Chameleon microRNAs in breast cancer: their elusive role as regulatory factors in cancer progression

Cesare Miglioli1,‡,*, Gaetan Bakalli1, Samuel Orso1, Mucyo Karemera1, Roberto Molinari2, Stéphane Guerrier1,3, Nabil Mili1,‡,*

1Geneva School of Economics and Management, University of Geneva.
2Department of Mathematics and Statistics, Auburn University.
3Faculty of Science, University of Geneva.
‡These authors contributed equally to this work.
∗Corresponding Address: Geneva School of Economics and Management, University of Geneva, Boulevard du Pont d’Arve 40, Geneva, Switzerland; cesare.miglioli@unige.ch and nabil.mili@unil.ch.

Abstract

Breast cancer is one of the most frequent cancers affecting women. Non-coding micro RNAs (miRNAs) seem to play an important role in the regulation of pathways involved in tumor occurrence and progression. Extending on the research in Haakensen et al., where significant miRNAs were selected as being associated with the progression from normal breast tissue to breast cancer, in this work we put forward 112 sets of miRNA combinations, each including at most 5 expressions with high accuracy in discriminating healthy breast tissue from breast carcinoma. Our results are based on a recently developed machine learning technique which, instead of selecting a single model (or combination of features), delivers a set of models with equivalent predictive capabilities that allow to interpret and visualize the interaction of these features. These results shed new light on the biological action of the selected miRNAs which can behave in different ways according to the miRNA network with which they interact. Indeed, these revealed connections may contribute to explain why, in some cases, different studies attribute opposite functions to the same miRNA. It is therefore possible to understand how the role of a genomic variable may change when considered in interaction with other sets of variables, as opposed to only considering its effect when it is evaluated within a unique combination of features. The approach proposed in this work provides a statistical basis for the notion of chameleon miRNAs and is inspired by the emerging field of systems biology.

Author Summary

• The notion of a single predictive genomic (statistical) model is replaced by that of a set of models that can be considered as exchangeable due to their indistinguishable (optimal) predictive abilities;

• Our results indicate that the role of miRNAs cannot be interpreted independently from the combination of features with which they interact and can therefore vary considerably when considered in a network of different combinations. Some miRNAs may act as chameleons and behave in opposite manners thereby showing some kind of antagonistic duality;

• Some miRNAs are exchangeable inside models with equivalent predictive ability and seem to point to latent biological functions.
Introduction

Non-coding RNA and breast cancer

Breast cancer (BC) is the second-most common cancer and second-leading cause of cancer mortality in American women. In the USA, its incidence in 2019 was roughly 268,600 and it is responsible for an estimated 41,760 deaths [2] to which one must add 62,930 new cases of Ductal Carcinoma In Situ (DCIS). Dysregulation of microRNAs (miRNAs) plays a key role in almost all cancers, including BC [3]. miRNAs are short endogenous noncoding RNAs that regulate their target messenger RNAs (mRNA) by promoting mRNA degradation or repressing translation. Chang et al. [3] found that increased expression of 12 mature miRNAs — hsa-miR-320a, hsa-miR-361-5p, hsa-miR-103a-3p, hsa-miR-21-5p, hsa-miR-374b-5p, hsa-miR-140-3p, hsa-miR-25-3p, hsa-miR-651-5p, hsa-miR-200c-3p, hsa-miR-30a-5p, hsa-miR-30c-5p, and hsa-let-7i-5p — all predicted improved BC survival. In a recent review, Adhami et al. [4] determined that two miRNAs (hsa-miR-21 and hsa-miR-210) were upregulated consistently and six miRNAs (hsa-miR-145, hsa-miR-139-5p, hsa-miR-195, hsa-miR-99a, hsa-miR-497 and hsa-miR-205) were downregulated in at least three studies. In another study, Haakensen et al. [1] identified some miRNA alterations during BC progression. These alterations were involved in the invasive signatures of BC including downregulation of hsa-miR-139-5p in aggressive subtypes and upregulation of hsa-miR-29c-5p in luminal subtypes. A total of 27 miRNAs were implicated in their proposed DCIS signature.

The latter study provided one of the main reasons to develop the work presented here. Indeed, Haakensen et al. [1] provide the following statement in their work: “hsa-miRNA-210-3p was significantly upregulated in both our analyses, but was downregulated in the same transition in Volinia et al [13] and is hence excluded from our proposed signature”. In the wake of this statement, hsa-miR-210 has previously been identified as a marker of poor prognosis in BC and other carcinomas [7]. In fact, Volinia et al. [13] found hsa-miR-210 to be downregulated in DCIS compared with normal breast tissue, but upregulated in invasive carcinomas compared with DCIS. In addition, Shao et al. [23] recently showed that hsa-miR-210 is associated with internal organ metastasis (liver, lung, and brain) in BC. In Haakensen's study, hsa-miR-210 was upregulated in DCIS compared to normal tissue and was not detected as significantly altered in any invasive subtype. Given the unclear role of this miRNA in breast carcinogenesis, Haakensen's study therefore discarded it from the list of miRNAs involved in BC progression.

Considering these studies, this work aims at addressing some questions that naturally arise from their conclusions. The first of these questions is as follows: why weren't there any concordant findings in many research studies regarding the role of miRNAs in the progression of BC? Are the different outcomes due to population selection, batch effect or to biological causes such as disease heterogeneity, overlapping of miRNA functions or network effects? A second question, that stems from the latter points, is the following: could an miRNA have either an activating effect or an inhibiting one in a given biological process (such as cancer progression) according to the other miRNAs with which it interacts? In other words, could a specific miRNA be upregulated in one study and downregulated in another as a result of the complex pathways in which it is involved (instead of this being the effect of the experimental conditions)? This work aims at investigating these questions more thoroughly and, for this purpose, we analyze the AHUS dataset using a recently proposed machine learning algorithm. The dataset is made available by Haakensen et al. on the open access ArrayExpress platform at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3759/?query=AHUS.
Genomic connections in systems biology

The use of mathematical and computational models in biology is referred to as systems biology (SB) [21]. SB is a field which, among others, focuses on the assumption that a discrete biological function can rarely be attributed to a single molecule [20]. Instead, most biological characteristics arise from complex interactions among the cell's numerous constituents, such as proteins, DNA, RNA and small molecules. Understanding the structure and the dynamics of complex intercellular networks that contribute to the structure and function of a living cell is therefore mandatory before assigning a function to any biological feature [19]. According to Barabási et al. [18], the inter- and intra-cellular connectivity implies that the impact of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but can spread along the links of the network and alter the activity of gene products that otherwise carry no defects. However the biological networks in which a single genomic variable is involved remain unknown. As a first step, one should then rely on a data-driven network built using statistical (and not biological) associations.

Research questions

Our study has three goals: (i) to investigate if, and to what level of accuracy, it is possible to use different combinations of miRNAs as biomarkers to discriminate normal breast tissue from breast carcinoma; (ii) to check how the behavior of these miRNAs varies according to the specific combination with which it interacts; (iii) to search for interchangeable miRNAs in these predictive models and by doing so, to decipher the biological targets of these variables.

