# PRADA: Benchmarks and performance

Hosseinali Asgharian
hosseinali.asgharian@ucsf.edu

Hani Goodarzi
hani.goodarzi@ucsf.edu

## 1 The choice of transcript sequences

PRADA using regular expression representations of binding sites (e.g. [TG]ATATA[GC] for RBMS1) to scan the target sequences of interest for presence and absence of matches. We have used repeat-masked RefSeq mRNA sequences by default; however, the use of 3'UTR sequences for analysis of cis-regulatory elements is common. Therefore, we compared the output of PRADA between mRNA and 3'UTR analysis, and as shown below, the results are largely consistent (**Fig 1**). This analysis highlights the robustness of our findings with respect to the choice of input sequences.

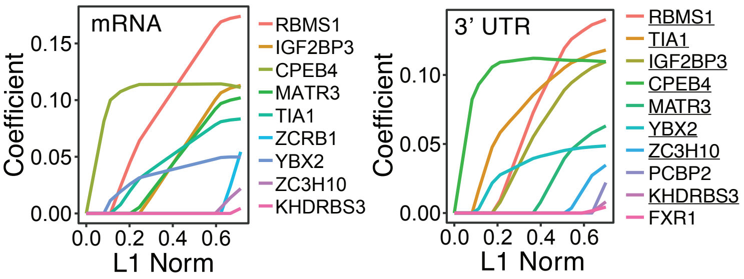

**Figure 1: The choice of input sequence**. (left) PRADA co-efficient as a function of L1-norm for mRNA RefSeq sequences. (right) A similar output using 3'UTR sequence for each RefSeq transcript. The RNA-binding proteins common between the two analyses are underlined. The top-ranking hits are identical between the two analyses and with the exception of a few RBPs, the two results largely overlap. The magnitude of the assigned coefficients is also quite similar (Euclidian distance=0.087, R = 0.88; P=0.0003).

## 2 The contribution of the selected RBPs to model performance (goodness of fit)

In order to measure the contribution of each selected feature (i.e. RBP) to overall performance of the model, we performed multiple independent analyses.

### 2.1 Contribution of features to $R^2$

We calculated the performance of the full models by calculating the correlation coefficient between the target values and the prediction of the model across all transcripts ($R = 0.31$, $P < 1e-16$). We then removed each RBP from the model by setting its assigned coefficient to 0 and re-calculating the R2 of the resulting model. As shown in **Fig 2A**, removing RBMS1 from the model resulted in the most substantial drop in $R^2$.

## 2.2 Increase in deviance in reduced models

If we have two models one of which ($M_2$) includes all of the predictors from the other model ($M_1$) plus k additional predictors, we can test if the larger model is significantly better than the smaller model, using the following equation:

$$Dev_{(M_2)} - Dev_{(M_1)} \sim \chi^2_k \tag{1}$$

Therefore, to evaluate the contribution of each feature (RBP), we removed it from the full model and measured the increase in deviance of the reduced model. We then calculated the associated p-value based on $\chi^2$ distribution. As shown in **Fig 2B**, removing RBMS1 from the model substantially and significantly increases the model's deviance.

## 2.3 Contribution of each feature to the Bayes Factor

A Bayesian approach to evaluating the goodness of fit involves measuring the drop in the model's Bayes Factor (BF) upon removing each feature. For this, we took advantage of the 'BayesFactor' package in R. The function *regressionBF* in this package, which allows evaluation of leave-one-out reduced models to the full model (with the parameter *whichModels*), evaluates each reduced model to the full model. As shown **Fig 2C**, the highest drop in BF results from removing RBMS1 from the model.

