UNIVERSIDAD NACIONAL DE CÓRDOBA

CLASIFICACIÓN ROBUSTA DE MUESTRAS DE CÁNCER DE MAMAS

Para optar al grado de:

Magister en Estadística Aplicada

Autor

Ingeniero Agrónomo

Alejandro Vargas Martínez

Diciembre, 2018

Comisión Asesora de tesis

Director

Dr. Cristóbal Fresno Rodríguez

Codirector

Dr. Julio Di Rienzo

Tribunal Evaluador

Dra. Ana Georgina Flesia
Facultad de Matemática, Astronomía, Física y Computación UNC

Dra. Cecilia Inés Bruno
Facultad de Ciencias Agropecuarias UNC

Mgter. María Laura Zingaretti
Centre for Research in Agricultural Genomics Universit at Autónoma de Barcelona
y Universidad Nacional de Villa María

Esta Tesis fue aprobada en la Escuela de graduados de Ciencias Económicas de

la Universidad Nacional de Córdoba para la obtención del grado académico de

Magíster en Estadística Aplicada el 20 de diciembre del 2018.

Dedicatoria

A Dios, a mi esposa, mis padres y mi hermano,

por todo el esfuerzo y apoyo brindado durante este proceso

permitiéndome alcanzar una meta más permitiéndome

superarme profesional y personalmente

Agradecimientos

A mi tutor Dr. Cristóbal Fresno Rodríguez y co-director Dr. Julio Di Rienzo por todo el asesoramiento estadístico y profesional, tiempo dedicado, motivación y apoyo brindado a lo largo de la Maestría y de la elaboración de este documento.

A la Universidad Nacional, en especial a la Escuela de Ciencias Agrarias por confiar en mi persona y brindarme la oportunidad de realizar los estudios de posgrado.

Al Ing. Claudio Vargas Rojas por todo el apoyo brindado durante el tiempo de estudio y estadía en Argentina.

A todas las personas que de alguna u otra forma colaboraron aportando al desarrollo de este trabajo de graduación.

Resumen en español

Distintos autores han expresado la importancia de conocer el efecto de la expresión génica de tejido normal en una muestra de cáncer de mama cuando se es analizada usando distintos clasificadores moleculares, debido a que se considera que está causando un grado de error al asignar una etiqueta a una muestra y, por lo tanto, la posibilidad de brindar una terapia inadecuada. El objetivo principal de este estudio fue el de evaluar el efecto de la expresión génica del tejido normal en magnitud y dirección sobre el diagnóstico de clases tumorales utilizando el clasificador molecular PAM50 como referencia. Para medir el efecto, se desarrolló una metodología estadística que estima el valor de proporción tumoral utilizando dos matrices de expresión génica de entrenamiento provenientes de muestras de pacientes con tumor y muestras de pacientes normales. La metodología propuesta fue evaluada utilizando las expresiones génicas de las muestras de pacientes sanos y las expresiones génicas de las muestras de los pacientes clasificados como “Normal-Like” de la base de entrenamiento. La función permitió estimar el valor de proporción de expresión tumoral presente en la muestra para luego aplicar la corrección de la expresión génica y generar una reclasificación utilizando PAM50. La cantidad de proporción de tejido normal presente en las muestras de cáncer de mama para cada una de las muestras de las cinco bases de datos públicas tuvo un impacto importante en la reasignación de la etiqueta luego de la corrección. Hubo cambios en muestras que pasaron de tener un diagnóstico favorable a uno menos favorable y viceversa. Sin embargo, factores como las expresiones génicas utilizadas para el entrenamiento del algoritmo que provienen de material biológico no puro y la misma heterogeneidad de la enfermedad, no permitieron tener una estimación más insesgada del vector de medias y de la matriz de varianza y covarianza de cada clase tumoral para una mejor estimación del valor de proporción tumoral mediante la metodología propuesta.

Robust classification of breast cancer samples

Key words: breast cancer, gene expression, molecular classifiers, normal tissue, deconvolution

Summary

Different authors have expressed the importance of knowing the effect of gene expression of normal tissue in a sample of breast cancer when it is analyzed in the different molecular classifiers, because it is causing a degree of error when assigning a label to a patient and therefore provide an inadequate therapy. The objective of this study was to evaluate the effect of normal tissue expression on magnitude and direction on the diagnosis of tumor classes using PAM50 molecular classifier as a reference. To measure the effect, a statistical methodology was developed. It estimates the value of tumor proportion in gene expression using two gene expression training databases with samples from patients with tumor and samples from healthy patients. The proposed methodology was evaluated using samples from healthy patients and samples from patients classified as "Normal-Like" from the training base. The function allowed us to estimate the proportion of tumor expression present in the sample and then apply the correction of gene expression using the estimated value. Finally, we perform a reclassification though PAM50. The proportion of normal tissue adjacent in sample of breast cancer obtained for each of the patients from the five public databases analyzed had an impact on the reassignment of the label after correction. There were changes in samples from having a favorable diagnosis at the beginning to a less favorable one and vice versa after reclassification. However, there were factors, such as the gene expressions used for training the algorithm like the use of non-pure biological material and the same heterogeneity of the disease that do not allow for a more unbiased estimate of the means vector and the variance and covariance matrix for each tumor class for a better estimate of the tumor proportion value using the proposed methodology.

INDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXO

1-Introducción

Cáncer de mama es el cáncer más común en mujeres en el mundo y cada vez son más las mujeres que están siendo diagnosticadas en comparación con otros tipos de cáncer. Los métodos clínicos y patológicos tradicionales hasta el día de hoy han sido la manera tradicional de asignar el subtipo, el grado en el que se encuentra el cáncer y la probabilidad de recurrencia de este. Estos métodos se basan en descriptores estándares y características físicas, como la edad del paciente, tamaño del tumor, características histológicas (grado del tumor) y el número de ganglios linfáticos axilares afectados (American Society Cancer 2017).

El cáncer de mama es una enfermedad extremadamente diversa y compleja. El tejido tumoral que la conforma es un tejido muy heterogéneo a nivel intertumoral que varía de paciente en paciente e intratumoral debido a la presencia de poblaciones de células distintas en un tumor individual además de subtipos de cáncer de mama específicos asociados a diferentes pronósticos. Dicha heterogeneidad se ha venido observando desde tiempo atrás y ha sido la base para la diferenciación de los subtipos (Polyak 2011).

Grupos de investigadores han desarrollado clasificadores moleculares en el área de la medicina, específicamente en el área de la oncología (Sotiriou et al, (2006)), como una alternativa a los métodos de diagnóstico tradicionales, con un alto potencial para proporcionar una mejor información sobre el pronóstico y la respuesta al tratamiento en pacientes con cáncer de mama, junto con el historial clínico de la persona (Bertucci, Birnbaum, 2008).

Un clasificador molecular requiere de una muestra biológica con los niveles de expresión de miles de genes en humanos para poder asignar bajo una regla de

asignación, una clase sobre el total de clases previamente detectadas en la enfermedad de estudio. Específicamente en el área de cáncer de mama se han desarrollado distintos clasificadores moleculares y se encuentran a nivel comercial. PAM50 es un clasificador molecular con fines de investigación y de libre acceso que se basa en la comparación entre el perfil genético de los pacientes (PGP) de 50 genes expresados y cinco tipos de perfiles genéticos intrínsecos (IGP) que representan subtipos de cáncer de mama (Basal, Her2-enriched, Luminal A, Luminal B y Normal). Para ello, asigna el subtipo que maximiza la correlación de Spearman entre el subtipo y el perfil de la muestra incógnita, es decir, muestra del paciente.

Dicho clasificador inicialmente hace un filtrado de información no requerida por el algoritmo de clasificación sobre las sondas que mapean los genes (aquellas sondas que no tuvieron suficiente señal de expresión), seguidamente hace una reducción a 50 genes únicos para luego asignar una etiqueta a la muestra al centroide más cercano (Perou et al. 2000, Perou et al. 2010). Estos clasificadores analizan la expresión génica total obtenida a través de la tecnología de los microarreglos. Estudios han demostrado que la variabilidad que existe a nivel clínico como a nivel molecular tienen un efecto sobre la clasificación del tumor en unas de las cinco (5) categorías descriptas a nivel molecular por Perou et al. (2000, 2010). Tales categorías son la Basal, Her2-enriched, Luminal A, Luminal B y Normal-Like cuando se analiza la expresión génica de cáncer de mama basada en microarreglos (Haibe-Kains et al. 2012).

Troester et al. 2009 citado por Elloumi et al. (2011), hacen referencia a que la variabilidad en este tipo de expresiones génicas a nivel molecular está mayormente asociado al tejido normal adherido al tejido canceroso de la biopsia. A nivel morfológico es indistinguible, pero a nivel molecular son distintos, debido a que

ambos tejidos tienen patrones de expresión molecular diferentes, por lo que podría ser una fuente que aumente el sesgo de los clasificadores moleculares al asignar una etiqueta a un paciente.

A nivel de laboratorios clínicos existen métodos que eliminan mediante la tecnología láser la mayor cantidad de tejido normal reconocible adherido a la biopsia. Sin embargo, esta práctica láser tiene un costo muy elevado. Una modelo simple de mezcla de la expresión génica del tumor y del tejido normal de una biopsia es la siguiente:

$$log_2\left(\frac{R}{G}\right)_{mezcla,i} = (\pi)\, log_2\left(\frac{R}{G}\right)_{tumor,i} + (1-\pi)\, log_2\left(\frac{R}{G}\right)_{normal,i}$$

donde:

- π ϵ [0,1] y representa la cantidad relativa de señal del tumor.
- (1- π) es la cantidad relativa de señal del tejido normal.
- R (Rojo) y G (Verde) son los valores de expresión del i-ésimo gen en los respectivos fluoróforos.
- $log_2\left(\frac{R}{G}\right)$ es el valor de expresión correspondiente a la mezcla, tumor o tejido normal del i-ésimo gen.

La presencia de tejido normal en el análisis de microarreglos se puede considerar como una fuente de error, la cual podría tener efectos sobre la expresión génica diferencial de los subtipos tumorales. Esta perturbación podría estar afectando los resultados finales de la clasificación, categorizando una muestra en un subtipo cuando en realidad pertenece a otro, y por ende el tratamiento no va a ser el correcto. Sin embargo, la magnitud y la dirección del efecto del tejido normal en clasificadores no ha sido probada aún (Elloumi et al. 2011).

El problema consiste en estimar, para una muestra de expresiones génicas obtenidas de un microarreglo de cRNA, el parámetro de la mezcla de dos distribuciones (tejido normal y tejido tumoral). Luego, corregir las expresiones génicas originales para obtener el valor de expresión propia del tumor. Con esto se

espera una reclasificación más precisa de los subtipos tumorales para el diagnóstico y un tratamiento basado en un diagnóstico más certero.

Desde un aspecto más general se pretende dar un mejor soporte a los clasificadores moleculares existentes en la investigación del cáncer de mama. Para ello se desarrolla una metodología que permita estimar el valor del parámetro de mezcla utilizando expresiones génicas de muestras de cáncer de mama y de pacientes sanos para entrenar al algoritmo.

Objetivo general

Estimar, a nivel de la expresión génica, la contaminación de tejido normal (proporción) de una biopsia de tejido mamario tumoral, para corregir el sesgo de los clasificadores moleculares en el diagnóstico de las clases tumorales.

Objetivos específicos

I. Desarrollar metodologías para estimar, en la expresión génica, la proporción de señal proveniente de tejido normal, mediante técnicas estadísticas que estudian mezclas de distribuciones.
II. Aplicar la metodología de corrección de la expresión génica, para evaluar su efectividad en muestras de tejido tumoral clasificadas por anatomía patológica.
III. Evaluación de la metodología sobre las expresiones génicas de los datos reales y de las expresiones génicas de los datos generados.

2-Revisión de literatura

2.1. - Heterogeneidad de tejido en cáncer de mama

El cáncer de mama es una enfermedad heterogénea a nivel de composición celular, información clínica y subtipos con distintos patrones de expresión génica que están asociados con el resultado final (Haibe-Kains et al. 2012). La muestra de tejido mamario con cáncer es un tejido con una alta heterogeneidad celular, esto debido a dos situaciones: La primera se debe a la heterogeneidad intratumoral que denota la coexistencia de subpoblaciones de células cancerosas que difieren en su genética, características fenotípicas o en el comportamiento dentro de un determinado tumor primario y entre un tumor primario dado y su metástasis. La segunda situación se genera al momento de la extracción de la muestra al paciente para luego ser analizada. Como se observa en la Figura 1, durante el procedimiento, al introducir la aguja, se corta tejido que sirve como muestra, al no ser tan precisa esta metodología, se extrae tejido mamario normal (sano o no tumoral) adherido al tejido mamario canceroso, generándose un factor de contaminación.

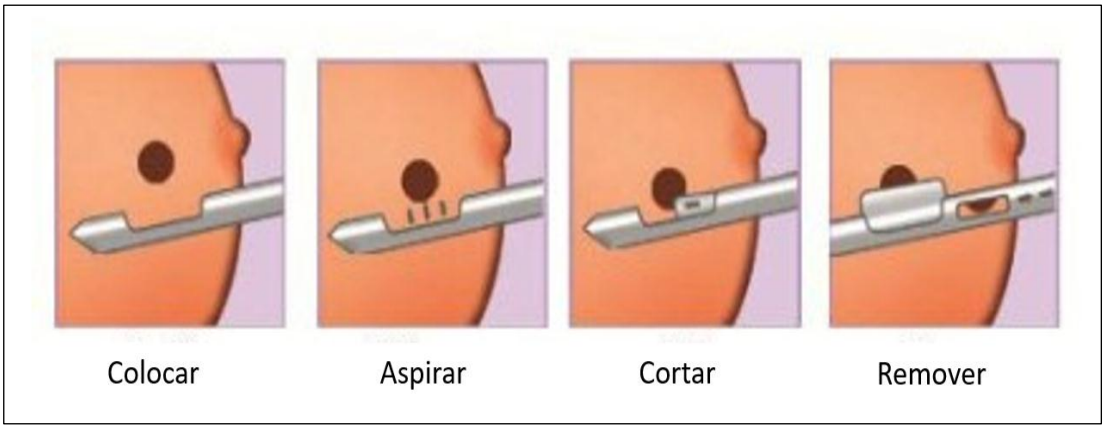

Figura 1. Esquema general del procedimiento quirúrgico para tomar una muestra de tejido mamario canceroso donde se introduce una aguja que permite extraer muestra de masa tumoral y removerla fuera del paciente para su posterior análisis histológico.

Una muestra biológica de un paciente que presenta una enfermedad autoinmune o condición patológica donde el sistema inmunitario es el encargado de atacar y destruir los propios órganos, típicamente contiene varios subconjuntos diferentes de células inmunitarias, y el proceso de diferenciar las expresiones génicas obtenidas a partir de tecnologías que miden las expresiones de miles de genes permite cuantificar sus proporciones relativas. Esencialmente, la expresión de cada gen en la muestra se modela como una combinación lineal de la expresión de ese gen en cada una de las celdas que comprenden esa muestra (Abbas et al. 2009). Esta diversidad ha generado un reto en la creación de clasificadores de tumores que sean clínicamente útiles tanto en la predicción como en el pronóstico.

2.2.- Microarreglos

El estudio de perfiles de expresión génica de células y de tejidos se ha convertido en una herramienta importante en la medicina. Los microarreglos de ADN son una de las opciones, entre las tecnologías disponibles, para la secuenciación del genoma que genera gran volumen de datos (Miranda y Bringas 2008). Las tareas de colección, manejo y análisis de estos datos de expresión de genes se han incrementado notablemente y han dado lugar a grandes sistemas de información con estructuras adecuadas para estos propósitos. Los microarreglos de ADN también conocidos como microarrays, biochips, microhileras, micromatrices, entre otras denominaciones, es una bio tecnología capaz de medir el nivel de expresión de miles de genes de un genoma simultáneamente en una muestra particular, estimando el número de ARN mensajero de cada gen con una especificidad relativamente alta comparado con otras tecnologías. Estas secuencias deben estar previamente caracterizadas y presentan el arreglo espacial de una matriz (filas y columnas). De esta manera cuando el “arreglo” es leído, se cuantificará la intensidad

de la fluorescencia de cada fluorocromo en un escáner especial y se apreciarán diferentes colores: verde cuando hay una sobre-expresión, rojo para supresión y amarillo cuando no hay cambios en la expresión entre la muestra normal y la muestra problema.

Un microarreglo de ácidos nucleicos es una plataforma sólida que puede ser de nylon, vidrio o plástico donde se encuentran inmovilizados un grupo limitado y finito de genes o fragmentos de genes (en promedio 50 bases de longitud) que pueden ir de cientos hasta miles, es decir, desde 100 elementos/cm^2 hasta 25,000 elementos/ cm^2 generando así los microarreglos de baja o alta densidad (Salcedo et al. 2003). Existen tecnologías para la construcción del diseño o arreglo, las cuales dependen de la macromolécula y logran en un tiempo corto, gran cantidad de información. Dichas tecnologías conocidas como bio chip y tiene su inicio hace más de una década en la compañía Affymax, que actualmente se llama Affymetrix. Actualmente, el mercado comercial cuenta con una gran gama de biochips para casos específicos, tales como Oncochip, Genechip, linfochip, nanochip, cardiochips, hepatochips, estafilochips, entre otros.

Se puede usar el perfil de expresión de alto rendimiento para comparar el nivel de transcripción de genes en condiciones clínicas con el fin de: 1) identificar biomarcadores de diagnóstico o pronóstico; 2) clasificar las enfermedades (Ej., Tumores con pronóstico diferente) que son indistinguibles por examen microscópico); 3) monitorear la respuesta a la terapia y 4) comprender los mecanismos implicados en la génesis de los procesos de enfermedad.

El desarrollo de métodos estadísticos para el análisis de conjuntos de datos con un elevado número de variables y un limitado número de mediciones ha ganado

importancia con el desarrollo de la tecnología de los microarreglos. Una de las principales características de los microarreglos es el gran volumen de datos que se generan, de ahí la importancia en saber y poder interpretarlos, utilizando poderosas técnicas estadísticas y bioinformáticas ofreciendo la capacidad de comparar y relacionar la información genética con una finalidad deductiva, brindando respuestas que van implícitas y que no parecen obvias a la vista de los resultados de los experimentos. En la mayoría de los estudios, el principal método estadístico para el análisis de las lecturas de las expresiones génicas generadas por el microarreglo para lograr una clasificación de un paciente a un subtipo, han sido los métodos de agrupamiento jerárquicos (Ghosh y Chibbaiyan 2002). Estos predictores multivariados se les denomina clasificadores moleculares y han mostrado tener una alta sensibilidad para identificar pacientes con alto riesgo de mortalidad y de recurrencia, así como pacientes con bajo riesgo que los métodos tradicionales que utilizan información clínica y patológica (van de Vijver et al. 2002).

El cuadro 1 muestra los métodos estadísticos para el análisis de los datos que puedan dar la mejor respuesta dependiendo de los objetivos del experimento orientados a la identificación de genes de comportamiento diferenciado entre las clases definidas o identificar genes con comportamientos similares sin que se conozca la clase a que pertenecen.

Cuadro 1. Resumen de los métodos estadísticos utilizados para la comparación, predicción o identificación de clases tumorales

Objetivos	Métodos estadísticos
Comparación de clases de tumores	t-test, F-test, Wilcoxon, Kruskal Wallis, SAM
Predicción de clases de tumores	kNN, DLDA, Naive, Bayes, QDA, LDA, LOCLDA, SVM
Identificar o clasificar de clases de tumores	k-means, SOM, HCL, SOTA

Fuente: (Miranda y Bringas 2008).

Actualmente en forma creciente se han publicado gran cantidad de artículos en los cuales se aplican los microarreglos en el área de la oncología (Salcedo et al. 2003). Los microarreglos han permitido dar un diagnóstico de cáncer evitando que los pacientes se les brinde tratamientos o terapias del tumor primario para evitar metástasis, emitiendo diagnósticos en diversas fases de la enfermedad para poder dar alternativas de tratamiento utilizando la clasificación de tumores humanos.

2.3 - Clasificadores moleculares

El estudio simultaneo de expresiones de una gran cantidad de genes obtenidas por el desarrollo de tecnologías de microarreglos de ADN se están utilizando ampliamente para el diagnóstico molecular de tumores de cáncer de mama, ya que proporcionan información rápida y reproducible sobre el nivel de expresión de un número elevado de genes que forman patrones , que luego se correlacionan con uno de los cinco subtipos descriptos a nivel molecular por Perou et al. (2000), para la clasificación del tumor (Cigudosa 2004). Estos cinco subtipos intrínsecos basados en perfiles de expresión génica inicialmente fueron categorizados en tres tipos Luminal, HER2 sobreexpresión y tumor fenotípico triple negativo. Investigadores definieron más subtipos dentro de los principales como Basal, Her2-sobreexpresado, Luminal A, Luminal B y Normal, entre ellos difieren en características clínicas, niveles de expresión génica, respuesta a los tratamientos y en el pronóstico. Los subtipos Luminal A y Luminal B son los que mejores pronósticos tienen en el sentido de identificar una célula tumoral cuando lo es, Her2 sobreexpresado es intermedio, mientas que Basal es el de peor pronóstico para un paciente.

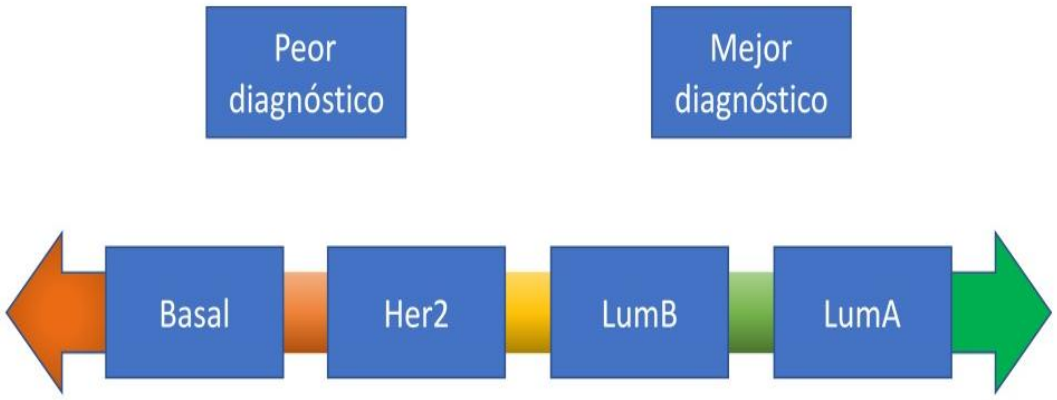

Figura 2. Clasificación del diagnóstico basado en los subtipos tumorales, donde se considera mejor diagnóstico cuando se detecta una célula tumoral (verdaderos) y peor diagnóstico cuando no detecta una célula tumoral (falsos negativos).
Fuente: (Dai et al. 2015)

El interés por identificar subtipos de cáncer de mama a nivel molecular radica en mejorar la asignación de una muestra a un subtipo tumoral, brindar un mejor pronóstico y por ende un tratamiento adecuado, evitando tratamientos innecesarios tales como quimioterapia o medicamentos muy tóxicos para la salud humana en aquellos pacientes que no lo requieran. Esto ha llevado a investigadores a la elaboración de predictores multivariados que evalúan la relación de una muestra de tejido mamario o de una biopsia con la expresión génica de grupos de genes, que son analizadas por diferentes tecnologías de microarreglos (microarreglos de cDNA y microarreglos de oligonucleótidos) asociados al cáncer de mama.