Materials and Methods

Genomic study

The results of our research are based on the AHUS dataset made available on the ArrayExpress platform by Haakensen et al. [1]. According to the authors, in order to collect this data the Akershus University Hospital sequentially collected breast tissue specimens from BC patients and from women undergoing surgery for breast reduction. These specimens were collected over 7 years between 2003 and 2009. miRNA-expression profiling was obtained for 55 invasive carcinomas and 70 normal breast tissue samples (including 29 tumor-adjacent normal tissue samples and 41 breast reduction samples) for a total of 125 as stated on the ArrayExpress platform. The samples were hybridized on Agilent 8x15K arrays (Agilent Technologies, Santa Clara, CA), catalogue number 4470B (v2) and 4470C (v3), and the features were extracted using Agilent Feature Extraction. Relevant information can be found in Haakensen et al. [1].

Statistical analysis

When considering the research goals defined earlier, the statistical tools used to achieve them need to be defined accordingly. Hence, the first step is to find “different combinations of miRNAs” which implies that we are not aiming to find a single statistical (machine learning) model to classify normal breast tissue and breast carcinoma. Indeed, we intend to find a variety of models (miRNA combinations) that all perform this classification task with a high level of accuracy and renders them equivalent in terms of predictive power. The idea of considering a multitude of models is not a common one but has been put forward in different settings (see e.g. Caruana et al. [55]) and was adequately stressed, for example, in Whittingham et al. [56] who state that “…/... further analysis should not be based on a single best model, but should explicitly acknowledge uncertainty among models that are similarly consistent with the data”. In fact, depending on the setting, the reliance on a single model can be rather risky and can often deliver contradicting results regarding if and how certain
variables contribute to explain or predict a given phenomenon of interest. In this perspective, we should choose an approach that allows to find a variety of "strong" models and that, in accordance with the subsequent research goals of this work, can be used to create miRNA networks highlighting how, for example, a specific miRNA can be used to detect (and can contribute differently to) breast carcinoma when considered with other miRNA combinations. In addition, in order to create networks that can also be interpreted from a biological perspective, we need these models also to be based on small (sparse) combinations of miRNAs.

There exist a wide variety of statistical and machine learning approaches to select and estimate models with few features (miRNAs) but, in most cases, these only select one model which therefore limits the possibility of considering how the impact of an miRNA can change when considered with another set of variables. For this reason this work proposes to use the "Sparse Wrapper AlGorithm" (SWAG) put forward in Molinari et al. [57] which is described in the following paragraphs.

**A wrapper method for sparse learning**

The SWAG is a method derived from the Panning algorithm presented in Guerrier et al. [9] for gene selection problems. The premise of this method is the assumption that, in order to adequately predict a certain outcome of interest (e.g. breast carcinoma), we only need an extremely small set of features and that there are many models (combinations of small sets of features) that can all have equivalent and high predictive power. Aside from allowing to understand if and how certain features can behave differently when considered in presence with other sets of features, the output of the SWAG also allows to facilitate replicability of results. Indeed, when a study proposes a single model (and hence a single combination of features) in order to detect or predict a certain response, this may not be usable for a research or medical structure that may not have the possibility of measuring all the selected features.

In order to respond to the above needs, the SWAG consists in a "greedy" wrapper algorithm that firstly requires the user to specify a model (or learning method), such as a logistic model, as well as the maximum number of variables \( p_{\text{max}} \) to be considered within such a model. The latter choice can be made, for example, based on prior knowledge of the problem and interpretability requirements (the smaller this number, the easier the output will be interpreted). Based on these choices and supposing there is a total of \( p \) features (e.g. biomarkers), the SWAG starts through a first screening step where \( p \) models are built, each using a distinct feature. At this stage, the out-of-sample prediction error of each model can be estimated via \( k \)-fold cross-validation (CV) repeated \( m \) times and the best of these models (in terms of lowest prediction error) can be selected thereby providing a list of features that, on their own, appear to be highly predictive for the considered response. The definition of "best" models will be given by the user through a parameter \( \alpha \) which represents a proportion (or percentile) and is usually chosen to be considerably small (i.e. between 0.01 and 0.1). With smaller values of \( \alpha \) implying a more strict selection of best models (hence the choice of only the most performing features), the SWAG then uses the features selected in the first step to progressively build higher-dimensional models (i.e. models with an increased number of feature combinations within them) until it reaches the maximum number \( p_{\text{max}} \). When building the models for a given dimension, the SWAG takes the best models from the previous step (i.e. the step that built models with one less feature than the current step) and randomly adds a distinct feature from the set of features selected at the first step. Having built \( m \) models at each step (where \( m \) is also chosen by the user), the final output of the SWAG is a set of “strong” models (i.e. models with high predictive power) where each is based on a combination of 1 to \( p_{\text{max}} \) features. A simplified representation of the SWAG is presented in Fig 1. With this output, it is then possible for the user to apply post-processing to select a subset of interest from this set of models. The SWAG is made available as an R package on the CRAN repository. At the same time, a development version can also be found at
Fig 1. SWAG flowchart.
A schematic representation of how the SWAG was calibrated for this work. The red set represents the first step which evaluates every one-dimensional model and selects the best expressions to be used in the general step represented by the blue set. The latter step evaluates and selects the best models of dimension 2 to $p_{\text{max}}$.

Implementation
The AHUS dataset is split into training and test subsets. The training subset contains 100 observations with a 56/44 split (normal tissue/invasive BC). The test subset has 25 observations with a 14/11 split. With CV, we estimate how accurately a predictive model will perform in practice. This is a well-established model validation technique for assessing how the results of a statistical analysis will generalize to an independent dataset. The aim is to estimate how accurately a model will perform in terms of prediction and the rationale of using this technique may be found in Fushiki [39] and Molinaro [10]. The caveats of CV are well explained in Bernau et al. [40] where the main setback can eventually consist in an overestimation of model performance in a broader application context.

Our research question is not about the validity of a given model selection method, but about the function of a specific miRNA in BC oncogenesis. To address this issue, we must not only select a set of models in which this specific miRNA is involved, but also determine the direction in which it acts (oncogenic or protective effect). Prior to this analysis, we have standardized the design matrix (miRNAs) to ensure a meaningful comparison across the different models. Then, to assess the role of the relevant miRNAs selected via the SWAG, we computed their $\beta$ coefficients by performing a logistic regression on each element of the set of models. The evaluation of the $\beta$ coefficients allows us to (i) identify either the oncogenic or protective effect of the variable and (ii) gain insight on its distribution. Positive values of $\beta$ mean that the miRNA associated with this coefficient has an activating effect on tumor progression (oncogenic effect); a negative value means the opposite (protective effect). Since the miRNAs can be included in different models delivered by the SWAG, one can compute an empirical distribution of the coefficients.

Single and associative effects on a binary variable
One of our research questions is whether a given miRNA has the same action (oncogenic or protective) when it is taken in isolation or when embedded within different models and feature combinations. In order to assess the biological action of the selected miRNAs, we compute both the single and associative effects of the $\beta$ coefficients for each of the selected miRNAs. A single effect is measured by the estimated value of a $\beta$ coefficient when considering a single miRNA in the logistic model. The associative effect is defined as all the different values that a $\beta$ coefficient takes when constructing the set of models with the SWAG. The associative effect of a specific miRNA therefore may be seen as an indicator of its biological impact in a broader context.

