-UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias de la Salud

Master Genes Identification in Breast Cancer: Towards a Better Diagnosis and Treatment
Proyecto de Investigación

Diego Adrián Barba Carrera
Joseth Paulina Adatty Molina

Medicina

Trabajo de Titulación como requisito
para la obtención del título de
Doctor en Medicina

Quito, 4 de octubre de 2019

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE CIENCIAS DE LA SALUD

HOJA DE CALIFICACIÓN
DE TRABAJO DE INTEGRACIÓN CURRICULAR

Master Genes Identification in Breast Cancer: Towards a Better Diagnosis and Treatment

Diego Adrián Barba Carrera
Joseth Paulina Adatty Molina

Calificación: **/ 10**

Nombre del profesor, Título académico **Yovani Marrero Ponce, BSc, MSc, PhD**

Firma del profesor:

Quito, 4 de octubre de 2019

Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante:

Nombres y apellidos: Diego Adrián Barba Carrera

Código: 00118535

Cédula de identidad: 1718133711

Firma del estudiante:

Nombres y apellidos: Joseth Paulina Adatty Molina

Código: 00116224

Cédula de identidad: 1716710056

Lugar y fecha: Quito, 4 de octubre de 2019

RESUMEN

Ocho genes máster han sido identificados como piezas ortogonales que involucran las vías de señalización de los pilares del cáncer. Entre estas incluyen proliferación, modificaciones metabólicas, resistencia a la muerte celular y otros más específicos de cada enfermedad. Los genes selectos tienen diferentes patrones de expresión, es decir, expresados al alta o a la baja en células cancerosas en comparación con las sanas. Los genes que se encuentran regulados a la baja son *WNT1 inducible signaling pathway protein 1 (WISP1, NM_003882), Aldehyde Dehydrogenase 4 Family Member A1 (ALDH4A1A, NM_003748)* y el *Transforming Growth Factor-β3 (TGF-β3)*. Mientras que los genes que se encuentran regulados al alta son *Small EDRK-rich factor 1A (SERF1A, AF073519), TSPYL5 (AL080059), Matrix Metalloproteinase 9 (MMP9, NM_004994), H-RAS-Like Suppressor (HRASLS, NM_020386)* y el *Small Integral Membrane Protein 5 (SMIM5, Contig24252_RC)*. Estos genes son decisivos para la predicción de la prognosis en el cáncer de mama.

Palabras clave: Cáncer de mama, transcriptoma, expresión genética, firma genética, prognosis, metástasis, aprendizaje automatizado, análisis de agrupamiento, selección de características, WEKA.

ABSTRACT

Eight master genes were identified at the head of the signaling pathways involving various hallmarks of cancer such as proliferation, metabolic modifications, resistance to cell death and others affected by the disease. The selected genes have different expression patterns predicted to be up-regulated or down-regulated in aggressive cancers. Down-regulated genes are WNT1 inducible signaling pathway protein 1 (*WISP1*, NM_003882), Aldehyde Dehydrogenase 4 Family Member A1 (*ALDH4A1*A, NM_003748) and the Transforming Growth Factor-β3 (*TGF-β3*). The up-regulated genes are Small EDRK-rich factor 1A (*SERF1A*, AF073519), *TSPYL5* (AL080059), Matrix Metalloproteinase 9 (*MMP9*, NM_004994), H-RAS-Like Suppressor (*HRASLS*, NM_020386) and the Small Integral Membrane Protein 5 (*SMIM5*, Contig24252_RC). These genes are key features for predicting the prognosis in breast cancer.

Key words: Breast cancer, Transcriptome, gene expression, gene signature, prognosis, metastasis, machine learning, cluster analysis, feature selection, WEKA.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

INTRODUCTION

Breast cancer (BC) is the most frequent tumor in women from developed and developing countries (Nandy, Gangopadhyay, & Mukhopadhyay, 2014). The risk factors at the origin of the disease are widely known, however anticipating cancer behavior to achieve an effective diagnosis, prognosis and treatment is hard. The difficultness is due to the tumor's genetic heterogeneity after the natural selection during its growth and pharmacotherapy. Furthermore, the ethnic profile of a tumor varies in its association to cancer susceptibility and evolution (Foulkes, Knoppers, & Turnbull, 2015). BC development can also be influenced by the life style of the patient as dietary habits, work and perturbation of circadian rhythm triggering and supporting the disease (Gilbertson, 2011; Gupta, Mutebi, & Bardia, 2015; Harris et al., 2016). The develop of a prognosis tool that groups tumor variation plus been able to simplify cancer complexity could mean a better diagnostic tool to foresee its aggressiveness, which is fundamental to apply a preventive, curative or palliative treatment (Nandy et al., 2014).

BC sub-types have helped to predict its evolution or responsiveness to different therapies (I. J. Beumer et al., 2016; Naoi & Noguchi, 2016). Sub-type classification is based on age, size, presence compromised lymph nodes, evidence of metastasis, biomarkers such as ERα (Estrogen Receptor-α), PR (Progesterone Receptor), HER2 (Human Epidermal Growth Factor Receptor 2), markers of cell division like Ki67, and immunohistochemistry (IHC) to distinguish between basal and luminal types (Sims, Howell, Howell, & Clarke, 2007). This basic classification of BC sub-type is informative regarding the aggressiveness, observing generally that woman younger than 35-40 years diagnosed with the disease have a poorer prognosis. The presence of hormone receptors could also determine the type of treatment being endocrine or chemotherapy sensitive (Bundred, 2001). For example, the luminal type positive for hormone receptors and negative for HER, show low levels of Ki67, present a

better prognosis and could be treated only with surgery or with adjuvant endocrine therapy. Moreover, this pattern shows lesser sensitivity to chemotherapy which generally works in highly proliferative cells (Győrffy et al., 2015; Sims et al., 2007). On the other side, basal type of tumors lacks the expression of ERα and PR and are HER2-positive. They are genetically instable, and are associated with poor prognosis (Sims et al., 2007). If the cancer by histological analysis shows greater tissue dissemination and high proliferation, the evidence show association with poor prognosis, but at the same time indicates a good sensitiveness for chemotherapy (Győrffy et al., 2015). At the end, using a multiple sub-type classification, results in a variable prognosis to each patient. Therefore, affecting the decision of the doctor at the time of choosing the treatment to be employed.