A nivel comercial existen clasificadores moleculares tales como Oncotype DX, MapQuant Dx y su versión simplificada, MammaPrint, Veridex 76gene, Theros Prosigna, Breast Cancer Index, EndoPredict y Immunohistochemistry 4 (IHC4), los cuales brindan una estimación individual por paciente del riesgo de la recurrencia de la enfermedad e información pronóstica independiente de la proporcionada por el estándar clínico y de las características patológicas. Diferentes tecnologías de microarreglos, métodos estadísticos, pacientes con características diferentes y el

tamaño de la muestra son algunas de las diferencias entre estos clasificadores moleculares que existen a nivel comercial (Cuadro 2), sin embargo, existen discrepancias en la asignación de los pacientes a subtipos de tumores (Cigudosa 2004).

Cuadro 2. Clasificadores moleculares comerciales para la asignación de subtipos tumorales y grado de recurrencia asociados a cáncer de mama disponibles

Clasificador Molecular	Proveedor	Técnica	Tipo de ensayo	Disponibilidad
PAM50	PROSIGNA	Microarreglo ADN	50 genes	Versión Académica
Mamma Print	Agendia	Microarreglo ADN	70 genes	Europa y Estados Unidos
Veridex 76 gene	Actualmente no disponible comercialmente	Microarreglo ADN	76 genes	Actualmente no disponible comercialmente
MapQuant	Ipsogen	Microarreglo ADN	97 genes	Europa
MapQuant Dx simplified	Ipsogen	qRT-PCR	8 genes	Europa
Oncotype DX	Genomic Health	qRT-PCR	21 genes Grado recurrencia	Europa y Estados Unidos
Theros	Biotheranostic	qRT-PCR	Ratio de 2 genes HOXB13 a IL17R (H/l) /índice grado molecular	Estados Unidos

Fuente: Elaboración propia

Estos predictores multivariados se les denomina clasificadores moleculares y han mostrado tener una alta sensibilidad para identificar pacientes con alto riesgo de mortalidad y de recurrencia, así como pacientes con bajo riesgo que los métodos tradicionales que utilizan información clínica y patológica (van de Vijver et al. 2002).

2.4 - Deconvolución de la expresión génica de un gen

La deconvolución es un enfoque *in silico* que permite analizar la expresión génica en muestras de tejidos heterogéneos. Las muestras de tejido generalmente contienen más de un tipo de célula. Esto significa que las mediciones hechas en una muestra medirán una combinación de señales de tipos de células ponderada por su abundancia (Järvstråt 2017). Las biopsias provenientes de cáncer de mama frecuentemente consisten en dos componentes distintos, epitelio glandular y tejido estromático que la rodea (Ahn et al. 2013). Ambos tejidos presentan patrones diferentes en su expresión génica (Elloumi et al. 2011). Las técnicas analíticas tradicionales que ignoran la presencia de la heterogeneidad de tejido presente en la muestra podrían sufrir de una inadecuada transcripción del perfil génico y estarían perdiendo información de genes que estén relacionados con la descripción del tipo de cáncer (Ahn et al. 2013). Esta relación tumor-estroma es una fuente de información que no ha sido estudiada y en el fondo contiene información de las expresiones de mezcla de la interacción de ambos tejidos que no pueden ser observadas directamente en el perfil génico global obtenido de la muestra (Wang et al. 2015). Para remover la presencia de tejido normal adherido al tejido tumoral, existen diferentes técnicas utilizadas actualmente por investigadores, entre las cuales están la separación física de muestras en subpoblaciones, incluida Citometría de flujo (FACS siglas en inglés), clasificación celular basada en cuentas magnéticas (MACS), microdisección usando una pipeta capilar, o microscopía de captura láser (Järvstråt 2017). La técnica que actualmente se está utilizando con mayor frecuencia es la técnica de láser, la cual genera micro disecciones que remueven físicamente los diferentes tipos de tejidos distintos al tumoral. Sin embargo, esta técnica es muy costosa dado el tiempo que se requiere y del equipamiento que se

necesita para llevarla a cabo (Ahn et al. 2013).Como un enfoque complementario a los métodos que físicamente separan las subpoblaciones de células, el análisis estadístico de los datos, generado a partir de muestras de tejidos complejas, permite separar tejido sano de tejido enfermo, es decir, teniendo en cuenta la heterogeneidad. El modelo más básico usa una combinación lineal de abundancias de las células junto con el patrón típico de expresión génica (Ahn et al. 2013). Se han propuesto varios enfoques estadísticos para deconvolucionar los perfiles de expresión génica obtenidos a partir de muestras heterogéneas de tejido en sub-perfiles específicos del tipo celular. La mayoría de los métodos se basan en un marco propuesto inicialmente por Venet et al. (2001), que incorpora el supuesto de linealidad de que la expresión de cada gen en una mezcla de tipos de células es un promedio ponderado de los valores de expresión que existirían para poblaciones puras de esos tipos de células. Los pesos (ponderadores) están determinados por la composición proporcional de los tipos de células en la mezcla y, por lo tanto, son los mismos para cada gen, pero difieren entre las mezclas de muestra. Dentro de los métodos más utilizados para deconvolucionar la heterogeneidad de expresión génica de una muestra de tejido canceroso analizado mediante microarreglos, los sistemas de ecuaciones lineales son los que están permitiendo poder cuantificar las proporciones de células en un tejido complejo.

Sea N_{ip} y T_{ip} las expresiones génicas del gen $p, con\ p = 1, \dots, P$ de tejido puro normal (N) y tejido puro tumoral (T), respectivamente, correspondiente a la muestra $i, con\ i = 1, \dots, N$. No se cuenta con la lectura de las expresiones de tumor puro T_{ip} para cada paciente, sino que se tienen las Y_{ip}, que corresponden a la expresión génica de la muestra de tumor (biopsia) i para el gen p correspondiente a un paciente. La ecuación lineal se representa de la siguiente forma:

$$Y_{ip} = \pi_{ip} T_{ip} + (1 - \pi_{ip}) N_{ip}$$

Donde π_{ip} representa la proporción de tejido tumoral y se asume que sea el mismo para todos los genes. Las expresiones de genes obtenidas a partir de microarreglos ya se encuentran normalizadas mediante la transformación de log_2 , por lo tanto, dichas expresiones siguen una distribución normal $N_{ip} \sim LN(\mu_{Np}, \Sigma_{Ng})$ y $T_{ip} \sim LN(\mu_{Tp}, \Sigma_{Tg})$ donde LN representa una distribución log_2Normal (Carvalho et al. 2007).

2.5 – Distribución Normal p-multivariante

La distribución normal univariada tiene como función de densidad:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}$$

y se escribe $X \sim N(\mu, \sigma^2)$, para indicar que x tiene distribución normal con media μ y varianza σ^2. Un vector aleatorio X es una colección de variables aleatorias y X tiene distribución Normal multivariada, denotado por $X \sim N(\mu, \Sigma)$, si su densidad es:

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu)^t \Sigma^{-1} (x - \mu)\right\}$$

donde μ es un vector de longitud p y Σ es un matriz de tamaño $p \times p$, simétrica y definida positiva. Luego $\mathbb{E}(X) = \mu$ y $\mathbb{V}(X) = \Sigma$.

Las propiedades principales son:

1- La distribución normal p -dimensional es simétrica entorno de μ.

2- La distribución normal p-dimensional tiene un único máximo en μ.

3- Si X es un vector aleatorio p-dimensional distribuido normalmente, la media del vector aleatorio normal es μ y su matriz de varianzas y covarianzas es Σ.

4- Si p variables aleatorias tienen distribución conjunta normal y están incorreladas son independientes.

5- Cualquier vector x normal p-dimensional con matriz Σ no singular puede convertirse mediante una transformación lineal en un vector z normal p-dimensional con vector de medias 0 y matriz de varianzas y covarianzas igual a la identidad (I). Llamaremos normal p-dimensional estándar a la densidad de z, que vendrá dada por:

$$f(z) = \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2} z^t z\right\} = \prod_{i=1}^{p} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} z_i^2\right\}$$

6- Las distribuciones marginales son normales.

7- Cualquier subconjunto de $h < p$ es normal $h - \text{dimensional}$

8- Si y es $(k\ x\ 1)$, $k \leq p$, el vector y $y = Ax$, donde A, es una matriz $(k\ x\ 1)$, es normal k-dimensional.

9- Al cortar con hiperplanos paralelos al definido por las p variables que forman la variable vectorial, x, se obtienen las curvas de nivel, cuya ecuación es:

$$(x - \mu)' V^{-1}(x - \mu) = cte.$$

Las curvas de nivel son, por tanto, elipsoides, y definen una medida de las distancias de un punto al centro de la distribución. Esta medida se denomina *distancia de Mahalanobis* y se representa por:

$$D^2 = (x - \mu)' V^{-1}(x - \mu)$$

10- La distancia de Mahalanobis se distribuye como una χ^2 con p grados de libertad.

2.5.1-<u>Estimación robusta del vector de medias y varianza</u>

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población normal de media μ y covarianza Σ, entonces los estimadores de máxima verosimilitud para μ y Σ serán:

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \;\; y \;\; \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$$

El método más usado para estimar un parámetro de tendencia central es el intervalo de confianza de la media, ya que el estimador mayormente utilizado es el la media muestral o promedio, como se mencionó anteriormente, es el estimador máximo verosímil y cuenta con la propiedad de ser insesgado. Sin embargo, este estimador se ve influenciado por valores extremos o cuando las muestras son de tamaño pequeño, generando errores en la estimación. Bajo circunstancias de tamaño muestral pequeño y presencia de valores extremos, pueden utilizarse métodos robustos para el cálculo de estadísticos de centralización o localización. Un buen estimador robusto es la mediana, ya que este estimador no se ve afectado por datos atípicos y para muestras pequeñas (Cuadro 3).

Cuadro 3. Métodos robustos para el cálculo de estadísticos de posición

Estimador de tendencia central	Estrategia
Media α-winsorizada muestral	Se sustituye un porcentaje de α, 20% generalmente, de valores extremos a cada lado de la muestra por el valor más próximo no sustituido.
Media α-recortada muestral	Se eliminan las k observaciones extremas de cada lado, en lugar de winsorizadas calculando la media aritmética de las observaciones restantes.
Mediana muestral	Divide la distribución en dos partes con el mismo número de elementos
Estimador de Huber	Se encuentra dentro de los denominados M-estimadores, que generalizan al estimador de máxima verosimilitud con buenas propiedades de robustez y eficiencia. En este caso se descartan las observaciones que sean mayores o menores a una constante

Fuente: Ramalle-Gómara y Andrés de Llano 2003)

2.6 - Estimación por máxima verosimilitud

El método de máxima verosimilitud es una técnica para estimar los valores de θ dada una muestra finita de datos, que escoge como estimadores de los parámetros aquellos valores que hacen máxima la probabilidad de que el modelo a estimar genere la muestra observada (Peña 2002).

Dado una muestra aleatoria simple de *n* elementos de una variable aleatoria p-dimensional X con función de densidad $f(\mathrm{X}|\theta)$, donde $\theta = (\theta_1, \dots, \theta_n)'$ es un vector de parámetros que tiene dimensión $r < np$. Si llamamos $\mathbb{X} = (x_1, \dots, x_n)'$ a los datos muestrales donde $x_i = (x_{i1}, \dots, x_{ip})'$ representa un individuo particular, entonces la función de densidad conjunta de la muestra será:

$$f(\mathrm{X}|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

Debido a la independencia de las observaciones. Además, si el parámetro θ es conocido, la función $f(\mathrm{X}|\theta)$ determina la probabilidad de aparición de la muestra.

La función de verosimilitud se define como la función de densidad conjunta de $X_1, X_2, \dots, X_n$ evaluada en $x_1, \dots, x_n$ y está dada por:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|\underline{X}) = f_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n|\theta)$$

La notación $L(\theta)$ indica que L es una función de θ y no de $(x_1, \dots, x_n)$, donde θ puede ser un escalar o un vector ($\theta = ((\theta_1, \dots, \theta_k))$.

Para cada muestra en particular $(x_1, \dots, x_n)$,la estimación de máxima verosimilitud de θ es el valor $\hat{\theta}_{MV}$ que maximiza la verosimilitud, es decir:

$$\mathcal{L}(\widehat{\theta}_{MV}|\underline{X}) = max_{\theta} l(\theta|\underline{X})$$

El estimador de máxima verosimilitud, $\hat{\theta}_{MV}(X_1, X_2, \ldots, X_n)$, es el valor de θ que indica la probabilidad de aparición de los valores muestrales observados y que se obtienen al calcular el valor máximo de la función $\mathcal{L}(\theta)$. Se asume que $\mathcal{L}(\theta)$, es diferenciable, entonces:

$$\frac{\partial l(\theta)}{\partial \theta_1} = 0$$
$$\vdots$$
$$\frac{\partial l(\theta)}{\partial \theta_r} = 0$$

Esto resulta en un número de ecuaciones con un número igual de variables, las cuales pueden resolverse simultáneamente y comprobar que realmente es un máximo, evaluando tal y como se expresa en esta expresión:

$$\frac{\partial^2}{\partial \theta_j^2} l(\theta) | \theta_j = \hat{\theta}_j \ < 0$$

En la práctica suele ser más práctico obtener el máximo del logaritmo de la función de verosimilitud y se define como función soporte, ambas funciones tienen el mismo máximo. Sea $l(\theta)$ la función de verosimilitud de la muestra aleatoria $X_1, X_2, \ldots, X_n$ de una población con $(\theta = (\theta_1, \ldots, \theta_k)$, la función de soporte tiene la siguiente expresión:

$$\mathcal{L}(\theta) = \ \mathrm{In} l(\theta) = \mathrm{In}\left[\prod_{i=1}^{n} f(x_i|\theta)\right] = \sum_{i=1}^{n} \mathrm{In}[f(x_i|\theta)]$$

Propiedades del estimador máximo verosímil

1-Invarianza: Si $\hat{\theta}_{MV}$ es el estimador máximo verosímil de θ, entonces $h(\hat{\theta}_{MV})$ es el estimador máximo verosímil de $h(\theta)$.

2-Consistencia: Bajo ciertas condiciones generales, $\hat{\theta}_{MV}$ es un estimador consistente de θ.

3-Insesgadez asintótica: Se verifica que $\lim_{n\to\infty} E[\hat{\theta}_{nMV}] = \theta$

4-Normalidad asintótica: Bajo ciertas condiciones generales

$$\sqrt{n}(\hat{\theta}_{MV} - \theta) \underset{\sim}{A} N(0, \sqrt{i(\theta)^{-1}}$$

Donde

$$i(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]$$

Es la cantidad de información de Fisher correspondiente a una observación.

Cuando se tiene n observaciones, se expresa de la siguiente forma:

$$\mathrm{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1, X_2, \dots, X_n;\theta)\right)^2\right] \overset{m.a.s}{=} n \cdot E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]$$

$$= n \cdot i(\theta)$$

Se tiene

$$\mathrm{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1, X_2, \dots, X_n;\theta)\right)^2\right] = -E\left[\left(\frac{\partial}{\partial\theta}\ln f(X_1, X_2, \dots, X_n;\theta)\right)^2\right]$$

La varianza asintótica de $\hat{\theta}_{MV}$ es:

$$Var[\hat{\theta}_{MV}] \underset{=}{A} \frac{1}{n \cdot i(\theta)} = \frac{1}{I(\theta)} = -\frac{1}{E\left[\frac{\partial}{\partial\theta}\ln f(X_1, X_2, \dots, X_n;\theta)\right]}$$

$$\approx -\frac{1}{\frac{\partial^2}{\partial\theta_j^2} l(\theta)|\theta \quad = \hat{\theta}_{MV}}$$

Para el caso de una normal multivariada se obtiene la estimación máximo verosímil de la siguiente manera.

Sean $X_1, X_2, \dots, X_n$ una muestra aleatoria de una población $X \sim N(\mu, \Sigma)$. La función de densidad conjunta de la muestra está dada por:

$$f_{X_1,X_2,\dots,X_n}(x_1, \dots, x_n|\mu, \Sigma) = \prod_{i=1}^{n} |\Sigma|^{-\frac{1}{2}}(2\pi)^{\frac{-p}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu)^t \Sigma^{-1}(x_i - \mu)\right\}$$

$$= |\Sigma|^{-\frac{n}{2}}(2\pi)^{\frac{-np}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^t \Sigma^{-1}(x_i - \mu)\right\}$$

y la función de verosimilitud para la muestra aleatoria de una población $X \sim N(\mu, \Sigma)$ está dada por:

$$L(\mu, \Sigma|\underline{X}) = f_{X_1,X_2,\dots,X_n}(x_1, \dots, x_n|\mu, \Sigma)$$

Para obtener el estimador máximo verosímil se utiliza la función de soporte, para eso se utiliza el logaritmo de la verosimilitud y tiene la siguiente expresión

$$L(\mu, \Sigma|\underline{X}) = -\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^t \Sigma^{-1}(y_i - \mu)$$

Los estimadores máximos verosímiles de μ y Σ resultan de la maximización de la log-verosimilitud y son respectivamente:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = \frac{1}{n} S$$

Donde

$$S = \sum_{i=1}^{n}(x_i - \hat{\mu})\,(x_i - \hat{\mu})'$$

3-Materiales y Métodos

3.1 - Datos reales

Los datos utilizados son dos microarreglos de genes llamados GEOBreastCancerData, la cual contiene información de expresión (medida por tecnología Affymetrix) de muestras de $\boldsymbol{n = 54675}$ sondas que mapean a un gen y $\boldsymbol{p = 78}$ muestras de mamas sanas de pacientes clasificados como sanos y $\boldsymbol{p = 641}$ muestras de biopsias de tejido de cáncer mamario de pacientes que presentan la enfermedad. Las expresiones fueron normalizadas y expresadas en logaritmo en base 2 utilizando el paquete de R llamado **ARSyN** (Nueda et al. 2012), el cual se encarga de remover ruido sistemático contenido en cada una de las muestras. Adicionalmente, la base de datos cuenta con las anotaciones e información clínica correspondiente.

Se utilizó la implementación de PAM50 de la librería R de **genefu** (Haibe-Kains et al. 2014) para obtener la etiqueta del subtipo tumoral de clasificación de cada muestra. Luego de haber utilizado PAM50 sobre las muestras de los pacientes que presentaron cáncer de la base de datos GEOBreastCancerData se obtuvo la clasificación según el subtipo de cáncer de mama al que pertenecía cada muestra (Cuadro 4).

Cuadro 4. Clasificación de las muestras de cáncer de mama de 641 pacientes en cinco subtipos de cáncer con el clasificador PAM50.

Subtipo de cáncer de mama	Cantidad de muestras
Basal	163
Her2-enriched	104
Luminal A	171
Luminal B	100
Normal-Like*	103

*Corresponde a una clase tumoral muy similar a tejido normal

Luego del filtrado de la información no requerida por PAM50 se generó una matriz con dimensiones de 126 sondas génicas (filas) y 641 muestras de pacientes (columnas), luego esta matriz se dividió en cinco matrices tomando en cuenta la asignación de las etiquetas a los pacientes. Estas matrices tienen dimensiones de la misma cantidad de sondas que mapean a los genes y la cantidad de columnas depende de la asignación de la cantidad de etiquetas asignadas a cada clase tumoral (cuadro 5). De la misma forma que se hizo con los pacientes con cáncer, se hizo con las muestras de los pacientes sanos, obteniendo como resultado una matriz de dimensión 126 sondas génicas (filas) x 78 pacientes (columnas).

De esta manera tenemos datos de muestras no apareados de expresiones génicas de pacientes con cáncer de mama y de muestras de expresiones génicas de tejido normal (el tejido sano categorizado como normal no correspondió al mismo paciente con cáncer de mama). Las cuales servirán para entrenar al algoritmo de estimación para obtener el valor de proporción de expresión tumoral en la muestra ($\boldsymbol{\pi}$).

3.1.2. – Alternativas de expresiones génicas de pacientes Normales

Para estimar el valor de proporción tumoral utilizando el algoritmo propuesto, como alternativa a las expresiones génicas de muestras de mamas sanas, se utilizaron como dos fuentes distintas, la primer correspondió a las muestras no apareadas de pacientes sanos de la base de entrenamiento y la segunda fueron las 103 muestras clasificadas como “Normal-Like” dado que este subtipo es un “proxy” a una expresión génica de una mama normal y tiene una mayor similaridad a las demás clases tumorales.

3.2.- Estimación de los parámetros iniciales de la mezcla requeridos por el algoritmo

Desde el punto de vista de las expresiones génicas, en una muestra de tejido tumoral mamario PAM50 utiliza 50 genes a los fines de asignar una etiqueta del subtipo tumoral a la muestra. Luego, el vector $\boldsymbol{Y}$ es de dimensión 50. No fue posible estimar los parámetros de la mezcla a partir de una única observación del vector multivariado. Sin embargo, cuando se evalúan los genes de la base de datos GEOBreastCancerData con el chip Affymetrix se obtienen lecturas de 126 sondas, que mapean a los 49 genes (La base de entrenamiento solamente tiene 49 genes de los 50 que utiliza PAM50). Esto quiere decir que hay genes que tienen lecturas de diferentes sondas, por lo que hay lecturas de genes repetidos. En particular hay 28 genes que están asociados a 105 sondas génicas y 21 a sondas génicas únicas. Así, hay 28 genes que tienen datos repetidos (Cuadro 5). Para ello haremos algunas suposiciones que implican una simplificación del problema.

a. Las densidades que conforman la mezcla son gaussianas (Carvalho et al. 2007).
b. Solo se consideran al mismo tiempo una mezcla de dos componentes: uno proveniente de tejido tumoral y otro proveniente de tejido normal (que se asume corresponde a la menor de las fracciones-contaminación).
c. El parámetro $\boldsymbol{\pi}$ es el mismo para todos los genes.
d. Se dispone de dos conjuntos de datos de entrenamiento que, por anatomía patológica, tiene confirmado el subtipo tumoral y el diagnóstico de tejido sano, además las expresiones génicas de los 49 de 50 genes que utiliza PAM50 de referencia.

Cuadro 5. Genes que comparten una misma sonda

Gen	Número de sondas	Gen	Número de sondas	Gen	Número de sondas
ACTR3B	2	CXXC5	3	MKI67	4
ANLN	2	EGFR	9	MMP11	5
BAG1	3	ERBB2	3	PGR	2
BCL2	4	ESR1	9	RRM2	2
BIRC5	4	FGFR4	4	SFRP1	4
BLVRA	5	FOXA1	2	SLC39A6	4
CCNB1	2	FOXC1	2	TMEM45B	2
CCNE1	2	KIF2C	2	TYMS	3
CDC6	2	MAPT	6		
CENPF	3	MDM2	10		

3.2.1-Estimación del vector de expresión génica media para cada subtipo tumoral

La estimación del vector de medias para la expresión génica de cada gen (49×1) para cada subtipo tumoral incluyendo el Normal-like, se realizó en dos etapas. En la primera etapa se utilizaron las 105 sondas génicas que mapean a 28 genes. Como primer paso se obtuvo la mediana como estimador robusto del subtipo tumoral de cada paciente, de esta manera se generó un vector de dimensión 28×1 correspondiente a la mediana de la expresión génica de los 28 genes. Luego, en la segunda etapa se trabajó con las sondas génicas únicas (21) que mapean a un gen en específico, se estimó la mediana para cada gen utilizando la información de cada paciente según el subtipo tumoral generando un vector de dimensión 21×1 correspondiente a la mediana de la expresión génica de los 21 genes, ya que, al ser sondas únicas, no se cuentan con repeticiones por gen. Por último, se unió ambos vectores para lograr obtener un vector de 49 genes con la mediana de la expresión génica para cada subtipo tumoral.

