frontiers in Oncology

ORIGINAL RESEARCH
published: 22 May 2020
doi: 10.3389/fonc.2020.00845

Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes

Soledad Ochoa[1,2], *Guillermo de Anda-Jáuregui*[1,3]* *and Enrique Hernández-Lemus*[1,4]*

[1] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2] Graduate Program in Biomedical Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Cátedras Conacyt para Jóvenes Investigadores', National Council on Science and Technology, Mexico City, Mexico, [4] Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico

Breast cancer is a disease that exhibits heterogeneity that goes from the genomic to the clinical levels. This heterogeneity is thought to be captured (at least partially) by the so-called breast cancer molecular subtypes. These molecular subtypes were initially defined based on the unsupervised clustering of gene expression and its correlate with histological, morphological, phenotypic and clinical features already known. Later, a 50-gene signature, PAM50, was defined in order to identify the biological subtype of a given sample within the clinical setting. The PAM50 signature was obtained by the use of unsupervised statistical methods, and therefore no limitation was set on the biological relevance (or lack of) of the selected genes beyond its predictive capacity. An open question that remains is what are the regulatory elements that drive the various expression behaviors of this set of genes in the different molecular subtypes. This question becomes more relevant as the measurement of more biological layers of regulation becomes accessible. In this work, we analyzed the gene expression regulation of the 50 genes in the PAM50 signature, in terms of (a) gene co-expression, (b) transcription factors, (c) micro-RNAs, and (d) methylation. Using data from the Cancer Genome Atlas (TCGA) for the Luminal A and B, Basal, and HER2-enriched molecular subtypes as well as normal tumor adjacent tissue, we identified predictors for gene expression through the use of an elastic net model. We compare and contrast the sets of identified regulators for the gene signature in each molecular subtype, and systematically compare them to current literature. We also identified a unique set of predictors for the expression of genes in the PAM50 signature associated with each of the molecular subtypes. Most selected predictors are exclusive for a PAM50 gene and predictors are not shared across subtypes. There are only 13 coding transcripts and 2 miRNAs selected for the four subtypes. *MiR-21* and *miR-10b* connect almost all the PAM50 genes in all the subtypes and normal tissue, but do it in an exclusive manner, suggesting a cancer switch from *miR-10b* coordination in normal tissue to *miR-21*. The PAM50 gene sets of selected predictors that enrich for a function across subtypes, support that different regulatory molecular mechanisms are taking place. With this study we aim to a wider understanding of the regulatory mechanisms that differentiate the expression of the PAM50 signature, which in turn could perhaps help understand the molecular basis of the differences between the molecular subtypes.

Keywords: multi-omic approaches, breast cancer subtypes, PAM50, elastic net, data integration

OPEN ACCESS

Edited by:
Chiara Romualdi,
University of Padova, Italy

Reviewed by:
Tanja Kunej,
University of Ljubljana, Slovenia
Valentina Silvestri,
Sapienza University of Rome, Italy

***Correspondence:**
Guillermo de Anda-Jáuregui
gdeanda@inmegen.edu.mx
Enrique Hernández-Lemus
ehernandez@inmegen.gob.mx

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *01 December 2019*
Accepted: *29 April 2020*
Published: *22 May 2020*

Citation:
Ochoa S, de Anda-Jáuregui G and Hernández-Lemus E (2020) Multi-Omic Regulation of the PAM50 Gene Signature in Breast Cancer Molecular Subtypes. Front. Oncol. 10:845. doi: 10.3389/fonc.2020.00845

1. INTRODUCTION

Breast cancer is the most common cause of cancer death among females (1). Breast tumors have been classified in molecular subtypes with distinctive clinical characteristics and a recognizable gene expression signature (2). Such signature has been reduced to 50 genes that achieve the best separation of subtypes, attaining the PAM50 classifier (3). However, the physiological implications of the difference in gene expression, if any, are not well-understood.

Given that gene expression is regulated by several interconnected mechanisms (4–7), differences across subtypes are expected for these mechanisms. Evidence of this was found in the form of distinguishable patterns of DNA methylation, mutation and miRNA expression that shape groups partially equivalent to the molecular subtypes (8). These patterns imply a link between the different omics and PAM50 gene expression, but do not clarify which genomic, epigenetic or post transcriptional changes drive the expression signature of such molecular subtypes. To advance in the identification of such drivers of molecular subtypes expression, we propose the use of a sparse model of PAM50 gene expression.

Sparse models achieve the selection of the best predictors of an independent variable by fitting penalized linear models. The penalization of the regression coefficients aim is to shrink them toward zero in such a way that predictors contributing lowly to prediction i.e., poorly associated with the independent variable, end up with null coefficient values and get filtered out of the model (9). Ridge Regression, Least Absolute Shrinkage and Selection Operator, and Elastic Network methods apply different penalizations. The elastic network approach selects groups of pairwise correlated variables instead of choosing a single predictor from the group (10, 11), augmenting the space of predictors of interest but also incrementing false positive rates (12).

Sparse models have been proposed for multi-omic sample classification (13, 14) and biomarker identification (15–17); but their capacity to simplify multi-omics co-interpretation has only been tested in the evaluation of the extent of different omics effects over a phenotype (18, 19). Here, the predictor selection capability of the elastic network approach is exploited to identify the CpGs, coding transcripts, and miRNAs most associated with the expression of the PAM50 genes in order to outline molecular differences behind the gene expression patterns characterizing breast cancer subtypes within a true multi-omic framework. The hypothesis is that PAM50 gene expression patterns are accompanied by distinctive regulatory elements, reflecting the way gene expression is controlled in the different breast cancer subtypes.

2. METHODS

2.1. Data Acquisition

Concurrent experimental samples of DNA methylation, transcript and miRNA expression were downloaded from the GDC (https://portal.gdc.cancer.gov/repository) at May 2019. Only samples with Illumina Human Methylation 450, RNA-seq and miRNA-seq measures were kept; filtering out samples quantified with the Illumina Human Methylation 27 BeadChip, which covers a smaller portion of the genome than the one we wanted to target. Subtype classification was also downloaded from the GDC trough TCGABiolinks R package (20).

After preprocessing them according to Aryee er al. (21), Tarazona et al. (22), and Tam et al. (23), and biomaRt v95, values of methylation for 384,575 probes and expression for 16,475 coding transcripts and 433 miRNA precursors were obtained for 45 unique samples of Her2, 395 LumA, 128 LumB, and 125 Basal subtypes, plus 75 samples of non-tumor (normal adjacent) tissue.

2.2. Elastic Network Implementation

The three different data types were concatenated and normalized to have mean = 0 and standard deviation = 1. Eighty percent of the samples for each subtype were used for training, leaving the rest for testing as in Liu et al. (13). Using the R package glmnet (24), elastic network models were fitted per subtype for each gene in the PAM50 classifier with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/enetGLMNET.R. The mixing parameter was held fixed at 0.5 because such value has shown a good performance (10), but shrinkage parameter (λ) was optimized between values from 0.001 and 1,000 through repeated cross-validation.

Cross-validation was repeated 100 times with k = 3-folds for the subtypes with <100 training samples (Her2+ subtype and normal tissue) and k = 5 for the more represented subtypes (Luminal A, Luminal B, and Basal). Chosen λ parameters were used to predict testing data and root mean squared error (RMSE) was calculated per model. Fitting was repeated with the same specifications, for only 40 samples per subtype to verify the effect of data set size.

2.3. Omics Comparison

For each PAM50 gene model, RMSE was calculated for the testing data either with (1) the complete set of selected predictors, (2) only with selected CpGs, (3) just with selected coding transcripts, or (4) solely with selected miRNAs. Omic's specific RMSE were evaluated by zeroing all coefficients not associated to the omic of interest in the already fitted models with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/RMSEperOmics.R, in an approach similar to the one used by Setty et al. (25) to search for key regulators. Obtained values shape RMSE distributions per omic which were compared via Kolmorogov–Smirnov test. This was done both per subtype per omic and mixing all the subtypes in a distribution per omic. *P*-values obtained were corrected for multiple testing with the FDR method.

