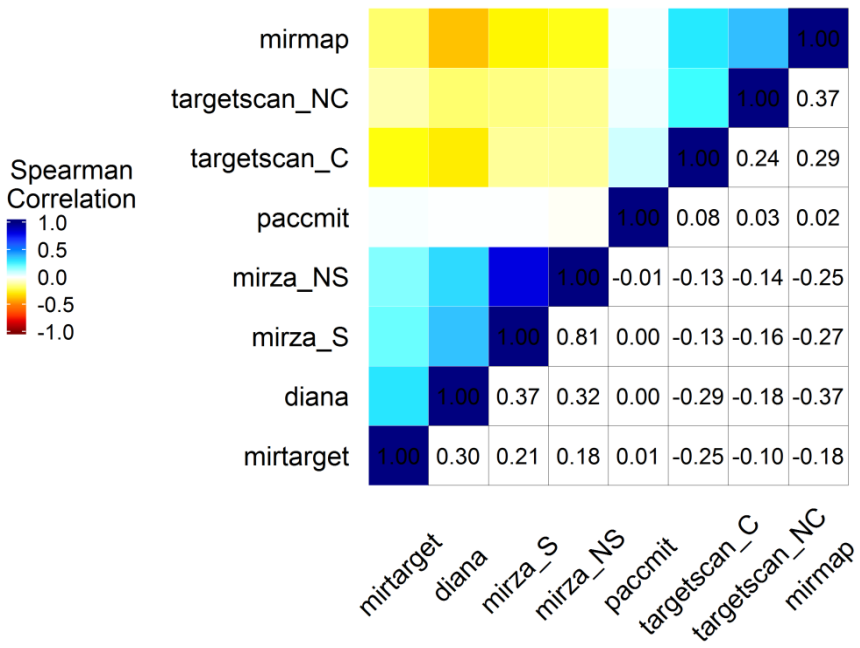
Supplementary Data

Evaluation of source data sets

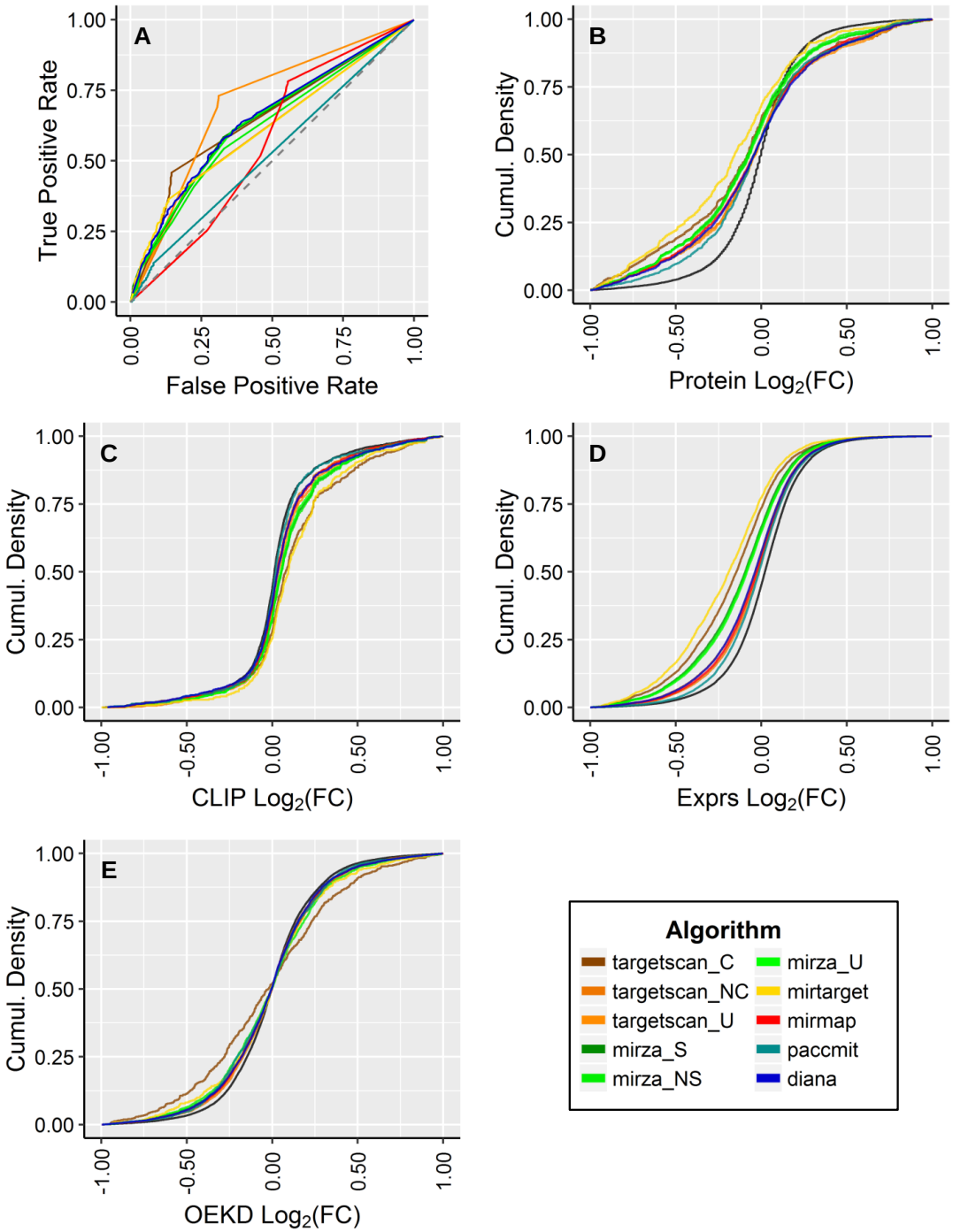
With the aim of building a comprehensive set of miRNA-gene interaction predictions, we assessed the degree of overlap among the established miRNA prediction algorithms used in this study (Diana microT-CDS (1), miRmap (2), MIRZA-G (with seed and no-seed requirement) (3), miRTarget3 (4), PACCMIT-CDS (5) and TargetScan v7.0 (6)) to find the most complementary data (Figure S1). A plot of the pairwise Spearman-rank-correlation coefficients comparing target predictions of the different algorithms) revealed the correlation among most of the input algorithms to be low, suggesting the potential to improve prediction classification by combining these non-overlapping datasets. For the initial analysis, the contributions of TargetScan conserved (_C) and non-conserved (_NC) were kept separate, as were the MIRZA seed and non-seed (_S and _NS, respectively, both without conservation) datasets. Among these, only the MIRZA data showed considerable overlap (coefficient 0.81) and for subsequent analysis, were merged into a single union set (_U). The TargetScan datasets were kept independent for the integration, but for subsequent performance comparisons, the union sets (mirza_U and targetscan_U) were used as a more impartial comparison of the data a user of these algorithms would see when performing an analysis.