These days, advanced molecular diagnostics based on the analysis of gene expression from the patient's breast biopsies, have emerged to help standard prognostic criteria in an early stage BC, when predicting recurrence (Fayanju, Park, & Lucci, 2018; Győrffy et al., 2015; Kwa, Makris, & Esteva, 2017). Gene expression signatures currently recommended in clinical guidelines are: i) Mammaprint™ by Agendia, a 70-gene test recommended for patients with stage II or node-negative plus ER-positive or negative BC with less than 5cm in diameter. Mammaprint™ was the first genomic test approved by the FDA in 2007. ii) The FDA approved Prosigna (Prediction Analysis of Microarrays 50, PAM50) in 2013 for the analysis of tumors positive for hormone receptors with or without lymph-node presence. iii) Even if not yet approved by the FDA, Oncotype DX a 21-gene test, for the possibility of recurrence after receiving chemotherapy, was one of the earliest molecular tests clinically analyzed (Kwa et al., 2017). Other multi-gene signatures such as EndoPredict, Immunohistochemistry 4 (IHC4), the Genomic Grade Index, among others are currently under development and testing, waiting for its analytical and clinical validation (Harris et al., 2016; Kwa et al., 2017; Naoi & Noguchi, 2016). The American Society of Clinical Oncology

(ASCO) recommend on its guidelines the use of Oncotype DX and Prosigna in patients with HR-positive, HER2-negative and node-negative BC. This is based on their capability to accurately and reliably measure the genotype of interest, in order to identify and predict the patient survival at an end-point of 5 – 10 years after surgery (Kwa et al., 2017). The ACSO's guidelines in 2016 made a moderate recommendation for Mammaprint as it should not be used to guide medical decision when using adjuvant chemotherapy, regarding patients with early stage BC, either ER-positive or ER-negative and lymph node status (Kwa et al., 2017). In 2017 the American Joint Committee for Cancer (AJCC) recommend the use of Oncotype DX™, Mammaprint™, EndoPredict™, Prosigna™ and Breast Cancer Index™, when approachable for patients with hormone receptor-positive, HER2-negative and Lymph node-negative tumors (Giuliano et al., 2017). The usage of genetic test has brought a better stratification for ER-positive, HER2-negative, lymph-node negative tumors. Nevertheless, a strong in silico analysis validating the strength, reproducibility and clinical impact of the genes considered on each signature, is necessary to apply in a multi-gene prognostic tools for an everyday clinical practice (Kwa et al., 2017).

Even if these genetic signatures show great prognosis efficacy, having a massive number of genes composing the signature brings incertitude due a lack of accuracy at a patient level. This could provide in many cases different risk categorizations and misleading tendencies (Dai et al., 2015; Sotiriou & Lambertini, 2017). Additionally, a huge number of genes in the signature to be analyze, implies a higher cost for the patients (Y. Li et al., 2017). This creates a need to find a reduce number of genes that makes the test more affordable and precise for each patient. A multiple effective estimation with a congruent analysis of the databases for the development of these signatures is crucial to lower the number of genes. Refining the amount of the genes will improve the precision of patient's prognosis, helping the physician to take the most important decisions such as the administration of adjuvant

chemotherapy and selecting the most appropriate drug according to the characteristics of the patient and the tumor.

This report focuses on the development of a new genetic signature that predicts the prognosis of BC using a comprehensive computational analysis of the distinct patterns of cancer gene expression and its validation in vitro. We analyze sets of genes already used for diagnosis and prognosis of BC in Mammaprint by means of machine learning and artificial intelligence analysis.

TOPIC DEVELOPMENT

Material and methods

Dry experiment.

Samples and clinical data.

78 (Learning Series, LS) and 19 (Test Series, TS) patients with lymph-node negative (pN0 or LN-) BC, 55-years old or younger, not having received adjuvant therapy and that were part of MP studies (van 't Veer et al., 2002) were used as training and prediction groups, respectively. These datasets contained the 70-gene prognosis profile. The expression pattern of the 70 genes in the 78 and 19 samples is shown as supporting information 1 (SI1) A and B, respectively.

The LS78 was used to identify a reduce groups of genes whose expression allowed distinguishing clinical outcomes. Of the 78 patients, 34 developed distant metastases within 5 years (bad-prognosis or poor prognosis group, designed here as Positive, and mean time to metastases 2.5 years). Patients (44) that remained free of disease after initial diagnosis for a period of at least 5 years were assigned to the good prognosis group (i.e. 'good outcome group' designed here as Negative, and mean follow-up of 8.7 years). TS19 was used to evaluate the predictive power of obtained models. This group consisted of 7 (Negative) patients who remained metastasis free for at least five years, and 12 patients (Positive) who developed distant metastases within five years.

Finally, to make a deep validation of the prognosis classifiers, an additional independent set of primary tumors from 295 young, LN- and LN+ BC patients was selected. These External Prediction Test (EPT295, see also SI1-C/B/D) (van de Vijver et al., 2002). Among the 295 patients, 180 had a poor-prognosis signature (positive label) and 115 had a good-prognosis signature (Negative label). All patients had stage I or II BC and were younger than 53 years old; 151 had LN- disease (EPT151), and 144 had LN+ disease (EPT144).

Data processing and feature selection.

Firstly, beginning from MP LS we carried out the following workflow for data reduction: 1) the MDs with NaN (Not-a-Number) values were dropped; and 2) a supervised feature selection with regard to the XXX values was performed using an in-house program. This program retains the best genes according to the Correlation Subset (Hall, 1999) and Relief-F (Robnik-Šikonja & Kononenko, 2003) methods available in the Waikato Environment for Knowledge Analysis (WEKA) software (v3.8). This supervised selection was repeated until obtaining an only matrix of until 50 genes.

Once the previous workflow concluded, the two selection methods aforementioned and a Shannon Entropy (SE)-based unsupervised method (Godden, Stahura, & Bajorath, 2000) were applied on the gene's matrix obtained. Thus, three subsets with the most suitable variables were obtained, so that: i) with the Correlation Subset method(Hall, 1999) a subset of features highly correlated with the prognosis and lowly correlated among them was retained; ii) with the Relief-F method(Robnik-Šikonja & Kononenko, 2003) the features that best distinguish among patients that are near each other were retained; and iii) with the SE-based method(Godden et al., 2000) the features that best yield distinct genes for structurally different patients were retained. The SE-based selection was performed with the IMMAN software, (Urias et al., 2015) following a discretization scheme equal to 78 bins. These three subsets were also shuffled among themselves (ensemble feature selection (Kuncheva, 2014)), obtaining four additional subsets. A wrapper-type selection (Kohavi & John, 1997) was also applied on all the previous subsets, by using the Best-First method as search strategy, the accuracy (Q%) as evaluation measure, and the k-Nearest Neighbor (k-NN),(Altman, 1992) Multilayer Perceptron (MLP),(Murtagh, 1991) Random Forest (RF) (Breiman, 2001) and Support Vector Machine (SVM), (Shevade, Keerthi, Bhattacharyya, & Murthy, 2000)

Logistic regression (LR)(Le Cessie & van Houwelingen, 1992), Linear Discriminant Analysis (LDA) learning methods.