2.4. Test vs. Reported Links Between Predictors and PAM50 Genes

Enrichment for previously reported regulatory links between PAM50 genes and CpGs, TFs, and miRNAs were tested by simple Fisher's Exact Test. Tests repeated by subtypes had *p*-values adjusted by FDR. Regulatory targets were taken from Illumina's annotation in the case of CpGs and from databases accessible through R packages in the case of TFs

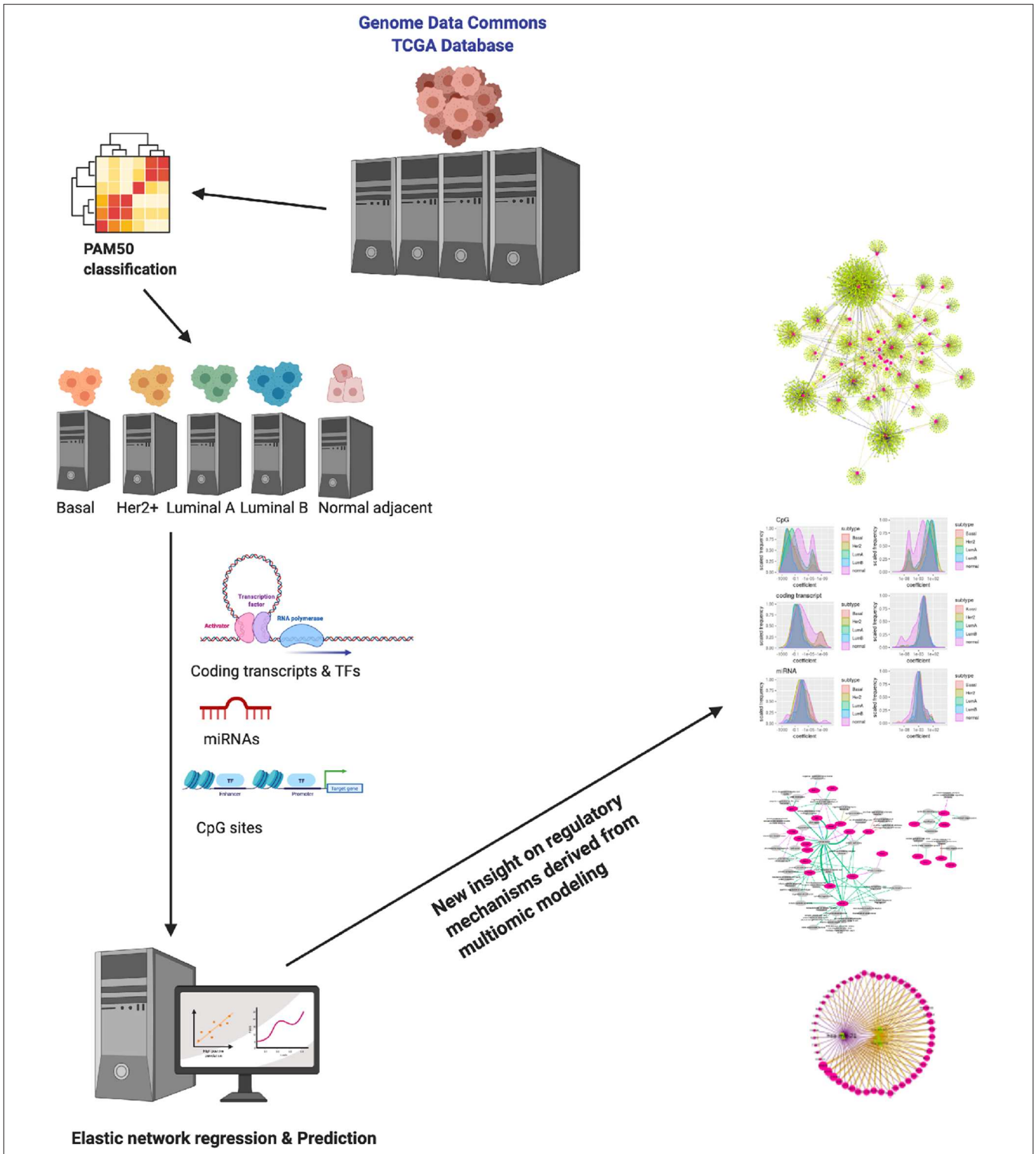

FIGURE 1 | Schematic depiction of this work. By analyzing multiomic data from the TCGA/Genome Data Commons collaboration for the different breast cancer molecular subtypes and healthy adjacent breast tissue via a generalized elastic network modeling, we have been able to derive some insight on the way the PAM50 genes are regulated (as predicted by the model). Results may shine some new light on the way PAM50 genes are able to capture intrinsic features of these phenotypes.

and miRNAs, with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/validateInteractions.R. tftargets https://github.com/slowkow/tftargets is the package used to retrieve TF targets. It queries both predicted and validated data from TRED(2007), ITFP(2008), ENCODE(2012), and TRRUST(2015) databases at the date specified in parentheses

next to each resource, plus the lists curated by Neph et al. (26) and Marbach et al. (27).

The package used to retrieve miRNA targets is multiMiR v2.2 (28), it queries DIANA-microT-CDS, ElMMo, MicroCosm, miRanda, miRDB, PicTar, PITA, TargetScan, miRecords, miRTarBase, and TarBase, also reporting both experimentally validated and predicted results. Universe size for enrichment tests were taken from these databases, constrained to regulators measured in the input datasets. The hypothesis is that models selected reported associations between a PAM50 gene and a regulator measured in the input dataset more than expected.

2.5. Analysis of the Selected Predictors

Selected predictors and associated coefficient values were loaded to Cytoscape to construct a network of PAM50 gene predictors per subtype. PAM50 genes are taken as targets while predictors are sources, this makes a directed network were out and indegree are estimated. Predictors with the largest outdegree were submitted to an analysis of differential expression and their coefficient value distributions were compared to the global miRNA distribution via Kolmorogov–Smirnov tests. The differential analysis of miRNA expression was done per subtype by limma's package *treat* function in order to control for both fold change and significance (29). A minimum fold change of 1.1 was used.

2.6. Gene Enrichment Analysis

Every set of predictors selected for a PAM50 gene was submitted to functional enrichment analysis with the R package *HTSanalyzeR* v2.13.1 (30) versus the GO-BP with the linked script https://github.com/CSB-IG/PAM50multiomics/blob/master/enrichment.R. Sets enriched across subtypes were further tested via Fisher's Exact Test with the alternative hypothesis that selection in one subtype is exclusive with regards to selection another subtype.

The code to perform all previous analyses (see **Figure 1**) can be found at the following GitHub repository: https://github.com/CSB-IG/PAM50multiomics

3. RESULTS

Elastic network models were fitted per gene, regressing PAM50 gene expression to DNA methylation, miRNA and coding transcript expression. Elastic networks model shrink the regression coefficients toward 0, filtering predictors by its strength of association with the variable of interest. This ability for feature selection was exploited versus unfiltered omic data to identify the CpGs, coding transcripts and miRNAs most related to the PAM50 genes in cancer subtypes and normal tissue.

We fitted five models for each PAM50 gene, one per subtype and one for the normal tissue, since differences are expected for each of the 5 phenotypes. Descriptors of models per subtype and omic are reported in **Table 1**.

The output of the model are lists of associations between PAM50 genes and the selected predictors. Each selected predictor has a coefficient of regression whose value reflects the extent of association with the PAM50 gene. Coefficients are never zero, since this value means predictors can be filtered out of the prediction; but can be both negative and positive indicating an opposite effect over the predicted value. Lists of associations shape networks like the one represented in **Figure 2**. Networks for the other subtypes and the normal tissue can be found at **Figures S1–S4**.

TABLE 1 | Size of input and output of the models per subtype: Basal, Her2+, Luminal A, Luminal B as well as normal (i.e. tumor-adjacent healthy tissue).

	Basal	Her2+	LumA	LumB	Normal
Samples	125	45	395	128	75
Selected CpGs	3,090	2,514	7,173	1,485	5,373
Known CpGs selected	9	0	21	12	0
Selected coding transcripts	1,525	591	3,115	888	2,340
Selected TFs	207	91	465	133	327
Selected TFs predicted by another software	15	2	49	11	19
Selected TFs experimentally observed	4	3	25	7	9
miRNAs	101	85	174	116	123
Selected miRNAs predicted by another software	7	3	7	2	4
Selected miRNAs experimentally observed	8	5	8	12	5

From observation of networks of selected predictors to PAM50 genes, it is evident that CpGs are the most selected predictors, followed by transcripts and with only a few miRNAs selected. It can also be seen that most predictors are exclusive of a PAM50 gene but all the PAM50 genes share predictors whose pattern of expression or methylation links one gene to another. This suggests the complete set of PAM50 expression is coordinated, independently of the gene being of luminal expression, basal, or any other signature.

3.1. Omics Contribute Differently to PAM50 Gene Expression Prediction in Normal Tissue and Cancer

In order to test the reliability of the fitted models, we checked the prediction error and the selection of previously reported associations. Regulation through DNA methylation, miRNA, or TF targeting is hence regarded as true positive and compared to model's results.

