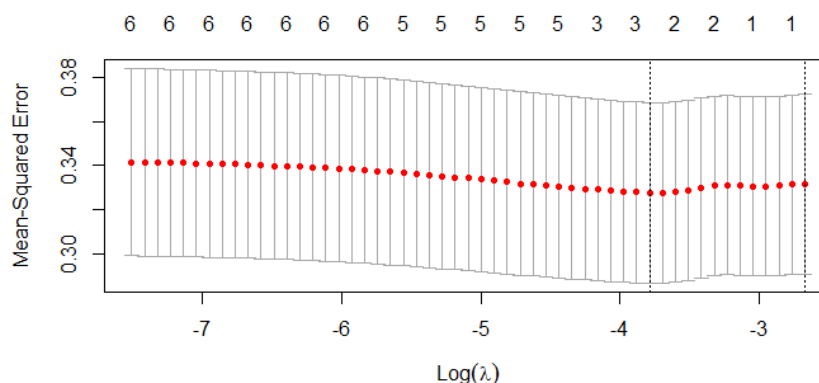


Supplemental Methods

To identify possible linear and non-linear combinations of candidate miRNAs, associated with the probability of developing distant metastases, the following statistical models and machine learning algorithms were also performed: (1) Cross-validated multivariable logistic model with Least Absolute Shrinkage and Selection Operator (LASSO) penalty; (2) Conditional Random Forest (CRF); (3) Support Vector Machine (SVM) as detailed below. It is worth mentioning that all miRNA variables (expression values) were first log-transformed (natural logarithm transformation) and then standardized (i.e., divided by their SD) before being used for statistical analysis.

Cross-validated multivariable logistic regressions with LASSO penalty

The consequence of imposing the LASSO penalty in a multivariable logistic regression model is to reduce (i.e., shrink) the coefficient values towards zero, allowing the less contributive variables to have a coefficient close to zero or equal to zero. This results in a natural selection of those predictors that produce the best weighted linear combination. The strength of using a logistic regression model is its interpretative simplicity and the possibility of expressing the relationship between predictors and disease through a simple mathematical formula. The choice of the penalty parameter value was determined as the one that achieves the lowest Mean Squared Error (MSE) after 10-fold cross validation of the data, as shown in the following figure:



The values of the penalty parameter (natural log value) are shown in X-axis whereas the corresponding 10-fold cross validated MSEs are shown in Y-axis. The number of nonzero coefficients detected at each penalty value is reported above this plot. This number starts at 6 (instead of 8) because it was found that when the penalty term was exactly zero, only 6 nonzero coefficients were estimated. The higher the value of the penalty parameter, the greater the shrinking effect. As shown, the optimal log penalty parameter value was found at -3.7854, which correspond to a penalty equal to 0.0227, and only three nonzero coefficients were detected for miR-3916, miR-3613-5p and miR-29c-3p. The estimated LASSO model was the following:

[Probability of metastases]= $\text{expit}(-1.403+0.0006065 \times \text{miR29c-3p}+0.1705286 \times \text{miR3613-5p}-0.3304211 \times \text{miR-3916})$

The Area Under the ROC Curve (AUC) computed by the LASSO using predicted probabilities was: 0.663 (95%CI: 0.573-0.754)

Conditional Random Forest

A decision tree algorithm recursively splits the data in subgroups, choosing the best binary split for each considered candidate predictor, to identify the most homogeneous (in terms of proportion of cases that developed metastases) and at the same time the most heterogeneous sets. The RF is a popular machine learning algorithm that consists in an ensemble of decision trees, where each tree is built on a random set of candidate variables and on a random set of bootstrap data portion. Each decision tree of the forest classifies the subjects in the out-of-bag (OOB) data (i.e. the data portion excluded for the tree building) as developing or not the metastases and each tree's classification is combined into a final classification through a "majority vote" mechanism. In the classical RF, the OOB data are also used to obtain a variable importance (VIMP) estimate, defined as the average over all trees of the difference in prediction error before and after the permutation of the values of each variable in the OOB data within each tree. However, the variable importance provided by the classical RF can be misleading in the case of correlated predictors (such as all miRNAs expressions) because a variable that has no effect of its own, but is correlated with a relevant predictor variable, can receive a high importance score. To this reason, although it is very computer intensive, the conditional RF (Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008) was performed instead of the classical RF so that the variable importance computed for a specific covariate is no longer affected by the correlation structure of all the other covariates. The relative VIMP (RVIMP) standardizes the importance values for ease of interpretation and is defined as the percent improvement with respect to the most important predictor. It was set that the conditional RF included a very large number of trees (10,000 trees) and that, for the building of each tree, 3 covariates were randomly selected.