Dado que la expresión de un gen es la suma ponderada de la expresión génica del tejido tumoral y de la expresión génica normal, la estimación de la media para cada gen utilizada en el algoritmo fue la siguiente:

$$E(Y_{ig}) = \pi E(T_{ig}) + (1-\pi)E(N_{ig})$$

Se obtuvo un vector de medianas de las expresiones génicas de cada subtipo de dimensión 49×1 utilizado para la estimación del parámetro de mezcla (π) en cada uno de los subtipos tumorales.

3.2.2-Estimación de la matriz de varianza y covarianza de la expresión génica de cada subtipo tumoral

De la misma forma metodológica del punto 3.2.1 para la estimación del vector de medianas de las expresiones génicas, primero se trabajó con las matrices de las expresiones génicas de los subtipos tumorales que contienen las sondas génicas con lecturas repetidas. Primero se obtuvo la mediana de las lecturas repetidas como estimador robusto de la expresión media de cada gen por paciente y luego se unió con las matrices de las expresiones génicas tumorales de sondas únicas. Teniendo una matriz con las expresiones génicas de los 49 genes de cada subtipo tumoral, a cada matriz se le calculó la matriz de varianza y covarianza de dimensión 49×49 genes.

Dado que la expresión de un gen es la suma ponderada de dos expresiones génicas, la estimación de la matriz de varianza y covarianza para 49 genes requeridos por PAM50 fue la siguiente:

$$Var(Y) = K_{\mathrm{p}\times 2\mathrm{n}} \times \Sigma_{2nx2n} \times K_{2n\times p}^{T}$$

Donde $K_{\mathrm{p}\times 2\mathrm{n}}$ es una matriz de dimensión 49 filas x 98 columnas que contiene el parámetro de la mezcla que se quiere estimar (el valor de 98 corresponde a 49

valores en la diagonal del parámetro de proporción tumoral ($\boldsymbol{\pi}$) a estimar y otros 49 valores de ($\mathbf{1} - \boldsymbol{\pi}$) que corresponde a la proporción de tejido normal) y sigma ($\boldsymbol{\Sigma}$) es la matriz de varianza y covarianza de dimensión 98×98, en donde la diagonal contiene la varianza para cada gen y afuera de la diagonal la covarianza entre los pares de genes. La estructura de la matriz de varianza y covarianza utilizada para estimación del parámetro de la mezcla en el algoritmo propuesto tuvo la siguiente forma:

$$Var(Y) =$$

$$\begin{bmatrix} \pi & 0 & 0 & 1-\pi & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \pi & 0 & 0 & 1-\pi \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} & & & \\ & 0 & & \\ & & & \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \pi & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi \\ 1-\pi & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1-\pi \end{bmatrix}$$

3.2.3-<u>Estimación del parámetro de mezcla por el método de máxima verosimilitud</u>

Dadas las $\boldsymbol{y_{ig}}$, que corresponden a la expresión génica observada de la mezcla para cada una de las matrices de expresión génicas utilizadas según el subtipo tumoral, el vector de medianas de las expresiones génicas normales $\boldsymbol{n_g}$ del gen $\boldsymbol{g}, g = 1,..,n$, el vector de medianas de las expresiones génicas tumorales $\boldsymbol{t_g}$ del gen $\boldsymbol{g}, g = 1,..,n,$ las matrices de varianzas y covarianzas de las expresiones génicas tumorales y expresiones génicas normales, se estimó el valor de proporción tumoral presente en la mezcla ($\boldsymbol{\pi}$) para cada subtipo tumoral (Basal, Luminal A, Luminal B, Her2-enriched, Normal-Like) y se determinó cuál de ellos maximizó la verosimilitud de la expresión génica observada utilizando los genes que tienen en común la base de datos GEOBreastCancerData y la base de datos real.

Para la estimación del parámetro por máxima verosimilitud se utilizó la función **optim()** de la librería **stats** con del método Brent de R (R Core Team 2018). Luego, se analizó la distribución del parámetro de mezcla estimado a lo largo de la cohorte de pacientes. Para poder utilizar dicha función, se definió la función de soporte de la siguiente manera:

Sea $Y = (y_1, \dots, y_n)^{'}$ una muestra aleatoria simple donde $y_i \sim N_p(\mu, \Sigma)$. La función de verosimilitud es:

$$l(\mu, \Sigma|Y) = \prod_{i=1}^{n} |\Sigma|^{-1/2} (2\pi)^{\frac{-p}{2}} \exp\left\{-\frac{1}{2}(y_i - \mu)^t \Sigma^{-1} (y_i - \mu)\right\}$$

$$l(\mu, \Sigma) = |\Sigma|^{-n/2} (2\pi)^{\frac{-np}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^t \Sigma^{-1} (y_i - \mu)\right\}$$

y la función de soporte será:

$$L(\mu, \Sigma|Y) = \ln l(\mu, \Sigma)$$

$$L(\mu, \Sigma) = -\frac{1}{2} pn \ln(2\pi) - \frac{1}{2} n \ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^t \Sigma^{-1} (y_i - \mu)$$

Definida la función de soporte, se definió un vector con valores iniciales, con nueve valores de proporción tumoral iniciando en 0.1 hasta 0.9 aumentando en 0.1, por último, se definió el parámetro a ser estimado.

Esta metodología además de permitir estimar el parámetro de mezcla de cada subtipo tumoral en la muestra sirvió como un clasificador ya que, al seleccionar el subtipo tumoral con el que se obtuvo la mayor verosimilitud, indirectamente le asignó una etiqueta con el subtipo tumoral al paciente. Sin embargo, esta alternativa de clasificador no se utilizó para ver el verdadero cambio como una reclasificación.

La figura 3 tiene una secuencia de cómo se llega a obtener el valor de proporción tumoral en las expresiones génicas de la muestra del paciente estimado por el algoritmo para cada uno de los subtipos tumoral, utilizando los valores iniciales estimados en los puntos 3.2.1 y 3.2.2.

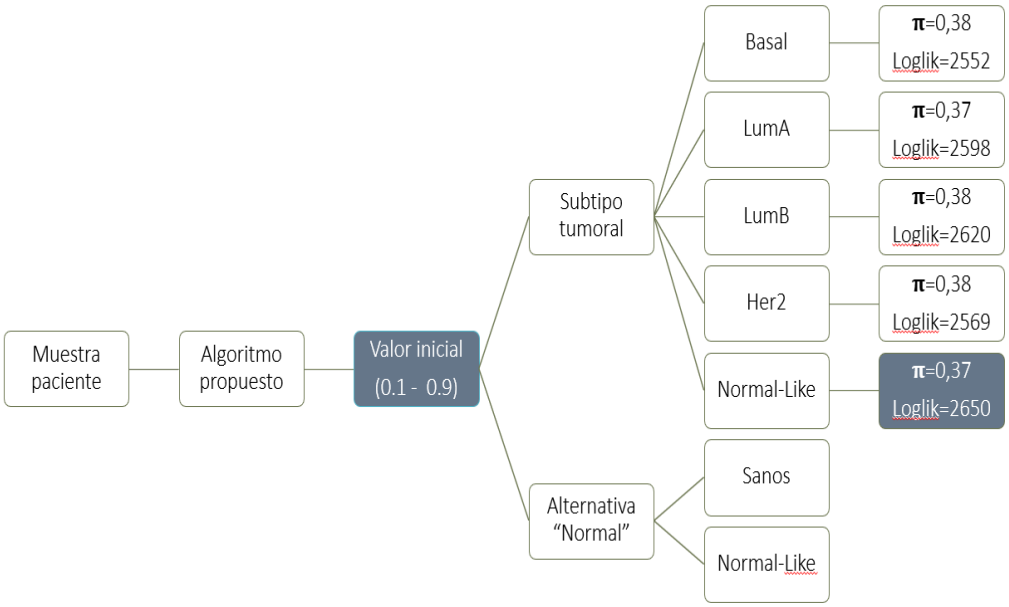

Figura 3. Secuencia de estimación del valor de proporción tumoral presente en la mezcla mediante el método de máxima verosimilitud.

Obtenidos los $\hat{\boldsymbol{\pi}}_{\boldsymbol{MV}}$ para cada muestra, se hizo la corrección con el valor $\hat{\boldsymbol{\pi}}_{\boldsymbol{MV}}$ estimado. Luego, se deconvolucionó las $\boldsymbol{y_{ig}}$ observadas para la muestra en una expresión génica corregida correspondiente al tejido tumoral.

$$t_{ig} = \{y_{ig} - (1 - \hat{\pi}_{MV})n_{ig}\}/\hat{\pi}_{MV}$$

Finalmente, con la matriz de expresiones génicas corregidas, se volvió a utilizar PAM50 para obtener una reclasificación de cada muestra y se estimó la magnitud y la dirección del cambio de clasificación de las muestras tumorales según la clasificación inicial obtenida por PAM50.

3.2.4-Evaluación del modelo de clasificación de subtipos tumorales

Inicialmente se evaluó el algoritmo propuesto que estima el valor de proporción tumoral en la mezcla para cada muestra sobre la base de entrenamiento GEOBreastCancerData. Luego, se evaluó el desempeño del algoritmo sobre cinco bases de datos reales reportadas por Haibe-Kains et al. (2012) (Cuadro 6) y se tuvo las siguientes consideraciones:

1. Se utilizaron tal como fueron reportadas por los autores, es decir, no se les realizó ningún pre-procesamiento para ser utilizadas.
2. Para los valores de las sondas que comparten un mismo gen (sondas repetidas) se les estimó la mediana de su expresión génica para resumir su valor y poder trabajar con un valor único.
3. Se hizo coincidir los genes de la base de entrenamiento con la base real y que a su vez los utilice PAM50, debido a que el chip del microarreglo no siempre fue el mismo por lo que los genes nunca fueron los mismos.
4. Cada base de datos se trabajó de forma independiente.

Cuadro 6. Bases de datos públicas de pacientes que presentan cáncer de mama.

Nombre de la base de datos	Cantidad de pacientes (#)	Fabricante del microarreglo	Genes de PAM50	Número de sondas	Fuente	Referencia
MAINZ	198	Affymetrix	44	22283	GEO: GSE11121	Schroeder et al. (2011a) y Schmidt et al. (2008)
TRANSBIG	200	Affymetrix	44	22283	GEO: GSE7390	Schroeder *et al*. (2011c) y Chin *et al*. (2006)
UNT	117	Affymetrix	50	44928	GEO: GSE2990	Schroeder *et al*. (2011d) y Sotiriou et al. (2006)
UPP	251	Affymetrix	50	44928	GEO: GSE3494	Schroeder et al. (2011e) y Miller et al. (2005)
VDX	334	Affymetrix	44	22283	GEO: GSE2034/ GSE5327	Schroeder et al. (2011f) y Minn et al. (2008)

Estimado el valor de proporción tumoral para cada muestra en cada una de las bases de datos reales, se procedió a analizar su distribución con el fin de ver como varió de paciente a paciente o si fue el mismo valor para todos los pacientes. Luego de estimar el valor de proporción tumoral en las expresiones observadas para cada muestra, se corrigió la expresión génica, para luego hacer una reclasificación mediante PAM50 y finalmente se comparó de forma sistemática como influyó la clasificación dado el grado de contaminación en los diferentes subtipos de PAM50.

PAM50 asigna un subtipo tumoral a la muestra de expresión génica observada a aquel subtipo que maximiza la correlación de Spearman entre el centroide de expresiones génicas propuesta por Perou et al. (2000) y la muestra incógnita. Basado en las correlaciones de Spearman que reporta PAM50, se analizaron dichos valores en las muestras que no cambiaron su clasificación con respecto a la inicial y aquellas muestras que sí cambiaron su clasificación con

respecto a la inicial, con el fin de observar la fuerza de asociación o asignación del subtipo tumoral.

Adicionalmente se calcularon siete medidas de desempeño (Cuadro 7) para cada una de las bases utilizadas, así como a la base de entrenamiento. Los siguientes términos son fundamentales para entender la prueba y los resultados de las medidas de desempeño en pruebas clínicas, donde:

- K representa la cantidad de clases que el clasificador puede asignar una etiqueta
- N es la cantidad de muestras o pacientes que contiene la base utilizada.
- VP es el paciente que tiene una etiqueta de enfermo y el clasificador le vuelve asignar la misma etiqueta (Verdadero positivo).
- VN es el paciente que tiene una etiqueta de sano y el clasificador le vuelve asignar la misma etiqueta (Verdadero negativo).
- FP es el paciente con etiqueta de sano y el clasificador le asigna etiqueta de enfermo (Falso positivo).
- FN es el paciente con etiqueta de enfermo y el clasificador le asigna etiqueta de sano (Falso negativo).

Cuadro 7. Medidas de desempeño de un clasificador para tablas a dos vías de clasificación (2×2).

Medida	Fórmula	Criterio de evaluación
Exactitud promedio	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i + VN_i}{N}$	Efectividad promedio para las clases tumorales
Micro-Especificidad	$\frac{\sum_{i=1}^{K} VN_i}{\sum_{i=1}^{K} VN_i + FP_i}$	Habilidad de la prueba a identificar correctamente los verdaderos negativos
Micro-Sensibilidad	$\frac{\sum_{i=1}^{K} VP_i}{\sum_{i=1}^{K} VP_i + FN_i}$	Habilidad de la prueba a identificar correctamente los verdaderos positivos
Micro F-Score	$2 * \frac{Micro-esp * Micro-Sen}{Micro-esp + Micro-Sen}$	Es el promedio entre la Especificidad y la Sensibilidad
Macro-Especificidad	$\frac{1}{K}\sum_{i=1}^{K}\frac{VN_i}{VN_i + FP_i}$	Habilidad de la prueba a identificar correctamente los verdaderos negativos
Macro-Sensibilidad	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i}{VP_i + FN_i}$	Habilidad de la prueba a identificar correctamente los verdaderos positivos
Macro F-Score	$2 * \frac{Macro-esp * Macro-Sen}{Macro-esp + Macro-Sen}$	Es el promedio entre la Especificidad y la Sensibilidad

3.3.-Simulación de datos

Para cada subtipo de cáncer de mama se generaron 150 muestras de tejido tumoral incluyendo el subtipo Normal Like. Se utilizaron los centroides o expresiones génicas medias de cada subtipo tumoral que utiliza PAM50 como referencia y la matriz de varianza y covarianza de cada subtipo tumoral obtenidas a partir de la base de datos de entrenamiento GEOBreastCancerData. Antes de generar las muestras, las expresiones génicas de la base de entrenamiento fueron corregidas utilizando el valor de proporción tumoral estimado para cada muestra por medio de la metodología propuesta y se reclasificó utilizando PAM50. A partir de esta nueva reclasificación se estimó la matriz de varianza y covarianza para cada subtipo tumoral siguiendo los pasos del punto 3.2.2. Finalmente se obtuvo una matriz de 750

muestras con pacientes que presentan algún subtipo tumoral. Además, se generaron 750 muestras de expresiones génicas de tejido Normal utilizando las expresiones génicas medias y matriz de varianza y covarianza estimada en los puntos 3.2.1 y 3.2.2 utilizando las expresiones génicas de los pacientes categorizados como Sanos de la base de datos de entrenamiento.

Seguidamente se hizo una combinación lineal de las dos cohortes de muestras generadas para obtener una sola matriz de expresiones génicas. La combinación lineal se hizo variando la cantidad de expresión génica tumoral y de expresión génica normal, pre-multiplicando la expresión génica por un valor de proporción establecido, simulando muestras que tengan menor proporción de expresión de tejido tumoral y más expresión de tejido normal hasta muestras que tengan mayor proporción de expresión de tejido tumoral y menos expresión de tejido normal. El valor de proporción tumoral propuesto fue de 0.1 y 0.9 aumentando en 0.1, por lo que se obtuvieron 9 matrices en total. Los valores base de expresión génica de las muestras fueron las mismas que las iniciales, solamente se varió la proporción de expresión en la muestra.

Se evaluó el algoritmo propuesto sobre cada una de las matrices anteriormente mencionadas y se estimó el valor de la proporción de tumor propuesto, luego se analizó el sesgo del valor de proporción estimado en relación al establecido, por último se corrigió el valor de expresión para cada una de las muestras utilizando el valor de proporción predicho y se volvió a utilizar PAM50 para una reclasificación y se observó como influyó la clasificación dado el grado de contaminación en los diferentes subtipos de PAM50.

Dado que se cuenta con una clasificación inicial de las muestras y se consideró como una clasificación verdadera, se evaluaron las medidas de desempeño propuestas en el apartado 3.2.4 para cada unas de las 9 matrices generadas anteriormente y ver como se comportaron según el valor de proporción propuesto y el valor de proporción estimado. De la misma manera que se evaluó en la base de entrenamiento y con las bases de datos públicas, se analizaron los valores de correlación de Spearman reportados por PAM50 en aquellas muestras que no cambiaron su clasificación con respecto a la inicial y con las muestras que cambiaron su clasificación respecto a la inicial.

4- Resultados

Para todas las validaciones que se realizaron evaluando el algoritmo propuesto para la estimación del valor de proporción de tumor en las muestras con cáncer, los resultados obtenidos convergieron a un mismo valor de proporción predicha independiente del valor de inicio que se le estableció a la función por medio del optimizador. Por tal razón no se indica un valor de inicio para la función ni en la validación con datos simulados, con la base de entrenamiento ni con las bases de datos reales.

4.1- Base de entrenamiento GEOBreastCancerData

La distribución de la expresión génica para las 78 muestras de las expresiones génicas de las muestras sanas y para las 103 muestras clasificados como Normal-like se muestran en la Figura 4. Los coeficientes de variación para las expresiones génicas de las muestras de pacientes Sanos y los clasificados como Normal-Like son 26.44 y 26.77% respectivamente.

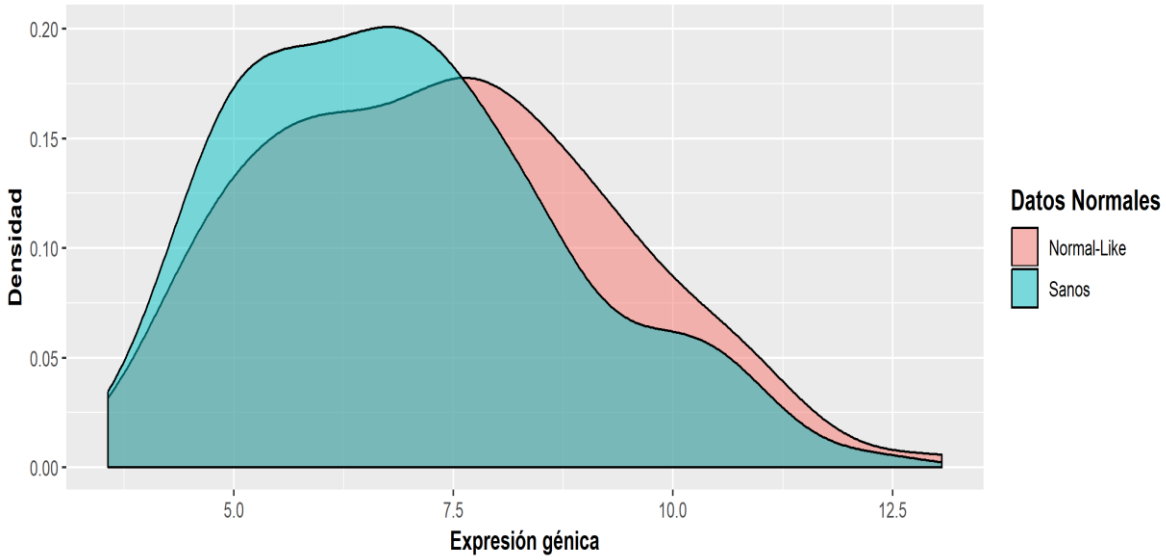

Figura 4. Función de Densidad para la expresión de 126 genes utilizando las muestras de pacientes sanos y normal-like como alternativa a las normales para la base datos pública GeoBreastCancerData.

El utilizar las expresiones génicas de las muestras clasificadas como Normal-Like como alternativa a las expresiones génicas de las muestras de pacientes Sanos para entrenar al algoritmo, genera una mejor estimación en los valores tumorales predichos. El cuadro 8 tiene las principales medidas de resumen obtenidas para el valor de proporción tumoral estimado por el algoritmo bajo las dos alternativas de expresiones génicas Normales.

Cuadro 8. Medidas de resumen del valor de proporción predicho por el algoritmo.

Clasificación	Mínimo	Q1	Mediana	Q3	Máximo	Media	Desvío estándar
Sanos	0.16	0.19	0.21	0.23	0.36	0.21	0.03
Normal-Like	0.48	0.53	0.53	0.54	0.58	0.53	0.01

En la Figura 5 se observa la función de distribución para la proporción tumoral predicha para las muestras de los pacientes Sanos y para las muestras de los pacientes Normal-Like observándose dos patrones muy distintos siguiendo una distribución normal para las muestras de pacientes sanos y una distribución t-student para las muestras de “Normal-like”. Al utilizar las expresiones génicas de las muestras sanas, la mayor parte de los valores predichos se concentran alrededor de proporciones de 0.2 existiendo cierta cantidad de muestras con valores mayores mientras que los valores predichos al utilizar las expresiones génicas de Normal-Like, la mayoría de las muestras se les detectó en promedio 0.5 de expresión génica tumoral.