The proportion of selected predictors can not be explained solely by the size of the omics taken as input (χ^2, p-value $<$ 2.2e-16, **Figure 3**), specifically, coding transcripts and miRNAs are overrepresented in the models (Fisher's Exact Test, p-value $<$ 2.2e-16). Concordantly, there are more true TF (Fisher's Exact Test, p-value $\leq$ 1.942846e-05) and miRNA (Fisher's Exact Test, p-value $\leq$ 7.573200e-11) relations than expected but less CpGs (Fisher's Exact Test, p-value $\leq$ 4.311267e-03). The exception is LumB subtype which has as many true positive CpGs as expected.

Given the difference between input and selected proportion of omics, we hypothesized a discrepant prediction power of

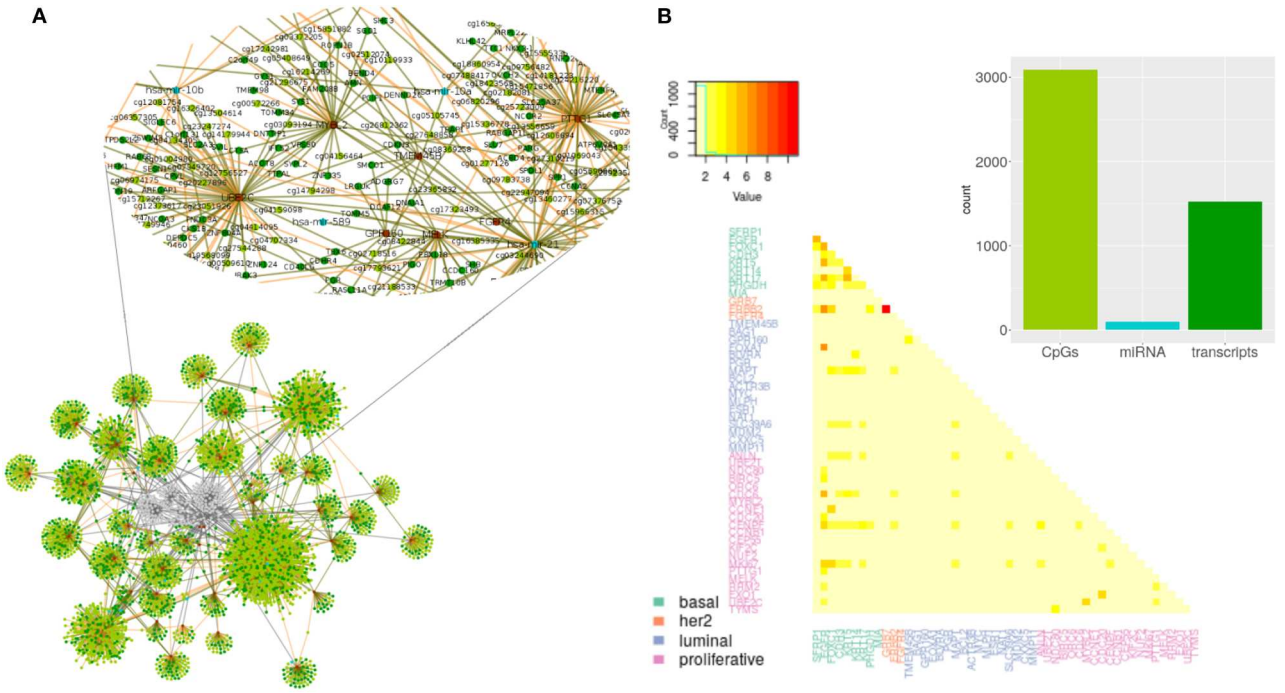

FIGURE 2 | Predictors selected per PAM50 gene for Basal subtype. **(A)** Topology of selected predictors and associated PAM50 genes. Brown circles are PAM50 genes. Transcripts are colored in dark green, miRNAs in blue, and CpGs in light green. Edges link each PAM50 gene with its selected predictors. The color of the line indicates the sign of the coefficient value associated with the predictor; negative values are in brown and positive ones in green. Zoom of the gray area shows the predictors selected for *MYBL2*. **(B)** Summary of the network. Barplot with the total representation of each omic plus heatmap of the count of predictors shared by PAM50 genes.

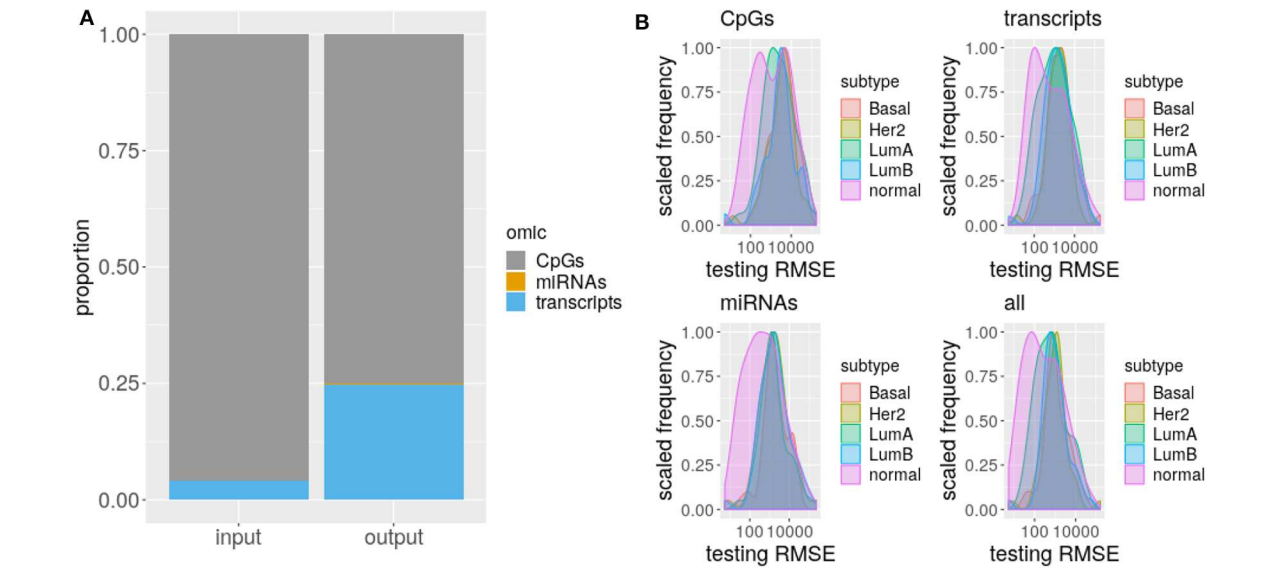

FIGURE 3 | Omics differ on selection and RMSE. **(A)** The proportion of each omic is shown for input and selected sets. The different omics are represented by a different color. **(B)** Distribution of testing RMSE per subtypes are displayed for single and multi-omics.

CpGs, coding transcripts, and miRNAs. To test this, we evaluated models carrying the complete set of selected predictors or just the predictors from each omic.

As RMSE is a standard measure to compare regression models that measures how far is the model prediction from the observed data in response variable units, then, the lower its value the better.

Normally, the error decreases the more independent predictors are included in the model, so we choose not to fit again with the selected predictor per omic, but to test the exact same model with the jointly fitted coefficient values, just zeroing predictor's coefficients from other than the omic of interest. This way, the RMSE distribution of a model containing only predictors of a given omic, represents how much of the total prediction is contributed by the predictors from that omic.

As suggested by the difference with the input proportions, DNA methylation is the less predictive omic for all the subtypes, thought this difference is not always significant (CpGs vs. coding transcripts ks. test p-value $\leq$ 0.03192 for LumB, Her2+, and Basal and CpGs vs. miRNAs ks. test p-value $\leq$ 0.02222 for Her2+ and Basal). This disagrees with the great prediction improvement reported by Huang et al. (16) for methylation data, a fact that could be driven by the much larger and heterogenous input data used here, that we believe captures better the heterogeneity of breast cancer subtypes. Meanwhile, coding transcript and miRNAs contribute the same, with no significant difference between their distributions for all the subtypes.

Remarkably, the error distribution obtained with the complete set of predictors significantly outperforms CpGs and some subtype miRNAs (ks.test p-value $\leq$ 0.02222 for LumA and Basal) but never outweighs coding transcripts. Single omics can not beat multi-omics error due to the design of the test, thus the outperforming of CpGs and miRNAs is unsurprising, what is startling is the complete statistical agreement between multi-omics prediction power and coding transcripts prediction power, which supports gene expression as the current best biomarker of molecular subtypes. We must note however that this may be related to (1) more info on RNA and (2) PAM50 was derived from expression signatures.

Finally, there is no significant difference across subtypes RMSE distributions for both single-omics and multi-omics, but CpGs (ks.test p-value $\leq$ 0.01601952), miRNAs (ks.test p-value $\leq$ 0.002834981), and multi-omics (ks.test p-value $\leq$ 0.03919459) distributions of normal tissue differ from the distribution of each subtype, suggesting these omics represent a distinct amount of PAM50 gene expression in normal tissue than in cancer, that is, the association of DNA methylation and miRNA expression with PAM50 gene expression is altered in cancer.

