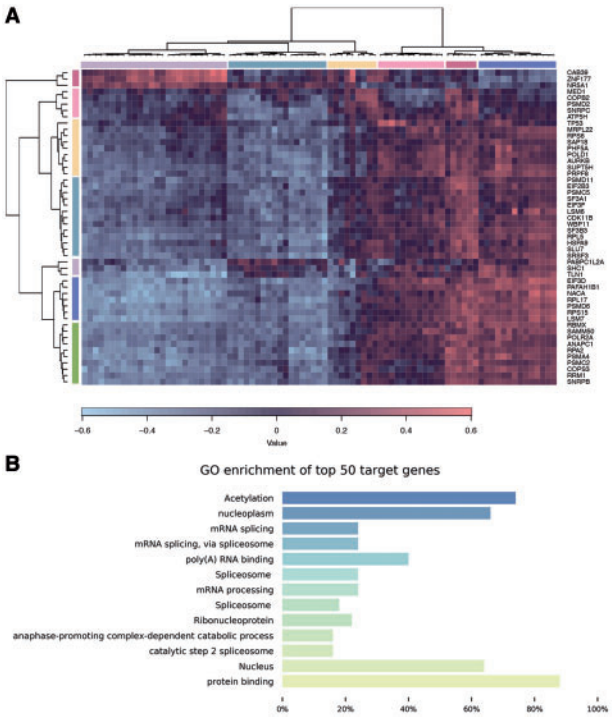

Fig. 4. Go enrichment of AP genes and heatmap of Pearson correlation coefficients between AP genes and their features. (**A**) A heatmap showing the Pearson correlation coefficients between top 50 PR genes (the rows) and a union of their features (the columns). Clustering was performed on both columns and rows. Different clusters are marked by sidebars with different colors. (**B**) The GO enrichment results of AP genes. The x-axis is the percentage of genes involved in this annotation category

meanwhile, 361/2647 (14%) of all PR genes are housekeeping genes. There exists a significant difference in the Spearman correlations between housekeeping PR genes and non-housekeeping PR genes (*t*-test *P*-value < 2.2e-16) (Supplementary Fig. S5). The housekeeping genes are significantly more accurately predicted.

These results partially explained the high performance of the AP genes. Top genes are associated protein processing and RNA processing related GO terms, and many of them appeared to be essential (housekeeping) across cell lines. Therefore, their function affects a numerous number of genes and proteins, and the activity of these genes could be uncovered by the expression levels of other genes, which made them more accurately predicted.

We also performed GO enrichment using DAVID on top 50 most frequent features (Supplementary Fig. S6A). Many of the enriched GO terms were related to cell adhesion and transmembrane signal transduction.

3.4 Copy number data complements the prediction power of gene expression data in predicting gene essentiality

As shown in Figure 5A, we tested four different combinations of expression features and copy number features:

1. The first method ranked all expression features by the combined correlation scores, then picked up top 10 as the predictive features.
2. The second method ranked all expression features by the combined correlation scores, then used top nine expression features plus the copy number feature of the PR gene itself (if available) as 10 predictive features.

