Grado en Estadística, Facultad de Ciencias

TRABAJO FIN DE GRADO

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

ANÁLISIS INTEGRADO DE DATOS GENÓMICOS PROCEDENTES DE PACIENTES CON MIELOMA MÚLTIPLE PARA LA IDENTIFICACIÓN DE GRUPOS DE ALTO RIESGO ASOCIADOS A LA PÉRDIDA DE FUNCIÓN DEL GEN SUPRESOR TUMORAL TP53.

AUTORA: LUCIA MORALEJA ALONSO

TUTORES:
LUIS ANTONIO CORCHETE SÁNCHEZ
NORMA CARMEN GUTIÉRREZ GUTIÉRREZ
JOSE MANUEL SÁNCHEZ SANTOS
ELENA SÁNCHEZ LUIS

ÍNDICE

Grado en Estadística, Facultad de Ciencias

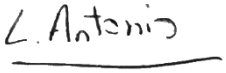
TRABAJO FIN DE GRADO

VNiVERSiDAD D SALAMANCA

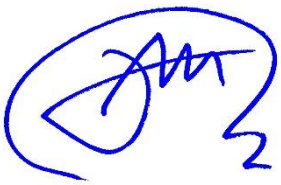
CAMPUS DE EXCELENCIA INTERNACIONAL

ANÁLISIS INTEGRADO DE DATOS GENÓMICOS PROCEDENTES DE PACIENTES CON MIELOMA MÚLTIPLE PARA LA IDENTIFICACIÓN DE GRUPOS DE ALTO RIESGO ASOCIADOS A LA PÉRDIDA DE FUNCIÓN DEL GEN SUPRESOR TUMORAL TP53.

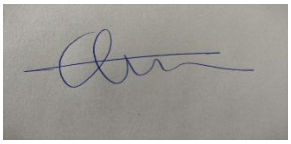
AUTORA: LUCIA MORALEJA ALONSO



LUIS ANTONIO CORCHETE SÁNCHEZ



JOSE MANUEL SÁNCHEZ SANTO



NORMA CARMEN GUTIÉRREZ GUTIÉRREZ

ELENA SÁNCHEZ LUIS

AGRADECIMIENTOS.

A mis padres, por inculcarnos siempre la importancia de una buena formación. Vosotros y Juan me hicisteis creer que soy capaz de todo. Este TFG es vuestro, como todos mis logros.

A mi abuela por acordarse de mi cada noche antes de un examen y ayudarme a su manera.

A toda mi familia, por demostrarme siempre que la solución está en los libros. Gracias por apoyarme incondicionalmente en esta y todas las etapas de mi vida.

A mis amigos de Estadística, por hacer de estos cuatro años un baile. Que pasen los años y que 1+1 siga sumando 7.

A Luis, por ser el mejor apoyo que podría tener en este proceso. Gracias por regalarme parte de tu tiempo pese a que todos sabemos que no te sobra.

ABSTRACT

Multiple myeloma (MM) is the second most common hematologic malignancy behind non-Hodking lymphoma and is characterized by the uncontrolled accumulation of clonal plasma cells within the bone marrow. The most recurrent symptoms are bone pain in bones such as the spine, sternum or ribs, as they are the richest source of bone marrow cell. Other symptoms derived from this patology, are bone fractures and kidney failure.

The profile of MM patients is characterized by an average age of 65 years and it is slightly more common in men than in women.

MGUS is a premalignant and asymptomatic stage that occurs in patients without clinical disorders that does not usually progress to multiple myeloma.

The prognosis of this disease is highly variable since it is a very heterogeneous disease that depends on many factors. One of these factors is the stage of the disease determined by the international staging system (ISS).

Treatments for MM have made great strides in recent years. Until the end of the 20th century, the available treatments were corticosteroids and other traditional chemotherapeutic agents. In the last decades proteasome inhibitors, immunomodulators ... etc have been introduced. These advances in therapies have led to an increase in patient survival, since there are multiple treatment options depending on the characteristics of each one.

Despite all these improvements, and the increased survival of patients, MM remains an incurable disease. Most of them present a pattern of recurrence with stages of remission followed by relapses until the patient stops responding to treatment.

MM is a disease with a lot of variability in the clinical course. This is caused by its great genomic and molecular complexity. The genomic alterations that have been found in patients have an impact on the prognosis of the disease. These alterations can be classified into translocations, copy number alterations, and point mutations. Translocations are a change in location of a segment of a gene to another place in the genome. Most translocations affect chromosome 14. These are seen in up to 60% of MM. The most frequent are t (11; 14), t (4; 14) and t (14; 16). Translocations are also detected in the MYC oncogene but to a lesser extent than the previous ones (up to 20% of cases).

Copy number abnormalities occur when cells gain or lose entire chromosomes or arms and regions. The chromosome that is most affected is chromosome 1, which is won in the q arm in 60% of the cases and the p arm is lost in 30% of them. The 17p deletion includes the suppressor gene locus TP53 and is the one associated with a worse prognosis.

Point mutations consist of the change of a single nucleotide of DNA. Mutations are very heterogeneous, but the most frequent are those that affect RAS oncogenes.

The aim of this work is to study the impact of 17p deletions and mutations of the TP53 gene on the gene profile of the most frequent primary translocations.

The data used is extracted thanks to novel techniques, including next-generation sequencing (NGS). This technology arises from a cluster of advances in the fields of physics, chemistry and biology and allows the complete and detailed study of the human genome. This fact makes it a technique closely linked to oncology, since cancer is a disease closely linked to the study of the genome.

RNA sequencing (RNA-seq) is of great importance in the field of biomedicine, due to its relevance in transcriptomics since the objective of transcriptomics is the study of gene expression. Transcription is the process by which a gene is expressed by generating an mRNA molecule from a DNA fragment. The development of transcriptomics has made it possible to determine how changes in gene expression affect the appearance and evolution of different diseases, as well as the evaluation of the efficacy of treatments.

DNA sequencing is the process that determines the order of DNA bases (guanine, adenine, cytosine and thymine) which is really important in diagnosing diseases by comparing healthy sequences with other mutated ones. Whole genome sequencing is an effective tool in identifying inherited disorders.

Exome sequencing is an important technique for discovering which genes are involved in certain diseases.

The data used in the study are from the 2019 CoMMpass database, which is the most recent published, and is composed of 1,150 patients from around the world who have just been diagnosed with MM and have not yet started treatment.

Within this database, all kinds of information are available for subsequent analysis. The copynumber is determined from the data files containing the locus of the gene and the value of the log ratio. To calculate the copynumber associated with the TP53 gene, this first file was crossed with the coordinates of the gene extracted from the Ensembl database using the bedtools tool. If the results for the achievement are less than -0.1 it means that the gene presents a deletion, while if they are greater than that value it means that it presents a gain.

To reach the objectives of the study, a series of analyzes are carried out. One of them is differential expression analysis, which determines whether there are genes that change their expression between two groups of interest. These groups can be, for example, cancer patients and healthy patients. This analysis is based on a hypothesis test in which it tries to test the null hypothesis that the gene does not change expression between the groups tested. One of the methods to analyze the differential expression is SAM, which can be performed in the R statistical program using the siggenes package. It returns some parameters such as the fdr or the Fold Change that are essential when interpreting these expression changes. If the fdr is less than 0.05 it means that the gene is significant and within these, those with a positive Fold Change are overexpressed, while those with a negative Fold Change are under-expressed. With the R edgeR package, differential expression analysis is performed as well as the necessary prior normalization.

The unsupervised analysis is another of the statistical techniques that will be used in the present work. It is a type of machine learning that determines which observations show a similar pattern. Two examples are the dendrogram and multidimensional-scaling. To determine these recurrence patterns, they are based on proximity matrices, that is, it is necessary to know the distances between the elements. This distance is commonly calculated with the Euclidean distance, although other metrics exist.

With the survival analysis, the prognosis and evolution of the disease in patients is evaluated since it studies the time that passes until the event of interest occurs, which can be death, relapse in the disease ... etc.

The censored data correspond to those patients for whom there is no information about the event because they are lost to follow-up and the truncations correspond to the patients who enter the analysis once the study has begun.

The survival function is a summary of lifetimes, which can be represented graphically by a Kaplan-Meier graph.

The Kaplan-Meier method is a non-parametric technique widely used in survival analysis since it consists of a graphical representation of a function called the survival curve, where each step represents the occurrence of the event in question.

When there are several interest groups, the Log-Rank test is used to compare whether there are differences between the survival curves of each of the groups. This is tested with a hypothesis test in which we try to test the null hypothesis that the curves are equal. One of the limitations of this test is that it only indicates whether there are differences but does not return a parameter that quantifies these differences. This parameter is returned by the Cox regression and is called the Hazard ratio. The Hazard ratio is a representation of how much an event is more likely to occur than it is not.

Another of the statistical techniques used in the study is machine learning, which is made up of a set of computational processes that perform a task for which they are not programmed. They adapt to the

data through experience to offer optimal results in the task that has been entrusted to them. This adaptation process is called training. The validation part is in which the training results are generalized to other data that have not been entered before.

Machine learning is classified into supervised and unsupervised learning. The first uses already tagged historical data while the second uses data that is not tagged. Supervised learning can in turn be divided into regression or classification problems according to whether the response it predicts is continuous or categorical respectively.

Support vector machines (SVM) is a machine learning algorithm that is based on the hyperplane concept, which separates the samples of the different classes. The maximum margin hyperplane is the one with the greatest distance to both classes. The soft margin is the one that allows some sample to be on the opposite side of what it should be because it does not belong to that class. Kernel functions are those that are used when the problem is not linear and therefore the set of observations cannot be separated by a hyperplane. Kernel functions are linear, radial, polynomial, and sigmoid.

The KNN algorithm is a non-parametric method that classifies an object based on its similarity to nearby objects. It is one of the simplest since it is based on the idea that a new object is classified in the most frequent class in which its k closest neighbors are classified. There is no predetermined value for k so it is defined according to the requirements of each case, and an odd value is usually chosen to avoid ties. For this algorithm it is necessary to know the distance between the observations, to know this value there are several metrics such as the Chebyshev distance or the Manhattan distance, the Euclidean distance being the most used.

Random Forest is another machine learning technique that can be used for both categorical variables and continuous variables, that is, it is multiclass, which is an advantage over other algorithms. It is a supervised learning method that uses already labeled data to learn how to classify other new, unlabeled data. It is based on the combination of a multitude of decision trees that each return one class prediction, the final prediction of the model being the one returned by the most trees, that is, the most “voted”. Decision trees are made up of several nodes in which decisions are made to classify observations based on predictor variables. It starts from a first node that branches out into others until reaching the leaves, where the decision is made since each leaf belongs to a class and the chosen one will be the class that that tree will offer.

Both Random Forest and KNN are carried out in the statistical program R using the caret library.

Statistical parameters such as sensitivity, specificity, precision, negative predictive value or accuracy are used to evaluate the performance of the algorithms.

Through overrepresentation analysis, we try to study the functionality of genes. It is carried out on the Webgestalt website, which offers different analyzes such as enrichment analysis, analysis based on network topology and over-representation analysis (ORA), the latter being the one used in the present work.

In the first instance, patient groups were obtained by manual stratification of the database. In the present work, the study of the 4 most common translocations was carried out, which are t (4; 14), t (11; 14), t (14; 16) and t (14; 20). To verify the presence of each translocation, the partners determined by Delly were found to coincide with the genomic positions described in the literature. To select the samples that present deletion and mutation of TP53, the file of non-synonymous mutations of CoMMpass was used and the position of the TP53 gene was searched in the Genome Browser of the UCSC and the samples of the file that had position within this genomic interval were selected . In this way, 84 mutations were detected, some of them in the same patient, that is, the same subject had several mutations.

In a second approach, Seq-FISH data provided directly by the CoMMpass database were collected and compared with the results obtained by the manual approach. Although the results were similar, it was finally decided to use those offered by Seq-FISH to avoid discrepancies with previous studies. Finally, the t (14; 20) was discarded since the groups in which said translocation was associated with the TP53 mutation or deletion presented very few samples.

Once the groups were defined, survival time was calculated and represented in a Kaplan Meier graph using the variable "disease progression" coded at 0 and 1, with 0 being no progression and 1 being progression, to study the evolution of the illness.

As results, it was obtained that the mutation in TP53 is associated with a lower survival of the patients as well as the presence of t (4; 14) and t (14; 16) was also related to this fact, on the contrary to the presence of t (11; 14) that did not imply a shorter survival.

After analyzing the prognostic groups, the gene signature was established for the different cytogenetic events, performing 4 contrasts (one for each of the translocations and another for TP53. In each of these contrasts, the interest group that is formed by patients who present translocation or alteration in TP53 but do not present any other event, and the control group, made up of those patients who did not present any of the events analyzed, the control group was made up of 318 patients for all contrasts.

Prior to the differential expression analysis, a multidimensional scaling was performed for each contrast in which the distribution of the patients was represented according to whether or not they presented the different cytogenetic alterations. For this, it was necessary to transform the normalized data to log base 2. The MDS was performed with the SIMFIT program, introducing the expression matrix in log2 in said program, excluding genes that presented less than one reading in their samples.