3.2. The Association Strength Distributions of Predictors Are Different for Each Subtype

The difference between omics extends to coefficient values, shown in **Figure 4**. Since coefficients represent the strength of association between predictors and PAM50 expression (16), coefficient values suggest that each omic has a specific association with PAM50 gene expression. Coefficient value distributions are significantly different between subtypes (ks.test p-value $\leq$ 2.82E-02) and omics (ks.test p-value $\leq$ 0.01535) with few exceptions for coding transcripts and miRNAs. Basal, Her2+, and LumB coding transcripts coefficients are not significantly different. Neither are miRNA coefficients of pairs LumA and normal tissue, LumB and Basal subtype, and Basal and Her2.

According to these distributions, DNA methylation has a strong but noisy association with PAM50 gene expression while miRNA (Fisher test p-values $\leq$ 0.001403597) and coding transcript (Fisher test p-values $\leq$ 1.086031e-29) association tends to be positive (**Figure S3**) and more stable. The elevated association between DNA methylation and PAM50 genes expression explains why so many CpGs get selected in spite of its low prediction power. A stronger association between DNA methylation and gene expression than between gene and miRNA expression had previously been found for ovarian cancer by Sohn et al. (18) using a different penalization modeling.

3.3. *miR-21* and *miR-10b* Are the Only Relevant Predictors Selected Across Subtypes

Next, we wanted to see how variable is actually the association between one predictor and the predicted PAM50 gene, that is, the specific coefficient values, not their distributions. For this, we wanted to focus on the predictors selected for a PAM50 gene across subtypes, shown in **Figure 5**. However, as noted before, models selected a great quantity of predictors exclusive for each gene, 93.45% of the selected CpGs, 74.24% of the coding transcript, and 81.37% the miRNAs are not shared between any two genes. In consequence, there are no CpGs associated with any gene for all the subtypes but there are 14 relations with coding transcripts and 51 with miRNAs satisfying this.

The 13 coding transcripts selected across subtypes as predictors of a specific PAM50 gene are trivial, since they just portray physical linkage. *ELP2* and *SLC39A6* are coded in opposite strands of the same locus while the rest of pairs are contiguous. Most of the associations, 84.77%, connect a PAM50 gene with a coding transcript in another chromosome, but these are not repeatedly selected across subtypes. It is worth mentioning that although all coefficients values are positive, even close predictors, like *YEATS4* and *SLC35E3* carry distinct coefficients.

Regarding miRNAs, there are only two miRNAs repeatedly selected among subtypes, *miR-10b* and *miR-21*. These are known breast cancer markers targeting some PAM50 genes (31). *Mir-21* has been experimentally linked with *BCL2*, *MYC*, *EGFR*, and *ERBB2* expression (32–35) and predicted to target *ESR1* and *FOXA1* (36, 37). On the other hand, *miR-10b* has been linked to *CDC6*, *EGFR*, and *SFRP1* (38, 39). There is no particular pattern among validated associations or coefficients, other than *miR-21* carrying mostly positive coefficient values and *miR-10b* selection extending up to normal tissue (for the full set of validated interactions please see **Supplementary Table S1**).

3.4. Micro-RNA *miR-21* and *miR-10b* Are Universal PAM50 Predictors in Cancer and Health

Next we wanted to check the role of *miR-21* and *miR-10b* per subtype. With this in mind, we revisited the models derived networks, that link PAM50 genes and predictors per subtype.

The networks show that genes overexpressed in each subtype get larger models. About 30% of the luminal genes have models

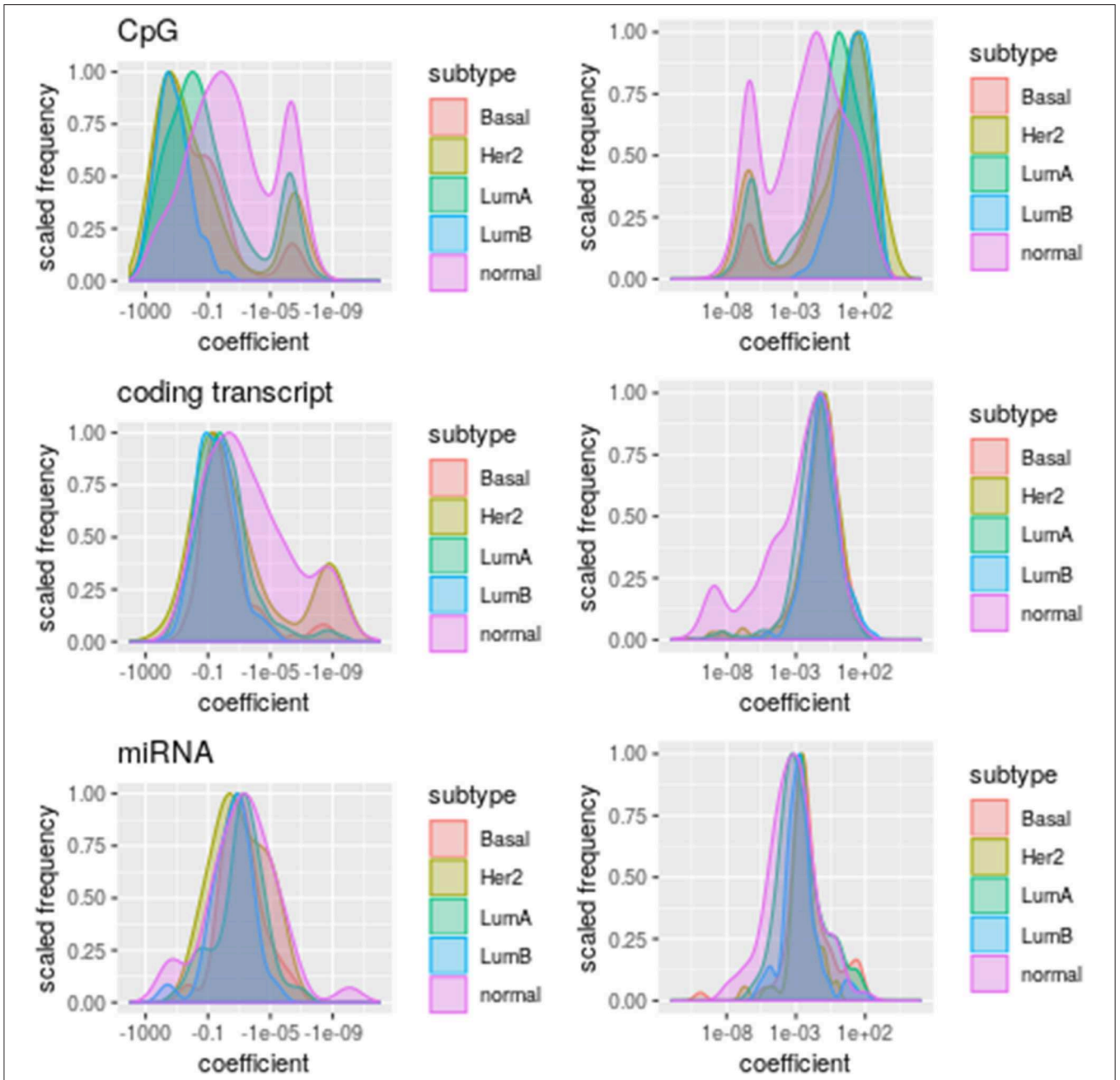

FIGURE 4 | Distribution of coefficient values per omic. The subtypes are represented by different colors. Plots at the **(Left)** column represent negative coefficients whereas those at the **(Right)** column stand for positive coefficient values.

larger than average for LumA subtype, while almost 90% of basal genes have the equivalent for Basal subtype. Her2+ subtype and normal tissue have no clear pattern, but for LumB subtype, half the luminal genes and 28% of the proliferative ones have increased size models.

Predictors that bridge between PAM50 genes can proceed from any omic, but CpGs are significantly underrepresented (Fisher test p-values $\leq$ 1.81E-88). CpGs are at most, selected for two subtypes as predictors of a specific PAM50 gene. There are just 24 CpGs in this situation, of which 15 are shared between Her2+ and another subtype or the normal tissue, including nine CpGs associated with *ERBB2* but placed in other loci than chromosome 17.