Figure S1: Spearman rank correlation evaluation of input algorithms



The miRNA prediction sets used as inputs for this study were first assessed according to different sources of ground truth data (Figure S2). Each has already demonstrated its performance through publication and usage and as a result, all were to be included for consideration in assembly of the reference data table. Use of different sources of ground truth data for experimentally determined miRNA:targene-gene intereractions revealed slight differences in ranking of the constituent data sets. The publication mining databases Tarbase and miRTarBase contain classification information regarding interaction pairs and sorting for the most reliable data sources (defined as “strong” evidence), the source data sets present with differing qualities of classifier, as indicated by the range of area under curve (AUC) measurements in the receiver operating characteristic (ROC) curve plots. When fold change data are incorporated, as in the case of the gene expression data of Linsley (7) and Grimson (8), the collected proteomics data, cross-linking and immunoprecipitation-(CLIP-) based reads, or the dual overexpression/knockdown experiments, rankings are similarly varied as demonstrated by increasing divergence from the cumulative distribution of the complete data set (Figure S2B, C, D, E). Among these measures, one data set in particular, PACCMIT appeared to perform poorly. Most of the data sets are based on target searches in the 3’ UTR regions of transcripts, whereas the PACCMIT used here contains only predictions based on coding regions of genes (3’UTR data were not available). The data were retained, however, since performance could be demonstrated to improve in analyses of proteomics data as performed in the original publication, and in the current study when consideration for the location of putative seed sites is taken into consideration.

Figure S2: Comparison of established miRNA prediction algorithms by validation data



Prediction algorithm performance is dependent on the source of validated miRNA:target data. The input data sets were compared against the five sources of validated interaction data collected for this study: classification from publication databases (A), cumulative distribution against $\log_2$ fold changes for gene expression after exogenous miRNA overexpression (B), fold changes for proteomics data (C), and fold changes for CLIP read data (D). Further gene expression studies were separately analysed where investigated miRNAs were both overexpressed and knocked down to identify genes likely to be regulated (E). Performance is observed to vary somewhat across the data sources, most notably for the special case of proteomics data, where only those genes with a 6nt seed site (6mer) in the coding region and not one in the 3'UTR were selected. The data illustrate that by exclusion of the protein coding region, the best performing algorithms (TargetScan and MIRZA) do not account for the data as well as some other algorithms, which appear to perform poorly when considering the entire transcript

As shown in the main text (Figure 1), since the source algorithms contain largely non-overlapping prediction data, there is the potential for information gain by their combination. Further, as illustrated above, the ground truth data used to assess the performance of an algorithm can influence the ranking of that database. For the present reference database construction then, all available experimentally validated data (as described in the main text) were combined for the training and the evaluation of the new model.

Assembling the training set for the machine learning task

The available training data of experimentally validated miRNA:mRNA interactions is imbalanced, with many more examples of positive than negative interactions. Imbalance is also seen for the number of data points contributed by each source of experimental data, with the most supplied from CLIP. To address the class imbalance and to prevent overwhelming the training set with a single data source, over- and under-sampling were employed to balance the data prior to training (Table S1).

Table S1 Contributions to model training by source

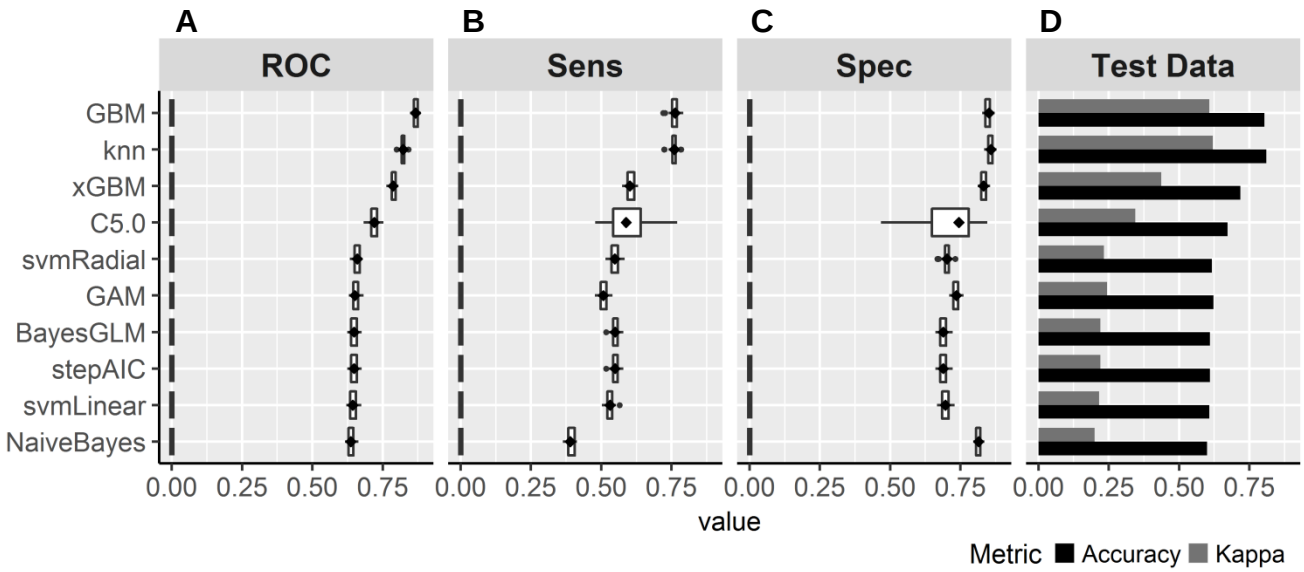
Type	Source	input	resampled	training
pos	CLIP	9666	3335	2331
	EXPRS	3803	3335	2379
	OEKD	1243	1144	2302
	PROTEO	1287	1182	2309
	PUBDB	4351	3335	2352
neg	CLIP	9683	3335	2328
	EXPRS	1577	1368	2370
	OEKD	628	625	2308
	PROTEO	649	643	2332
	PUBDB	433	433	2335