Machine Learning Modelling.

Initially Exploration with LDA. To obtain the binary predictions with statistical models developed using prognosis classes for the LS, we followed the criterion that patient with experimental xxx were classified as poor prognosis (i.e., Positive), while patients with xxx were classified as good prognosis (i.e., Positive). The dependent variable was then assigned a value of 1 or -1 when the patient had XXX greater than or lower than the threshold of 5 years, respectively. Statistical analysis was carried out with STATISTICA package ("STATISTICA (data analysis software system)," 2001). The LDA was used to find the classifier functions (van de Waterbeemd, 1995) by using the forward stepwise and best subset methods for the attribute selection. The tolerance parameter was set to 0.01. By using the models, one patient can be classified as either positive, if $\Delta P\% > 0$, being $\Delta P\% = [P(+) - P(-)] \times 100$, or Negative otherwise. P (+) and P (-) are the probabilities with which the equations classify a patient as positive and negative, respectively.

The quality of the models was determined according to Wilks' λ, the square of the Mahalanobis distance D2, Fisher ratio (F), significance level (p) and the percentage of good classification (accuracy, Q). Therefore, parameters like sensitivity 'hit rate' (SE), specificity (SP), false positive rate (fprate) (also called false alarm rate) and Matthews' correlation coefficient (MCC) were taken into account (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). Also, the principle of parsimony (Occam's razor) was considered, in that models with high statistical significance but having as few parameters as possible were preferred. However, the main criterion to select the best model is based on the prediction statistics for a Test Set that were never used in the process of model development (Tropsha, 2010).

Individual (base) Models. Subsequently, several individual classification models were built using the learning methods aforementioned, which are available in the Weka software (v3.8). In detail, the k-NN method was configured with k-values (number of neighbors) between 3 and 6 (determined by cross-validation), and the ratio 1/distance was used as distance weighting scheme. Moreover, the Puk function (Üstün, Melssen, & Buydens, 2006) was used as kernel in the SVM. The remaining learning methods were utilized with the respective default configurations. These configurations were also employed in the wrapper-type selection explained above. It is important to highlight that on the subsets of MDs obtained from the wrapper-type selection, only the learning method used in the wrapper was employed to build the model. The training assessment was performed through 10-fold cross-validation. The best and less redundant base (individual) models were retained to build ensemble (consensus) models.

Ensemble (Multi-classification models). The ensemble models were built using the Bagging, (Breiman, 1996) Additive Regression(Friedman, 2002) and Stacking(Wolpert, 1992) procedures; while the consensus models were developed using the Maximum (Max), Minimum (Min), Average (AVE) and Weighted Average (WA) aggregation functions. The weightings for the WA function were calculated according to Equation 9 described in ref. (García-Jacas et al., 2018) whit the ensemble models with a training performance better than the one achieved by the best individual model, as well as all the consensus models built, were assessed on the TS and ETS sets. In this way, the best models based on the LS were selected according to their training and test results.

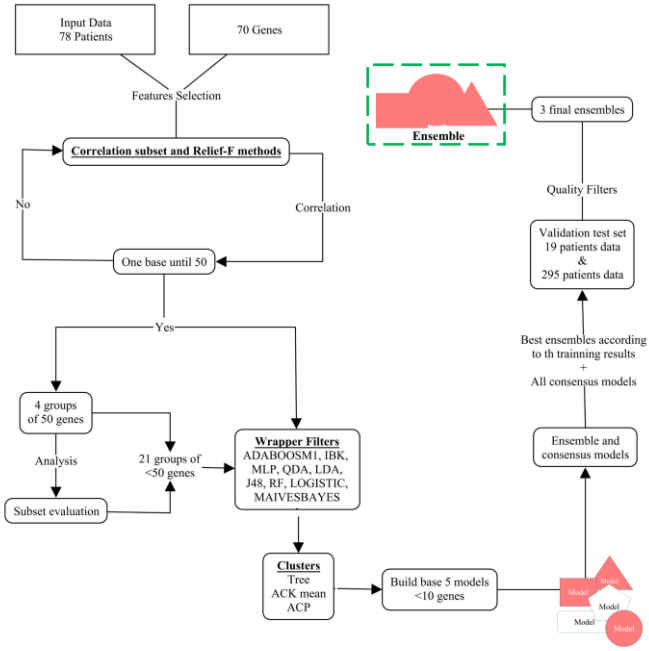

Figure 1: Workflow analysis of the 70 genes towards the ensemble models. Applicability domain.

An applicability domain (AD) analysis was carried out to know the behavior of the models to perform reliable predictions.(Dragos, Gilles, & Alexandre, 2009; Roy, Kar, & Ambure, 2015; Sahigara et al., 2012) This analysis is often performed using only one method to determine the structures that fall within the AD of a model (e.g., Liew et al.(Liew, Lim, & Yap, 2011) only used an approach based on range, and Aranda et al.(Aranda, Garro Martinez, Castro, & Duchowicz, 2016) only used the leverage method). However, there are several AD methods that characterize of distinct way the interpolation space stated by the MDs used.(Sahigara et al., 2012) So, inspired on the idea of consensus-based decision, five approaches (i.e., City-block, Euclidean, Mahalanobis, Range and Density – see refs.(J, N, & T, 2005; Netzeva et al., 2005) for a detailed description) were considered in a recent study to determine the reliability of the predictions.(García-Jacas, Martinez-Mayorga, Marrero-Ponce, & Medina-Franco, 2017) More specifically, if the prediction for a molecule lies outside of the bounds in at least two AD methods, then it is considered as unreliable. This approach will be

the one used in this report. These AD methods are available in the Ambit Discovery software (2008) and they were used with their respective default configurations.

Clustering analysis.

Clustering analysis (CAs) are simple and extremely useful non-supervised data mining methods to explore relationships that exist among objects and allocate to the same classes the similar ones, on the basis of predefined similarity (or dissimilarity) measures (Barnard & Downs, 1992; Brown & Martin, 1996). First k-nearest neighbours cluster analysis (k-NNCA), also known as hierarchical agglomerative clustering, was performed by using Ward's Linkage and the Euclidean distance as amalgamation rule and proximity function, respectively, to have preliminary insight on the ''possible'' number of clusters that naturally exist in the examined data, to be later used in the k-Means Cluster Analysis (k-MCAs).
To evaluate the statistical quality of data partitions in the clusters a standard analysis of variance (ANOVA) for each dimension (variable) was performed. The values of the standard deviation (SS) between and within clusters, of the respective Fisher's ratio and their p level of significance, were examined (Johnson & Wichern, 1988; Mc Farland & Gans, 1995). Statistical analysis was carried out with STATISTICA package ("STATISTICA (data analysis software system)," 2001).