For the three translocations, a differentiation was observed in the distribution of the groups, clearly appreciating how the patients presenting each of them had a different distribution from the control group. This difference is not so clear in TP53, since all the patients (with or without the alteration) were distributed homogeneously in the graph.

After studying the structure of the data, a search was made for the gene expression signature for each of the groups. Genes with a FDR less than 0.05 (significant genes) were selected to later cross those of the 4 contrasts in a Venn diagram and see which ones are unique to each comparison.

A box plot was made for the two most overexpressed and the two most under-expressed genes of each contrast. That is, within those that present a positive Fold Change (overexpressed genes) the two that have a lower FDR. And within those that present a negative Fold Change (under-expressed genes) the two that present a lower FDR.

The next step was to perform an analysis of gene overrepresentation in biological pathways and functions (ORA) with the unique genes of each contrast to determine possible cellular functions of these genes. It was carried out in Webgestalt, extracting the 20 routes and functions with the highest overrepresentation. The KEGG, Reactome and Gene Ontology (GO) databases were considered and genome protein-coding was used as a reference set.

Graphically, it is represented by a bar graph, which corresponds to the pathways and the size of each of them is represented on the other axis, that is, how many genes form them. As a secondary axis, the -log10 (fdr) that indicates whether the path is significant or not.

The last step was to predict the progression of the disease with machine learning techniques. To do this, the database was first divided into two matrices, the training matrix and the validation matrix. The first one consisted of 2/3 of the samples while the validation one was made up of the remaining third of the samples, chosen randomly. In order for there to be enough patients with progression and without progression in both matrices, first of all, the data set was divided into patients with progression and those without progression, and 2/3 of the samples showing progression and progression were randomly chosen. 2/3 of the samples that did not show progression. When added together, the training matrix was formed, which was made up of 409 samples (2/3 of 613, which is the number of samples in the database). The validation matrix was made up of the 204 remaining samples.

This process was carried out for the exclusive genes of each contrast, so that 8 matrices were obtained, one for training and the other for validation for each of the 4 contrasts.

Once the training and validation matrices were formed, the prediction was carried out with the different machine learning algorithms.

With the weighted SVM technique it is necessary to first calculate the weights of each of the groups to weight them. The results obtained were a specificity and precision of zero, which indicates that the algorithm does not capture at all well the patients who do not present the progression. The sensitivity obtained was 1, which indicates that, however, it does capture patients with progression well. Globally, the accuracy was 0.5147, which is not considered good, but is acceptable.

When making the predictions with all the genes, the problem of overfitting arises, so it was decided to make a selection of the important genes with the Boruta package to make the prediction only with them.

Boruta returns the important genes of each contrast. In the case of TP53, 9 important genes were obtained, for t (4; 14) there were 8 genes returned as important, in the case of t (11; 14) 9 important genes were obtained and finally for t (14; 16) boruta returned 7 important. New training and validation matrices were created, the same as the previous ones but this time only with the genes considered important, and the prediction was performed again.

This time the specificity and precision were no longer 0, so the algorithm avoiding overfitting captures and better classifies patients without progression.

To use the KNN algorithm, it is necessary to recode the group variable since it has to be a factor type variable, so that 0 (no progression) became “g0” and 1 (progression) became “g1”.

As results, moderately high values were obtained for specificity, which indicates that it is robust to patients without progression. Accuracy values are around 0.55, which can be considered acceptable without being good at all.

When predicting with this algorithm for the important variables only, the sensitivity was significantly improved, and the accuracy remained similar.

In the case of the Random Forest algorithm, it is also necessary for the group variable to be a factor, which is why the “g0” and “g1” coding were kept. The results were good in both sensitivity and specificity, so it captures both groups of patients well.

All this analysis led us to the conclusions that on the one hand, the presence of mutation in TP53 and the presence of t (4; 14) and t (11; 14) imply a lower survival and a worse prognosis and evolution of the disease, and that nevertheless, the t (14; 16) is not related to this fact.

On the other hand, after analyzing the results obtained with the 3 machine learning algorithms, it cannot be concluded that none of them is more robust when it comes to predicting disease progression since the results are similar, except in the case of SVM, which when making the prediction with all genes does not make a good prediction, because it does not capture progression-free patients.

INTRODUCCIÓN

Como indica el título del trabajo de fin de grado, su objetivo fundamental es el integrar una serie de análisis y herramientas estadísticas y bioinformática para identificar grupos de individuos con un alto riesgo de padecer mieloma múltiple debido a la pérdida de la función biológica de un gen supresor tumoral llamado TP53.

Para llegar hasta este objetivo, iremos desde el conocimiento previo de la enfermedad hasta la metodología estadística empleada, pasando por la explicación de la compleja tecnología que se utiliza para la obtención de los datos que se analizaron en este trabajo.

Por eso, en el primer capítulo se exponen las características clínicas y biológicas de la enfermedad de la que trata el trabajo, el mieloma múltiple (MM). En él se presta particular atención a las alteraciones citogenéticas ya que serán las que definan los grupos de pacientes con MM sobre los que se llevará a cabo el análisis estadístico.

El segundo capítulo está dedicado a explicar los aspectos básicos de la tecnología de secuenciación de próxima generación o *next generation sequencing* (NGS) con la que se generarán los datos que se usarán posteriormente en las herramientas y procedimientos estadísticos.

En el capítulo tres se recoge la metodología estadística utilizada en este trabajo. Entre los métodos y técnicas descritos podemos encontrar el Análisis de Supervivencia, Análisis no Supervisado y Machine Learning.

Finalmente, los resultados del análisis se detallan en el capítulo 4 y las conclusiones derivadas de éstos en el capítulo 5.

Capítulo 1.- Contexto médico y biológico.

1.1 Mieloma Múltiple

El mieloma múltiple (MM) es una neoplasia hematológica caracterizada por la acumulación incontrolada de células plasmáticas (CP) clonales en el interior de la médula ósea (MO). Estas CPs producen grandes cantidades de una inmunoglobulina monoclonal, también denominada componente monoclonal o paraproteína. Las manifestaciones clínicas de esta neoplasia, que se derivan del daño orgánico que ocasionan las CP y la paraproteína, se engloban bajo el acrónimo CRAB (hiperCalcemia, insuficiencia Renal, Anemia, lesiones óseas=Bone lesions).

El síntoma más frecuente es el dolor óseo (80%), que principalmente se localiza en huesos con abundante médula ósea como la columna vertebral, el esternón, las costillas y parte proximal de las extremidades. En ocasiones se llegan a producir fracturas óseas y en fases avanzadas de la enfermedad puede aparecer hipercalcemia. La infiltración de las CP en la MO da lugar a un desplazamiento de las otras series hematopoyéticas, provocando principalmente anemia (70%). La insuficiencia renal, que se observa en torno al 20% de los pacientes, está motivada por el acúmulo del componente monoclonal en los túbulos renales, dando lugar al llamado "riñón del mieloma". La expansión de la inmunoglobulina monoclonal ocasiona una disminución de las Igs policlonales originando un estado de inmunodeficiencia que favorece la aparición de infecciones. De hecho, esta es una de las principales causas de mortalidad de los pacientes con MM (AEAL, 2017).

El MM representa un 10% de todos los cánceres hematológicos, por lo que se trata de la segunda neoplasia hematológica más frecuente, por detrás del linfoma no hodgkin. En España se diagnostican 4 nuevos casos por cada 100.000 habitantes/año. La edad media en el momento del diagnóstico es de 65 años, y menos del 15% de los casos corresponden a menores de 50 años. Es ligeramente más frecuente en hombres que en mujeres y en personas de raza negra (AECC, 2021). La etiología del MM aún es desconocida y no se ha demostrado que sea una enfermedad hereditaria.

Cuando la transformación neoplásica de las células plasmáticas no produce trastornos clínicos, se habla de una gammapatía monoclonal de significado incierto (GMSI), que es una etapa premaligna y asintomática que no requiere tratamiento. Si bien la mayoría de las GMSI no se transforman a MM, este riesgo de progresión persiste indefinidamente con una tasa de transformación anual del 1%.

El pronóstico del MM es muy variable ya que depende de numerosos factores. Uno de los más reconocidos es el estadio en el que se encuentra la enfermedad, que se determina según el Sistema Internacional de Estadificación (ISS) basado en los valores séricos de la beta2-microglobulina y la albúmina. Más recientemente se han incorporado a este índice pronóstico las alteraciones citogenéticas de alto riesgo (MMRF, 2021).

Hasta finales del siglo pasado, los tratamientos disponibles para el MM eran los corticosteroides, agentes alquilantes y otros quimioterápicos tradicionales. El tratamiento del MM ha experimentado grandes avances en las últimas décadas gracias a la introducción de nuevos agentes terapéuticos como los inhibidores del proteasoma (IP), los inmunomoduladores (IMIDs) y los anticuerpos monoclonales.

Esto ha contribuido a mejorar la supervivencia de los pacientes con MM, llegándose a una mediana de 7-8 años.

En la actualidad, existen diversas modalidades terapéuticas cuya elección depende fundamentalmente de que el paciente esté en condiciones, por su edad y estado general, de someterse a un trasplante autólogo de células hematopoyéticas (TAPH).

Los pacientes de nuevo diagnóstico y candidatos a TAPH reciben un tratamiento de inducción que consiste habitualmente en la combinación de bortezomib, lenalidomida y dexametasona (régimen que recibe el nombre de VRD) con el fin de alcanzar la mejor respuesta posible. Posteriormente, los pacientes se someten a un TAPH que consiste en la administración de altas dosis de quimioterapia con melfalán, seguida de la infusión de células progenitoras hematopoyéticas autólogas. Dependiendo de la

respuesta alcanzada después del TAPH, los pacientes reciben un tratamiento de mantenimiento, habitualmente con lenalidomida.

Los pacientes no candidatos a TAPH suelen recibir distintas combinaciones de los fármacos indicados anteriormente, según las características de cada paciente. En concreto, la combinación de lenalidomida más dexametasona es una excelente opción, sobre todo en pacientes mayores y frágiles

Aunque la supervivencia de los pacientes ha mejorado en los últimos años, el MM sigue siendo una enfermedad incurable. En la mayoría de ellos se observa un patrón de recurrencia con etapas de remisión seguidas de recaídas hasta que la enfermedad se vuelve refractaria, es decir, deja de responder al tratamiento (Krzeminski et al., 2016). Es por ello que gran parte de la investigación está orientada a conseguir terapias que retrasen lo máximo posible la progresión de la enfermedad.

1.2 Alteraciones genéticas en el MM

El mieloma múltiple (MM) es una neoplasia de células plasmáticas caracterizada por una gran complejidad genómica y molecular, lo que explica en buena medida la variabilidad observada en la evolución clínica y la respuesta al tratamiento. A diferencia de lo que sucede en algunas leucemias y linfomas, no se han encontrado anomalías cromosómicas ni genéticas específicas, aunque sí se han descrito numerosas alteraciones citogenéticas recurrentes con importantes repercusiones en el pronóstico.

Las alteraciones genómicas del MM se pueden clasificar en translocaciones, alteraciones en el número de copias ("copy number variation") y mutaciones puntuales.

Traslocaciones cromosómicas

Una translocación se define como un cambio de localización de un segmento de un cromosoma a otro lugar del genoma. La mayoría de las traslocaciones afectan al cromosoma 14, concretamente al gen de la cadena pesada de las inmunoglobulinas (*IGH*) situado en el locus 14q34. Las traslocaciones de *IGH* a diferentes regiones del genoma se observan hasta en el 60% de los MM. La consecuencia del *producto de fusión resultante es una desregulación de los oncogenes que se sitúan bajo el control del* "enhancer" de *IGH*. Las más frecuentes son la t(11;14), detectada en el 15-20% de los casos, que origina un aumento de la expresión de la ciclina D1; la t(4;14) que aparece aproximadamente en el 15% de los MM y que tiene como resultado la desregulación simultánea de 2 genes, *FGFR3* y *NSD2* (anteriormente conocido como MMSET); y la t(14;16), observada como mucho en el 5% de los MM, que conlleva un aumento de la expresión del oncogén MAF. Se considera que las traslocaciones de *IGH* son eventos oncogénicos primarios en el desarrollo del MM. Las traslocaciones t(4;14) y la t(14;16) se incluyen en la mayoría de las clasificaciones pronósticas como marcadores genéticos de mal pronóstico. No obstante, la introducción de los IP en los esquemas terapéuticos ha mejorado la supervivencia de los pacientes con estas traslocaciones. En cambio, la t(11;14) no tiene una repercusión negativa en la supervivencia de los pacientes con MM.

Aunque en una proporción mucho menor que las traslocaciones del gen *IGH*, en el MM se detectan traslocaciones del oncogen *MYC* hasta en un 20% de los casos. A diferencia de las traslocaciones de *IGH,* los reordenamientos de *MYC* constituyen eventos oncogénicos secundarios. Aunque no existe unanimidad en cuanto al valor pronóstico de las alteraciones del oncogén *MYC*, los últimos estudios demuestran supervivencias más cortas en los pacientes con reordenamientos de *MYC*.