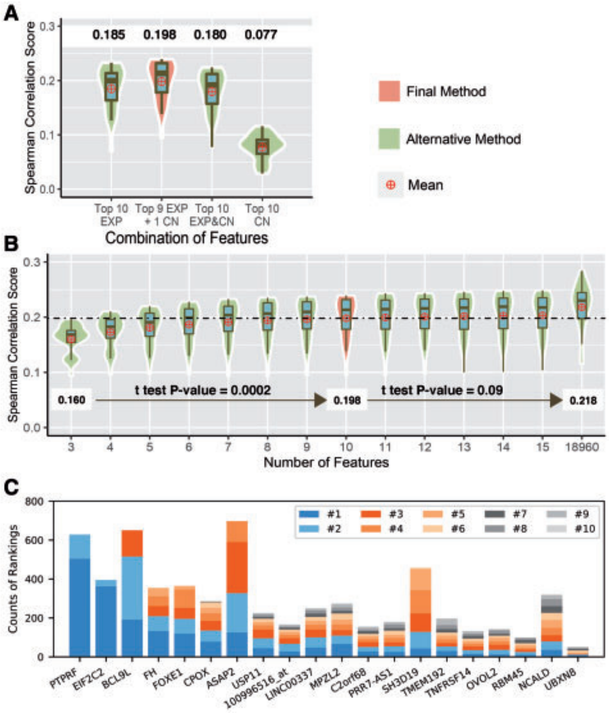

Fig. 5. Cross-validation results and rankings of features. The final method is highlighted in red, while alternative methods are in green. The dashed horizontal line in B marks the mean of Spearman coefficients of the final method. (**A**) The performance of different combinations of features. ‘Top 10 EXP’ means that the 10 predictive features were top 10 expression features ranked by the combined correlation scores. ‘Top 9 EXP +1 CN’ means that the 10 predictive features were composed of top nine expression features plus the copy number profile of the PR gene itself (if available). ‘Top 10 EXP&CN’ means that we mixed expression features and copy number features together, then ranked them by their combined correlation scores, then picked up top 10 features in this mixed list. ‘Top 10 CN’ means that the 10 predictive features are top 10 copy number features ranked by the combined correlation scores. (**B**) The performance of different numbers of features. A model with k features includes (k–1) expression features and one copy number feature (if available). 18 960 is the maximum number of features we could use. (**C**) A summary of the rankings of features. The features are ordered by the ratio of how many times it has top two combined correlation scores over how many times it was selected as a predictive feature. The label ‘100996516_at’ is the name of the probe, which does not map to a gene. See also Supplementary Figures S1 and S3

3. The third method pooled the expression data and copy number data together, resulting in a list of mixed features. Then the local scores, global scores and combined correlation scores were calculated in the same way as mentioned above. After ranking the mixed features by combined correlation scores, we chose the top 10 features in this ranked list as predictive features.

$$18\,960\ (\text{expression features}) + 23\,288\ (\text{copy number features}) = 42\,248\ (\text{mixed features}) \tag{8}$$

4. The fourth method ranked all copy number features by the combined correlation scores then picked up top 10 as the predictive features.

The second method outperformed the others, which proved that adding copy number information of the PR gene itself could improve the prediction performance, but adding copy number of other genes would have the opposite effect. In our experiment, the number of features is overwhelmingly large compared to the number of examples. A traditional feature selection scheme will inevitably lead to selection of noisy features that happen to/coincidentally correlate with the outcome. The selection of self-CNV is, instead, guided more by biological insight than a purely data-driven approach. This insight is that the copy number of a gene itself will greatly affect the result when the gene is targeted or is to be suppressed. The known relationship between cancers and copy number alteration (CNA) can explain the prediction power of the copy number of the PR gene itself. CNAs frequently appear in cancer cells and are believed to contribute to the progression of cancers (Bowcock, 2014), because regions subjected to copy number change could harbor key genes that have the potential to trigger the development or change the characteristics of cancers. HER2-positive breast cancer is breast cancer with overexpressed HER2 protein or amplified *HER2* gene (Ng *et al.*, 2015; Lee *et al.*, 2014; Slamon *et al.*, 1987). For HER2-positive breast cancers, targeting HER2 will make a significant change to the proliferation of cancer cells, which implies that the essentiality of *HER2* is high. Meanwhile, the essentiality of *HER2* is lower for HER2-negative individuals. Therefore, it is reasonable to assume the non-random co-occurrence of high/low gene essentiality and different patterns of copy numbers. This non-random co-occurrence between the feature and the gold standard is a potential reason that explains the predictive power of copy number features. Using copy number data of other genes did not result in better performance. This is probably because the copy number change of a certain gene is often accompanied by the copy number change of other genes especially its flanking genes because the genome of the cancer cells is fragile and unstable (Nijhawan *et al.*, 2012). The copy number features are therefore prone to false positive choices.