As a matter of fact, according to Cox [17], these effects are typically different. For a random variable considered both alone and conditionally on a confounding variable $W$, the single and associative inferences may have opposite signs by the so-called Yule–Simpson effect [16]. This effect, if observed, may be explained in two ways: (i) the existence of sub-populations or (ii) the influence of a finite set of latent classes $W$ (such as biological functions) within the population under study. An example of this phenomenon with some mathematical explanations can be found in Cox [17] and Boehm et al. [38]. Splitting the population into defined subgroups may, to some extent, dodge the first pitfall (i.e., the existence of sub-populations). However, BC heterogeneity is large and has been documented in terms of
different histological subtypes, treatment sensitivity profiles, and clinical outcomes. Furthermore, the heterogeneous expression of the oestrogen receptor, progesterone receptor, and HER2 has been reported in different areas of the same tumour. Molecular profiling studies have confirmed that spatial and temporal intratumour heterogeneity of BCs exist at a level beyond common expectations [41]. Splitting BC populations into subtypes may then be a misleading precaution. The second pitfall (i.e. the existence of latent biological functions shared by many genomic features) is even more elusive. Having not a single but a paradigmatic set of predictive models may help get around this hurdle. We addressed this challenge (differentiating the effect of sub-populations from that of latent biological variables on single and associative coefficients) by mixing the 55 invasive carcinomas into one category. If single and associative coefficients retain the same sign throughout the 112 selected models, we can conclude with some confidence that the existence of sub-populations (BC sub-types) has no effect on the oncogenic or protective effect of the relevant miRNAs inside the AHUS dataset. On the contrary, if single and associative coefficients have opposite signs, then one can assume that the effect of the relevant miRNA differs according to its environment.

**Horizontal and vertical organizations**

We make the heuristic hypothesis that the human genome as a whole and its sub-units (such as non-coding RNAs) can be interpreted as semiotic systems. To give meaning to the miRNA-based net-like structures that we build through our statistical analysis, we borrow the notions of syntagm and paradigm from structural semiotic analysis, inspired by de Saussure theory [43]. A simple and useful introduction to semiotics may be found in Chandler [77]. De Saussure emphasized that meaning (in our case, oncogenic or protective effect) arises from differences between signifiers (in our case, miRNAs). These differences are of two kinds: syntagmatic (concerning positioning within a model) and paradigmatic (concerning substitution within a given model). These two dimensions are often presented as axes, where the horizontal axis is the syntagmatic and the vertical axis is the paradigmatic. The plane of the syntagm is that of the combination of signifiers (i.e. selected miRNAs) within a statistical model whilst the plane of the paradigm is that of the selection of signifiers. Whilst syntagmatic relations are combination possibilities, paradigmatic relations are functional contrasts. The meaning of a signifier is determined by both its paradigmatic and syntagmatic relations. According to this conception, the set made of all the selected models may be seen as the set of syntagmatic "sentences" selected by the SWAG, and the set made of the selected miRNAs as the set of paradigmatic Omics features. In this study, the horizontal syntagmatic axis will be used to tackle the second research question (to check how the behavior of the miRNAs varies according to the specific combination with which it interacts). The vertical paradigmatic axis will be used to address the third research question (to search for interchangeable miRNAs in these predictive models and by doing so, to decipher the biological targets of these miRNAs).

**Results**

**Breast cancer / normal tissue discrimination**

When applying SWAG to the AHUS data, a total of 45 miRNAs were selected, making a set of 112 indistinguishable models of sizes 4 to 5. These 112 models and the 45 selected miRNAs are displayed in the supporting information material. They perform similarly or outperform the lasso, a standard model selection method used in genomics [42], with less than half the number of miRNAs selected by the latter. This is evident from the comparison in terms of accuracy, sensitivity, specificity as well as positive and negative predictive values at the standard logistic cut-off level of 0.5 (cf. Fig 2). Indeed all the 112 SWAG models have an equal
or greater accuracy than the lasso. Moreover a large majority of SWAG models (i.e. 97 out of 112) dominate lasso in any of the considered metrics. This can be inferred also visually in Fig 2 by looking at the barplots on the right of each specific SWAG range (i.e. the interval between the smallest and largest values among all SWAG models), with the vertical red line representing the corresponding lasso value. As a whole, these results suggest that SWAG more precisely targets the set of miRNAs involved in BC progression. Furthermore, Tab 1 reports two specific SWAG models, among the 112 selected ones, that achieve a perfect out-of-sample classification in terms of area under the curve (AUC). It is therefore possible to discriminate BC from normal breast tissue with extreme accuracy using miRNAs as biomarkers. The added value of the SWAG compared to lasso is that (i) it produces a set of equivalent models instead of a single one and (ii) the number of selected variables per model is smaller by a factor of two, making the models more easily interpretable.

**Fig 2. Comparison between Lasso and SWAG.**
We compare accuracy, sensitivity, specificity, negative predictive value (NPP), positive predictive value (PPP), number (#) of miRNAs of the lasso with the ranges (i.e. smallest-to-largest value intervals) of the same metrics for the 112 SWAG models. On the right of each SWAG range, a barplot illustrates the distribution of the specific metrics for the 112 considered models with a vertical red line representing the corresponding value for the lasso. All evaluations have been made at the standard 0.5 cut-off of logistic regression.

**Table 1. Best out-of-sample SWAG models that achieve a perfect classification in terms of area under the curve (i.e. AUC = 1).**

| Non-coding miRNAs |
|-------------------|
| Model 1  | hsa-miR-1274a hsa-miR-21 **hsa-miR-92a** hsa-miR-328 hsa-miR-140-3p |
| Model 2  | hsa-miR-1274a hsa-miR-21 **hsa-miR-92a** hsa-miR-328 **hsa-miR-30b** |

The non-coding miRNAs are displayed in order of presence in the SWAG chain. The chameleon miRNAs are presented in bold.

Among the 45 selected miRNAs, 8 were present in more than 16 % of all SWAG models. These 8 miRNAs are displayed in Tab 2, with their respective pairwise Spearman correlations, illustrated in Fig 3. As a non-parametric measure of rank correlation, Spearman correlation assesses how well the relationship between two variables can be described using a monotonic function. It is used in our study as an index of result consistency. Two miRNAs having similar effects on cancer progression should be positively correlated. The rational of this statement is that correlation and mutual information are closely related [76]. Selected miRNAs in each of the 112 statistical models are presented with single and associative $\beta$ coefficients which are in most cases either positive (oncogenic effect) or negative (protective effect on tumor progression). However, one of them stands out from the others, hsa-miR-92a, whose associative $\beta$ is positive and whose single one is negative (cf. Tab 3). This point will be discussed later.
Table 2. Model occurrence rate of the most frequently selected miRNAs.

| Non-coding miRNA | Model occurrence rate (%) |
|------------------|---------------------------|
| hsa-miR-1274a    | 75.0                      |
| hsa-miR-21       | 74.1                      |
| hsa-miR-139-3p   | 44.6                      |
| hsa-miR-125b-2*  | 39.3                      |
| hsa-miR-92a      | 25.0                      |
| hsa-miR-449a     | 22.3                      |
| hsa-miR-155      | 18.8                      |
| hsa-miR-200c     | 16.1                      |

Fig 3. Spearman correlation ($\hat{\rho}$) matrix for the 8 most frequent miRNAs selected by the SWAG.

The upper triangular part shows the estimator $\hat{\rho}$ between miRNAs and their respective p-value computed via non-parametric bootstrap. The color of the boxes indicates the direction of $\hat{\rho}$ (blue for positive correlation and red for negative). The lower triangular part illustrates the bootstrap distribution of $\hat{\rho}$ via the density plot, with the dark circle being the estimator of $\hat{\rho}$ and the horizontal black line its 95% confidence interval. The star on the boxes’ upper-left indicates the level for which the correlation is significant.