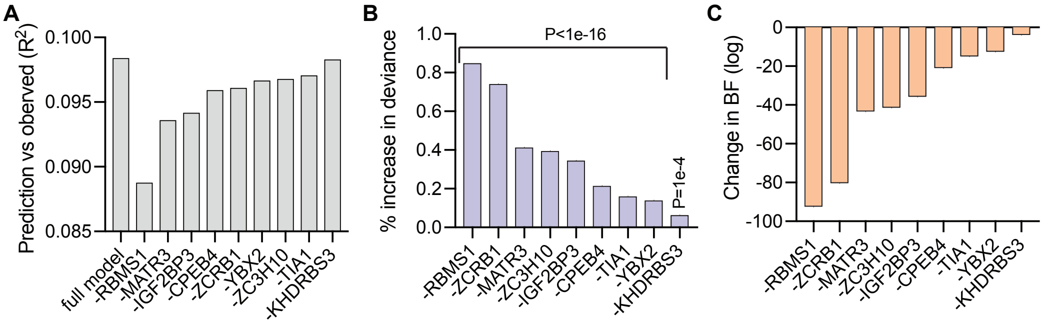

**Figure 2: RBMS1 contributes substantially and significantly to the model's goodness of fit. (A)** Comparing the full model's $R^2$ (left-most bar) to reduced models from which the RBP listed on x-axis is removed. **(B)** The increase in deviance that results from removing each RBP from the model relative to the deviance of the full models (in %). Also shown are the P values associated with each bar based on $\chi^2$. All Ps, except for the last one, were smaller than the smallest number in R and were set to $1e-16$ by default. **(C)** Similar to (B), but instead measuring the change in $ln(BF)$ in the reduced models compared to the full model. The x-axes are sorted in all three plots.

## 2.4 Evaluating RBMS1 against models from randomly sampled datasets

### 2.4.1 *Randomly generated datasets*

The generate random datasets, we started with the gene expression dataset of colon cancer lines (Fig 1 in the main text). The values for each gene were $z$-normalized across samples, and then the values permutated across each sample and transformed back to expression values using the mean and standard deviation for each gene in the original dataset (in practice, since $z$-scores across genes for each sample were largely Gaussian, this is akin to sampling from a normal distribution with the same mean and SD as the original data). This process randomizes the expression of the RBPs, as well as their targets. We then compared the 'poorly' and 'highly' metastatic labels to calculate the resulting log fold-change. PRADA was then used to identify the associated RBPs that best describe the model. Once the models were trained, we then evaluated the contribution of the top RBP to the goodness of fit (as described in section 2.1). We recorded the $(R^2_{full} - R^2_{reduced})$ for each model from 100 randomly generated dataset and compared the results to that of RBMS1. As shown in Fig 3A, RBMS1 far surpasses any of the randomly generated models.

### 2.4.2 *Random sample assignment*

To ensure that our null distribution is not overly artificial, we sought to also use a conservative sampling approach as well. In this case, the gene expression values were maintained but the cell lines were randomly shuffled between the 'poorly' and 'highly' metastatic labels. The resulting models were evaluated similar to above. This is a conservative estimate because specific random groupings of the 11 cell lines, could in certain configurations be biologically meaningful (e.g. MSI vs. MSS lines). And we expect biologically meaningful comparisons to result in models that identify RBPs associated with the underlying molecular mechanisms. Nevertheless, under this conservative scenario, RBMS1 was among the top $5\%$ of RBPs sampled from these groupings (**Fig 3B)**.

Together these results establish RBMS1 as an indispensable feature to the performance of the models and its goodness of fit.