3.5 Previously confirmed cancers biomarkers were picked out by the feature selection method

After feature selection, each PR gene was assigned 10 predictive features (see Supplementary Table S3 for the predictive features of all PR genes). We ranked the expression features with respect to how many times each was among 10 predictive features (Supplementary Table S4). We found most of the top features were genes previously identified as cancer biomarkers (Table 1). *BCL9L* (B-cell CLL/lymphoma 9-like, BCL9-2) was predictive for 650/2647 genes (24.6%) and has been associated with a variety of human cancers, such as colon, pancreatic (Sannino *et al.*, 2016a), leukemia (Sannino *et al.*, 2016b), intestinal (Zatula *et al.*, 2014) and breast (Zatula *et al.*, 2014). *PTPRF* was predictive for 628/2647 genes (23.7%). *PTPN12* ranked right after *PTPRF* for being predictive for 503/2647 gene (~19%). PTPN12 and PTPRF are both members of the protein tyrosine phosphatases (PTPs) family. Proteins of this family have key roles in many biological processes including cell proliferation, migration and differentiation (Li *et al.*, 2015). Various studies proved that the aberration of *PTPRF* is associated with tumorigenesis and malignancy (Soulières *et al.*, 2015). *PTPN12* is known for playing vital roles in cell migration and adhesion (Luo *et al.*, 2014) and has important roles in ovarian (Villa-Moruzzi, 2011, 2013), hepatocellular (Luo *et al.*, 2014) and breast cancers and is a potential therapeutic biomarker (Harris *et al.*, 2014; Tonks, 2006). In addition to the above-mentioned genes, other top features are potential biomarkers or tumor suppressors (Table 1). This result showed that our feature selection process was efficient in detecting predictive biomarkers.

Table 1. Top 10 features and their related cancers

Feature	# of PR genes (%)	Known related cancers
ASAP2	697 (26.3%)	—
BCL9L	650 (24.6%)	Colon cancer, pancreatic cancer (Sannino *et al.*, 2016a), leukemia (Sannino *et al.*, 2016b)
PTPRF	628 (23.7%)	Breast cancer, lung cancer (Bera *et al.*, 2014; Liu *et al.*, 2015; Soulières *et al.*, 2015), prostate cancer (Trojan *et al.*, 2005), colorectal cancer (Bera *et al.*, 2014; Bujko *et al.*, 2015), gastric cancer, hepatocellular carcinoma (Soulières *et al.*, 2015)
PTPN12	503 (19.0%)	Breast cancer (Harris *et al.*, 2014; Luo *et al.*, 2014; Tonks, 2006), hepatocellular carcinoma (Luo *et al.*, 2014), ovarian cancer (Villa-Moruzzi, 2013, 2011)
ANXA1	480 (18.1%)	Bladder cancer (Yu *et al.*, 2014), lung cancer, pancreatic cancer, colorectal cancer, liver cancer (Guo *et al.*, 2013)
AJUBA	479 (18.1%)	Malignant mesothelioma (Tanaka *et al.*, 2015)
CYTIP	471 (17.8%)	—
SH3D19	457 (17.3%)	—
CMTM4	418 (15.8%)	Clear cell renal cell carcinoma (Li *et al.*, 2015)
EIF2C2	395 (14.9%)	Bladder carcinoma (Zhang *et al.*, 2015), colon cancer (Li *et al.*, 2010), myeloma (Zhou *et al.*, 2010)

The first column is the gene symbol of the CCLE-EXP gene. The second column is the number of PR genes that those features were predictive for. The percentage is second column/total*100%. The third column listed a few cancers that were shown in published studies to be closely related to the CCLE-EXP genes.

3.6 Prediction accuracy of ten features resembled that of the whole feature set

We tested the prediction performance of different numbers of features using 5-fold cross-validation (Fig. 5B). The alternative numbers we tested ranged from 3 to 18 960, which is the total number of expression features. The performance kept increasing at the beginning, but the increasing rate of performance became much slower as more features were added. Three features and 10 features performed significantly different (t-test P-value = 0.0001689), but the difference between using 10 features and 18 960 features was not statistically significant (t-test P-value = 0.08603). It is also worth noting that the difference between 5 and 10 features was not statistically significant (t-test P-value = 0.09318). This showed the possibility that fewer features can be used while still capable of capturing most of the information and getting reasonable prediction accuracy. Most importantly, the computational time, in terms of linear SVM, grows linearly with the number of examples and a subset of 10 genes has great advantage over the whole ~20 000 gene features from a model construction perspective.