The collection of experimentally-validated interactions was first trimmed to retain only those for which prediction data were available from the input algorithms (column “input” Table S1). As mentioned, the positive interaction data generally outnumber the negative, and these data were downsampled as shown (column “resampled”). The resampling proportions were determined as follows. A total count was obtained of all experimentally-validated interactions (33,320). The final balanced set would consist of five sources (CLIP, EXPRS, OEKD, PROTEO, PUBDB), each with two types of interaction (positive and negative) thus resulting in 10 separate groups. The total count was divided by 9.99 (approximately 10) to yield a group target size of 3,335 interactions required for each data source and each type of interaction. Where the available data exceeded this value, entries were under-sampled, and where lower, they were over-sampled. The final training set

(column “train” Table S1) indicates the 70/30 split of training versus test data used for model building. The method resulted in retention of 18,375 interactions (or 56%) of the original input data.

Multiple machine learning models suitable for classification tasks were considered in the present study. Models for binary classification from the input features were generated using the R caret package (9). The full set of experimentally-validated data were split 70:30 into training and test sets respectively. In the training data set, 5 repeats of 10-fold cross validation were applied to control for overfitting, and data were pre-processed with centring and scaling. For each round of cross-validation, the training data were further sub-divided into pseudo-training and pseudo-test sets to train the model and evaluate its performance. Tuning was according to parameters appropriate to each model: Naïve Bayes (nb), Bayesian generalised linear models (bayesglm), generalised linear models with stepwise AIC (glmStepAIC), k-nearest neighbours (knn), gradient boosting machine (gbm), extreme gradient boosting machine (xgbTree), C5.0 (C50), and support vector machines with radial and linear kernels (svmRadial and svmLinear respectively). These methods were chosen on the basis of efficiency, flexibility, and assessment of contribution of predictors (that is, whether the class prediction is best achieved through inclusion of all predictors/input data sets, or only a subset).

Figure S3: Comparison of machine learning algorithms identify most suitable models.



Machine learning algorithms quantitatively combined the input data sets into a reference set with better classification performance. Performance of a selection of machine learning algorithms was compared across the 10-fold cross validation resampling of the training sets according to the area under the ROC curve (ROC), the sensitivity (Sens) and the specificity (Spec). The models were further tested against unseen data (Test Data) to measure accuracy and Kappa values. The top two models, GBM and knn, were similar in performance and resulting scores were averaged to generate the final reference table. Models listed: gradient boosting machine (GBM), k-nearest neighbours (knn), extreme gradient boosting machine (xGBM), C5.0, cost-sensitive support vector machine with a radial kernel (svmRadial), boosted generalised additive model (GAM), Bayesian generalised linear model (BayesGLM), stepwise linear regression with AIC (StepAIC), support vector machine with linear kernel (svmLinear), and Naive Bayesian models

Results of the cross-validation on the training set were analyzed for several metrics: area under the curve (AUC) of the receiver operator characteristic (ROC) curve (Figure S3A), the sensitivity (Figure S3B) and the specificity (Figure S3C). The final model was further assessed against the test dataset (30% of the original collection of experimentally-validated interactions), data which were unseen by the model during the training/cross-validation (Figure S3D). Across the applied assessments, the two best-performing machine learning algorithms were knn and GBM; within the experimental error the two models were found to be indistinguishable and thus the final reference dataset used for positive and negative prediction in the remainder of the study was taken as the arithmetic mean of class probabilities from these two.