Oncogenic or protective role of miRNAs: the syntagmatic axis

The values of the single and associative coefficients of the eight most frequently selected miRNAs are displayed in Fig 4 and Tab 3. Coefficients related to miRNAs that are present in at least 10% of models and that show discordant values between single and associative $\beta$s are presented in Tab 4. Among the eight most frequently selected miRNAs, the median value of the $\beta$s considered along with their range makes it possible to identify three classes of miRNA (cf. Fig 4 and Tab 3): (i) oncogenic miRNAs with single and associative positive values of $\beta$ (i.e. hsa-miR-1274a, hsa-miR-21, hsa-miR-449a, hsa-miR-155, hsa-miR-200c); (ii) protective miRNAs (hsa-miR-139-3p, hsa-miR-125b-2*); (iii) an undefined miRNA (hsa-miR-92a) with a positive single coefficient and a negative associative one. An indication of the consistency of these results lies in the Spearman correlation coefficients $\hat{\rho}$ (cf. Fig 3): hsa-miR-1274a is significantly positively correlated (at the level $\alpha = 5\%$) to its oncogenic partner hsa-miR-155 and negatively correlated to protective miRNAs hsa-miR-125b-2* and hsa-miR-139-3p. The same consistency can be observed for hsa-miR-21. The two protective miRNAs, hsa-miR-139-3p and hsa-miR-125b-2*, are also significantly positively correlated with each other, and significantly negatively correlated to oncogenic miRNAs such as hsa-miR-449a and hsa-miR-1274a.

The case of hsa-miR-92a which displays discordant single and associative coefficients is not isolated. Among the miRNAs selected in at least 10% of the models, three show a similar behaviour to that of hsa-miR-92a, hsa-miR-320d, hsa-miR-193a-5p and hsa-miR-30b.

Using a bioinformatics-based interaction analysis of hsa-miR-92a-3p and key genes in tamoxifen-resistant BC cells, Cun et al. [69] found that hsa-miR-92a-3p was higher in BC serum or tissue than in healthy volunteer serum or adjacent normal tissue. Hence, high expression of hsa-miR-92a-3p seems to predict poor prognosis for BC patients according to this meta-analysis study. This finding is in contradiction with previous results published by Nilsson et al. [70] that suggest that downregulation of hsa-miR-92a-3p is associated with aggressive BC features and increased tumour macrophage infiltration. In relation to
hsa-miR-320d action in BC, Cava et al. [86] found that its downregulation favours BC progression. To the best of our knowledge, no other study has investigated the role of hsa-miR-320d in BC progression therefore it is not possible to compare our result with other data coming from recent literature. According to Maltseva et al. [87], hsa-miR-193a-5p is less expressed in inflammatory BC patients and is known to play a suppressive role in BC. This statement is in contradiction with the findings in Li et al. [88] that state that long non-coding RNA small nucleolar RNA host gene 1 (SNG1) activates HOXA1 expression via sponging hsa-miR-193a-5p in BC progression. Finally, the role of hsa-miR-30b has been shown to be versatile, as a recent review points out [89]. Members of the hsa-miR-30 family play a role in the regulation of tumorigenesis, interference with tumor invasion and metastasis as well as reversal of drug resistance. Nevertheless some hsa-miR-30 family members have independent protective effects on the prognosis of BC patients. In conclusion, a versatile role of miRNAs in BC progression is quite a common finding in recent literature. Our results give a statistical basis to this allegation.

**Fig 4. Distribution of β coefficients for the most frequently selected miRNAs.**
We present the single effect (i.e. the estimated value of a β coefficient when considering a single miRNA in the logistic model), the median and range of the associative effect (i.e. the different values that a miRNA specific β coefficient takes in each of the 112 SWAG models) for each of the most frequently selected miRNAs displayed in both Tab 3 and Tab 4.

**Table 3. Single and associative coefficients (median values and range) for the eight most frequently selected miRNAs.**

| miRNA           | Single β | Median associative β | Associative β Range     |
|-----------------|----------|-----------------------|-------------------------|
| hsa miR-1274a   | 1.427    | 2.120                 | (0.768;3.392)           |
| hsa–miR–21      | 1.996    | 3.174                 | (1.858;4.880)           |
| hsa miR-139-3p  | −2.191   | −0.979                | (−1.799;−0.443)         |
| hsa–miR–125b–2* | −1.106   | −1.510                | (−2.451;−1.003)         |
| hsa–miR–92a     | −0.736   | 0.939                 | (0.095;1.315)           |
| hsa–miR–449a    | 3.672    | 1.228                 | (0.379;2.644)           |
| hsa–miR–155     | 2.973    | 1.806                 | (0.294;1.920)           |
| hsa miR-200c    | 1.628    | 1.600                 | (0.907;2.153)           |

**Table 4. Single and associative coefficients (median values and range) for the "chameleon" miRNAs present in at least 10% of the models.**

| miRNA           | Single β | Median associative β | Associative β Range     |
|-----------------|----------|-----------------------|-------------------------|
| hsa-miR-92a     | −0.736   | 0.939                 | (0.095;1.315)           |
| hsa-miR-320d    | −1.064   | 1.174                 | (−0.412;2.543)          |
| hsa-miR-193a-5p | −1.734   | 1.114                 | (−1.318,1.225)          |
| hsa-miR-30b     | 0.694    | −0.753                | (−1.422;0.595)          |

Associative ranges and not confidence intervals are shown since some of the coefficients display a bi-modal distribution.
An example of paradigmatic substitution: hsa-miR-140-3p and hsa-miR-30b

Among the 112 selected models, two of them show a perfect discriminating power (cf. Tab 1). These two differ only by one miRNA: hsa-miR-140-3p and hsa-miR-30b respectively. From a linguistic point of view, these two miRNAs can be seen as synonyms, meaning that they can be swapped without affecting the meaning of the "sentence" (the predictive power of the model). As stated previously, miRNAs are short endogenous noncoding RNAs that regulate their target messenger RNAs by promoting their degradation or by repressing their translation [44]. It is therefore intuitive to look at the set of target miRNAs associated with each of these regulatory factors and to determine their intersection. By doing so, we can create a list of common target genes. We acquired information on the targets of these two miRNAs from the http://mirbase.org/ platform [90], [91], [92], [93], [94], and [95]. The targets of these two miRNAs were obtained by crossing information from the TargetScanvert database (http://www.targetscan.org/) [83], [85] and the miRDB database (http://mirdb.org/) [84]. The results are shown in Tab 5.