Meanwhile, coding transcripts and miRNAs fulfill this role more often (Fisher test p-values $\leq$ 5.84E-03) than solely input proportions would explain. This is no surprise since both pertain to the same level of molecular features, that of transcripts, as the PAM50 gene expression signature; as such, coding transcript and miRNA may be subject to the same biomolecular pressures. The stunning observation is that one miRNA can link almost all of

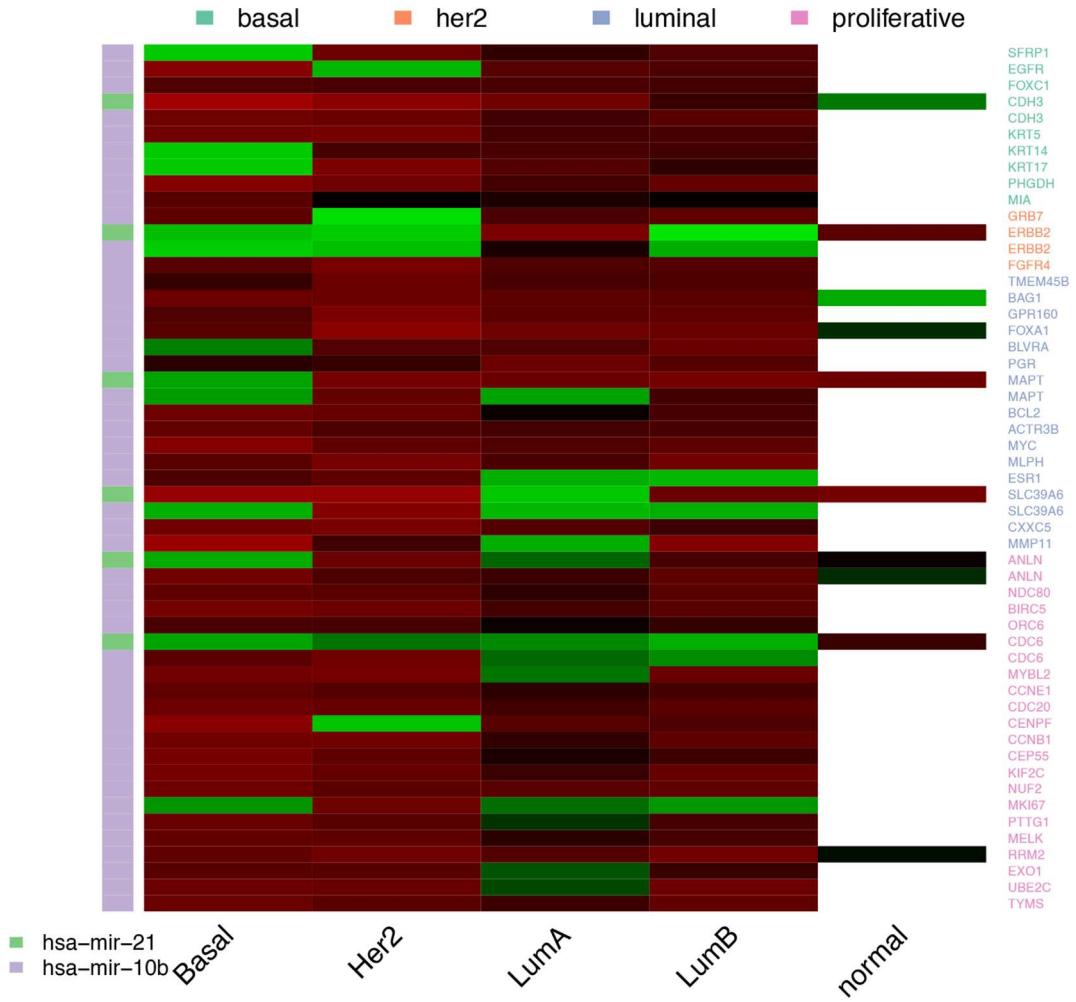

FIGURE 5 | miRNAs selected across the four cancer subtypes. miRNAs are represented by the color bar at the left. The highest positive coefficient value appears in bright red, the lowest negative coefficient is in bright green. Near zero coefficients look black. The white cells in the column of normal tissue means the predictors was not selected.

the PAM50 genes for all the cases (**Figure 6**). The outstanding miRNAs are again *miR-21* and *miR-10b*.

For normal tissue *miR-10b* was selected as predictor of all PAM50 genes while *miR-21* is linked to only four genes. On the contrary, *miR-21* is connected to most genes in the all the breast cancer subtypes, while *miR-10b* is poorly linked. For LumA subtype, shown in **Figure 6B**, both *miR-10b* and *miR-10a* are highly connected, but still can not reach genes like *FOXC1*, which is connected instead with *miR-21*.

Both *miR-10a* and *miR-10b* are members of the miR-10 family encoded within the Hox genes genomic clusters; *miR-10a* resides upstream from *HOXB4* and *miR-10b* upstream from *HOXD4* (40). Due to their relatedness they will be referred as *miR-10a/b*.

The hub-like behavior of these miRNAs agrees with previous observations of our group of highly connected miRNAs per subtype (41), which are important for network cohesion (42). Although the coefficients networks maintain a large connected component when removing *miR-10a/b* and *miR-21*, tens to hundreds of predictors are needed to link all the PAM50 genes; when only one of these miRNAs is required to achieve the same.

Given that each miRNA has the potential to target hundreds of genes (43), *miR-10a/b* and *miR-21* are not so exceptional in this regard. However, as explained earlier, only a fraction of PAM50 genes have a regulatory relation with these miRNAs, suggesting most of the detected associations are indirect. Indirectness is consistent with the low values of the coefficients, which range from −0.2938690 to 0.4333184, when miRNAs coefficient values range within two orders of magnitude higher. Coefficient value distributions of *miR-10a/b* and *miR-21* are also significantly different than the rest of miRNA coefficients (ks.test p-value $\leq$ 9.068e-05).

3.5. PAM50 Genes Enrich for Different Functions per Subtype

The selection of predictors we have presented is based on a statistical association with the pattern of expression of a

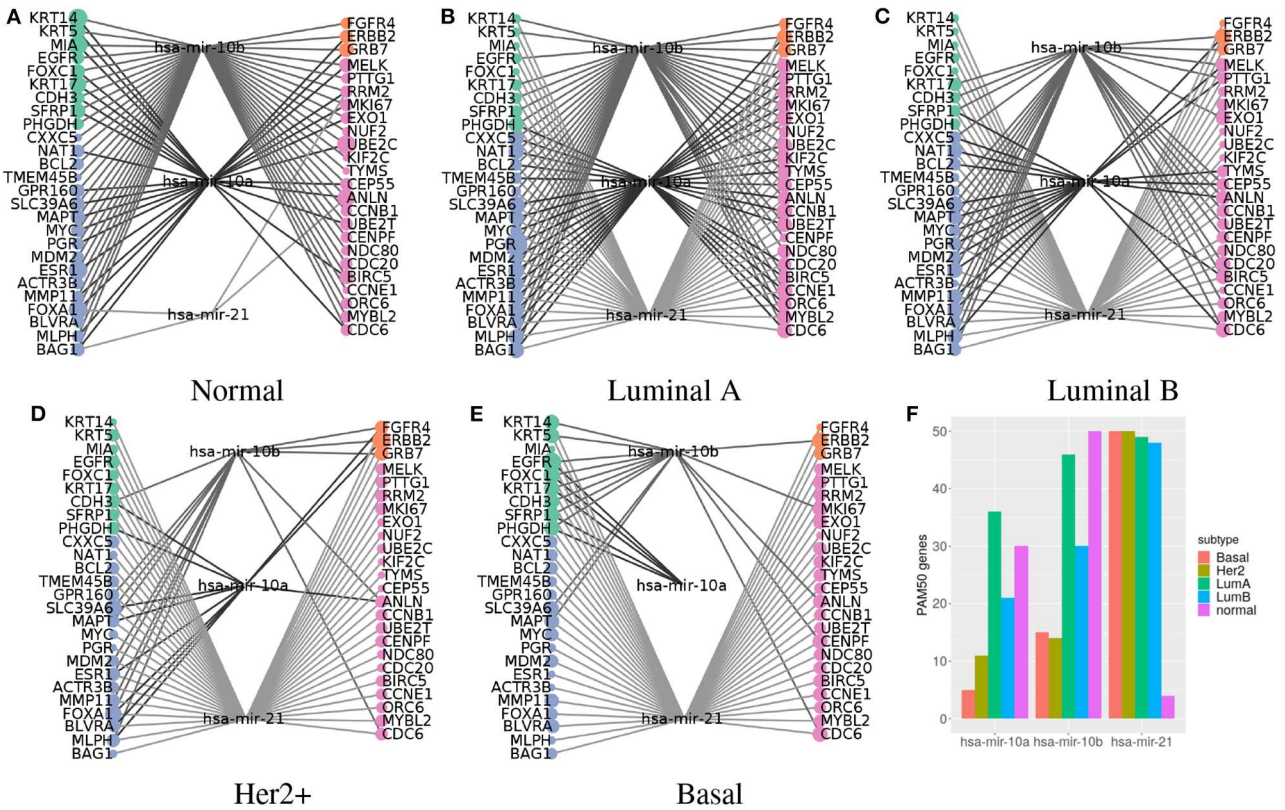

FIGURE 6 | Predictors connecting most PAM50 genes transition from *miR-10a/b* in normal tissue to *miR-21* in cancer subtypes. **(A–E)** PAM50 genes are at the sides, colored according to their pattern of expression: green for basal expression, orange for Her2 enriched, blue for luminal expression, and pink for proliferation favorable. Node size represents the number of predictors selected. Predictors are in the middle. Lines connect PAM50 genes to the associated predictors. Lines are in a gray gradient to distinguish different predictors. **(F)** The bar chart represents the number of PAM50 genes connected to the miRNAs per subtype.