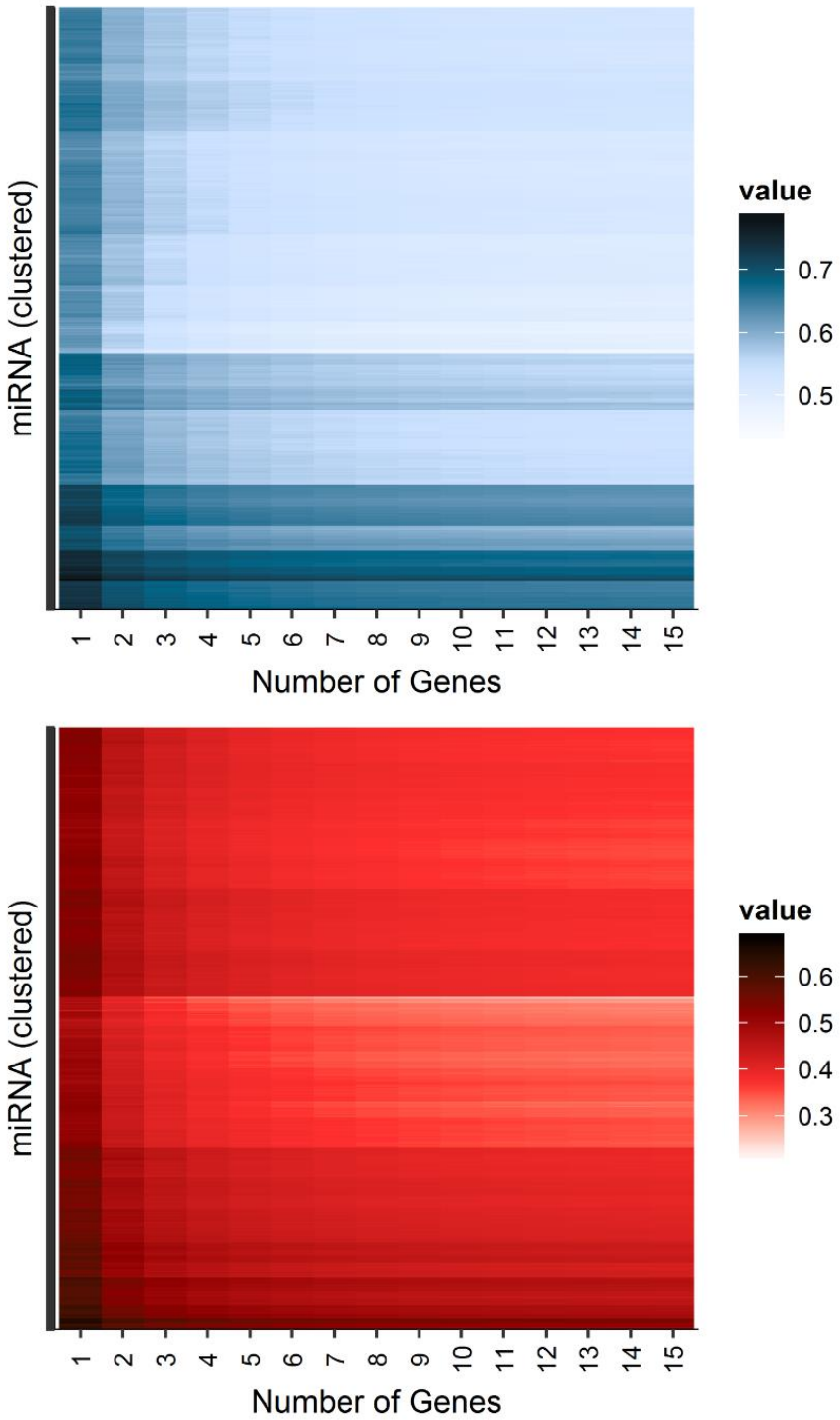
Analysis of raw combinatorial scores

The first stage of the analysis is the generation of a raw score, taken as the geometric mean of the individual prediction scores from the combined reference database. Although similar results were obtained using the arithmetic mean, the geometric mean was ultimately used as a better reflection of combining probabilities (that is, the class prediction probabilities from the individual interaction scores). These were not simply multiplied together, as such a combination would only be appropriate for independent events and we presume that the class probability of one miRNA:target

gene may influence the probability of the class determination of the same miRNA with another gene target.

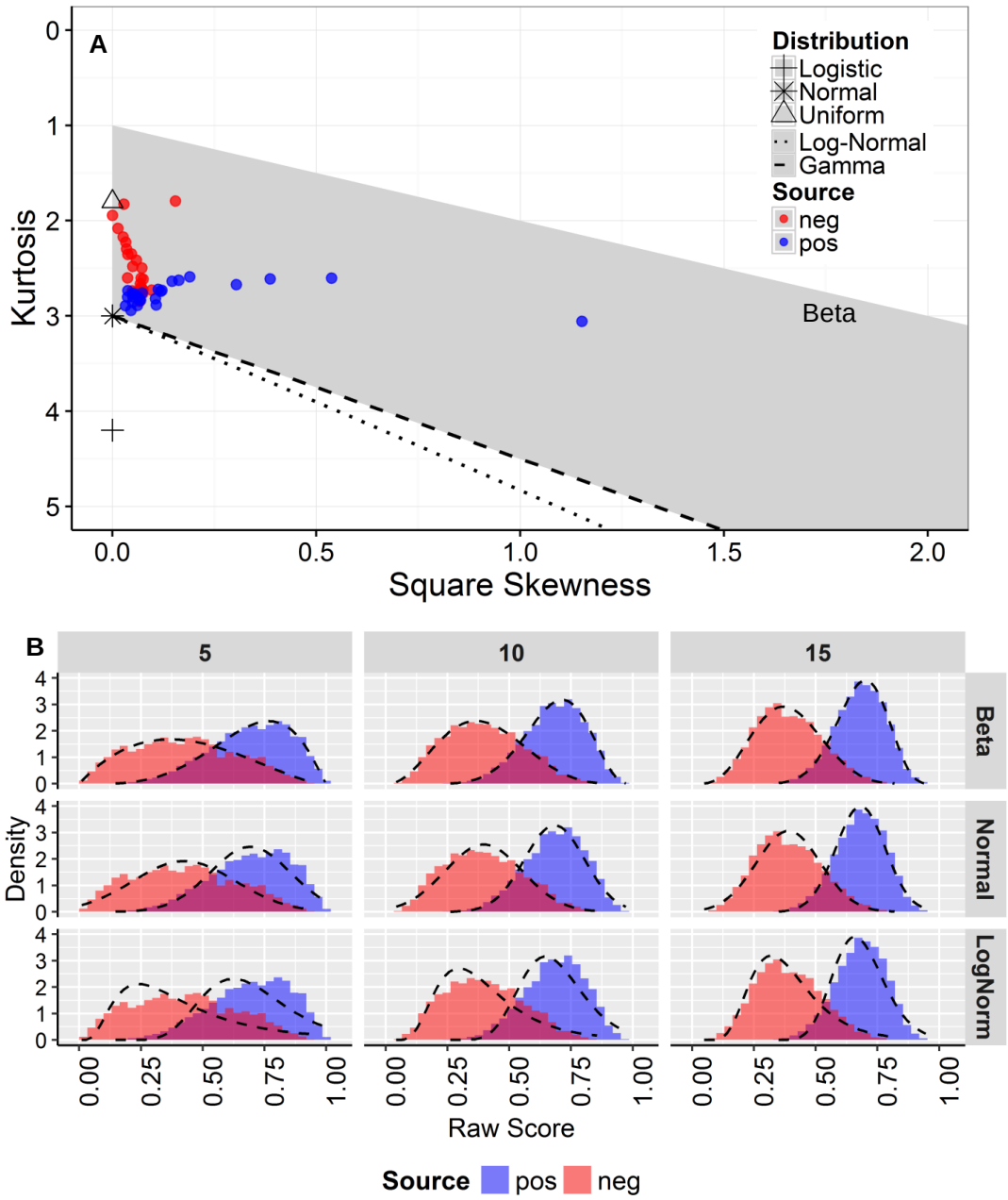
The next stages were undertaken to characterise the behaviour of randomly sampled data. With as thorough a description of the random data as possible, results returned from the analysis which differ considerably from the random distribution may be meaningful (in the context of the computational analysis and incognizant of biological considerations such as tissue expression). From the analyses of raw scores determined by randomly sampling interacting genes per miRNA, patterns were evident, indicating dependence of the resulting score on the miRNA under consideration, and the number of genes chosen (Figure S4). In an effort to describe the population of scores across all possible miRNA:gene clusters, the raw data were tested against possible parametric distributions (Figure S5). A plot of kurtosis versus the square of skewness of the data after Cullen and Frey (10) is used to suggest distributions which bear gross similarities in shape with known distributions. From the region of the plot in which the positive and negative data for the gene analysis sizes from 1 to 25 genes, a beta distribution is consistent, however, normal and log-normal were also considered, given their proximity to many of the data points (Figure S5A). Maximum likelihood estimation using the R `fitdistrplus` package fitdist function was employed to determine the parameters of the best fitting normal, log-normal, and beta distributions fitting the positive and negative data (Figure S5B). As a further comparison, quantile-quantile plots of positive and negative data across the tested distributions and gene analysis sizes were generated, further confirming the suitability of the beta distribution, over normal or log-normal distributions (Figure S6).

