VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS
1218 ~ 2018

# CENTRO DE INVESTIGACIÓN DEL CÁNCER

# FACVLTAD DE CIENCIAS

# GRADO DE ESTADÍSTICA

Trabajo de fin de Grado

**Flujos de trabajo sistemáticos para el diseño y análisis computacional de microarrays de proteínas.**

Autora: Helena Fidalgo Gómez

Tutores: Dr. Manuel Fuentes García

Dr. José Manuel Sánchez Santos

CiC
CENTRO DE INVESTIGACIÓN DEL CÁNCER

*Salamanca, 2019*

CENTRO DE INVESTIGACIÓN DEL CÁNCER

FACULTAD DE CIENCIAS

GRADO DE ESTADÍSTICA

Trabajo de fin de Grado

# Flujos de trabajo sistemáticos para el diseño y análisis computacional de microarrays de proteínas.

Autora: Helena Fidalgo Gómez

Tutores: Dr. Manuel Fuentes García

José Manuel Sánchez Santos

Dr. Manuel Fuentes García

Dr. José Manuel Sánchez Santos

Helena Fidalgo Gómez

*Salamanca, 2019*

# *Índice:*

En cumplimiento a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (LOPD), no se pueden divulgar las bases de datos empleadas en el presente trabajo, debido a que estos pertenecen a pacientes del Hospital Clínico de Salamanca.

A su vez, parte del contenido de este trabajo está relacionado con el Proyecto de Investigación: *FIS17/ 01930*, elaborado y estudiado en el Laboratorio de Proteómica y Genómica del Centro de Investigación del Cáncer, por tanto, todos los derechos y toda propiedad intelectual e industrial corresponden a este laboratorio.

# *1. Introducción:*

La Proteómica, se define como el conjunto de las diferentes técnicas relacionadas por su disposición de aportar información sobre el proteoma, lo que comprende desde las identidades de las proteínas constituyentes, especificar la función de una proteína determinada, hasta su ubicación intracelular (Cazzulo, 2014). La técnica que se emplea para estudiar la composición del proteoma se denomina perfil proteico y existen dos métodos para poder obtener este perfil proteico. Ambos se basan en separar las proteínas de su proteoma: la electroforesis de proteínas y la espectrometría de masas (Lovric, 2011).

Actualmente existe una nueva alternativa para diseccionar el proteoma: la tecnología del microarray, concretamente los microarrays de proteínas. La tecnología del microarray representa una nueva herramienta para abordar estudios biológicos de alto rendimiento y se basa en técnicas conocidas como de alta capacidad de trabajo *(high-throughput).* Al analizarse cantidades masivas de datos, donde se deben de ajustar las fuentes de variabilidad con el fin de identificar las principales proteínas involucradas entre las muchas otras que forman la base de datos inicial, la Estadística es una herramienta crucial dentro de la tecnología del microarray (Valledor & Meijón, 2014).

## *1.1. Introducción al Proteoma:*

Todo organismo tiene su propio genoma, el cual contiene la información biológica necesaria para mantener vivo a ese organismo. La palabra "genoma" procede de las palabras: "Genes" y "Cromosomas"; se define como la mezcla completa de cromosomas y sus determinados genes en una especie biológica concreta (Cazzulo, 2014).

Se denomina Genotipo al conjunto de toda la información genética de un organismo, tanto unicelular como pluricelular. El ácido desoxirribonucleico *(ADN)* contiene toda esta información genética y es quien la transmite de generación en generación. El ADN está compuesto por cuatro moléculas llamadas nucleótidos hechos a base de Adenina, Timina, Citosina y Guanina (también conocidas como bases nitrogenadas), que se enlazan unas con otras formando una larga cadena. El ADN está formado por dos de estas cadenas que se unen de manera complementaria por enlaces electrostáticos. La Adenina se une con la Timina y la Guanina con la Citosina, formando así los pares de bases (el genoma humano mide aproximadamente 3.200.000.000 de pares de bases y está dividido en 23 porciones más pequeñas llamadas cromosomas). Un gen es un trozo de genoma que se ha comprobado que desempeña cierta función en la vida de la célula. El uso de esta información contenida en el genoma, requiere de la actividad coordinada de enzimas y proteínas, que participan en una compleja serie de reacciones bioquímicas conocida como expresión del genoma (Brown, 2008).

Por tanto, se entiende por “Genómica” al estudio completo del genoma en diferentes organismos. Se basa en el análisis de la secuencia del ADN y la ubicación de los genes secuenciados en diferentes cromosomas. Para que la célula utilice la información biológica que existe dentro de su genoma, se deben expresar de forma coordinada grupos de genes, en los que cada gen representa una sola unidad de información (Cazzulo, 2014).

El producto inicial de la expresión del genoma se conoce como transcriptoma (un conjunto de moléculas de ácido ribonucleico *(ARN)* derivadas de los genes que codifican las proteínas). El segundo producto de la expresión del genoma es el Proteoma, el conjunto de proteínas que componen la célula y que determina las reacciones bioquímicas que puede llevar a cabo esa determinada célula**.** Las proteínas que forman el proteoma son sintetizadas por traducción de las moléculas individuales de ARN presentes en el transcriptoma (Figura A).

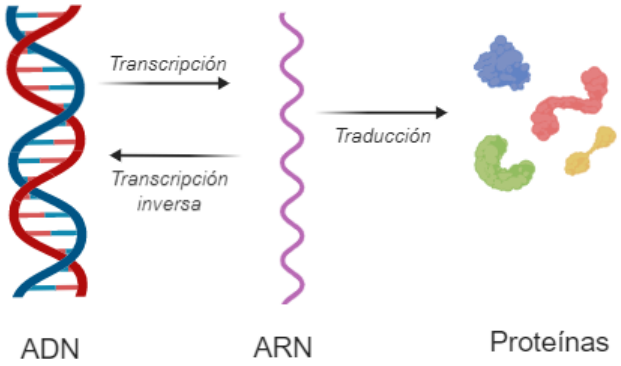



*Figura A. Ejemplo de Dogma Central de la Biología, de DNA a RNA a Proteína. Imagen creada con BioRender.com*

El hecho de estudiar miles de secuencias de ARN en una célula, es decir, el transcriptoma, permite determinar cuándo y dónde está activado o desactivado cada uno de los genes que componen las células y tejidos de un determinado organismo (Tyers & Mann, 2003).

Existen varias clases de ARN y entre todas ellas destaca el ARN mensajero *(ARNm)*, el cuál desempeña un papel crucial en la elaboración de proteínas. En este proceso el ARNm se transcribe a partir de genes; luego, los transcritos de ARNm se entregan a los ribosomas (las máquinas moleculares ubicadas en el citoplasma de la célula); estos "traducen" la secuencia de las letras químicas (nucleótidos) en el ARNm y ensamblan componentes básicos llamados aminoácidos para formar proteínas; el resultado final de la expresión del genoma es el proteoma (Brown, 2008).

El Proteoma está formado por todas las proteínas que componen la célula en un determinado momento, por ejemplo: cuando se analizan los proteomas de diferentes tipos de células (en mamíferos) se puede comprobar que apenas existen diferencias en proteínas abundantes (se consideran proteínas abundantes a aquellas que tengan un número de copias superior a 50.000 por célula), lo que quiere decir que gran parte de

ellas son proteínas que llevan a cabo actividades bioquímicas generales que ocurren en todas las células (Cazzulo, 2014).

Una vez que se obtiene una secuencia específica de ADN, existen diferentes métodos para localizar los genes presentes. Al obtener, por primera vez, la secuencia de un genoma, el principal objetivo es localizar las posiciones de todos los genes. Por ejemplo: para los genes que codifican proteínas se utilizarían los marcos de lectura abiertos (*Open Reading Frame, ORF*), es decir, se analizaría la secuencia de ARN específica, comprendida entre un codón de inicio y un codón de terminación, lo cual es complicado y puede conducir a errores debido a la presencia de intrones (un intrón es una región del ADN) (Brown, 2008). Las células decodifican el ARNm al leer sus nucleótidos en grupos de tres, estos se conocen como codones.

El número de proteínas que produce un organismo no es igual al número de genes. Los organismos pluricelulares, son capaces de producir un número de proteínas bastante superior al número de genes. Cada célula es capaz de producir solamente ciertas proteínas, debido a las modificaciones en la cromatina que son específicas de cada tipo celular, lo cual conduce a la falta de transcripción de una gran cantidad de genes (Cazzulo, 2014).

Además el proteoma cambia continuamente, ajustándose a las necesidades y condiciones ambientales que necesita la célula en determinados momentos. Por tanto, la identidad y abundancia de cada proteína que compone el proteoma simboliza un equilibrio entre la síntesis de nuevas proteínas y la degradación de las proteínas existentes (Brown, 2008).

### *1.1.1. La Proteómica:*

En genómica uno de los objetivos principales es establecer la composición del genoma, incluyendo la información sobre polimorfismos y mutaciones comunes. En la genómica funcional, el objetivo principal de los científicos es el de analizar la expresión de los genes, con el fin de poder describir las funciones e interacciones entre los genes y las proteínas, e incluso alguno de ellos, considera la proteómica como parte de la genómica funcional (Lovric, 2011).

Mientras que en genómica se estudia y se analiza el genoma de un organismo, en proteómica se estudia y se analiza su proteoma, el cuál difiere de una célula a otra y de un momento a otro, el principal objetivo de este área es analizar todo el proteoma de un determinado organismo, a través de una serie de experimentos o incluso, en un solo experimento (Lovric, 2011).

Por lo tanto, la Proteómica, se define como el conjunto de las diferentes técnicas relacionadas por su disposición para aportar información sobre el proteoma, lo que comprende desde las identidades de las proteínas constituyentes, especificar la función de una proteína determinada, hasta su ubicación intracelular. Gracias al estudio del

proteoma en diferentes situaciones metabólicas y patologías, se pueden identificar proteínas que se relacionan con determinados estadios fisiológicos, bien por su presencia, su alteración o su ausencia. Por ejemplo, si se analizase el proteoma en una patología concreta se podrían hallar proteínas que ayudarían a diagnosticar cierta enfermedad o analizar su evolución. Estas proteínas reciben el nombre de *biomarcadores* (Cazzulo, 2014).

La gran mayoría de los estudios de Proteómica se centran en correlacionar ciertas funciones con la expresión o modificación de proteínas específicas. Solo algunos de ellos tienen como objetivo describir los proteomas completos o compararlos entre distintas especies. Para una correlación funcional se necesita analizar las características proteicas más importantes de relevancia funcional. Estos análisis se pueden resumir en cuatro puntos (Lovric, 2011):

i. Identificación y cuantificación del nivel proteico.
ii. Identificación y cuantificación de modificaciones proteicas.
iii. Identificación y cuantificación de la localización proteica subcelular.
iv. Identificación y cuantificación de interacciones proteicas.

Históricamente, el estudio de la expresión proteica fue el primer parámetro analizado por los científicos. Esto implicaba una cierta forma de "cuantificación": presente/no presente de una proteína (que como mínimo se podía encontrar una diferencia de 3 a 10 veces en el nivel de expresión).

Las modificaciones post-traduccionales *(PTMs)* son muy importantes para la función de las proteínas y la Proteómica es la única ciencia que intenta analizarlas. Una modificación post-traduccional es un cambio químico que sufren las proteínas después de su síntesis por los ribosomas, muchas proteínas no podrían ejercer su función sin esta modificación química. Sin embargo los enfoques actuales son incapaces de analizar todas las modificaciones post-traduccionales posibles (Brown, 2008).

La identificación de redes de interacciones de proteínas es otro de los objetivos más desafiantes dentro de la proteómica. Generalmente, en un solo estudio, el principal objetivo es identificar todas las partes que interactúan con una sola proteína, en cambio, varios estudios, tomados juntos, se pueden usar para identificar todas las interacciones dentro de un único módulo de señalización. Ninguna proteína puede ejercer función sola, siempre tiene que haber una interacción con alguna otra (Brown, 2008).

La Proteómica estudia todo este tipo de problemas y la técnica que se emplea para estudiar la composición del proteoma se denomina *perfil proteico o proteómica de expresión*.

Para obtener este perfil proteico se utilizan dos métodos basados en separar las proteínas de su proteoma: la electroforesis de proteínas y la espectrometría de masas de proteínas. Por un lado la electroforesis de proteínas en gel es el método estándar para separar las proteínas de una muestra, pero tiene una serie de limitaciones, que son cubiertas por la

espectrometría de masas por láser de desorción/ionización asistida por matriz (*MALDI-TOF*) (Brown, 2008). Este método permite obtener una base de datos donde se encuentra la relación masa/carga *(m/z)* de las formas ionizadas que se producen cuando se exponen las moléculas del compuesto a un campo de alta energía, siendo "m" la masa molecular del compuesto y "z" un múltiplo entero de la carga elemental *(e)* del electrón. Esta relación masa/carga permite conocer la masa molecular y a su vez deducir la composición de aminoácidos (Ong & Mann, 2005) .

A su vez se puede ir obteniendo información relevante acerca de la actividad del genoma, identificando pares y grupos de proteínas que interactúan entre sí. Esta información puede llegar a ser valiosa cuando se trata de asignar una función a un nuevo gen o a una nueva proteína, ya que una interacción con una segunda proteína puede indicar la función de esta nueva proteína. Una forma de estudiar todas estas interacciones a nivel global, es a través de mapas de interacción proteica, este método permite vincular el proteoma con la bioquímica celular (Brown, 2008).

Los mapas de interacción proteica o redes de interactomas, proporcionan todas las interacciones que se producen entre los componentes de un proteoma. Cada una de estas redes está construida en torno a una pequeña cantidad de proteínas que muestran diversas interacciones y nodos, junto con una cantidad mucho mayor de proteínas con escasas conexiones individuales (Brown, 2008).

Actualmente existe una nueva alternativa para diseccionar el proteoma: la tecnología del microarray, concretamente los microarrays de proteínas. Destacan por el estudio de las interacciones proteína-proteína a gran escala, el estudio de proteínas diferencialmente expresadas con bajos niveles de expresión y la búsqueda de biomarcadores (Cazzulo, 2014).

### *1.1.2. Métodos de análisis estadístico en Proteómica:*

Una de las disciplinas más relevantes en proteómica, es la proteómica cuantitativa que se basa en la determinación de la abundancia de las distintas especies proteicas y la comparación de estas entre distintos individuos, poblaciones, orígenes… Existen diferentes métodos para realizar proteómica cuantitativa y todos tienen un objetivo común: establecer perfiles proteicos que pertenezcan a individuos o poblaciones (Valledor & Meijón, 2014).

Aunque se ha avanzado mucho en el análisis de enriquecimiento, validación de bases de datos, conocimiento y uso de los espectrómetros de masas… la proteómica cuantitativa considera el diseño experimental y el análisis de datos como dos de sus puntos más débiles. A través del diseño experimental se construye la base del desarrollo de todo ensayo así como el análisis estadístico de sus datos. Estos datos, al ser datos proteómicos, cuentan con una serie de particularidades, por tanto, el análisis estadístico

que se emplea cubre desde la estadística univariante con métodos tanto paramétricos como no paramétricos, hasta la estadística multivariante (Valledor & Meijón, 2014).

Para llevar a cabo un adecuado diseño experimental hay que tener en cuenta una serie factores:

- Las fuentes de variación existentes: biológicas, técnicas…
- Cuántas muestras como máximo podrían procesarse en paralelo en el laboratorio.
- Las pruebas estadísticos que se pueden aplicar.
- El empleo de un número mínimo de muestras con el fin de no malgastar muestras, recursos y tiempo.

Existen diferentes tipos de réplicas que se pueden relacionar con el tipo de ruido-variación dentro del propio sistema. Dentro de los casos en los que existe una alta variación relacionada a la metodología del análisis, las réplicas técnicas o analíticas se pueden definir como medidas repetidas de una misma muestra biológica. Por último, las muestras biológicas se definen como muestras que pertenecen a diferentes individuos sometidos a un mismo tratamiento (Valledor & Meijón, 2014).

A la hora de analizar datos procedentes de microarrays, un problema al que se enfrenta la estadística es que el número de genes/proteínas es mucho más grande que el número de sujetos, por lo tanto, en términos estadísticos estamos ante el caso de miles de variables frente a muy pocas muestras (Rivas-Lopez, Sánchez-Santos, & De Las Rivas, 2005).

Para llegar a un punto común que satisfaga tanto a la Biología como a la Estadística, el número de réplicas biológicas a realizar dependerá, en gran medida, del poder estadístico que se desea conseguir. Esta potencia se define como la capacidad del test para detectar la variación *(1-β, donde β es la tasa de falsos negativos, FDR)* y depende del número de réplicas, de la varianza intratratamiento (que suele ser muy alta), de la magnitud de la variación a estudiar y del nivel de significación deseado α (Valledor & Meijón, 2014).

Una cuestión fundamental, dentro del diseño experimental, es la importancia del muestreo y el procesado de muestras, el mismo protocolo de extracción de proteínas que se ha aplicado a un individuo, sujeto, población… se debe mantener y aplicar a todos por igual a lo largo del desarrollo del experimento (Valledor & Meijón, 2014).

Dentro del análisis estadístico univariante, se encuentran los test paramétricos como la prueba *t-Student, ANOVA…* y sus equivalentes no paramétricos como *la prueba U de Mann-Whitney, Kruskal-Wallis…*. Estos test se caracterizan por tener una hipótesis nula, previamente definida, que se desea contrastar y que se considera verdadera hasta que una prueba estadística indique lo contrario. Gracias a estas pruebas se pueden estudiar diferencias entre distintos tratamientos pero no permiten analizar un conjunto de variables, agruparlas o diferenciarlas (Valledor & Meijón, 2014).

Las técnicas necesarias para poder diferenciar entre muestras o agruparlas… son técnicas multivariante que sirven para reducir la complejidad de la muestra. Gracias a ellas se pueden indicar con exactitud los *spots* (celdas del microarray) que son importantes dentro del análisis, empleando tanto las diferencias individuales entre *spots*, como la relación de covarianza entre todos los *spots* analizados. Todos estos *spots* se pueden utilizar para definir distintos procesos biológicos (Valledor & Meijón, 2014).

Dentro de los métodos de estadística multivariante, como el análisis de componentes principales, esta técnicas permiten reducir las dimensiones de la matriz de datos, ya que al estar compuesta por miles de *spots* hace muy difícil su tratamiento con otros test estadísticos. Estos métodos permiten reducir la matriz en una serie de componentes que recogen gran parte de la variación original de la muestra, lo que posibilita la simplificación de la muestra permitiendo observar interacciones y relaciones existentes entre distintos spots (Valledor & Meijón, 2014).

Técnicas como los árboles de agrupamiento o los *Heatmaps*, son también métodos multivariantes muy utilizados en proteómica, el hecho de agrupar todos los *spots* que se comportan de forma similar dentro de la muestra permite definir grupos y detectar muestras atípicas. Gracias a los mapas de calor (*heatmaps*) se pueden visualizar todas las variaciones en los niveles de acumulación de los *spots* (Valledor & Meijón, 2014).

Aunque las técnicas multivariantes dan una visión global de los datos, es necesario emplear una estrategia univariante en la definición de marcadores biológicos, por lo que se calcularía el nivel de significación para tener en cuenta la tasa de falsos positivos que podemos encontrar en la muestra (*FDR o Bonferroni*). Por lo tanto, la mejor estrategia que se puede seguir, para analizar este tipo de datos, es aplicar ambos tipos de métodos. Además el hecho de aplicar estas dos técnicas, permitiría tener una mayor información sobre los datos (Valledor & Meijón, 2014).

## *1.2. La tecnología del microarray:*

La tecnología del microarray es una nueva herramienta para abordar estudios biológicos de alto rendimiento, son técnicas conocidas como "de alta capacidad de trabajo" o *high-throughput*. Un microarray se puede definir como una plataforma sólida donde se han impreso miles de muestras diferentes, por ejemplo, de: ARN, ADN, Proteínas, Metabolitos… lo que permite el análisis simultaneo de diferentes moléculas. Según *Schrenzel et al. (*2009), los microarrays consisten en una colección de moléculas ancladas con una disposición determinada sobre un soporte sólido y que actúan de agentes de captura específicos para las moléculas que se quieren analizar o detectar. Por lo tanto, existen varios tipos de moléculas que se utilizan como agentes de captura en esta tecnología (Figura C). Esta tecnología presenta diferencias dependiendo de la naturaleza de la molécula que participa en el ensayo, por lo tanto, la tecnología empleada en microarrays de ADN será diferente que la de los microarrays de proteínas.

Gracias a los microarrays de ADN se pueden realizar análisis de expresión génica a nivel genómico, sin embargo la determinación de la cantidad de ARNm mediante

microarrays de ADN no proporciona suficiente información sobre las proteínas que se traducen a partir de esa molécula, esto se debe a las modificaciones post-traduccionales a lo largo de su proceso de síntesis (Tomizaki, Usui, & Mihara, 2010).

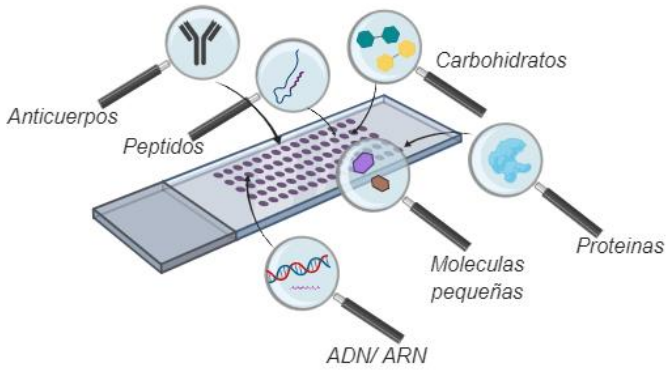



*Figura C. Representación de las diferentes moléculas que pueden ser empleadas como agentes de captura en un microarray. Imagen con BioRender.com*

Este límite, por parte de los microarrays de ADN, queda cubierto por los microarrays de proteínas, los cuales permiten la detección, caracterización y cuantificación de las proteínas, posibilitan estudiar el tipo de función que desempeñan y sus interacciones con otras proteínas u otras moléculas. Este tipo de tecnología es fundamental dentro del campo de la Proteómica, siendo de gran valor dentro de la investigación biológica básica.