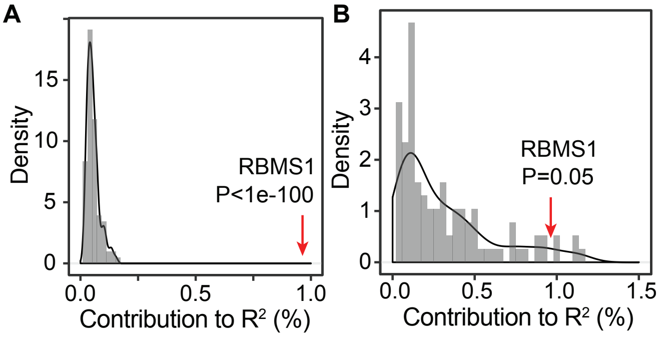

**Figure 3: Comparing the contribution of RBMS1 to the full model against the top RBP from randomly generated datasets**. **(A)** 100 modeled were trained on randomly generated data and the contribution of their top feature (i.e. RBP) to $R^2$ was recorded. The red arrow indicates the same measure for RBMS1 from the original model. **(B)** A similar analysis, but the random datasets were generated by shuffling cell line labels and keeping the gene expression profiles constant.

## 3 Evaluating PRADA's robustness

The stability of predictions by computational algorithms are measured through a variety of approaches; however, robustness estimates are the most commonly used metric. Robustness measure the consistency of outcomes when the starting data is varied, either through sub-sampling or jackknifing. To test the robustness of RBPs nominated by PRADA, we sub-sampled $2/3$ of the transcripts and re-ran PRADA to identify the top RBPs. We used two independent measures to then compared the resulting coefficients to those from the complete input data: (i) spearman correlation between coefficients of the two models, and (ii) the Euclidian distance between the coefficient vectors (excluding the intercept). We repeated this 100 times to generate a distribution for these measures. As shown in **Fig 4**, we observed high correlation coefficients and low Euclidian distances, emphasizing the robustness and stability of PRADA. As a point of comparison, we also tested unmodified Lasso regression (without the $\Delta Exp(RBP)$ in the denominator of the penalty term) using a similar strategy. As shown below, unlike PRADA, sub-sampled datasets result in widely different coefficients and result in low correlations and high Euclidian distances. In fact, the instability of Lasso model, which were used in an earlier study to identify master regulators (Perron et al., 2018), was the main motivation for developing PRADA.

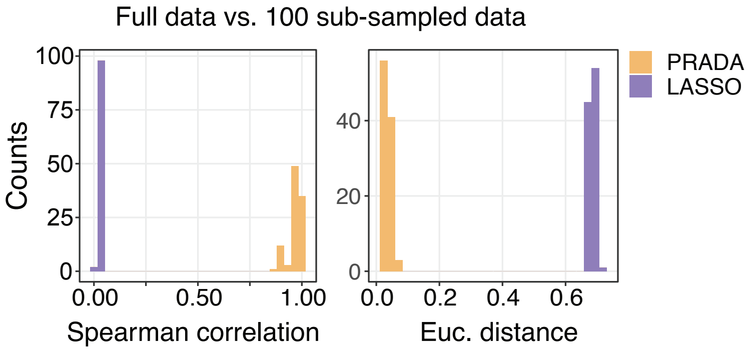

**Figure 4: Robustness of PRADA and its comparison to Lasso**. First transcripts coding for RNA-binding proteins were separated from the logFC profile comparing poorly and highly metastatic colon cancer lines (Fig 1 in main text). The remainder of the transcripts were sub-sampled by randomly selecting $66\%$ of the transcripts. The RBP transcripts were then appended and PRADA was used to identify RBPs with the highest absolute coefficients. This sampling process was repeated 100 times and each time the spearman correlation between the coefficients from the full dataset and those of the sub-sampled datasets was recorded (left). Similarly, the Euclidian distance between the coefficient vectors was also calculated (right). This process was carried out for models fit using Lasso instead of PRADA (purple bars).