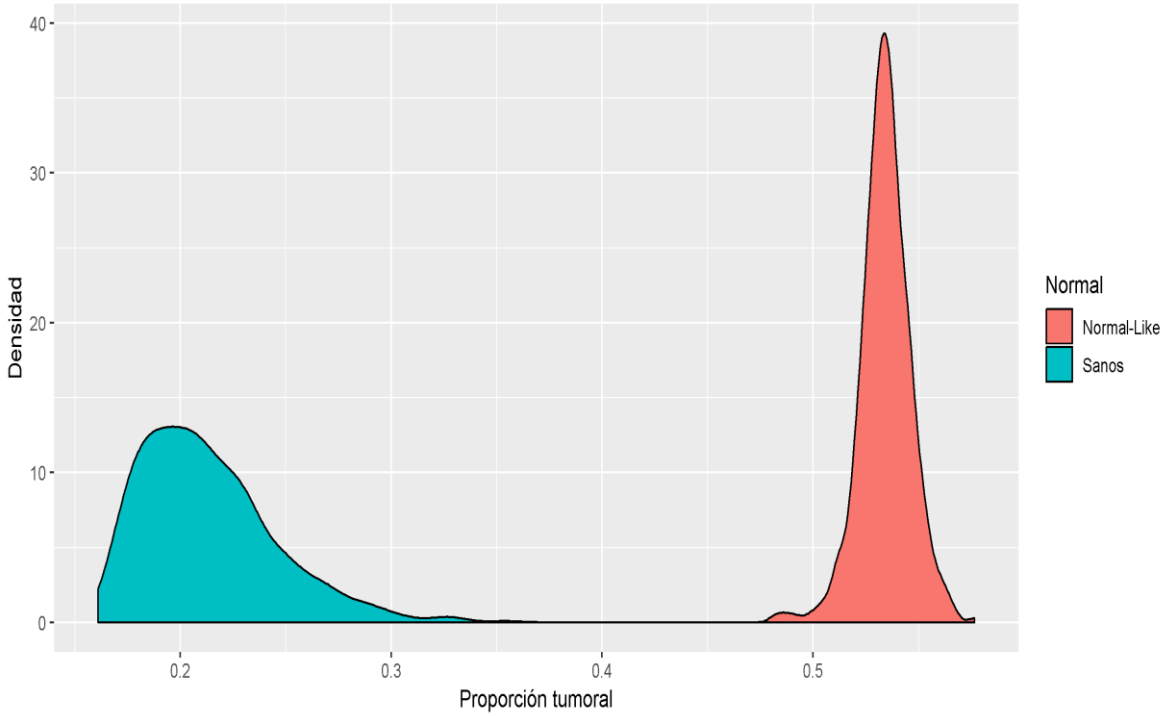

Figura 5. Función de densidad para la proporción tumoral detectada por el algoritmo utilizando las muestras de pacientes sanos y normal-like como normales para la base GeoBreastCancerData.

El cuadro 9 tiene la reclasificación de las muestras luego de haber corregido la matriz de expresión génica utilizando los valores de proporción tumoral predicho observándose que, al utilizar las expresiones génicas de los pacientes Sanos, los subtipos Basal, LumB y Normal tuvieron un aumento en la cantidad de muestras mientras que para Her2 y LumA disminuyo la cantidad de muestras con respecto a la clasificación inicial. Lo contrario sucedió al utilizar las expresiones génicas de los pacientes “Normal-Like” donde las muestras de los subtipos Basal y LumA aumentaron y las muestras de Her2 y LumB disminuyeron con respecto a la clasificación inicial. No se indica la clasificación obtenida para el subtipo Normal utilizando la alternativa Normal-Like dado que las muestras iniciales fueron utilizadas para entrenar al algoritmo.

Cuadro 9. Clasificación de muestras de la base GeoBreastCancerData utilizando PAM50.

Clasificación	Subtipos de cáncer				
	Basal	**Her2**	**LumA**	**LumB**	**Normal**
Inicial	163	104	171	100	103
Sano	169	84	152	125	111
Normal-Like	176	92	189	81	-

Independientemente de la alternativa de expresiones génicas normales utilizada, el cambio o reasignación de las etiquetas de las muestras con respecto a la clasificación inicial, hubo cambios de muestras que inicialmente se clasificaron a un subtipo tumoral con un diagnóstico malo a un subtipo con mejor diagnóstico y viceversa. Como se observa en la matriz de contingencia del cuadro 10, para la alternativa Normal-Like no existe una reclasificación ya que al utilizarlos como pacientes Normales para entrenar el algoritmo al estimarse la matriz de varianza y covarianza se estarían utilizando los mismos datos.

Cuadro 10. Matriz de contingencia para la clasificación inicial y reclasificación de muestras de la base GeoBreastCancerData utilizando PAM50

Alternativa Normal	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	133	10	1	19	6
	Her2	2	73	3	6	0
	LumA	0	1	135	6	10
	LumB	22	16	19	61	7
	Normal	6	4	13	8	80
Normal-Like	Basal	163	6	0	7	-
	Her2	0	90	0	2	-
	LumA	0	6	171	12	-
	LumB	0	2	0	79	-
	Normal	-	-	-	-	-

El utilizar las expresiones génicas de Normal-Like como alternativa a muestras Normales, las medidas de desempeño obtenidas que se observan en el

cuadro 11 son mayores en comparación a la alternativa de las expresiones génicas de muestras sanas. La precisión a nivel micro y macro son las medidas que mayor porcentaje tuvieron.

Cuadro 11. Medidas de desempeño del clasificador utilizando la base GEOBreastCancerData

Clasificación	Medidas de desempeño (%)						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	90	94	75	83	94	74	83
Normal-Like	97	98	93	96	98	91	94

La figura 6 muestra las correlaciones de Spearman con la que fueron asignadas las etiquetas de los subtipos a las muestras utilizando PAM50. Independientemente de las alternativas de las expresiones génicas utilizadas como Normales y si hubo cambio de subtipo tumoral en la muestra con respecto a la clasificación inicial, las correlaciones de spearman de las muestras para cada subtipo tumoral estuvieron por debajo de 0.5. Aquellas muestras que cambiaron de subtipo luego de haber corregido su expresión génica y ser reclasificados, obtuvieron correlaciones por debajo de 0.2, generándose una incertidumbre en la

nueva etiqueta del subtipo tumoral a la muestra.

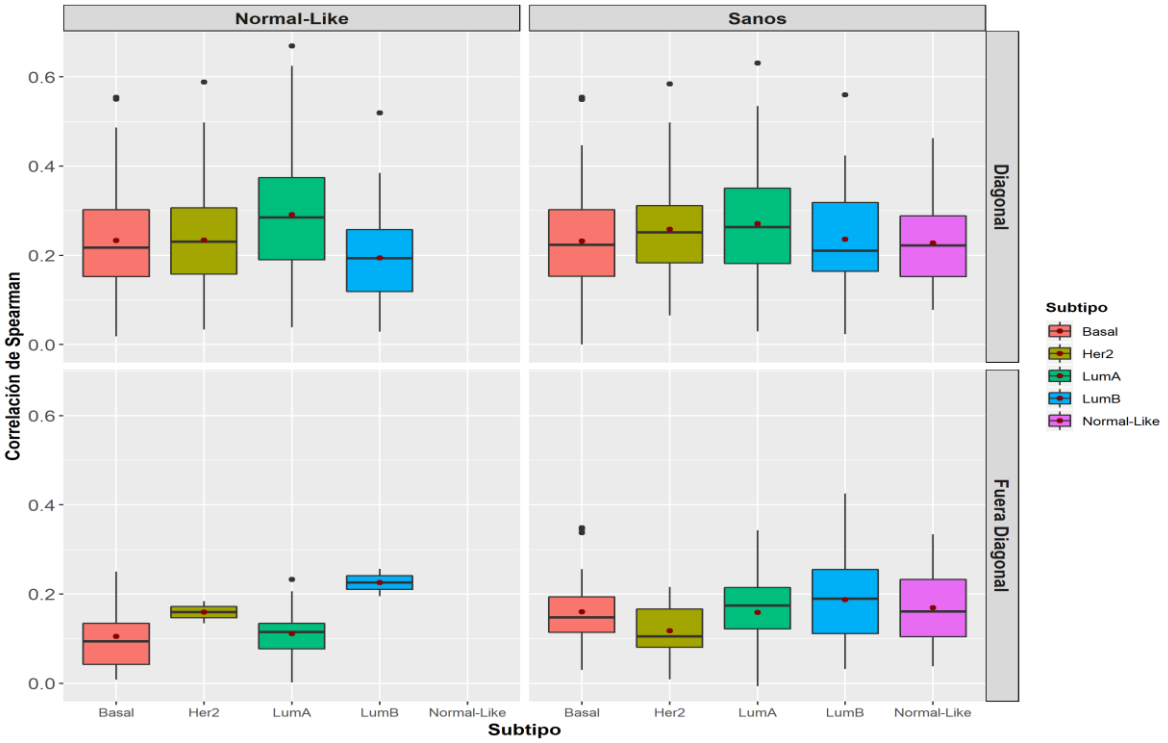

Figura 6. Diagrama de cajas para las correlaciones de Spearman dadas por PAM50 a las muestras de la base GeoBreastCancerData que cambiaron su clasificación con respecto a la inicial y para las muestras que no cambiaron su clasificación.

4.2- Simulación de datos

La figura 7 muestra las medidas de resumen de los valores predichos de la proporción tumoral que se utilizó para generar las expresiones génicas de las muestras de las clases tumorales y de las muestras de sanos mediante la combinación lineal propuesta. Solamente se consideraron las muestras de los pacientes sanos de la base de datos GeoBreastCancerData.

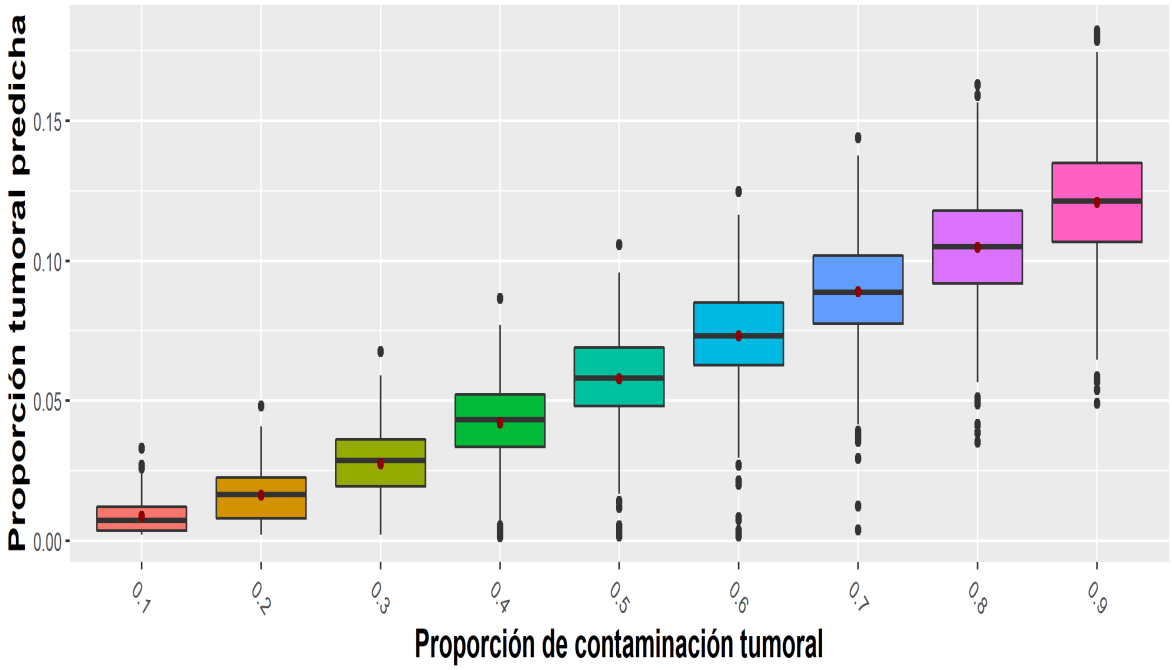

Figura 7. Diagrama de cajas para los valores de proporción tumoral predichos por el algoritmo para las expresiones génicas simuladas.

Al ver la distribución de los valores predichos de proporción tumoral para cada valor que se utilizó para fijar el contenido de tejido tumoral en la base de datos, la figura 8 muestra que las proporciones tumorales siguen una distribución normal y solamente cuando se contaminó con un valor de 0.1 se logra ver una densidad con un pico máximo donde representa que la mayoría de los pacientes tuvieron un valor de 0.009 en promedio y para el resto de los valores de contaminación predichas presentan valores en un rango amplio.

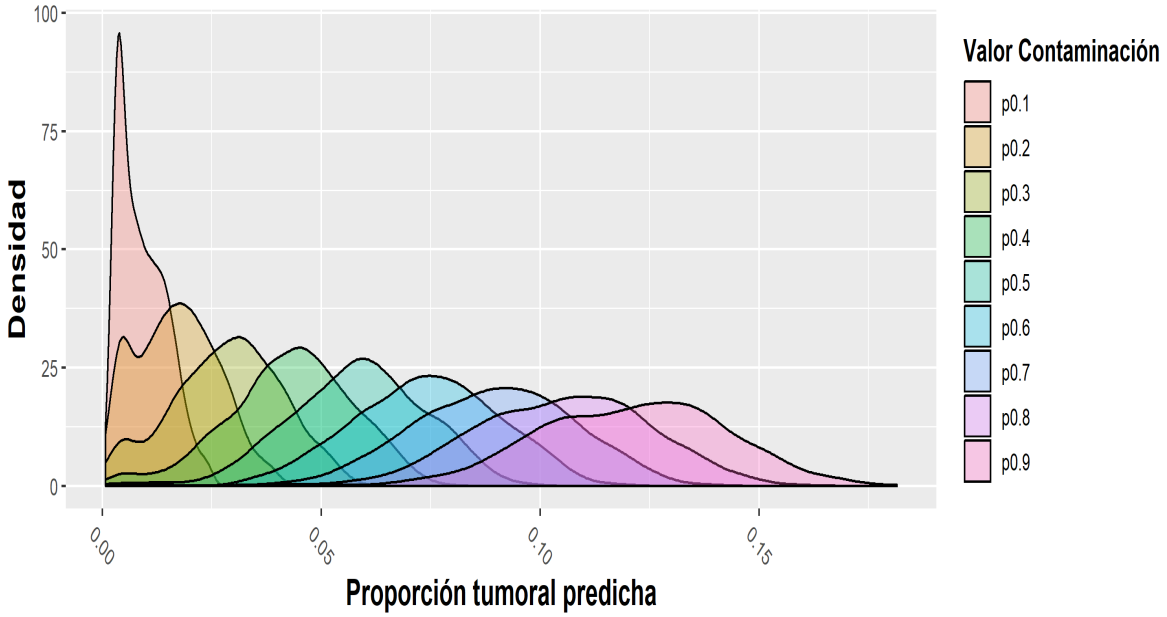

Figura 8. Función de densidad para la proporción tumoral obtenidas para la combinación lineal de las expresiones génicas tumorales y expresiones génicas normales simuladas.

Con respecto a los valores de proporción tumorales obtenidos al evaluar el modelo utilizando el algoritmo para obtener el valor fijado y posteriormente corrigiendo la base de datos simulada con una cohorte de pacientes normales junto con un valor de contaminación para obtener la clasificación inicial, a valores bajos del valor proporción tumoral se obtiene una reclasificación no deseada, si bien las diferencia de los pacientes reclasificados con respecto a la inicial no es muy amplia, el cambio de etiquetas de subtipos reclasificados si es considerable debido a que hay cambios en todas las direcciones y en magnitudes (ver Anexo 5). Conforme aumenta el valor de proporción tumoral, las clasificaciones fueron mejorando, luego de un valor de 0.6 tanto las clasificaciones de los pacientes como los cambios de etiquetas en los pacientes para los distintos subtipos tumorales se fueron acercando al valor inicial propuesto, pero no alcanzando el 100% de la clasificación inicial.

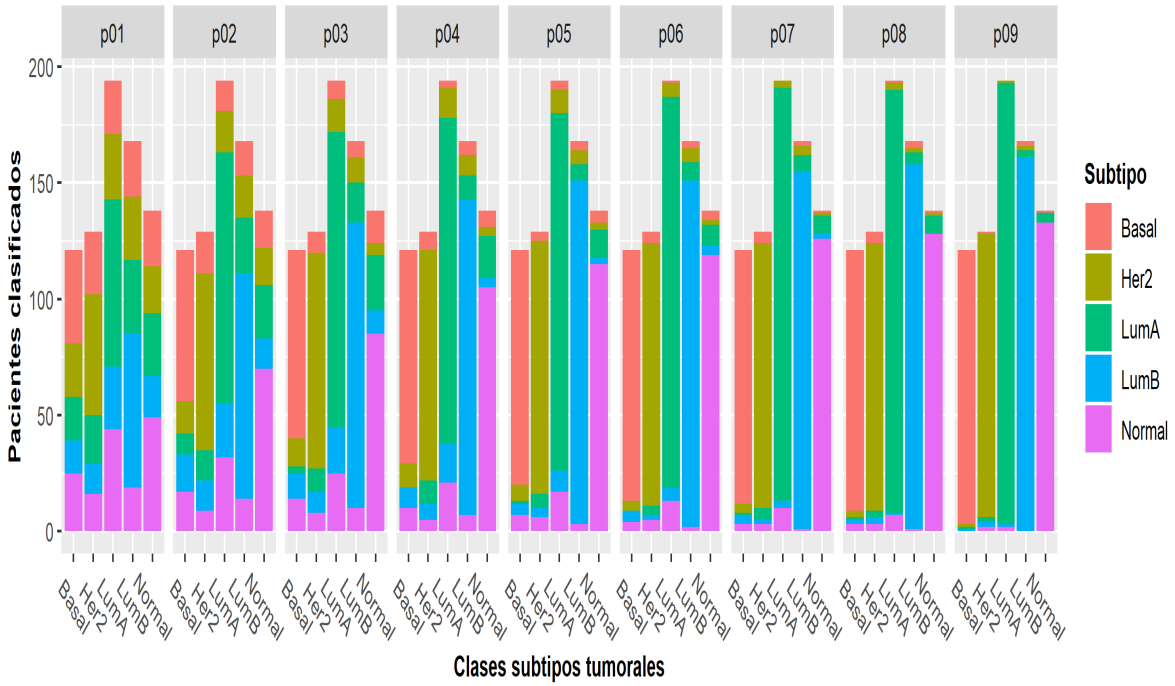

Figura 9. Gráfico de barras para la reclasificación de las muestras en relación con la clasificación inicial bajo los distintos valores de proporción tumoral propuesto.

A medida que las clasificaciones fueron mejorando de la misma forma las medidas de desempeño calculadas a partir de los resultados fueron aumentando (figura 10). La exactitud promedio, que representa la efectividad promedio de las clases en acertar correctamente, fue aumentando desde un 75% con valores de proporción tumoral de un 0.1 hasta un 99% cuando el valor fijado fue de 0.9. Con respecto a la precisión que detecta la habilidad de acertar verdaderos negativos se obtuvo porcentajes finales de 99 para nivel micro y macro análisis y porcentajes finales de sensibilidad de 96 por ciento de forma micro y macro. Al hacer un balance entre la precisión y sensibilidad se obtienen valores de 98 por ciento para micro y macro análisis.

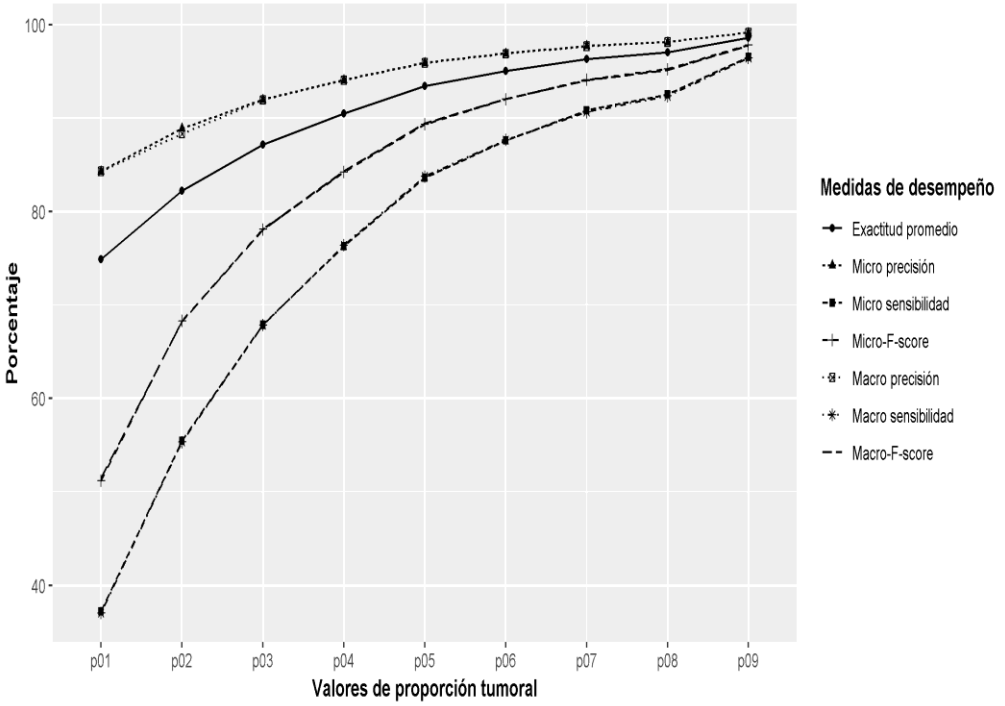

Figura 10. Medidas de desempeño del clasificador PAM50 luego de la corrección de la expresión para la combinación lineal de las expresiones génicas tumorales y expresiones génicas normales simuladas.

En la figura 11 se muestran las correlaciones máximas de Spearman obtenidas por PAM50 al asignar una etiqueta a la cohorte de pacientes dado el valor de proporción tumoral fijado. Las correlaciones obtenidas para los pacientes que no cambiaron su clasificación con respecto a la inicial su rango oscila en 0.2 promedio, inicialmente un poco variable, conforme se fue aumentando la proporción tumoral en la contaminación las correlaciones se fueron estabilizando. Lo mismo sucedió con aquellos pacientes en los que su clasificación cambió con respecto a la inicial, siendo muy variables al inicio con correlaciones de 0.2 en promedio para luego estabilizarse a los 0.1.

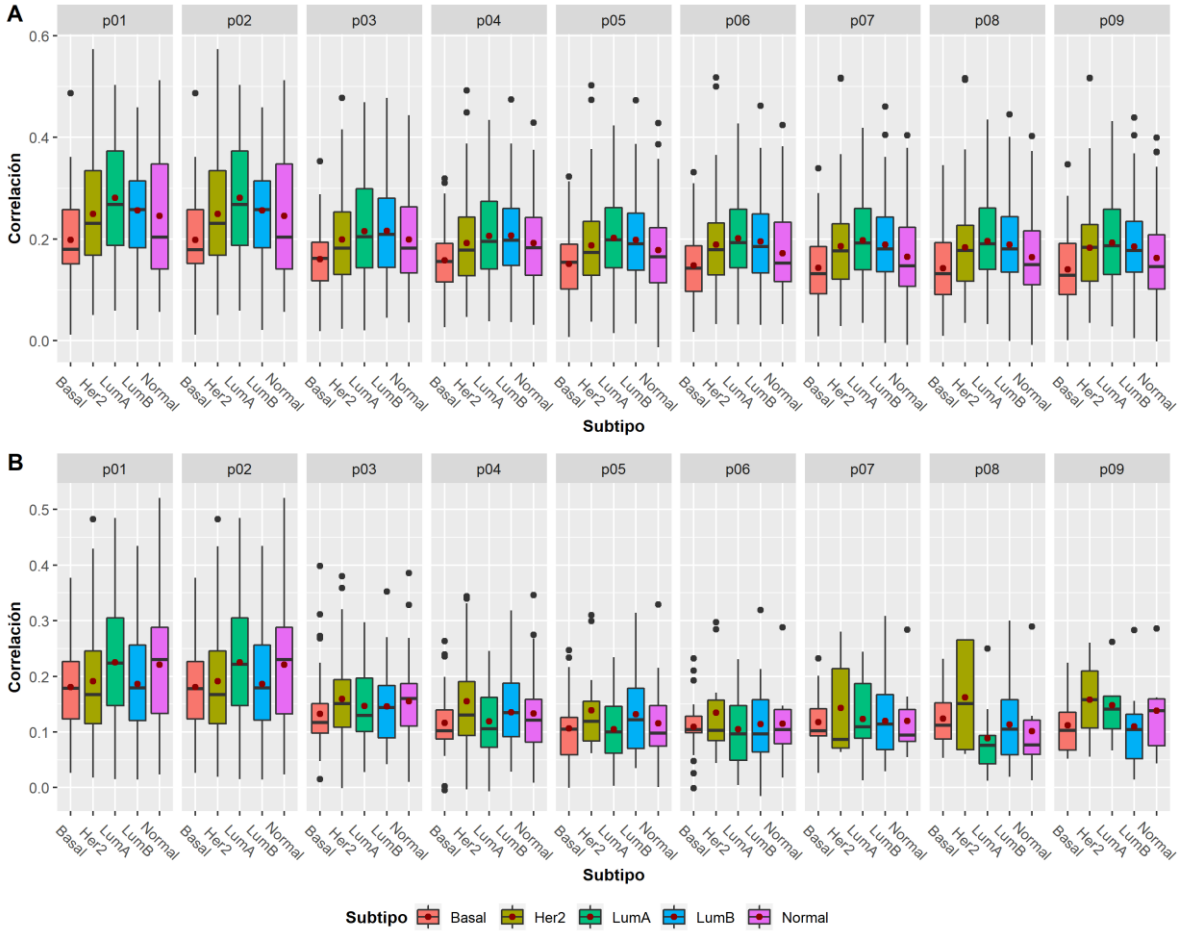

Figura 11. Correlaciones máximas de Spearman obtenidas por PAM50 para las muestras que no cambiaron de etiqueta con respecto al inicial (A) y para las muestras que cambiaron de etiqueta (B).