Figure S4: Random samples of raw miRNA:target gene prediction scores show dependence on the number of genes in the analysis and the identity of the miRNA



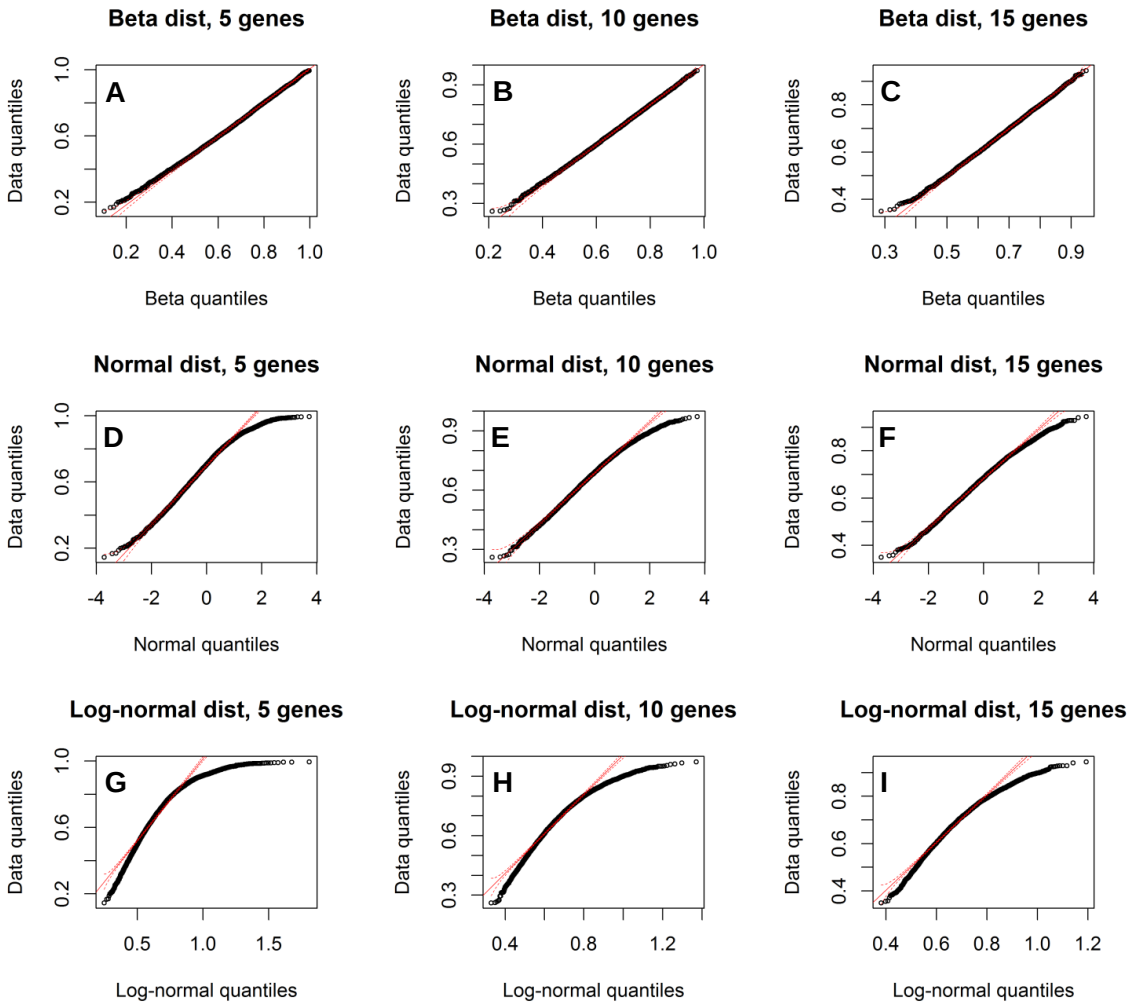
Positive (top, blue) and negative (bottom, red) prediction reference data sets were used for sampling of raw scores for miRNAs with gene combinations ranging from 1 to 15 genes. Patterns are evident in the score distributions, indicating dependence on the size of the analysis, and which miRNA is under study.

Figure S5: Raw score moments identify suitable parametric distributions



Random sampling of raw scores in combinations corresponding to 1-25 genes were subjected to 2,500 bootstrap replicates. The kurtosis of the resulting datasets was plotted against the square of the skewness after Cullen and Frey, using a modified version of the descdist function from the R `fitdistrplus` package (A). Points corresponding to analyses using the positive (blue) and negative (red) data fall within the beta region of the plot. Fits of the data to normal, log-normal, and beta distributions were overlaid on histograms of the data (B) revealing the poor fits of the normal and log-normal distributions, in contrast to the closer fit of the beta.

Figure S6: Quantile-quantile plotting identifies a suitable parametric distribution to model random data



To better quantify the fitting of the models to the experimental data, quantile-quantile plots were generated for the positive data. Plots for 5 (A, D, G), 10 (B, E, H), and 15 (C, F, I) randomly chosen genes were tested against the beta (A, B, C), normal (D, E, F) and log-normal (G, H, I) distributions. Across the range of gene selection sizes, the beta distribution better models the random gene selection. Plots generated using the R package car. Plots continued on next page