## 4 Evaluating PRADA on real data

To test PRADA on known data, we selected the two RNA-binding proteins HNRNPA2B1 and YBX1. We have previously shown that YBX1 regulates mRNA stability, and more importantly had carried out YBX1 HITS-CLIP that provides the list of real targets and a high-confidence binding site (Goodarzi et al., 2015). In this study, we had also carried out YBX1 knockdown and gene expression profiling that can be used as an input for PRADA (Goodarzi et al., 2015). Similarly, for HNRNPA2B1, we had generated a CLIP dataset (Goodarzi et al., 2012) and knock-down RNA-seq data in two different cell lines, namely HeLa and MDA-MB-231 cells (Alarcon et al., 2015). PRADA already incorporated binding site predictions for these two RBPs based on their *in vivo* binding, so we simply provided the log fold-change values from the knock-down experiments as the input to PRADA and asked whether these factors will be identified by PRADA as regulators of mRNA stability. It should be noted that since both YBX1 and HNRNPA2B1 are broad regulators, it is expected additional regulatory modulations will take place downstream to them that may also be captured by PRADA; nevertheless, our expectation was that PRADA will be identify YBX1 and HNRNPA2B1 as top contenders in their respective datasets.

YBX1 is an enhancer of RNA stability and as shown below in **Fig 5A**, PRADA has assigned a positive and significant co-efficient to YBX1 that partially explains gene expression changes resulting from its knockdown. Similarly, HNRNPA2B1 is also assigned a significant co-efficient in both of its knockdown datasets, in MDA and HeLa cells respectively (**Fig 5B-C**).

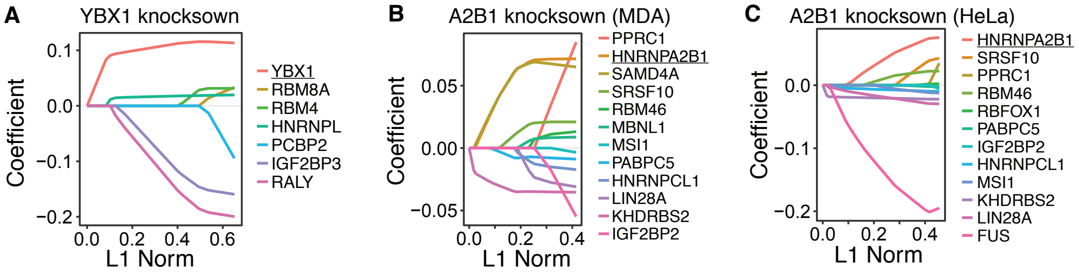

**Figure 5: PRADA identifies YBX1 and HNRNPA2B1 as regulators of RNA stability in their respective knockdown gene expression datasets**. **(A)** A previously published dataset comparing YBX1 knockdown and control cells in MDA-MB-231 (Goodarzi et al., 2015) was used as known data to evaluate PRADA. As shown here, YBX1 is assigned a significant positive co-efficient by PRADA, which is consistent with its role as an enhancer of RNA stability. **(B-C)** A similar analysis for HNRNPA2B1 using RNA-seq data comparing HNRNPA2B1 knockdown to control cells in two different cell lines (Alarcon et al., 2015). PRADA has identified HNRNPA2B1 as a positive regulator of gene expression in this dataset as well, which is consistent with its role as an enhancer of RNA stability. Importantly, there is substantial overlap between the factors nominated by PRADA in the two cell lines, which further emphasizes both the consistency of HNRNPA2B1 across cell lines and the stability of the models from PRADA. In each plot, the known RBP of interest is underlined.