4.3- Bases de datos públicas

Los análisis de las bases de datos públicas de manera individual se encuentran del anexo 5 al anexo 8.

4.3.1 – Análisis general de las bases públicas

Se analizaron las cinco bases de datos con las expresiones génicas de cada muestra ($n = 1130$). La figura 12 muestra las funciones de densidad obtenidas para los valores de proporción tumoral estimada por el algoritmo observándose que al utilizar las expresiones génicas de las muestras clasificados como “Normal-Like” de

la base de datos GeoBreastCancerData se lograron valores mayores en comparación al utilizar las muestras de pacientes Sanos.

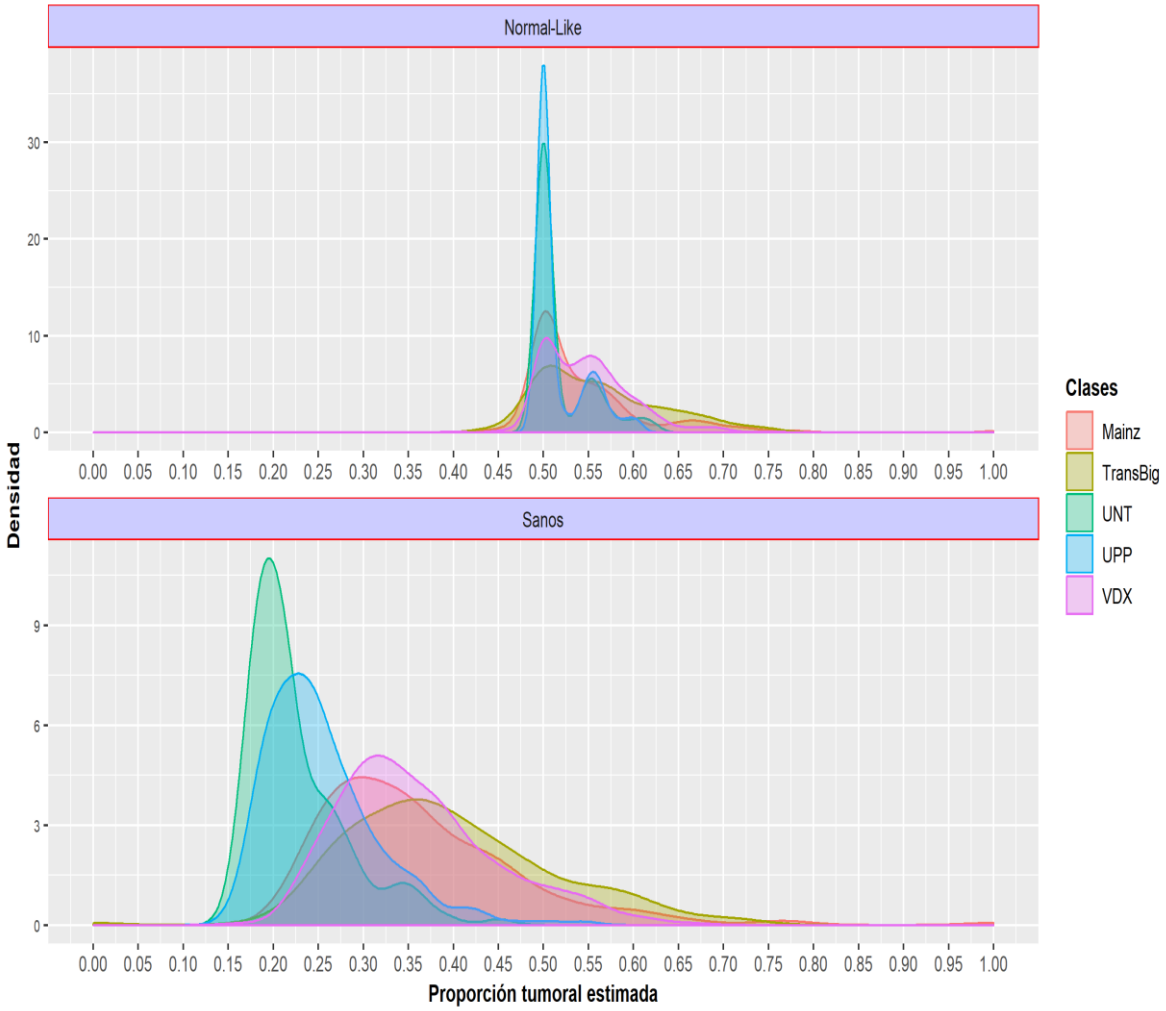

Figura 12. Funciones de densidad para la proporción tumoral estimada utilizando expresiones génicas de muestras clasificadas como Sanos y Normal-Like como alternativa a expresiones génicas Normales.

El cuadro 12 presenta la clasificación dada por PAM50 en frecuencia absoluta y en paréntesis la frecuencia relativa de las $n = 1130$ muestras observándose que la cantidad de muestras que inicialmente fueron etiquetados como LumA y Her2 disminuyeron luego de la corrección de la expresión génica de cada muestra, mientras que los subtipos Basal y Normal-Like la cantidad de muestras se mantuvieron en cantidad y LumB aumentó en uno por ciento.

Cuadro 12. Clasificación de las $n = 1130$ muestras con expresiones génicas utilizando PAM50.

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal-Like
Inicial	226 (20.0)	176 (15.6)	340 (30.1)	233 (20.2)	155 (13.7)
Sanos	227 (20.1)	180 (15.9)	357 (31.6)	217 (19.2)	149 (13.2)
Normal-Like	228 (20.2)	197 (17.4)	311 (27.5)	238 (21.1)	156 (13.8)

Las medidas de desempeño que se muestran en el cuadro 13 obtenidos de manera global fueron muy buenas en ambos casos, siendo por arriba del 95 por ciento y menores al 99. Sin embargo, los resultados obtenidos al utilizar las expresiones génicas de las muestras clasificadas como “Normal-Like” fueron más bajas en comparación con los resultados al utilizar muestras Sanas, reflejando que hubo un mayor cambio en la reclasificación de las muestras con respecto a la clasificación inicial.

Cuadro 13. Medidas de desempeño del clasificador PAM50 luego de la corrección de las expresiones génicas.

Clasificación	Medidas de desempeño (%)						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	98	99	95	97	99	95	97
Normal-Like	97	98	95	95	97	98	95

Los porcentajes de cambio expresados en el cuadro 14 de “Clasificación de las $n = 1130$ muestras utilizando PAM50.” no reflejan el verdadero cambio que existió luego de la reclasificación. El cuadro 15 es una matriz de contingencia para la clasificación inicial y la reclasificación de las muestras, observándose los verdaderos movimientos de las muestras dado la nueva asignación del subtipo tumoral luego de haber aplicado la corrección sobre la expresión génica con el valor de proporción estimado. Al utilizar las expresiones génicas de las muestras sanos como alternativa a las normales, los mayores cambios se dan para las clases tumorales LumB donde

cambian quince muestras a LumA y para Normal-Like diez muestras cambiaron a LumA. Lo contrario sucede cuando se utilizan las expresiones génicas clasificadas como “Normal-Like”, la clase tumoral LumA tuvo dos cambios importantes en las muestras, el primero fueron quince muestras que pasaron a LumB y diecinueve muestras a Normal-Like. En el anexo 9 se muestran de manera gráfica los cambios de etiqueta de las muestras luego de haber corregido la matriz de expresión génica.

Cuadro 14. Matriz de contingencia para la clasificación inicial y reclasificación de las muestras utilizando las distintas alternativas de expresiones génicas Normales.

Tipo de Normal	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal-Like
Sano	Basal	221	1	0	1	4
	Her2	4	171	0	5	0
	LumA	0	1	331	15	10
	LumB	0	3	2	212	0
	Normal	1	1	7	0	141
Normal-Like	Basal	219	0	0	2	7
	Her2	7	174	4	10	2
	LumA	0	0	302	1	8
	LumB	0	2	15	220	1
	Normal	0	0	19	0	137

De manera global, se calcularon todas las correlaciones de Spearman obtenidas para cada una de las muestras de cada base de datos. Las correlaciones con las que se les reasignó una etiqueta a las muestras que no cambiaron de etiqueta con respecto a la clasificación inicial fueron mayores a 0.5 pero no mayores a 0.75 en promedio, mientras que para aquellas muestras que cambiaron de etiqueta, las correlaciones estuvieron entre 0.2 y 0.4 en promedio. Ambas situaciones se presentaron independientemente de las alternativas de expresión génica utilizadas como Normales. La figura 13 muestra las correlaciones de

Spearman de manera gráfica obtenidas tanto para las muestras que se mantuvieron en la diagonal y fuera de ellos utilizando ambas muestras de pacientes Normales.

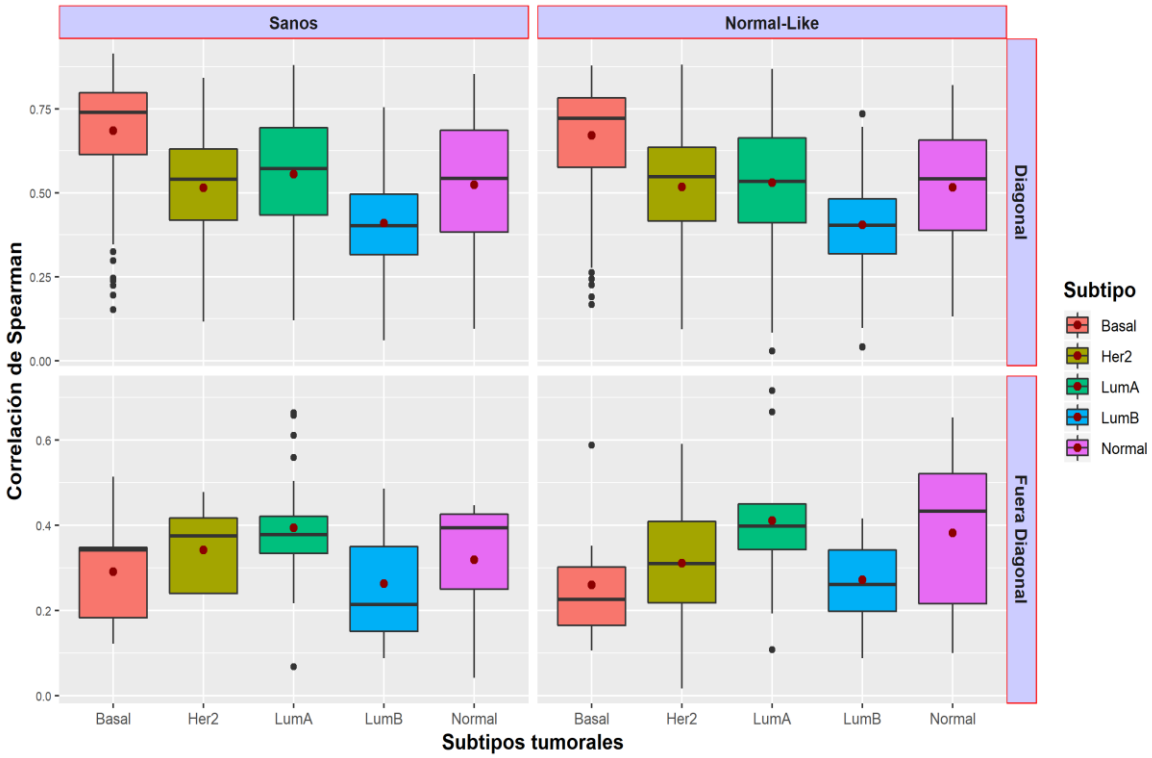

Figura 13. Correlaciones máximas de Spearman obtenidas por PAM50 para las muestras que no cambiaron de subtipo tumoral con respecto al inicial (Diagonal) y para las muestras que cambiaron de subtipo tumoral (Fuera Diagonal).

5-Discusión

5.1 - Análisis del uso de distintas alternativas de expresiones génicas normales y simulación de expresiones génicas.

El conocimiento del microambiente en el que se desarrolla un cáncer es importante en la comprensión de la biología de la enfermedad. Slaughter et al. (1953) citado por Aran et al. (2017) estudió el tejido normal adherido a una biopsia de cáncer de mama en cientos de muestras para analizar las características biológicas estableciendo que dicho tejido se encuentra en un estado intermedio pre-neoplásico compuesto de células morfológicamente normales pero alteradas molecularmente. Estudios más recientes realizados por Graham et al. (2010) llegaron a la conclusión que el efecto del microambiente alrededor del tumor es esencial para desarrollar respuestas terapéuticas y métodos quirúrgicos. Sin embargo, a nivel molecular los patrones de expresión génica de tejido normal y epitelio en cánceres de mama en humanos no han sido ampliamente estudiados (Aran et al. 2017).

El subtipo “Normal-Like” presenta características muy similares al subtipo Luminal A y Basal, se encuentra en un diagnostico intermedio entre los luminales y Basal y presenta una buena prognosis, sin embargo, es un poco más ligera que Luminal A. Weigelt et al. (2010) sostiene que este subtipo no existe y que es considerado como un subtipo con una alta contaminación de tejido normal, adiposo, entre otros, apoyado en que se han realizado disecciones manuales para eliminar tejido circundante obteniendo como resultado cero casos encontrados para esta categorización. Grahamn (2002) determinó que los perfiles de expresión génica del estroma de pacientes normales y de epitelio de pacientes con cáncer de mama no son estadísticamente distintos.

Elloumi et al. (2010) utilizaron 48 muestras de pacientes clasificados como “Normal-Like” para obtener la mediana de la expresión génica para generar perfiles prototipos normales como base para generar nuevas expresiones controlando el grado de expresión normal en la muestra y obtuvo resultados similares en clasificaciones bajo distintos escenarios, variando de cero a 50 por ciento al compararlos con los obtenidos en el mismo estudio con muestras pareadas (muestras con cáncer y sanas de un mismo paciente). Por tal razón es que se consideró en este trabajo de investigación el uso de las expresiones génicas de muestras clasificadas como “Normal-like” como alternativa al uso de las expresiones génicas Normales, cuando no se cuentan con datos de expresiones génicas de pacientes sanos.

Otra razón que se consideró para usar las expresiones génicas de muestras “Normal-Like” se basó en la variación y el rango intercuantil que presentan las expresiones génicas que tienen las distintas alternativas de muestras considerados como normales. Variaciones bajas para datos Normales tienden a tener un efecto similar a través de las muestras cuando se requiere asignar una etiqueta a una muestra por medio de un clasificador mientras que variaciones altas permiten tener una mejor clasificación dado que permite una mejor dinámica en la distinción y asignación entre las distintas clases tumorales (Elloumi et al. 2010).

Con relación a la generación de muestras de expresiones génicas de distintas clases tumorales y muestras de expresiones génicas normales, analizando los valores de proporción tumorales predichos por el algoritmo basado en el valor de proporción tumoral fijado para la combinación lineal de las muestras de pacientes, los resultados obtenidos fueron los mismos independientemente del valor de inicio fijado, por lo que la estimación de la proporción de expresión tumoral utilizando el

método de máxima verosimilitud asegura una estabilidad al buscar el valor que maximiza la función. Con respecto a la estimación de la proporción tumoral fijada para realizar la combinación lineal y lo estimado por la función no existió concordancia, ya que la discrepancia entre lo predicho y lo estimado fue muy grande.

Sin embargo, cuando se realiza la corrección de la matriz de expresión génica con los valores estimados, se observa como la reclasificación de las muestras se va semejando a la propuesta inicialmente conforme se va aumentando la proporción de tejido tumoral existiendo una relación con las medidas de desempeño calculadas. A partir de los valores que estima el algoritmo con una proporción tumoral de 0.5, las medidas de desempeño empiezan a mostrar valores por arriba del 80% en la detección de verdaderos positivos y 95% en la detección de verdaderos negativos, hasta llegar a porcentajes de 96% en sensibilidad y 99% de especificidad para valores altos de proporción tumoral fijada. El balance entre ambas medidas de desempeño llega a ser de 89% tanto a nivel Micro como Macro cuando el algoritmo detecta 0.6 de promedio de proporción tumoral cuando el verdadero valor fue de 0.5 y cuando el verdadero valor fue de 0.9 el algoritmo detectó 0.12 como valor predicho sin embargo el balance para las medidas de desempeño fue de 98% para ambos niveles.

5.2 - Análisis de bases de datos públicas

Al analizar la función de distribucion de las expresiones génicas de las muestras sanas y de las muestras clasificadas como “Normal-like”, se observa un solapamiento en la mayoría de los valores de expresión mostrando no existir una marcada variación para dichas alternativas, sin embargo, los resultados obtenidos

de las proporciones estimadas para cada una de las muestras en cada base de datos se comportaron totalmente distintos.

Utilizando las expresiones génicas de las muestras Sanas, los valores predichos para las muestras de las bases UNT y UPP la diferencia de medias es de 0.03 de proporción tumoral y para las bases VDX, TransBig y Mainz la diferencia de medias es de 0.15 en promedio de proporción tumoral, comparando entre ambos grupos de bases la diferencia promedio para la proporción tumoral predicha es de 0.13 $(p \leq 0.05)$.

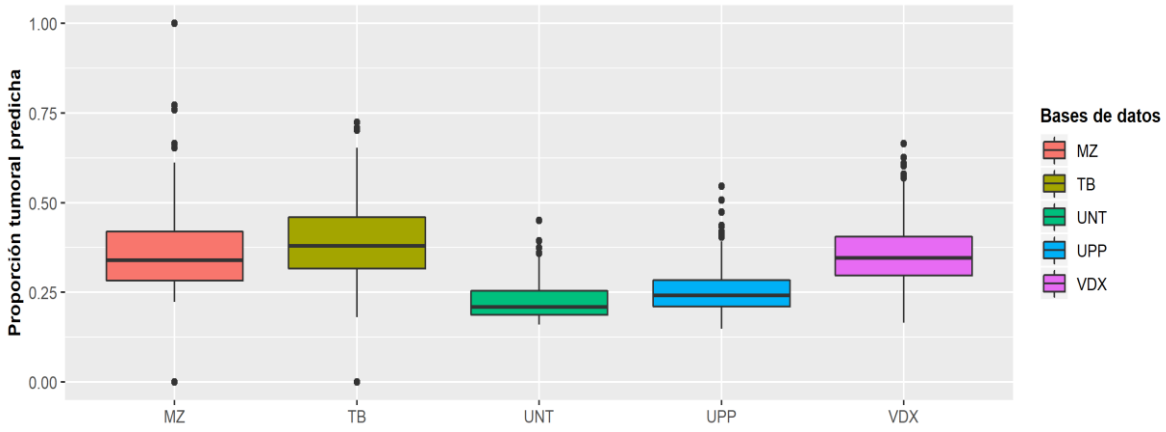

Figura 14. Diagrama de cajas de las proporciones tumorales predichas por el algoritmo utilizando muestras de pacientes Sanos como alternativa a pacientes Normales.

Para el caso de las expresiones de las muestras clasificados como “Normal-Like” el comportamiento de los valores de proporción tumoral predicha tuvieron el mismo comportamiento que en el caso de las muestras de los pacientes sanos, solamente la magnitud de la proporción tumoral predicha fue lo que cambió. Para las bases UPP y UNT no hubo diferencias entre los valores $(p \geq 0.05)$, mientras que para las bases VDX, Mainz y TransBig la diferencia promedio fue de 0.01 de proporción tumoral predicha. Entre ambos grupos la diferencia que existe es de 0.03 $(p \leq 0.05)$.

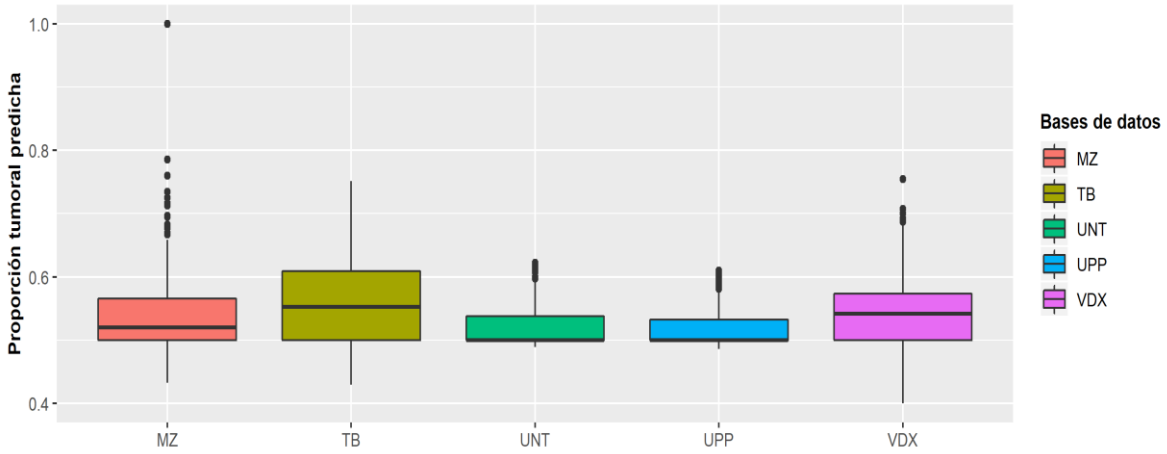

Figura 15. Diagrama de cajas de las proporciones tumorales predichas por el algoritmo utilizando muestras de pacientes “Normal-Like” como alternativa a pacientes Normales.

Se procedió a utilizar otros criterios de calidad para evaluar el desempeño del algoritmo propuesto dado que no se cuenta con matrices de expresión génica que se sepan que las expresiones génicas del material biológico provengan de una muestra pura, es decir que sea cien por ciento tumores por lo tanto no se conoce el verdadero subtipo de cáncer a la cual pertenece la muestra del paciente y por la heterogeneidad que presenta el desarrollo la misma enfermedad.