Anomalías en el número de copias (CNA)

Habitualmente la célula mielomatosa gana y pierde tanto cromosomas completos como brazos y regiones cromosómicas, que hacen que casi todos los MM sean aneuploides, es decir no tengan una dotación diploide normal. Los MM hiperdiploides se caracterizan por la presencia de trisomías que afectan especialmente a los cromosomas impares y por una frecuencia baja de alteraciones estructurales. El cromosoma que más se ve afectado por ganancias y pérdidas es el cromosoma 1, en el que el brazo

q se gana hasta en el 60% de los casos y regiones de 1p se pierden en un 30% de los pacientes, aproximadamente. Para las ganancias de 1q que implican la adquisición de más de 3 copias se ha acuñado el término de amplificación de 1q. Otras CNA frecuentes son la monosomía del cromosoma 13 y la deleción de 17p13 que incluye el locus del gen supresor TP53. La deleción de 17p se observa tan solo en el 8-10% de los casos en el momento del diagnóstico, si bien su frecuencia aumenta significativamente en los pacientes con MM en estadios avanzados y en progresión.

La deleción de 17p es probablemente uno de los marcadores genéticos asociados a peor pronóstico. Las alteraciones del cromosoma 1 también se han asociado con menor supervivencia, especialmente las pérdidas de 1p. En el caso de las alteraciones de 1q, es la amplificación, definida como la presencia de más de tres copias, la que peor pronóstico acarrea. Por el contrario, los cariotipos hiperdiploides que presentan trisomías como únicas alteraciones se asocian consistentemente con las supervivencias más prolongadas.

Mutaciones puntuales

Las mutaciones puntuales consisten en el cambio de en un único nucleótido del ADN. La secuenciación mediante técnicas de nueva generación (NGS) del exoma y del genoma completo de miles de pacientes con MM ha confirmado la heterogeneidad mutacional del MM. Las mutaciones somáticas más frecuentes son las que afectan a los oncogenes *RAS*, fundamentalmente *KRAS* y *NRAS*. Las mutaciones de *FAM46C*, *DIS3* y *TP53* aparecen con una frecuencia de aproximadamente el 10% cada una de ellas. Si se consideran las rutas biológicas en lugar de genes individuales, las vías de señalización que con mayor frecuencia presentan mutaciones en el MM son la vía RAS/MAPK (*KRAS, NRAS* y *BRAF*) en el 40% de los casos, la vía NFkB (*TRAF3*, *CYLD* y *LTB*) en el 20%, y las vías de reparación de ADN (*TP53, ATM* y *ATR*) en el 15%. Otros genes recurrentemente mutados en MM son *PRDM1*, *IRF4* y *SP140*, implicados en la diferenciación del linaje B, y los genes supresores de tumores *DIS3* y *FAM46C* cuyo papel en la patogenia del MM es poco conocido.

1.3 Estudio CoMMpass

La heterogeneidad y variabilidad genética del MM exige un gran esfuerzo por parte de la comunidad científica a la hora de realizar análisis genómicos en miles de pacientes con MM, para entender mejor las bases biológicas de la enfermedad y así poder aspirar a encontrar estrategias terapéuticas individualizadas en función del contexto genético de cada paciente.

En este sentido, la finalidad del estudio CoMMpass, ("Relating Clinical Outcomes in MM to Personal Assessment of Genetic Profile"), promovido por la "Múltiple Myeloma Research Foundation" (MMRF), es proporcionar la mayor cantidad de información genómica de más de 1000 pacientes diagnosticados de MM, realizando un seguimiento de la respuesta al tratamiento y de la evolución a lo largo de un tiempo prolongado.

La MMRF es una organización fundada en 1998 que tiene por objetivo acelerar la cura del mieloma múltiple, y el estudio CoMMpass es la piedra angular de la iniciativa para la medicina personalizada emprendida por la MMRF en los últimos años.

En el estudio CoMMpass participaron centros hospitalarios de 4 países, entre los que se encuentra España. El Hospital Universitario de Salamanca centralizó la recepción y procesamiento de las muestras de diversos hospitales nacionales antes de proceder a enviar el material necesario de cada paciente para el análisis genómico en el "Translational Genomics Research Institute (TGen)" de Phoenix (Arizona).

Una de las características principales de esta base es el largo tiempo de seguimiento al que son sometidos los participantes, ya que la evolución de cada paciente para analizar la respuesta al tratamiento es registrada cada 6 meses durante 8 años.

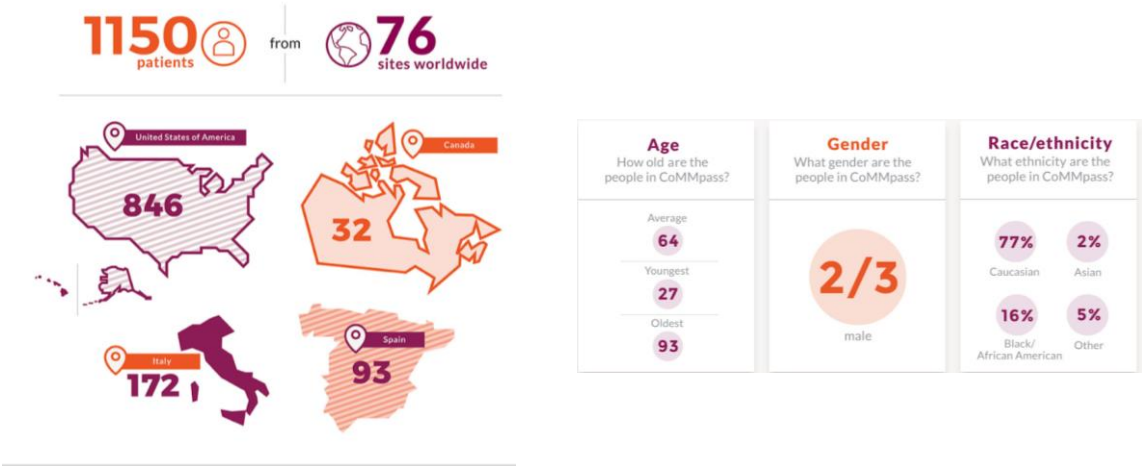


Figura 1. Datos descriptivos de la base CoMMpass.

Al estudiar un número tan grande de pacientes, la probabilidad de que cualquier nuevo diagnosticado de MM sea un caso similar a alguno de los pacientes estudiado en CoMMpass es muy elevada.

1.4 Definiciones biológicas

El ADN es un ácido nucleico que se halla en el núcleo de las células y contiene la información genética hereditaria. Está compuesto por 4 bases nitrogenadas que son adenina, guanina, citosina y timina. Así, las bases nitrogenadas son las unidades básicas que forman los ácidos nucleicos y el orden en que se encuentran determinará nuestro código genético.

La función principal del ADN es suministrar la información genética. Esto se consigue mediante la replicación, proceso por el cual la información genética se transfiere de una célula a las células hijas; la codificación de las proteínas; el metabolismo celular y la mutación que determina nuestra evolución como especie. (Valencia, 2017).

El ARN o ácido ribonucleico es otro tipo de ácido nucleico. Su principal función es la síntesis de proteínas. Está formado por las bases nitrogenadas adenina, guanina, citosina y uracilo. El ARN se clasifica en ARN mensajero (ARNm) que es el encargado de transportar la información útil para la síntesis de proteínas; ARN de transferencia (ARNt) que transporta los aminoácidos para esa síntesis y ARN ribosómico (ARNr) que ayuda a leer los ARNm (Biología, 2021).

Mientras que el ADN contiene la información genética, el ARN es quien permite que esa información sea comprendida por las células. Otra de las diferencias entre ADN y ARN es que el primero está formado por una hebra doble de nucleótios, mientras que la del segundo es una cadena simple.

1.5 Objetivos

El objetivo principal de este trabajo es estudiar el impacto que tienen las deleciones del cromosoma 17p y las mutaciones del gen TP53 en el perfil génico de las translocaciones primarias más frecuentes que tienen lugar en el mieloma múltiple.

A partir de este objetivo inicial, surge un sub-objetivo que es el establecimiento de las características biológicas de las distintas firmas mediante análisis de sobrerrepresentación de vías.

Otro objetivo secundario es la predicción de la progresión para las dichas firmas.

Capítulo 2.- Datos

2.1 Secuenciación de próxima generación (NGS)

La secuenciación de próxima generación (del inglés, *next generation sequencing*,NGS) es una tecnología revolucionaria cuya aparición se debe a un cúmulo de avances en los campos de la física, química y la biología, como fueron el desarrollo de la secuenciación de Sanger, la generación del marcaje con fluorescencia y la automatización y masificación del proceso de secuenciación.

Esta tecnología permite el estudio completo y detallado del genoma y del transcriptoma, motivo por el que ha adquirido una gran importancia en el campo de la oncología al ser el cáncer una enfermedad estrechamente ligada al estudio del genoma.

Las limitaciones o desventajas de la NGS surgen de la implementación de la infraestructura necesaria, fundamentalmente por el enorme coste en capacidad de almacenamiento del ordenador, ya que se manejan grandes cantidades de datos (Behjati & Tarpey, 2013). Sin embargo, los beneficios obtenidos por el uso de las técnicas de NGS superan con creces a las limitaciones, ya que a nivel génico no solo permite la identificación de variantes de un nucleótido, sino también la detección de cambios en el número de copias, y de eventos genómicos como traslocaciones, inserciones e inversiones de grupos de nucleótidos. A nivel de ARN la NGS también presenta notables beneficios como el descubrimiento de nuevas variantes de ARN, la determinación de eventos de splicing (corte y empalme) alternativo o la cuantificación de ARNm para el análisis de expresión génica. Además podemos encontrar múltiples aplicaciones en la identificación de nuevos patógenos, el estudio del microbioma, o el análisis de factores epigenéticos y de las interacciones ADN-proteína.

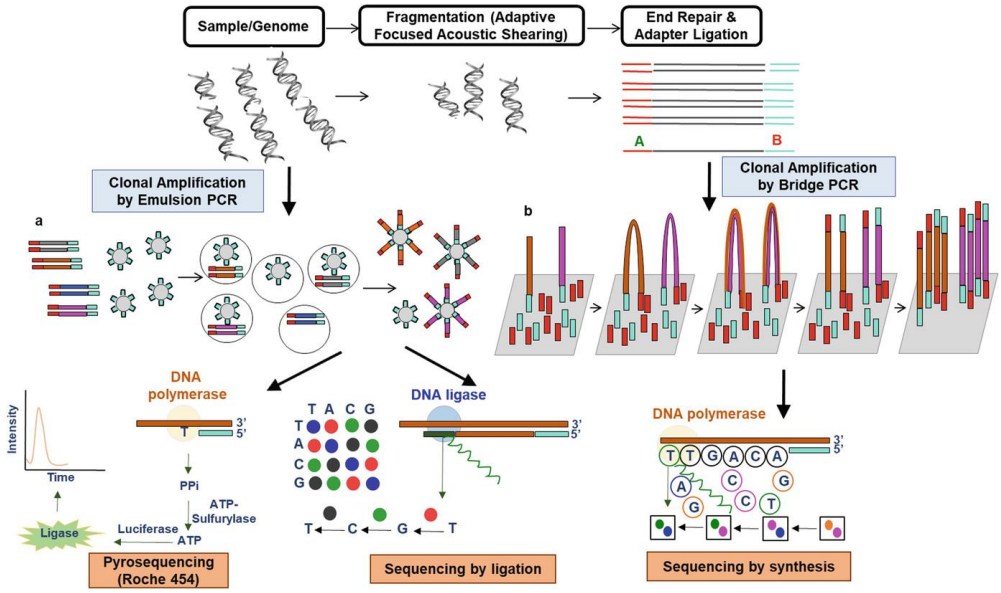


Figura 2.- Representación de los pasos involucrados en la secuenciación de DNA utilizando plataformas de NGS. Extraída de (Gupta & Verma, 2019).

2.1.1 Secuenciación del RNA (RNA-Seq)

El análisis estadístico de datos multi-ómicos es de gran importancia debido a la cantidad de aplicaciones que tiene en el campo de la biomedicina. Una de las ciencias ómicas que ha tenido un extraordinario desarrollo en las últimas décadas ha sido la transcriptómica, cuyo objetivo es el estudio de la expresión génica. Se considera que un gen se expresa cuando al partir un fragmento de ADN codificante se genera una molécula de ARNm en un proceso conocido como transcripción. La

transcripción es un proceso que tiene lugar en el núcleo celular, y es un proceso dinámico que depende de distintos factores como el tipo de tejido, la presencia de estímulos externos e internos, enfermedades, etc. Actualmente, la transcriptómica goza de gran relevancia en diferentes campos de la biología, ya que su desarrollo ha permitido determinar cómo afectan los cambios de expresión génica en la aparición y avance de diferentes enfermedades, o el estudio de la eficacia de un tratamiento a la hora de combatirlas (Carrasco, 2015).

Las técnicas de secuenciación de RNA (RNA-seq) son una aproximación para cuantificar y evaluar datos transcriptómicos, al igual que otras tecnologías surgidas previamente, como los microarrays, Sin embargo, la RNA-seq ha conseguido superar las limitaciones que presentaban los microarrays, al no ser una matriz cerrada en la que solamente un conjunto de genes diana podrán ser determinados, sin considerar las características transcripcionales específicas de las muestras analizadas.

El término RNA-seq se acuñó por primera vez en el año 2008 por (Lister et al., 2008). pero los orígenes de la técnica RNA-seq se remontan al año 1975 con la metodología de secuenciación desarrollada por Sanger y Coulson.