### *1.2.2. Parámetros a tener en cuenta:*

A la hora de elaborar un microarray de proteínas hay que tener un cierto esquema, donde cada uno de los pasos son cruciales para no cometer fallos (López et al., 2006) (Figura D):

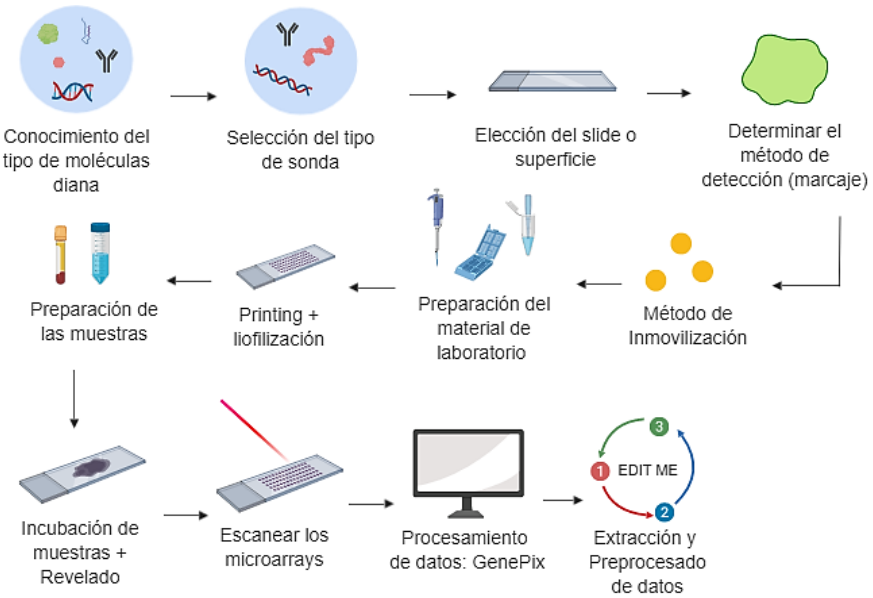



*Figura D. Esquema del flujo de trabajo que conlleva el diseño y elaboración de un microarray. Imagen creada con BioRender.com*

El diseño de este tipo de microarray, es más complicado que otros, debido a la complejidad y diversidad estructural de las proteínas, ya que estas no tienen ni una estructura homogénea ni un patrón de unión específico; cada proteína tiene unas propiedades bioquímicas diferentes, lo que complica los estudios de reactividad, funcionalidad e interacción entre proteínas. La principal dificultad de este tipo de tecnología es la adquisición y unión de proteínas a superficies donde puedan interaccionar con otras proteínas o ligandos, y donde permitan detectar, si hay o no, interacción (Guerrero Picó, 2015).

En cuanto a la elección del tipo de sonda, al ser de naturaleza proteica, conlleva una mayor dificultad, ya que solamente con la búsqueda bibliográfica en bases de datos de proteínas (como por ejemplo *Protein Data Bank*), no es suficiente, puesto que para obtener dicha molécula se debe realizar la expresión de su gen y su posterior separación y purificación. (Corrales, Calvete, & Sociedad Española de Proteómica, 2014).

La elección del soporte o superficie donde se van a colocar las proteínas se debe tener en cuenta en todo momento, ya que condiciona tanto el formato final del array como el método de detección. Esto también afecta a la tecnología de inmovilización elegida, la cual debe tener en cuenta la naturaleza del compuesto que se desea inmovilizar y la superficie a la que se va a unir, permitiendo la funcionalidad y accesibilidad de las proteínas que se van a depositar en el array. Dentro de todos los tipos de soporte que existen, hay una serie de conceptos clave que deben ser comunes a todos ellos: estabilidad química, mínimas uniones inespecíficas, baja señal de fondo, alto ratio superficie/volumen y compatibilidad con los distintos sistemas de detección (López *et al.*, 2005).

Actualmente nuevos soportes como superficies de membrana y de vidrio, no requieren el empleo de agentes bloqueantes para eliminar el ruido de fondo, además previenen el contacto directo de la proteína con la superficie mediante la introducción de grupos funcionales como polietilenglicol *(PEG)* (López et al., 2006).

Una vez escogido el tipo de soporte y de sonda, el siguiente paso es el marcaje o método de detección. Existen diversos métodos de marcaje, con etiquetado *(label-dependent)* y sin etiquetado *(label-free).* Dentro de los procesos con etiquetado, destacan entre otros, la detección por fluorescencia y quimioluminiscencia (Tomizaki et al., 2010). Los más utilizados hoy en día son: el marcaje fluorimétrico (empleando fluoróforos), colorimétrico (utilizando moléculas que catalizan una reacción la cual está asociada a un cambio de color), radioactivo (usando moléculas radiactivas) o quimio luminiscente. A pesar de las diversas técnicas de marcaje, destaca el marcaje fluorimétrico, se basa en el uso de proteínas (como la biotina) que actúan como puentes entre la sonda y la enzima del revelado, la cual está unida con una determinada molécula que reconoce estas proteínas (Guerrero Picó, 2015)(Figura E).

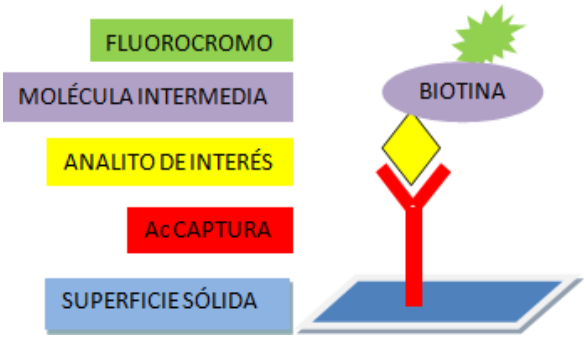



*Figura E. Ejemplo de un microarray plano con marcaje de la proteína a través de una molécula intermedia. Imagen adaptada de (R. P. González, s. f.)*

Si el marcaje es fluorimétrico existen múltiples herramientas para poder obtener una captura de las zonas de trabajo y obtener imágenes bidimensionales. Un claro ejemplo son los escáneres de fluorescencia comerciales, en cambio si el marcaje es colorimétrico se utilizarán escáneres de documentos convencionales (Guerrero Picó, 2015).

Los métodos de inmovilización son el último paso a tener en cuenta antes de llevar a cabo el printing. Estos métodos consisten en determinar cuál será el mejor método a través del cual se fusionará la sonda a la superficie o soporte escogido. Para llevar a cabo una unión adecuada, en algunos casos se requiere la activación de la superficie, también se necesita la modificación de las moléculas sonda mediante la aplicación de diferentes grupos químicos o de una proteína (biotina) (Aboytes et al., 2003).

Para favorecer la inmovilización de las proteínas es frecuente recurrir a moléculas adaptadoras de afinidad, que consisten fundamentalmente en proteínas, poliaminoácidos o polipéptidos. Algunas de estas proteínas, que se utilizan también como controles dentro de un microarray son: Biotina (*unión a estreptavidina*); Proteína G (*unión a anticuerpo Fc*); Proteína A (*unión a anticuerpo Fc*); GST, Glutation S-transferasa (*unión a anti-GST*); MBP, proteína de unión a la maltosa (*unión a anti-MBP*); TRX, tiorredoxina reductasa (*unión a anti-TRX*), GFP, proteína verde fluorescente (*unión a anti-GFP*) (López et al., 2006).

Diseñado el microarray, se prepara todo el material de laboratorio necesario y se lleva a cabo el *printing*. Existen dos formas diferentes: síntesis *in situ* y *spotting*. La primera es propia de los microarrays de ADN, y no se puede llevar a cabo en el resto de microarrays, en cambio la segunda, se utiliza tanto en microarrays de ADN como de proteínas. La modalidad de fabricación *in situ* se basa en la unión del oligonucleótido a través de la deposición de nucleótidos de forma cíclica. La modalidad de fabricación *spotting* se basa en colocar la disolución que contiene las moléculas sondas, ya sintetizadas mediante un robot, esto se puede llevar a cabo por contacto o sin contacto.

Realizado el *printing* y la liofilización (para la deshidratación de la sonda), se preparan las muestras en el laboratorio (muestras de sangre, liquido encéfalo raquídeo…) para más adelante llevar a cabo la incubación y revelado de estas.

Incubadas las muestras, se emplea un escáner, los escáneres de fluorescencia usan láseres para iluminar uno a uno todos los pixeles hasta que todos ellos se hayan escaneado y grabado como un archivo de imagen de alta resolución, en esta imagen se analiza la intensidad de todos los *spots* que forma el array. La intensidad está relacionada con la cantidad de moléculas diana o *target*, reconocidas por el agente de captura inmovilizado en la superficie. Estas imágenes son analizadas en un proceso de extracción de datos donde se mide la fluorescencia relativa de dos etiquetas fluoróforos de muestras de prueba y control, identificados en diferentes *spots* que forman la imagen, a través de ella se extraen los indicadores numéricos de la unión relativa de las moléculas de prueba y control en cada uno de los *spots* (Handran, s. f.) (Figura F).

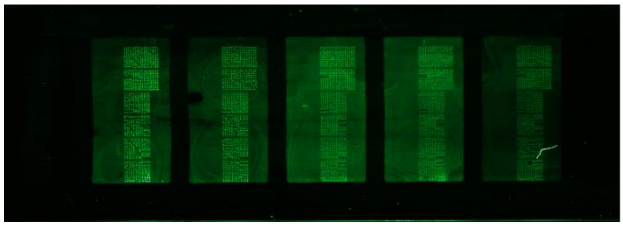


*Figura F. Ejemplo de microarray analizado a través del software GenePix®. Imagen creada con SensoVation AG.*

### *1.2.3. Tipos de microarray:*

Gracias a la base de datos que se obtendrá y junto a diversos métodos bioinformáticos, se obtiene una matriz de datos final, que corresponde a los diferentes *spots* que forman el microarray.

Existe una gran variedad de microarrays de proteínas que se pueden clasificar en función de:

A. Según el formato del microarray: arrays planos o de microesferas.
B. Según el contenido del microarray: analíticas, funcionales o de fase inversa.
C. Según su método de detección.

(Juanes-Velasco et al. 2018)

(Debido a la extensión del trabajo, este capítulo se desarrolla por completo en los Anexos teóricos, en este apartado se explicará el tipo de microarray utilizado en el proyecto).

A. <u>Microarrays según el formato:</u>

Los microarrays de proteínas se pueden clasificar en dos grupos dependiendo de la estructura física del soporte:

i. *Arrays planos:* en este tipo de array, el contenido es inmovilizado en microspots, dispuestos en una superficie de dos dimensiones, sobre una matriz sólida (Merbl & Kirschner, 2011), generalmente cada centímetro cuadrado de esta superficie

sólida, está compuesto por 1000 spots. En estos arrays se busca principalmente la robustez y reproducibilidad del ensayo, lo cual está relacionado con una serie de parámetros que afectan al array como por ejemplo la superficie, el método de detección, el ruido de fondo (*background signal*), etc. (Ellington, Kullo, Bailey, & Klee, 2010).

Además, es necesario tener en cuenta la señal que proporciona el ruido de fondo ya que afecta a la señal propia de los *spots.* Esta señal inespecífica debe evaluarse y controlarse para detectar correctamente la señal de interacción entre el agente y la proteína.

Otro aspecto clave para estos arrays, es la adquisición de imágenes del array utilizando escáneres de fluorescencia, para analizar a continuación esta imagen con el software GenePix®, donde las intensidades de fluorescencia de cada microspot se calculan en función de su señal de píxeles, canales de emisión… Finalmente los datos de cada uno de los spots que componen el array, son exportados en forma de hoja de datos (García-Valiente et al., 2019).

ii. *Arrays de microesferas.*

B. Microarrays según el contenido:

Según la naturaleza del agente capturador, existen diferentes tipos de microarrays de proteínas:

- *Microarrays analíticos o Microarrays de anticuerpos:* el modelo más representativo de microarrays de proteínas analíticas es el microarray de anticuerpos, en el que las proteínas se detectan después de la captura del anticuerpo mediante el marcado directo de las proteínas (Reymond Sutandy, Qian, Chen, & Zhu, 2013).

  Estos pueden llegar a tener de cientos a miles de anticuerpos impresos, y generalmente su uso se basa en la identificación de biomarcadores en cáncer u otras patologías (Corrales et al., 2014).

- *Microarrays funcionales o Microarrays de proteínas recombinantes.*
- *Microarrays de fase reversa o de lisados celulares.*

C. Microarrays según su método de detección:

Teniendo en cuenta el método de detección que se aplicará al array, se diferencia entre métodos con etiquetas *(label-dependent)* y sin etiquetas o con etiquetado libre *(label-free)* (González-González, Jara-Acevedo, Matarraz, Jara-Acevedo, & Paradinas, 2011).

I. Métodos basados en etiquetas *(label-dependent)* :

II. Métodos basados en etiquetado libre *(label- free)*:

- Resonancia de plasmones superficiales (SPR): ángulo de incidencia, se puede analizar si existe, o no, algún tipo de interacción biomolecular.

### *1.2.4. Aplicaciones de los microarrays de proteínas:*

El área de la medicina y la salud son campos privilegiados, donde investigar y desarrollar la tecnología micro- y la nanotecnología, prometen resultados revolucionarios. Gracias a estos avances se pueden crear estructuras de escala nanométrica con un alto potencial de diagnóstico y terapéutico. Dentro de las principales líneas de investigación y desarrollo de la nanotecnología, con su respectiva aplicación en medicina y salud, destaca la creación de productos médicos que facilitan la administración de fármacos, el *screening* de nuevas drogas, métodos diagnósticos, la identificación de nuevos biomarcadores… (A. D. González, s. f.).

Los microarrays de proteínas, tienen diversas aplicaciones. Por ejemplo: los microarrays de anticuerpos se utilizan para el análisis del proteoma de fluidos, el uso de este tipo de microarray permite la detección de proteínas presentes en vías de señalización y proporcionan información directa sobre los estados de fosforilación de las proteínas; los microarrays de dominios recombinantes se emplean, fundamentalmente, en el estudio de la funcionalidad, señalización, búsqueda de sustratos o identificación de biomarcadores. Por último, los microarrays de fase reversa se aplican en numerosos estudios relacionados, tanto con cáncer como con enfermedades infecciosas, se utilizan para analizar las interacciones proteína-proteína, efectos cuantitativos del RNAi combinatorial en modelos de línea celular, análisis de interacción entre tumor-estroma, identificación de biomarcadores…en los últimos años se han comenzado a utilizarse en diversas aplicaciones y ensayos clínicos (Corrales et al., 2014).

# *2. Objetivos:*

El objetivo principal del presente trabajo es diseñar y desarrollar estrategias computacionales para el análisis sistemático de microarrays de proteínas. Para alcanzar este objetivo general se establecieron los siguientes subobjetivos:

i. *Subobjetivo 1:* Diseño y lectura del microarray.
ii. *Subobjetivo 2:* Desarrollo de diferentes algoritmos en relación con el preprocesado de datos.
iii. *Subobjetivo 3:* Creación de un árbol de decisión.
iv. *Subobjetivo 4:* Visualización y análisis de datos.
v. *Subobjetivo 5:* Empleo de diferentes técnicas bioinformáticas con el fin de relacionar los datos con diferentes bases de datos como KEGG, GO… y hallar las redes de interacción y las rutas de señalización.

# *3. Materiales y métodos:*

Para este proyecto se elaboraron microarrays planos de proteínas printados con anticuerpos, donde el método de detección empleado fue: etiquetas fluorescentes convencionales. En microarrays de detección de proteínas con etiquetado, *label-dependent*, especialmente en aquellos que involucran anticuerpos, el principal método de detección es la *biotina* marcada con un fluoróforo, la cual se combina con las proteínas de estudio, formando un complejo que es detectado por los anticuerpos printados en el microarray. En microarrays analíticos o de anticuerpos, las moléculas diana serían los anticuerpos asociados a una determinada enfermedad (Tomizaki et al., 2010), en este caso, los microarrays están compuestos por proteínas y anticuerpos asociados a la Leucemia Linfocítica Crónica *(LLC).* El esquema que siguen estos microarrays es: (Figura H)

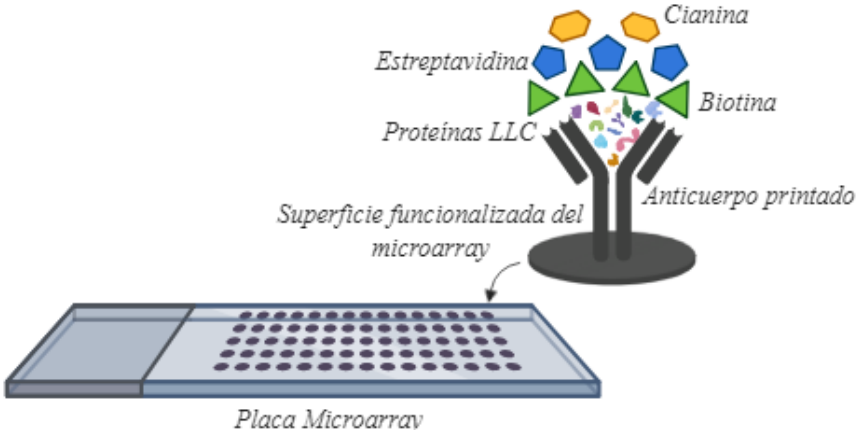



*Figura H. Esquema nanométrico del ensayo de análisis masivo de muestras de LLC. Imagen creada en BioRender.com*

El proceso de *printing* (la deposición ordenada y orientada de los anticuerpos sobre la superficie funcionalizada de los arrays), se lleva a cabo por *ArrayJet®Printer Marathon v.1.4.* Mediante nanoinyección sin contacto. Para llevar a cabo la impresión se requiere una placa de 384 pocillos, donde los puntos generados reciben el nombre de *spots.* En estas placas se incluyen tanto controles positivos como controles negativos.

Para poder normalizar las determinaciones, en la placa se incluyen Mastermix *(MM)*, alícuotas de las soluciones en las que se encuentran suspendidos los anticuerpos. Debido a que estos están suspendidos en diferentes soluciones, se utilizaron 12 MM diferentes:

i. MM1: $PBS\ 1x$
*ii.* MM2: $PBS\ 1x+< 0.1\%\ Azida + 1\%\ BSA$
*iii.* MM3: $PBS\ 1x+< 0.1\%\ Azida + 1\%\ BSA + 50\%\ Glicerol$
*iv.* MM4: $PBS\ 1x+ < 0.1\%\ Azida + 0.05\%\ BSA + 50\%\ Glicerol$
*v.* MM5: $PBS\ 1x+ < 0.1\%\ Gelatina + 0.1\%\ Azida$
*vi.* MM6: $100\ mM\ Hepes\ (pH\ 7.5) + 100\ mM\ NaCl + 100\frac{ug}{ml}\ BSA + 50\%\ Glicerol$
*vii.* MM7: $PBS\ 1x + 1\%\ BSA + 50\%\ Glicerol + 0.02\%\ Azida$

*viii.* MM8: $PBS\ 1x + 150\ mM\ NaCl + 50\%\ Glicerol$
*ix.* MM9: $PBS\ 1x + 1M\ Urea$
*x.* MM10: $PBS\ 1x + 0.2\%\ Gelatina$
*xi.* MM11: $PBS\ 1x + 0.1\%\ Azida\ sódica$
*xii.* MM12: $PBS\ 1x + 0.1\%\ Gelatina + 0.2\%\ BSA$

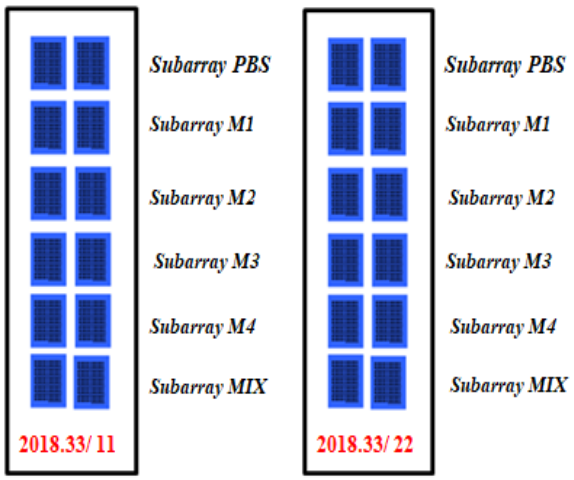
Una vez incubadas las muestras en la superficie del microarray, se emplea el escáner *SensoSpot Fluorescence* (SensoVation AG), para escanearlos. Se ajustan los parámetros de lectura al fluoróforo utilizado y se obtiene una imagen en formato TIFF. Esta imagen se analiza a través del software *GenePix®Pro 6.0.* La longitud de onda que se utilizará será de 532 nm. A lo largo del trabajo se analizan 10 microarrays diferentes, estos se clasifican en grupos diferentes en función del diseño de las muestras, cada una de las muestras representa a un paciente con LLC (estas muestras son siempre diferentes), según avance el proceso del diseño el número de réplicas irá disminuyendo y el número de muestras aumentando. Todos los microarrays contienen subarrays de buffer fosfato salino *(PBS)*, ya que esto forma parte del control de calidad del microarray*:*

- El primer conjunto está formado por 4 microarrays de proteínas con un total de 9 muestras: *(Figura I).*

  En cada uno de los *slides* se observan 3 muestras diferentes, junto con una réplica de cada una de las muestras, es decir, un total de 5 subarrays (la réplica de la muestra 3 se encuentra en el cristal número 2, y el último *slide* solamente contiene la réplica de la muestra 8 y 3 réplicas de la muestra 9). Este conjunto de microarrays se caracteriza por tener una réplica de cada una de las muestras que lo forman, con el objetivo de poder tener una mayor validez y confiabilidad en los resultados.


- El segundo grupo de arrays está formado por 2 microarrays de proteínas con un total de 4 muestras diferentes:*(Figura J).* En cada uno de los *slides* hay 4 subarrays diferentes que representan a cada una de las muestras, además se añadió un subarray más que contenía un *MIX* de todas las muestras que contenían el array.

Ambos arrays son iguales, es decir, comparten diseño y son las mismas muestras.
En este caso la réplica no se colocó debajo de su muestra, si no que la "réplica" se colocó en el otro cristal, con el objetivo de comparar si los resultados que se obtenían en el primer microarrays, eran resultados similares a los que se encontraban en el segundo microarray.




- El tercer conjunto de arrays está formado por 4 microarrays de proteínas con un total de 17 muestras diferentes y 3 *MIX*: *(Figura K).*

En cada uno de los *slides* se observan 5 subarrays diferentes que corresponden a diferentes muestras o a uno de los *MIX*. Este *MIX* está formado por las 17 muestras que forman este conjunto de microarrays, el objetivo de crear un subarray que contenga un *MIX* de las muestras es comprobar en cuál o cuáles muestras, la proteína printada con anticuerpos da positivo dentro del *MIX*. En este caso no existen réplicas, solamente se trabaja con una única muestra que corresponde a un único paciente.


Es necesario añadir, que la posición dentro del microarray que se haya escogido para cada una de las muestras, les afecta. No se obtendrá el mismo resultado colocando la muestra en otro subarray con una posición más alta o más baja.