Al analizar las medidas de desempeño del algoritmo de manera global se obtiene una tasa micro de precisión y sensibilidad $98.2\% \leq \text{micro} - \text{f} - \text{precisión} \leq 98.7\%$ y $95.20\% \leq \text{micro} - \text{f} - \text{sensibilidad} \leq 95.22\%$ dependiendo de las expresiones génicas utilizadas como alternativa Normales y cuando se analizó de manera individual para cada base de datos para cada una de las alternativas utilizadas, el 95% las muestras fueron etiquetados correctamente utilizando las expresiones génicas de Sanos como alternativa a las normales y un 93% fueron etiquetados correctamente utilizando las muestras clasificadas como “Normal-Like” luego de la corrección de la matriz de expresión génica utilizando PAM50, observándose resultados muy positivos con errores bajos de cometer una mala clasificación con un 5% y 7% de detectar falsos negativos, sin embargo se puede

considerar alto por la importancia de la enfermedad. En relación con la sensibilidad, el algoritmo propuesto mostró valores promedio de 98% para las bases de datos reales obteniendo un valor de índice de falsos negativos relativamente bajo para estos tipos de prueba siendo de un 2%.

Elloumi et al. (2010) al utilizar muestras apareadas para ver cómo influye la proporción de tejido normal en muestras con cáncer de mama, observó que al aumentar la proporción de tejido normal en las muestras, la asignación de las etiquetas y el pronóstico cambió utilizando distintos clasificadores moleculares. Para el caso de PAM50 las muestras se reclasificaron de subtipos más agresivos a menos agresivos conforme se aumentó la cantidad de tejido normal y para el resto de los clasificadores moleculares la reclasificación se dio en distintas direcciones. Luego obtuvo el mismo resultado cuando probó el efecto del aumento de la proporción de tejido normal en muestras de bases de datos públicas, donde las asignaciones de las nuevas etiquetas de subtipos se movieron de subtipos menos agresivos a subtipos más agresivos.

La reasignación de las etiquetas en las muestras utilizando la matriz de expresión génica corregida luego de haber estimado para cada muestra un valor de proporción tumoral independientemente de la alternativa normal utilizada para entrenar al algoritmo, generaron cambios en ambas direcciones tal como se ha mencionado que la magnitud y la dirección de cambio en predictores genómicos pueda que se dé en una sola dirección (por ejemplo, pacientes que pasen de un diagnóstico malo a uno bueno) o que sea impredecible con direcciones de cambio inconsistentes.

Dependiendo de la alternativa de las expresiones génicas “Normales”, hubo dos clases que presentaron un mayor movimiento en la reasignación de las etiquetas a las muestras. Utilizando las muestras de Sanos la clase LumB tuvo quince muestras que pasaron a ser LumA y cuando se utilizó las expresiones de “Normal-Like” diez muestras cambiaron a LumA, contrario cuando se utilizan las expresiones génicas clasificadas como “Normal-Like”, la clase tumoral LumA tuvo dos cambios de muestras grandes, el primero fueron quince muestras que pasaron a ser LumB y diecinueve muestras como Normal-Like.

Sorlie y colegas encontraron grupos similares para Basal y Normal-Like bajo un agrupamiento utilizando el set de genes de cáncer intrínseco propuesto por ellos mismos (Sorlie et al. 2003), haciendo indicar que el aspecto de subtipos Basales y Normal-Like son más similares al tejido epitelial normal en comparación con los otros subtipos de cáncer de mama. La similaridad descrita entre los subtipos Basal y Normal-Like también fue demostrada por Perou *et al*. (2000). Bajo esta fundamentación se puede explicar cómo existe ese cambió tan abrupto en pacientes que inicialmente fueron clasificados como Normal-Like y luego de la corrección pasaron a ser Basal o Her2.

Sin embargo, al analizar los valores de correlaciones de Spearman con que se asignaron las nuevas etiquetas a las muestras que cambiaron de un subtipo a otro no fueron mayor a 0.5, evidenciando que la asociación entre el perfil de la muestra y el centroide más cercano fue muy débil. Las muestras que no cambiaron su etiqueta con respecto a la inicial en cada una de las bases de datos reales evaluadas, las correlaciones de Spearman obtenidas se mantuvieron en un rango de 0.5 a 0.8. La regla que utiliza PAM50 para asignar la etiqueta a la muestra con respecto al centroide de mayor similaridad no es confiables en el sentido que

pueden existir correlaciones cercanas a cero o incluso negativas y aun así asigna una clase a la muestra. Bajo este argumento no se puede refutar el hecho que los nuevos pacientes que cambiaron de etiquetas tengan una asociación con la etiqueta asignada dada la correlación tan baja, pero fue la que maximizó la asignación de la etiqueta.

Zhaoqi *et al*., (2014) llegan a una serie de conclusiones luego de realizar un conjunto de pruebas con diferentes clasificadores moleculares utilizando un total de 1975 pacientes clasificados en distintos subtipos. Una de ellas fue que la habilidad de predecir por parte de los clasificadores moleculares está directamente relacionada con la calidad de la base de datos y otra conclusión fue que el subtipo Normal-Like tiene altas correlaciones entre sus genes lo que permite ser una clase informativa de buena calidad sugiriendo que la expresión génica de este grupo contiene un alto valor predictivo.

Los cambios en dirección y en magnitud obtenidos en este trabajo están directamente relacionados con la base de datos que se utilizó en el algoritmo para ser entrenada dado hoy en día no existen bases de datos reales que se sepa que 100% de la expresión génica corresponda a tumor puramente (Ma et al. 2003), además dado al alto costo que tiene el utilizar la tecnología de microarreglos son muy pocas las bases de datos que contienen muestras pareadas que corresponden a una misma persona tener una muestra de tumor y otra de tejido sano.

Muy poco se sabe acerca de la expresión génica en tejido premaligno y los estudios que se han enfocado a la histología de tejido normal son limitados debido a la dificultad de obtener tejido fresco homogéneo, por tal razón Grahamn en el 2002 decidió estudiar la expresión de tejido normal en pacientes con cáncer y encontró

que aun a nivel de tejido normal existen heterogeneidad en poblaciones celulares entre ER+ y ER-. Finak, 2009 demostró que las variaciones en la expresión génica entre grupos de muestras provenientes de disecciones de tejido mamario o asociadas una muestra de tumor no están asociadas con características clínicas, pero pueden explicarse por tejido y variabilidad específica del paciente reforzando lo obtenido por Ma et al. 2003, que demostraron la falta de diferencias significativas entre la obtención de tejido normal por reducción mamaria y epitelio adyacente al cáncer (tres muestras) utilizando microarreglos de ADNc.

En consecuencia, los estudios de expresión génica que examinan los perfiles génicos completos de muestras de pacientes con cáncer de mama tratan estas muestras como una expresión génica homogénea, generando una evaluación inexacta. El usar los perfiles de expresión génica trae una serie de desventajas entre ellas se puede destacar la gran variación en expresión que pueda existir de paciente en paciente, la procedencia de las muestras y la condición experimental bajo la cual fue obtenida la muestra, entre otras puede generar mucho ruido y por ende un error en la asignación de las etiquetas a los pacientes.

6-Conclusiones

El método propuesto consistió en la deconvolución computacional de expresiones génicas provenientes de microarreglos de una mezcla compuesta por dos poblaciones (normal y tumor) de una muestra de cáncer de mama. Los perfiles de expresión génica utilizados como pacientes Normales para entrenar al algoritmo tienen un efecto considerable en la estimación del valor de proporción tumoral tanto para los pacientes simulados como en los pacientes de las bases de datos reales.

Las expresiones génicas de las muestras utilizadas para la estimación del vector de medias y de las matrices de varianza y covarianza para cada subtipo tumoral no son muestras biológicas totalmente puras y tampoco se conoce con certeza real la etiqueta correspondiente al subtipo tumoral por lo que fue un factor que influyó en la estimación del valor de proporción tumoral en la muestra.

El algoritmo propuesto estima el valor de proporción tumoral para cada muestra que permite ser utilizado en la corrección de su expresión génica y luego utilizar esta para una reasignación de una clase tumoral utilizando PAM50.

El contenido de tejido normal en una muestra de cáncer de mama tiene un efecto en la asignación de una etiqueta por medio de un clasificador molecular. El cambio de etiquetas en función de la agresividad de los subtipos tumorales se dio en ambas direcciones, existiendo pacientes que inicialmente fueron etiquetados con subtipos con un peor pronóstico para luego de la reclasificación a tener un subtipo con mejor pronóstico y viceversa con aquellos pacientes que inicialmente fueron etiquetados como un pronóstico bueno.

Dado que no se conoce la etiqueta real de la muestra del paciente, pero si una clasificación inicial basada en la expresión génica perturbada, las medidas de

desempeño obtenidas en la reclasificación de las muestras de los pacientes fueron casi del 100%. Cuando se utilizó las expresiones génicas de “Normal-like” se obtuvo valores más bajos (97%) que utilizar Sanos, evidenciando cambios en la reclasificación debido a valores mayores estimados para la proporción tumoral.

Las correlaciones de Spearman con las que PAM50 asignó la nueva etiqueta a los pacientes que cambiaron de subtipo fueron muy bajas independientemente de la base de datos pública utilizada y de la fuente las muestras de los pacientes utilizados como Sanos generando un problema por la ambigüedad que existe en la forma en que asigna PAM50 la etiqueta, esto trae consecuencias en la toma de decisiones a nivel clínico como de terapia que le corresponde a cada paciente.

La heterogeneidad de las poblaciones de distintos tejidos celulares presentes en las muestras de cáncer de mama tiene un impacto importante en la asignación de algún subtipo tumoral utilizando un clasificador molecular y su relación con una terapia asociada.

7-Bibliografía

Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF (2009) Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. PLOS ONE 4(7): e6098. https://doi.org/10.1371/journal.pone.0006098.

Abba, M. C., Sun, H., Hawkins, K. A., Drake, J. A., Hu, Y., Nunez, M. I., … Aldaz, C. M. (2007). Breast Cancer Molecular Signatures as Determined by SAGE: Correlation with Lymph Node Status. Molecular Cancer Research: MCR, 5(9), 881–890. http://doi.org/10.1158/1541-7786.MCR-07-0055.

Ahn, J., Yuan, Y., Parmigiani, G., Suraokar, M. B., Diao, L., Wistuba, I. I., & Wang, W. (2013). DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. Bioinformatics, 29(15), 1865–1871. http://doi.org/10.1093/bioinformatics/btt301

American Cancer Society (2017). Breast Cancer Facts & Figures 2017-2018. Atlanta: American Cancer Society, Inc.

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskostky, B., Krings, G., Goga, A., Sirota, M., Butte, J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. Nature Communications. 8 (1077).

Francois Bertucci, Daniel Birnbaum. (2008). Reasons for breast cancer heterogeneity. Journal of biology, 7(2), 6.

Carvalho, B., Bengtsson, H., Speed, T., Irizarry, R. (2007) Exploration, normalization, and genotype calls of highdensity oligonucleotide SNP array data. Biostatistics, 8, 485–499.

Chin, K. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell, 10(6):529541.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research, 5(10), 2929–2943.

Elloumi, F., Hu, Z., Li, Y., Parker, J. S., Gulley, M. L., Amos, K. D., & Troester, M. A. (2011). Systematic Bias in Genomic Classification Due to Contaminating Non-neoplastic Tissue in Breast Tumor Samples. BMC Medical Genomics, 4, 54. http://doi.org/10.1186/1755-8794-4-54

Elloumi, F., Hu, Z., Li, Y., Parker, J. S., Gulley, M. L., Amos, K. D., & Troester, M. A. (2011). Systematic Bias in Genomic Classification Due to Contaminating Non-neoplastic Tissue in Breast Tumor Samples. BMC Medical Genomics, 4, 54. http://doi.org/10.1186/1755-8794-4-54

Ghosh, D., Chinnaiyan, A. (2002). Mixture modelling of gene expression data from microarray experiments, Bioinformatics, Volume 18, Issue 2, 1 February 2002, Pages 275–286.

Godino, F. (2014). Estimación de matrices de covarianza: Nuevas Perspectivas. Departamento de estadística, investigación operativa y cálculo numérico de la facultad de ciencias-UNED. España.

Graham, K. (2002). Identification of distinct gene expression signatures in histologically normal epithelium from patients with different breast cancer risk (Tesis Doctoral). Boston University, Estados Unidos.

Graham, K., De las Morenas, A., Tripathi, A., King, C., Kavanah, M., Mendez, J., Stone, M., Slama, J., Miller, M., Antoine, G., Willers, H., Sebastiani, P., Rosenberg, C. (2010). Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. Br J Cancer 102, 1284–1293, https://doi.org/10.1038/sj.bjc.6605576.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., & Sotiriou, C. (2012). A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. JNCI Journal of the National Cancer Institute, 104(4), 311–325.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., y Speed, T. P. (2003). Summaries of aymetrix genechip probe level data. Nucleic acids research, 31(4):e15-e15.

Järvstråt, L. (2017). Bioinformatic approaches to gene expression in leukemia. Networks and deconvolution Lund: Lund University, Faculty of Medicine.

Ma XJ, Salunga R, Tuggle JT, Gaudet J, Enright E, McQuary P, Payette T, Pistone M, Stecker K, Zhang BM, Zhou YX, Varnholt H, Smith B, Gadd M, Chatfield E, Kessler J, Baer TM, Erlander MG and Sgroi DC. (2003). Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci USA, 100, 10, 5974-9.

Miller, D., Smeds, J.,George, J., Vega, V., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E., Bergh, J. (2005) "An expression signature for p53 status in human breast cancer predicts mutation sta- tus, transcriptional effects, and patient survival" Proceedings of the National Academy of Sciences of the United States of America 102 (38):13550-13555.

Minn, A. J. et al. (2005). Genes that mediate breast cancer metastasis to lung. Nature, 436(7050):518-524.

Miranda, J., Bringas, R. (2008). Análisis de datos de microarreglos de ADN. Parte II: Cuantificación y análisis de la expresión génica. Biotecnología Aplicada. 25:290-300.

Nueda, Maria J., Ferrer, A, Conesa, A. (2012). “ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments”. Biostatistics, 13 (3), 553–566, https://doi.org/10.1093/biostatistics/kxr042

Peña, Daniel. (2002). Análisis de Datos Multivariante. Mc Graw-Hill. España.

Penrose, R. (1955): “A Generalized Inverse for Matrices.” Proceedings of the Cambridge Philosophical Society, Vol. 51, pp. 406 – 413.

Perou CM, Parker JS, Prat A, Ellis MJ, Bernard PS (2010). “Clinical implementation of the intrinsic subtypes of breast cancer." The lancet oncology, 11(8), 718-719.

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jerey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. (2000). “Molecular portraits of human breast tumours." Nature, 406(6797), 747-752.

Polyak K. (2011). Heterogeneity in breast cancer. The Journal of clinical investigation, 121(10), 3786-8.

Ramalle-Gómara, E. y Andrés de Llano, J.M. (2003). Utilización de métodos robustos en la estadística inferencial. Atención Primaria, Volumen 32(3), 177-182.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Salcedo, M., Vázquez, G., Hidalgo, A., Pérez, C., Piña, P., Santillán, K., Alatorre, B., Arreola, H., López, R., Montoya, C., Navarro, B., Cerón, T. (2003). Microarreglos en oncología. Revista Especializada en Ciencias de la Salud; 6(1):19-25.

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., y Quackenbush, J. (2011a). breastCancerMAINZ: Gene expression dataset published by Schmidt et al. [2008] (MAINZ). R package version 1.3.1.

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., y Quackenbush, J. (2011c). breastCancerTRANSBIG: Gene expression dataset published by Desmedt et al. [2007] (TRANSBIG). R package version 1.3.1.

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., y Quackenbush, J. (2011d). breastCancerUNT: Gene expression dataset published by Sotiriou et al. [2007] (UNT). R package version 1.3.1.

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., y Quackenbush, J. (2011e). breastCancerUPP: Gene expression dataset published by Miller et al. [2005] (UPP). R package version 1.3.1.

Schroeder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., y Quackenbush, J. (2011f). breastCancerVDX: Gene expression datasets published by Wang et al. [2005] and Minn et al. [2007] (VDX). R package version 1.3.1.

Sorlie T, Tibshirani R, Parker J, Hastie T, JS, M., Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lnning PE, Brown PO, Dale ALB and Botstein D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA, 100, 14, 8418-8423.

Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin

Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, Mauro Delorenzi, Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis, JNCI: Journal of the National Cancer Institute, Volume 98, Issue 4, 15 February 2006, Pages 262–272, https://doi.org/10.1093/jnci/djj052 Troester MA, Lee MH, Carter M, Fan C, Cowan DW, Perez ER, Pirone JR, Perou CM, Jerry DJ, Schneider SS: Activation of host wound responses in breast cancer microenvironment. Clin Cancer Res 2009, 15(22):7020-7028.

Van de Vijver, M., He, Y., van't Veer, L., Dai H, Hart A, Voskuil D, et al. (2002) A gene expression signature as a predictor of survival in breast cancer. N Engl. J Med 2002; 347:1999–2009.

Wang, N., Gong, T., Clarke, R., Chen, L., Shih, I.-M., Zhang, Z., … Wang, Y. (2015). UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. Bioinformatics, 31(1), 137–139. http://doi.org/10.1093/bioinformatics/btu607.

Weigelt B, Mackay A, A'hern R, Natrajan R, Tan DS, Dowsett M, Ashworth A, Reis-Filho JS (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol; 11: 339-349 [PMID: 20181526 DOI: 10.1016/S1470-2045(10)70008-5].

Zhaoqi, L. Xiang-Sun, Z., Shihua, Z. (2014). Breast tumor subgroups reveal diverse clinicar prognostic power. DOI: 10.1038/srep04002

8-Anexos

Anexo 1. Cuadro de valores de proporción tumoral predichos para la combinación lineal de las expresiones génicas tumorales y expresiones génicas normales simuladas

Proporción tumoral	Mínimo	Q1	Mediana	Q3	Máximo	Media	Desvio estándar
0.1	0.002	0.004	0.008	0.013	0.025	0.009	0.005
0.2	0.002	0.01	0.017	0.023	0.041	0.017	0.009
0.3	0.002	0.02	0.029	0.037	0.058	0.028	0.012
0.4	0.002	0.034	0.044	0.053	0.078	0.043	0.014
0.5	0.002	0.048	0.059	0.069	0.097	0.058	0.015
0.6	0.001	0.062	0.074	0.085	0.118	0.074	0.017
0.7	0.017	0.076	0.088	0.102	0.139	0.089	0.018
0.8	0.028	0.091	0106	0.119	0.161	0.105	0.019
0.9	0.038	0.106	0.122	0.136	0.182	0.121	0.021

Anexo 2. Clasificación de muestras luego de la corrección de la expresión génica sobre cada una de las bases simuladas proveniente de la combinación lineal de las expresiones génicas tumorales y expresiones génicas normales simuladas

Clasificación	Subtipos de cáncer				
	Basal	Her2	LumA	LumB	Normal
Inicial	121	129	194	168	138
p-0.1	138	150	171	138	153
p-0.2	127	142	177	162	142
p-0.3	119	135	181	173	142
p-0.4	116	135	178	173	148
p-0.5	118	135	180	169	148
p-0.6	121	131	189	166	143
p-0.7	117	126	199	165	143
p-0.8	122	124	199	163	142
p-0.9	122	126	200	165	137

Anexo 3. Matriz de contingencia para la clasificación de pacientes de la base de datos simulada utilizando PAM50.

Proporción tumoral propuesta	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
p-01	Basal	40	27	23	24	24
	Her2	23	52	28	27	20
	LumA	19	21	72	32	27
	LumB	14	13	27	66	18
	Normal	25	16	44	19	49
p-02	Basal	65	18	13	15	16
	Her2	14	76	18	18	16
	LumA	9	13	108	24	23
	LumB	16	13	23	97	13
	Normal	17	9	32	14	70
p-03	Basal	81	9	8	7	14
	Her2	12	93	14	11	5
	LumA	3	10	127	17	24
	LumB	11	9	20	123	10
	Normal	14	8	25	10	85
p-04	Basal	92	8	3	6	7
	Her2	10	99	13	9	4
	LumA	0	10	140	10	18
	LumB	9	7	17	136	4
	Normal	10	5	21	7	105
p-05	Basal	101	4	4	4	5
	Her2	7	109	10	6	3
	LumA	1	6	154	7	12
	LumB	5	4	9	148	3
	Normal	7	6	17	3	115
p-06	Basal	108	5	1	3	4
	Her2	4	113	6	6	2
	LumA	0	4	168	8	9
	LumB	5	2	6	149	4
	Normal	4	5	13	2	119
p-07	Basal	109	5	0	2	1
	Her2	4	114	3	4	1
	LumA	1	5	178	7	8
	LumB	4	2	3	154	2
	Normal	3	3	10	1	126
p-08	Basal	112	5	1	3	1
	Her2	3	115	3	2	1
	LumA	1	3	182	5	8
	LumB	2	3	1	157	0
	Normal	3	3	7	1	128
p-09	Basal	118	1	0	2	1
	Her2	1	122	1	2	0
	LumA	1	2	190	3	4
	LumB	1	2	1	161	0
	Normal	0	2	2	0	133

Anexo 4. Medidas de desempeño del clasificador para la combinación lineal de las expresiones génicas tumorales y expresiones génicas normales simuladas

Proporción tumoral propuesta	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
0.1	75	84	37	51	84	37	51
0.2	82	89	55	68	88	55	68
0.3	87	92	68	78	92	68	78
0.4	91	94	76	84	94	76	84
0.5	93	96	84	89	96	84	89
0.6	95	97	88	92	97	88	92
0.7	96	98	91	94	98	90	94
0.8	97	98	93	95	98	92	95
0.9	99	99	97	98	99	96	98

Anexo 5. Análisis de las muestras de la base de datos VDX

En la figura 16 se observa la distribución de los valores predichos para la proporción tumoral para cada alternativa de datos utilizados en el algoritmo y se logra observar una alta variabilidad para los valores obtenidos con las expresiones génicas de las muestras Sanos mientras que para Normal-Like los valores de proporción se concentran alrededor de valores como 0.5 y 0.6 con una variabilidad menor.

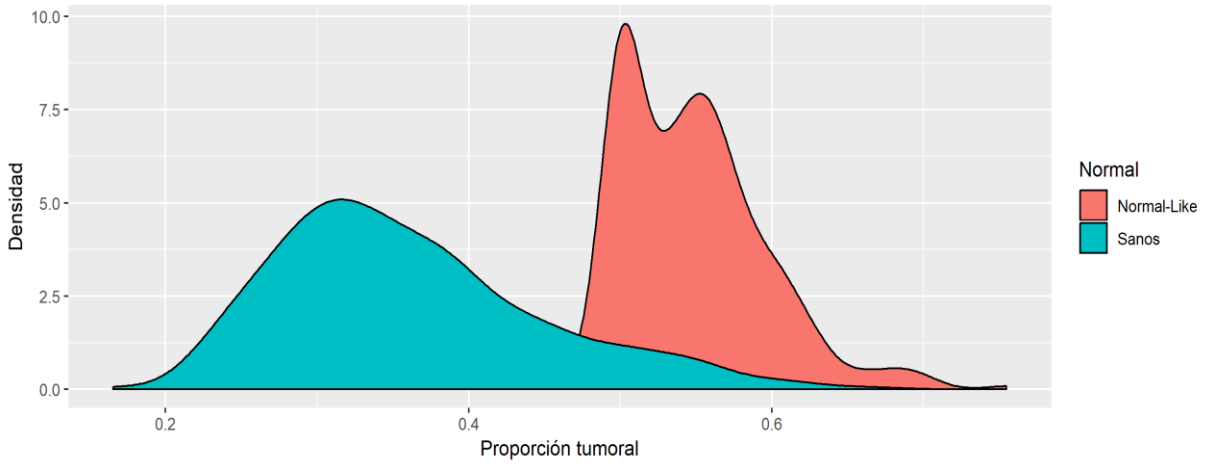

Figura 16. Función de densidad para la proporción tumoral predicha utilizando distintas alternativas de expresiones génicas Normales.

