Supplemental Figures

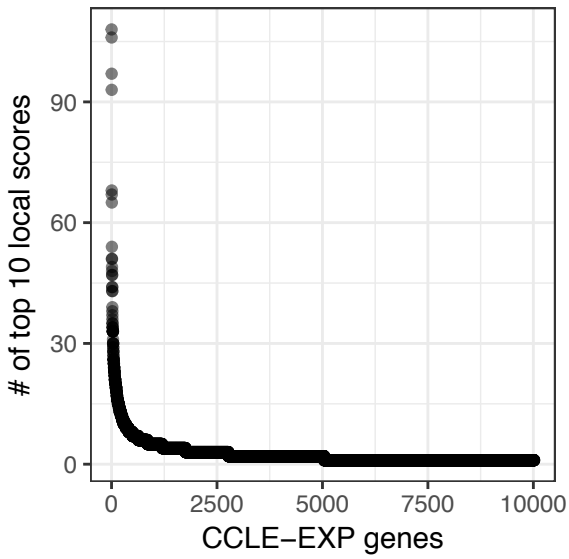

Figure S1 The distribution of the number of top 10 local scores for each CCLE-EXP genes. Related to Methods. The x-axis is the index of CCLE-EXP genes (18960 in total). The y-axis shows how many times a specific CCLE-EXP gene ranks top 10 with respect to local scores. Zeros (y=0) are omitted in this figure, therefore the tail of the x-axis is smaller than 18960. The highest y value is 108. Here 108 is the value of *max{count of occurrence in top 10}* in Equation 5 in the main text. The y values are the *count of occurrence in top 10* in Equation 5. This figure was generated with the all the training data. In cross validations, these numbers will change, because different training samples are used.

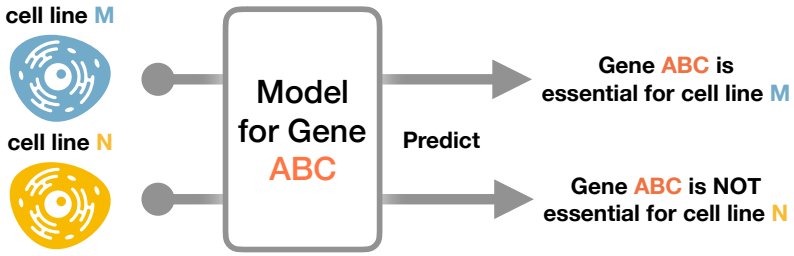

Figure S2 The schematic diagram of the gene specific model. Related to Methods. The gene essentiality predicting models described in this manuscript are gene specific models. The features are different for each gene-specific model. The predictions are also made separately for each candidate driver gene. Here gene ABC is the candidate driver gene. The essentiality of this gene in uncharacterized cell lines can be predicted by the model for gene ABC.

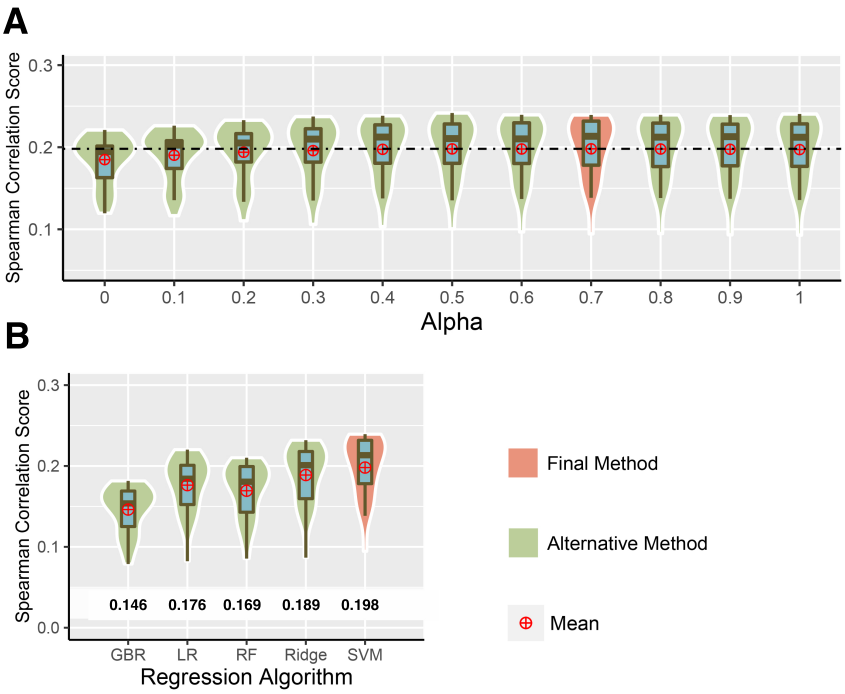

Figure S3 Cross-validation results of weight parameter α and 5 machine learning algorithms. Related to Methods and Figure 5.