Todas las imágenes en formato TIFF se analizaron a través del software *GenePix®Pro 6.0 (Anexos, Figuras 1-3, imágenes en formato TIFF);* al cargar la imagen en este

software se deben de ajustar una serie de parámetros en función de la intensidad de los microarrays, en este caso, para que cada uno de los microarrays de los diferentes conjuntos estuvieran bajo las mismas condiciones, se cargó la imagen de todos a la vez, ajustando las mismas condiciones de intensidad y una malla. Dicha malla permite saber que contiene un determinado *spot* y se debe de ajustar a todas las intensidades de cada uno ellos. Por ejemplo, si se toma como referencia el array número 8 del grupo 3, se carga la malla y se ajusta la intensidad, el resultado sería el siguiente: *(Figura L)*

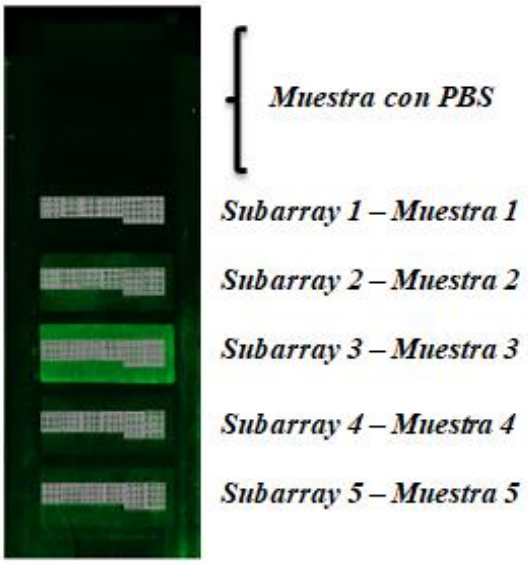



*Figura L. Imagen del microarray 8 del conjunto 3, analizada a través del software GenePix®Pro 6.0.con malla cargada e intensidades ajustadas. Imagen creada en GenePix®Pro 6.0.*

Además, este software, a través de las intensidades de los microarrays, permite realizar un primer control de calidad del array, con el objetivo de comprobar si todos los controles que lo componen funcionan correctamente. Los microarrays son fabricados de manera que puedan aportar información acerca del rendimiento de la hibridación y del escaneo de la imagen, por lo tanto, cada uno de estos arrays contiene una serie de controles que representan los marcadores de una determinada muestra. El valor de intensidad que se obtiene de todos estos controles, se utilizan para generar estadísticos de control de calidad. (Maestre, s. f.)

Tomando como ejemplo el microarray 8 del conjunto 3, esta vez sin malla, solamente con las intensidades de cada *spot*, se pueden analizar las intensidades que corresponden a controles o interacciones entre proteína-anticuerpo, que representan cada uno de los *spots.*

Por lo tanto, dependiendo de las intensidades de cada uno de los *spots* y de la malla establecida con anterioridad, se puede establecer una primera clasificación con el objetivo de analizar si los controles funcionan correctamente: *(Figura M)*

1. Control positivo: da señal positiva.
2. Control negativo: no da señal.

3. Posible resultado positivo: da señal positiva.
4. Posible resultado negativo: no da a penas señal.

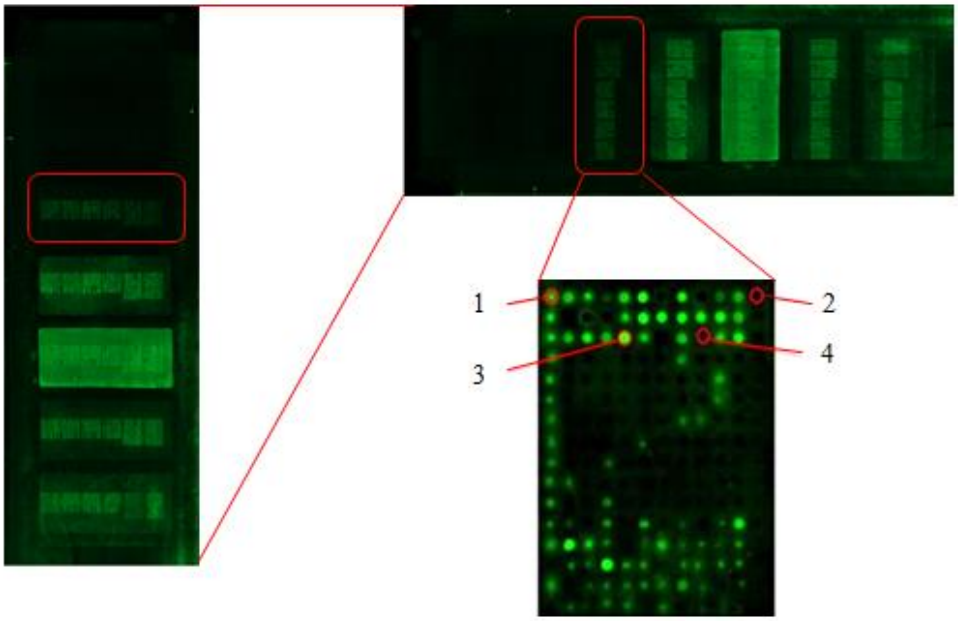



*Figura M. Esquema de un primer control de calidad del microarray, tomando como referencia uno de los 6 bloques del microarray 8 del conjunto 3. Imagen creada en GenePix®Pro 6.0.*

Una vez ajustada las intensidades y la malla para todos los microarrays de un determinado conjunto, se extraen indicadores numéricos, producto de la unión relativa entre las sustancias de prueba y las de control de cada *spot*. Esto implica un procesamiento cuantitativo para calcular las relaciones de intensidad de fluorescencia en diferentes longitudes de onda, lo que genera una gran cantidad de datos que requieren una interpretación adicional.

El software *GenePix®Pro 6.0*.genera una base de datos con 41 variables diferentes y más de 4.500 datos por cada microarray, el número de datos depende del número de proteínas printadas con anticuerpos que se hayan incluido en el microarray. Por lo tanto, algunas de las variables que componen estas bases de datos son:

- BLOCK: es el número de bloque al que pertenece un determinado *spot*, asignado por el software. Este valor aumenta de izquierda a derecha y de arriba abajo. Un bloque se puede definir como una estructura básica que se basa en el número establecido entre filas y columnas de los spots del array.

- COLUMN/ ROW: es el número de columna o fila de una entidad.

- NAME/ ID: es el nombre y el número de identificación definido por el usuario de cada punto del microarray. Esta opción estará en blanco si no se ha cargado y aplicado una malla que contenga los nombres y la identificación de cada *spot.*

- X/Y: es la ubicación física del *spot* en la imagen cargada, la coordenada (0,0) se ubica en la esquina superior izquierda del área de escaneo.

- DIAMETER: es el diámetro que se ha ajustado anteriormente a la intensidad del spot en cuestión, es valor viene medido en micras.

- F532 MEDIAN, F532 MEAN: es la mediana o los valores de intensidad media en longitud de onda 2 (532 nm) de cada uno de los *spots.*

- F532 SD: es la desviación estándar de los valores de intensidad en la longitud de onda 2 (532 nm) de cada uno de los *spots.* Un valor muy alto de desviación estándar puede significar un problema técnico en el microarray.

- B532: es un parámetro constante que afecta a todo el microarray, representa la intensidad de fondo en la longitud de onda 2 (532 nm) que afecta a un determinado *slide*.

- B532 Median, B532 Mean, B532 SD: es la mediana, la media o la desviación estándar de la intensidad de fondo en la longitud de onda 2 (532 nm) de los píxeles que cumplan los criterios para ser asignados como píxeles de ruido de fondo.

El resto de variables son: Ratio of Medians (2/532), Ratio of Means (2/532), Median of Ratios (2/532), F532 CV, B532 CV, Mean of Ratios (2/532), Ratios SD (2/532), Index Rgn Ratio (2/532), Rgn R2 (2/532), F532 Total Intensity, SNR 532, Normalize, F532 Median - B532, Sum of Medians (2/532), F532 Mean - B532, Sum of Means (2/532), Log Ratio (2/532), entre otras (Handran, s. f.).

Al trabajar con 10 microarrays diferentes, se analizarán 10 bases de datos cada una de ellas con 41 variables y con 4800 datos, se trabajarán con más de 215 proteínas diferentes. Cada uno de estos datos corresponde a un *spot* que pertenece a una determinada proteína printada con un anticuerpo.

Todos los microarrays están diseñados de cierta forma para que cada proteína esté representada en una fila dentro todos los subarrays. Esta fila contiene 6 *spots* donde se encuentra la proteína, por ejemplo: la proteína *CASPASA2* se encuentra en todos los microarrays y en todas las muestras de un mismo microarray (subarrays), de tal forma que, dentro de una misma muestra, esta proteína se ubica en 6 spots colocados en una misma fila.

Todas las proteínas están suspendidas en una Mastermix determinada: *4EBP1-MM6, a-Tubulina-MM10, ABL-MM11, Acetyl-Histona3-MM6, Acetyl-Histona4-MM6, Acetyl-P53-MM6, AIF-MM6, AKT-MM6, ANTI CHEMOKINE CXCR2-MM1…S6-MM6, SOCS-3-MM5, SOCS6-MM5, SOD2-MM11, SQSTM1-MM5, STAT3-MM5, TGF-BETA-MM6, TUBERIN/TSC2-MM6, VEGF Receptor 2-MM1…*

Además, para poder estudiar la señal de intensidad propia de las Mastermix, dentro del microarray existen *spots* que contienen únicamente una determinada *MM*, de tal forma que, en cada uno de los subarrays, hay *spots* con *MM1*, *MM2*, *MM3, …MM12.*

Por último, dentro del *slide*, en todas las muestras, hay controles positivos y negativos, con el objetivo de estudiar la calidad del microarray. Algunos de estos controles son:

- Controles positivos: *Anti Halo, Anti Biotin, Suero, IgG, Anti Human IgG, Biotin.*
- Controles negativos: *Anti GST, MM1, …, MM12, GST, BSA + BS3, PBS.*

Esta lista de proteínas, MM, controles…es constante para todos los 3 conjuntos de microarrays, siempre se trabajará con las mismas moléculas con la finalidad de encontrar aquellas proteínas printadas con anticuerpos que sean representativas de la Leucemia Linfocítica Crónica.

Para averiguar qué proteínas tienen interacción con el anticuerpo, y dan señal positiva "verdadera", estos datos deben pasar por un preprocesado, el cual se divide en 3 conceptos:

I. *Background Correction.*
II. *Filtering*
III. *House-keeping.*

*I. Background Correction:*

Debido a que existe cierta cantidad de hibridación no específica contenida en el *background* y que afecta a la sensibilidad y especificidad del microarray, la corrección del *background* permite minimizar los efectos que dicha hibridación causa sobre cada señal de los *spots.*

En este tipo de microarrays, la corrección del ruido de fondo se divide en dos partes, para poder obtener la señal de interacción "verdadera o limpia" entre la proteína y el anticuerpo, primero se debe eliminar la señal de fondo que proporciona el software al analizar la imagen, y una vez obtenido ese valor, restar la señal de intensidad que proporciona la *MM* sobre ese *spot:*

$$\delta_{kij} - MM_{xki}$$

Siendo:

$\delta_{ki}$ = señal de intensidad verdadera/ limpia del *spot* $j$ en el array $k$ $(k = 3, 4, 5, 6, 8, 11, 12, 16, 18, 22)$ y en la muestra $i$ $(i = 1, 2, 3, \dots)$.

$MM_{xki}$ = señal de intensidad de la Mastermix $x$ $(x = 1, 2, \dots, 12)$ en el array $k$ y en la muestra $i$.

Por lo tanto, para eliminar el ruido de fondo generado por el software, la primera parte del algoritmo se podrá definir como:

$$\delta_{kij} = \left(Me\ \delta_{kij} - \beta_k\right) - 2\sigma\ \gamma_{kij}$$

Siendo:

$\delta_{kij}$ = señal de intensidad verdadera del *spot j* en el array $k$ ($k = 3, 4, 5, 6, 8, 11, 12, 16, 18, 22$) y en la muestra $i$ ($i = 1, 2, 3, \dots$).

$Me\ \delta_{kij}$ = la mediana del valor de señal de intensidad de fondo del *spot j* en el array $k$ ($k = 3, 4, 5, 6, 8, 11, 12, 16, 18, 22$) y en la muestra $i$ ($i = 1, 2, 3, \dots$).

$\beta_k$ = parámetro de señal de intensidad de fondo constante que afecta a todo el microarray.

$2\sigma\ \gamma_{ki}$ = la varianza de señal de intensidad de fondo que hace referencia al *spot j* del array $k$ y en la muestra $i$.

Uno de los principales problemas que presentan los microarrays es la gran variabilidad que tienen sus datos, debido a esto se trabaja con la mediana de los datos y no con su media, la señal de intensidad de una misma proteína en una misma muestra puede variar de un *spot* a otro con una diferencia de más de 5000 píxeles de intensidad, por esta razón se trabaja con la mediana de la intensidad del spot y no con su media.

Además, aunque se hayan analizado los conjuntos de microarrays con los mismos parámetros (con el software *GenePix®Pro 6.0*), el parámetro de señal de intensidad de fondo es diferente para cada array del mismo conjunto, debido a esto se debe diferenciar este parámetro entre unos y otros. *(Gráfico 1).*

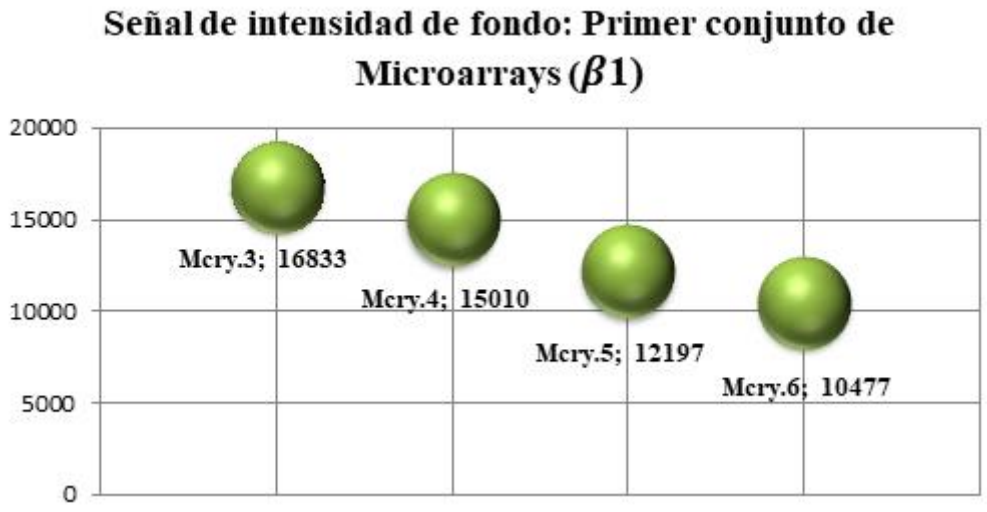



*Gráfico 1. Gráfico de las diferentes señales de intensidad constante ($\beta_k, k = 1$), que afectan al primer conjunto de microarrays. En el eje Y se observa la señal de intensidad de fondo y en el eje X el parámetro de intensidad constante de cada subarray.*

La varianza hace referencia a la dispersión de los datos en cuanto a la señal de intensidad de fondo; en la tecnología del microarray la dispersión de los datos es muy alta, y debido a este hecho, es mucho más complicado tratar con los datos.

Una vez se hayan calculado la señal de ruido que proporcionaba el software, se debe de eliminar el valor añadido que proporcionan las Mastermix *(MM)* a la señal de intensidad.

Dentro de los microarrays se colocaron *spots* que solamente contenían la *MM*, gracias a esto, el valor que pertenece a una determinada *MM* se puede eliminar más fácilmente. Cada una de estas Mastermix está compuesta de forma diferente, la *MM1* solamente contiene *PBS* mientras que la *MM12* contiene *PBS, Gelatina* y *BSA.* A medida que aumenta el número de referencia de la Mastermix, aumentan los compuestos que contiene, por lo que (teóricamente) según aumentan estos compuestos, debería aumentar también la señal de intensidad de la *MM.*

Cada una de estas señales de intensidad varía de una muestra a otra, el valor que se encuentra en los *spots* de *MM1* de la Muestra 1 en el Array 3 no son los mismos valores de los *spots* de *MM1* de la Muestra 2 en el Array 3, por lo tanto, siempre de deben diferenciar entre unas muestras y otras.

Teóricamente se ha visto que según aumentaban los compuestos, debería aumentar la señal de intensidad, pero en la práctica esto no es así. Teniendo en cuenta que cada Mastermix también está afectada por la señal de ruido que provoca el software:

$$MM_{xkij} = \left(Me\ MM_{xkij} - \beta_k\right) - 2\sigma\ \gamma_{kij}$$

Siendo:

$Me\ MM_{xkij} =$ mediana de la señal del *spot j* que hace referencia a la intensidad de la Mastermix $x$ $(x = 1, 2, \dots, 12)$ en el microarray $k$, en la muestra $i$.

$\beta_k$ = parámetro de señal de intensidad de fondo constante que afecta a todo el microarray.

$2\sigma\ \gamma_{kij}$ = la varianza de la señal del *spot j* que hace referencia a la intensidad de fondo del microarray $k$ y en la muestra $i$.

Como existen *4 spots* que hacen referencia a cada una de las *MM*, se escogerá el máximo de estos valores, es decir, si se toman como referencia el valor de la MM1 en el Microarray 8 en la Muestra 2, se restan los valores para obtener el valor de intensidad propia de esta MM1, el resultado será: *(Tabla 1. Señal de intensidad de la MM1).*

| Spots *MM1* | Señal de intensidad ($MM_{182j}$) |
|---|---|
| $MM_{1821}$ | 19840 |
| $MM_{1822}$ | 23395 |
| $MM_{1823}$ | 20889 |
| $MM_{1824}$ | 25767 |

$Max\ MM_{182j} = 25767 \sim 26000$

Además, una vez que se obtiene el valor final, se aproxima a su valor entero más próximo con el fin de asegurar que, se está eliminando al 100% la señal de intensidad que es propia de la *MM* y que no pertenece a la señal de interacción entre la proteína y el anticuerpo.

Calculando ese valor para cada una de las *MM* se observa que muchas de estas Mastermix tienen valores máximos que no superan al valor máximo de la *MM1* (que es la que menos compuestos contiene), por tanto, se encuentran 3 casos:

$$\begin{cases} Si\ Max\ MM_1 > Max\ MM_x, & Max\ MM_1 \\ Si\ Max\ MM_1 < Max\ MM_x, & Max\ MM_x \\ Si\ Max\ MM_x \leq 0, & Max\ MM_x = 0 \end{cases}$$

Si existe una determinada *MM* que tiene valores negativos, quiere decir que estos valores se sobre expresaron al analizar la imagen, y su señal de intensidad superó al límite, que tiene por defecto el software *GenePix®Pro 6.0*, por lo tanto, el valor máximo de esa *MM* será 0.

Una vez obtenidos el valor de intensidad propio de cada una de las *MM* y eliminado el ruido que afectaba a la señal de interacción entre la proteína y el anticuerpo, se resta el valor de la Mastermix donde se encuentran suspendidas cada una de las moléculas.

*II. Filtering:*

En el siguiente paso se realizará un filtrado de datos; se han eliminado ya la señal que pertenecía a la *MM* y aquella que, proporcionada el software, se trabaja con la señal de intensidad de interacción limpia que existe entre la proteína y el anticuerpo.

Si esta señal es negativa, quiere decir que no existe interacción entre la proteína y el anticuerpo, en cambio sí es positiva quiere decir que sí que hay interacción. Se filtrarán todas las bases de datos para trabajar solamente con aquellas proteínas que tengan señal positiva, el resto con señal negativa se eliminarán de la base de datos.

El hecho de diseñar el microarray con 6 *spots* que hacen referencia a una misma proteína en todas las muestras, tiene un objetivo. En este paso, solamente pasarán el siguiente filtro, aquellas proteínas que tenga como mínimo el 50% de sus *spots* con señal positiva, con el fin de asegurar que todas las proteínas que hayan pasado todos estos filtros realmente tengan señal de interacción con el anticuerpo, ya que si se aceptasen todos aquellos *spots* con señal positiva seguramente se admitirían como positivas muchas proteínas que no deberían formar parte de la lista final.

Si existe un total de 6 *spots* por proteína, el 50% serán 3 *spots*, por lo que solo se aceptarán aquellas que tengan un número mínimo de 3 réplicas.

Una vez pasados estos dos conceptos, se comenzará a trabajar con una lista final de proteínas, se hallarán la media y la desviación típica de cada una de las intensidades pertenecientes a una determinada proteína para cada una de las muestras que componen los microarrays.

Por lo tanto, se obtendrán un total de 39 bases de datos diferentes que harán referencia a pacientes con Leucemia Linfocítica Crónica, donde sus proteínas dan señal de interacción con determinados anticuerpos.

Por ejemplo, la lista de proteínas definitivas de la muestra 6 del microarray 5 está compuesta por 95 proteínas de 217 que contenía la lista original *(Tabla 2. Lista de proteínas del subarray 6, microarray 5).*

| *Lista de proteínas - subarray 6, microarray 5* | | |
|---|---|---|
| *4EBP1* | *MDM2* | *MAX* |
| *a-Tubulina* | *MEK1/2* | *MCL1* |
| *AIOLOS* | *Mouse mAb CATHEPSIN* | *MDM2 - 2* |
| *AK1* | *N-FATc1* | *SOCS-3* |
| *Anti Biotin 1/200* | *NEUREGULIN-3* | *SOCS6* |
| *anti biotin* | *NFKappaBP65* | *SQSTM1* |
| *ANTI CHEMOKINE* | *NOXA* | *STAT3* |
| *ANTI FOS* | *OPN sc-21742* | *TRAP* |
| *ANTI JUN* | *p-AKT (T308)* | *VEGF Receptor 2* |
| *Anti-GM-CSF* | *P-AS160* | *BIOTIN 10 pg* |
| *Anti-human IL-5* | *P-BLNK* | *HSP60* |
| *Anti-human IL-6* | *P-CRAF (ser 338)* | *HSP90* |
| *Anti-human IL-8* | *p-ERK* | *IGF-IrB* |
| *ATF-4* | *P-GAB1 (tyr 627)* | *IKAROS* |
| *ATF-6* | *P-GSK3B* | *IkB-a(H-4)* |
| *ATIII sc-271987* | *P-MEK1/2* | *Integrinalpha6* |
| *BCL-2* | *p-PDGFR-B* | *IRF4* |
| *BSA+BS3* | *P-PRAS40 (t246)* | *Rb pAb anti IL-4* |
| *C-EBPalpha* | *P-STAT3* | *Rb pAb anti IL-6* |
| *CALNEXIN* | *P-TYR* | *Rb pAb Anti TNF a* |
| *CASPASA2* | *p21* | *S6* |
| *CASPASA9* | *PDGFR-B* | *HEXOKINASE II* |
| *CD 21 sc-13135* | *Phospho-H2AX (ser139)* | *HGF-a* |
| *CD19 Clone* | *PhosphoHistona3(ser10)* | *HIF1a* |
| *CDC2 (POH1)* | *PIDD* | *HSF-1* |
| *Cleaved PARP* | *PIM-1* | *Rb pAb anti CSF* |
| *CRBN* | *PIM1(C93F2)* | *Rb pAb Anti IFN g* |
| *CXCR4* | *PIM2* | *Rb pAb anti IL-2* |
| *CyclinB1* | *PUMAa* | *FLG(C-15)* |
| *CyclinD2* | *Rabbit mAb ACTINA(C-2)* | *HCAM sc-9960* |
| *ERK2* | *Rabbit pAb MCL1* | *HDAC6* |
| *FGF-13* | *RAIDD* | |

Dicho listado contiene los siguientes controles: *Anti Biotin 1/200, anti biotin, BSA+BS3, BIOTIN 10 pg.*

El resto de las listas se encuentran en: *(Anexos, Tablas 1 – 52, Lista de proteínas).* La mayor lista pertenece a las muestras del segundo conjunto de microarrays, el número 11 y 22.