La RNA-seq se apoya en la tecnología NGS y se utiliza para cuantificar la cantidad de RNA en una muestra biológica en un momento concreto analizando el transcriptoma celular (Wang, Gerstein, & Snyder, 2009).

Entre las aplicaciones más frecuentes se encuentran la detección de transcritos codificantes, y el contraste de expresión génica diferencial entre dos o más condiciones.

El análisis de RNA-seq se compone de varios pasos. El primero es la fragmentación del RNA en porciones que son las lecturas. Cada una de ellas es secuenciada y alineada frente a un genoma o un transcriptoma de referencia, de modo que se considera que, un gen o transcrito concreto presentará mayor o menor expresión en función del número de lecturas que hayan alineado contra la referencia del mismo. La determinación del número de lecturas que corresponde a cada gen se lleva a cabo mediante el paso de contaje, que consiste en recoger los datos de conteo para resumirlos en tablas, de manera que se puedan emplear en la realización de diferentes análisis, por ejemplo, en la identificación de genes diferencialmente expresados (DE) (Carrasco, 2015). Es importante tener en cuenta que la longitud del gen puede influir en el número de lecturas, ya que un gen de mayor longitud puede presentar más lecturas que otro con longitud reducida, por lo que, en estudios comparativos donde se consideran múltiples genes es interesante el escalado de la expresión génica en función de la longitud de cada uno de los genes analizados a través de un proceso conocido como normalización.

En este contexto normalizar no significa transformar los datos para que sigan una distribución normal, sino que se busca que todas las muestras estén en la misma escala, para que ninguna tenga más probabilidad de tener genes diferencialmente expresados por su profundidad de secuenciación. Es decir, con la normalización se busca que las distribuciones de los datos sean comparables, reduciendo a su vez la probabilidad de errores de tipo I o falsos positivos.

Una de las técnicas de normalización más comúnmente empleadas en RNA-seq es la propuesta por el algoritmo EdgeR y desarrollada por (Robinson et al.,2010) conocida como la media recortada de M valores, o TMM (del inglés *Trimmed Mean of M-values*). Sin embargo, existen otros métodos más clásicos utilizados en la normalización, como por ejemplo *FPKM (Fragments Per Kilobase Million)* o *RPKM (Reads Per Kilobase Million)* entre otros. En el caso de *RPKM* el procedimiento consiste en contar el total de lecturas en una muestra y dividirlo entre un millón obteniendo un factor de escala por millón para después dividir los recuentos de lecturas entre dicho factor. Normalizando lo anterior para la profundidad de secuenciación se obtienen las lecturas por millón *RPM* (*reads per million*) y dividiendo este valor entre la longitud del gen se obtiene finalmente el RPKM.

En el caso de FPKM el procedimiento es muy similar al RPKM. La diferencia entre ambos es que mientras que en RPKM cada lectura corresponde a un fragmento, en FPKM dos lecturas pueden corresponder a un mismo fragmento y no cuenta este fragmento dos veces.

2.1.2 Secuenciación del DNA

Las primeras secuencias de ADN se obtuvieron en los años 70 pero requirieron métodos muy laboriosos para conseguirlas. Uno de los primeros fue la secuenciación con dideoxinucleótidos desarrollado por Sanger y Coulson. Actualmente, gracias al desarrollo de la NGS, se dispone de una tecnología de secuenciación que hace que el proceso sea mucho más rápido y sencillo.

La secuenciación de ADN es un procedimiento por el cual se determina el orden físico de las bases en el ADN (adenina, guanina, citosina y timina). El conocimiento de estas secuencias es muy importante en la investigación orientada al diagnóstico médico, entre otras áreas, ya que comparando secuencias sanas con otras mutadas se pueden diagnosticar diversas patologías y ayudar a la elección del tratamiento más adecuado (Chmielecki & Meyersom, 2013).

Los métodos de secuenciación más frecuentes son la secuenciación del genoma completo, considerada una gran herramienta en la investigación, pues es capaz de identificar trastornos hereditarios, así como identificar mutaciones involucradas en la progresión del cáncer. Otro método común es la re- secuenciación dirigida, por el cual solo se analizan regiones de interés (Illumina, 2021). Esta secuenciación puede estar orientada al estudio de los exones de todo el genoma, o bien centrarse en piezas concretas como es el caso de los paneles génicos.

2.1.3 Secuenciación del exoma

La secuenciación del exoma (del inglés *whole exome sequencing* o WES) es una técnica empleada para la secuenciación de las regiones genómicas codificantes, conocidas como exones. Los exones son las regiones de ADN que codifican proteínas y que en los humanos corresponden al 1% del genoma (Ng et al., 2009). Esta técnica permite la detección tanto de mutaciones somáticas o germinales, como variantes en el número de copias del ADN. Dado su menor tamaño en comparación con la secuenciación de genoma completo, su empleo puede ser muy útil en la toma decisiones en la clínica, como la determinación de la elección de tratamientos en pacientes con una determinada patología, o en diagnóstico prenatal. Es por esto una herramienta importante en el descubrimiento de genes que intervienen en ciertas enfermedades, como por ejemplo, de genes impulsores de cáncer.

El cáncer es una enfermedad poligénica y cada tumor contiene unas mutaciones somáticas distintas en diversos oncogenes y genes supresores tumorales, por ello es complicado identificar cuáles son los genes que causan dicha enfermedad. Existen múltiples estudios del exoma del cáncer con el fin de resolver estas incógnitas.

Capítulo 3.- Material y métodos

3.1 Preprocesamiento de los datos

Los datos utilizados en este estudio son los de la base CoMMpass de 2019 (la más reciente de la que disponemos) y está formada por alrededor de 1150 pacientes recién diagnosticados y que aún no han empezado con el tratamiento.

Para la determinación del número de copias (*copynumber*) se emplearon los datos disponibles de secuenciación de exoma completo, segmentados por ventanas genómicas con un número de copias similar. Estos datos aparecen recogidos en archivos de texto delimitado por tabuladores con la extensión "seg". Dicho archivo contiene los intervalos genómicos de cada uno de los segmentos detectados y el valor del $\log_2$ del ratio del número de copias de cada una de las muestras tumorales. El valor del $\log_2$ del ratio del número de copias fue calculado como el cociente entre el número de copias de la muestra tumoral y la referencia o muestra control (LR). Éstos son archivos de texto delimitados por tabulaciones sin encabezados y con extensión ".seg" que se cruzan con la herramienta *bedtools*, mediante la consola de Ubuntu en Linux.

Para el cálculo del número de copias asociado al gen *TP53*, se realizó un cruce de intervalos genómicos entre el archivo seg inicial y las coordenadas de este gen extraídas de la base de datos Ensembl en su versión del genoma humano hg19 o GRCh37. Este análisis se llevó a cabo con la herramienta *bedtools* en la consola de Ubuntu. La determinación del estado del número de copias de *TP53* se estableció fijando un criterio de corte de +/- 0.1 en el valor del LR, de modo que si el valor es menor a -0.1 significa que el gen presenta una deleción en esa muestra. Por el contrario, si el LR es mayor que 0.1, se considera que el gen presenta una ganancia.

Cabe decir que *bedtools* es una herramienta imprescindible en el análisis genómico. Se desarrolló en el laboratorio Quinlan de la Universidad de Utah con la participación de científicos de varios centros de investigación del mundo. Esta herramienta permite cruzar, fusionar o complementar intervalos genómicos de distintos archivos (A & N, 2021).

En este caso la función a utilizar es *bedtools intersect* para detectar superposiciones entre los dos conjuntos de datos con el siguiente código:

```
bedtools intersect -wa -wb -a archivo1.bed -b archivo2.bed > Resultado.bed
```

3.2 Análisis de expresión génica

3.2.1 Conceptos generales

Uno de los objetivos de la transcriptómica es la determinación de la expresión diferencial entre dos o más condiciones a contraste. Mediante estos análisis es posible determinar si existen genes que cambien su expresión entre dos grupos de interés como por ejemplo entre pacientes con cáncer y un grupo control sano. El análisis de expresión génica diferencial se basa en un contraste de hipótesis donde la hipótesis nula (H_0) supone que la expresión del gen no cambia entre los grupos contrastados (gen no significativo), mientras que la hipótesis alternativa (H_1) dice que la expresión del gen sí cambia entre los grupos (gen significativo).

Un método estadístico clásico que se ha venido utilizando en los análisis de microarrays para analizar expresión génica diferencial es SAM *(Significance Analysis of Microarrays)* y se realiza con el paquete *siggenes* del programa estadístico de software libre R. Devuelve varios parámetros como el FDR y el fold change, indispensables para la interpretación posterior de los resultados. Brevemente, SAM utiliza una prueba *t* de Student modificada para calcular las diferencias entre los grupos contrastados y analiza si esa diferencia es significativa mediante el uso de permutaciones. Finalmente, SAM lleva a cabo un ajuste del *p*-valor obtenido de la prueba de permutaciones para evitar la aparición de errores de tipo I o falsos positivos. Este ajuste lo hace mediante el cálculo del *q*-valor, que, es una optimización propuesta

por (Storey, 2010) (46) del *False Discovery Rate* (FDR) propuesto por Benjamini y Hochberg en 1995 (Amat Rodrigo, 2016).

Así, para controlar el error de tipo I global, es decir, controlar el número de falsos positivos en este tipo de estudios masivos (genes significativos que no lo son) se suele estimar el FDR o se realiza algún otro tipo de ajuste para múltiples comparaciones, como el propio *q*-valor, el ajuste de Bonferroni o el FWER (del inglés *Family-wise error rate*), entre otros. El FDR en concreto es la proporción esperada de falsos positivos dentro de los resultados positivos (significativos), es decir, la proporción de test significativos que realmente no lo son. Es diferente del *p*-valor de cada contraste, ya que el *p*-valor es la probabilidad de que haya falsos positivos en cada una de las pruebas.

El FDR se controla estableciendo un límite de modo que, dentro de un conjunto de test considerados significativos, la proporción de falsos positivos no supere un valor concreto establecido como dicho límite.

El *Fold Change* (FC), por su parte, hace referencia a la ratio de expresión génica entre una condición final o problema y otra condición inicial o control. Este parámetro es comúnmente usado en análisis de datos de expresión génica, haciendo referencia al cambio de intensidad del gen. Así, el FC suele ser utilizado para determinar qué genes se sobreexpresan o se infraexpresan en cada condición.

Aunque el método SAM es muy popular en el estudio de microarrays, su aplicación en el análisis de datos procedentes de técnicas de secuenciación masiva no es apropiada, dado que este método asume una distribución normal de los datos y la RNA-seq no suele cumplir estos supuestos. Por este motivo se han desarrollado otro conjunto de técnicas cuya asunción subyacente es la aplicación de modelos de distribución basados en una distribución binomial negativa. En este sentido nos encontramos con algoritmos como el propuesto en el paquete *edgeR* de *Bioconductor* de R

3.2.2 edgeR

El paquete edgeR de R utiliza diferentes métodos estadísticos basados en la distribución binomial negativa.

La distribución binomial negativa es una distribución de probabilidad discreta considerada como una ampliación de la distribución geométrica y es usada en procesos en los que se repite un ensayo hasta conseguir el primer éxito.

Sea $\boldsymbol{X}$ una variable aleatoria que sigue una distribución binomial negativa $\boldsymbol{X}\sim BN(r,p)$ siendo r el número de resultados favorables y p la probabilidad de obtener un resultado A que es constante en todas las pruebas

Su función de probabilidad es:

$$f(k) = p(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

El paquete *edgeR* de Bioconductor normaliza mediante el método TMM (Trimmed Mean of M-values), es decir, la media truncada de M valores (Robinson & Oshlack, 2010). Para poder realizar la normalización es necesario que el objeto sea de la clase *DGEList*, que es un tipo de objeto formado por una tabla que contiene los contajes de lecturas por gen en cada muestra, además de los indicadores de grupo, tamaño de las librerías, que se utilizarán como factores de normalización, y una tabla con las características y anotaciones de cada uno de los genes analizados.

Tras generar el objeto *DGEList*, el siguiente paso es normalizar con la función ***calcNormFactors:*** *Calculate Normalization Factors to Align Columns of Count Matrix*. La función calcula los factores de normalización para escalar los tamaños de la biblioteca sin procesar. Este paso, como se viene indicando, se realizó especificando en el comando "Method" el método "TMM" **.**

Esto produce una matriz de datos normalizada sobre la que se realizará el resto del análisis aguas abajo con el paquete *edgeR*.

Hay que tener en cuenta también que, antes de empezar con el análisis de expresión diferencial, se tiene que depurar o filtrar el conjunto de datos. Esto consiste en eliminar aquellos genes que presentan un número bajo de lecturas alineadas que resulte irrelevante para el análisis en cuestión. En este caso, se consideraron para ser eliminados los que tengan menos de 4 conteos en todas las muestras (Carrasco, 2015).