A. Cross-validation results of weight parameter α. Different weights could be assigned to local and global scores by adjusting the value of the weight parameter α. To find the most optimized value for α, we performed 5-fold cross-validation to test 11 different α values ranging from 0.0 to 1.0, with a step of 0.1. When α = 0, the combined correlation scores equal local scores. When α=1, the combined correlation scores equal global scores. Namely, larger α values gave more weight to global scores. As shown in the figure, α = 0 achieved the worst performance. When α = 0.7, the Spearman correlation was the highest. **B**. Cross-validation results of 5 machine learning algorithms. GBR: gradient boosting regression, LR: linear regression, RF: random forest, Ridge: Ridge regression. We hypothesized that a linear model is suitable for our regression model. The provided dataset contained a limited number of samples, which might be unable to support a more sophisticated non-linear model. Also, introduction of features with high global scores aimed at identifying “nexus” genes that might affect the most essential pathways, and the expression level of any one of the genes should be a strong signal for our prediction by themselves. A linear model would be good enough for this scenario. The final results of cross-validation are shown in this figure. Linear regression, ridge regression, and linear support vector machine regression performed much better than nonlinear models such as gradient boosting regression tree and random forest regression, which echoed with our hypothesis. The linear support vector machine regression method performed marginally better than the other two linear models and thus became our final model of choice.

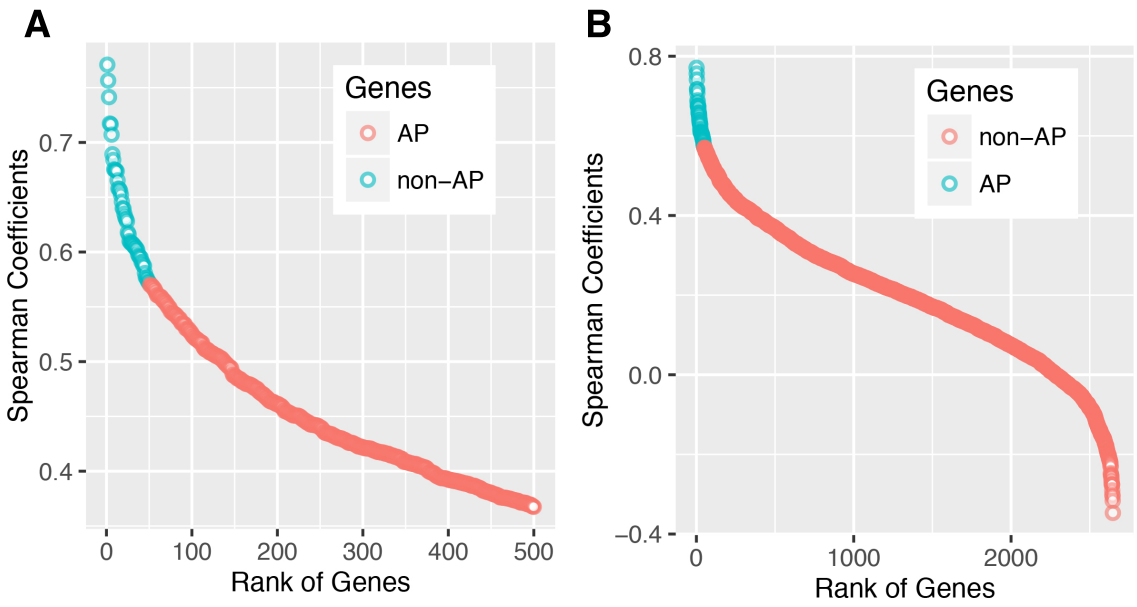

Figure S4 Top PR genes and their Spearman coefficients. Related to Figure 2.

A. The PR genes were sorted according to their Spearman coefficients. **B**. This is a zoomed-in view of A, focusing on top 500 PR genes. Two elbows are shown in A. AP genes (top 50 genes) are located before the first elbow.

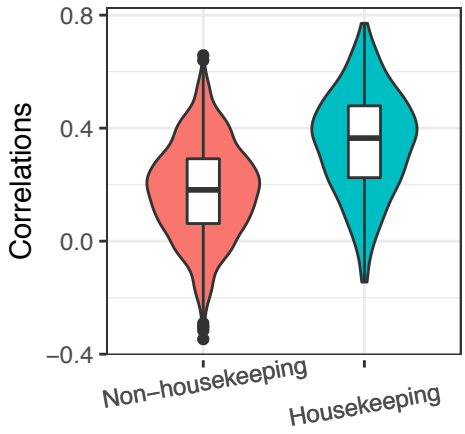

Figure S5 Comparison between the Spearman correlation scores of housekeeping PR genes and non-housekeeping PR genes. Related to Figure 4.

This figure shows the different distribution of the Spearman correlations of the non-housekeeping PR genes and housekeeping PR genes. The difference in sample sizes between these two groups are not reflected in this plot.