Las medianas y las deviaciones de las intensidades de cada una de las proteínas vienen representadas en diferentes histogramas: *(Gráfico 2)*

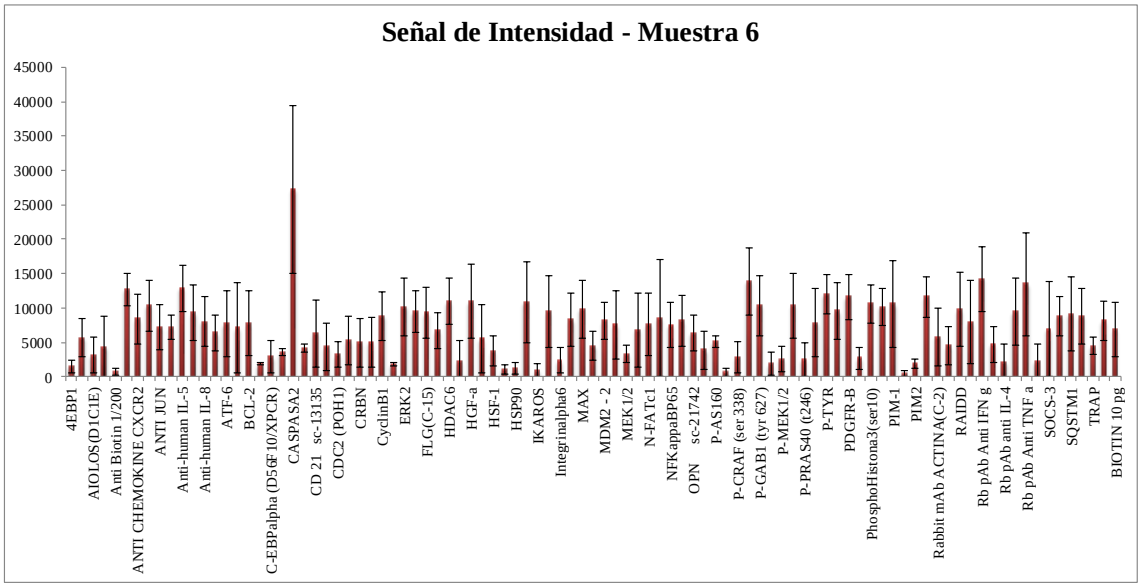



*Gráfico 2. Histograma con las proteínas resultantes de la muestra 6 junto con la media y desviación de intensidad. En el eje X vienen representadas las proteínas y el en eje Y la señal de intensidad.*

En este subarray se puede ver que la señal de intensidad más alta pertenece a *CASPASA2*, el resto de las proteínas tienen una señal inferior a 15000 píxeles.

Cada uno de los subarrays contiene su propia lista de proteínas printadas con anticuerpos, juntos con sus determinados gráficos de intensidades. *(Anexos, Gráficos 1-44, Señal de intensidad de los subarrays ).*

*III. House-keeping.*

El mantenimiento biológico de las células afecta a la expresión génica de ciertos genes y por lo tanto a las proteínas, aunque existen determinadas proteínas que no sufren cambios a lo largo de este mantenimiento biológico. Estas proteínas reciben el nombre de *"house-keeping proteins"* y por definición no reflejan ningún cambio en sus niveles de expresión durante el desarrollo celular, el tratamiento o anomalías presentes durante una determinada enfermedad. (Khimani et al., 2005)

Todas estas proteínas de mantenimiento pueden utilizarse para estimar los niveles de expresión relativos a otras proteínas, en los microarrays que se han utilizado a lo largo del trabajo existen varias proteínas que cumplen estas características: *BIOTIN, Anti Biotin, Tubulina, Anti-b-tubulina, GAPDH…*

Por lo tanto, se utilizaron como proteínas de mantenimiento todas aquellas que pasasen los filtros anteriores y que estuvieran presentes en los 10 microarrays. Las proteínas resultantes para normalizar las intensidades fueron: *BIOTIN y Anti Biotin.*

En los pasos anteriores, se hallaron la intensidad propia de cada *spot* y un cierto número de ellos, que correspondían a una proteína específica. Por tanto, cada uno de los

subarrays contiene una lista con un número mínimo de 3 spots de una determinada proteína con su intensidad. Para realizar los ratios se toman cada una de estas intensidades, calculadas anteriormente y se divide por la media de la intensidad de la proteína de mantenimiento.

Una vez realizado este paso, se halla la media de todas las intensidades, para así obtener por cada proteína un único valor medio de intensidad *(Tabla 3. Media de Intensidades escalonadas, de la muestra 6).*

| *Name* | *Muestra 1* | *Muestra 2* | *Muestra 3* | *Muestra 4* | *Muestra 5* | *Muestra 6* | *Muestra 7* | *Muestra 8* | *Muestra 9* | *Biotin 10 pg* |
|---|---|---|---|---|---|---|---|---|---|---|
| *4EBP1* | 0 | 0 | 0 | 0 | 0 | 0,23 | 0 | 0 | 0 | 1 |
| *a-Tubulina* | 0 | 0 | 0 | 0 | 0 | 0,83 | 2,127716 | 0 | 0 | 1 |
| *Acetyl-Histo* | 0 | 0 | 0 | 0 | 0 | 0,00 | 0 | 0 | 0,62168189 | 1 |
| *Acetyl-Histo* | 0 | 0 | 0 | 0 | 0 | 0,00 | 0 | 0 | 0,69430759 | 1 |
| *AIF* | 0 | 0 | 0 | 0 | 0 | 0,00 | 0 | 0 | 0,78793706 | 1 |
| *AIOLOS* | 0 | 0 | 0 | 0 | 0 | 0,46 | 1,40176789 | 0 | 0 | 1 |
| *AK1* | 0 | 0 | 0 | 0 | 0 | 0,62 | 1,48913601 | 0 | 0,58239598 | 1 |
| *AKT* | 0 | 0 | 0 | 0 | 0 | 0,00 | 0 | 0 | 0,56906154 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| *SQSTM1* | 0 | 0 | 0 | 0 | 0 | 1,33 | 1,78557171 | 2,13066472 | 0,50226612 | 1 |
| *STAT3* | 0 | 0 | 0 | 0 | 0 | 1,28 | 0,93988024 | 4,26693718 | 0 | 1 |
| *TRAP* | 0 | 0 | 0 | 0 | 0 | 0,66 | 0 | 0 | 0 | 1 |
| *TSC2* | 0 | 0 | 0 | 0 | 0 | 0,00 | 1,66421443 | 0 | 0 | 1 |
| *VEGF* | 0 | 1,72 | 0 | 0,18 | 0 | 1,17 | 1,40387796 | 0 | 0,54555557 | 1 |

Escalados los valores de todas las intensidades frente a la media de intensidad de esta proteína se obtiene: Ratio frente a BIOTIN *(Gráfico 3)*, el resto de gráficos con sus respectivos ratios se encuentran en *(Anexos, Gráficos 47-50, Ratio frente a BIOTIN).*

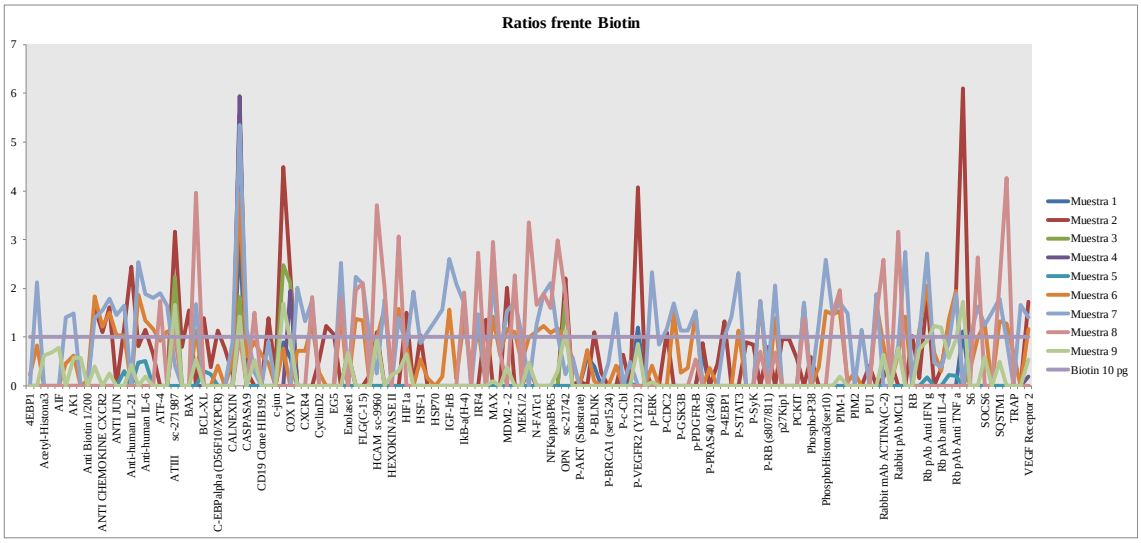



*Gráfico 3. Resultado de la normalización de los datos frente a la proteína de mantenimiento BIOTIN para el primer conjunto de microarrays. En el eje Y viene representado la señal de intensidad normalizada y en el eje X todas las proteínas. En la leyenda se pueden observar todas las muestras.*

Como se puede comprobar en el gráfico la muestra con mayor intensidad, dentro del primer conjunto de microarrays, corresponde a la muestra 2 y a la muestra 4. Aunque en esta última muestra la única proteína que pasó todos los conceptos dentro del preprocesado de datos fue *CASPASA9*, el hecho de que cierta muestra tenga las intensidades más altas no quiere decir que tenga el mayor número de proteínas. La

muestra con mayor número de proteínas corresponde a la muestra 6 (con 95 proteínas) y a la muestra 7 (con 94 proteínas), aunque la muestra 7 tiene señal de intensidad superior a la muestra 6. Si se realiza un *Ranking Test* (prueba que muestra aquellas proteínas con mayor señal de intensidad por muestra, con el objetivo de comprobar si siguen cierta progresión) con los resultados obtenidos en el gráfico anterior, tomando las proteínas con mayor intensidad para cada una de las muestras:

*i.* Muestra 1: ***CASPASA2***> *P-VEGFR2 (Y1212)*> *ATIII sc-271987*> *Rabbit pAb p-BIK* > *Cleaved PARP (asp214).*

*ii.* Muestra 2: *Rabbit pAb p-BIK*> ***CASPASA2***> *Cleaved PARP (asp214)*> *P-VEGFR2 (Y1212)*> *ATIII sc-271987.*

*iii.* Muestra 3: *Cleaved PARP (asp214)*> *ATIII sc-271987*> *COX IV*> ***CASPASA2*** > *OPN sc-21742.*

iv. Muestra 4: ***CASPASA2***> *COX IV*> *VEGF Receptor 2.*

*v.* Muestra 5: ***CASPASA2***> *Anti-human IL-6*> *Anti-human IL-5*> *BCL-XL*> *Anti-GM-CSF.*

*vi.* Muestra 6: ***CASPASA2***> *Rb pAb Anti IFN g*> *p-ERK*> *Rb pAb Anti TNF a*> *Anti-human IL-5.*

*vii.* Muestra 7: ***CASPASA2***> *RAIDD*> *Rb pAb Anti IFN g*> *IGF-IrB*> *Phospho Histona3 (ser10).*

*viii.* Muestra 8: *STAT3*> *BCL-2*> *HCAM sc-9960*> *Mouse mAb CATHEPSIN K*> *Rabbit pAb MCL1.*

*ix.* Muestra 9: *Rabbit pAb p-BIK*> *Cleaved PARP (asp214)*> *ATIII sc-271987*> ***CASPASA2***> *Rb pAb anti IL-2.*

Se puede comprobar que en cada una de las muestras siempre hay alguna proteína con los niveles más altos y que está presente en la gran mayoría de subarrays, por ejemplo, la ***CASPASA9***, que es una de las proteínas con mayor intensidad y que está presente en casi todas las muestras. También se puede comprobar como la muestra número 8 es la más diferente de todas, entre todos los subarrays hay proteínas que se repiten en cambio no hay ninguna proteína común en el *Racking Test*, que se encuentre en la muestra 8.

Para el resto de microarrays se normalizaron los datos frente a esta proteína de mantenimiento, y se realizaron los ratios englobando las muestras por conjunto de microarrays. Para el segundo conjunto de microarrays (número 11 y 22), al ser las mismas muestras en diferentes *slides*, se escalaron los datos y se calcularon los ratios para cada microarray por separado *(Anexos, Gráficos 48 y 49)*, además con el objetivo de comprobar si los microarrays funcionaban correctamente, se analizó el *MIX* de cada uno de estos cristales, estudiando que muestra/muestras eran las que habían favorecido la presencia de una determinada proteína dentro del *MIX (Anexos, Gráfica 45 y 46).*

Por último, para el tercer conjunto de microarrays, al ser 17 muestras y 3 MIX, se normalizaron todos estos datos frente a la proteína de mantenimiento *(BIOTIN)*, y se realizaron los ratios de forma independiente: por un lado, las intensidades normalizadas de las muestras y por otro lado las intensidades normalizadas de los *MIX*. También se estudió qué muestras eran las que habían favorecido la presencia de una determinada

proteína. Establecido el preprocesado de datos, el siguiente paso fue la creación de un árbol de decisión para facilitar el análisis de los microarrays: *(Figura N).*

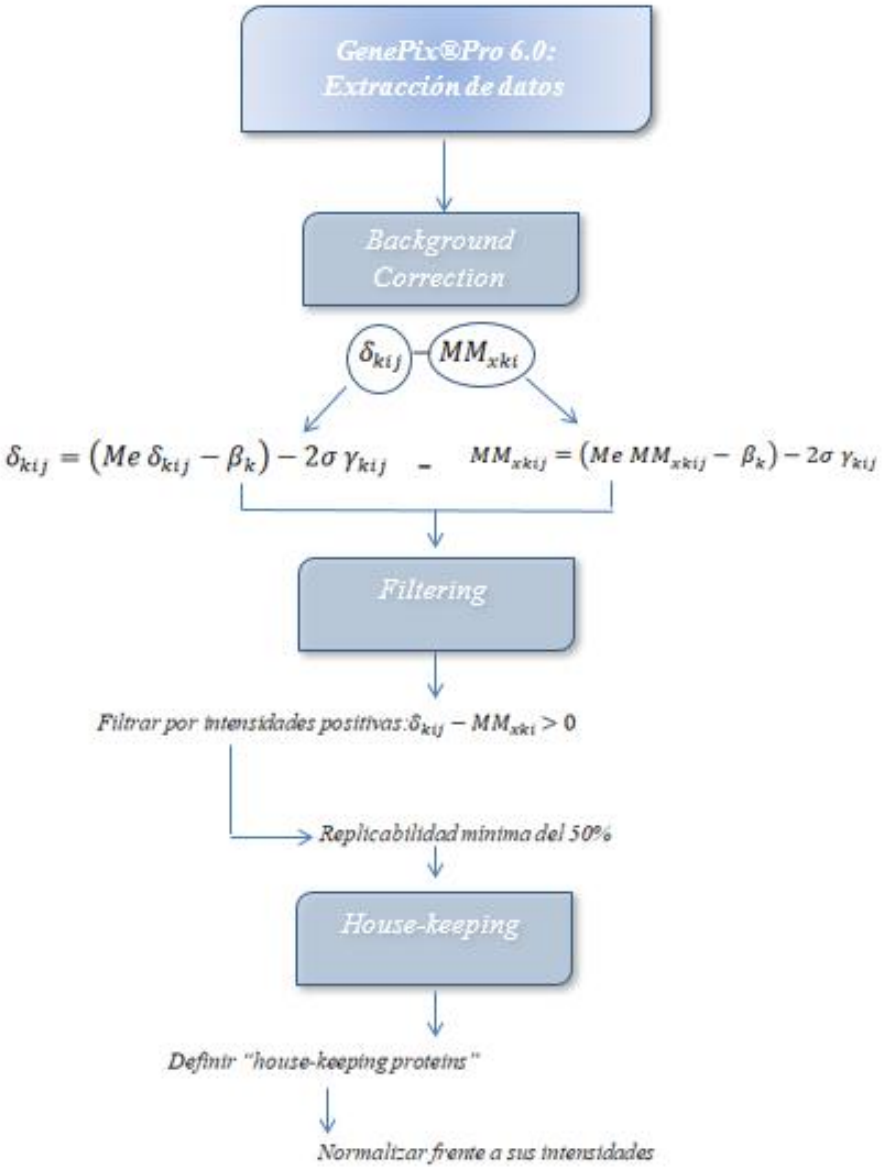



*Figura N. Árbol de decisión frente al pre procesado de datos, a partir de la obtención de la base de datos en GenePix®Pro 6.0.*

Al tener las listas definitivas de cada una de las muestras *(Anexos, Tablas 1-52, Listas de proteínas)*, a través de diferentes técnicas estadísticas, se puede determinar que muestras son más similares, como se agrupan las proteínas y diferentes formas de visualizar todos los datos.

En los gráficos anteriores se interpreta la señal de intensidad de las proteínas, lo cual permite trabajar los datos de forma cuantitativa. Para obtener una mayor información acerca de estos datos, se calcularon el número de *spots* que daban señal positiva para

cada una de las proteínas, con el objetivo de analizar si gran parte de ellas daban señal con el número máximo de *spots* (6 *spots*) o menos.

Por lo tanto, no solo se trabaja con los datos de forma cuantitativa *(Gráfico 6 y 7)* sino también de forma cualitativa *(Gráfico 8)*. Tomando de ejemplo uno de los microarrays con mayor número de proteínas: microarray 11, con 193 proteínas diferentes que dan señal positiva:

- Análisis Cuantitativo de los datos: *Cluster Dendrogram/ Heatmap.*

Realizando un dendrograma (diagrama de árbol) del microarray 11, el cual diferencia en 3 grupos las 193 proteínas originales. El nivel de similitud se mide en el centro del gráfico y las proteínas se especifican alrededor del grupo al que pertenecen. Se formaron conglomerados utilizando el método de *WARD*, este criterio de análisis jerárquico se conoce como "Criterio de Varianza Mínima de Ward": minimiza el total dentro de la varianza del clúster. En cada paso, el par de clúster con distancia mínima entre ellos son mezclados, además para implementar este método, en cada uno de estos se debe encontrar el par de clúster que llevan al incremento mínimo del total de la varianza del clúster después de mezclarlos. («Ward's Method (Minimum variance method) - Statistics How To», s. f.)

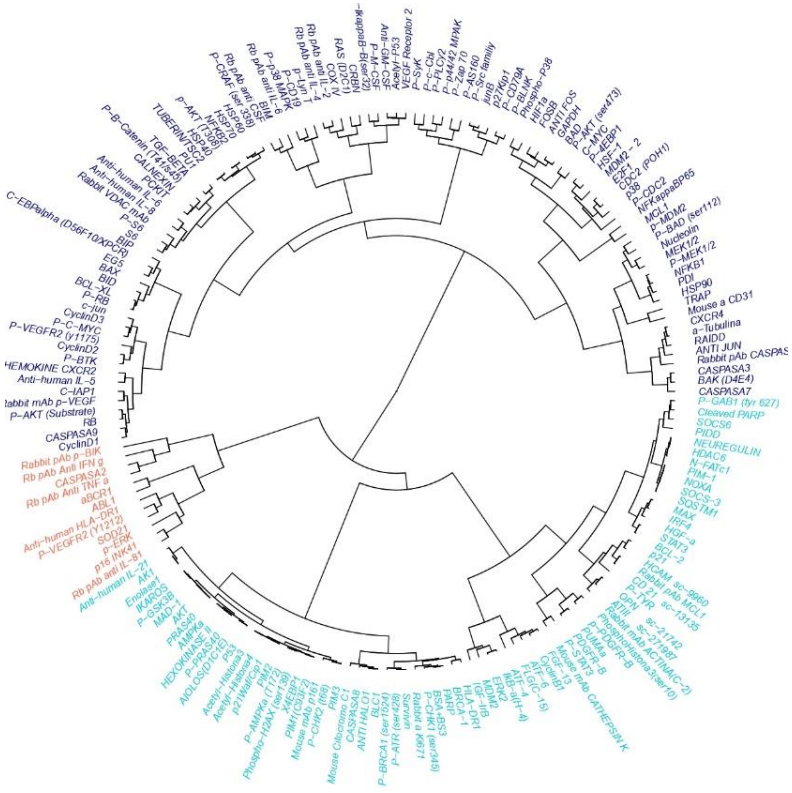


*Gráfico 6. Dendrograma a partir del método de WARD, microarray 11. Gráfico realizado con el paquete "ape" del software R.*

El resto de dendrogramas de cada uno de los microarrays, se encuentran en: *(Anexos, Gráficos 51-54, -Cluster Dendograma-).*

A partir de este algoritmo de clúster se realizó un *Heatmap* utilizando los datos anteriores *(Gráfico 7):*

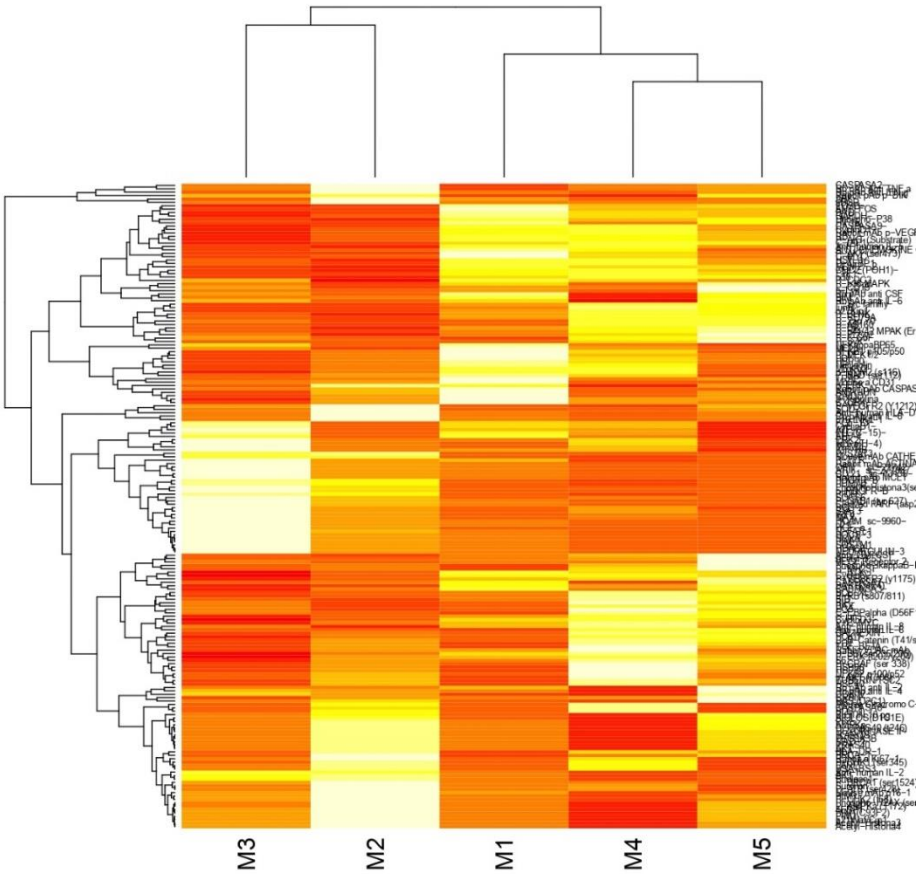



*Gráfico 7. Heatmap del microarray 11, diferenciando entre las 193 proteínas y las 5 muestras. Gráfico realizado con el paquete "ggplot2" del software RStudio.*

A la derecha del gráfico se pueden ver las proteínas que contiene la base de datos del microarray 11, y a la izquierda la agrupación de todas ellas. A través de dicho gráfico, se observa como las muestras se agrupan y presentan perfiles proteícos similares.