WET experiment.

Cell culture.

The expression profile of the 8 genes analyzed were tested in 3 different cell lines: Fibroblasts, MCF-7 and MDA-MB-231. Dr. Patricia Luz from "Los Andes University" in Chile donated the primary Fibroblasts obtained from the cell culture of human early birth foreskin with informed consent; MDA-MB-231 and MCF-7 are ATCC cell lines. All cells were cultured with Dulbecco's Modified Eagle Medium (DMEM 10%) High Glucose (11965; Gibco by Life Technologies, ThermoFisher Scientific, MA, U.S) supplemented with

10% Fetal Bovine Serum (FCS), and Penicillin- Streptomycin (10,000 U/mL) according to the manufacturer recommendations. Cells were incubated at 37°C with 5% CO2. The Cells passage was performed when cells achieved 80% confluence in a T75 flask with Filter (Corning, U.S.) to prevent overgrowth and changes in their gene expression profile. Fibroblasts were maintained in culture for 45 days till 6 passages during this time.

RNA extraction.

The total RNA was isolated from 2 million cells using PureLink™ RNA Mini Kit (12183025; Ambion by Life Technologies, ThermoFisher Scientific, MA, U.S), according to the manufacturer's instructions. RNA quantification was assessed using the Epoch Microplate Spectrophotometer (BioTek Instrument, Winooski, VT, USA). RNA was diluted with RNase-free H2O to reach a final concentration of 35ng/□l. A dilution of the RNA was performed to obtain aliquots and preserve its integrity with a 75□l sample stored at -20ºC and a 25□l sample backup stored -80ºC, assays were not repeated more than three times from a single aliquot to assure mRNA quality.

Expression profiles of the 8 genes quantified by RT-qPCR.

Real Time quantitative PCR was performed in a StepOne Real-Time PCR System (Marca). The SuperScript® III Platinum® SYBER ® Green One-Step qRT-PCR Kit with Rox was used in the quantification reaction (11745500; Invitrogen by Life Technologies, ThermoFisher Scientific, MA, U.S) in a final 10µl volume. The analyzed genes and primer sequence are listed in the SI2. As a standard to the primers, a housekeeping gene was used: PUM1 (F:5'AGTGGGGGACTAGGCGTTAG3'; R:5'GTTTTCATCACTGTCTGCATCC3'), a positive control was also used with ZNF154 (F:5'-GGTTTTTATTTTAGGTTTGA–3'; R:5'-AAATCTATAAAAACTACATTACCTAAAATACTCTA–3'). The qPCR reaction followed the beneath conditions: pre-incubation at 95 °C for 5'; denaturation at 95 °C for 15'';

annealing and extension at 60’ °C for 30’’ (45 cycles for denaturation and extension stages); a holding stage at 40ºC for 1’; melting at 95 °C for 15’’; 60 °C for 1’; and 95 °C for 15’’. The expression profiles of the 8 genes were obtained estimating the 2-ΔCT and its geometric means, duplicates were performed for each gene and cell line.

Statistical analysis.

Data was analyzed using GraphPad Prism 6. We performed an un-paired, non-parametric t test with Mann-Whitney to analyze the differences among conditions regarding the gene expression of the panel, using an alpha-value of 0.05 (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$).

Results

All essays outcomes were divided in two groups. The first group compares each gene expression within the three cell lines: Fibroblasts, MCF-7, MDA-MB-231. The second one, relates the expression of all genes within one cell line secluded. Figure 1 represents the first group of essays with the geometric means distribution of neat cyclin times (*CT*) of PUM1 expression in each cell line with no difference within them; it also shows geometric means distribution of *2-ΔCT* of 8 genes.

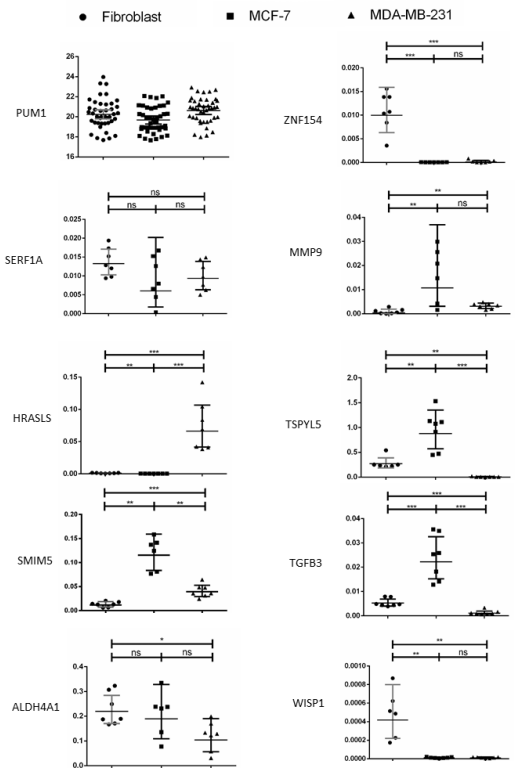

Figure 2: Individual gene expression of the Master-Gene gene panel tested in normal (fibroblasts) and phenotypically different breast cancer cells (MCF-7 and MDA-MB-231). Assays were performed 7 times for each gene in the three different cell lines (● fibroblast, ■ MDF-7 and ▲MDA-MB-231). The 2-ΔCT fold by qPCR was performed with PUM1 as the housekeeping. a. PUM1, CTs of expression are plotted for the three cell lines analyzed. The geometric mean of the 2-ΔCT fold for each gene and assay is shown in the scatter plot b. ZNF154, c. SERF1A, d. MMP9, e. HRASLS, f. TSPYL5, g. SMIM5, h. TGFβ3, i. ALDH41 and j. WISP1. Statistical analysis for all conditions: Mean ±SD; un-paired, non-parametric t test with Mann-Whitney (*p < 0.05, **p < 0.01, ***p < 0.001).

Positive control *ZNF-154* gene expression is superior in Fibroblasts than in MCF-7 and MDA-MB-231 ($p < 0.001$). *SERF1A* shows no significant difference comparing the three cell lines. *MMP9* is up-regulated in Fibroblasts in contrast to MCF-7 and MDA-MB-231 ($p < 0.01$). *HRASLS* is up-regulated in MDA-MB-231 match up to Fibroblasts and MCF-7 ($p < 0.001$); there is also a significant difference between the last two cells ($p < 0.01$). *TSPYL5* is down-regulated in MDA-MB-231 judged against to Fibroblasts ($p < 0.01$) and MCF-7 ($p < 0.001$); there is a significant difference flanked by Fibroblasts and MCF-7 ($p < 0.01$). *SMIM5*'s MCF-7 expression is higher compared to both Fibroblasts and MDA-MB-231($p < 0.01$). In addition, there is statistical difference in MDA-MB-231 and

Fibroblasts ($p < 0.001$). *TGF-β3* shows a pronounced expression in MCF-7 related to the remaining cell lines, with statistical difference among the three of them. *ALDH4A1* is down-regulated in MDA-MB-231 and MCF-7 associated to Fibroblasts, however the only proper difference is between MDA-MB-231 and Fibroblasts ($p < 0.05$). *WISP1* shows elevated expression in Fibroblasts weight against MCF-7 and MDA-MB-231 ($p < 0.01$).