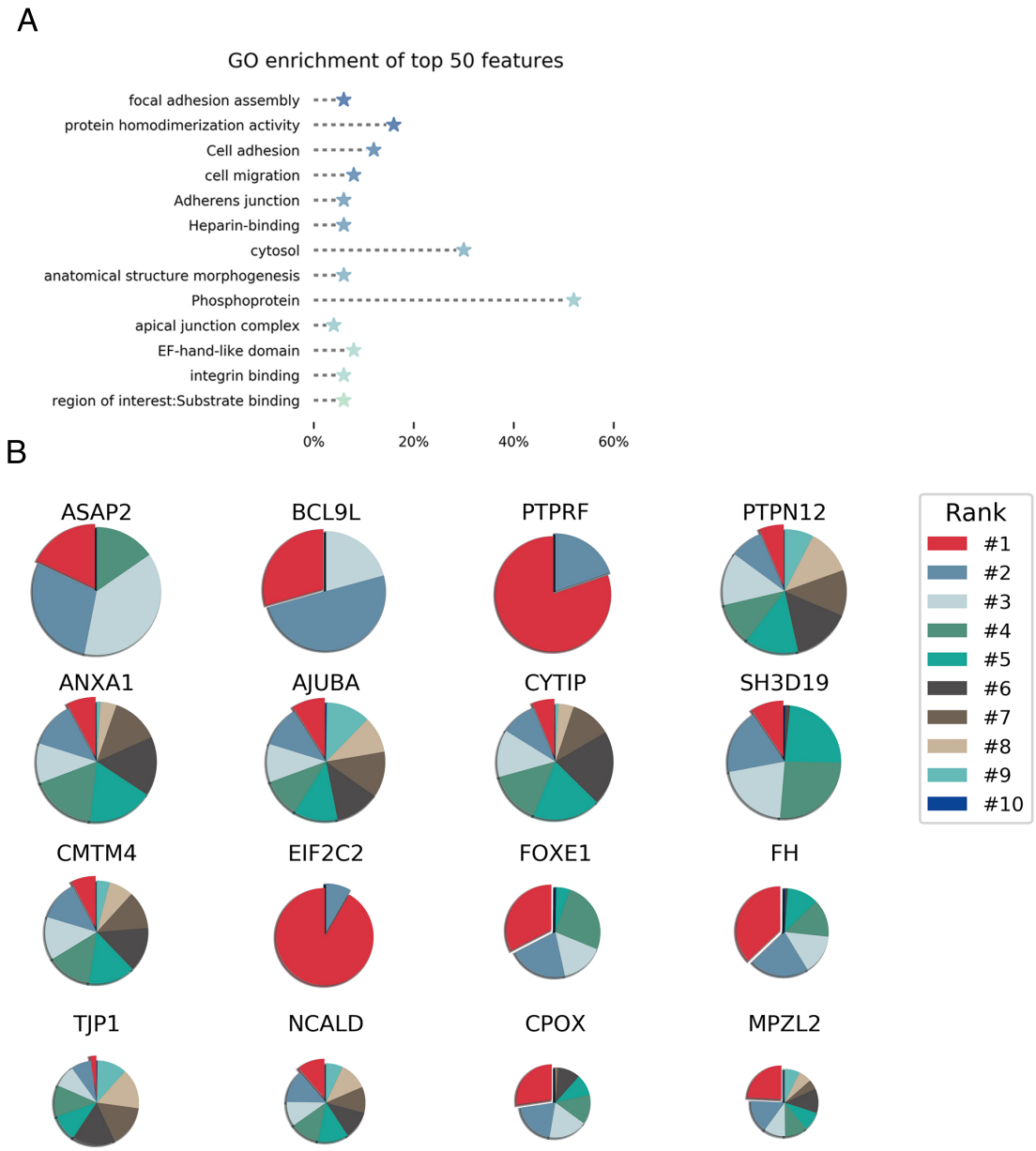

Figure S6 Rankings of top 16 expression features. Related to Figure 5.

A. The GO enrichment result of top 50 features. **B**. A summary of the ranking of top 16 expression features (CCLE-EXP genes). The radiuses of the pies are proportional to the number of PR genes that a feature was assigned to.

Supplemental Tables

Table S1: Top PR genes and their Spearman correlations. Related to Figure 2. (Table_S1.xls)

Table S2: Cell line performance. Related to Figure 2. (Table_S2.xls)

Table S3: PR genes and predictive features. Related to Figure 2. (Table_S3.xls)

This is a list of all PR genes and the 10 predictive features assigned to them. In the first table the name of the features are official gene IDs. In the second table, the features names are probe IDs used in the raw dataset.

Table S4: Top features and the summary of their rankings. Related to Figure 5. (Table_S4.xls)

The first column is the name of the CCLE-EXP gene, the second column is the total number of PR genes for which this CCLE-EXP gene is predictive. The third column is how many times this CCLE-EXP gene ranked No.1 in 10 predictive features.