### Table 5. hsa-miR-140-3p and hsa-miR-30b common targets.

| Gene symbol | Gene name |
|-------------|-----------|
| NDST1       | N-deacetylase and N-sulfotransferase 1 |
| CCNT2       | Cyclin T2 |
| USP49       | Ubiquitin specific peptidase 49 |
| ZC3H6       | Zinc finger CCCH-type containing 6 |
| DTX4        | Deltex E3 ubiquitin ligase 4 |
| CDK6        | Cyclin dependent kinase 6 |
| SRGAP3      | SLIT-ROBO Rho GTPase activating protein 3 |
| RAB21       | RAB21, member RAS oncogene family |
| P2RY2       | Purinergic receptor P2Y2 |
| GLG1        | Golgi glycoprotein 1 |
| KCNB1       | Potassium voltage-gated channel subfamily B member 1 |
| TNKS        | Tankyrase |
| UBN2        | Ubinuclein 2 |
| RFT1        | RFT1 homolog |
| ARID2       | AT-rich interaction domain 2 |
| TYRO3       | TYRO3 protein tyrosine kinase |
| EIF5A2      | Eukaryotic translation initiation factor 5A2 |
| HPCAL4      | Hippocalcin like 4 |

These targets were selected by crossing information coming from the http://mirdb.org/ and http://www.targetscan.org/ databases

Among these target genes common to hsa-miR-140-3p and hsa-miR-30b, some are well known to play a pivotal role in cancer progression. In this direction, one can cite USP49 [96], CDK6 [97], RAB21 [98], P2RY2 [99], TNKS [100], ARID2 [101], TYRO3 [102], EIF5A2 [103]. This set of common targets may explain why these two miRNAs are "synonyms" and can be exchanged in predictive models without any harm. The latent functions of these putative
target genes is shown in Tab 6.

**Table 6. Function of hsa-miR-140-3p and hsa-miR-30b targets involved in cancer pathophysiology.**

| Gene symbol | Gene Function                                      |
|-------------|---------------------------------------------------|
| USP49       | Histone H2B lysine deubiquitination / mRNA splicing |
| CDK6        | Cell dedifferentiation / cell division             |
| RAB21       | Rab protein signal transduction                    |
| P2RY2       | Cellular ion homeostasis / cellular response to ATP |
| TNKS        | Cell division / mitotic spindle organisation       |
| ARID2       | Negative regulation of cell migration and cell population proliferation |
| TYRO3       | Apoptotic cell clearance / cell adhesion and migration |
| EIF5A2      | mRNA transport / regulation of cell population proliferation |

Source: [https://ensembl.org/](https://ensembl.org/) and [https://uniprot.org/](https://uniprot.org/)

**Discussion**

With regards to our research questions, we can firstly conclude that it is possible to differentiate normal breast tissue from breast carcinoma by using miRNAs as biomarkers with reasonable sensitivity and specificity. Secondly, some selected miRNAs behave in opposite ways according to the models in which they are embedded. We decided to call these miRNAs, whose action is conditioned by their insertion in a syntagmatic axis, *chameleon* micro RNAs. Thirdly, any model selection method such as the one used for this work (SWAG) that gives the opportunity to build "horizontal" and "vertical" axes can point to latent biological functions and help researchers develop new hypotheses. In our case, with regards to hsa-miR-140-3p and hsa-miR-30b, some latent cell functions such as cell division and differentiation, mRNA splicing and transport as well as cellular ion homeostasis appear to be highly relevant. According to Stepanenko *et al.* [29], cancer evolution is a stochastic process both at the genome and gene levels. Most tumors contain multiple genetic subclones, evolving in either succession or in parallel, either in a linear or branching manner, with heterogeneous genome and gene alterations, extensively rewired signaling networks, and addicted to multiple oncogenes easily switching with each other during cancer progression and medical intervention. Hundreds of discovered cancer genes or gene products are classified according to whether they function in an oncogenic or protective manner in a cancer cell. However, there are many cancer "*gene-chameleons*", which behave in opposite manners in different experimental settings showing what Stepanenko calls "antagonistic duality". These statements find confirmation in our study. This antagonistic duality affects not only genes, but also miRNAs. For this subgroup, the distribution of the $\beta$ coefficients, either single or associative, include the value zero thereby indicating an ambiguous or dualistic effect. These results are in line with the most recent literature about their action in BC progression. Further research taking into account this versatile effect according to net-like structures is needed. Our research places itself within the emerging field of artificial intelligence [11]. With the advent of Big Data and the ever-increasing storage and computing power, the challenge has shifted from collecting data to turning it into meaningful and actionable insights. This challenge requires that we leave on the side of the road statistical methods that select genomic items taken in isolation and that we favour methods that scrutinize biological...
systems. By producing net-like combinations of equivalent models, it is possible to shed light on the latent biological confounding variables that are usually ignored and may reverse the effect of the considered Omics feature.

**Added value**

The added value of our research is fourfold:

- Predictive models with indistinguishable (optimal) predictive abilities are not unique, but belong to a set of equivalent and, in some sense, exchangeable models.

- Our results indicate that miRNAs are not isolated items but are integrated into two-dimension statistical axes. Their function cannot be inferred independently of the other components of the syntagmatic or horizontal axis.

- Some miRNAs are exchangeable in terms of predictive ability and point to latent biological functions.

- Conflicting results in the literature suggest that a protective or an oncogenic effect cannot be definitely assigned to any miRNA. Data-driven nets may help biologists in building new hypotheses and experimental designs in order to decipher the function of non-coding RNA, which may act as *chameleon* molecules according to the organization in which they are embedded.

**Supporting Information**

Below we state all the material that is auxiliary to the main content of the article and can be found in a separate file.

- S1 Table 1. List of all the 45 miRNAs selected by the SWAG.
- S1 Table 2. List of all the models with 4 miRNAs selected by the SWAG.
- S1 Table 3. List of all the models with 5 miRNAs selected by the SWAG. (Part 1)
- S1 Table 4. List of all the models with 5 miRNAs selected by the SWAG. (Part 2)
- S1 Table 5. List of all the models with 5 miRNAs selected by the SWAG. (Part 3)

**Acknowledgments**

We thank Haakensen *et al.* for having made available the AHUS data set on the free access ArrayExpress platform at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3759/?query=AHUS. The statistical analysis performed in this study is based on their data.

**Conflict of Interest Statement**

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Competing interests:** The authors declare that they have no competing interests.
References

1. Haakensen VD, Nygaard V, Greger L, Aure MR, Fromm B, Bukholm IR, Lüders T, Chin SE, Git A, Caldas C, Kristensen VN, Brazma A, Borresen-Dale AL, Hovig E, Helland Å. Subtype-specific micro-RNA expression signatures in breast cancer progression. Int J Cancer. 2016 Sep 1;139(5):1117-28. doi: 10.1002/ijc.30142.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7–34. doi:10.3322/caac.21551.

3. Chang JT, Wang F, Chapin W, Huang RS. Identification of MicroRNAs as Breast Cancer Prognosis Markers through the Cancer Genome Atlas. PLoS ONE 11(12): e0168284. https://doi.org/10.1371/journal.pone.0168284.

4. Adhami M, Haghdoost AA, Sadeghi B, Malekpour Afshar R. Candidate miRNAs in human breast cancer biomarkers: a systematic review. Breast Cancer. 2018 Mar;25(2):198-205. doi: 10.1007/s12282-017-0814-8.

5. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics 2018;19:232.https://doi.org/10.1186/s12859-018-2217-z

6. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices BMC Bioinformatics 2012;13:328. https://doi.org/10.1186/1471-2105-13-328.

7. Bavelloni A, Ramazzotti G, Poli A, et al. MiRNA-210: A Current Overview. Anticancer Res. 2017;37(12):6511-6521. doi:10.21873/anticancerres.12107

8. Salem O, Erdem N, Jung J, et al. The highly expressed 5’isomiR of hsa-miR-140-3p contributes to the tumor-suppressive effects of miR-140 by reducing breast cancer proliferation and migration. BMC Genomics 2016;17:566. Published 2016 Aug 8. doi:10.1186/s12864-016-2869-x

9. Guerrier S, Mili N, Molinari R, Orso S, Avella-Medina M, Ma Y. A Predictive Based Regression Algorithm for Gene Network Selection. Front Genet. 2016;7:97. Published 2016 Jun 15. doi:10.3389/fgene.2016.00097

10. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005; 21:3301–3307.