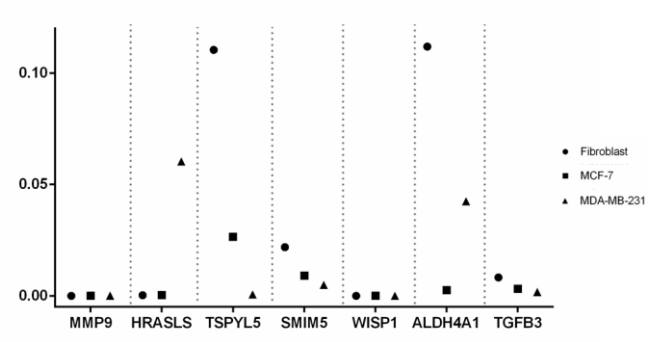

Figure 3: The Master-Gene gene panel tested in normal (fibroblasts) and phenotypically different breast cancer cells (MCF-7 and MDA-MB-231). Assays were performed 3 times for each cell lines (● fibroblast, ■ MDF-7 and ▲MDA-MB-231). The 2-ΔCT fold by qPCR was performed with PUM1 as the housekeeping. a. The Scatter Plot of the Geometric Mean for each gene of the three assays by cell line is shown. Individual analysis of the gene expression by assay for each cell line is presented in the following graphs b. MMP9, c. HRASLS, d. TSPYL5, e. SMIM5, f. WISP1, g. ALDH4A1 and j. TGFβ3. Statistical analysis for all conditions: Mean ±SD; un-paired, non-parametric t test with Mann-Whitney (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

Figure 3 exemplifies the second group of essays with the geometrical mean of *2-ΔCT* summary plot of all gene's expression in each cell line. The genes expression pattern of MCF-7 compared to Fibroblasts shows up-regulation of *HRASLS*, *WISP1* and down-regulation of *MMP9*, *TSPYL5*, *SMIM5*, *ALDH4A1*, and *TGF-β3*. While MDA-MB-231 contrasted to Fibroblasts illustrates up-regulation of *MMP9*, *HRASLS*, *WISP1* and down-regulation of *TSPYL5*, *SMIM5*, *ALDH4A1*, and *TGF-β3*.

CONCLUSIONS

After the identification of the seven Master-Genes, we analyzed their expression in human BC cell lines MCF-7 and MDA-MB-231 together with non-cancerous foreskin Fibroblasts as a control. Mammaprint™ is principally recommended for ER/PR positive, HER2-negative tumors. We choose as an ideal hormone sensitive BC model, the MCF-7 cell line for being ER+, PR+ HER-, Ki67 low, endocrine and chemotherapy responsive cell line (Holliday & Speirs, 2011). Then, we also choose MDA-MB-231 as a highly aggressive Caludin-low ER-, PR-, HER-, positive to Ki67 with an intermediate response to chemotherapy as a contrast to the MCF-7 phenotype (Dai, Cheng, Bai, & Li, 2017).

PUM1 was chosen as a housekeeping gene, it encodes the RNA-binding protein known as PUF, which is characterized by a highly conserved C-terminal RNA-binging domain, composed of eight tandem repeats. The protein binds to related sequence motifs in the 3' untranslated region (3'UTR) of specific target mRNAs and repress their translation. *PUM1* constitutive transcription has a role in cell development, specification and differentiation of tissues in a large scale of organisms, varying from yeast to humans and plants (Spassov & Jurecic, 2003). *PUM1* is the most stable gene for the normalization of expression in invasive breast tumor studies, shown in the calculations made with both NormFinder and BestKeeper (Kılıç, Çelebiler, & Sakızlı, 2014). Moreover, *ZNF154* was used as a positive control thanks to the evidence of highly methylated promoters across a wide spectrum of human solid epithelial tumors, including BC (Margolin et al., 2016; Sanchez-Vega et al., 2013).

Small EDRK-rich Factor 1A gene, (*SERF1A*, AF073519), has a role associated to autophagy activation and avoid the accumulation of dysfunctional

proteins, however, it's role in oncological diseases is disconcerted (Tian, Roepman, van't Veer, et al., 2010; Veltman et al., 2005). Experimentally, we can mention that there is not significant difference between the expression in the three cell lines. This means that even though a gene signature could include a gene like this, it not possible to know why it behaves like this on the cell cultures, because of its unknown function.

Aldehyde Dehydrogenase 4 Family Member A1, (*ALDH4A1*, NM_003748), encodes a mitochondrial-matrix NAD+-dependent enzyme that catalyzes the second step of the proline degradation pathway, which is done by the oxidation of L-glutamate-γ-semialdehyde (GSA) to L-glutamate (Srivastava et al., 2012; Yoon, Nakamura, & Arakawa, 2004). In BC, down-regulation of this gene is associated with the sustained angiogenesis by alternating the metabolism under hypoxia in the microenvironment (Tian, Roepman, van't Veer, et al., 2010; van 't Veer et al., 2002). We found a lower expression in MDA-MB-231 versus Fibroblasts. It is interesting to know that this gene cannot be differentiated in both Fibroblasts and MDA-MB-231 from MCF7. This could be due to its role on the expression of *p53* generating a protective factor against oxidative stress seen with higher levels of expression (Pors & Moreb, 2014). Meaning that depending of the cell phenotype this gene could have different function that correlates to the level of expression in different type of tumors.

Testis-specific Y-encoded-like protein 5, (*TSPYL5*, AL080059), can override senescence-like proliferation arrest, oncogene-induced senescence and contribute to cell transformation. This is given by its role as a negative regulator of *p53* function and modulator of cell growth. (Shen et al., 2015). P53 binds to the N-terminal domain of USP7, however TSPYL5 binds to the same domain to reduce the stability of P53 by competing with it. There is evidence of this interaction with USP7 in MCF7 cells from BC and primary human fibroblasts (Cheng, Niu, Yang, Wang, & Lu, 2013;

Epping et al., 2011; Gao et al., 2013). More fascinating, *TSPYL5* is been known to work as a tumor suppressor in various tissues. Therefore, when hyper-methylated it helps to development gastric and colon cancer, hepatocarcinoma, and glioma. Our results suggest a similar mechanism of action on BC MDA-MB-231 cells even though there is no evidence of the epigenetic state of the gene in this situation. There is evidence of a cell type such as MDA-MB-231 on gastric cancer with low expression of *TSPYL5* (Jung, Park, Bang, & Kim, 2007; Qiu et al., 2016; Shen et al., 2015).