En función de la intensidad del color, se puede analizar la presencia o ausencia de las proteína en las diferentes muestras. Si el color es blanco quiere decir que no hay presencia de esa proteina en una determinada muestra, en cambio cuanto más fuerte sea el color, mayor intensidad tendrá.

El resto de *Heatmaps* de cada uno de los microarrays, se encuentran en: *(Anexos, Gráficos 55-58, -Cluster Dendograma-).*

- Análisis Cualitativo de los datos: *ScatterPlot.*

Para analizar si el comportamiento de una misma muestra es igual en diferentes *slides*, se compara el número de *spots* para una proteína en concreto en la muestra, pero ubicadas en diferentes microarrays. Tomando como ejemplo la muestra 3 del segundo conjunto de microarrays *(Figura J)*, teóricamente si las muestras son iguales y ambas se

encuentran ubicadas en la misma posición en los dos microarrays y han sido escaneadas con los mismos ajustes, ambas muestras deberían tener un número similar de proteínas y de *spots*. Para comprobarlo se realizará un *ScatterPlo*t comparando el número de *spots* de ambas muestras *(Gráfico 8):*

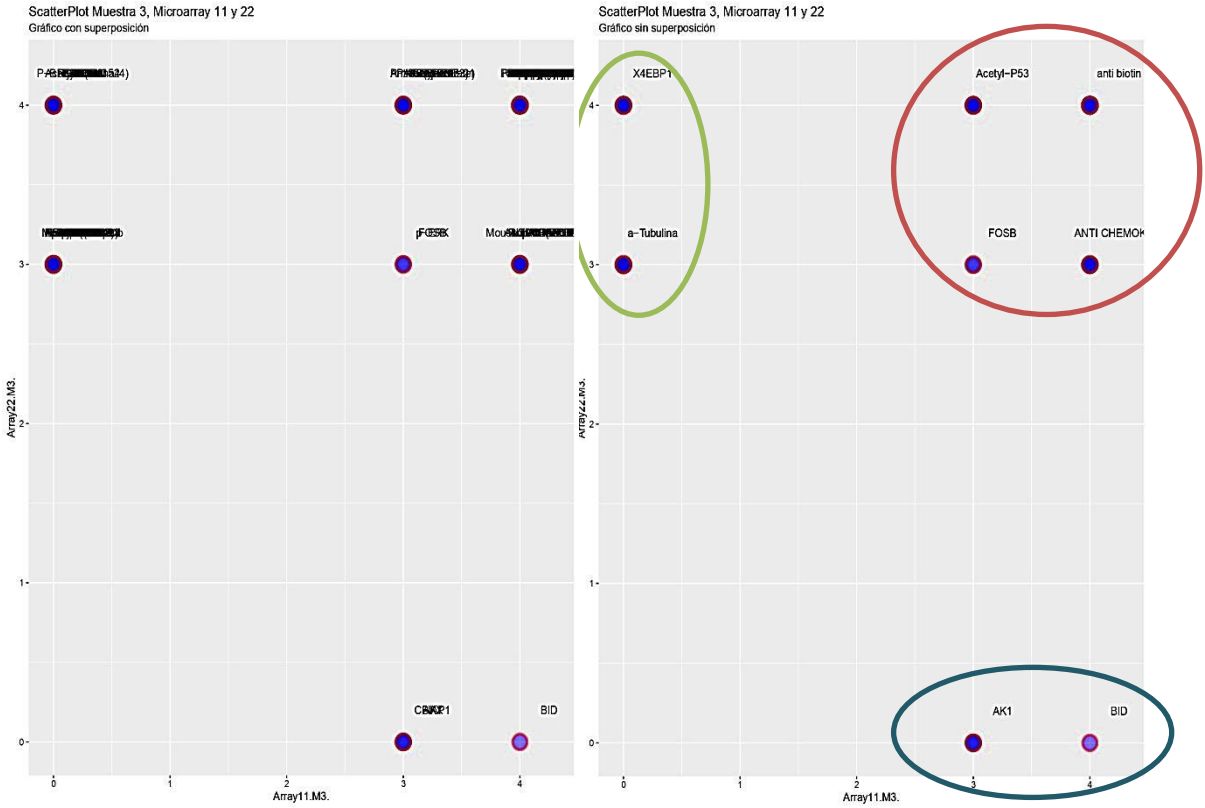



*Gráfico 8. ScatterPlot del microarray 11 y 22, diferenciando entre la muestra 3 de los dos microarrays, diferenciando entre las proteínas comunes. En el eje de ordenadas se encuentra el número de spots que pertenecen al microarray 22, en el eje de abscisas los del microarray 11. Gráfico realizado con el paquete "ggplot2" del software RStudio.*

El gráfico muestra una mayor intensidad en cada uno de los puntos, en función de todas las proteínas que esté representando en esa coordenada. En la muestra 3 del microarray 11 hay un total de 111 proteínas diferentes, en cambio en el microarray 22, dan señal positiva 143. Solamente hay 3 proteínas del microarray 11, que no están presentes en el microarray 22, en definitiva, a pesar de que ambas muestras tengan las mismas condiciones, existen diferencias entre ellas.

También se puede comprobar que la gran mayoría de proteínas dan señal con un número máximo de 3 /4 spots. Aunque existan 6 spots que correspondan a una determinada proteína, solamente dan señal positiva un máximo de 4. Por tanto, se puede considerar que el 100% de los datos corresponde a 4 *spots*.

Para tener una mayor fiabilidad de los resultados obtenidos en los microarrays, se cruzaron todas las proteínas resultantes de los microarrays, con la base de datos proveniente de otro estudio de pacientes con LLC, mediante espectrometría de masas,

con el fin de hallar la intersección entre ambas. Las proteínas que estuvieran presentes en ambas tecnologías tendrían mayor probabilidad de ser propias de esta patología.

La base de datos correspondiente a la espectrometría de masas se caracteriza por presentar un total de 2979 proteínas y 40 variables, algunas de estas son: número de péptidos *(Peptid Count)*, localización cromosómica, nombre del gen al que pertenecen, descripción de cada una de las proteínas, nomenclatura de la proteína en UniProt *(Canonical Name).*

Para poder cruzar las bases de datos obtenidas a través de la tecnología del microarray, es necesario obtener el nombre canónico de cada una de las proteínas, ya que es el nombre de acceso a la proteína que se encuentra en el espectrómetro. El nombre canónico lo proporcionan bases de datos como UniProt o NeXtProt *(Anexos, Tablas 53-74, Cruce de resultados con UniProt/ NeXtProt y por espectrometría de masas).*

UniProt se trata de una base de datos de acceso libre que proporciona la secuencia de proteínas, así como información funcional, la gran mayoría de entradas se derivan de proyectos de secuenciación del genoma. A partir de ella se puede obtener información sobre la función biológica de las proteínas derivadas de la literatura de investigación (Pichler, Warner, & Magrane, 2018).

NeXtProt se basa en una plataforma de conocimiento en línea sobre proteínas humanas, proporciona información acerca de su función, ubicación subcelular, expresión e interacción y papel en patologías. La mayor parte de la información se obtiene de la base de datos UniProt Swiss-Prot (Gaudet et al., 2017).

Tomando una de las listas finales con mayor número de proteínas resultantes (142 proteínas de un total de 217), que corresponde con la base de datos del microarray 22 de la muestra 3 *(Anexos, Tabla 63, Cruce de resultados con UniProt/ NeXtProt y por espectrometría de masas).*

Con los nombres canónicos de cada una de ellas, se obtiene la intersección entre ambas tecnologías, por ejemplo, tomando los datos de la muestra 3 del microarray 22, se encuentra 38 proteínas comunes, lo que permitirá comenzar a trabajar con una variable más: el número de péptidos *(Peptid Count).*

Para la integración de todos los datos, se crea una nueva base de datos que contenga datos cuantitativos y cualitativos de ambas técnicas proteómicas. En aquellas proteínas que no se encuentren mediante espectrometría de masas, se les asignará un valor de 0.

Actualmente, en el ámbito de la Proteómica, es fundamental visualizar de forma conjunta los resultados obtenidos en diferentes técnicas proteómicas. A través de gráficos *Circos*, se puede observar la intersección de los resultados de la tecnología de los microarrays de proteínas junto con los resultados de espectrometría de masas. Circos utiliza un diseño de ideograma circular para facilitar la visualización de las relaciones entre las muestras, la visualización de la identificación y el análisis de similitud y diferencias entre ambas técnicas y permite presentar los datos en forma de gráficos de

dispersión, líneas, histogramas, texto…(Krzywinski et al., 2009). Por ejemplo, en el microarray 11, se visualiza tanto los datos cuantitativos como cualitativos obtenidos en ambas técnicas *(Gráfico 9):*

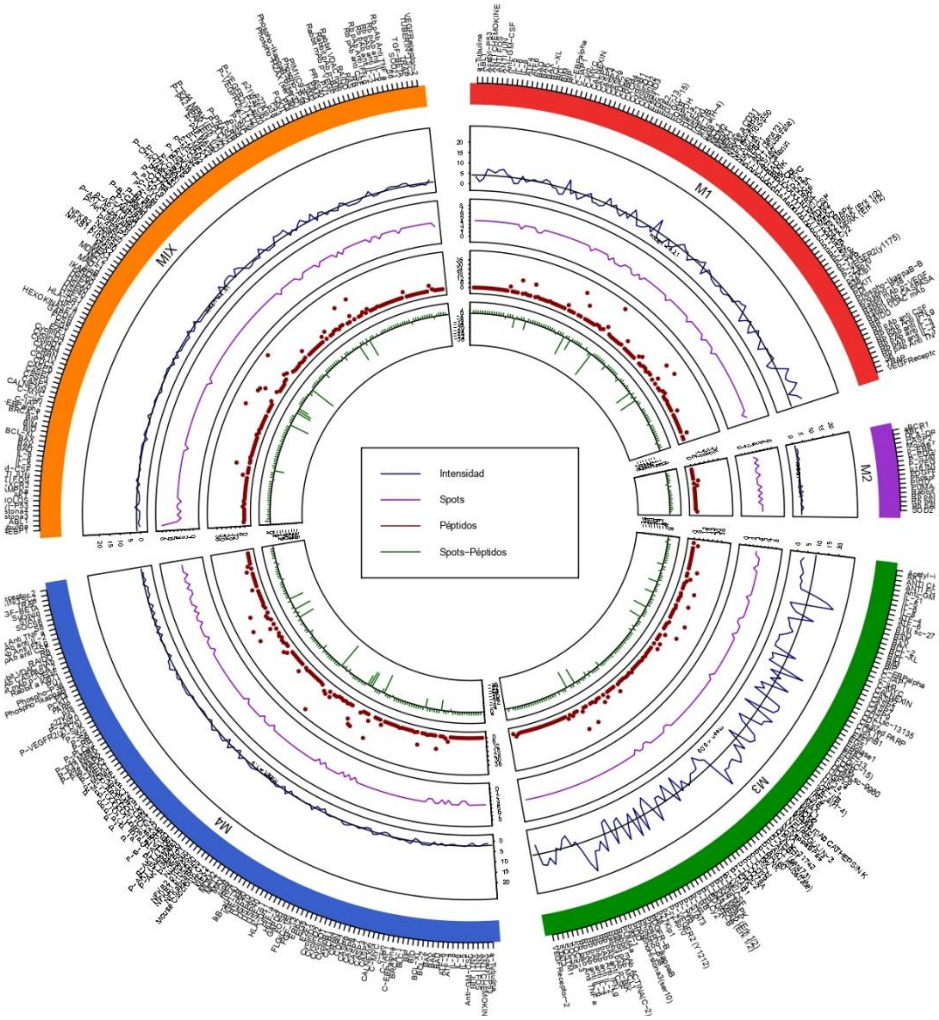



*Gráfico 9. CircoPlot del microarray 11, visualizando la integración de ambas técnicas. Gráfico realizado con el paquete "circlize" del software R.*

Desgranando el gráfico desde la capa más externa a la más interna, se obtiene la señal de intensidad escalada de cada una de las muestras, el número de *spots*, el número de péptidos proporcionado por el *MS/MS* y la correlación entre ambas técnicas. Se destaca la capa más externa, ya que es el número de proteínas que contiene cada subarray, y en función de esta, variarán el resto de capas; atendiendo al número de proteínas encontradas, el subarray 5 *(MIX)* contiene 130 proteínas, seguido por el subarray 4 con 126, el más pequeño de todos pertenece al subarray 2 con solamente 19 proteínas. El resto de *Circos*, propios de otros microarrays, se encuentran en: *(Anexos, Gráficos 59-62, CircoPlot).*

A raíz de este gráfico se analizó la relación entre el número de péptidos y el número de *spots*, con el fin de comprobar si el número de *spots* de las proteínas encontradas en los microarrays se relacionan con el número de péptidos. Para ello se obtiene una nueva

base de datos solamente con las proteínas correlacionadas con el espectrómetro de masas, ya que solamente estas proporcionan el número de péptidos. Se aplicarán dos test no paramétricos, el test de Kruskal Wallis y el test de Spearman con el propósito de contrastar:

> $H_0 =$ *no existe dependencia entre las dos variables, el número de spots y de péptidos no está relacionado.*
>
> $H_1 =$ *existe dependencia entre las dos variables, las proteínas muestran mayor número de péptidos en función del número de spots que tengan.*

El test de Kruskal Wallis es una prueba estadística no paramétrica que evalúa las diferencias entre 3 o más grupos muestreados independientemente en una sola variable continua no distribuida normalmente (McKight & Najab, 2010).

Tomando el subarray 5 del microarray 11, ya que presenta mayor número de proteínas correlacionadas: 38 proteínas. Se aplica el test de *Kruskal Wallis*:

```
> library("car", lib.loc="~/R/win-library/3.5")
> library("carData", lib.loc="~/R/win-library/3.5")
> Test.MIX.array.11 <- read.csv("C:/Users/HELENA/Desktop/Test MI
X array 11.csv", sep =",", header = T, dec = ".")
> Tt<-Test.MIX.array.11
> attach(Tt)
The following objects are masked from Tt (pos = 3):

    Name, Peptidos, Spots
> kruskal.test(Peptidos ~ Spots, data = Tt)

  Kruskal-Wallis rank sum test

data:  Peptidos by Spots
Kruskal-Wallis chi-squared = 1.9193, df = 2, p-value = 0.383
```

Al tener un p-valor superior a 0.05 (*p-value* = 0.383 > 0.05) se acepta $H_0$ , las dos variables son independientes, las proteínas no muestran mayor número de péptidos en función del número de *spots* que tengan. Se aplicó también el test de Spearman con el fin de asegurar que realmente no existe dependencia.

La prueba de Spearman se aplica para analizar la asociación monótona entre 2 variables, es una medida de correlación, donde la la hipótesis a contrastar es si existe relación entre las variables. Se caracteriza por ordenar los datos y reemplazarlos por su respectivo rango, el coeficiente de correlación varía entre el -1 y el 1 $(-1 < \rho < 1)$, cuanto más se aproxime a 0 menor será la relación que exista entre las variables, un $\rho$ negativo próximo a -1 se puede interpretar como una relación lineal alta negativa, en cambio, si es positivo, se interpretará como una relación lineal alta positiva.

Realizando un gráfico que represente la matriz de correlaciones entre las dos variables *(Gráfico 10)*, se observa que realmente no existe correlación entre las dos variables, ya que la señal de intensidad entre las dos variables se encuentra en el intervalo $(-0.2, 0.2)$.

```
> library(corrplot)
> M<-cor(Tt)
> corrplot(M, method = "pie")
```

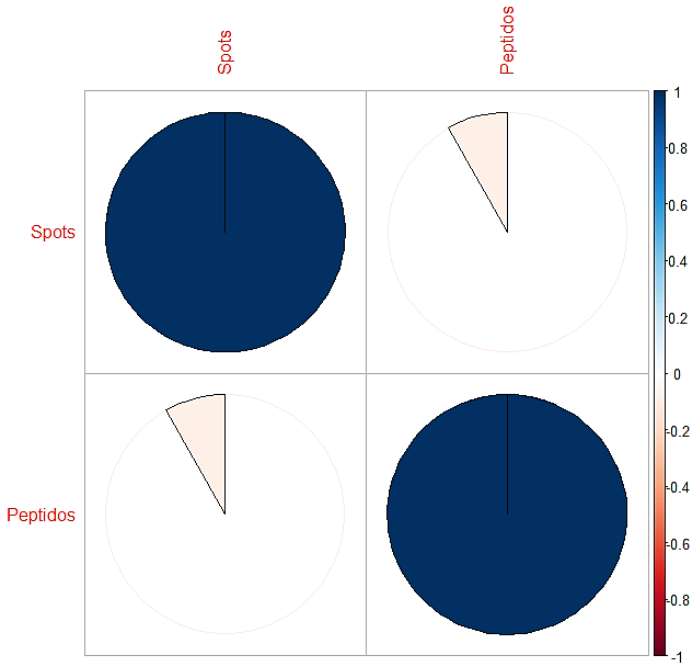



*Gráfico 10. Representación gráfica de la matriz de correlación de la muestra 5 (MIX), a la derecha se encuentra un baremo de color, el cual muestra que en 0 no hay ningún tipo de correlación y en 1/-1 una correlación alta entre las variables. Gráfico realizado con el paquete "corrplot" del software RStudio.*

```
> cor(Peptidos, Spots, method = "spearman")
[1] -0.1278912
> cor.test(Peptidos, Spots, method = "spearman")

  Spearman's rank correlation rho

data:  Peptidos and Spots
S = 10308, p-value = 0.4442
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho = -0.1278912
```

El coeficiente de correlación de Spearman da como resultado ρ = -0.128 y no es significativo ya que su p-value=0.4442 > 0.05, debido a que ρ está próximo a 0, quiere decir que no hay relación monótona entre estas dos variables. Una vez aplicado estos dos test estadísticos se puede afirmar que no hay evidencias suficientes para admitir la dependencia entre estas dos variables, por tanto, se considerará la independencia entre ambas.

Debido a la gran cantidad de proteínas encontradas en cada uno de los microarrays, con el objetivo de disminuir la dimensionalidad de estos y poder agrupar las proteínas en grupos que expliquen aproximadamente lo mismo que las proteínas originales, se aplicó un PCA *(Principal Component Analysis)*. Esta técnica multivariable simplifica la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información, los componentes se ordenan por la cantidad de varianza original que

describen («RPubs - Análisis de Componentes Principales (Principal Component Analy sis, PCA) y t-SNE», s. f.). Tomando todas las proteínas del microarray 11, el resultado que se obtiene al realizar un PCA *(Gráfico 11):*

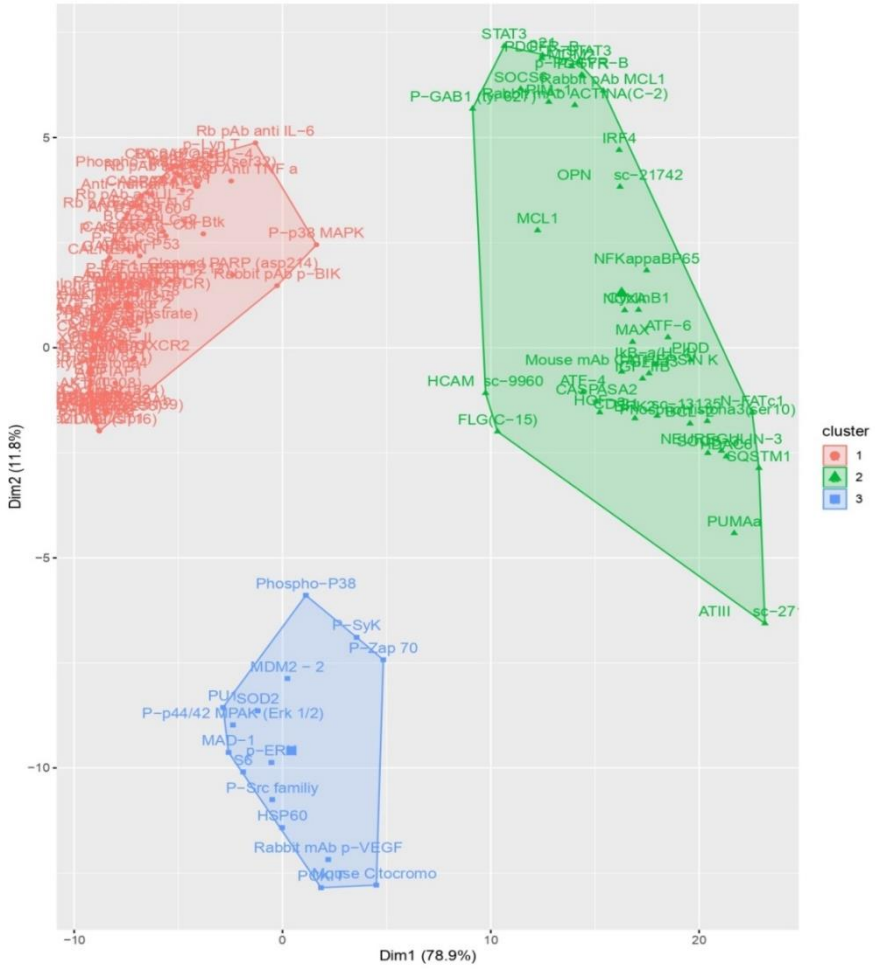



*Gráfico 11. Análisis de Componentes Principales, del microarray 22, con las 150 proteínas originales que lo componen. Gráfico realizado con el paquete "ggplot2" del software RStudio.*

El PCA muestra 3 grupos principales de proteínas, independientemente de las muestras ubicadas en dicho microarray, lo mismo ocurre con el PCA para el microarray 22, *(Anexos, Gráfico 64, PCA microarray 22)*:

i. Cascada del BCR (Receptor de la célula B) – clúster verde.
ii. Factores de transcripción relacionados con dicha cascada – clúster azul.
iii. Otras rutas de activación del linfocito B (que no son BCR) – clúster rojo.

A raíz de este PCA se empieza a analizar la relación entre las proteínas y la razón por la que han dado señal positiva en el microarray. Para estudiar la relación entre dichas proteínas se emplearán diversas técnicas y herramientas bioinformáticas para crear las redes de interacción, hallar la localización celular, sus vías de señalización…,a través de KEGG, GO, OMIM, entre otras. El primer análisis se realizará con la herramienta de búsqueda STRING, es una base de datos que contiene información de numerosas fuentes, incluidos datos experimentales, métodos de predicción computacional y textos

públicos. A través de las redes de interacción proteína-proteína se pueden estudiar los procesos celulares a nivel de sistema, estas redes se pueden usar para filtrar y evaluar datos genómicos-funcionales, además de proporcionar una plataforma intuitiva con el fin de obtener información sobre las propiedades estructurales, funcionales y evolutivas de las proteínas.(Schwartz, Yu, Gardenour, Finley, & Ideker, 2009)

Para que STRING reconozca el archivo, se necesita renombrar los nombres de las proteínas con los "*Query Terms*" *(Anexos, Tabla 75, "Query Terms").* Los *"Query Terms"* hacen referencia al termino propio de la proteínas.