Rank	miRNAs	Conditional VIMP	Conditional RVIMP
1	miR-3916	0.016267073	100.0%
2	miR-3613-5p	0.004595122	28.2%
3	miR-29c-3p	0.003540244	21.8%
4	miR-4286	0.003281707	20.2%
5	miR-422a-5p	0.002136585	13.1%
6	miR-504-5p	0.001869512	11.5%
7	miR-29-5p	0.001347561	8.3%
8	miR-27a-5p	-0.000354878	0.0%

According to conditional RF algorithm, miR-3916, miR-3613-5p, miR-29c-3p and miR-4286 were identified as the most associated miRNAs (conditional RVIMP>20%) to discriminate metastatic from non metastatic patients.

Support Vector Machine

SVM is a supervised machine learning algorithm allowing to solve binary classification issues, classifying both linear and non-linear data. It uses kernel functions to identify elements in a space without using coordinates but calculating the inner product between the images of all data pairs in the function space. This operation, called kernel trick, is often

computationally cheaper than the coordinate calculation. The most used kernel functions are the linear, radial and the polynomial (i.e. sigmoid) ones. SVM aims at generating a hyperplane with the main purpose to optimize the distance between the nearest points belonging to two different classes, known as support vectors. To reduce the risk of overfitting, the full dataset was randomly split into training (taking about 2/3 of the data) and test (taking the other 1/3 of the data) subsets. All SVMs were performed on training dataset and their performances were evaluated in test dataset. SVMs were performed both using the whole panel of 8 miRNAs and the “best 4 miRNAs”, found by the conditional RF algorithm. For each SVM, the optimal values for the “cost” and “gamma” (internal) parameters were detected following 10-fold cross-validation technique of the training dataset. The AUC was calculated using individual predicted probabilities of developing metastases, estimated from SVM (built on the training dataset).

Kernel	miRNAs	Best cost parameter	Best gamma parameter	Apparent AUC (95%CI) [training dataset]	Validated AUC (95%CI) [test dataset]
Linear	All 8 miRNAs	2^{-15}	2^{-3}	0.632 (0.523-0.741)	0.684 (0.536-0.831)
	The top 4 miRNAs only (i.e. miR3916, miR3613-5p, miR29c-3p and miR4286)	2^{13}	2^{-2}	0.594 (0.475-0.712)	0.591 (0.435-0.748)
Radial	All 8 miRNAs	2^4	2^{-3}	0.998 (0.993-1.000)	0.700 (0.527-0.872)
	The top 4 miRNAs only (i.e. miR3916, miR3613-5p, miR29c-3p and miR4286)	2^4	2^{-4}	0.799 (0.702-0.897)	0.651 (0.495-0.807)
Sigmoid	All 8 miRNAs	2^{-3}	1	0.617 (0.508-0.725)	0.614 (0.478-0.751)
	The top 4 miRNAs only (i.e. miR3916, miR3613-5p, miR29c-3p and miR4286)	2^{-2}	2^{-3}	0.525 (0.394-0.656)	0.503 (0.355-0.651)

As shown, the radial-basis kernel provided the best performance both using all the 8 miRNAs included in the panel and using the “best 4 miRNAs” only. It is worth noting that the very high AUC achieved in the training dataset using all miRNAs (i.e. AUC=0.998) is only apparent because in the test dataset this measure was reduced to 0.700. Instead, the AUC achieved using only the “best 4 miRNAs” achieved an optimistic AUC of 0.799 (in the training dataset) but it was more consistent with that estimated in the test dataset (AUC=0.651), which is, by definition, more conservative. To this reason, and following the principle of parsimony, the SVM built using the “best 4 miRNAs” only along with a radial-basis kernel was preferred.

All these supplemental statistical analyses were carried out using the R Foundation for Statistical Computing software (ver. 4.0, packages: “glmnet”, “party” and “e1071”).