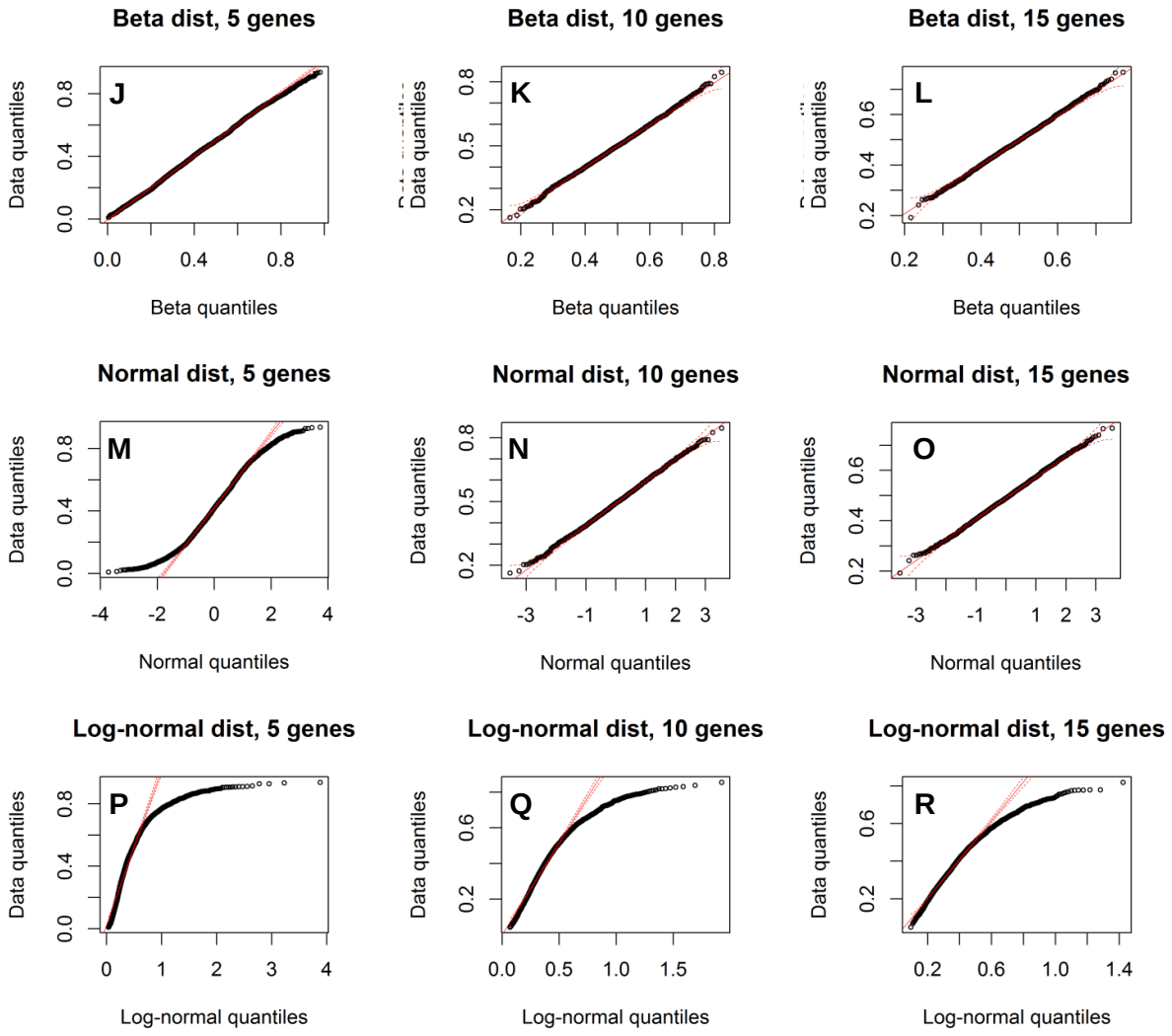


Figure S5 (continued): Corresponding plots for the negative prediction data for 5 genes (J, M, P), 10 genes (K, N, Q), and 15 genes (L, O, R) for the beta (J, K, L), normal (M, N, O) and log-normal (P, Q, R), distributions

Application of metaMIR to investigation of miRNA-mediated gene co-regulation

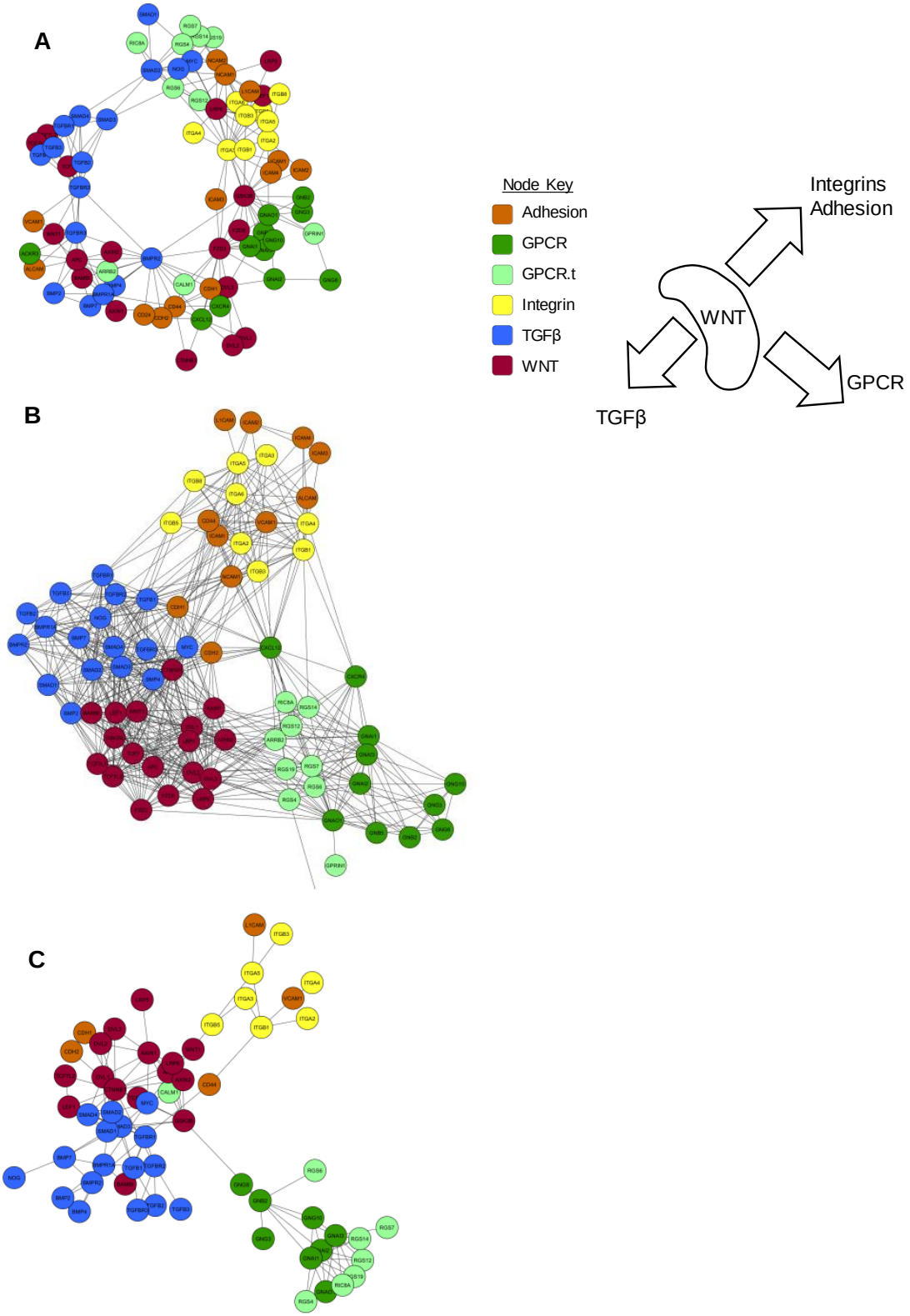
The main function of metaMIR is the identification of miRNA candidates that are predicted to regulate clusters of genes simultaneously. As described in the main text, the number of genes that can be analysed in a single pass is bound at both high and low ends to limit both the noise of the results (in the number of randomly identified clusters returned) and the computational expense of larger analyses. In addition to the focused examples described in the main text, a general analysis was performed, still in the domain of neural development, to compare the distances obtained from the metric used in the metaMIR clustering algorithm, with results from other networks.In the main text, focused examples of the investigation of Hippo/YAP signalling and NC-development-related transcription factors and TGFβ are presented. Here in addition, a broader analysis was also conducted employing the clustering function of metaMIR with genes involved in several signalling pathways and constructs related to neural development Table S2. These were adhesion molecules, G-protein coupled receptors (GPCRs), and TGFβ and Wnt signalling. Resulting networks were inspected with Cytoscape 3.3.0 (11) and subjected to the same automated network layout profiles for comparison Figure S7.