El cuadro 15 muestra las clasificaciones de las muestras dependiendo de qué alternativa de expresiones génicas se utilizaron como muestras Normales. Las clases Basal, Her2 y Normal su cambio siempre fue en una misma dirección, pero en distinta magnitud y para las clases LumA y LumB su comportamiento fue distinto en dirección.

Cuadro 15. Clasificación de las muestras utilizando PAM50 para la base de datos VDX

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal-Like
Inicial	91 (26.5)	51 (14.8)	108 (31.4)	67 (19.5)	27 (7.8)
Sanos	93 (27.0)	52 (15.1)	118 (34.3)	57 (16.6)	24 (7.0)
Normal-Like	92 (26.7)	59 (17.2)	95 (27.6)	73 (21.2)	25 (7.3)

Al utilizar las muestras de sanos se obtienen mejores resultados al reclasificar tanto para detectar verdaderos positivos como verdaderos negativos en comparación con Normal-Like, donde se obtuvieron valores más bajos, mostrando que hubo un cambio en la clasificación de las muestras con respecto a la clasificación inicial dada por PAM50 (Cuadro 16).

Cuadro 16. Medidas de desempeño del clasificador luego de la corrección de la matriz de expresión génica

Clasificación	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	98	99	95	97	99	93	96
Normal-Like	97	98	93	95	98	92	95

Utilizando ambas alternativas propuestas como normales, se obtuvo un cambio en la reclasificación de las etiquetas de las muestras luego de haber corregido la base de datos con el valor de proporción estimado por el algoritmo. Como se observa en el cuadro 17, dos muestras nuevas pasaron de ser Her2 y Normal para ser clasificados como Basal, en comparación con Her2 que tuvo un aumento de una muestra. Sin embargo, una muestra que inicialmente fue clasificada como Her2 pasó a ser Basal y dos LumB pasaron a ser Her2. La clase que más sufrió movimiento fue LumA, con un aumento de diez muestras, ocho provenientes de LumB y tres de Normales y uno que inicialmente era LumA es ahora Normal y las clases LumB y Normal sufrieron una disminución de diez y tres muestras respectivamente. Al utilizar la alternativa de Normal-Like, el mayor cambio se dio para la clase LumA al disminuir su clasificación inicial de 108 muestras a 95, y LumB que sufrió un aumento de seis muestras donde la mayoría inicialmente fueron LumA, sin embargo, tres muestras que inicialmente fueron etiquetados como LumB pasaron a ser Her2. En promedio 4 muestras para ambas situaciones cambió de etiquetas

Normales a otros subtipos de mayor riesgo y hubo nueve muestras que se reclasificaron como LumB.

Cuadro 17. Matriz de contingencia para la clasificación inicial y reclasificación de muestras de la base VDX utilizando distintas alternativas para expresiones génicas Normales.

Tipo de Normal	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	91	1	0	0	1
	Her2	0	50	0	2	0
	LumA	0	0	107	8	3
	LumB	0	0	0	57	0
	Normal	0	0	1	0	23
Normal-Like	Basal	90	0	0	0	2
	Her2	1	50	3	3	2
	LumA	0	0	94	0	1
	LumB	0	1	8	64	0
	Normal	0	0	3	0	22

Independientemente de la fuente que se utilizó como datos Normales, el comportamiento de las correlaciones de Spearman con las que se asignó la etiqueta en PAM50 para las muestras que se mantuvieron en la diagonal fue muy similar para cada una de las clases y se mantuvo entre 0.35 y 0.75 dependiendo de la clase (Figura 16).

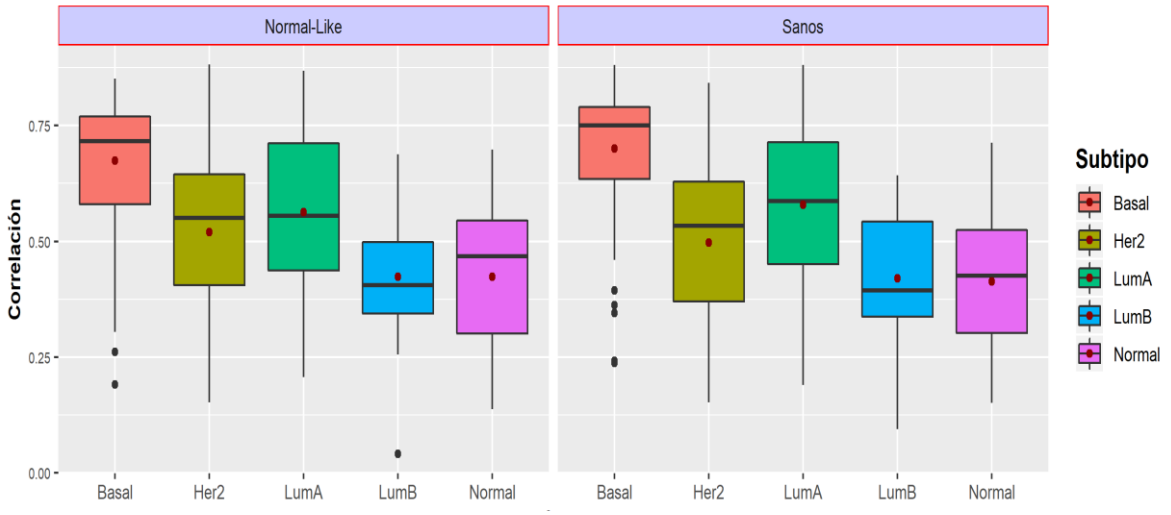

Figura 16. Diagrama de cajas para las correlaciones de Spearman dadas por PAM50 de las muestras que no cambiaron su clasificación con respecto a la inicial para la base VDX.

La figura 17 muestra el diagrama de las correlaciones de Spearman con las que se asignó una etiqueta a las muestras utilizando PAM50 fue muy baja independiente de la fuente de datos Normales.

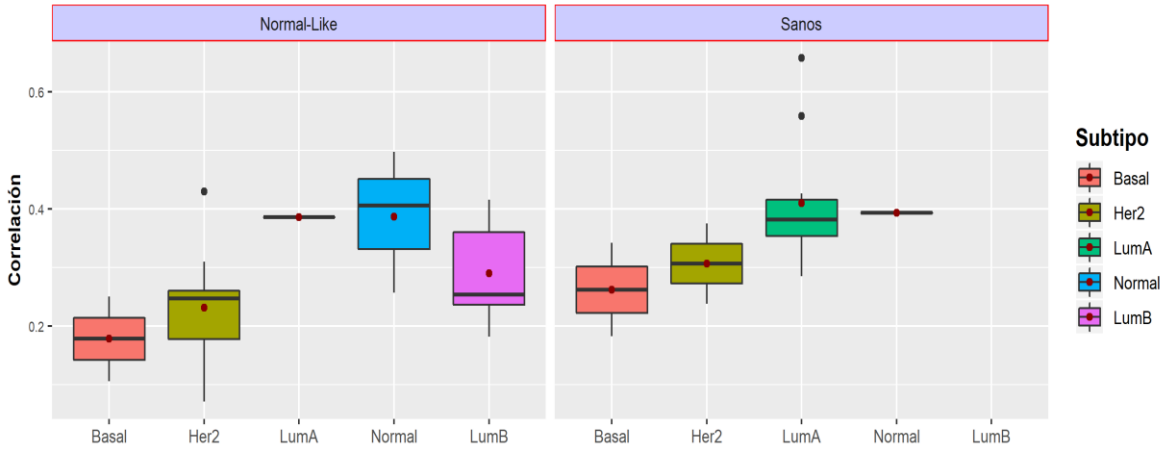

Figura 17. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que cambiaron su clasificación con respecto a la inicial para la base VDX.

Anexo 5. Análisis de las muestras de la base de datos TransBig

Existe una amplia diferencia para las distribuciones de los valores tumorales predichos detectados por el algoritmo propuesto para las muestras de la base de datos TransBig. En la figura 18 se observa como los valores de proporción tumoral estimados usando las muestras de Normal-Like son mucho menos variables que los estimados usando Sanos, pero con una mayor concentración alrededor de 0.5 de proporción tumoral y en sanos los valores oscilan entre 0.2 y 0.6 con algunos valores extremos.

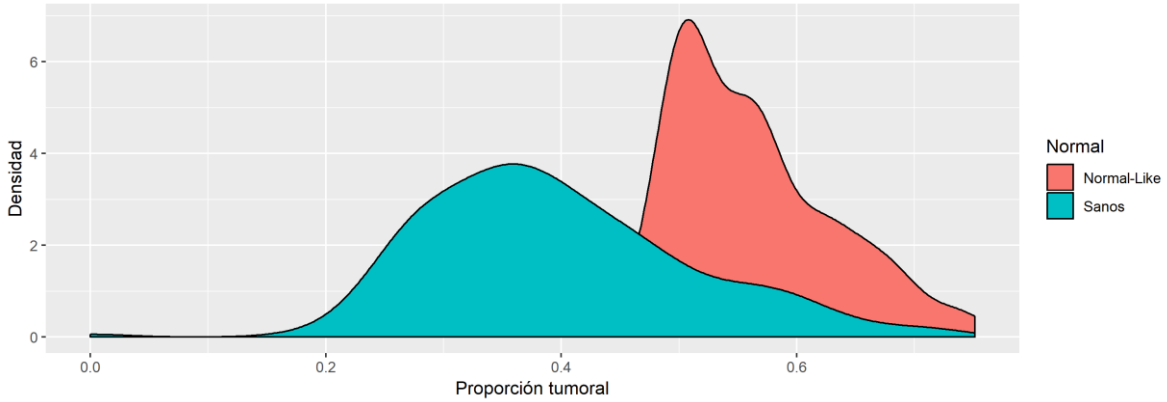

Figura 18. Función de densidad para la proporción tumoral utilizando distintas alternativas de pacientes sanos.

Para las clases Basal y Her2 independientemente de que naturaleza sean expresiones génicas normales, su cambio en la clasificación siempre fue un aumento y para las clases LumA, LumB y Normal su comportamiento fue distinto en dirección. El cuadro 18 muestra la clasificación de las muestras asignadas a las diferentes clases tumorales.

Cuadro 18. Clasificación de muestras utilizando PAM50 para la base de datos TransBig

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal
Inicial	46 (23.2)	26 (13.1)	66 (33.3)	37 (18.7)	23 (11.6)
Sanos	46 (23.2)	28 (14.1)	70 (35.4)	35 (17.7)	19 (09.6)
Normal-Like	45 (22.7)	30 (15.2)	60 (30.3)	42 (21.2)	21(10.6)

El cuadro 19 muestra en las medidas de desempeño obtenidas según la fuente de datos Normales utilizada, obteniendo que la detección de verdaderos positivos y verdaderos negativos fueron muy similares para ambas situaciones indicándonos que no hubo un cambio significativo en la clasificación inicial con respecto a la reclasificación.

Cuadro 19. Medidas de desempeño del clasificador luego de la corrección de las expresiones génicas de las muestras de TransBig

Clasificación	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	97	98	93	96	98	92	95
Normal-Like	97	98	93	95	98	93	96

Al utilizar las muestras de sanos, el mayor cambio en la reclasificación se dio en las muestras Normales, ya que al inicio había veintitrés muestras, luego de la reclasificación cuatro de ellos fueron asignados como LumA y uno como Normal y hubo una nueva muestra Normal que inicialmente fue LumA. En cuanto al utilizar las expresiones génicas de Normal-Like, el cambio de etiquetas en las muestras fue similar en la mayoría de las clases siendo LumA el que presentó una disminución de muestras clasificadas con respecto al inicio, ya que cuatro muestras se reclasificaron como LumB, dos muestras como Normales y uno como Her2 (Cuadro 20).

Cuadro 20. Matriz de contingencia para la clasificación inicial y reclasificación de muestras de la base TransBig utilizando distintas alternativas para pacientes Normales.

	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	45	0	0	0	1
	Her2	1	25	0	2	0
	LumA	0	0	64	2	4
	LumB	0	1	1	33	0
	Normal	0	0	1	0	18
Normal-Like	Basal	43	0	0	0	2
	Her2	3	26	1	0	0
	LumA	0	0	59	0	1
	LumB	0	0	4	37	1
	Normal	0	0	2	0	19

Las correlaciones de Spearman con las que se le asignó un subtipo tumoral a cada uno de las muestras reclasificadas que no cambiaron su clasificación con respecto a la inicial, independientemente de la alternativa de datos que se utilizó como Normales fue de 0.5 en promedio, solamente LumB fue asignado con una correlación de 0.35 como promedio de todos los pacientes. La figura 18 muestra el diagrama de cajas para las correlaciones de Spearman con las que PAM50 asigno una muestra a una clase a los pacientes que no cambiaron con respecto a la clasificación inicial.

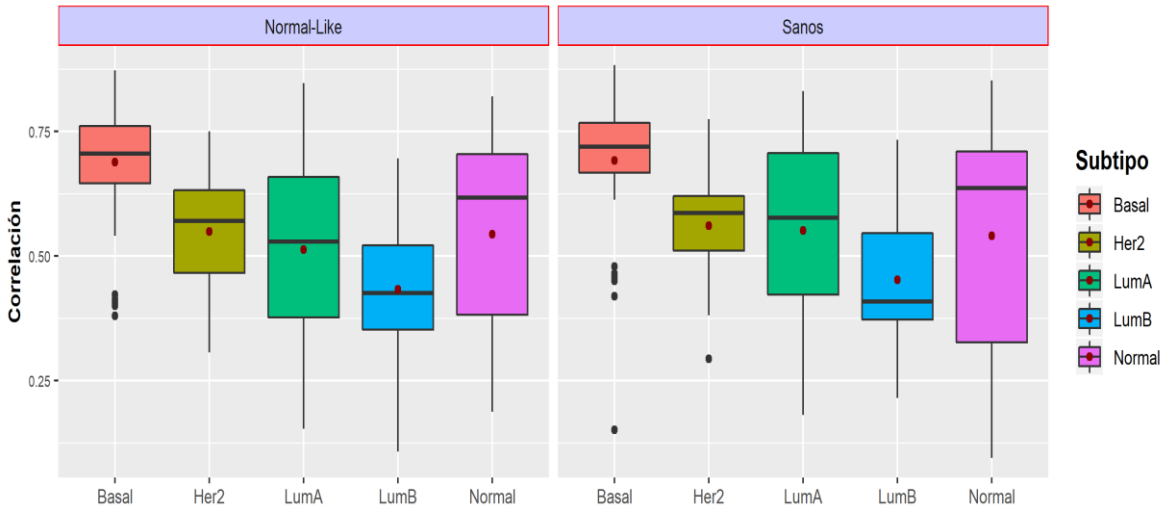

Figura 19. Diagrama de cajas para las correlaciones de Spearman dadas por PAM50 de las muestras que no cambiaron su clasificación con respecto a la inicial para la base TransBig.

Como se muestra en la figura 20, las muestras que cambiaron de clasificación con respecto a la inicial, PAM50 asignó a una nueva clase tumoral con correlaciones de Spearman de 0.3 en promedio. Hubo un caso en una muestra que fue asignado como Normal con una correlación menor a 0.1.

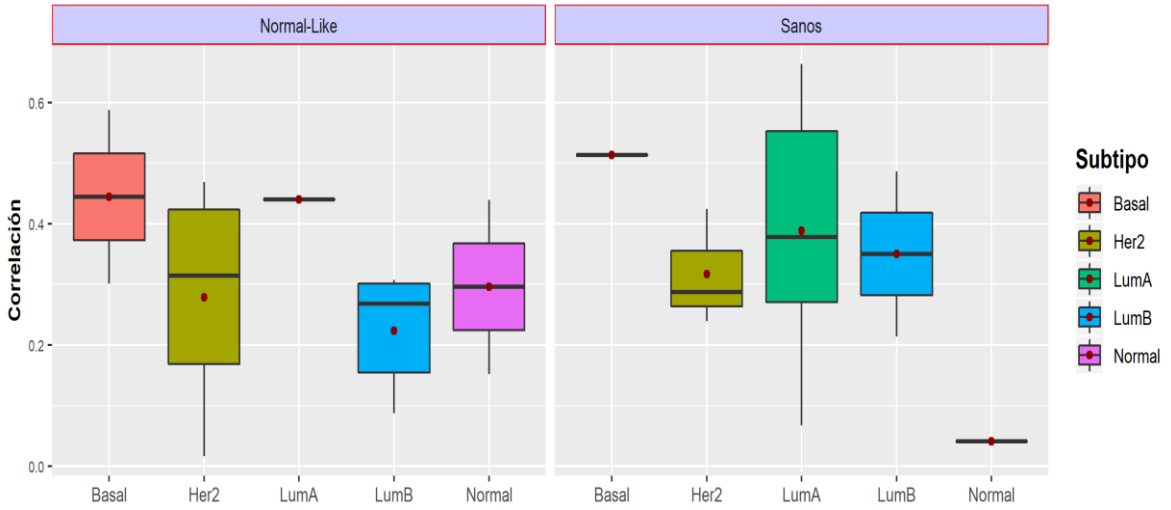

Figura 20. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que cambiaron su clasificación con respecto a la inicial para la base TransBig.

Anexo 6. Análisis de las muestras de la base de datos UPP

Existen distribuciones muy marcadas para las proporciones tumorales estimadas para cada una de las alternativas utilizadas como muestras normales en el algoritmo para las muestras de UPP (ver figura 21). Al utilizar sanos se puede observar una amplia variación detectándose valores que oscilan desde 0.10 hasta 0.45, viéndose una pequeña concentración de datos alrededor de 0.23 mientras que para Normal-Like la variación es menor, pero con dos grupos marcados muy claramente, el de mayor concentración de muestras se da alrededor de 0.50 mientras que para el otro grupo se da para un valor de 0.56 de proporción tumoral.

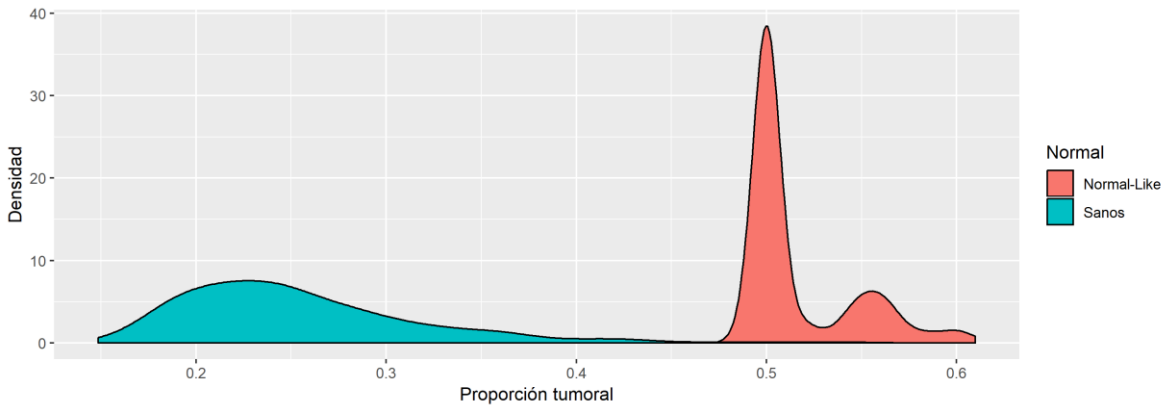

Figura 21. Función de densidad para la proporción tumoral utilizando distintas alternativas de expresiones génicas normales.

El cuadro 21 muestra la clasificación de muestras para las clases Basal y Normal, el utilizar Sanos o Normal Like no genera un cambio en la cantidad de muestras a como si lo fue para las clases Her2, LumA y LumB.

Cuadro 21. Clasificación de muestras utilizando PAM50 para las muestras de la base de datos UPP.

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal
Inicial	34 (13.5)	49 (19.5)	69 (27.5)	53 (21.1)	46 (18.3)
Sanos	34 (13.5)	48 (19.1)	73 (29.1)	50 (19.9)	46 (18.3)
Normal-Like	34 (13.5)	55 (21.9)	66 (26.3)	49 (19.5)	47 (18.7)

En el cuadro 22 de medidas de desempeño del clasificador luego de la corrección de la matriz de expresión génica, para la de detección de verdaderos positivos se obtienen valores bajos y entre ambas condiciones existe un mayor movimiento de pacientes utilizando las muestras de Normal-Like ya que los valores de sensibilidad son menores que los obtenidos con la alternativa de Sanos.

Cuadro 22. Medidas de desempeño del clasificador luego de la corrección de los datos

Clasificación	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	98	99	96	97	99	96	97
Normal-Like	96	98	91	94	98	91	94

A pesar de que no hubo un cambio muy marcado en la cantidad de muestras clasificadas para las clases Normal y Basal (Cuadro 23), se logra determinar que si hubo cambios en la asignación de etiquetas de las muestras. Al utilizar Sanos se ve como en las clases mencionadas anteriormente, dos muestras cambiaron de Normal a LumA y Basal y una muestra de Basal paso a ser Her2 y para el caso de utilizar las muestras de Normal-Like, la clase Normal hubo siete muestras que cambiaron de etiqueta mientras que ocho cambiaron a ser Normal. Para el resto de las clases en ambas alternativas, LumB fue el subtupo que tuvo una reasignación de etiquetas mayormente donde al menos cuatro muestras cambiaron a otra clase.

Cuadro 23. Matriz de contingencia para la clasificación inicial y reclasificación de muestras de la base UPP utilizando distintas alternativas para muestras Normales.

	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	33	0	0	0	1
	Her2	1	47	0	0	0
	LumA	0	1	67	4	1
	LumB	0	1	0	49	0
	Normal	0	0	2	0	44
Normal-Like	Basal	32	0	0	0	2
	Her2	2	49	0	4	0
	LumA	0	0	60	1	5
	LumB	0	0	1	48	0
	Normal	0	0	8	0	39

En la figura 22 se muestran los diagramas de cajas de las correlaciones de Spearman con las que PAM50 asigna una etiqueta a una muestra y se visualiza que independientemente de las expresiones génicas utilizadas como alternativa a Normal, las correlaciones de Spearman obtenidas a las muestras que no cambiaron su condición de subtipo estuvieron en entre 0.4 y 0.6 promedio, considerándose correlaciones no muy fuertes.