A continuación, se realizó el análisis de expresión diferencial mediante la función exactTest del paquete edgeR. Esta función implementa un test exacto propuesto por (Robinson, McCarthy, & Smyth, 2010). Consistente en una generalización del test binomial exacto, que contrasta la diferencia de medias entre dos grupos de variables aleatorias binomiales negativas. Los *p*-valores obtenidos fueron ajustados mediante el FDR de Benjamini y Hochberg. Se consideraron como genes diferencialmente expresados aquellos que alcanzaron un FDR menor a 0.05 (5%).

3.4 Análisis no supervisado

El análisis no supervisado es un tipo de aprendizaje automático que estudia la estructura intrínseca de los datos con la finalidad de detectar qué elementos presentan un patrón similar. Trata problemas de dos tipos principalmente, de agrupación y de reducción de la dimensionalidad. En el presente trabajo se va a utilizar el dendrograma y multidimensional- scaling.

3.4.1 Dendrograma

El *Clustering Jerárquico* es una técnica de agrupamiento de datos que se basa en comparar las distancias entre los elementos sujetos a estudio. Los grupos de datos resultantes se llaman *clusters*.

La manera de representar un clustering jerárquico es el dendrograma. El dendrograma es un diagrama de árbol generado a partir del cálculo de las distancias entre cada par de elementos, de manera que se van formando conglomerados con ellos. Así, para proceder a este tipo de análisis, en primer lugar, se calcula la distancia entre cada par de elementos y luego se van fusionando con los más cercanos mediante alguna técnica de unión conocida como *linkage.* La estrategia de control irrevocable (greedy) regula que cada vez que se unen dos clusters no se reconsidera otra posible unión (Berzal, 2021)

Esta metodología requiere el uso de una medida de la distancia para establecer el grado de similitud entre los distintos elementos a agrupar. En este trabajo, se decidió emplear la distancia Euclídea, que indica la separación entre dos puntos en un espacio que satisface los axiomas y teoremas de la geometría de Euclides y corresponde con la distancia en línea recta entre dos objetos. Sean $X=(x_1,...,x_n)$ e $Y=(y_1,...,y_n)$

$$d_E(X,Y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2}$$

donde d es la distancia, x e y son las coordenadas de cada observación y n la observación a la que se hace referencia.

La distancia o similitud entre cada par de clases se representa en el eje Y del gráfico mientras que las observaciones se especifican en el eje X.

La mínima distancia entre los agrupamientos de datos se denomina *Single-linkage* mientras que cuando esta distancia es máxima se utiliza el término *Complete-linkage.* La *Average-linkage* es la distancia promedio entre dos grupos (Clements, 2019)

Existen dos enfoques dentro de la agrupación jerárquica. El divisivo y el aglomerativo. El primero engloba todos los puntos de los datos en un solo grupo, que se irá dividiendo en otros grupos más pequeños. El proceso acaba cuando cada uno de estos nuevos grupos contenga una sola muestra. En el tipo aglomerativo se da el proceso contrario. Se comienza con cada muestra como un grupo diferente y a partir de ahí, se van fusionando con las muestras más cercanas hasta terminar en un único grupo.

3.4.2 Multidimensional-Scaling

El *MultiDimensional Scaling* (MDS) es un procedimiento que permite analizar matrices de proximidad y que proporciona un diagrama visual, de dos dimensiones generalmente, de distancias entre conjuntos de objetos, de modo que aquellos objetos que presentan distancias menores son más similares. También se puede usar como un método de reducción de la dimensión. El primer modelo de escalamiento multidimensional se desarrolló en 1958 por Torgerson (Torgerson, 1958).

Existen dos subtipos de *MDS:* métrico y no métrico. El primero se utiliza para similitudes cuantitativas, mientras que el segundo se usa cuando las similitudes son cualitativas y las distancias están en una escala ordinal, de ahí que dicho subtipo también sea llamado MDS ordinal.

Es necesario el cálculo de la distancia entre los puntos, que al igual que para el dendrograma, se suele utilizar la euclídea. Dichas distancias se introducen en un matriz de distancias **D** de tamaño *n x n*.

Las distancias de la diagonal son todas igual a 0 ($d_{ii}=0$) y además se satisface que $d_{ij}= d_{ji}$ para todo i y j (simetría).

El objetivo del MDS es representar las distancias observadas mediante unas variables $y_1,\ldots,y_k$ donde k tiene que ser inferior a n.

A partir de una matriz (**D**) de distancias *n x n*, es decir, una matriz cuadrada se trata de obtener una matriz *n x k*, con los valores de las *k* variables en los *n* individuos, de tal manera que la distancia euclídea entre los elementos al medirlos entre dichas variables reproduzca aproximadamente la matriz de partida **D** (Borg & Groenen, 1997).

3.5 Análisis de supervivencia

El análisis de supervivencia es una técnica estadística que estudia el tiempo que transcurre hasta la aparición de un determinado evento de interés, como puede ser la muerte debida a una enfermedad o la recaída tras recibir un tratamiento para una enfermedad, etc.

El análisis de supervivencia aborda tres situaciones clásicas que resuelve la Inferencia Estadística:

* Estudio univariante, que es la descripción y resumen de los tiempos de vida, lo que permite sacar patrones para futuros pacientes.

*Estudio bivariante, que es la comparación de patrones de supervivencia entre dos grupos.

* Estudio multivariante, que se basa en la construcción de un modelo capaz de predecir el tiempo de vida de los pacientes en función de un conjunto de variables predictoras (Gomez & Cobo, 2004).

Una característica del análisis de supervivencia que hay que considerar es la censura, ya que implica limitaciones e imprecisiones de los datos. Los datos censurados corresponden a aquellos pacientes en los que no existe información acerca de la ocurrencia del evento. Esto puede darse bien, porque ha habido una pérdida de seguimiento, o bien, porque el estudio haya concluido sin que el evento se haya producido. De este modo, a pesar de que los datos censurados no dan información completa sobre la ocurrencia del evento, ofrecen una información parcial del mismo que bien interpretada, puede ser de gran utilidad. Al igual que la censura, otra causa de imprecisión de los datos es el truncamiento. Los truncamientos son entradas al estudio después del hecho que define el origen. Es decir, no se conoce el origen de todos los individuos, ya que cuando entran al estudio, la enfermedad ya habría empezado.

Uno de los puntos clave en el análisis de supervivencia es la estimación de la función de supervivencia. Dicha función se calcula para cada instante de tiempo t y mide la probabilidad de que un paciente sobreviva al instante t.

Para la representación gráfica de la función de supervivencia pueden utilizarse el estimador de Kaplan-Meier o el método actuarial (Gomez & Cobo, 2004). El segundo método presenta la desventaja

de que predefine los intervalos de tiempo y por tanto sus resultados dependen bastante de esa elección

3.5.1 Método Kaplan Meier

El método Kaplan Meier, propuesto en 1958, es una de las técnicas no paramétricas de análisis de supervivencia más utilizadas. Consiste en una representación gráfica mediante una función escalonada también llamada curva de supervivencia, en la que cada escalón corresponde a la ocurrencia del evento de interés en la población de estudio.

Este método se basa en dos suposiciones:

*el destino de los individuos que se retiran del estudio es similar al de los que se quedan, es decir, los individuos no censurados representan a los censurados

* la respuesta dependerá del periodo de tiempo en el que cada individuo entra al estudio.

De este modo, calcula la proporción acumulada de individuos que sobrevive para cada tiempo individual de supervivencia.

Con la siguiente fórmula se calcula la proporción de individuos que sobrevive a cada intervalo:

$$\frac{n-r}{n-r+1}$$

siendo n el tamaño de muestra y r las observaciones no censuradas ordenadas.

Para poder aplicar la fórmula se debe conocer el valor de r, para ello se ordenan los tiempos de supervivencia de menor a mayor (tanto los censurados como los no censurados) y se enumeran. El valor de r corresponde con la enumeración de los no censurados, por lo que son solo esos valores los que serán de interés.

Una vez aplicada la fórmula de los pacientes que sobreviven a cada intervalo, se multiplica el valor resultante de la primera censura por el de la siguiente, obteniendo así el primer valor de la proporción acumulada que sobrevive. Este valor se vuelve a multiplicar por el obtenido de la fórmula en la siguiente censura y así sucesivamente hasta tener todos los valores de la probabilidad acumulada.

Esta proporción nos indica la probabilidad de sobrevivir hasta un instante determinado, que es la probabilidad de sobrevivir hasta el periodo anterior (t-1) por la probabilidad de sobrevivir durante el intervalo (t-1,t)

Tabla 1.- Ejemplo del método Kaplan Meier extraído de (Pita Fernández, 2001).

Tiempo supervivencia	*Nº orden*	*Orden no censuradas*	$\frac{n-r}{n-r+1}$	*Proporción supervivencia*
3+	*1*	-	-	-
6	*2*	*2*	*4/5= 0.8*	*0.8*
7	*3*	*3*	*3/4= 0.75*	*0.8*0.75= 0.6*
7+	*4*	-	-	-

8	*5*	*5*	*1/2= 0.5*	*0.6*0.5= 0.3*
10	*6*	*6*	*0*	*0*

Los resultados obtenidos en la tabla anterior se representan en un gráfico, de forma que la proporción de supervivencia queda reflejada en el eje de ordenadas y la variable temporal expresada en semanas, meses,…etc. en el eje de abscisas. A este gráfico se le puede añadir la mediana de supervivencia que corresponde a una probabilidad de supervivencia de 0.5 y que al proyectarla sobre el eje de abscisas muestra gráficamente el tiempo al que sobreviven el 50% de los individuos

Cuando se dispone de dos o más grupos distintos, uno de los análisis más interesantes es la comparación de curvas de supervivencia para comprobar si existen diferencias significativas entre ellas.

Para realizar esta comparación, es común utilizar la prueba de Mantel-Cox o Log-Rank que es una modificación del test Chi-cuadrado de bondad de ajuste y se define como una prueba estadística no paramétrica basada en un contraste de hipótesis que trata de probar si hay diferencias entre dos curvas de supervivencia, obtenidas mediante el método de Kaplan Meier, comparando el número de eventos observados en cada grupo con el número esperado de los mismos (Carreño Serra, 2006)

La hipótesis nula es la de que no hay diferencias en la supervivencia de ambos grupos.

H_0: las curvas son iguales

H_1: las curvas no son iguales

Se utiliza una prueba Chi-cuadrado con las pérdidas observadas y esperadas.

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

siendo O_1 y O_2 el número total de pérdidas para los grupos 1 y 2 respectivamente y E_1 y E_2 el número total de pérdidas observadas en los grupos 1 y 2.

Este test utiliza unos grados de libertad de k-1. El valor de k corresponde con el número de grupos que se comparan, en este caso son dos grupos por comparación, por lo que se establece un grado de libertad. Los grados de libertad son el número de observaciones que pueden variar al estimar los parámetros estadísticos (Zaforas, 2017)

Si el valor obtenido de la prueba Chi-cuadrado es significativo, entonces hay diferencias entre las curvas de supervivencia de los diferentes grupos (Pita Fernández, 2001).

Sin embargo, esta prueba no propone un parámetro que cuantifique las diferencias entre los grupos, solo informa si dichas diferencias existen. Para la obtención de este parámetro se realiza la regresión de Cox.

El modelo de regresión de Cox es un método semiparamétrico que estima un parámetro denominado *Hazard ratio* considerado como la razón entre ambas funciones de riesgo. El *Hazard ratio* es el cociente entre dos funciones de riesgo, por ejemplo, entre los pacientes que reciben tratamiento y los que no. Es una representación de cuánto más probable es que ocurra el evento frente a que no ocurra en un grupo frente al otro grupo (Molina Arias, 2015). Para entender este concepto es necesario conocer previamente la definición de función de riesgo. La función de riesgo *h(t)* es aquella que mide la probabilidad de que, en un periodo de tiempo concreto, ocurra el evento que se estudia en los individuos a los que aún no les ha ocurrido (Llopis Pérez, 2013a)

La regresión de Cox trabaja con la función de riesgo. Es un instrumento que ayuda a detectar relaciones entre el riesgo de que ocurra el evento de interés y alguna de las variables independientes, detectando si la curva de supervivencia o la función de riesgo cambian para los diferentes valores de las variables independientes o explicativas (Llopis Pérez, 2013b)

$$h(t/x) = h_0(t)\, exp\,(X'B)$$

donde B es un vector de coeficientes de regresión desconocidos que parametrizan el modelo y h_0 es la tasa de riesgo de línea base.

3.6 Machine Learning

Los algoritmos de *machine learning* son procesos computacionales que usan unos datos de entrada para desempeñar una tarea sin estar programados para ello mezclando técnicas de distintas disciplinas como Inteligencia Artificial, Estadística, Informática, etc. Se adaptan continuamente mediante la repetición o experiencia para cada vez ofrecer unos mejores resultados en la tarea encomendada. Este proceso de adaptación es denominado entrenamiento. De este modo, *machine learning* se podría definir como un conjunto de algoritmos que entrenan a las computadoras para que aprendan de datos de ejemplo y puedan realizar tareas sin estar específicamente programadas para ello.

El aprendizaje también engloba la parte en la que el algoritmo se generaliza para que también pueda proporcionar el resultado deseado en datos nuevos, que no han sido usados antes, distintos de los de entrenamiento.

Este tipo de técnicas se han aplicado con éxito en diversos campos desde las finanzas hasta la biología computacional pasando por aplicaciones médicas y biomédicas.