PAM50 gene. The covariation sustaining such an association may respond to how a specific group of predictors is able to attain some biological function. To test this, functional enrichment was done with the set of selected predictors per gene per subtype, versus Gene Ontology Biological Processes categories (GO-BP) (**Figure 7**).

Only two PAM50 genes are enriched for some process in the Basal subtype, *FOXC1* (basal cluster) and *ANLN*(proliferative cluster). Neither the *ANLN* enrichment for telomere protection nor the *FOXC1* linkage to transforming growth factor response are within these genes immediate annotated processes. Though *FOXC1* is actually related with *TGFβ* since both are able to regulate EMT (44).

In the case of Her2+, just *ORC6* (proliferative cluster) is enriched for the totally unexpected process of synapse assembly, but, despite the significant *p*-value, we must notice that this is based on only two genes.

LumA is the most enriched subtype. This is not surprising since it has the largest number of selected coding transcripts, which is the starting material for enrichment. The 20 enriched genes are mostly linked to distinct cellular division aspects. The exception are the three keratins, genes with basal expression, which are connected through their normal processes, suggesting selected predictors respond to the normal gene's function. *MYC* and *UBE2T* are linked to rather wide categories (45) while *MLPH* associates with other than its normal processes. The remaining 14 genes are connected through categories consistent with their proliferative expression, which again alludes to a selection that followed the normal function of the genes. This is again consistent with the available evidence.

For LumB subtype, *MELK* and *CCNB1* enrich for cell division as would be normally expected; while *MYBL2* is unintuitively linked to negative regulation of epithelial cell proliferation, which however, has been reported (46). Finally, the normal tissue shows different cell division aspects coherent with the proliferative expression of its enriched genes.

Altogether, few genes have predictors with significant enrichment extended across subtypes. Eight genes enriched in two subtypes, including *CCNB1*, *MKI67*, and *UBE2C*, that connect with the same processes, the expected ones, for the two subtypes. *MELK* also connects with its normal process for two subtypes but in LumA and LumB subtypes plus normal tissue. *ANLN*, *CEP55*, *KRT17*, *MYBL2*, and *ORC6*, enrich for different processes across subtypes, that is, a fifth of the genes with any kind of enrichment, but five of the nine genes enriched for more than one subtype.

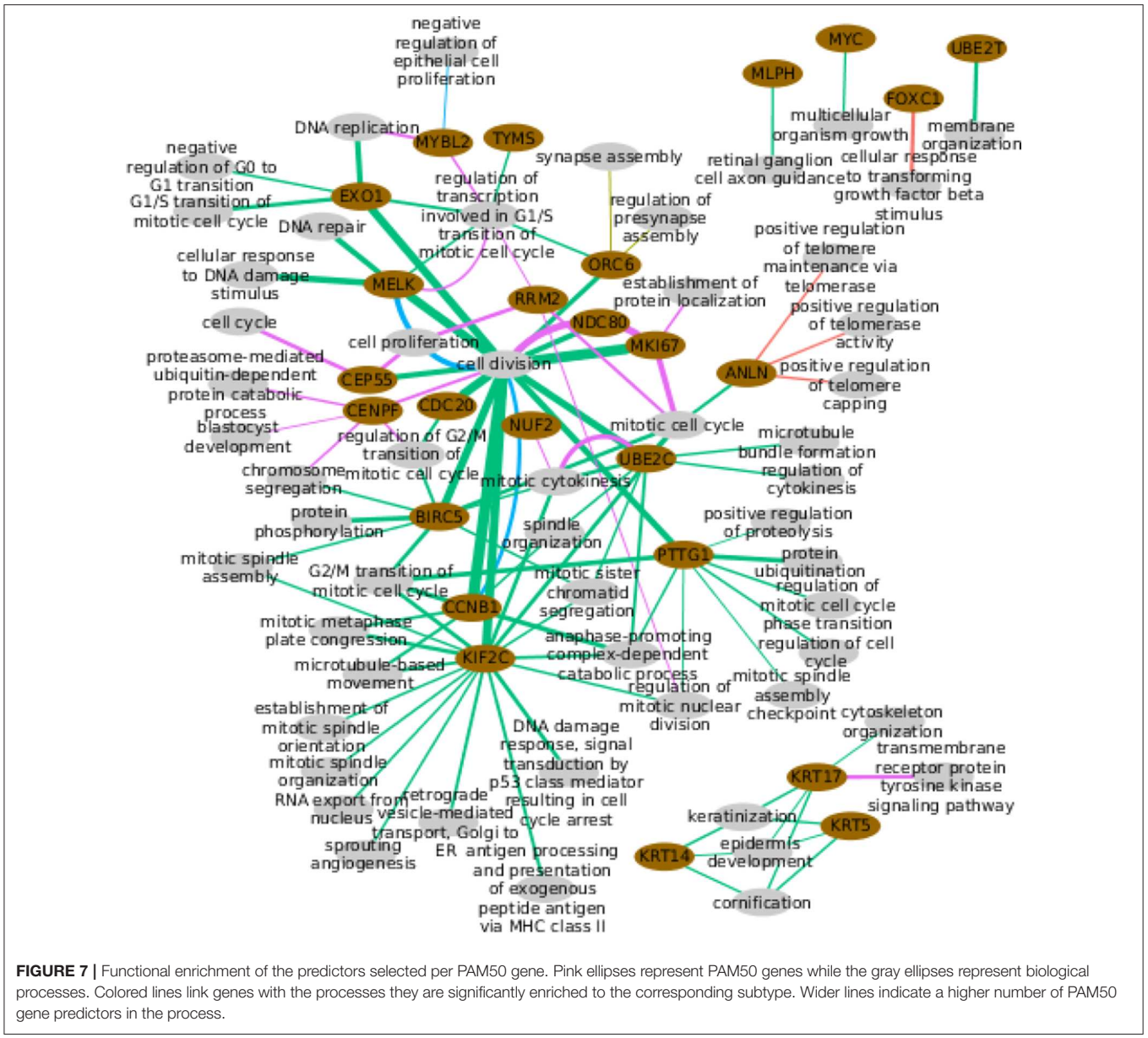

FIGURE 7 | Functional enrichment of the predictors selected per PAM50 gene. Pink ellipses represent PAM50 genes while the gray ellipses represent biological processes. Colored lines link genes with the processes they are significantly enriched to the corresponding subtype. Wider lines indicate a higher number of PAM50 gene predictors in the process.

To further test the functional enrichment per subtype, we compared the sets of predictors selected per subtype for each one of the 9 genes that enrich for several subtypes. Genes enriched for cell division across subtypes, *CCNB1*, *MKI67*, and *MELK* connect to the process via distinct sets of selected predictors. From the beginning, these genes bear different predictors (Fisher's Exact Test H1: less, p-value $\leq$ 1.281e-09), with a small intersection whose removal does not change the significant enrichment for cell division. This reflects the robustness of the process, which is so important that distinct subsets of the 603 genes annotated in the category are enough to call it.

The other two genes enriched for the same process across subtypes, *UBE2C* for mitotic cytokinesis and, *MELK* for regulation of transcription involved in G1/S transition of mitotic cell cycle, lost the functional enrichment when the predictors selected in both LumA and normal tissue (the intersection) were removed. This implies LumA mitotic cytokinesis and regulation of transcription may be involved in G1/S transition of mitotic cell cycle relying on the normal tissue mechanism.

The quantity of shared predictors between the sets selected for *CEP55*, indicates that predictor selection in the LumA subtype is exclusive for normal tissue selection (Fisher's Exact Test H1: less, p-value = 1.141e-10). This means that the differential enrichment between LumA and normal tissue is sustained by different predictors, suggesting *CEP55* fulfills divergent roles in these phenotypes. This matches differences observed between cancer and normal tissue (47) but, to our knowledge, not reported for LumA subtype.

The same reasoning supports *KRT17* and *ORC6* divergent roles across subtypes. It is odd that *KRT17* is linked to kinase signaling for normal tissue and not for a breast cancer subtype, when this has been described for another cancer (48) but this may be associated to tumor incidence over adjacent tissue (49). For *ANLN* and *MYBL2*, selection exclusion between subtypes is not significant, meaning that differential enrichment of these genes could settle on the same predictors, suggesting functional diversity.