3.7 Most expression features are only predictive for a small set of PR genes

We counted the occurrence of each expression feature in 10 predictive features and found that some expression features appear to be predictive for as many as ~25% of PR genes (Table 1), such as *ASAP2* (~26.3%), *BCL9L* (~24.6%) and *PTPRF* (~23.7%).

Most expression features which appeared at least once in 10 predictive features were only predictive for a very small set of genes; 49.4% of all predictive expression features only appeared once. ~90% of all predictive features were predictive for no >~1.5% of all PR genes. Therefore, features were usually not globally predictive. The vast majority of features were specifically associated with a very small set of PR genes.

Some features had very special patterns of ranking (Fig. 5C and Supplementary Fig. S6B). For example, *PTPRF* was selected as predictive features for 628 genes and its combined correlation scores all ranked top two in those 628 genes. *EIF2C2* appeared 395 times in 10 predictive features, and 91.65% of the time it ranked #1. In contrast, for *PTPN12* the rankings were more evenly distributed (#1: 6.36%, #2: 8.55%, #3: 13.72%, #4: 10.93% and so on). We noticed that those features such as *PTPRF* and *EIF2C2* had very few neighbors in the functional network. *PTPRF* had only 30 nearest neighbors. *EIF2C2* and *BCL9L* had no neighbors under our criteria (though *EIF2C2* and *BCL9L* do interact with numerous genes). In contrast, *FH* had 839 neighbors and *CPOX* had 215 neighbors. This raises the possibility that features with fewer neighbors had higher rankings due to the reason that the network it is involved in is small and all the genes in this network are strongly associated with each other, so the feature is highly predictive for those genes involved in that network. Another reason could be that they were associated with RNA and protein processing functions. For example, *EIF2C2* encodes AGO2 (Argonaute-2) which is expressed ubiquitously in most parts of the body and plays a key role in RNA interference (Meister, 2013) and DNA repair (Ye *et al.*, 2015). Thus, the expression levels of such genes could affect or be affected by a large set of genes. This could result in a relatively high global score. As our model puts more weight to global scores than local scores (7:3), those genes were more likely to have higher rankings.

4 Discussion

We developed a model that used the gene expression and copy number profiles of cancer cells to predict the essentiality of a list of genes. The accuracy of our prediction was measured by the Spearman correlations between the predicted scores and the Achilles gene essentiality scores. The Achilles essentiality scores of genes in cancer cells were measured by the influence of genes on the proliferation rate of the cancer cells. The essentiality scores were provided by Project Achilles. The expression and copy number profiles were provided by CCLE. We selected 10 features from over 40 000 expression/copy number features to predict the essentiality of a set of prioritized genes in cancer cells. The 10 predictive features were composed of nine expression features and one copy number feature.

We defined top 50 most accurately predicted PR genes as Accurately Predicted (AP) genes. These genes' essentiality could be accurately predicted from expression and copy number features that we selected. The evaluation of our method was done in 44 testing cancer cell lines, and as shown in Figure 2C there were no extreme values or outliers. These facts indicated that the prediction accuracy of AP genes was genuinely stable in different cancer cell lines. Some of the AP genes and other accurately predicted genes are identified as housekeeping genes. We further investigated the genes which are non-housekeeping but still accurately predicted, and found that those genes are likely to be cancer-specific essential genes. POLD1, a non-housekeeping gene, has a Spearman score of 0.66. It was associated with colorectal cancers (Palles *et al.*, 2013). TP53 is a well-known non-housekeeping oncogene that is accurately predicted with a score of 0.62. This shows the exciting prospect that the top predicted genes (including AP genes) can be future research subjects, as they are potential cancer specific essential/housekeeping genes.