Introducidos todos los datos, se obtiene un gráfico de interacciones similar al *Gráfico 13:*

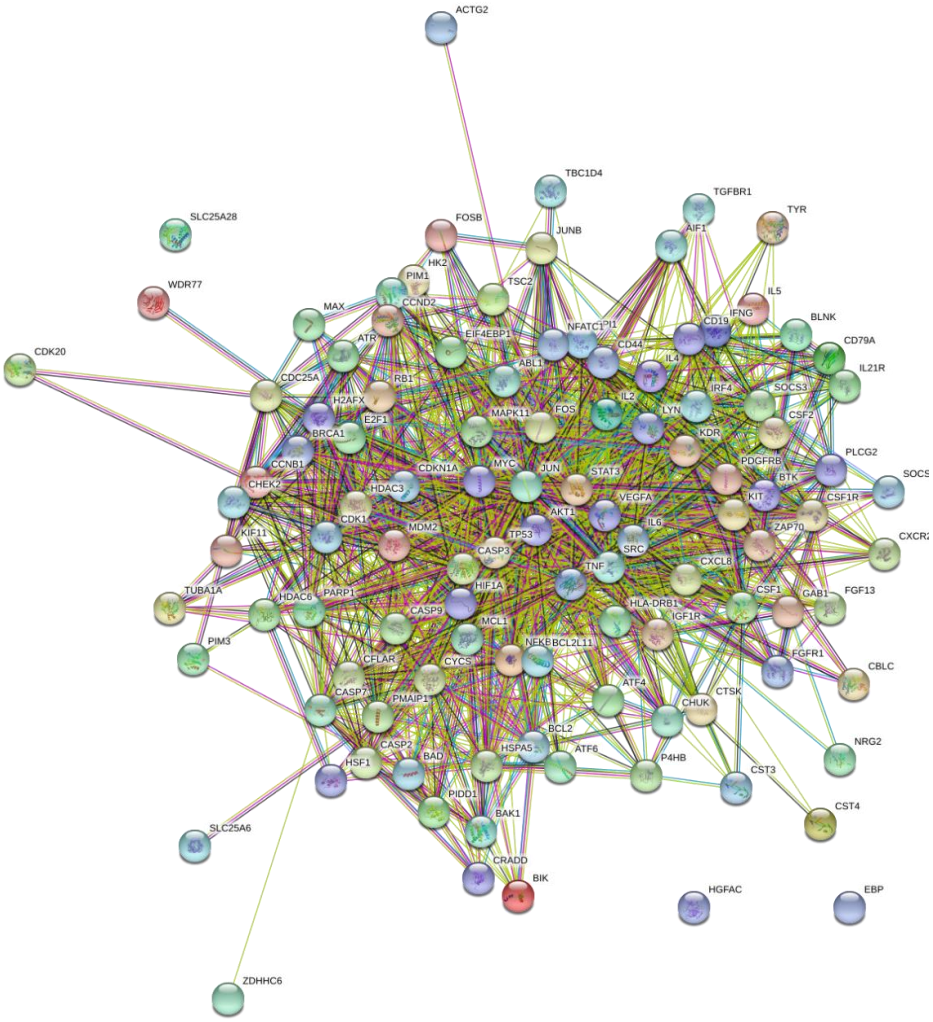



*Gráfico13. Red de interacción entre proteínas creada en STRIG, muestra 3 microarray 22.*

Los nodos de la red representan las proteínas analizadas. Los nodos coloreados representan una interacción directa entre proteínas, en cambio, los nodos blancos representan una relación indirecta. Cada una de las aristas representa asociaciones entre proteínas, las asociaciones están destinadas a ser específicas y significativas. Dependiendo del color de la arista, las relaciones pueden ser:

i. *Interacciones conocidas (azul claro y purpura):* el azul representa interacciones conocidas de bases provenientes de bases de proteínas curadas y el violeta simboliza interacciones conocidas determinadas de forma experimental.

*ii.* *Interacciones predichas (azul oscuro, rojo y verde):* el azul oscuro representa coocurrencia de genes, el rojo simboliza fusiones genéticas y el verde genes vecinos *("gene neighborhood").*

*iii.* *Otras interacciones (verde pistacho, negro y violeta):* las asociaciones de color negro representan la co-expresión, el violeta la homología de proteínas y el verde pistacho representa asociaciones entre proteínas basadas en publicaciones *(PubMed).*

En esta red de interacción predominan las asociaciones conocidas e interacciones basadas en publicaciones. La gran mayoría de proteínas se encuentran relacionadas entre sí, seleccionando uno de los nodos que se encuentran en el centro de la red y que mayor número de interacciones tiene: *(Gráfico 14)*

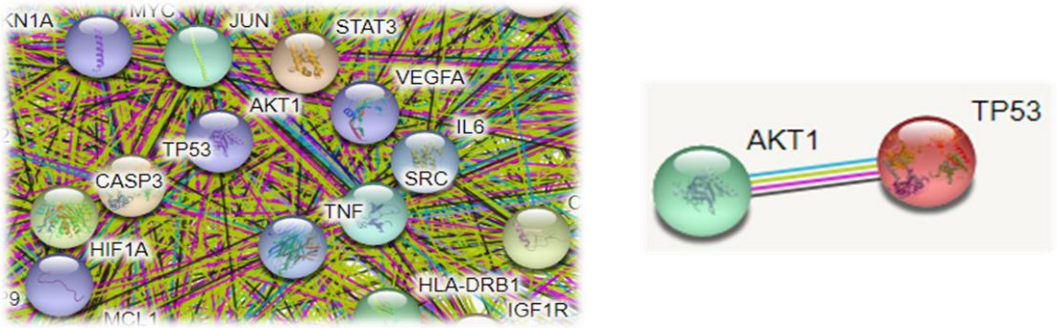



*Gráfico14. Red de interacción entre proteínas creada en STRIG, proteína (AKT1-TP53), muestra 3 microarray 22.*

*AKT1* se encuentra en el centro de la red, y tiene múltiples interacciones con otros nodos; esta regula diversos procesos, como el metabolismo, la supervivencia celular, …. Dicho nodo se encuentra muy próximo a *TP53, STAT3, VEGFA, TSC2*…lo que quiere decir que existe cierta interacción entre ellas , por ejemplo, tomando la proteína *TP53* se observa que tiene interacciones conocidas basadas en bases de datos curadas y determinadas experimentalmente, junto a publicaciones *(Anexos, Tabla 76, Publicaciones de referencia: AKT1/TP53)* y de co-expresión, todas estas interacciones se consideran evidencias que sugieren un enlace funcional (tiene una puntuación combinada -*combined score*-, entre las 4 interacciones, de 0.994 sobre 1), siendo *TP53* un antígeno tumoral celular. A través de *Reactome_Pathways*, se observa las vías en las que ambas proteínas, se encuentran implicadas *(Anexos, Tabla 77, Vías de señalización).* Todas las funciones moleculares se pueden ver en: *(Anexos, tabla 78, Función molecular, GO_Ontology).* Analizando las funciones de ambas proteínas, hace sentido que estén relacionadas; además, teniendo en cuenta las palabras clave halladas en UniProt *(Anexos, Tabla 79, UniProt_keywords),* en la cual aparecen términos como apoptosis, enfermedad, proto-oncogén, entre otras.

Para contrastar dicha información se examinarán las rutas de señalización a través de *Reactome.*

*Reactome_Pathway* es una biblioteca gratuita, de código abierto y curada, proporciona herramientas bioinformáticas de visualización, interpretación y análisis del conocimiento en diferentes vías de señalización.(Fabregat et al., 2017). En el *Gráfico 13* se puede observar las rutas de señalización de las proteínas, correspondientes a la muestra 3, para comprobar los nombres de las rutas:*(Anexos, Pathways names, tabla 80)*

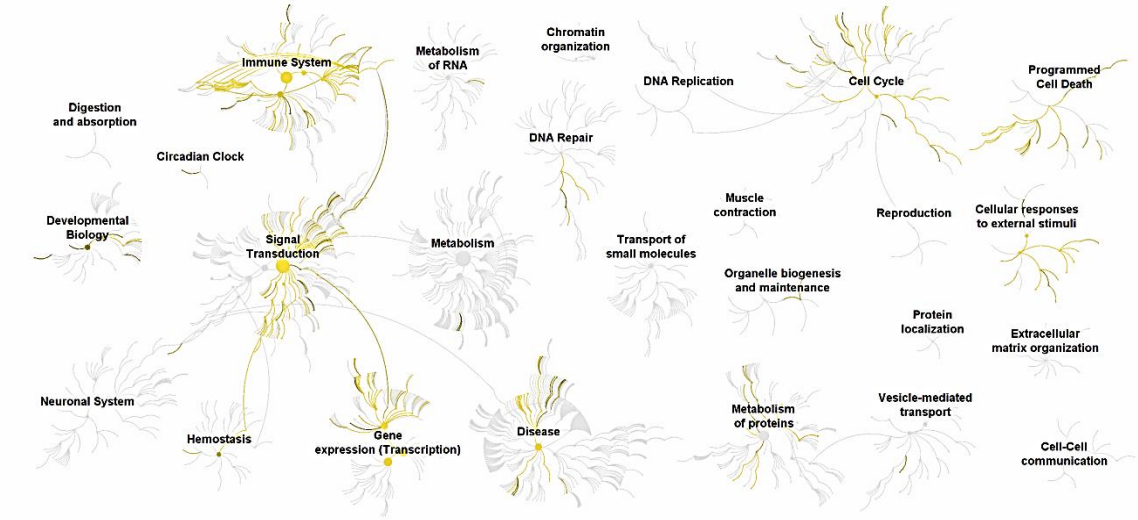



*Gráfico 13. Gráficos con las diferentes vías de señalización de la muestra 3, microarray 22. Gráfico proporcionado por Reactome.org.*

Se encuentran señaladas vías del sistemna inmune, pertenecientes al ciclo celular, hemostasis, muerte celular programadas…Si se tiene en cuenta la jerarquía de eventos, el proceso sería el indicado en la tabla 3.A/ 3.B: *(Tabla 3.A. Jerarquía de eventos, resultado obtenido en Reactome_Pathway: primera parte de la tabla).*

| ***Event Hierarchy*** | ***Count*** | ***FDR*** |
|---|---|---|
| *Cell Cycle* | (22/622) | 5,11E-05 |
| *Cell-cell communication* | (1/130) | 7,80E-10 |
| *Chromatin Organization* | (4/240) | 3,03E-01 |
| *Circadian Clock* | (2/68) | 1,86E-01 |
| *Development Biology* | (19/1054) | 6,71E-27 |
| *Digestion and absortion* | - | - |
| *Disease* | (34/1177) | 1,10E-05 |
| *DNA Repair* | (7/313) | 7,23E-02 |
| *DNA Replication* | (1/127) | 7,72E-01 |
| *Extracellular matrix organization* | (6/301) | 1,38E-01 |
| *Gene Expression (Transcription)* | (44/1498) | 8,64E-08 |
| *Hemostasis* | (15/722) | 4,64E-02 |
| *Inmune System* | (57/2233) | 2,73E-08 |
| *Metabolism* | (13/2132) | 9,98E-01 |
| *Metabolism of proteins* | (30/2010) | 7,98E-02 |
| *Metabolism or RNA* | (4/675) | 9,57E-01 |
| *Muscle Contraction* | (2/209) | 6,99E-01 |
| *Neuronal System* | (4/419) | 7,19E-01 |
| *Organelle biogenesis and maintenance* | (7/298) | 6,71E-02 |
| *Programmed Cell Death* | (16/183) | 2,83E-08 |

*(Tabla 3.B. Jerarquía de eventos, resultado obtenido en Reactome_Pathway: segunda parte de la tabla).*

| ***Event Hierarchy*** | ***Count*** | ***FDR*** |
|---|---|---|
| *Protein localization* | - | - |
| *Reproduction* | (3/144) | 1,47E-01 |
| *Signal Transduction* | (70/2758) | 1,00E-10 |
| *Transport of small molecules* | (1/731) | 1,00E+00 |
| *Vesicle-mediated transport* | (6/761) | 8,81E-01 |

Si se selecciona la vía de señalización correspondiente a patología (*Disease*), lo cual presenta un FDR bajo, se puede observar como parten de ella diferentes redes. Las cuales forman parte del sistema inmunológico, neurodegenerativas, infecciosas… Cada una de las aristas representa una ruta de señalización diferente: *(Gráfico 14)*

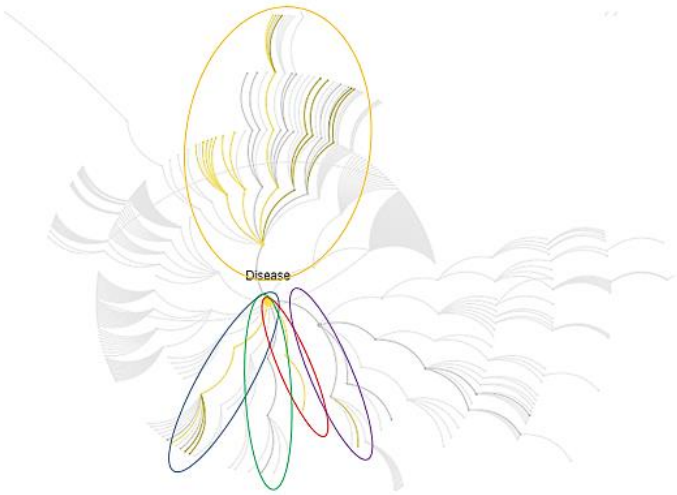



*Gráfico 14. Gráfico con las diferentes rutas dentro de la vía de señalización de la patología. Gráfico proporcionado por Reactome.org.*

1. Enfermedades propias del Sistema Inmune (4/27) FDR:1.79E-3:
   Enfermedades asociadas con la cascada de señalización de TLR.
2. Enfermedades Neurodegenerativas (4/22) FDR: 9.61E-4:
   Desregulación de *CDK5* que desencadena múltiples rutas de señalización en modelos de Alzheimer.
3. Trastornos del desarrollo (1/13) FDR: 1.4E-1:
   Pérdida de la función de *MECP2* en el síndrome de *Rett.*
4. Enfermedades infecciosas (5/463) FDR: 6.25E-1:
   Captación y acción de toxinas bacterianas.
5. Enfermedades propias de señal de transducción celular (26/403) FDR: 1.25E-10:
   Señalización por *EGFR, FGRF, P13K/AKT, NOTCH1, TGF-B Receptor Complex* en cáncer y señalización del oncogén *MAPK.*

Dentro de las opciones anteriores la vía con menor FDR corresponde a la ruta de señalización de las enfermedades relacionadas con la señal de transducción celular, la cual engloba muchas más. Entre ellas destaca P13K/AKT por la señalización en cáncer con un p-valor de 4.5E-13 *(< 0.005)* y un FDR de 8.43E-11. De 113 proteínas que

contiene esta vía, 16 están presentes en la base de datos *(Anexos, Tabla 81, Vía de señalización en cáncer: PI3K /AKT).*

La vía *PI3K/AKT* es una de las rutas más desreguladas en las mayor parte de los cánceres. La sobreexpresión anormal de *AKT* es una característica común en los cánceres, tanto tempranos como avanzados. (Mundi, Sachdev, McCourt, & Kalinsky, 2016). Por tanto, es un importante centro de señalización celular; el cual contribuye a la alteración de la apoptosis y la progresión del ciclo celular en las células cancerosas mediante la inhibición de *BAD* y *CASPASE9*. (Mundi et al., 2016). Tras el análisis de la ruta de *AKT* en *Reactome* se observan nodos de diferentes colores, los nodos con mayor intensidad representan proteínas que se encuentran en la base de datos cargada, por lo tanto, se puede afirmar que proteínas como *CDKN, BAD, MDM2, BCL2…*pertenecen a vías de señalización en cáncer: *(Gráfico 15)*

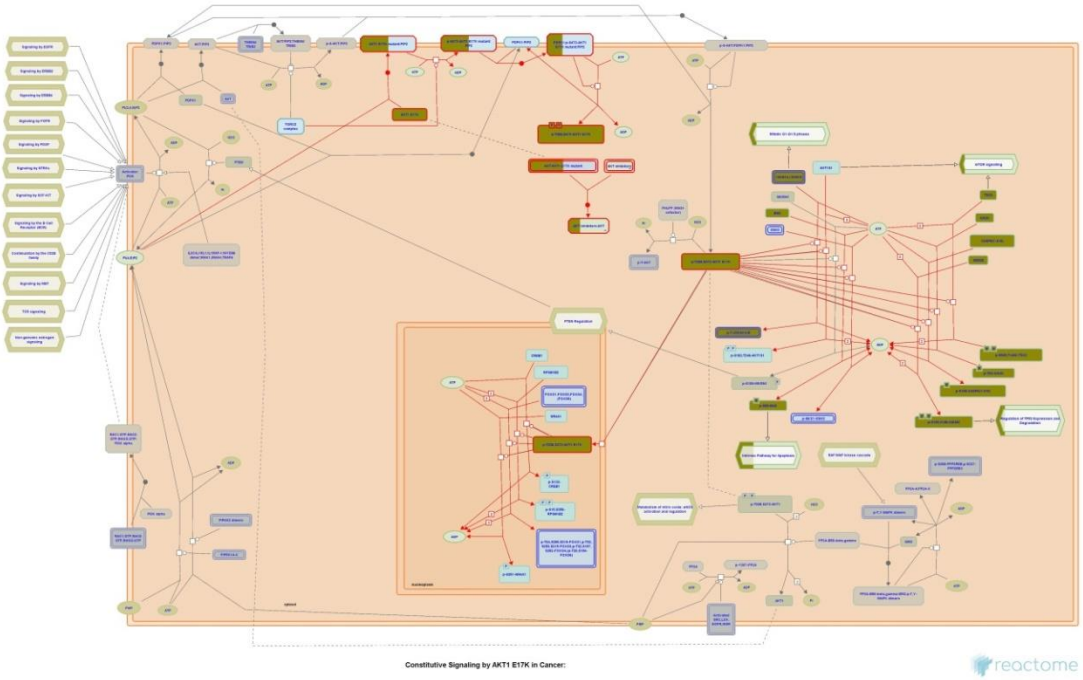



*Gráfico 15. Ruta de señalización de AKT, mostrando las proteínas que estan presentes en la base de datos y forman parte de la ruta de señalización. Gráfico proporcionado por Reactome.org.*

A través de *Enrichr* se puede comprobar que proteínas están presentes una determinada patología. En *Reactome* se ha comprobó anteriormente que había diferentes rutas de señalización que pertenecían a patologías, ahora se estudiarán en qué enfermedades están implicadas.

Un análisis de enriquecimiento analiza conjuntos de genes generados por experimentos de todo el genoma. *Enrichr* contiene una colección de diversas bibliotecas de conjuntos de genes, disponibles para su análisis y descarga; actualmente tiene entre 180-184 conjuntos de genes anotados de 102 bibliotecas de conjuntos de genes. Es un recurso integral para conjuntos de genes curados y un motor de búsqueda que acumula conocimiento biológico.

La opción *OMIM Disease* es una biblioteca que contiene genes humanos y trastornos genéticos, con especial atención a la relación entre la variación genética molecular y la expresión fenotípica, la cual reconoce proteínas que son propias de una determinada enfermedad *(Tabla 5. Proteínas implicadas en una determinada enfermedad, resultado obtenido en Enrichr.com.)*

| *Term* | *P-value* | *Z-score* | *Combined Score* | *Genes* |
|---|---|---|---|---|
| *Lymphoma* | 0,00033 | -1,7905 | 14,36 | *MYC;IL21R;BCL2* |
| *Breast_cancer* | 0,00068 | -1,7776 | 12,96 | *CHEK2;AKT1;TP53* |
| *Leukemia* | 0,00141 | -1,4293 | 9,38 | *PDGFRB;KIT;BCL2;ABL1* |
| *Colorectal_Cancer* | 0,00194 | -1,1702 | 7,31 | *CHEK2;AKT1;TP53* |
| *Immunodeficiency* | 0,01297 | -0,2176 | 0,95 | *NFKBIA;IL2* |
| *Diabetes* | 0,37332 | 2,2786 | -2,25 | *IL6* |
| *Anemia* | 0,3161 | 2,2136 | -2,55 | *IFNG* |
| *Blood* | 0,20076 | 3,0134 | -4,84 | *CD44* |
| *Ovarian_Cancer* | 0,00284 | 0,8379 | -4,91 | *AKT1;BRCA1* |
| *Prostate_Cancer* | 0,17032 | 3,0906 | -5,47 | *CHEK2* |
| *Schizophrenia* | 0,17032 | 3,5816 | -6,34 | *AKT1* |
| *Asthma* | 0,13335 | 3,7889 | -7,63 | *TNF* |
| *Migraine* | 0,09475 | 3,8979 | -9,19 | *TNF* |
| *Macular_Degeneration* | 0,10595 | 4,3356 | -9,73 | *CST3* |
| *Malaria* | 0,08341 | 4,2549 | -10,57 | *TNF* |
| *Hypogonadism* | 0,0891 | 4,4235 | -10,7 | *FGFR1* |
| *Dementia* | 0,07193 | 4,3156 | -11,36 | *TNF* |
| *Rheumatoid_arthritis* | 0,07769 | 4,9944 | -12,76 | *IL6* |
| *Pancreatic_Cancer* | 0,06614 | 4,7538 | -12,91 | *TP53* |
| *Ectodermal_Dysplasia* | 0,06614 | 4,9256 | -13,38 | *NFKBIA* |
| *Skin/Hair/Eye_Pigmentation* | 0,07193 | 5,7686 | -15,18 | *TYR* |

El *p-valor* se calcula a partir de la prueba estadística de Fisher, que es una prueba de proporción que asume una distribución e independencia binomial para la probabilidad de que cualquier gen pertenezca a cualquier conjunto. El conjunto de genes *MYC, IL2IR, BCL2* tienen un p-valor menor a 0.005, por tanto, se puede afirmar que sí están y son propias del Linfoma. El *Z-Score* se calcula para evaluar la desviación del rango esperado.

Por último, *Combined Score* se calcula tomando el *p-valor* de la prueba de Fisher y multiplicando la desviación del rango esperado proporcionado por *Z-Score.* Finalmente, teniendo en cuenta ambas pruebas, solo se consideran aquellas patologías, cuyos genes tengan una puntuación combinada positiva.

Además, *Enrichr* proporciona diferentes herramientas de visualización *(Gráfico 16)*, para esta tabla se ha creado un Clustergrama, dicho gráfico permite visualizar y analizar datos de alta dimensión agrupados jerárquicamente, interactivos y compartibles.

El *Gráfico 16* indica los diferentes genes que determinan una cierta enfermedad, mostrando su puntuación combinada (dato que se muestra en la tabla anterior REF), muestra aquellas patologías con una puntuación combinada positiva .


*Gráfico 16. Clustergrama de la opción: OMIM Disease de Enrichr, muestra los genes implicados en una cierta patología, junto con su puntuación combinada, en el eje X se encuentran las proteína y en el eje Y las patológias, con su puntuación combinada. Gráfico proporcionado por Enrichr.com.*

Por último, se realizará un análisis *DAVID.* Esta es una herramienta bioinformática que integra anotaciones genómicas funcionales con resúmenes gráficos intuitivos, las listas con identificadores de genes o proteínas, se anotan y se resumen de acuerdo con los datos categóricos compartidos para asociar una colección de genes con un término biológico funcional de forma sistemática *(GO_Ontology),* dominio de proteínas *(Protein_Domain)* y miembros de rutas de señalización bioquímicas *(Biocheminal Pathway Membership)*(Sherman et al., 2007) .