4. DISCUSSION

Sparse penalized models have already proven useful to discover molecular mechanisms, cluster samples, and predict outcomes such as survival (50). Penalization permits the fitting of models otherwise unattainable given the relatively small sample sizes and huge number of variables measured by the omics. Here, the elastic network approach was used for integrated interpretation of different omics measuring DNA methylation and expression of both coding transcripts and miRNAs.

However, a large training set is always preferable, and not all breast cancer subtypes have been extensively sampled, which is reflected in the models. For Luminal A, the most frequent and sampled subtype, the highest number of predictors were selected by the models; while Her2+, with only 45 samples, got the lowest number of selected predictors. To assure comparability across subtypes we trained the models again, but now using the same number of samples, 40 samples, for all the subtypes. Patterns found with this subset persist in the analysis of the whole set of data, supporting comparability (**Figures S5–S8**). Nevertheless, the absence of predictors found for LumA in the smaller subtype's models due to a lack of representation can not be ruled out. This could specifically affect the functional enrichment of PAM50 neighborhoods of predictors and so, the functional divergence between subtypes is not definitive but should be experimentally tested.

Multi-omic modeling of PAM50 gene expression is no better than the sole use of coding transcripts, supporting gene expression as the best biomarker of molecular subtypes. However, our point in using the sparse model was not to predict PAM50 but to identify the molecular differences associated with PAM50 signatures that may lead to functional differences.

At the global level, a reduced prediction power of DNA methylation and miRNAs containing models was observed for all subtypes vs. normal tissue, indicating that the influence of this omics on PAM50 gene expression is reduced for cancer. Although this may be born out of incomplete knowledge or incipient technology, an alteration of these omics has been effectively reported; specifically, a generalized hypomethylation has been observed for breast and other cancers (51).

Different predictors were expected per cancer subtype, but the exclusivity of predictors from all the omics was surprisingly high. Only 13 coding transcripts and 2 miRNAs were selected for the four subtypes. The lack of CpGs selected across subtypes is consistent with the high strength of association it has with PAM50 gene expression. If the pattern of expression is different between subtypes, the highly associated CpGs should be different.

The ubiquitous selection of *miR-10b* and *miR-21* across subtypes suggests a central role for these miRNAs in breast cancer, which is actually supported by the literature. Proliferation, cell migration, and *in vivo* tumor growth of MCF7 and MDA-MB-231 cell lines implanted in nude mice is inhibited through antagomiR-21 (52) demonstrating the relevance of this miRNA, at least for luminal A and triple negative subtypes. In turn, both sub and overexpression of *miR-10* are oncogenic. *MiR-10b* overexpression enhances cell migration and invasion by targeting *HOXD10*; while subexpression of *miR-10b-3p*, coded in the same *miR-10b* locus, participates in breast cancer onset by upregulating the cell cycle regulators *BUB1*, *PLK1*, and *CCNA2* (53).

Coherent with the ubiquitous selection of *miR-21* breast cancer subtypes and its replacement by *miR-10a/b* in normal tissue. *MiR-21* is significantly overexpressed for all cancer subtypes while *miR-10b* is underexpressed, as previous reports say (31). *Mir-10a* is significantly underexpressed in Basal and Her2+ subtypes and slightly overexpressed in luminal subtypes, but this is not significant in LumB case. The proposal is that when *miR-10b* coordinates PAM50 genes, normal tissue expression is predicted; when *miR-10b* is sub expressed and *miR-21* is overexpressed, this second miRNA gains *miR-10b* place, coordinating cancer expression of the PAM50 genes. Since *miR-10b* has a known role in metastasis (31), it would be interesting to observe the dynamics of the networks throughout the evolution of the disease.

Additionally, the small coefficients associated with these miRNAs are consistent with indirect associations. Considering all these pieces, the transition from hub *miR-10a/b* in normal tissue to *miR-21* in breast cancer through the luminal subtypes, evokes a switch between two master regulators. Master regulators are genes needed for the specification of a lineage by its capacity to regulate downstream genes either directly or not, whose missexpression can re-specify the fate of cells (54).

Nonetheless, sparse models can not select regulators naively, they need to feed on known regulators (16, 25, 55). Then, the regulatory capacity of selected predictor can not be stated, leaving *miR-10a/b* and *miR-21* just as universal predictors of PAM50 genes.

Another limitation of the study is the absence of an estimator of significance or accuracy intrinsic to the methodology (56). Regression models quality is described in terms of RMSE, without an indication of how well the selected predictors describe PAM50 expression. A ROC curve is not feasible, since models would have to be turned into the classification setting, and even this is unreachable, because true negative regulators can not be ascertained, as non regulators could simply be regulators yet to discover.

Finally, it is important to mention that applying the same shrinkage to inherently different molecular levels, like CpG methylation and transcript expression, could shrink to zero all the coefficients of subtler effect predictors (13). Thus, the next implementation of sparse multiomic models on PAM50 expression should adopt multiple penalizations, which could

even ameliorate the bias on subtype representation (57). Distinct values for the mixing parameter should also be probed, as well as data decomposition into latent variables (58).

Future Directions

Apart from exploration of alternative frameworks, the immediate follow up should be the experimental assessment of the observations described here. Specifically, silencing and expression of *miR-10a/b* and *miR-21* need to be tested for each breast cancer subtype. Disection of interaction between the miRNAs and the PAM50 genes is required too.

Then, more omics could be included in the models. Copy number variation is the first candidate to be incorporated since it is already available in the databases and has a proven effect on Her2+ subtype, in particular regarding the effect of the *Her2* amplicon since it has been associated to regulation of growth and survival processes. But single nucleotide variation and chromatin accessibility are also available for some samples.

Other phenotypes with discriminant patterns of expression could benefit from sparse modeling. There could be significant predictors linked to the glioblastoma subtypes as was observed for breast cancer. Predictors represent potential regulators of the mechanisms behind subtype heterogeneity and, as such, are interesting markers of cancer. In this sense, predictor selection across stages, not subtypes, could illuminate the driving forces behind disease development. Alternative methods like A–JIVE (59) and sPLS (60) would have also exciting outcomes in this settings.

A relevant mid to long term future direction will be the implementation of experimental assays to test for multi-omic synergistic or cooperative phenomena, aiming at providing some mechanistic clues of the biological functions behind. There is however a strong challenge on this given the combinatorial mixture of effects that may be complex to disentangle. Some promissory (yet preliminary) advances are starting to arise.

5. CONCLUSION

Holistic studies of cancer are needed to dissect its complexity. Initiatives like The Cancer Genome Atlas have delivered the distinct molecular perspectives that need to be interpreted as a whole. The elastic net models subject of this work, approach such an integration in a rather simplistic linear form. Yet, the methodology is powerful enough to prove the intuition that PAM50 gene expression patterns are accompanied by distinctive potentially regulatory elements. Predictors are selected in an almost exclusive manner, heavily dictated by the omic of origin, with CpGs strongly associated to PAM50 expression not selected across subtypes. The way *miR-10a/b* and *miR-21*, the only relevant predictors selected for all subtypes, are connected and differentially expressed, suggest an specific regulatory difference between breast cancer and normal tissue that merits further research.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Genome Data Commons site https://bit.ly/2Itoi2e. The code to perform all previous analyses can be found at the following GitHub repository: https://github.com/CSB-IG/PAM50multiomics.