In this study, the features we selected were solely used as biomarkers for gene essentiality, but, as shown above, many previously discovered cancer biomarkers appeared in our top feature list. There were still some top features for which we could not confirm their relationship with cancers. Those genes might have vital roles in cancer cells that worth further exploration. In the feature selection procedure, we gave a higher weight to global scores (0.7) than local scores (0.3). However, as these two scores are on different scales (global scores $\in [0,\ 1]$, local scores $\in [-1,\ 1]$), and the identity in distribution cannot be proved either, a higher weight score does not imply higher importance. Nevertheless, when used alone, the global scores have better prediction performance than local scores (Supplementary Fig. S3A). This result indicates that global score is marginally more important than local scores in the feature selection process.

The support from CCLE and Project Achilles made the DREAM challenge and this study possible. However, the Achilles gene essentiality scores obtained by the loss-of-function screen were noisy in nature, as shRNA experiments could involve incomplete knockdown and nonspecific or off-target knockdowns (Svoboda, 2007). Our algorithm has the potential to improve its performance by integrating data from separate resources, such as data generated by CRISPR-Cas9 technology, to compensate for the noise (Smith *et al.*, 2017). In particular, the study of the driver genes (epi-drivers) that are not mutated in their DNA sequence, but their epigenetic changes control the metastatic status of cells (Chatterjee *et al.*, 2017), will require additional datasets and algorithms to predict, which is a promising future research direction.

Funding

This work has been supported by grants from National Science Foundation 1452656 and Alzheimer's Association BAND-15-367116 (YFG) and National Institutes of Health P30ES17885 and U24CA210967 (GSO).

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bera,R. *et al.* (2014) Functional genomics identified a novel protein tyrosine phos-phatase receptor type F-mediated growth inhibition in hepatocarcinogenesis. *Hepatology*, **59**, 2238–2250.

Bertrand,D. *et al.* (2015) Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.*, **43**, e44.

Bowcock,A.M. (2014) Invited review DNA copy number changes as diagnostic tools for lung cancer. *Thorax*, **69**, 495–496.

Bujko,M. *et al.* (2015) Expression changes of cell-cell adhesion-related genes in colorectal tumors. *Oncol. Lett.*, **9**, 2463–2470.

Carter,H. *et al.* (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, **69**, 6660–6667.

Chatterjee,A. *et al.* (2017) Epigenetic drivers of tumourigenesis and cancer metas-tasis. *Semin. Cancer Biol.* doi:10.1016/j.semcancer.2017.08.004.

Cowley,G.S. *et al.* (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependen--cies. *Sci. Data*, **1**, 140035.

Guo,C. *et al.* (2013) Potential role of Anxa1 in cancer. *Future Oncol.*, **9**, 1773–1793.

Harris,I.S. *et al*. (2014) PTPN12 promotes resistance to oxidative stress and sup-ports tumorigenesis by regulating FOXO signaling. *Oncogene*, **33**, 1047–1054.

Hart,T. *et al*. (2014) Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol*., **10**, 733.

Huang,D.W. *et al*. (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.

Huang,D.W. *et al*. (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Lee,H.J. *et al*. (2014) HER2 heterogeneity affects trastuzumab responses and survival in patients with HER2-positive metastatic breast cancer. *Am. J. Clin. Pathol*., **142**, 755–766.

Li,H.-D. *et al*. (2015) Functional networks of highest-connected splice isoforms: from the chromosome 17 Human Proteome Project. *J. Proteome Res*., **14**, 3484–3491.

Li,J. *et al*. (2015) Loss of PTPN12 stimulates progression of ErbB2-dependent breast cancer by enhancing cell survival, migration, and epithelial-to-mesenchymal transition. *Mol. Cell. Biol*., **35**, 4069–4082.

Li,L. *et al*. (2010) Argonaute proteins: potential biomarkers for human colon can-cer. *BMC Cancer*, **10**, 38.

Li,T. *et al*. (2015) CMTM4 is frequently downregulated and functions as a tumour suppressor in clear cell renal cell carcinoma. *J. Exp. Clin. Cancer Res*., **34**, 122.