11. Kantarjian H, Yu PP. Artificial Intelligence, Big Data, and Cancer. JAMA Oncol. 2015;1(5):573-574. doi:10.1001/jamaoncol.2015.1203

12. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003 Aug 5;100(16):9440-5. doi: 10.1073/pnas.1530509100

13. Volinia S, Galasso M, Sana ME, et al. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci U S A. 2012;109(8):3024–3029. doi:10.1073/pnas.1200010109

14. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. Front Cell Dev Biol. 2014 Aug 19;2:38. doi: 10.3389/fcell.2014.00038.
15. Xu J, Yang P, Xue S, Sharma B, Sanchez-Martin M, Wang F, Beaty KA, Dehan E, Parikh B. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives Hum Genet. 2019 Feb;138(2):109-124. doi: 10.1007/s00439-019-01970-5.

16. Yule GU. Notes on the theory of association of attributes in statistics. Biometrika 1903 (2): 121–34.

17. Cox DR. Conditional and marginal association for binary random variables. Biometrika, Volume 90, Issue 4, December 2003, Pages 982–984, https://doi.org/10.1093/biomet/90.4.982

18. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56-68. doi: 10.1038/nrg2918.

19. Di Carlo S, Politano G, Savino A, Benso A. A systematic analysis of a mi-RNA inter-pathway regulatory motif. J Clin Bioinforma. 2013 Oct 24;3(1):20. doi: 10.1186/2043-9113-3-20.

20. Alon U. Biological networks: the tinkerer as an engineer. Science. 2003, 301 (5641): 1866-1867. 10.1126/science.1089072.

21. Nielsen J. Systems biology of metabolism. Annu. Rev. Biochem. 2017, 86:245-275. https://doi.org/10.1146/annurev-biochem-061516-044757.

22. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007.

23. Shao B, Wang X, Zhang L, Li D, Liu X, Song G, Cao H, Zhu J, Li H. Plasma microRNAs Predict Chemoresistance in Patients With Metastatic Breast Cancer. Technol Cancer Res Treat. 2019 Jan 1;18:1533033819828709. doi: 10.1177/1533033819828709.

24. Liu J, Wang P, Zhang P, Zhang X, Du H, Liu Q, Huang B, Qian C, Zhang S, Zhu W, Yang X, Xiao Y, Liu Z, Luo D. An integrative bioinformatics analysis identified miR-375 as a candidate key regulator of malignant breast cancer. J Appl Genet. 2019 Nov;60(3-4):335-346. doi: 10.1007/s13353-019-00507-w.

25. Ward A, Balwierz A, Zhang JD, Küblbeck M, Pawitan Y, Hielscher T, Wiemann S, Sahin Ö. Re-expression of microRNA-375 reverses both tamoxifen resistance and accompanying EMT-like properties in breast cancer. Oncogene. 2013 Feb 28;32(9):1173-82. doi: 10.1038/onc.2012.128.

26. Mansoori B, Mohammadi A, Ghasabi M, Shirjang S, Dehghan R, Montazeri V, Holmskov U, Kazemi T, Duijf P, Gjerstorff M, Baradaran B. miR-142-3p as tumor suppressor miRNA in the regulation of tumorigenicity, invasion and migration of human breast cancer by targeting Bach-1 expression. J Cell Physiol. 2019 Jun;234(6):9816-9825. doi: 10.1002/jcp.27670.

27. Cao XC, Yu Y, Hou LK, Sun XH, Ge J, Zhang B, Wang X. miR-142-3p inhibits cancer cell proliferation by targeting CDC25C. Cell Prolif. 2016 Feb;49(1):58-68. doi: 10.1111/cpr.12235.
28. Glaich O, Parikh S, Bell RE, Mekahel K, Donyo M, Leader Y, Shayevitch R, Sheinboim D, Yannai S, Hollander D, Melamed Z, Lev-Maor G, Ast G, Levy C. DNA methylation directs microRNA biogenesis in mammalian cells. Nat Commun. 2019 Dec 11;10(1):5657. doi: 10.1038/s41467-019-13527-1.

29. Stepanenko AA, Vassetzky YS, Kavsan VM. Antagonistic functional duality of cancer genes. Gene. 2013 Oct 25;529(2):199-207. doi: 10.1016/j.gene.2013.07.047.

30. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics. 2002;18 Suppl 2:S231-S240. doi:10.1093/bioinformatics/18.suppl-2.s231.

31. Wang L, Mo Q, Wang J. MIrExpress: A Database for Gene Coexpression Correlation in Immune Cells Based on Mutual Information and Pearson Correlation. J Immunol Res. 2015;2015:140819. doi:10.1155/2015/140819

32. Xiao S, Zhu H, Luo J, Wu Z, Xie M. miR-425-5p is associated with poor prognosis in patients with breast cancer and promotes cancer cell progression by targeting PTEN. Oncol Rep. 2019 Dec;42(6):2550-2560. doi: 10.3892/or.2019.7371.

33. Guo YH, Tang FY, Wang Y, et al. Low concentration of formononetin promotes proliferation of estrogen receptor-positive cells through an ERα-miR-375-PTEN-ERK1/2-bcl-2 pathway. OncoTarget. 2017 Oct 19; 8(59):100045-100055. doi: 10.18632/oncotarget.21923.

34. Jiang D, Zhou B, Xiong Y, Cai H. miR-135 regulated breast cancer proliferation and epithelial mesenchymal transition acts by the Wnt/β-catenin signaling pathway. Int J Mol Med. 2019 Apr;43(4):1623-1634. doi: 10.3892/ijmm.2019.4081.

35. Tang J, Ma W, Zeng Q, Tan J, Cao K, Luo L. Identification of miRNA-Based Signature as a Novel Potential Prognostic Biomarker in Patients with Breast Cancer. Dis Markers. 2019 Dec 30;2019:3815952. doi: 10.1155/2019/3815952.

36. Zehentmayr F, Hauser-Kronberger C, Zellinger B, et al. Hsa-miR-375 is a predictor of local control in early stage breast cancer. Clin Epigenetics. 2016;8:28. Published 2016 Mar 8. doi:10.1186/s13148-016-0198-1

37. Schwickert A, Weghake E, Brüggemann K, et al. microRNA miR-142-3p Inhibits Breast Cancer Cell Invasiveness by Synchronous Targeting of WASL, Integrin Alpha V, and Additional Cytoskeletal Elements. PLoS One. 2015;10(12):e0143993. Published 2015 Dec 10. doi:10.1371/journal.pone.0143993

38. Boehm L, Reich BJ, Bandyopadhyay D. Bridging conditional and marginal inference for spatially referenced binary data. Biometrics. 2013;69(2):545-554. doi.org/10.1111/biom.12027

39. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput 21, 137–146 (2011). doi.org/10.1007/s11222-009-9153-8

40. Bernau C, Riester M, Boulesteix AL, et al. Cross-study validation for the assessment of prediction algorithms. Bioinformatics. 2014; 30+12):i105-i122. doi:10.1093/bioinformatics/btu279
41. Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. Nat Rev Clin Oncol. 2015;12(7):381-394. doi:10.1038/nrclinonc.2015.73

42. Wang H, Lengerich BJ, Aragam B, Xing EP. Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics. 2019; 35(7):1181-1187. doi:10.1093/bioinformatics/bty750.