En el ámbito de la oncología, la utilización de las técnicas de *machine learning* se ha extendido a casi todas las áreas del campo, modelando la respuesta tumoral, planificando el tratamiento, etc.

El aprendizaje automático tiene su origen en el siglo XVIII aunque en la historia moderna fue Arthur Samuel el que acuñó el término de "aprendizaje automático". En 1958 surgen las primeras arquitecturas de redes neuronales impulsadas por Rosenblatt y su desarrollo del perceptrón, aunque contaba con limitaciones que se fueron solventando en años posteriores con el desarrollo del perceptrón multicapa. Posteriormente se desarrollaron los árboles de decisión en 1986 por Quinlan y las máquinas de vector soporte (El Naqa & Murphy, 2015)

El *machine learning* se puede dividir en dos grandes áreas. Estas son:

- Aprendizaje supervisado: Parte de datos históricos ya etiquetados sobre los que se buscan patrones para realizar predicciones futuras. Es decir, cuenta con un aprendizaje previo. Algunos algoritmos utilizados en aprendizaje supervisado son Support Vector Machine *(SVM)*, Random Forest *(RF)*, K-Nearest Neighbors *(KNN)* o Partial Least Squares *(PLS)*. El aprendizaje supervisado puede dividirse, en función de la respuesta predicha, en problemas de regresión y problemas de clasificación. La diferencia entre ambos es que los problemas de regresión predicen un valor real o continuo, mientras que los de clasificación predicen la categoría a la que pertenece un objeto (Sancho Caparrini, 2017)
- Aprendizaje no supervisado: en este caso se usan datos que no están etiquetados, y se exploran para intentar encontrar estructuras o criterios de organización. Dentro del aprendizaje no supervisado se encuentran los métodos de clustering o el multidimensional-scaling, entre otros (González, 2021).

Podría considerarse otra área del *machine learning*, el aprendizaje semi-supervisado. Este incluye tanto datos etiquetados como datos sin etiquetar. La parte etiquetada se usa como aprendizaje para la parte no etiquetada.

Aunque el *machine learning* tiene múltiples aplicaciones, la más extendida es la minería de datos. Se usa con el objetivo de resolver los problemas a la hora de realizar análisis en los que se intenta establecer relaciones entre varias características, mejorando la eficiencia de los sistemas.

A partir del conjunto de datos original se crean dos matrices, una es la de entrenamiento con un k% de las muestras, y otra de validación con el (100-k)% restante de las muestras. La primera de ellas será la que entrenará al conjunto de datos para que se puedan hacer predicciones sobre la segunda de ellas (El Naqa & Murphy, 2015)

3.6.1 Support Vector Machine

Las Máquinas de Vector Soporte o *Support Vector Machine (SVM)* es una nueva técnica de *machine learning*, desarrollada inicialmente por Vladimir Vapnik, que asigna etiquetas a objetos a partir de ejemplos de los que ha aprendido. Se basa en cuatro conceptos que son el hiperplano separador, el hiperplano de margen máximo, el margen blando o *soft margin* y las funciones *kernel.*

El hiperplano separador es aquel que separa las muestras de las distintas clases. Existen un conjunto de hiperplanos que separan las dos clases, de ahí surge el concepto de hiperplano de margen máximo, que es aquel hiperplano separador con la máxima distancia a ambas clases. El margen se define como la distancia que existe entre el hiperplano y el vector de expresión más cercano.

El margen blando es aquel que permite que alguna observación quede en el lado contrario del hiperplano del que debería estar porque no pertenece a esa clase, sin afectar al resultado final. Esto requiere de unos parámetros que lo regulen para que no haya demasiadas clasificaciones erróneas en el modelo (Noble, 2006).

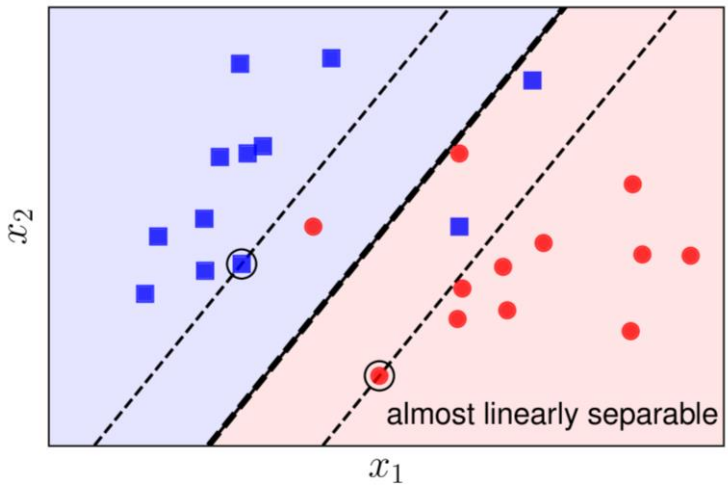


Figura 3.- Ejemplo de margen blando extraído de (Nguyen, 2017)

En el gráfico anterior se puede observar un ejemplo de cómo funciona el margen blando. Permite que alguna observación de la clase azul quede en el lado del hiperplano que no les corresponde. Y lo mismo para las observaciones rojas.

Las SVM son algoritmos que, dentro de un conjunto de entrenamiento en el que las observaciones están clasificadas en una de las dos categorías, construyen un modelo para que nuevas observaciones sean asignadas a esa o a la otra categoría (Vapnik, 1995). Este método ofrece una precisión mayor que cualquier otro algoritmo de clasificación.

El usuario debe realizar una validación cruzada (la validación cruzada sirve para comprobar que la calidad de la predicción del modelo en el conjunto de entrenamiento es buena y por tanto se puede usar para el conjunto de validación (Delgado, 2018) para determinar el ajuste óptimo de los parámetros ya que el rendimiento de este algoritmo es muy sensible a cómo se configuran los parámetros del coste y el kernel. El proceso de selección de parámetros es conocido como selección del modelo, y presenta un problema, que es el tiempo que tarda en realizarlo.

Nos encontramos con parámetros que están asociados con el uso de SVM que pueden afectar a los resultados. Dentro de estos parámetros están incluidos la elección de las funciones *kernel*, la desviación estándar del kernel gaussiano, etc (Srivastava & Bhambhu, 2005).

En este algoritmo es fundamental el concepto de hiperplano, ya que SVM busca un hiperplano con la máxima distancia entre los puntos de ambas categorías, quedando cada una de ellas a un lado de dicho hiperplano. Este hiperplano es un límite de decisión a la hora de clasificar nuevos datos. El vector que contiene los puntos más cercanos a dicho hiperplano es llamado *vector de soporte.*

Dicho hiperplano se dice que es afín, lo que significa que no tiene por qué pasar por el origen. La ecuación de la recta del hiperplano para dos dimensiones es

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

Pero puede generalizarse para p- dimensiones como:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = 0$$

Los puntos de $\mathbf{X}= (x_1, x_2, \ldots, x_p)$ que cumplen la ecuación pertenecen al hiperplano. Cuando no se satisface la igualdad anterior para alguno de los puntos de **X**, significa que dicho punto se encuentra a un lado o a otro del hiperplano, según el signo de la ecuación.

$y_k = 1$ si pertenece a la clase A

$y_k = -1$ si pertenece a la clase B

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p > 0$ si $y_i=1$

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p < 0$ si $y_i = -1$

El clasificador óptimo, llamado *maximal margin hyperplane (hiperplano óptimo de separación)* es el que está más alejado de todas las observaciones de entrenamiento. Se obtiene mediante el cálculo de la distancia en perpendicular de cada observación al hiperplano. La distancia más pequeña es la que indica cómo de alejado se encuentra el hiperplano de los datos, y recibe el nombre de margen. El hiperplano óptimo de separación es el que obtiene un mayor margen.

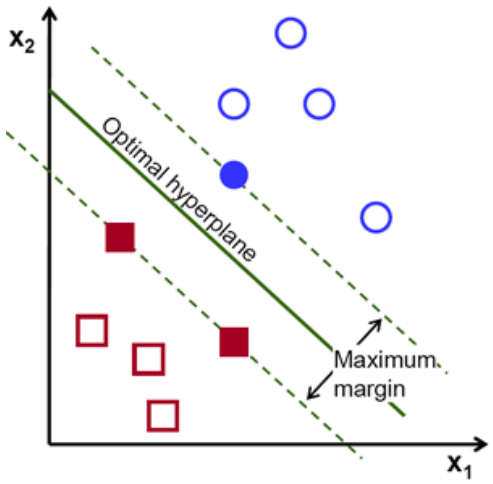


Figura 4.- Hiperplano óptimo. Extraído de (Zaforas, 2017)

En R, los algoritmos necesarios para utilizar este método están contenidos en las librerías *e1071* y *LiblinearR*. La librería que se usará en este trabajo es *e1701*.

Los SVM se pueden clasificar en casos perfectamente separables linealmente, casos cuasi-separables linealmente y casos no separables linealmente.

El paquete *e1071* utiliza la técnica de casos cuasi-separables.

Para los ejemplos cuasi-separables no existe una forma lineal que separe los datos, y por tanto, no se puede obtener un *maximal margin hyperplane.* Se usan los hiperplanos de separación de margen blando *(soft margin classifiers)* que permiten que alguna observación esté en el lado incorrecto del margen o del hiperplano sin que afecte al resultado final. Como es evidente, el SVM no se implementa para que permita muchas calificaciones erróneas, de modo que el usuario tiene que especificar unos parámetros que controlen cuantas observaciones incumplan el hiperplano de separación y hasta qué distancia pueden llegar. El *soft margin* realiza una compensación entre la violación del hiperplano y la extensión del margen (Noble, 2006).

Cuando el problema no es lineal, es decir, el conjunto de observaciones no se puede separar con un hiperplano, el método *support vector classifier* tiene limitaciones y en estos casos se utiliza el método de las máquinas de vector soporte con aumento de la dimensión mediante las funciones *Kernel* que son funciones no lineales que transforman el problema a un problema lineal, y una vez obtenida la solución se devuelve al espacio original. Fundamentalmente, añaden una nueva dimensión elevando al cuadrado los valores de expresión originales, de modo que realiza una clasificación (n+1)-dimensional de unos datos originalmente n-dimensionales (Noble, 2006).

Antes de realizar el análisis hay que determinar el mejor kernel (Amat Rodrigo, 2017).Las funciones kernel más populares son lineal, polinómica, sigmoide y RBF (*Radial basis Function)*. Esta última es la más comúnmente usada ya que presenta menos dificultades numéricas, tiene menos hiperparámetros que la polinómica y además mapea muestras no lineales a un espacio dimensional.

3.6.2 K-nearest neighbors

El algoritmo de los K-vecinos más cercanos *(KNN)*, surge por primera vez en 1951 en un informe de la Facultad de Medicina de Aviación de la fuerza aérea de EEUU en el que Fix y Hodges introdujeron un método no paramétrico para clasificar patrones, que recibió el nombre de K-nearest neighbor.

Este método clasifica un objeto según la similitud que tiene con los objetos vecinos, de los que sí se conoce la clase (javatpoint, 2021). El método de los K-vecinos más cercanos (*KNN)* es un método no paramétrico y es uno de los algoritmos de clasificación más sencillos. Surge de la necesidad de realizar análisis discriminantes cuando las estimaciones paramétricas de densidades de probabilidad no se conocen o son difíciles de determinar.

Se basa en la idea de que un nuevo objeto se clasifica en la clase más frecuente a la que pertenecen sus vecinos más cercanos.

Una representación gráfica de los K vecinos más cercanos es el diagrama de Voronoi.

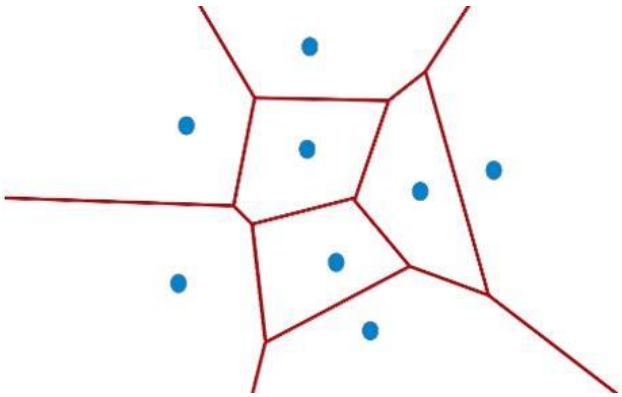


Figura 5.- Diagrama de Voronoi. Extraído de (Toledano Díaz, 2019)

Definir el valor de k es un aspecto fundamental del algoritmo. Se ha demostrado que no hay un número óptimo de vecinos que se adapte a cualquier conjunto de datos, es decir, no existe un valor predeterminado para k que se pueda usar en cualquier caso. Por tanto, se tendrá que definir para cada conjunto según los requisitos de este. Un valor pequeño de k tendrá bajo sesgo, pero alta varianza, mientras que cuando se dispone de un alto número de vecinos la situación será la inversa, es decir, sesgo más alto y varianza más baja. Se suele elegir un valor de k impar para evitar empates.