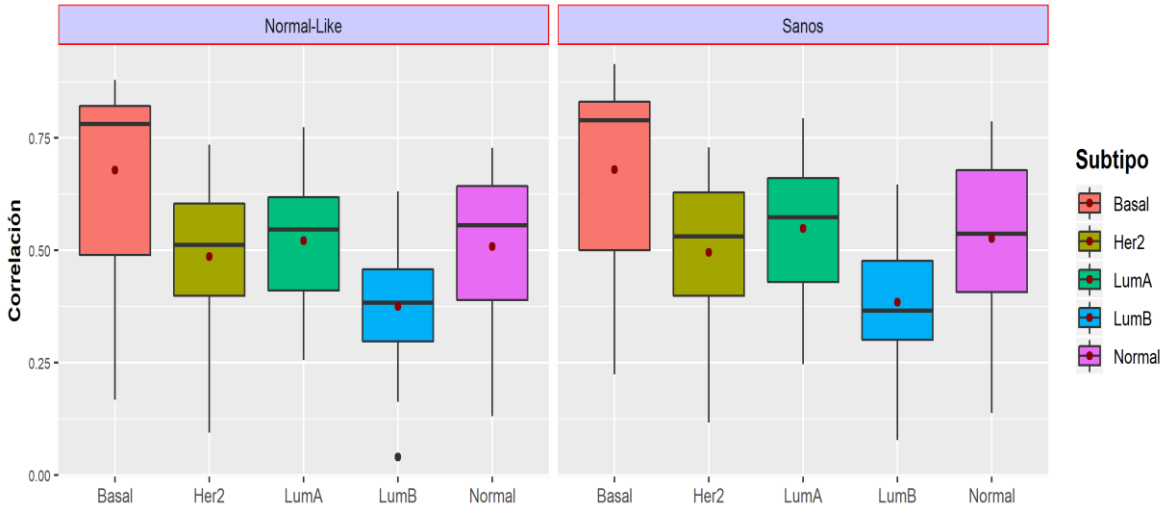

Figura 22. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que no cambiaron su clasificación con respecto a la inicial para la base UPP.

Aquellas muestras que cambiaron su etiqueta luego de la corrección de la expresión génica, las correlaciones de Spearman para la asignación de la nueva etiqueta (figura 23) fueron en su mayoría correlaciones por debajo de 0.4. Cuando se utilizó las expresiones génicas de las muestras Sanos, solamente la clase LumA tubo más de una muestra que cambio de etiqueta.

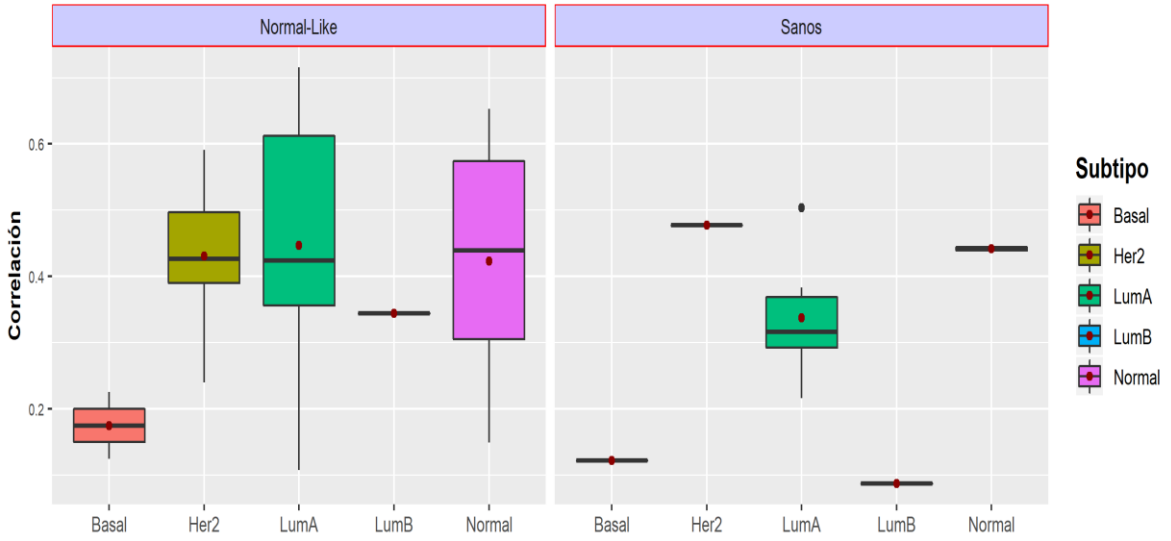

Figura 23. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que cambiaron su clasificación con respecto a la inicial para la base UPP.

Anexo 7. Análisis de las muestras de la base de datos Mainz

Con respecto a las muestras de la base de datos Mainz, la variación en las proporciones tumorales estimadas es mayor cuando se utiliza las muestras de los pacientes Sanos en comparación con las muestras clasificadas como Normal-Like. Existen valores de proporción altos haciendo que la distribución tenga una cola más pesada a la derecha (figura 24). Para los datos de Normal-Like se observan dos grupos importantes, uno con mayor cantidad de muestras que presentan en promedio valores de proporción de 0.5 aproximadamente y otro grupo con menor cantidad de muestras que presenta en promedio 0.65 de proporción tumoral detectada.

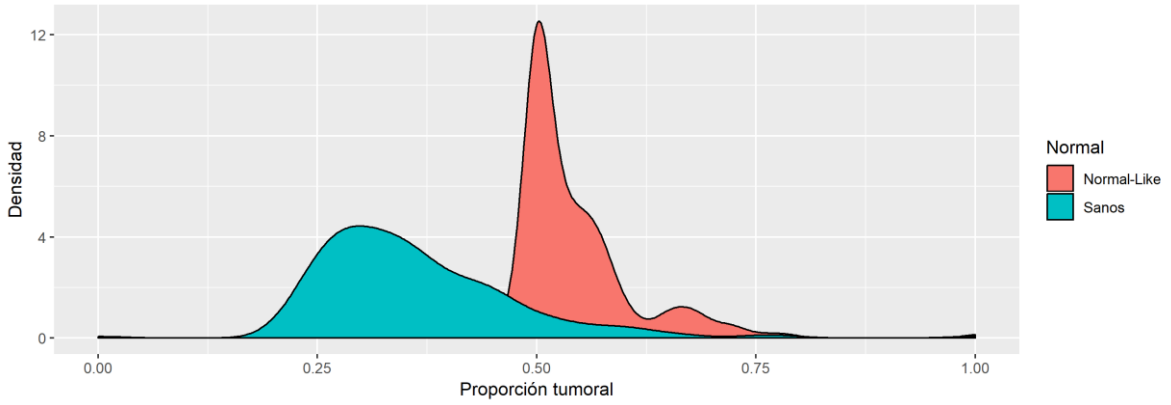

Figura 24. Densidad para la proporción tumoral utilizando distintas fuentes de pacientes sanos.

En el cuadro 24 de la clasificación de los muestras, la reasignación de muestras utilizando distintas alternativas como pacientes Normales hace que se comporte muy similar, manteniéndose el porcentaje sin cambios evidentes, la única clase que tuvo un cambio fue LumA y Normal al utilizar Normal-Like.

Cuadro 24. Clasificación de muestras utilizando PAM50 para la base de datos Mainz

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal
Inicial	33 (16.5)	28 (14.0)	57 (28.5)	49 (24.5)	33 (16.5)
Sanos	32 (16.5)	29 (14.5)	57 (28.5)	49 (25.5)	33 (16.5)

Normal-Like	35 (17.5)	30 (15.0)	51 (25.5)	48 (24.0)	36 (18.0)

En relación con las medidas de desempeño reflejadas en el cuadro 25 nos reafirma lo anterior, obteniendo resultados de muy poco cambio, sin embargo, los valores para Normal-Like son más bajos que los Sanos, evidenciando que existe un ligero cambio en las etiquetas de las muestras con relación a la obtenida inicialmente.

Cuadro 25. Medidas de desempeño del clasificador luego de la corrección la matriz de expresiones génicas

Clasificación	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	98	99	95	97	99	94	96
Normal-Like	97	98	93	96	98	94	96

Si bien la cantidad de muestras se mantiene similar con respecto a la clasificación inicial, para ciertas clases tumorales las muestras sufrieron un cambio en el subtipo de cáncer que inicialmente se les asignó. La matriz de contingencia del cuadro 26 refleja un cambio importante que fue la disminución de muestras de LumB y el aumento de muestras Normales

Cuadro 26. Matriz de contingencia para la clasificación inicial y reclasificación de muestras de la base Mainz utilizando distintas fuentes para pacientes Normales.

	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	31	0	0	0	1
	Her2	1	27	0	1	0
	LumA	0	0	54	1	1
	LumB	0	1	1	47	0
	Normal	1	0	2	0	30
Normal-Like	Basal	33	0	0	1	1
	Her2	0	27	0	3	0
	LumA	0	0	50	0	1
	LumB	0	1	2	45	0
	Normal	0	0	5	0	31

La figura 25 muestra el diagrama de cajas para las correlaciones con las que se asignaron las etiquetas de los subtipos tumorales utilizando PAM50. Se obtuvo valores promedio de 0.5 hacia arriba para ambas situaciones (Sanos y Normal-Like).

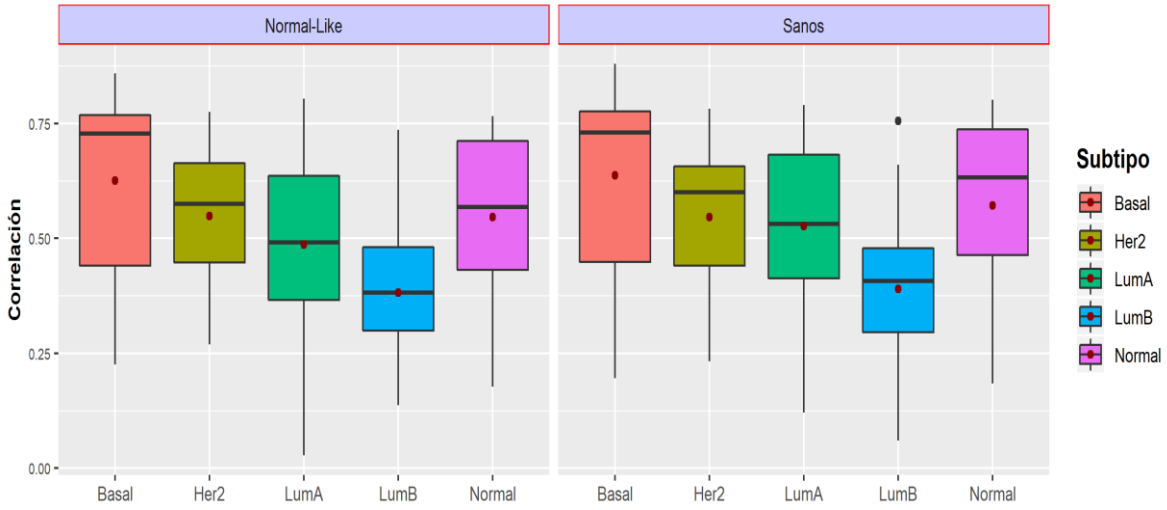

Figura 25. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que no cambiaron su clasificación con respecto a la inicial para la base Mainz.

Existe una diferencia entre las correlaciones que se obtuvieron utilizando la condición de Normal-Like y Sanos (figura 26), para la primera condición las correlaciones fueron entre 0.2 y 0.4 en promedio, siendo estas consideradas muy

bajas, mientras que para Sanos, las correlaciones estuvieron en 0.4 promedio para todos los subtipos tumorales.

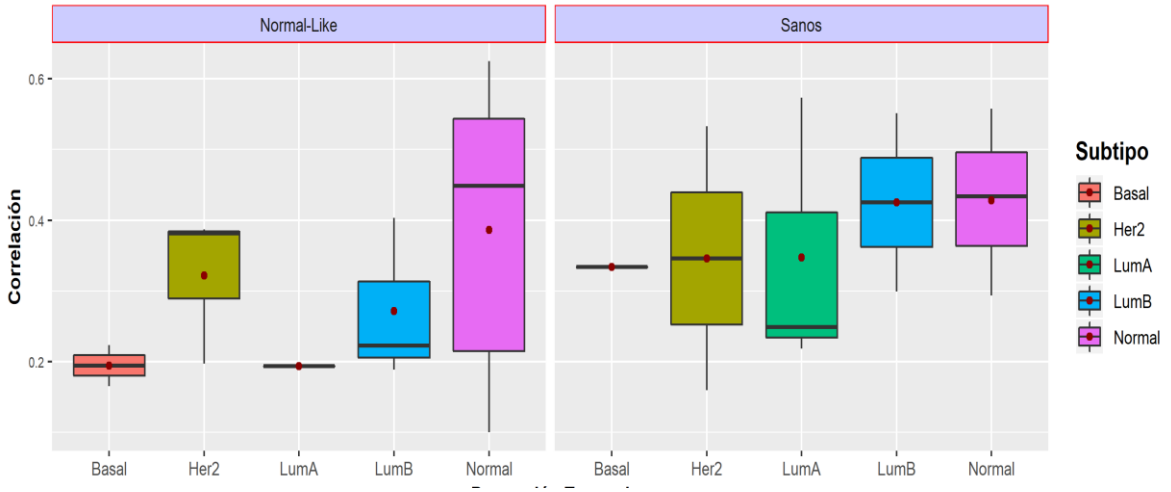

Figura 26. Diagrama de cajas para las correlaciones dadas por PAM50 de las muestras que cambiaron su clasificación con respecto a la inicial para la base Mainz.

Anexo 8. Análisis de las muestras de la base de datos UNT

Las distribuciones para las proporciones tumorales estimadas para cada una de las muestras de la base de datos UNT bajo cada una de las condiciones utilizadas en el algoritmo mostraron comportamientos totalmente distintos. En la figura 26 se puede observar que al utilizar expresiones génicas de muestras sanos se puede observar una amplia variación viéndose una pequeña concentración de datos alrededor de 0.20 mientras que utilizar Normal-Like la variación es menor, pero con dos grupos marcados muy claramente, el de mayor concentración de muestras se da alrededor de 0.50 mientras que para el otro grupo se da para un valor de 0.55 de proporción tumoral.

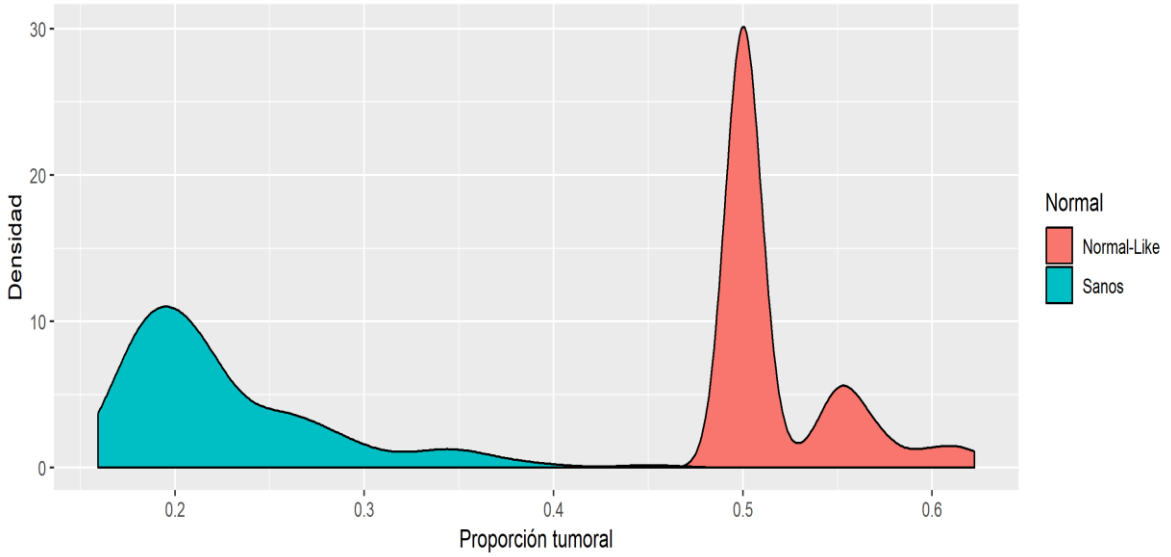

Figura 28. Densidad para la proporción tumoral utilizando distintas fuentes de muestras de pacientes sanos.

El cuadro 27 se tiene la clasificación de muestras utilizando PAM50 para la base de datos UNT. Independientemente de la alternativa de muestras Normales el cambio en la asignación de las etiquetas se mantiene muy similar para todas las clases.

Cuadro 27. Clasificación de las muestras utilizando PAM50 para la base de datos UNT.

Clasificación	Subtipos de cáncer (%)				
	Basal	Her2	LumA	LumB	Normal
Inicial	22 (16.1)	22 (16.1)	40 (29.2)	27 (19.7)	26 (19.0)
Sanos	22 (16.1)	23 (16.8)	39 (28.5)	26 (19.0)	27 (19.7)
Normal-Like	22 (16.1)	23 (16.8)	39 (28.5)	26 (19.0)	27 (19.7)

Las medidas de desempeño obtenidas luego de la reclasificación reflejadas en el cuadro 28, se puede notar que los cambios ocurridos fueron los mismo para ambas bases, dado que se obtuvieron valores iguales.

Cuadro 28. Medidas de desempeño del clasificador luego de la corrección de la matriz de expresiones génicas

Clasificación	Medidas de desempeño						
	Exactitud promedio	Micro precisión	Micro sensibilidad	Micro f-score	Macro precisión	Macro sensibilidad	Macro f-score
Sanos	99	99	98	99	99	98	99
Normal-Like	99	99	98	99	99	98	99

El cuadro 29 tiene la matriz de contingencia para la clasificación inicial y reclasificación de muestras, observándose cambios de etiquetas de muestras en el mismo paciente independientemente de la alternativa de las expresiones génicas como normales. Inicialmente tuvo la asignación como LumB y luego de la reclasificación pasó a ser Basal.

Cuadro 29. Matriz de contingencia para la clasificación inicial y reclasificación de pacientes de la base UNT utilizando distintas fuentes para pacientes Normales.

	Reclasificado	Inicial				
		Basal	Her2	LumA	LumB	Normal
Sano	Basal	21	0	0	1	0
	Her2	1	22	0	0	0
	LumA	0	0	39	0	0
	LumB	0	0	0	26	0
	Normal	0	0	1	0	26
Normal-Like	Basal	21	0	0	1	0
	Her2	1	22	0	0	0
	LumA	0	0	39	0	0
	LumB	0	0	0	26	0
	Normal	0	0	1	0	26

Las muestras que no cambiaron de subtipo a pesar de aplicar una corrección en sus expresiones génicas utilizando las diferentes fuentes de pacientes Normales tuvo valores de correlación de Spearman entre 0.5 y 0.75 comportándose de manera muy similar según las muestras utilizadas como pacientes Normales, la cual se puede observar en la figura 28 donde se tienen los diagramas de cajas para las correlaciones dadas por PAM50.

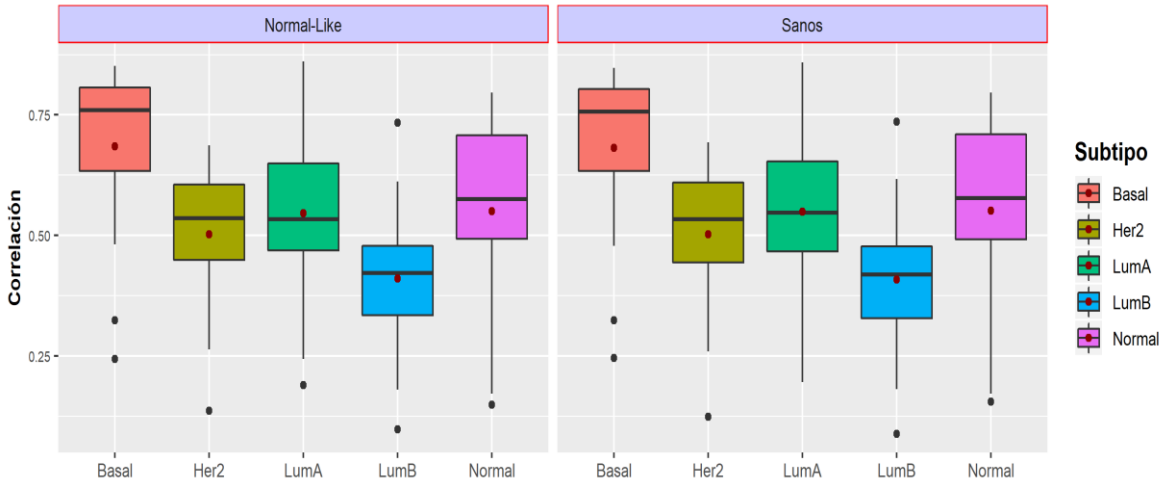

Figura 29. Diagrama de cajas para las correlaciones de Spearman dadas por PAM50 a las muestras que no cambiaron su clasificación con respecto a la inicial para la base UNT.

En la figura 30 se tienen los diagramas de cajas para las correlaciones obtenidas para aquellas muestras que cambiaron su clasificación con respecto al inicial luego de haber corregido su expresión eliminando el contenido proporción de tejido normal estimado bajo las diferentes fuentes de pacientes Normales, la asignación de las etiquetas tuvo una correlación no mayor a 0.4 y por arriba de 0.1.

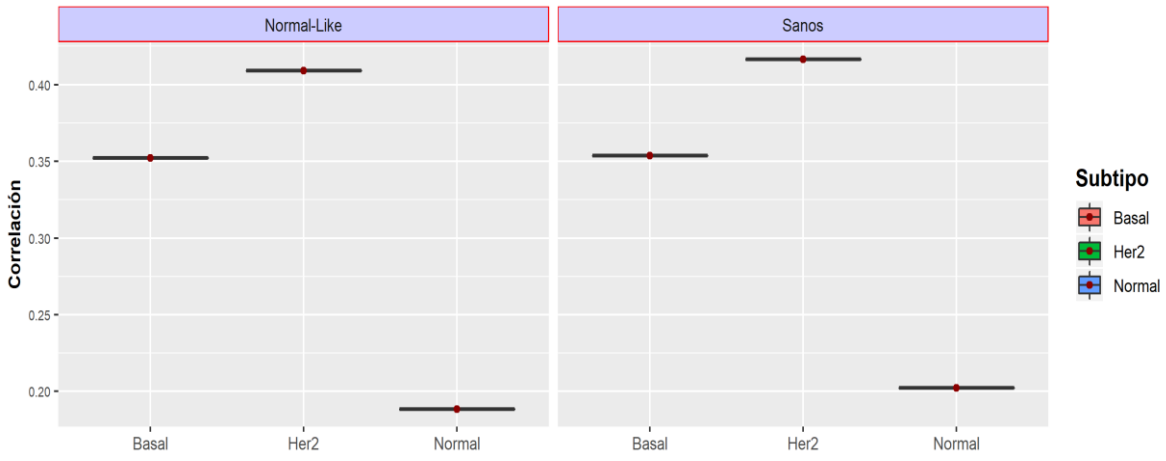

Figura 30. Diagrama de cajas para las correlaciones de Spearman dadas por PAM50 de las muestras que cambiaron su clasificación con respecto a la inicial para la base UNT.

Anexo 9. Cambio de muestras para las cinco bases de datos en dirección y magnitud. Están acomodadas según fueron analizadas en los resultados.