43. De Saussure F, Harris R (transl.). Course in General Linguistics. New York : Bloomsbury Academic, 2013

44. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med. 2014;20(8):460-469. doi:10.1016/j.molmed.201406.005.

45. Chang JTH, Wang F, Chapin W, Huang RS. Identification of MicroRNAs as Breast Cancer Prognosis Markers Through the Cancer Genome Atlas. PLoS One. 2016 Dec 13;11(12):e0168284. doi: 10.1371/journal.pone.0168284.

46. Salem O, Erdem N, Jung J et al. The highly expressed 5’isomiR of hsa-miR-140-3p contributes to the tumor-suppressive effects of miR-140 by reducing breast cancer proliferation and migration. BMC Genomics 17, 566 (2016). https://doi.org/10.1186/s12864-016-2869-x

47. Hughes TR, de Boer CG. Mapping yeast transcriptional networks. Genetics. 2013; 195(1):9-36. doi:10.1534/genetics.113.153262

48. Österlund T, Bordel S, Nielsen J. Controllability analysis of transcriptional regulatory networks reveals circular control patterns among transcription factors. Integr Biol (Camb). 2015;7(5):560-568. doi:10.1039/c4ib00247d

49. Sun T, Song Y, Yu H, Luo X. Identification of lncRNA TRPM2-AS/miR-140-3p/PYCR1 axis's proliferates and anti-apoptotic effect on breast cancer using co-expression network analysis. Cancer Biol Ther. 2019;20(6):760-773. doi:10.1080/15384047.2018.1564563

50. Troschel FM, Böhly N, Borrmann K, et al. miR-142-3p attenuates breast cancer stem cell characteristics and decreases radioresistance in vitro Tumour Biol. 2018;40(8):1010428318791887. doi: 10.1177/1010428318791887

51. Zhang X, Wu M, Chong QY, et al. Amplification of hsa-miRNA-191/425 locus promotes breast cancer proliferation and metastasis by targeting DICER1. Carcinogenesis. 2018;39(12):1506-1516. doi:10.1093/carcin/bgy102

52. Naoem LD, Muthaiyan M, Venkatesan A. Identification of dysregulated miRNAs in triple negative breasts cancer: A meta-analysis approach. J Cell Physiol. 2019;234(7):11768-11779. doi:10.1002/jcp.27839

53. Xu B, Zhang X, Wang S, Shi B. MiR-449a suppresses cell migration and invasion by targeting PLAGL2 in breast cancer. Pathol Res Pract. 2018;214(5):790-795. doi:10.1016/j.prp.2017.12.02

54. Shi W, Bruce J, Lee M, et al. MiR-449a promotes breast cancer progression by targeting CRIP2. Oncotarget. 2016;7(14):18906-18918. doi:10.18632/oncotarget.7753
55. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. ICML ’04: Proceedings of the twenty-first international conference on Machine learning. July 2004; doi10.1145/1015330.1015432

56. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? Journal of animal ecology. 2006; 75(5), 1182-1189. doi: 10.1111/j.1365-2656.2006.01141.x

57. Molinari R, Bakalli G, Guerrier S, Miglioli C, Orso S, Scaillet O. SWAG: A Wrapper Method for Sparse Learning arXiv. 2020; 2006.12837

58. Cheng X, Chen J, Huang Z. miR-372 promotes breast cancer cell proliferation by directly targeting LATS2. Exp Ther Med. 2018;15(3):2812-2817. doi:10.3892/etm.2018.5761

59. Zhao YX, Liu HC, Ying WY, et al. microRNA 372 inhibits proliferation and induces apoptosis in human breast cancer cells by directly targeting E2F1. Mol Med Rep. 2017;16(6):8069-8075. doi:10.3892/mmr.2017.7591

60. Schwickert A, Weghake E, Brüggemann K, et al. microRNA miR-142-3p Inhibits Breast Cancer Cell Invasiveness by Synchronous Targeting of WASL, Integrin Alpha V, and Additional Cytoskeletal Elements. PLoS One. 2015;10(12):e0143993. Published 2015 Dec 10. doi:10.1371/journal.pone.0143993

61. Cao XC, Yu Y, Hou LK, et al. miR-142-3p inhibits cancer cell proliferation by targeting CDC25C. Cell Prolif. 2016;49(1):58-68. doi:10.1111/cpr.12235

62. Wang C, Schwab LP, Fan M, Seagroves TN, Buolamwini JK. Chemoprevention activity of dipyridamole in the MMTV-PyMT transgenic mouse model of breast cancer. Cancer Prev Res (Phila). 2013;6(5):437-447. doi:10.1158/1940-6207.CAPR-12-0345

63. Feliciano A, Castellvi J, Artero-Castro A, et al. miR-125b acts as a tumor suppressor in breast tumorigenesis via its novel direct targets ENPEP, CK2-α, CCNJ, and MEGF9. PLoS One. 2013;8(10):e76247. Published 2013 Oct 3. doi:10.1371/journal.pone.0076247

64. Madhavan D, Peng C, Wallwiener M, et al. Circulating miRNAs with prognostic value in metastatic breast cancer and for early detection of metastasis. Carcinogenesis. 2016;37(5):461-470. doi:10.1093/carcin/bgw008

65. Zheng W, Luo Z, van der Laan MJ. Marginal Structural Models with Counterfactual Effect Modifiers. Int J Biostat. 2018;14(1)://ijb.2018.14.issue-1/ijb-2018-0039/ijb-2018-0039.xml. Published 2018 Jun 8. doi:10.1515/ijb-2018-0039

66. Hajage D, De Rycke Y, Chauvet G, Tubach F. Estimation of conditional and marginal odds ratios using the prognostic score. Stat Med. 2017;36(4):687-716. doi:10.1002/sim.7170

67. Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. Genet Epidemiol. 2017;41(5):427-436. doi:10.1002/gepi.22046
68. van Oudenhoven FM, Swinkels SHN, Ibrahim JG, Rizopoulos D. A marginal estimate for the overall treatment effect on a survival outcome within the joint modeling framework. Stat Med. 2020;10.1002/sim.8713. doi:10.1002/sim.8713

69. Cun J, Yang Q. Bioinformatics-based interaction analysis of miR-92a-3p and key genes in tamoxifen-resistant breast cancer cells. Biomed Pharmacother. 2018;107:117-128. doi:10.1016/j.biopha.2018.07.158

70. Nilsson S, Möller C, Jirström K, et al. Downregulation of miR-92a is associated with aggressive breast cancer features and increased tumour macrophage infiltration. PLoS One. 2012;7(4):e36051. doi:10.1371/journal.pone.0036051

71. Moi L, Braaten T, Al-Shibli K, Lund E, Busund LR. Differential expression of the miR-17-92 cluster and miR-17 family in breast cancer according to tumor type; results from the Norwegian Women and Cancer (NOWAC) study. J Transl Med. 2019;17(1):334. Published 2019 Oct 3. doi:10.1186/s12967-019-2086-x

72. Danza K, De Summa S, Pinto R, et al. TGFbeta and miRNA regulation in familial and sporadic breast cancer. Oncotarget. 2017;8(31):50715-50723. Published 2017 Jan 30. doi:10.18632/oncotarget.14899

73. Vimalraj S, Miranda PJ, Ramyakrishna B, Selvarumugan N. Regulation of breast cancer and bone metastasis by microRNAs. Dis Markers. 2013;35(5):369-387. doi:10.1155/2013/451248.