A la hora de ejecución de este algoritmo, en primer lugar, se dividen los datos en muestras de validación o prueba y muestras de entrenamiento. Sea ω la verdadera clase de una muestra de entrenamiento, y $\widehat{\omega}$ la clase predicha para la muestra de prueba. (ω, $\widehat{\omega}$=1,2...Ω) donde Ω es el número total de clases. Durante el entrenamiento se usa la clase verdadera ω para entrenar al clasificador y en las pruebas se predicen las clases $\widehat{\omega}$ de las muestras de prueba.

El método KNN se considera un método de clasificación supervisado, ya que utiliza las etiquetas de clase de los datos de entrenamiento.

En ocasiones, para obtener un mayor rendimiento del clasificador, se transforman antes los valores utilizando métodos como la estandarización.

La estandarización elimina los efectos causados por las distintas escalas de medida, ya que las variables pueden estar en unidades diferentes. Transforma los distintos valores de la matriz en valores Z utilizando la media y la desviación estándar mediante la siguiente relación:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Donde x_{ij} es el valor de la muestra i para la característica j, μ_j es la media de todo x_{ij} y σ_j es la desviación estándar de todo x_{ij}. Si los valores siguen una distribución gaussiana, el histograma de los valores Z representará una distribución Normal con media 0 y desviación típica 1.

Las predicciones de clase no se realizan en muestras que hayan sido utilizadas para el entrenamiento del clasificador ya que la precisión quedaría sesgada. Es por eso que se realizan en muestras que quedaron fuera del proceso de aprendizaje.

La validación cruzada o *cross-validation* determina la precisión de la clasificación, evaluando el rendimiento de los clasificadores (Peterson, 2009).

Para calcular la distancia con los datos más próximos el algoritmo *KNN* se suele basar en la distancia euclídea entre las muestras de prueba y las muestras de entrenamiento. Se obtiene con la siguiente fórmula. Se define como la distancia recta más corta entre dos puntos (P,Q), donde p_i son las coordenadas de P y q_i son las coordenadas de Q.

$$d_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Si bien es cierto que es la más común, no es la única usada en el algoritmo KNN. Existen otras métricas, como por ejemplo:

La distancia de Chebyshev entre dos puntos corresponde a la mayor de sus diferencias.

$$D_{Chev}(p,q) = max(|p_i - q_i|)$$

Siendo p_i las coordenadas del punto p y q_i las coordenadas del punto q

La distancia Manhattan se calcula como la suma de las diferencias absolutas de las coordenadas de cada uno de sus puntos.

$$D_M(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

Donde p_i corresponde a las coordenadas de p y q_i corresponde a las coordenadas de q.

3.6.3 Random Forest

Los bosques aleatorios *(Random Forest)* fueron introducidos por Leo Breiman (Breiman, 2001) inspirado en trabajos anteriores de Amit y German (Amit & Geman, 1997). Se pueden usar tanto con variables categóricas como con variables continuas. En el primer caso el problema se denomina de clasificación mientras que en el segundo se denomina de regresión, es decir, puede ser multiclase, lo que es una ventaja frente a otros algoritmos de *machine learning*. Entre otras ventajas de esta técnica podemos encontrar que es relativamente rápida tanto para entrenarla como para predecir, solo depende de uno o dos parámetros de ajuste y que se puede usar en problemas de gran dimensión (Cutler, Cutler, & Stevens, 2012)

Es un método basado en la creación y combinación de multitud de árboles de decisión que devuelven una predicción de clase, siendo la predicción final del modelo la de la clase que más árboles hayan predicho (Yiu, 2019)

Random Forest es un algoritmo de aprendizaje supervisado, es decir, utiliza datos etiquetados para aprender a clasificar datos no etiquetados. Esta técnica incluye una combinación de predictores de árbol, cada uno de ellos vota a una clase, siendo la clase elegida la votada por más árboles de decisión.

Los árboles de decisión están compuestos por varios nodos en los que se van tomando decisiones de clasificación de individuos en función de variables predictoras. Parten de un primer nodo que se va ramificando sucesivamente hasta llegar a las hojas. Cada hoja corresponde a una clase, por lo tanto, la hoja elegida es la clase que ofrecerá este árbol. Después se fusionan las decisiones de los distintos árboles para encontrar una respuesta.

En los problemas de clasificación se usa el índice Gini, que determina cuál de las ramas del árbol es más probable que ocurra. Dicho índice viene dado por la siguiente fórmula:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

C representa el número de clases y p_i representa la frecuencia relativa de la clase que se está observando en el conjunto de datos.

La entropía también puede determinar cómo se ramifican los nodos de un árbol de decisión usando la probabilidad de un resultado concreto para tomar una decisión sobre la forma de ramificación del nodo (Schott, 2019)

$$Entropy = \sum_{i=1}^{C} p_i * log_2(p_i)$$

Esta técnica se lleva a cabo con la librería *caret (classification and regression training)* del software R que incluye distintas funciones que facilitan el uso de diversos métodos de clasificación y regresión.

3.6.4 Evaluación del rendimiento

El rendimiento de cada algoritmo puede ser evaluado mediante distintos parámetros estadísticos, como sensibilidad, especificidad, precisión o exactitud, entre otros.

La sensibilidad es la capacidad de clasificar verdaderos positivos, es decir, la capacidad de detectar el suceso en los sujetos en los que ha ocurrido, mientras que la especificidad se define como la capacidad de clasificar correctamente los casos negativos, o, dicho de otra forma, la capacidad de detectar la ausencia del suceso en los pacientes en los que no se ha dado.

$$\text{Sensibilidad} = \frac{VP}{VP+FN}$$

donde VP son los verdaderos positivos y FN los falsos negativos.

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN+FP}}$$

siendo VN los verdaderos negativos y FP los falsos positivos.

El valor predictivo positivo (o precisión) se corresponde con la proporción de pacientes con resultado positivo en la prueba bien clasificados, mientras que el valor predictivo negativo (VPN) es la proporción de pacientes con resultado negativo en la prueba que están bien clasificados, es decir, la probabilidad de ser negativo si la prueba ha resultado negativa. Ambos parámetros se pueden representar con las siguientes fórmulas

$$\text{Precisión} = \frac{\text{VP}}{\text{VP+FP}}$$

$$\text{VPN} = \frac{\text{VN}}{\text{VN+FN}}$$

3.7 Análisis de sobrerrepresentación: Webgestalt

WebGestalt es una herramienta web de análisis de enriquecimiento funcional que incorpora tres métodos complementarios, el análisis de sobrerrepresentación (ORA), el análisis de enriquecimiento de conjuntos de genes (GSEA) y el análisis basado en topología de red (NTA). Utiliza como fuentes bases de datos como KEGG (*Kyoto Encyclopedia of Genes and Genomes*) y Reactome, que son dos bases de datos de rutas biológicas que describen la funcionalidad de los genes. Se diferencian en que la primera fue iniciada por el programa del genoma humano japonés mientras que la segunda es americana.

KEGG incluye información sobre genes, proteínas, rutas metabólicas, interacciones moleculares...etc

Para representar los resultados obtenidos en este análisis se suelen utilizar gráficos de doble eje, donde se muestra por un lado el tamaño *(Count)* es decir, el número de genes que están incluidos en esa vía y por otro lado el -$\log_{10}$(fdr), que es el indicador de la significación estadística de la vía. Cuando este valor es menor que 0.05, la vía es significativamente sobrerrepresentada en genes del conjunto de datos estudiado.

Esta herramienta también permite el estudio de listas de genes agrupadas en función de sus características biológicas, y representadas por los términos *Biological process*, *cellular component* y *molecular function* de la base de datos *geneontology*. Geneontology es la mayor fuente de información sobre las funciones de los genes (Geneontology, 2021).

El análisis que lleva a cabo WebGestalt se basa en la distribución hipergeométrica que es una distribución de probabilidad discreta en la que el muestreo es sin reposición y por tanto las probabilidades no se mantienen constante(Liao et al., 2019).

La función de probabilidad del test hipergeométrico es la siguiente:

$$P[X = x] = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Capítulo 4.- Resultados y discusión

Se realizó una primera estratificación manual de los pacientes de la serie CoMMpass en función de los datos de traslocaciones obtenidos por secuenciación masiva de DNA generados a través del programa Delly. En el presente trabajo se llevó a cabo el estudio de las 4 translocaciones más frecuentes en MM: t(4;14), t(11;14), t(14;16) y t(14;20). Para comprobar la presencia de cada translocación se buscó específicamente que los *partners* determinados por Delly coincidiesen con las posiciones genómicas descritas en la literatura, la cual, fue obtenida de la base de datos *genecards.*

En los cuatro casos, el *partner* del cromosoma 14 se encuentra en el locus IGH 14q32, en la posición *chr14:106,032,614 - 107,288,051 (GRCh37/hg19).*Así, de acuerdo con esta aproximación, 176 pacientes presentaron una traslocación t(4;14), de los que 132 coincidieron con los *partners* previamente descritos para MM. Este proceso fue repetido para el resto de las traslocaciones.

En el caso de la deleción y mutación de *TP53* el proceso fue ligeramente diferente. Para seleccionar las muestras que presentan mutación en *TP53* se utilizó el archivo mutaciones no sinónimas provisto por CoMMpass. A continuación se buscó en el *Genome Browser* de la *UCSC* la posición del gen *TP53* y se seleccionaron las muestras del archivo de mutaciones que tienen la posición dentro del siguiente intervalo genómico: *chr17:7,571,720 - 7,590,868 (-200,+200).* De este modo, se detectaron 84 mutaciones, algunas de ellas en el mismo paciente, es decir, un mismo sujeto presentó varias mutaciones en *TP53*. En todos los casos se empleó la versión del genoma humano hg19.

En una segunda aproximación se recogieron los datos Seq-FISH proporcionados directamente por la base CoMMpass. La seq-FISH consiste en una estimación y aproximación de los datos de hibridación fluorescente *in situ* (FISH) a partir de datos de secuenciación. Los resultados obtenidos por nuestra aproximación manual, y la provista por la base se comparan en la tabla 2.

Tabla 2. Comparación resultados método manual y SeqFISH.

	método manual	Seq-FISH
4;14	67	83
11;14	101	124
14;16	23	23
14;20	10	10

Se puede ver como para las dos últimas translocaciones se obtienen los mismos resultados por los dos métodos pero en el caso de la t(4;14) y la t(11;14) no todas las muestras que presentan la translocación son captadas por el método manual.

Para continuar con el análisis posterior se decidió trabajar con los datos provistos por la SeqFISH para evitar discrepancias con otros estudios previos (A Comparison of Clinical FISH and Sequencing Based FISH Estimates in Multiple Myeloma: An Mmrf Commpass Analysis, Chase Miller, BSc, Jennifer Yesil, MS, Mary Derome, MS, Andrea Donnelly, Jean Marrian, Kyle McBride, MS, Mmrf CoMMpass Network, Daniel Auclair, PhD, Jonathan J Keats, PhD Blood (2016) 128 (22): 374.). Así, un resumen de las características genómicas de la cohorte de pacientes de la serie CoMMpass en función del Seq-FISH se recoge en la siguiente tabla (tabla 3), donde se definen los grupos para cada

translocación con el número de muestras que presentan solo la translocación sin otros eventos, las que presentan la translocación y mutación en tp53...etc.

Tabla 3. Grupos de pacientes

	solo Tx	Tx + mut *TP53*	Tx + del *TP53*	Trx + del + mut *TP53*	sin eventos (ni Tx ni del ni mut *TP53*)	resto
4;14	72	2	6	3	469	61
11;14	108	3	5	8	439	50
14;16	20	2	2	1	508	80
14;20	9	0	1	0	518	85

Tx: traslocación.

A la vista de estos datos, se decidió descartar la translocación t(14;20) para el análisis ya que los grupos en los que dicha translocación se asoció con la mutación o deleción de *TP53* presentaron muy pocas muestras (menos de 2).

Una vez definidos los grupos se calculó el tiempo de supervivencia libre de la enfermedad tanto para los pacientes que presentaron *TP53*, como para aquellos que presentaron las distintas traslocaciones seleccionadas. Para ello se realizó un análisis de supervivencia diferenciando en un gráfico de Kaplan-Meier las curvas de supervivencia para los pacientes que tienen mutación y/o deleción en *TP53* y los que no tienen *TP53*. Se utilizó la variable "progresión de la enfermedad" codificada con 0 y 1, correspondiendo el cero a la no progresión y el uno a la progresión.