A premise of the clustering application is that there is sufficient information in patterns of miRNA predictions per gene to permit their unsupervised organization. The data of the reference table represent a regulatory information space, however its use of these data was able to partially recapitulate functional protein-protein networks, as shown by organizational groupings in comparison to STRING (12) and BioGRID (13) networks. Some of the grouping observed includes homologues, which by sequence similarity might be expected to have similar patterns of miRNA interaction (such as for TGFBR1, -2, and -3 or BMP2, -4, and -7). Across all transcripts, any such sequence similarities can at best be local, given the diverse range of sequence lengths (coding 207 – 8532 nt; 3'UTR 0 – 32870 nt).

Table S2: Groupings of genes used for analysis.

Grouping		Genes
Adhesion		*ALCAM, ICAM,NCAM2, CD44, ICAM1, ICAM4, L1CAM, CD24, ICAM2, NCAM1, VCAM1, CDH1, CDH2*
Integrins		*ITGA2, ITGA3, ITGA4, ITGA5, ITGA6, ITGB1, ITGB3, ITGB5, ITGB8*
Transforming Growth Factor β		*TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, TGFBR3, SMAD1, SMAD2, SMAD3, SMAD4, BMP2, BMP4, BMP7, BMPR1A, BMPR2, NOG, MYC*
WNT		*WNT1, FZD3, FZD6, LRP5, LRP6, BAMBI, DVL1, DVL2, DVL3, GSK3B, CTNNB1, AXIN1, AXIN2, APC, TCF7, TCF7L1, TCF7L2, LEF1*
G-Protein Coupled Receptors	Reception	*CXCR4, ACKR3, CXCL12, GNAO1, GNAI1, GNAI2, GNAI3, GNB2, GNB5, GNG3, GNG8, GNG10*
	Transduction	*ARRB2, GPRIN1, RIC8A, CALM1, RGS4, RGS6, RGS7, RGS12, RGS14, RGS19*
	Nuclear Effector	*AJUBA, WWTR1, YAP1, TEAD1, TEAD2, TP73*

Figure S7: MetaMIR clustering yields a network similar to function protein networks



Clustering results were extracted from metaMIR analysis (A) of genes involved in multiple pathways and functions associated with neural development. The resulting network was subjected to the same network layout algorithm and visually compared to networks using common genes from STRING (B) and BioGRID (C). See text for details.

Expanded analysis of inputs and reference dataset using protein-based data

As noted in the methods of the main text, the input prediction algorithms used in this study primarily focused on the 3'UTR of the target transcripts. Only two, Diana microT-CDS and PACCMIT-CDS, included data for predictions where the coding region of the transcript may be targeted.