74. Wang B, Li J, Sun M, Sun L, Zhang X. miRNA expression in breast cancer varies with lymph node metastasis and other clinicopathologic features. IUBMB Life. 2014;66(5):371-377. doi:10.1002/iub.1273

75. Wu H, Li S. Non-coding RNA MT1JP exerts anti-cancer effects in breast cancer cells by regulating miR-92-3p. Gen Physiol Biophys. 2020;39(1):59-67.

76. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012;13:328. Published 2012 Dec 9. doi:10.1186/1471-2105-13-328

77. Chandler D Semiotics: The Basics London: Routledge, 2017. https://doi.org/10.4324/9781315311050

78. Schopman NC, Heynen S, Haasnoot J, Berkhout B. A miRNA-tRNA mix-up: tRNA origin of proposed miRNA. RNA Biol. 2010;7(5):573-576. doi:10.4161/rna.7.4.13141.

79. Wang H, Tan Z, Hu H, et al. microRNA-21 promotes breast cancer proliferation and metastasis by targeting LZTFL1. BMC Cancer. 2019;19(1):738. Published 2019 Jul 27. doi:10.1186/s12885-019-5951-3

80. Petrović N. miR-21 Might be Involved in Breast Cancer Promotion and Invasion Rather than in Initial Events of Breast Cancer Development. Mol Diagn Ther. 2016;20(2):97-110. doi:10.1007/s40291-016-0186-3.
81. Zhang W, Xu J, Wang K, Tang X, He J. miR-139-3p suppresses the invasion and migration properties of breast cancer cells by targeting RAB1A. Oncol Rep. 2019 Nov;42(5):1699-1708. doi: 10.3892/or.2019.7297. Epub 2019 Aug 29. PMID: 31485677; PMCID: PMC6775813.

82. Zheng Y, Wang M, Wang S, et al. LncRNA MEG3 rs3087918 was associated with a decreased breast cancer risk in a Chinese population: a case-control study. BMC Cancer. 2020;20(1):659. Published 2020 Jul 15. doi:10.1186/s12885-020-07145-0

83. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015 Aug 12:e05005. doi: 10.7554/eLife.05005. PMID: 26267216; PMCID: PMC4532895.

84. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020 Jan 8;48(D1):D127-D131. doi: 10.1093/nar/gkj757. PMID: 31504780; PMCID: PMC6943051.

85. Chiang HR, Schoenfeld LW, Ruby JG, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev. 2010;24(10):992-1009. doi:10.1101/gad.1884710

86. Cava C, Bertoli G, Ripamonti M, et al. Integration of mRNA expression profile, copy number alterations, and microRNA expression levels in breast cancer to improve grade definition. PLoS One. 2014;9(5):e97681. Published 2014 May 27. doi:10.1371/journal.pone.0097681

87. Maltseva DV, Galatenko VV, Samatov TR, et al. miRNome of inflammatory breast cancer. BMC Res Notes. 2014;7:871. Published 2014 Dec 4. doi:10.1186/1756-0500-7-871

88. Li J, Zeng T, Li W, et al. Long non-coding RNA SNHG1 activates HOXA1 expression via sponging miR-193a-5p in breast cancer progression. Aging (Albany NY). 2020;12(11):10223-10234. doi:10.18632/aging.103123

89. Yang SJ, Yang SY, Wang DD, et al. The miR-30 family: Versatile players in breast cancer. Tumour Biol. 2017;39(3):1010428317692204. doi:10.1177/1010428317692204.

90. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155-D162. doi:10.1093/nar/gky1141

91. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68-D73. doi:10.1093/nar/gkt1181

92. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(Database issue):D152-D157. doi:10.1093/nar/gkq1027

93. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008;36(Database issue):D154-D158. doi:10.1093/nar/gkm952

94. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue):D140-D144. doi:10.1093/nar/gkj112
95. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res. 2004;32(Database issue):D109-D111. doi:10.1093/nar/gkh023

96. Luo K, Li Y, Yin Y, et al. USP49 negatively regulates tumorigenesis and chemoresistance through FKBP51-AKT signaling. EMBO J. 2017;36(10):1434-1446. doi:10.15252/embj.201695669

97. Liu J, Duan Z, Guo W, et al. Targeting the BRD4/FOXO3a/CDK6 axis sensitizes AKT inhibition in luminal breast cancer. Nat Commun. 2018;9(1):5200. Published 2018 Dec 5. doi:10.1038/s41467-018-07258-y.

98. Lin S, Cao C, Meng Y, et al. Comprehensive analysis of the value of RAB family genes in prognosis of breast invasive carcinoma. Biosci Rep. 2020;40(5):BSR20201103. doi:10.1042/BSR20201103

99. Martínez-Ramírez AS, Garay E, García-Carrancá A, Vázquez-Cuevas FG. The P2RY2 Receptor Induces Carcinoma Cell Migration and EMT Through Cross-Talk With Epidermal Growth Factor Receptor. J Cell Biochem. 2016;117(4):1016-1026. doi:10.1002/jcb.25390

100. Gay-Bellile M, Romero P, Cayre A, et al. ERCC1 and telomere status in breast tumours treated with neoadjuvant chemotherapy and their association with patient prognosis. J Pathol Clin Res. 2016;2(4):234-246. Published 2016 Jul 13. doi:10.1002/cjp2.52

101. Pan D, Kobayashi A, Jiang P, et al. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. Science. 2018;359(6377):770-775. doi:10.1126/science.aao1710

102. Smart SK, Vasileiadi E, Wang X, DeRyckere D, Graham DK. The Emerging Role of TYRO3 as a Therapeutic Target in Cancer. Cancers (Basel). 2018;10(12):474. Published 2018 Nov 29. doi:10.3390/cancers10120474

103. Liu Y, Du F, Chen W, Yao M, Lv K, Fu P. EIF5A2 is a novel chemoresistance gene in breast cancer. Breast Cancer. 2015;22(6):602-607. doi:10.1007/s12282-014-0526-2
First Screening

Compute the cv-error for logistic regression with one miRNA as variable

Subset of highly predictive miRNA

General Screening

Compute the cv-error for logistic regression with \( p \) miRNAs as variables

Structure 1 + Structure 2 + Subset of highly predictive miRNA

additive structure

Repeat General Screening up to fixed maximum model dimension defined

Figure 1
| Metric      | lasso | SWAG     |
|------------|-------|----------|
| Accuracy   | 0.80  | [0.80,0.96] |
| Sensitivity| 0.55  | [0.55,0.91]  |
| Specificity| 1     | [0.93,1]   |
| NPP        | 0.74  | [0.74,0.93] |
| PPP        | 1     | [0.89,1]   |
| # miRNAs   | 12    | [4,5]     |

Figure 2
Figure 3
Figure 4

Empirical distribution of $\beta$'s

Percentage of model including mi-RNA:
- [100%, 85%)
- [85%, 60%)
- [60%, 45%)
- [45%, 30%)
- [30%, 15%)
- [15%, 0%]

Associative $\beta$ median
Single $\beta$
Range for Associative $\beta$