Para acceder al análisis *DAVID* se utilizaron los nombres canónicos que asignan UniProt y NeXtProt a las proteínas, se toma el nombre canónico de todas las proteínas que componen la muestra 3 del microarray 22 *(Anexos, Tabla 63, -Canonical Name-).*

*DAVID* reconoce 122 proteínas de un total de 142. Estas 122 proteínas son agrupadas dependiendo de la función que desempeñen en el organismo, su ubicación celular, la patología a la que pertenecen, entre otras.

Para analizar la ubicación celular de todas las proteínas, *GOTERM_CC_Direct* proporciona dicha información. *DAVID* reconoce 48 ubicaciones celulares diferentes, junto con el número de proteínas que contiene cada una de ellas y su *p-value, Fold Enrichment, Bonferroni, Benjamini, FDR (Tabla 5. Ubicación células de las proteínas, resultado obtenido en DAVID.com).* La hipótesis a contrastar será si X proteínas pertenecen a Y ubicación celular o no.

| *Term* | *Count* | *%* | *PValue* | *Fold Enrichm* | *Bonferroni* | *Benjamini* | *FDR* |
|---|---|---|---|---|---|---|---|
| *Cytosol* | 65 | 53,28 | 3,00E-18 | 2,9532 | 7,01E-16 | 7,01E-16 | 3,85E-15 |
| *Nucleus* | 66 | 54,10 | 1,82E-08 | 1,8357 | 4,27E-06 | 2,13E-06 | 2,34E-05 |
| *Nucleoplasm* | 43 | 35,25 | 7,15E-08 | 2,3263 | 1,67E-05 | 5,58E-06 | 9,20E-05 |
| *Cytoplasm* | 63 | 51,64 | 8,89E-08 | 1,8170 | 2,08E-05 | 5,20E-06 | 1,14E-04 |
| *Protein Complex* | 15 | 12,30 | 5,55E-07 | 5,4834 | 1,30E-04 | 2,60E-05 | 7,13E-04 |
| *Membrane Raft* | 10 | 8,20 | 8,96E-06 | 7,3112 | 2,09E-03 | 3,49E-04 | 1,15E-02 |
| *PML body* | 7 | 5,74 | 4,62E-05 | 10,7580 | 1,07E-02 | 1,54E-03 | 5,94E-02 |
| *Mitochondrion* | 23 | 18,85 | 5,16E-05 | 2,6026 | 1,20E-02 | 1,51E-03 | 6,63E-02 |
| *Perinuclear region of cytoplasm* | 15 | 12,30 | 5,91E-05 | 3,6380 | 1,37E-02 | 1,54E-03 | 7,60E-02 |
| *External side of plasma membrane* | 9 | 7,38 | 8,37E-05 | 6,3639 | 1,94E-02 | 1,96E-03 | 1,07E-01 |
| *Extrinsic component of cytoplasmic* | 6 | 4,92 | 8,52E-05 | 13,2893 | 1,97E-02 | 1,81E-03 | 1,09E-01 |
| *Nuclear chromatin* | 8 | 6,56 | 2,91E-04 | 6,2430 | 6,58E-02 | 5,66E-03 | 3,73E-01 |
| *Mast cell granule* | 4 | 3,28 | 3,40E-04 | 28,6879 | 7,64E-02 | 6,10E-03 | 4,36E-01 |
| *Mitochondrial outer membrane* | 7 | 5,74 | 4,59E-04 | 7,0757 | 1,02E-01 | 7,64E-03 | 5,88E-01 |
| *Membrane* | 28 | 22,95 | 9,38E-04 | 1,9169 | 1,97E-01 | 1,45E-02 | 1,20E+00 |
| *Endoplasmic Reticulum* | 15 | 12,30 | 0,0010983 | 2,7285 | 2,27E-01 | 1,59E-02 | 1,40E+00 |
| *Extracellular space* | 19 | 15,57 | 0,0030796 | 2,1244 | 5,14E-01 | 4,16E-02 | 3,89E+00 |
| *Cell* | 5 | 4,10 | 0,004312 | 7,5306 | 6,36E-01 | 5,46E-02 | 5,40E+00 |
| *Late endosome* | 5 | 4,10 | 0,0086514 | 6,1726 | 8,69E-01 | 1,01E-01 | 1,06E+01 |
| *Transcription factor complex* | 6 | 4,92 | 0,0090428 | 4,6822 | 8,81E-01 | 1,01E-01 | 1,10E+01 |
| *Multivesicular body* | 3 | 2,46 | 0,0091133 | 20,5379 | 8,83E-01 | 9,70E-02 | 1,11E+01 |
| *Golgi apparatus* | 13 | 10,66 | 0,0114721 | 2,2688 | 9,33E-01 | 1,15E-01 | 1,38E+01 |
| *CSF1-CSF1R complex* | 2 | 1,64 | 0,0131264 | 150,6116 | 9,55E-01 | 1,26E-01 | 1,56E+01 |
| *Extracellular exosome* | 29 | 23,77 | 0,014552 | 1,5538 | 9,68E-01 | 1,33E-01 | 1,72E+01 |
| *Extracellular region* | 19 | 15,57 | 0,0181239 | 1,7774 | 9,86E-01 | 1,57E-01 | 2,10E+01 |
| *Myelin sheath* | 5 | 4,10 | 0,0181315 | 4,9543 | 9,86E-01 | 1,52E-01 | 2,10E+01 |
| *Rb-E2F complex* | 2 | 1,64 | 0,0196254 | 100,4077 | 9,90E-01 | 1,58E-01 | 2,25E+01 |
| *B cell receptor complex* | 2 | 1,64 | 0,0196254 | 100,4077 | 9,90E-01 | 1,58E-01 | 2,25E+01 |
| *Cell surface* | 9 | 7,38 | 0,0269315 | 2,5009 | 9,98E-01 | 2,04E-01 | 2,96E+01 |
| *Intracellular* | 16 | 13,11 | 0,0286238 | 1,8091 | 9,99E-01 | 2,09E-01 | 3,12E+01 |
| *Melanosome* | 4 | 3,28 | 0,0291079 | 5,9648 | 9,99E-01 | 2,06E-01 | 3,16E+01 |
| *Cytoplasmic side of plasma membra* | 3 | 2,46 | 0,0311929 | 10,7580 | 9,99E-01 | 2,13E-01 | 3,35E+01 |
| *I-kappaB/NF-kappaB complex* | 2 | 1,64 | 0,0324964 | 60,2446 | 1,00E+00 | 2,15E-01 | 3,46E+01 |
| *Perikaryon* | 4 | 3,28 | 0,0329166 | 5,6835 | 1,00E+00 | 2,11E-01 | 3,50E+01 |
| *Spindle microtubule* | 3 | 2,46 | 0,0339831 | 10,2690 | 1,00E+00 | 2,12E-01 | 3,59E+01 |
| *Focal adhesion* | 7 | 5,74 | 0,0447431 | 2,6964 | 1,00E+00 | 2,64E-01 | 4,45E+01 |
| *Death-inducing signaling complex* | 2 | 1,64 | 0,0451998 | 43,0319 | 1,00E+00 | 2,60E-01 | 4,48E+01 |
| *Plasma membrane* | 36 | 29,51 | 0,0564361 | 1,3157 | 1,00E+00 | 3,07E-01 | 5,26E+01 |
| *Pore complex* | 2 | 1,64 | 0,0577377 | 33,4692 | 1,00E+00 | 3,07E-01 | 5,34E+01 |
| *Axon* | 5 | 4,10 | 0,0591889 | 3,3922 | 1,00E+00 | 3,07E-01 | 5,44E+01 |
| *Lysosome* | 5 | 4,10 | 0,0623724 | 3,3321 | 1,00E+00 | 3,14E-01 | 5,63E+01 |
| *Cytoplasmic, membrane-bounded ve* | 4 | 3,28 | 0,0630018 | 4,3656 | 1,00E+00 | 3,10E-01 | 5,67E+01 |
| *Nuclear Membrane* | 5 | 4,10 | 0,0648184 | 3,2885 | 1,00E+00 | 3,12E-01 | 5,77E+01 |
| *Caveola* | 3 | 2,46 | 0,068397 | 6,9513 | 1,00E+00 | 3,20E-01 | 5,98E+01 |
| *Cytoplasmic vesicle* | 5 | 4,10 | 0,0698597 | 3,2045 | 1,00E+00 | 3,20E-01 | 6,06E+01 |
| *Endoplasmic reticulum chaperone co* | 2 | 1,64 | 0,0701123 | 27,3839 | 1,00E+00 | 3,15E-01 | 6,07E+01 |
| *Neuron projection* | 5 | 4,10 | 0,0715839 | 3,1775 | 1,00E+00 | 3,15E-01 | 6,15E+01 |
| *Cyclin-dependent protein kinase hol* | 2 | 1,64 | 0,0943802 | 20,0815 | 1,00E+00 | 3,90E-01 | 7,20E+01 |

El procedimiento estadístico para aceptar o rechazar la $H_0$ se denomina test global, y generalemente se utiliza el *p-valor*, en cambio, si se quiere especificar cuál de todas las $H_{0,i}$ debe ser rechazada, se utilizarán test de hipótesis múltiples.

El método de *Fisher* es un test global y se aplica generalmente cuando se quiere contrastar la misma hipótesis en estudios independientes. Es a partir de la membrana plasmática donde el *p-valor* aumenta, y es mayor a 0.05, en cambio, para las primeras ubicaciones, el *p-valor* es muy bajo. Por ejemplo, el nucleoplasma tiene un *p-valor* =

0.0000000715 < 0.05, esto quiere decir que se rechazaría la hipótesis nula. La gran mayoría de proteínas se encuentran en el citoplasma.

*Fold Enrichment* representa la cantidad a la que los genes/ proteínas, están representados en un gran conjunto de genes o proteínas. Por lo tanto a mayor número, mayor representación.

Los métodos de *Bonferroni* y *Benjamini-Hochberg*, son procedimientos que se utilizan en test múltiples, donde se propone rechazar $H_0$ si se rechaza alguna $H_{0,i}$.

La corrección de *Bonferroni* es un método que se utiliza para contrarrestar el problema de las comparaciones múltiples. Si se prueban múltiples hipótesis, aumenta la posibilidad de un evento raro y, por ello, aumenta la probabilidad de rechazar incorrectamente la $H_0$. El test corrije este aumento al comprobar cada hipótesis individual a un nivel de significancia de $\alpha/m$, donde $\alpha$ es el nivel alfa general deseado y $m$ es el número de hipótesis. Por tanto $m$ será el número total de hipótesis nulas y $m_0$ el número de hipótesis nulas verdaderas; la tasa de error familiar *(FWER)* es la probabilidad de cometer, al menos, un error de tipo I.

El valor q de *Benjamini-Hochberg*, disminuye la tasa de descubrimiento falso, ayuda a evitar errores de tipo I, falsos positivos. El objetivo de este método es controlar el FDR mediante la corrección secuencial de *Bonferroni* modificada para las pruebas de hipótesis múltiple. La probabilidad estadística de rechazar incorrectamente un verdadero $H_0$ se inflará significativamente junto con el número de hipótesis probadas simultáneamente.

*FDR*, muestra la probabilidad de rechazar la hipótesis nula siendo esta verdadera, por lo tanto, cuanto más pequeña sea esta probabilidad, menor probabilidad de cometer un error tipo I, por ejemplo, el citosol tiene un FDR de 3.85E-15 (una probabilidad muy baja de confundir la ubicación celular de las 65 proteínas se encuentran en el citosol), en cambio, para la vesícula endoplasmática tiene un FDR mucho más alto, de 6.06E+01, lo que quiere decir que tiene una probabilidad más alta de confundir su ubicación celular y cometer el error de seleccionar una ubicación donde no se encuentren esas 5 proteínas.

# *5. Conclusiones y Perspectivas:*

Teniendo en cuenta los objetivos que se establecieron anteriormente, a lo largo del presente trabajo se ha desarrollado una estrategia computacional para el análisis sistemático de microarrays de proteínas, donde a partir del software *GenePix®Pro 6.0*. se obtiene un primer control de calidad de los datos junto con una base de datos. A partir de dicho conjunto de datos, se diseñaron diferentes ecuaciones con el fin de obtener un listado definitivo de proteínas con su señal de intensidad.

A partir de dichas ecuaciones, que corresponden a una parte diferente del preprocesado de datos, se creó un árbol de decisión que esquematizaba todos los pasos y ecuaciones *(Figura N).*

Con el fin de hallar un diseño óptimo (en relación al preprocesado de datos), se emplearon procesos diferentes: se aplicó un punto de corte a la señal de intensidad de las proteínas, a partir de la *MAD (Median Absolut Desviation)* pero como no se respetaba la replicabilidad de los *spots* y debido a la gran variabilidad de estos, dicho punto de corte no era representativo de los datos y muy pocas proteínas superaban este *punto de corte;* también se utilizó el coeficiente de variación (*CV*) para establecer otro punto de corte diferente, con la intención de que fuera menor y más proteínas pudieran superarlo, pero debido a la gran variabilidad que presentaba la señal de intensidad, ocurría lo mismo que en el proceso anterior; por último, no se tuvieron en cuenta la replicabilidad de los *spots,* es decir, no se tuvo en cuenta que al menos el 50% de los *spots* que pertenecían a una determinada proteína, pasasen el filtrado, se admitieron todos los *spots* con intensidad positiva y se comprobó si en todos los subarrays los spots con replicabilidad menor al 50% se comportaban igual y seguían cierta progresión. Finalmente se observó que en cada uno de los subarrays los datos se comportaban de forma totalmente independiente, por tanto, tampoco se pudo aplicar dicho proceso.

Obtenidos los listados de proteínas definitivos y utilizando una estadística sencilla, se obtiene una primera impresión de los datos: los mejores resultados corresponden a los microarrays 11 y 22, donde los datos se comportan de forma más similar en todas las muestras y la señal de intensidad no muestra mucha variabilidad entre unos subarrays y otros. El PCA para estos dos microarrays, muestra los mejores resultados.

Además, como contenían los listados con mayor número de proteínas, se tomó el subarray 3 del microarray 22, (ya que, al tener mayor número de datos, era el más representativo de todos) para aplicar las diferentes técnicas y análisis. A través del dendrograma y del método de Ward, se obtuvo el número óptimo de clúster en los que se dividían las proteínas, con el "*Heatmap*" se comprobó qué muestras eran más similares y como era la señal de intensidad de las proteínas que contenían. El PCA, utilizando de nuevo el método de Ward y la distancia euclídea, separa las proteínas en 3 clúster diferentes, cada uno de ellos caracterizado por una vía de señalización principal diferente.

Utilizando la base de datos del espectrómetro de masas se obtuvo el número de péptidos de ciertas proteínas y a través de un *CircoPlot* se pudieron englobar todos los datos (tanto cuantitativos como cualitativos) en un solo gráfico, donde a partir del cual se estudió la relación entre el número de péptidos y *spots,* con test estadísticos no paramétricos: *Kruskal Wallis* y el coeficiente de correlación de *Spearman.*

A partir de los resultados obtenidos con el PCA, se comprobó que entre las proteínas existía cierta relación. Utilizando diferentes métodos bioinformáticos y relacionando los términos con bases de datos de *KEGG, GO, Reactome, OMIM_DISEASE* (a partir de los

nombres canónicos y nombres propios hallados y relacionados en UniProt y NeXtProt), se resolvieron ciertas cuestiones.

Teniendo en cuenta el gráfico de señalización del receptor de células B (ya que se parte de muestras de linfocitos B incubadas en cada uno de los subarrays), se pueden contrastar algunas de las conclusiones descritas anteriormente *(Figura F)*:

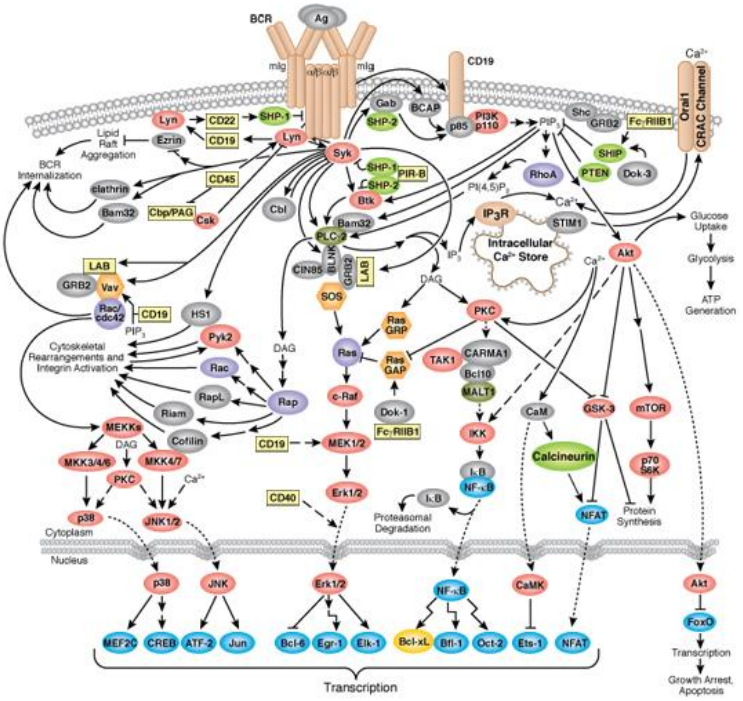



*Figura F. Gráfico de señalización del receptor de células B . Figura obtenida en Cell Signaling Technology.*

En este gráfico se puede observar como las proteínas se encuentran relacionadas unas con otras; este resultado se contrastó con la herramienta STRING *(Gráfico 13)* donde se obtuvo una red de interacción entre proteínas, en la cual todas éstas, interaccionaban entre ellas. Esta herramienta (al igual que DAVID), también ofrece una descripción de la proteína y su función donde se puede comprobar (por ejemplo) que proteínas como *Jun, NFat, Bcl, ATF, …* son factores de transcripción, y que proteínas son Quinasas, como *Lyn, Syk, MEK ½, p38 (Anexos, Tabla 78, Función molecular/ Tabla 79, Palabras clave de UniProt).*

Además, en el gráfico aparece la vía de la apoptosis, vía que aparece en *Reactome* y que incluye a las mismas proteínas que también se incluyen en este gráfico: *AKT, MEK ½, MDM2, Btk, …*en la jerarquía de eventos que se hallaron con esta herramienta, aparecen muchas otras vías donde se encontrarían otro tipo de proteínas que también aparecen en el gráfico. Debido a que uno de los eventos, que aparecía en Reactome, con menor FDR, era *Disease* (enfermedad/ afección), en Enrichr se revelaron ciertas proteínas que indicaban una determinada patología. Una de las patologías era el Linfoma, cáncer que afecta al sistema inmunitario (por esta razón aparecía la vía de señalización del Sistema Inmune *(Gráfico 14)* en Reactome). Existen diferentes tipos de linfomas, uno de ellos

son los Linfomas no Hodgkin que afectan a las células B, debido a esto también aparece la Leucemia como una de las principales patologías (a parte de otros tipos de cáncer).

Por lo tanto, en este subarray, se puede concluir que existen ciertas proteínas que indican que el paciente no solo tiene Leucemia Linfocítica Crónica, si no también es probable que tenga otro tipo de afecciones como cáncer de mama y/o cáncer colorrectal.

Por último, en DAVID, se mostró la ubicación celular de todas las proteínas que componían la muestra 3; en la *Figura F* se observa como *p38, Jun, ATF, NFat, CREB,* entre otros, se encuentran en el Núcleo, siendo estas factores de transcripción que se obtuvieron en el PCA (clúster azul) con 66 proteínas encontradas en esa ubicación; en la membrana plasmática, es donde se encuentra la cascada de BCR (clúster verde) que aparecía en el PCA; por último en el gráfico aparece el citoplasma, y en DAVID esta ubicación contiene 63 proteínas diferentes con un p-valor muy bajo, en el PCA se puede encontrar en el clúster rojo, con otras rutas de activación, que no son BCR. *(Anexos, Ubicación celular de las proteínas, tabla 82)* . Estas ubicaciones son las que menor p-valor y FDR tienen, y asimismo aparecen en el gráfico.

Relacionados los datos con cada una de las bases de datos anteriores: KEGG, GO, OMIM, … a través de términos propios de UniProt y NeXtProt, se alcanzan las siguientes conclusiones finales:

i. En STRING se comprobó cómo gran parte de las proteínas estaban relacionadas, especialmente aquellas que se encontraban en el centro de la red.
ii. Gracias a los términos clave de UniProt y a su función (información obtenida en GO_Ontology), se tuvo una primera descripción de los datos. Debido a que algunos términos indicaban apoptosis y determinadas patologías, se utilizó Reactome para analizar las vías de señalización de todas estas (Reactome Pathways), donde se examinó una determinada vía, la cual indicaba que los datos podían pertenecer a un paciente con cáncer.
iii. Gracias a Enrichr, se relacionaron todos estos con la base de datos OMIM_Disease, donde se confirmó que los datos no solo reafirmaban el cáncer, si no a qué tipo de cáncer correspondían.
iv. A través de DAVID y *GOTERM_CC_Direct*, se analizó su ubicación celular, donde se constató que la gran mayoría de proteínas se encontraban en el núcleo y citoplasma.

Actualmente el flujo de trabajo, diseñado y desarrollado en el presente trabajo, es el que se utiliza para analizar los microarrays de proteínas, elaborados en el Centro de Investigación del Cáncer en Salamanca.

# 6. *Bibliografía*:


- Aboytes, K., Humphreys, J., Reis, S. & Ward, B. (2003). *Slide Coating and DNA Inmobilization Chemistries in A Beginner's Guide to Microarrays*. Springer US. 1-41.
- Brown, T. (2008). *Genomas, transcriptomas y proteomas. Genomas/ Genome*. Ed. Médica Panamericana.
- Casado-Vela, J., Gonzalez-Gonzalez, M., Matarraz, S., Martinez-Esteso, M. J., Vilella, M., Sayagues, J. M., … Carlos Lacal, J. (2013). *Protein Arrays: Recent Achievements and their Application to Study the Human Proteome* [Text].
- Cazzulo, J. J. (2014). De la Genómica a la Proteómica. *Manual de Proteómica, Sociedad Española de Proteómica, INTECH, UNSAM-CONICET, Argentina*, 13-20.
- Corrales, F., Calvete, J. J., & Sociedad Española de Proteómica. (2014). *Manual de proteómica*. Sociedad Española de Proteómica.
- Ellington, A. A., Kullo, I. J., Bailey, K. R., & Klee, G. G. (2010). Antibody-Based Protein Multiplex Platforms: Technical and Operational Challenges. *Clinical Chemistry*, *56*(2), 186-193.
- García-Valiente, R., Fernández-García, J., Carabias-Sánchez, J., Landeira-Viñuela, A., Góngora, R., Gonzalez-Gonzalez, M., & Fuentes, M. (2019). A Systematic Analysis Workflow for High-Density Customized Protein Microarrays in Biomarker Screening. En X. Wang & M. Kuruc (Eds.), *Functional Proteomics: Methods and Protocols* (pp. 107-122).
- Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Britan, A., Cusin, I., Domagalski, M., … Bairoch, A. (2017). The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research*, *45*(D1), D177-D182.
- González, A. D. (s. f.). *MICRO Y NANOTECNOLOGÍA EN MEDICINA: LOS CHIPS O MICROARRAYS DE ADN*. *7*.