## 5 Evaluating model performance

Finally, to assess the performance of PRADA using commonly used machine-learning metrics, we took advantage of simulated data to generate known positive and negative test cases. For this, we used the YBX1 dataset used above to simulate data with realistic distributions. For this, we split the log fold-changes in the YBX1 knockdown data into transcripts that are bound by (targets) versus others (non-targets). We used YBX1 HITS-CLIP data to separate targets and non-targets, which is the gold-standard as it captures precise RNA-protein interactions *in vivo*. To simulate positive knockdown data, we randomly selected an RBP and sampled its predicted targets from the YBX1 target distribution and the rest of transcripts were sampled from the non-target distribution. For sampling, we used a Gaussian kernel, whereby we sampled values from logFC vectors (with replacement) and added a randomly generated value from a normal distribution with mean $0$ and $SD$ of the density bandwidth. Once logFC data was simulated for targets and non-targets, we also simulated the knockdown of the test RBP by sampling from a normal distribution with mean $-1$ and SD of $0.5$ (i.e. $2$-fold reduction in expression on average). For the negative set, RBPs were selected and knocked down, but the logFC values were sampled from the entire distribution. This process was repeated until 100 positive and 100 negative datasets were simulated. We then ran PRADA on these sets and asked whether the test RBP was nominated among the top contenders (we set a cut-off of $1$ for $L1$-norm to select the top RBPs). As shown in **Fig 6A**, PRADA achieves an accuracy of $97.5\%$ with balanced specificity and sensitivity.

Since an earlier study had used Lasso regression in a similar setting (Perron et al., 2018), we repeated the same process for Lasso as well. As shown in **Fig 6B**, we observed a markedly reduced performance. Accuracy of Lasso was below $90\%$ and its sensitivity was especially low compared to PRADA.

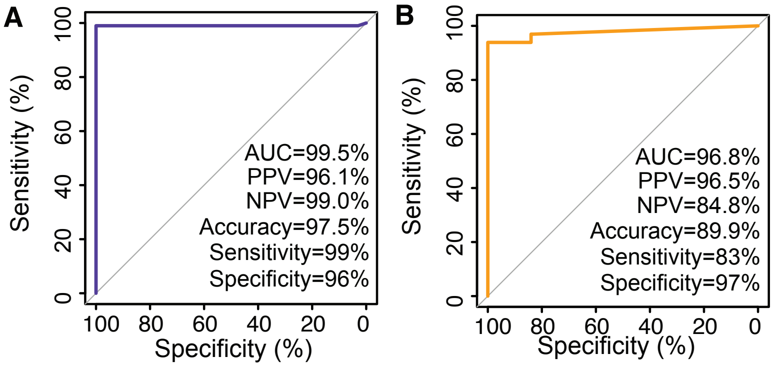

**Figure 6: ROC analysis for PRADA on simulated positive and negative data**. **(A)** ROC curve for PRADA, along with other performance metrics, from a randomly generated dataset that simulates RBP knockdowns. **(B)** A similar analysis using Lasso regression instead of PRADA. The R package pROC was used to generate the plots and caret was to calculate other performance metrics: AUC (Area Under the Curve), positive predictive value (PPV), negative predictive value (NPV), accuracy, sensitivity, and specificity.

Together, these results indicate that PRADA provides a rational, stable, and biologically meaningful approach for detecting post-transcriptional regulators of gene expression.

## References

Alarcon, C. R., Goodarzi, H., Lee, H., Liu, X., Tavazoie, S., & Tavazoie, S. F. (2015, September). HNRNPA2B1 Is a Mediator of m(6)A-Dependent Nuclear RNA Processing Events. *Cell*, *162*(6), 1299–308. doi: 10.1016/j.cell.2015.08.011

Goodarzi, H., Liu, X., Nguyen, H. C. B., Zhang, S., Fish, L., & Tavazoie, S. F. (2015, May). Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. *Cell*, *161*(4), 790–802. doi: 10.1016/j.cell.2015.02.053

Goodarzi, H., Najafabadi, H. S., Oikonomou, P., Greco, T. M., Fish, L., Salavati, R., ... Tavazoie, S. (2012, April). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, *485*(7397), 264–268. doi: 10.1038/nature11013

Perron, G., Jandaghi, P., Solanki, S., Safisamghabadi, M., Storoz, C., Karimzadeh, M., ... Riazalhosseini, Y. (2018, May). A General Framework for Interrogation of mRNA Stability Programs Identifies RNA-Binding Proteins that Govern Cancer Transcriptomes. *Cell Reports*, *23*(6), 1639–1650. doi: 10.1016/j.celrep.2018.04.031