AUTHOR CONTRIBUTIONS

SO organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. GA-J contributed to design of the study, generated programming code, and contributed to the writing of the manuscript. EH-L conceived the study, contributed to design of the study, provided funding, discussed findings, and reviewed the writing of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México. EH-L is recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

ACKNOWLEDGMENTS

This paper constitutes a partial fulfilment of the Graduate Program in Biomedical Sciences of the National Autonomous University of México (UNAM) requirements of SO (María de la Soledad Ochoa-Méndez). She acknowledges the scholarship and support provided by the National Council of Science and Technology (CONACyT) and UNAM. **Figure 1** was generated using Biorender (https://biorender.com/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.00845/full#supplementary-material

Figures S1–S4 depict the topology of the networks for the non-basal subtypes that were not shown. **Table S1** contains a list of all validated interactions.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin*. (2018) 68:394–424. doi: 10.3322/caac.21492
2. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of

breast cancer. *Breast*. (2015) 24:S26–35. doi: 10.1016/j.breast.2015.07.008
3. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. *Nature*. (2000) 406:747. doi: 10.1038/35021093
4. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. (2012) 22:1658–67. doi: 10.1101/gr.136838.111
5. Vimalraj S, Miranda P, Ramyakrishna B, Selvamurugan N. Regulation of breast cancer and bone metastasis by microRNAs. *Dis Mark*. (2013) 35:369–87. doi: 10.1155/2013/451248
6. Cao J, Luo Z, Cheng Q, Xu Q, Zhang Y, Wang F, et al. Three-dimensional regulation of transcription. *Protein Cell*. (2015) 6:241–53. doi: 10.1007/s13238-015-0135-7
7. Liu X, Chen X, Yu X, Tao Y, Bode AM, Dong Z, et al. Regulation of microRNAs by epigenetics and their interplay involved in cancer. *J Exp Clin Cancer Res*. (2013) 32:96. doi: 10.1186/1756-9966-32-96
8. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. (2012) 490:61–70. doi: 10.1038/nature11412
9. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol. 112. New York, NY: Springer (2013). doi: 10.1007/978-1-4614-7138-7
10. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x
11. Neto EC, Bare JC, Margolin AA. Simulation studies as designed experiments: the comparison of penalized regression models in the "large p, small n" setting. *PLoS ONE*. (2014) 9:e107957. doi: 10.1371/journal.pone.0107957
12. Kirpich A, Ainsworth EA, Wedow JM, Newman JR, Michailidis G, McIntyre LM. Variable selection in omics data: a practical evaluation of small sample sizes. *PLoS ONE*. (2018) 13:e0197910. doi: 10.1371/journal.pone.0197910
13. Liu J, Liang G, Siegmund KD, Lewinger JP. Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*. (2018) 19:369. doi: 10.1186/s12859-018-2401-1
14. Tini G, Marchetti L, Priami C, Scott-Boyer MP. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinformatics*. (2019) 20:1269–79. doi: 10.1093/bib/bbx167
15. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A. -Omics biomarker identification pipeline for translational medicine. *J Transl Med*. (2019) 17:155. doi: 10.1186/s12967-019-1912-5
16. Huang S, Xu W, Hu P, Lakowski TM. Integrative analysis reveals subtype-specific regulatory determinants in triple negative breast cancer. *Cancers*. (2019) 11:507. doi: 10.3390/cancers11040507
17. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omic assays. *Bioinformatics*. (2019) 35:3055–62. doi: 10.1093/bioinformatics/bty1054
18. Sohn KA, Kim D, Lim J, Kim JH. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst Biol*. (2013) 7:S9. doi: 10.1186/1752-0509-7-S6-S9
19. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. (2013) 7:523. doi: 10.1214/12-AOAS597
20. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. (2016) 44:e71. doi: 10.1093/nar/gkv1507
21. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. (2014) 30:1363–9. doi: 10.1093/bioinformatics/btu049
22. Tarazona S, Furió-Tarí- P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res*. (2015) 43:e140. doi: 10.1093/nar/gkv711
23. Tam S, Tsao MS, McPherson JD. Optimization of miRNA-seq data preprocessing. *Brief Bioinformatics*. (2015) 16:950–63. doi: 10.1093/bib/bbv019
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. (2010) 33:1–22. doi: 10.18637/jss.v033.i01
25. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol*. (2012) 8:605. doi: 10.1038/msb.2012.37
26. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. *Cell*. (2012) 150:1274–86. doi: 10.1016/j.cell.2012.04.040
27. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods*. (2016) 13:366–70. doi: 10.1038/nmeth.3799
28. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Res*. (2014) 42:e133. doi: 10.1093/nar/gku631
29. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*. (2009) 25:765–71. doi: 10.1093/bioinformatics/btp053
30. Wang X, Terfve C, Rose JC, Markowetz F. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*. (2011) 27:879–80. doi: 10.1093/bioinformatics/btr028
31. O'Day E, Lal A. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Res*. (2010) 12:201. doi: 10.1186/bcr2484
32. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY. miR-21-mediated tumor growth. *Oncogene*. (2007) 26:2799–803. doi: 10.1038/sj.onc.1210083
33. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, et al. Estradiol-regulated microRNAs control estradiol response in breast cancer cells. *Nucleic Acids Res*. (2009) 37:4850–61. doi: 10.1093/nar/gkp500
34. Barker A, Giles KM, Epis MR, Zhang PM, Kalinowski F, Leedman PJ. Regulation of ErbB receptor signalling in cancer cells by microRNA. *Curr Opin Pharmacol*. (2010) 10:655–61. doi: 10.1016/j.coph.2010.08.011
35. Huang TH, Wu F, Loeb GB, Hsu R, Heidersbach A, Brincat A, et al. Up-regulation of miR-21 by HER2/neu signaling promotes cell invasion. *J Biol Chem*. (2009) 284:18515–24. doi: 10.1074/jbc.M109.006676
36. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*. (2007) 8:69. doi: 10.1186/1471-2105-8-69
37. Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*. (2009) 37:W273–6. doi: 10.1093/nar/gkp292
38. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*. (2011) 8:559–64. doi: 10.1038/nmeth.1608
39. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. *Genome Biol*. (2003) 5:R1. doi: 10.1186/gb-2003-5-1-r1
40. Lund AH. miR-10 in development and cancer. *Cell Death Differ*. (2010) 17:209–14. doi: 10.1038/cdd.2009.58
41. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, highly connected microRNAs control functionality in breast cancer networks. *Int J Genomics*. (2018) 2018:9585383. doi: 10.1155/2018/9585383
42. Drago-García D, Espinal-Enríquez J, Hernández-Lemus E. Network analysis of EMT and MET micro-RNA regulation in breast cancer. *Sci Rep*. (2017) 7:13534. doi: 10.1038/s41598-017-13903-1
43. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. (2013) 153:654–65. doi: 10.1016/j.cell.2013.03.043
44. Yu M, Bardia A, Wittner BS, Stott SL, Smas ME, Ting DT, et al. Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science*. (2013) 339:580–4. doi: 10.1126/science.1228522
45. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. (2019) 47:D419–26. doi: 10.1093/nar/gky1038

46. Martin FT, Dwyer RM, Kelly J, Khan S, Murphy JM, Curran C, et al. Potential role of mesenchymal stem cells (MSCs) in the breast tumour microenvironment: stimulation of epithelial to mesenchymal transition (EMT). *Breast Cancer Res Treat*. (2010) 124:317–26. doi: 10.1007/s10549-010-0734-1
47. Jeffery J, Sinha D, Srihari S, Kalimutho M, Khanna KK. Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. *Oncogene*. (2016) 35:683–90. doi: 10.1038/onc.2015.128
48. Sankar S, Tanner JM, Bell R, Chaturvedi A, Randall RL, Beckerle MC, et al. A novel role for keratin 17 in coordinating oncogenic transformation and cellular adhesion in Ewing sarcoma. *Mol Cell Biol*. (2013) 33:4448–60. doi: 10.1128/MCB.00241-13
49. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*. (2017) 8:1077. doi: 10.1038/s41467-017-01027-z
50. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*. (2016) 17(Suppl. 2):15. doi: 10.1186/s12859-015-0857-9
51. Vidal Ocabo E, Sayols S, Moran S, Guillaumet-Adkins A, Schroeder MP, Royo R, et al. A DNA methylation map of human cancer at single base-pair resolution. *Oncogene*. (2017) 36:5648–57. (2017). doi: 10.1038/onc.2017.176
52. Wang SE, Lin RJ. MicroRNA and HER2-overexpressing cancer. *MicroRNA*. (2013) 2:137–47. doi: 10.2174/2211536611302999011
53. Biagioni F, Bossel Ben-Moshe N, Fontemaggi G, Yarden Y, Domany E, Blandino G. The locus of microRNA-10b: a critical target for breast cancer insurgence and dissemination. *Cell Cycle*. (2013) 12:2371–5. doi: 10.4161/cc.25380
54. Chan SSK, Kyba M. What is a master regulator? *J Stem Cell Res Ther*. (2013) 3:114. doi: 10.4172/2157-7633.1000e114
55. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. (2012) 28:2458–66. doi: 10.1093/bioinformatics/bts476
56. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, et al. Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS Genet*. (2015) 11:e1005689. doi: 10.1371/journal.pgen.1005689
57. Lee G, Bang L, Kim SY, Kim D, Sohn KA. Identifying subtype-specific associations between gene expression and DNA methylation profiles in breast cancer. *BMC Med Genomics*. (2017) 10:28. doi: 10.1186/s12920-017-0268-z
58. Lê Cao KA, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*. (2009) 10:34. doi: 10.1186/1471-2105-10-34
59. Feng Q, Jiang M, Hannig J, Marron J. Angle-based joint and individual variation explained. *J Multivar Anal*. (2018) 166:241–65. doi: 10.1016/j.jmva.2018.03.008
60. Rohart F, Gautier B, Singh A, Le Cao KA. mixOmics: An R package for-omics feature selection and multiple data integration. *PLoS Comput Biol*. (2017) 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.