- González, R. P. (s. f.). *Desarrollo de microarrays de anticuerpos en formato esfera para la validación de biomarcadores de artrosis*. 38.
- González-González, M., Jara-Acevedo, R., Matarraz, S., Jara-Acevedo, M., & Paradinas, S. (2011). *Nanotecnología en proteómica: arrays de proteínas y nuevos sistemas de detección*. 6.
- Guerrero Picó, I. (2015). *Desarrollo de microarrays de altas prestaciones basados en atrapamiento con un hidrogel. Aplicación a inmunoensayo y ensayos genéticos.*
- Handran, S. (s. f.). *Biological Relevance of GenePix Results*. 9.
- Juanes-Velasco, P., Carabias-Sanchez, J., Garcia-Valiente, R., Fernandez-García, J., Gongora, R., Gonzalez-Gonzalez, M., & Fuentes, M. (2018). Microarrays as Platform for Multiplex Assays in Biomarker and Drug Discovery. *Rapid Test - Advances in Design, Format and Diagnostic Applications*.
- Khimani, A. H., Mhashilkar, A. M., Mikulskis, A., O'Malley, M., Liao, J., Golenko, E. E., … Lott, S. T. (2005). Housekeeping genes in cancer: normalization of array data. *BioTechniques*, *38*(5), 739-745.
- López, M., Mallorquín, P., & Vega García, M. (2006). *Aplicaciones de los microarrays y biochips en salud humana: informe de vigilancia tecnológica*. Madrid: Genoma España.
- Lovric, J. (2011). *Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis*.
- Maestre, J. G. (s. f.). *ANÁLISIS DE DATOS DE MICROARRAYS*. 73.
- Merbl, Y., & Kirschner, M. W. (2011). Protein microarrays for genome-wide posttranslational modification analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *3*(3), 347-356. https://doi.org/10.1002/wsbm.120

- Nanotecnología en Proteómica: microarrays de proteínas y nuevos sistemas de detección.(s. f.).de https://www.institutoroche.es/biotecnologia/69/nanotecnologia_en_proteomica_microarrays_de_proteinas_y_nuevos_sistemas_de_deteccion
- Ong, S.-E., & Mann, M. (2005). Mass spectrometry–based proteomics turns quantitative. *Nature Chemical Biology*, *1*, 252.
- Opiniones de Proteoma. (s. f.). https://www.datuopinion.com/proteoma
- Pichler, K., Warner, K., & Magrane, M. (2018). SPIN: Submitting Sequences Determined at Protein Level to UniProt. *Current Protocols in Bioinformatics*,
- Reymond Sutandy, F., Qian, J., Chen, C.-S., & Zhu, H. (2013). Overview of Protein Microarrays. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, *0 27*, Unit-27.1.
- Rivas-Lopez, M. J., Sánchez-Santos, J. M., & De las Rivas, J. (2005). Estructura y análisis de microarrays. *Boletín de Estadística e Investigación Operativa. BEIO*, *21*(2), 10-15.
- Schwartz, A. S., Yu, J., Gardenour, K. R., Finley, R. L., & Ideker, T. (2009). Cost effective strategies for completing the Interactome. *Nature methods*, *6*(1), 55-61.
- Tomizaki, K., Usui, K., & Mihara, H. (2010). Protein–protein interactions and selection: array-based techniques for screening disease-associated biomarkers in predictive/early diagnosis. *The FEBS Journal*, *277*(9), 1996-2005.
- Tyers, M., & Mann, M. (2003). From genomics to proteomics. *Nature*, *422*(6928), 193. https://doi.org/10.1038/nature01510
- Valledor, L., & Meijón, M. (2014). Métodos de análisis Estadísticos en Proteómica. *Manual de Proteómica, Sociedad Española de Proteómica, INTECH, UNSAM-CONICET, Argentina*, 73-110.

# *Summary*

Proteomics is defined as the set of different techniques related by their willingness to provide information on the proteome, including from the identities of constituent proteins, specifying the function of a particular protein, to its intracellular location. The technique used to study the composition of the proteome is called protein profile and there are two methods to obtain this protein profile, which are based on separating proteins from their proteome: protein electrophoresis and mass spectrometry.

There is now a new alternative for dissecting the proteome: microarray technology, specifically protein microarrays. Microarray technology represents a new tool for high-performance biological studies and is based on techniques known as high-throughput. When analyzing massive amounts of data, where the sources of variability must be adjusted in order to identify the main proteins involved among the many others that make up the initial database, Statistics is a crucial tool within microarray technology.

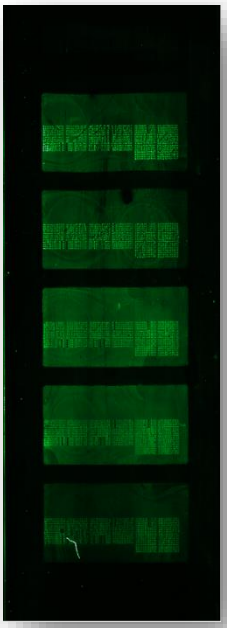
The statistical analysis methods used range from univariate statistics with both parametric and non-parametric methods to multivariate statistics. When analyzing data from microarrays, a problem faced by statistics is that the number of genes/proteins is much greater than the number of subjects, therefore, in statistical terms we are faced with the case of thousands of variables versus very few samples.

The type of microarray used to carry out the present work were: flat microarrays of proteins printed with antibodies, where conventional fluorescent labels were used as detection method. In label-dependent protein detection microarrays, especially those involving antibodies, the main detection method is biotin marked with a fluorophore, which is combined with the study proteins, forming a complex that is detected by the antibodies printed on the microarray. In this case, the microarrays are composed of proteins and antibodies associated with Chronic Lymphocytic Leukemia *(CLL).*

10 different microarrays were designed and analyzed, which are classified into different groups depending on the design of the samples, each of these represents a patient with LLC (these samples are always different), as the design process advances, the number of replicas will decrease and the number of samples will increase.

It is necessary to add that the position within the microarray chosen for each of the samples affects them. The same result will not be obtained by placing the sample in another subarray with a higher or lower position.

All the images in TIFF format were analyzed through the GenePix®Pro 6.0 software; when loading the image into this software, a series of parameters must be adjusted according to the intensity of the microarrays. In this case, so that each of the microarrays of the different sets were under the same conditions, the image of all was loaded at the same time, adjusting the same conditions of intensity and a mesh. This mesh allows to know that it contains a certain spot and it must be adjusted to all the intensities of each one of them.



Microarrays are manufactured in such a way that they can provide information about the performance of hybridization and image scanning, therefore, each of these arrays contains a series of controls that represent the markers of a given sample. The intensity value obtained from all these controls are used to generate quality control statistics: a positive control would give a positive signal and a negative control would not give a signal.

The *GenePix®Pro 6.0* software extracts numerical indicators, a product of the relative union between the test substances and the control substances of each spot. This involves quantitative processing to calculate fluorescence intensity ratios at different wavelengths, which generates a large amount of data that requires additional interpretation. 10 different databases were generated with 41 variables and more than 4,500 data per microarray, the number of data depends on the number of proteins printed with antibodies that have been included in the microarray.

All microarrays are designed in a certain way so that each protein is represented in a row inside all subarrays, this row contains 6 spots where the protein is found. All the proteins are suspended in a determined Mastermix; to be able to study the signal of own intensity of the Mastermix, inside the microarray there are spots that contain only a determined *MM*.

To find out which proteins interact with the antibody, and give a "true" positive signal, these data must go through a preprocessing process:

*I. Background Correction:*

Since there is a certain amount of non-specific hybridization contained in the background that affects the sensitivity and specificity of the microarray, background correction minimizes the effects that such hybridization causes on each signal of the spots. In this type of microarrays, the correction of the background noise is divided into two parts, in order to obtain the signal of interaction "true or clean" between the protein and the antibody, first must eliminate the background signal provided by the software when analyzing the image, and once that value is obtained, subtract the signal intensity that provides the MM on that spot.

*II. Filtering:*

Once the signal that belonged to the *MM* and the one that provided the software has been eliminated, we work with the signal of intensity of clean interaction that exists between the protein and the antibody. If this signal is negative, it means that there is no interaction between the protein and the antibody, but if it is positive, it means that there is interaction. All the databases will be filtered to work only with those proteins that have a positive signal, the rest with a negative signal will be removed from the database.

Those proteins that have at least 50% of their *spots* with a positive signal will be accepted, in order to ensure that all the proteins that have passed all these filters really have a signal of interaction with the antibody; if there is a total of 6 *spots* per protein, 50% will be 3 *spots*, so only those that have a minimum number of 3 replicates will be accepted.

Once these two phases of data preprocessing have been completed, a definitive list of proteins is obtained, containing each subarray.

*III. House-keeping:*

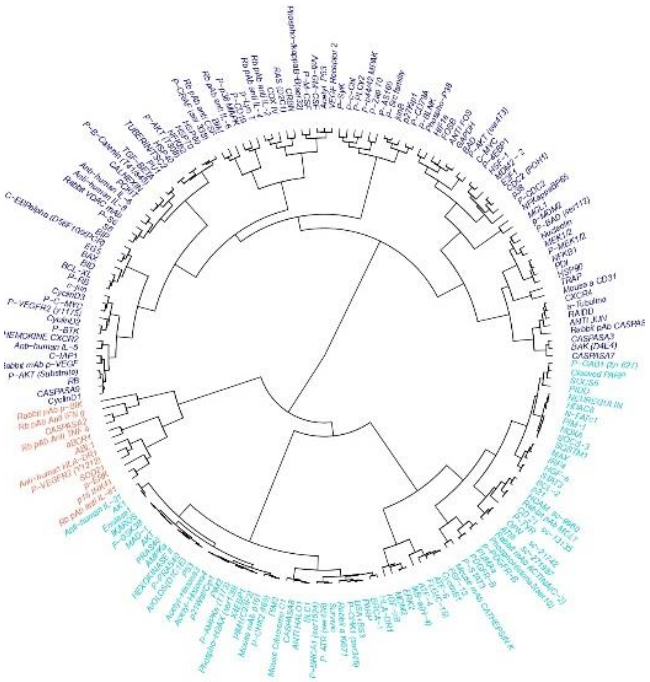
The biological maintenance of cells affects the gene expression of certain genes and therefore proteins, although there are certain proteins that do not undergo changes throughout this biological maintenance. These proteins are called "*house-keeping proteins*" and by definition do not reflect any change in their levels of expression during cell development, treatment or anomalies present during a given disease. All these maintenance proteins can be used to estimate the expression levels relative to other proteins, therefore, all those that passed the previous filters and that were present in the 10 microarrays were used as maintenance proteins: *BIOTIN.*

Once the data preprocessing is finished, it is obtained which proteins interact with an antibody and its signal intensity. By having definitive lists of each one of the samples, through different statistical techniques, it can be determined which samples are more similar, how proteins are grouped and different ways of visualizing all the data.

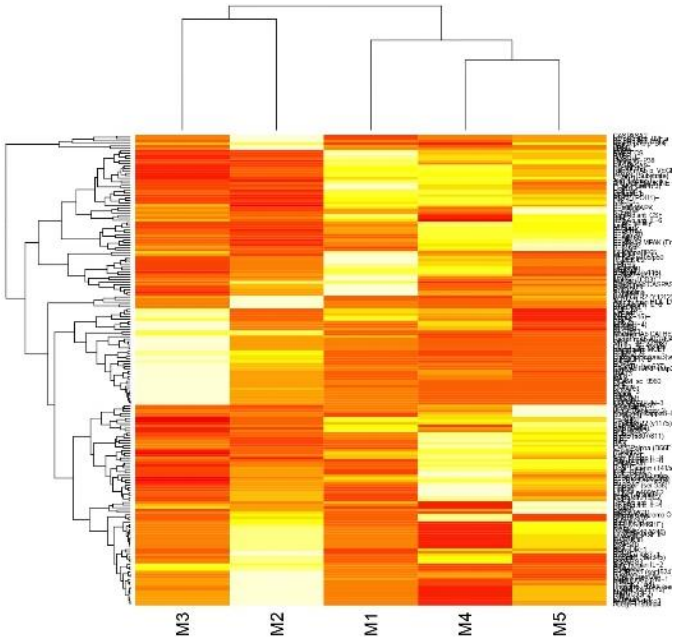
In order to perform a quantitative analysis of the data, the following was carried out *Cluster Dendrogram:*

Clusters were formed using the WARD method, this hierarchical analysis criterion is known as "Ward Minimum Variance Criterion", which minimizes the total within the cluster variance.

Difference in 3 groups the original 193 proteins, the level of similarity is measured in the center of the graph and the proteins are specified around the group to which they belong.



In order to analyze which samples had a more similar protein profile, a *Heatmap* was carried out*:*



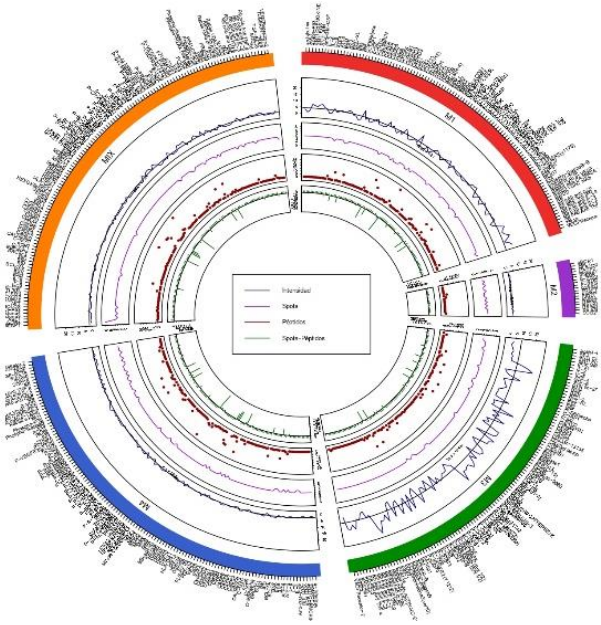

Depending on the intensity of the color, you can analyze the presence or absence of the protein in X sample, if the intensity is white means that there is no presence of that protein in a given sample, however, the stronger the color, the greater intensity that protein will have in the sample.

It can be seen that samples are more similar to others thanks to the Dendogram that is at the top of the graph.

To perform a qualitative analysis of the ScatterPlot data, the number of spots giving a positive signal for each of the proteins was calculated. By comparing the number of spots of the same protein in the same sample, but in different microarrays, it is possible to analyse whether these behave the same in different slides. Theoretically if the samples are the same and both are located in the same position in the two microarrays and have been scanned with the same adjustments, both samples should have a similar number of proteins and spots; to verify it, this type of graphs are used.

In order to have greater reliability of the resulting proteins in the microarrays, all the resulting proteins from each of the databases were crossed with a database obtained through protein mass spectrometry. This database was also obtained from a patient with Chronic Lymphocytic Leukemia, therefore, the proteins that were present in both technologies were much more likely to be typical of this pathology: we look for the intersection between both databases. This will allow working with one more variable: the number of peptides (*Peptid Count*): for all those proteins common to both databases, their number of peptides will be added to the original database, the rest of the proteins will be given a number of 0 peptides.

Currently, in Proteomics, a visualization tool to facilitate the identification and analysis of the similarities and differences that arise when comparing the Proteome is through Circus graphs. This allows to integrate in a single graph all the information available about all the subarrays contained in each microarray: all the proteins that make up each sample, together with the signal intensity of each of them, the number of *spots*, the number of peptides provided by the *MS/MS* and the difference between *spots* and peptides.



As a result of this result, the relationship between the number of peptides and the number of spots was analysed. The main objective is to check whether proteins printed with antibodies show a greater number of peptides depending on the number of spots

they have. Two non-parametric tests were applied, the *Kruskal Wallis* test and the *Spearman* test with the purpose of contrasting if there was dependence or not between both variables.

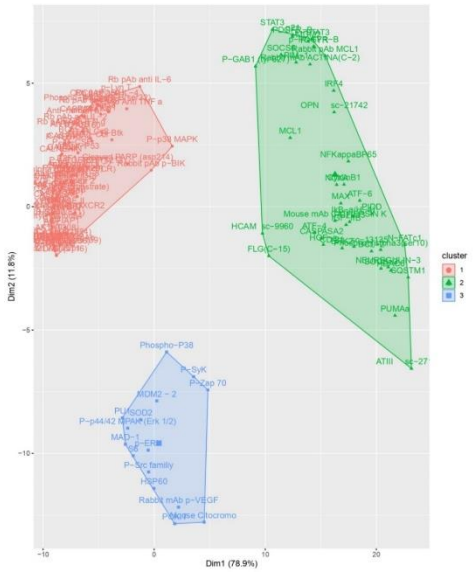
In the Kruskal Wallis test a p-value = 0.383 > 0.05 was obtained, with which it was confirmed that the two variables are independent, the proteins do not show a greater number of peptides depending on the number of spots they have. The *Spearman* test was also applied in order to ensure that dependence does not really exist. The Spearman correlation coefficient results in $\rho = -0.128$ and is not significant since its p-value=0.442 > 0.05, because $\rho$ is close to 0, meaning that there is no monotonous relationship between these two variables.

Once these two statistical tests have been applied, it can be stated that there is not enough evidence to admit the dependence between these two variables; therefore, the independence between them will be considered.

Due to the large amount of proteins found in each of the microarrays, in order to decrease their dimensionality and to be able to group the proteins in groups that explain approximately the same as the original proteins, a PCA (Principal Component Analysis) was applied. it was discovered that the graph encompassed the proteins according to their signaling in the B cell receptor.

Three main routes stand out:

iv. BCR Cascade (B Cell Receiver) - Green Cluster.
v. Transcription factors related to this cascade - blue cluster.
vi. Other B-lymphocyte activation pathways (other than BCR) - red cluster.




It is from this graph that we begin to analyze the relationship between proteins and the reason why they have given a positive signal in the microarray. In order to study the relationship between these proteins, various bioinformatic techniques and tools will be used to create the interaction networks, find the cell location, its signalling pathways..., through KEGG, GO, OMIM,…

In STRING we see how much of the proteins are related, especially those that are in the center of the network. This tool relates the data with other databases, taking into account the keywords found in UniProt, in which terms such as apoptosis, disease, proto-

oncogene, ... The signaling pathways through Reactome will be examined to get more information about these data.

In *Reactome_Pathway* different signalling pathways were found, among them "Disease" with an FDR 1.10E-05; it is observed how different networks start from this pathway, which represent diseases of the Immune System with an FDR:1.79E-3, Neurodegenerative with an FDR: 9.61E-4, Developmental disorders with an FDR: 1.4E-1, Infectious Diseases with an FDR: 6.25E-1, Cell transduction signal diseases with an FDR: 1.25E-10. Bearing in mind the above options, the route with the lowest FDR corresponds to the signalling route for diseases related to the cell transduction signal, which includes many more. Among them, P13K/AKT stands out for signalling in cancer with a p-value of 4.5E-13 (< 0.005) and an FDR of 8.43E-11.

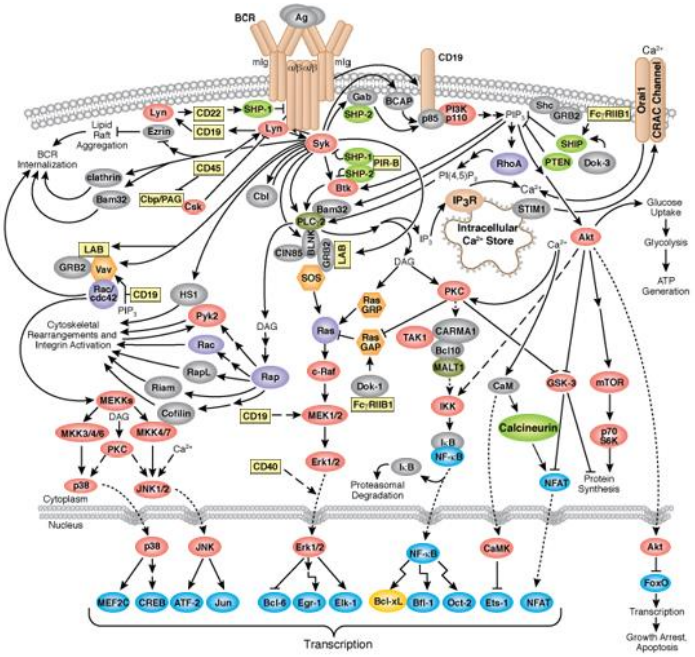
Through Enrichr it can be verified that genes are involved in a certain pathology, with the option OMIM_Disease, it was confirmed that the data not only reaffirmed the cancer, but also what type of cancer they corresponded to:

| *Term* | *P-value* | *Z-score* | *Combined Score* | *Genes* |
|---|---|---|---|---|
| *Lymphoma* | 0,00033 | -1,7905 | 14,36 | *MYC;IL21R;BCL2* |
| *Breast_cancer* | 0,00068 | -1,7776 | 12,96 | *CHEK2;AKT1;TP53* |
| *Leukemia* | 0,00141 | -1,4293 | 9,38 | *PDGFRB;KIT;BCL2;ABL1* |
| *Colorectal_Cancer* | 0,00194 | -1,1702 | 7,31 | *CHEK2;AKT1;TP53* |
| *Immunodeficiency* | 0,01297 | -0,2176 | 0,95 | *NFKBIA;IL2* |

Lymphoma and Breast Cancer being the pathologies with the highest combined score.

Finally, a DAVID analysis will be carried out to study the cellular location of all these proteins, with the cytosol and nucleus being the locations with the highest p-value and lowest FDR.

If all this information is contrasted with the B-cell receptor signalling graph (since samples of B lymphocytes incubated in each of the subarrays are taken as a starting point), found in Cell Signaling Technology:



Where you can see how proteins are related to each other, this result was contrasted with the STRING tool where it was obtained a network of interaction between proteins, in which all of them interacted with each other.

In the graph appears the apoptosis pathway, which appears in Reactome and includes the same proteins that are also included in this graph: AKT, MEK ½, MDM2, Btk, ...in the hierarchy of events that were found with this tool, there appear many other pathways where you would find other types of proteins that also appear in the graph.

In Enrichr, all the proteins that belonged to these pathways were related to the OMIM_Disease database, where it was confirmed that the data indicate that the patient not only has Chronic Lymphocytic Leukaemia, but it is also probable that he has other types of affections such as breast cancer and/or colorectal cancer.

Through DAVID, was shown the cellular location of all proteins in the graph above is observed as p38, Jun, ATF, NFat, CREB, ... are found in the Nucleus, these are transcription factors that were obtained in the PCA (blue cluster); In the plasma membrane, is where the cascade of BCR (green cluster) that appeared in the PCA is found; finally in the graph appears the cytoplasm, and in DAVID this location presents a very low p-value (in the PCA could be found in the red cluster, with other routes of activation, which are not BCR).

Currently, the workflow, designed in the present work, is the one carried out to analyse all the protein microarrays of the Cancer Research Centre of Salamanca, in which both the replicability of the samples and the replicability of the spots are respected.