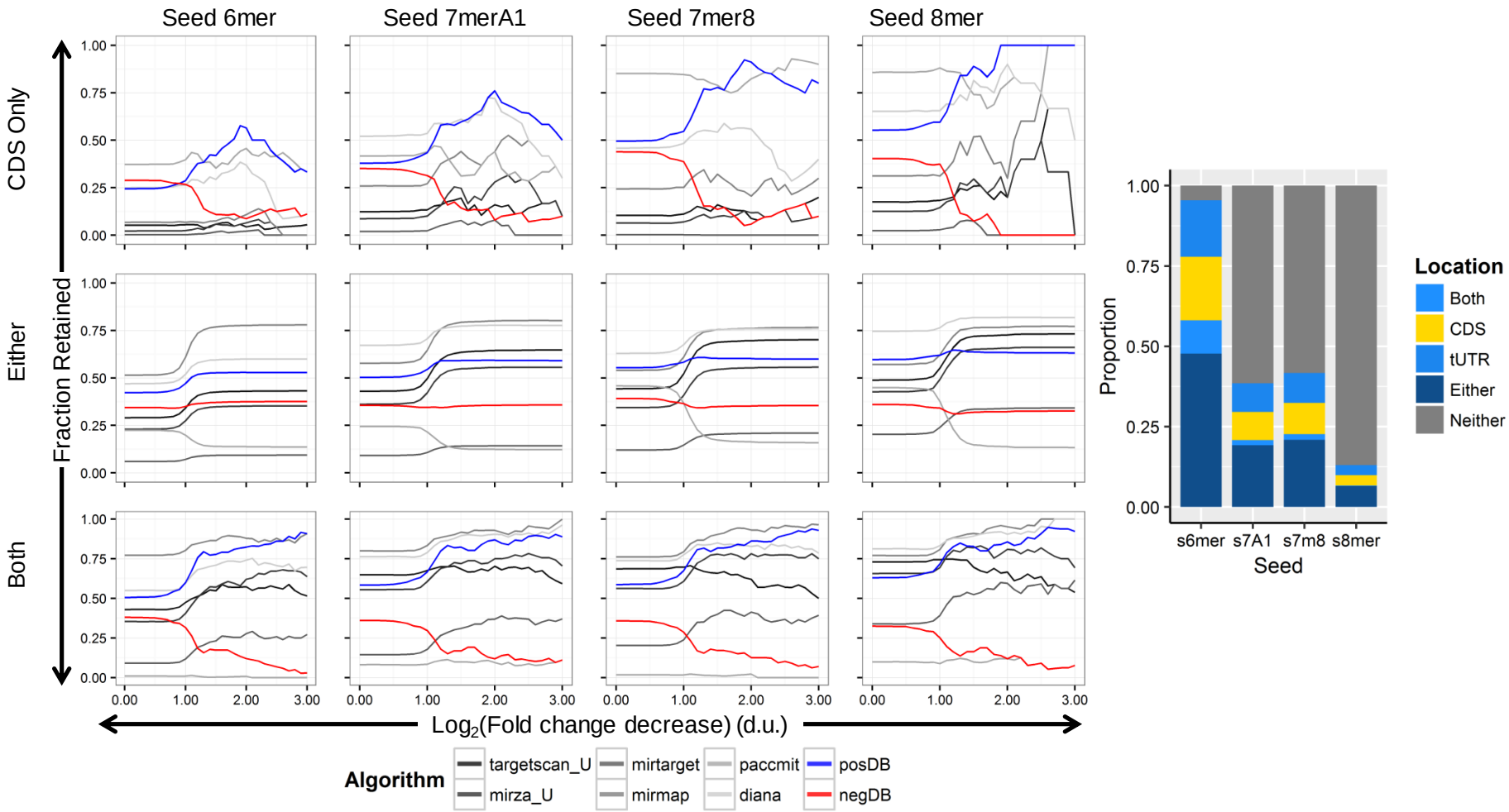
The experimentally-validated proteomics data were subjected to an expanded analysis to assess the potential influence on performance measure of the inputs and resulting reference database. Taking fold changes of protein expression as a measure of miRNA targeting, any interaction effects will reflect regulation in all potential transcript regions, and this regulation will be uncoupled from changes or lack of changes in transcript expression, in cases where transcript levels are unchanged, but protein expression is perturbed.

As a further stratification of the data, consideration was made for the location of potential seed match sites. None of the input algorithms depends exclusively on seed site matching, however the presence of a seed match of varying size can be taken as an indication of miRNA binding.

In the present study, an analysis was conducted to assess algorithm performance, indicated by prediction of the greatest influence on protein expression, subdivided by presence of seed match location (Figure S8). Seeds considered ranged from the most inclusive 6mer seed (matching nucleotides 2-7), the 7merA1 (matching miRNA nucleotides 2-7 with an A opposite position 1), 7mer8 (matching miRNA nucleotides 2-8), and 8mer (matching nucleotides 2-8 with an A opposite position 1. When only considering transcripts which contain a seed match in the coding region, input algorithm rankings are again adjusted, with PACCMIT and Diana typically the best performing (Figure S8, top row). For more comprehensive analysis, cases where both the coding region and the 3'UTR contain a site (middle row, Figure s6), and cases where either one of the other may contain a site are also indicated (bottom row, Figure S8). A further analysis, searching transcripts for putative seed sites reveals that the case of a seed match appearing only in the coding region and not in the 3'UTR is not uncommon, thereby suggesting that predictions based solely on the 3'UTR may miss some interactions.

Overlaid on these, the positive and negative reference databases demonstrate good performance, comparable to the best performing inputs. Such performance is advantageous, but non-essential to the performance of the reference set which is intended for classification, rather than fold change prediction tasks.

Figure S8: Inclusion of transcript coding sequences in miRNA target prediction analysis can improve results evaluated by proteomics data.



Transcripts for proteins identified in the proteomics data were screened for the major classes of miRNA seed sites. The proteomics data were first trimmed to retain transcripts with the indicated seed match position (coding region – CDS – only, either the CDS or 3' untranslated region – 3'UTR, or both). For the plots, the data were progressively trimmed such that only those transcripts exhibiting a fold change decrease above the threshold (x-axis) were kept. These proteomic data were compared to the data in the respective prediction algorithm data tables (to reveal which algorithms predict the highest measured protein fold changes). Indicated seeds are defined as: 6mer (match miRNA nucleotides 2-7), 7A1 (match 2-7, A opposite position 1), 7m8 (match 2-8) and 8mer (match 2-8, A opposite position 1). Databases which have included CDS prediction data, including the reference database of this study, outperform those that do not in cases of a CDS site and no 3'UTR. Assessment of the frequency suggests this occurrence comprises a notable proportion of seed-based target matches.

REFERENCES

1. Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N., Vlachos, I.S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T. and Hatzigeorgiou, A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows, *Nucleic Acids Res.,* **41,** 73.
2. Vejnar, C.E. and Zdobnov, E.M. (2012) MiRmap: comprehensive prediction of microRNA target repression strength, *Nucleic Acids Res.,* **40,** 11673–11683.
3. Gumienny, R. and Zavolan, M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G, *Nucleic acids research,* **43,** 1380–1391.
4. Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-Ligation studies, *Bioinformatics (Oxford, England)*, **32**, 1316-22.
5. Marín, R.M., Sulc, M. and Vanícek, J. (2013) Searching the coding region for microRNA targets, *RNA (New York, N.Y.),* **19,** 467–474.
6. Agarwal, V., Bell, G.W., Nam, J.-W. and Bartel, D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs, *eLife,* **4**.
7. Linsley, P.S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M.M., Bartz, S.R., Johnson, J.M., Cummins, J.M., Raymond, C.K. and Dai, H. *et al.* (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression, *Molecular and cellular biology,* **27,** 2240–2252.
8. Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing, *Mol. Cell,* **27,** 91–105.
9. Kuhn, M. (2008) Building Predictive Models in R Using thecaret Package, *Journal of Statistical Software,* **28**.
10. Cullen, A.C. and Frey, H.C. (1999) Probabilistic techniques in exposure assessment. A handbook for dealing with variability and uncertainty in models and inputs. Plenum Press, New York.
11. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome research,* **13,** 2498–2504.
12. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A. and Simonovic, M. *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic acids research,* **37,** 6.
13. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N. and O'Donnell, L. *et al.* (2015) The BioGRID interaction database: 2015 update, *Nucleic acids research,* **43,** 8.