WNT1 inducible signaling pathway protein 1 (*WISP1*, NM_003882), is a cysteine rich secreted matricellular protein that can modulate cellular responses, such as cell growth, differentiation and survival (Berschneider B, 2010). It has been associated with different types of cancer including breast, lung, and colon, gastric, brain, pancreatic, sarcoma, ovarian, esophageal, head and neck (Davies, Watkins, Mansel, & Jiang, 2007; Gurbuz & Chiquet-Ehrismann, 2015). It works as an oncogene regulated by the Wnt-1-β-catenin pathway, in which the mutations on *Wnt-1* produce accumulation of β-catenin in the cytoplasm. Then, the *CREB*-binding site in the *WISP1* gene is activated by PKA because of the increase of cAMP by β-catenin. In this context, the transcriptions of *WISP1* is elevated promoting an excess in cell proliferation (Chiang et al., 2015). There is a mechanism suggesting that *WISP1* can activate the antiapoptotic Akt/PKB signaling pathway. Which works by inhibition of the mitochondrial release of cytochrome *c* and up-regulation of the antiapoptotic *Bcl-X_L*, and so blocking *p53* dependent cell death through sequestering *p53*-mediated BH3 domain-only molecules such as PUMA or Noxa (Su, Overholtzer, Besser, & Levine, 2002). Besides, all oncogenic characteristics liked to elevated levels of expression, in 2007 was published evidence that the levels of *WISP1* are down-regulated when the tumor has poor prognosis, presence of nodal involvement,

bad clinical outcome, advanced TNM scores, and a more advance grade in tumor differentiation (Davies et al., 2007). After, in 2017 it is first described that Wnt-derived BC tumors have an aberrant re-activation post Wnt-inhibition. Making the reactivation process the driver of BC recurrence and progression (Pohl et al., 2017). This mechanism could explain the down regulation in cancer cells seen in our results compared to Fibroblasts.

H-RAS-Like Suppressor (*HRASLS*, NM_020386), family of enzymes consist in three members in humans with a phospholipase A/acyltransferase (PLA/AT) activity regulating the function of RAS family of GTPases specially K-RAS. *HRASLS* could be implicated in the regulation of multiple cellular processes, among them growth, proliferation, cell death, invasiveness and angiogenesis (Han, Jeong, & Jang, 2017). It also reduces the palmitoylation of H-RAS which induces its deactivation. Additionally, the gene has an important role in fat metabolism in adipocytes through the PGE2/cAMP pathway and lipid degradation (Han et al., 2017; Tsai, Chen, Wang, & Lee, 2015). Even if the role of *HRASLS* has been linked to the regulation of RAS family members, its complete molecular interactions with other mechanism in cancer has not been completely elucidated. Interestingly, the expression of *HRASLS* in BC has been related to a bad prognosis through our and others bioinformatic analysis, which opens the field to deeper investigation (Beumer et al., 2016; van 't Veer et al., 2002). In correlation with our results, *HRASLS* is one of the most frequently genes overexpressed in BC (Di Benedetto et al., 2015).

The Transforming Growth Factor-β (*TGF-β*) superfamily are cytokines composed by three described isoforms from the genes of *TGF-β1, -β2, and -β3*. They function as ligands that bind to a heteromeric complex of transmembrane TGF-β serine/threonine kinase type I and II receptors. Consequently, this activates the Smad

pathway involving receptor-regulated Smads 2 and -3 and its mediator Smad 4 to form heteromeric complexes. The final response is evoked in the nucleus as transcriptional responses (de Kruijf et al., 2013). In breast cancer, all isoforms of TGF-β have a dual role raging from a tumor suppressor function to an oncogenic role. Specifically, *TGF-β3* has a role in the maturation of myoepithelium, suppression of milk and epithelial–mesenchymal transition (EMT). More important, there is increased expression in breast cancer cells with positive hormone receptors, compared to normal tissues (Hachim, Hachim, Dai, Ali, & Lebrun, 2018). We validated this with evidence of increased expression of the gene in MCF7 cells. The down-regulation of the gene is only seen on MDA-MB-231 cells and is associated with bad prognosis (Tian, Roepman, van't Veer, et al., 2010).

Matrix Metalloproteinase 9 (*MMP9*, NM_004994) comes from a family of endopeptidases that has a gelatin-binding structural class which is composed by a domain called minimal-domain (Egeblad & Werb, 2002). It is secreted as a zymogen, but when activated is capable of digesting different types of collagen and other elements of the extracellular matrix for example elastin, fibrillin, laminin, gelatin. This allows the protein to participate in different processes both physiological an pathologic, such as neurite growth, embryonic development, angiogenesis, ovulation, mammary gland involution, wound healing, osteoblastic bone formation, and osteoclastic bone resorption (Perrotta, Sciangula, Aquila, & Mazzulla, 2016; Visse & Nagase, 2003). More important, MMP9 participates in different hallmarks of cancer. It promotes cellular growth signals, regulates apoptosis, increases angiogenesis, facilitates metastasis and intravasation, suppresses the immune system (Bergers et al., 2000; Coussens, Tinkle, Hanahan, & Werb, 2000; Egeblad & Werb, 2002; Fang et al., 2000; Kim, Yu, Kovalski, & Ossowski; Sheu et al., 2001; Vu et al., 1998). All this

supports the idea of up-regulation in BC and its close relation with lymph node metastasis and tumor staging. Thus, making it to become an indicator of bad prognosis (Li, Qiu, Li, & Wang, 2017).

Small Integral Membrane Protein 5 (*SMIM5*, Contig24252_RC) encodes a transmembrane protein with unknown functions, but related to bad prognosis in MammaPrint's BC signature (Marchionni L, 2008; Tian, Roepman, van't Veer, et al., 2010). In our results, we can infer the high expression of this gene is associated with both cancer cell phenotypes.

This study provides insights with repercussions for the future BC population, as well as for the study of its prognosis and treatment techniques according to the cell's hostility of the tumor. We have shown the expression of 8 genes in different cell lines *in vitro*, which corroborate the explanation of the Hallmarks of cancer and are up regulated or down regulated according to its function. Our analysis also shows that gene expression is related to specific phenotypes of the cell that could determine the prognosis of the patient. This leaves us to deliberate about the importance of an appropriate survival prediction for the near future, not only in BC, but in other neoplastic diseases in order to make personalized cancer therapy more feasible (Pukazhendhi & Glück, 2014). Nowadays there are gene signatures that help to determine BC prognosis using a massive number of genes and redundancy on its information at the time of predicting the prognosis. Such situation is unfortunate for any patient due to its cost and precision. We determined that it is possible to generate a gene signature with minor number of genes that achieve the same force of prediction as others existing. We also, proved the strength of it with an *in vitro* cell culture essay, that is comparable to a trial with BC biopsy samples. All this as background explains the reason why these seven genes are a practical signature for BC prognosis.