Liu,P.-J. *et al*. (2015) In-depth proteomic analysis of six types of exudative pleural effusions for nonsmall cell lung cancer biomarker discovery. *Mol. Cell. Proteomics*, **14**, 917–932.

Luo,R.-Z. *et al*. (2014) Decreased expression of PTPN12 correlates with tumor recurrence and poor survival of patients with hepatocellular carcinoma. *PLoS One*, **9**, e85592.

Meister,G. (2013) Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet*., **14**, 447–459.

Ng,C.K.Y. *et al*. (2015) Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous HER2 gene amplifica-tion. *Genome Biol*., **16**, 107.

Nijhawan,D. *et al*. (2012) Cancer vulnerabilities unveiled by genomic loss. *Cell*, **150**, 842–854.

Palles,C. *et al*. (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet*., **45**, 136–144.

Reich,M. *et al*. (2006) GenePattern 2.0. *Nat. Genet*., **38**, 500–501.

Sanchez-Garcia,F. *et al*. (2014) Integration of genomic data enables selective discovery of breast cancer drivers. *Cell*, **159**, 1461–1475.

Sannino,G. *et al*. (2016) Role of BCL9L in transforming growth factor-β (TGF-β)-induced epithelial-to-mesenchymal-transition (EMT) and metastasis of pan-creatic cancer. *Oncotarget*, **7**, 73725–73738.

Slamon,D.J. *et al*. (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, **235**, 177–182.

Smith,I. *et al*. (2017) Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol*., **15**, e2003213.

Soulières,D. *et al*. (2015) PTPRF expression as a potential prognostic/predictive marker for treatment with erlotinib in non-small-cell lung cancer. *J. Thorac. Oncol*., **10**, 1364–1369.

Stratton,M.R. *et al*. (2009) The cancer genome. *Nature*, **458**, 719–724.

Svoboda,P. (2007) Off-targeting and other non-specific effects of RNAi experi-ments in mammalian cells. *Curr. Opin. Mol. Ther*., **9**, 248–257.

Tanaka,I. *et al*. (2015) LIM-domain protein AJUBA suppresses malignant meso-thelioma cell proliferation via Hippo signaling cascade. *Oncogene*, **34**, 73–83.

Tonks,N.K. (2006) Protein tyrosine phosphatases: from genes, to function, to disease. *Nat. Rev. Mol. Cell Biol*., **7**, 833–846.

Trojan,L. *et al*. (2005) Identification of metastasis-associated genes in prostate cancer by genetic profiling of human prostate cancer cell lines. *Anticancer Res*., **25**, 183–191.

Tsherniak,A. *et al*. (2017) Defining a Cancer Dependency Map. *Cell*, **170**, 564–576.e16.

Villa-Moruzzi,E. (2013) PTPN12 controls PTEN and the AKT signalling to FAK and HER2 in migrating ovarian cancer cells. *Mol. Cell. Biochem*., **375**, 151–157.

Villa-Moruzzi,E. (2011) Tyrosine phosphatases in the HER2-directed motility of ovarian cancer cells: involvement of PTPN12, ERK5 and FAK. *Anal. Cell. Pathol*., **34**, 101–112.

Vogelstein,B. *et al*. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Ye,Z. *et al*. (2015) Argonaute 2: a novel rising star in cancer research. *J. Cancer*, **6**, 877–882.

Yu,S. *et al*. (2014) Correlation of ANXA1 expression with drug resistance and relapse in bladder cancer. *Int. J. Clin. Exp. Pathol*., **7**, 5538–5548.

Zatula,N. *et al*. (2014) The BCL9-2 proto-oncogene governs estrogen receptor alpha expression in breast tumorigenesis. *Oncotarget*, **5**, 6770–6787.

Zhang,Z. *et al*. (2015) EIF2C, Dicer, and Drosha are up-regulated along tumor progression and associated with poor prognosis in bladder carcinoma. *Tumour Biol*., **36**, 5071–5079.

Zhou,Y. *et al*. (2010) High-risk myeloma is associated with global elevation of miRNAs and overexpression of EIF2C2/AGO2. *Proc. Natl. Acad. Sci. USA*, **107**, 7904–7909.