REFERENCES

Bergers, G., Brekken, R., McMahon, G., Vu, T. H., Itoh, T., Tamaki, K., . . . Hanahan, D. (2000). Matrix metalloproteinase-9 triggers the angiogenic switch during carcinogenesis. *Nature cell biology, 2*(10), 737-744. doi:10.1038/35036374

Berschneider B, K. M. (2010). WNT1 inducible signaling pathway protein 1 (WISP1): a novel mediator linking development and disease. *Int J Biochem Cell Biol, 43*(3), 306-309. doi:10.1016/j.biocel.2010.11.013

Beumer, I., Witteveen, A., Delahaye, L., Wehkamp, D., Snel, M., Dreezen, C., . . . Glas, A. (2016). Equivalence of MammaPrint array types in clinical trials and diagnostics. *Breast Cancer Research and Treatment, 156*(2), 279-287. doi:10.1007/s10549-016-3764-5

Cheng, C., Niu, C., Yang, Y., Wang, Y., & Lu, M. (2013). Expression of HAUSP in gliomas correlates with disease progression and survival of patients. *Oncol Rep, 29*(5), 1730-1736. doi:10.3892/or.2013.2342

Chiang, K.-C., Yeh, C.-N., Chung, L.-C., Feng, T.-H., Sun, C.-C., Chen, M.-F., . . . Juang, H.-H. (2015). WNT-1 inducible signaling pathway protein-1 enhances growth and tumorigenesis in human breast cancer. *Scientific Reports, 5*, 8686. doi:10.1038/srep08686

Coussens, L. M., Tinkle, C. L., Hanahan, D., & Werb, Z. (2000). MMP-9 Supplied by Bone Marrow–Derived Cells Contributes to Skin Carcinogenesis. *Cell, 103*(3), 481-490. doi:http://dx.doi.org/10.1016/S0092-8674(00)00139-2

Dai, X., Cheng, H., Bai, Z., & Li, J. (2017). Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. *Journal of Cancer, 8*(16), 3131-3141. doi:10.7150/jca.18457

Davies, S. R., Watkins, G., Mansel, R. E., & Jiang, W. G. (2007). Differential Expression and Prognostic Implications of the CCN Family Members WISP-1, WISP-2, and WISP-3 in Human Breast Cancer. *Annals of Surgical Oncology, 14*(6), 1909-1918. doi:10.1245/s10434-007-9376-x

de Kruijf, E. M., Dekker, T. J. A., Hawinkels, L. J. A. C., Putter, H., Smit, V. T. H. B. M., Kroep, J. R., . . . Mesker, W. E. (2013). The prognostic role of TGF-β signaling pathway in breast cancer patients. *Annals of Oncology, 24*(2), 384-390. doi:10.1093/annonc/mds333

Di Benedetto, M., Toullec, A., Buteau-Lozano, H., Abdelkarim, M., Vacher, S., Velasco, G., . . . Perrot-Applanat, M. (2015). MDA-MB-231 breast cancer cells overexpressing single VEGF isoforms display distinct colonisation characteristics. *British Journal Of Cancer, 113*, 773. doi:10.1038/bjc.2015.267

https://www.nature.com/articles/bjc2015267#supplementary-information

Egeblad, M., & Werb, Z. (2002). New functions for the matrix metalloproteinases in cancer progression. *Nat Rev Cancer, 2*(3), 161-174. doi:http://www.nature.com/nrc/journal/v2/n3/suppinfo/nrc745_S1.html

Epping, M. T., Meijer, L. A. T., Krijgsman, O., Bos, J. L., Pandolfi, P. P., & Bernards, R. (2011). TSPYL5 suppresses p53 levels and function by physical interaction with USP7. *Nat Cell Biol, 13*(1), 102-108. doi:http://www.nature.com/ncb/journal/v13/n1/abs/ncb2142.html#supplementary-information

Fang, J., Shing, Y., Wiederschain, D., Yan, L., Butterfield, C., Jackson, G., . . . Moses, M. A. (2000). Matrix metalloproteinase-2 is required for the switch to the angiogenic

phenotype in a tumor model. *Proceedings of the National Academy of Sciences of the United States of America, 97*(8), 3884-3889.

Gao, Y., Koppen, A., Rakhshandehroo, M., Tasdelen, I., van de Graaf, S. F., van Loosdregt, J., . . . Kalkhoven, E. (2013). Early adipogenesis is regulated through USP7-mediated deubiquitination of the histone acetyltransferase TIP60. *4*, 2656. doi:10.1038/ncomms3656

http://dharmasastra.live.cf.private.springer.com/articles/ncomms3656#supplementary-information

Gurbuz, I., & Chiquet-Ehrismann, R. (2015). CCN4/WISP1 (WNT1 inducible signaling pathway protein 1): A focus on its role in cancer. *The International Journal of Biochemistry & Cell Biology, 62*, 142-146. doi:http://dx.doi.org/10.1016/j.biocel.2015.03.007

Hachim, M. Y., Hachim, I. Y., Dai, M., Ali, S., & Lebrun, J.-J. (2018). Differential expression of TGFβ isoforms in breast cancer highlights different roles during breast cancer progression. *Tumor Biology, 40*(1), 1010428317748254. doi:10.1177/1010428317748254

Han, C. W., Jeong, M. S., & Jang, S. B. (2017). Molecular interaction between K-Ras and H-REV107 in the Ras signaling pathway. *Biochemical and Biophysical Research Communications, 491*(2), 257-264. doi:https://doi.org/10.1016/j.bbrc.2017.07.120

Holliday, D. L., & Speirs, V. (2011). Choosing the right cell line for breast cancer research. *Breast Cancer Research, 13*, 215. doi:10.1186/bcr2889

Jung, Y., Park, J., Bang, Y.-J., & Kim, T.-Y. (2007). Gene silencing of TSPYL5 mediated by aberrant promoter methylation in gastric cancers. *Lab Invest, 88*(2), 153-160. doi:http://www.nature.com/labinvest/journal/v88/n2/suppinfo/3700706s1.html

Kılıç, Y., Çelebiler, A. Ç., & Sakızlı, M. (2014). Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer. *Clinical and Translational Oncology, 16*(2), 184-190. doi:10.1007/s12094-013-1058-5

Kim, J., Yu, W., Kovalski, K., & Ossowski, L. Requirement for Specific Proteases in Cancer Cell Intravasation as Revealed by a Novel Semiquantitative PCR-Based Assay. *Cell, 94*(3), 353-362. doi:10.1016/S0092-8674(00)81478-6

Li, H., Qiu, Z., Li, F., & Wang, C. (2017). The relationship between MMP-2 and MMP-9 expression levels with breast cancer incidence and prognosis. *Oncology Letters, 14*(5), 5865-5870. doi:10.3892/ol.2017.6924

Marchionni L, W. R., Marinopoulos SS, et al. . (2008). Impact of Gene Expression Profiling Tests on Breast Cancer Outcomes. *Evidence Reports/Technology Assessments*(160).

Margolin, G., Petrykowska, H. M., Jameel, N., Bell, D. W., Young, A. C., & Elnitski, L. (2016). Robust Detection of DNA Hypermethylation of ZNF154 as a Pan-Cancer Locus with in Silico Modeling for Blood-Based Diagnostic Development. *J Mol Diagn, 18*(2), 283-298. doi:10.1016/j.jmoldx.2015.11.004

Perrotta, I., Sciangula, A., Aquila, S., & Mazzulla, S. (2016). Matrix Metalloproteinase-9 Expression in Calcified Human Aortic Valves: A Histopathologic, Immunohistochemical, and Ultrastructural Study. *Applied Immunohistochemistry & Molecular Morphology, 24*(2), 128-137. doi:10.1097/pai.0000000000000144

Pohl, S.-G., Brook, N., Agostino, M., Arfuso, F., Kumar, A. P., & Dharmarajan, A. (2017). Wnt signaling in triple-negative breast cancer. *Oncogenesis, 6*, e310. doi:10.1038/oncsis.2017.14

Pors, K., & Moreb, J. S. (2014). Aldehyde dehydrogenases in cancer: an opportunity for biomarker and drug development? *Drug Discovery Today, 19*(12), 1953-1963. doi:http://dx.doi.org/10.1016/j.drudis.2014.09.009

Pukazhendhi, G., & Glück, S. (2014). Circulating tumor cells in breast cancer. *Journal of Carcinogenesis, 13*, 8. doi:10.4103/1477-3163.135578

Qiu, X., Hu, B., Huang, Y., Deng, Y., Wang, X., & Zheng, F. (2016). Hypermethylation of ACP1, BMP4, and TSPYL5 in Hepatocellular Carcinoma and Their Potential Clinical Significance. *Digestive Diseases and Sciences, 61*(1), 149-157. doi:10.1007/s10620-015-3878-3

Sanchez-Vega, F., Gotea, V., Petrykowska, H. M., Margolin, G., Krivak, T. C., DeLoia, J. A., . . . Elnitski, L. (2013). Recurrent patterns of DNA methylation in the ZNF154, CASP8, and VHL promoters across a wide spectrum of human solid epithelial tumors and cancer cell lines. *Epigenetics, 8*(12), 1355-1372. doi:10.4161/epi.26701

Shen, J., LeFave, C., Sirosh, I., Siegel, A. B., Tycko, B., & Santella, R. M. (2015). Integrative epigenomic and genomic filtering for methylation markers in hepatocellular carcinomas. *BMC Medical Genomics, 8*, 28. doi:10.1186/s12920-015-0105-1

Sheu, B.-C., Hsu, S.-M., Ho, H.-N., Lien, H.-C., Huang, S.-C., & Lin, R.-H. (2001). A Novel Role of Metalloproteinase in Cancer-mediated Immunosuppression. *Cancer Research, 61*(1), 237-242.

Spassov, D., & Jurecic, R. (2003). The PUF Family of RNA-binding Proteins: Does Evolutionarily Conserved Structure Equal Conserved Function? *IUBMB Life, 55*(7), 359-366. doi:doi:10.1080/15216540310001603093

Srivastava, D., Singh, R. K., Moxley, M. A., Henzl, M. T., Becker, D. F., & Tanner, J. J. (2012). The Three-Dimensional Structural Basis of Type II Hyperprolinemia. *Journal of Molecular Biology, 420*(3), 176-189. doi:http://dx.doi.org/10.1016/j.jmb.2012.04.010

Su, F., Overholtzer, M., Besser, D., & Levine, A. J. (2002). WISP-1 attenuates p53-mediated apoptosis in response to DNA damage through activation of the Akt kinase. *Genes & Development, 16*(1), 46-57. doi:10.1101/gad.942902

Tian, S., Roepman, P., van't Veer, L. J., Bernards, R., De Snoo, F., & Glas, A. M. (2010). Biological Functions of the Genes in the Mammaprint Breast Cancer Profile Reflect the Hallmarks of Cancer. *Biomarker Insights, 5*, BMI.S6184. doi:10.4137/bmi.S6184

Tian, S., Roepman, P., van't Veer, L. J., Bernards, R., de Snoo, F., & Glas, A. M. (2010). Biological Functions of the Genes in the Mammaprint Breast Cancer Profile Reflect the Hallmarks of Cancer. *Biomarker Insights, 5*, 129-138. doi:10.4137/BMI.S6184

Tsai, F.-M., Chen, M.-L., Wang, L.-K., & Lee, M.-C. (2015). H-rev107 Regulates Cytochrome P450 Reductase Activity and Increases Lipid Accumulation. *PLOS ONE, 10*(9), e0138586. doi:10.1371/journal.pone.0138586

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., . . . Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536. doi:http://www.nature.com/nature/journal/v415/n6871/suppinfo/415530a_S1.html

Veltman, I., Veltman, J., Janssen, I., Hulsbergen-van de Kaa, C., Oosterhuis, W., Schneider, D., . . . van Kessel, A. G. (2005). Identification of recurrent chromosomal aberrations in germ cell tumors of neonates and infants using genomewide array-based comparative genomic hybridization. *Genes, Chromosomes and Cancer, 43*(4), 367-376. doi:10.1002/gcc.20208

Visse, R., & Nagase, H. (2003). Matrix Metalloproteinases and Tissue Inhibitors of Metalloproteinases. *Structure, Function, and Biochemistry, 92*(8), 827-839. doi:10.1161/01.res.0000070112.80711.3d

Vu, T. H., Shipley, J. M., Bergers, G., Berger, J. E., Helms, J. A., Hanahan, D., . . . Werb, Z. (1998). MMP-9/Gelatinase B Is a Key Regulator of Growth Plate Angiogenesis and Apoptosis of Hypertrophic Chondrocytes. *Cell, 93*(3), 411-422.

Yoon, K.-A., Nakamura, Y., & Arakawa, H. (2004). Identification of ALDH4 as a p53-inducible gene and its protective role in cellular stresses. *J Hum Genet, 49*(3), 134-140.