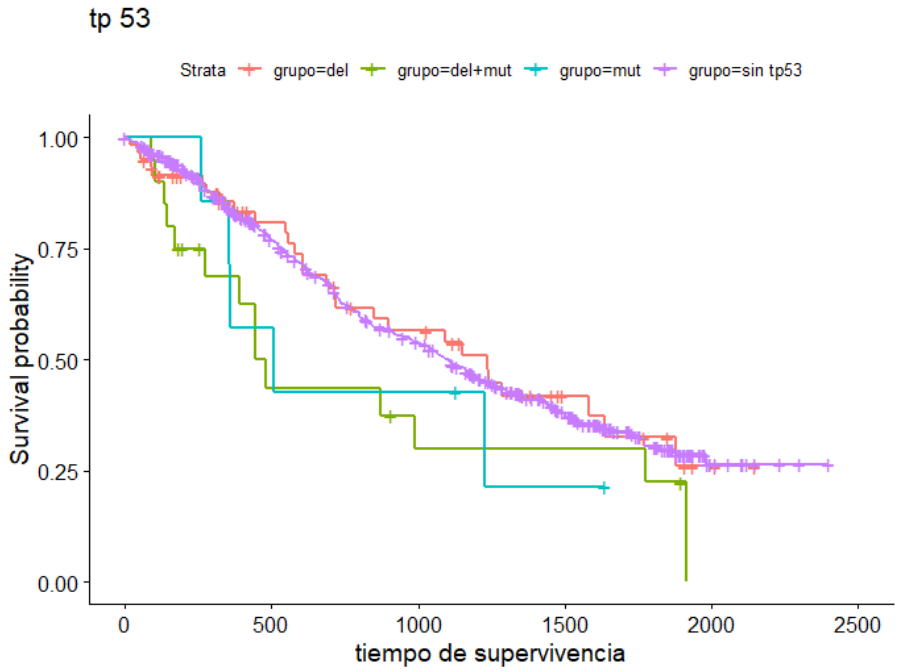


Figura 6. Gráfico de Kaplan Meier para la alteración de TP53

En la figura anterior se observa como los grupos con mutación no sinónima en tp53 presentan un peor pronóstico que el resto, ya que las curvas para los pacientes sin tp53 y los que presentan únicamente deleción son similares y con una supervivencia libre de progresión relativamente más larga.

En lo relativo a la traslocación t(4;14), el resultado que se obtuvo en el análisis de supervivencia se recoge en la siguiente figura:

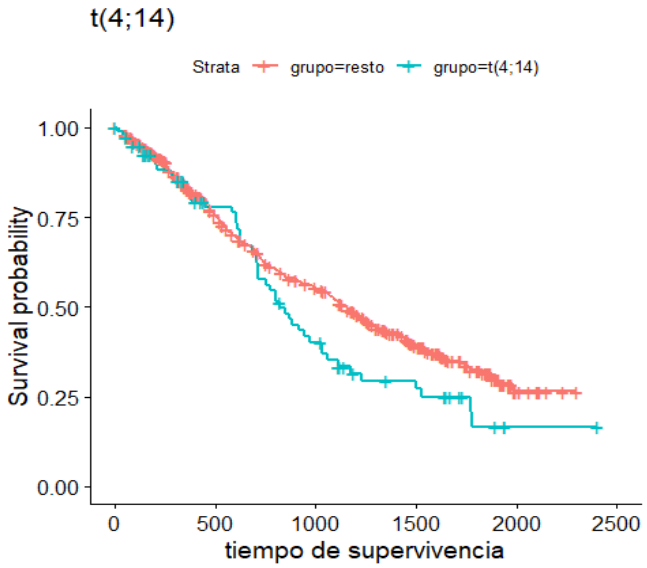


Figura 7. Gráfico de Kaplan Meier para t(4;14)

Es significativo que la presencia de traslocación t(4;14) está ligada a una supervivencia menor en los pacientes con MM, no obstante, este es un hecho ampliamente conocido ya que los pacientes diagnosticados con esta alteración son considerados de alto riesgo. Esto no ocurre con la presencia de otras traslocaciones como puede ser la t(11;14), que como se puede observar en la figura 8 su presencia no implica una menor supervivencia de los pacientes.

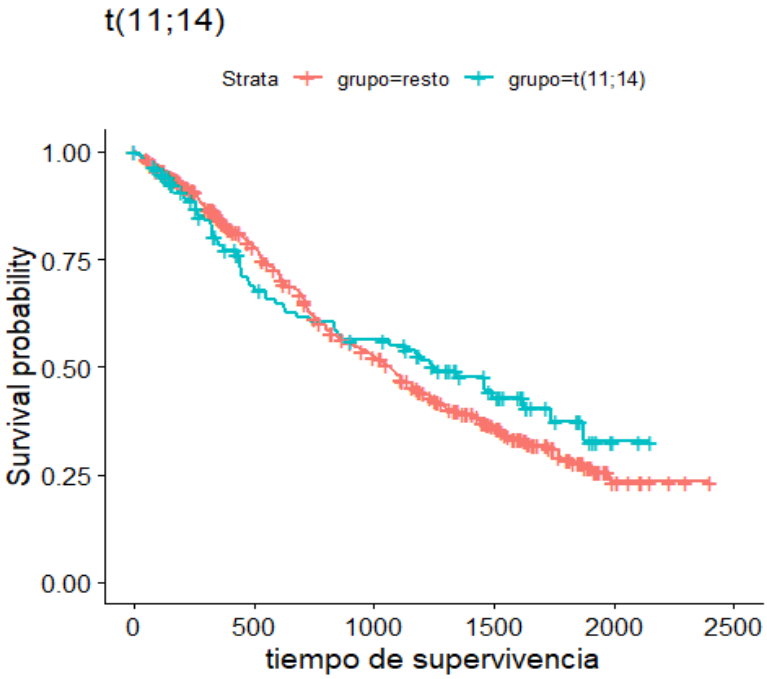


Figura 8. Gráfico de Kaplan Meier para t(11;14)

Finalmente, en el caso de la t(14;16) también existe constancia de que su presencia es de muy mal pronóstico, hecho que se confirma también sobre nuestra serie, donde los pacientes con esta traslocación presentan una supervivencia notablemente inferior a los que no la presentan (Figura 9).

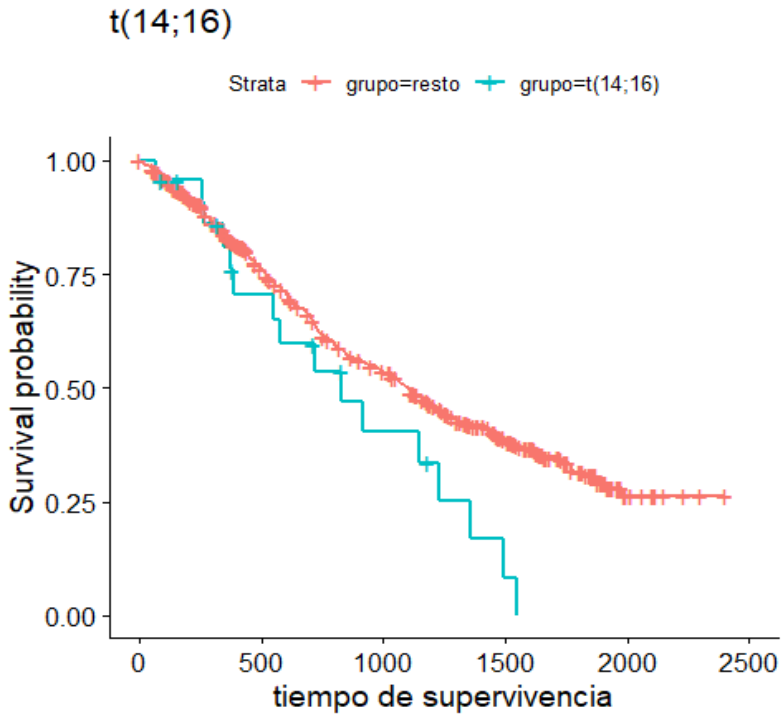


Figura 9. Gráfico de Kaplan Meier para t(14;16)

Una vez estudiados los grupos pronóstico de las translocaciones y de la alteración de *TP53* , se estableció la firma génica para los diferentes eventos citogenéticos analizados, realizando 4 contrastes (uno para cada una de las tres translocaciones y otro para la alteración de tp53). En cada uno de estos contrastes se comparan dos grupos de pacientes, el grupo de interés o grupo problema, formado por los pacientes que presentaron la translocación o alteración en tp53, pero no presentaron ningún otro evento; y el grupo control, compuesto por los pacientes que no presentaron ninguno de los eventos analizados. Este último lo compusieron 318 pacientes en todos los contrastes. El grupo problema por su parte, presentó variabilidad en cuanto a su número de componentes en función del evento citogenético contrastado. En la siguiente tabla se resume cuántos pacientes lo forman en cada caso. Por ejemplo, para la translocación 4;14, 72 pacientes presentaron dicha translocación y ningún otro evento.

Tabla 4. Pacientes de cada alteración.

Tipo de evento	Pacientes
4;14	72
11;14	108
14;16	20
deleción + mutación en tp53	20

Previo al estudio de la expresión génica diferencial se realizó un análisis no supervisado mediante multidimensional-scaling para cada contraste para ver la distribución de los pacientes según presenten o no las diferentes alteraciones citogenéticas. Para llevar a cabo este análisis, es necesario transformar a logaritmo en base 2 los datos de contajes normalizados. Además, se precisa de la adición de una pseudocuenta al conjunto de datos completo con el fin de evitar problemas en la transformación, al presentar algunos genes cero lecturas en algunas muestras. Este problema puede resolverse en Excel con la siguiente fórmula: `=log(valor +1 ; 2)`

El multidimensional-scaling se llevó a cabo con el programa informático SIMFIT. Para ello, se introdujo en dicho programa una matriz con la expresión de los genes en las muestras de cada comparación (se excluyen los genes que presenten menos de una lectura en las muestras). La matriz tiene que tener extensión "*.sim*".

Este primer multidimensional scaling proporciona las coordenadas de cada muestra, que hay que guardar y en Excel, asociar cada una de ellas a la muestra correspondiente y al grupo al que pertenecen. Este nuevo conjunto de datos vuelve a ser introducido en SIMFIT en formato MANOVA, haciendo distinción por colores según el grupo.

Para la translocación t(4;14) el gráfico resultante es el siguiente:

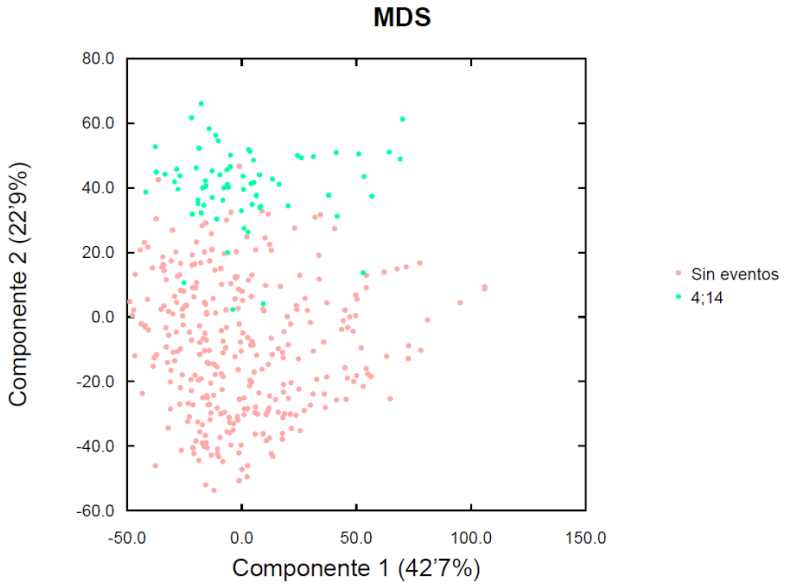


Figura 10. Multidimensional-Scaling para t(4;14)

En los ejes se indica la proporción de varianza que captura cada componente. Se observa una tendencia a la separación sobre el eje de ordenadas o segunda componente de los grupos sin eventos y con la t(4;14), lo que podría indicar que ambos grupos presentan un alto número de diferencias a nivel de expresión génica considerando el conjunto total de genes estudiados.

De manera análoga se llevó a cabo el gráfico del Multidimensional-Scaling para la traslocación t(11;14) donde se aprecia una clara diferenciación entre el grupo que presenta la traslocación t(11;14) y los pacientes sin eventos (Figura 11).

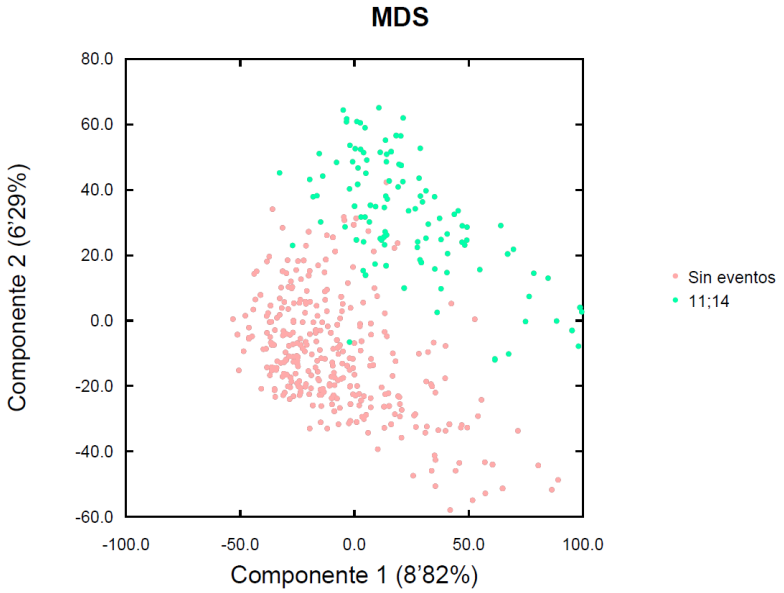


Figura 11. Multidimensional-Scaling para t(11;14)

Finalmente, se muestra el gráfico para el contraste de los pacientes que muestran la translocación t(14;16) frente a aquellos que no presentan ningún evento:

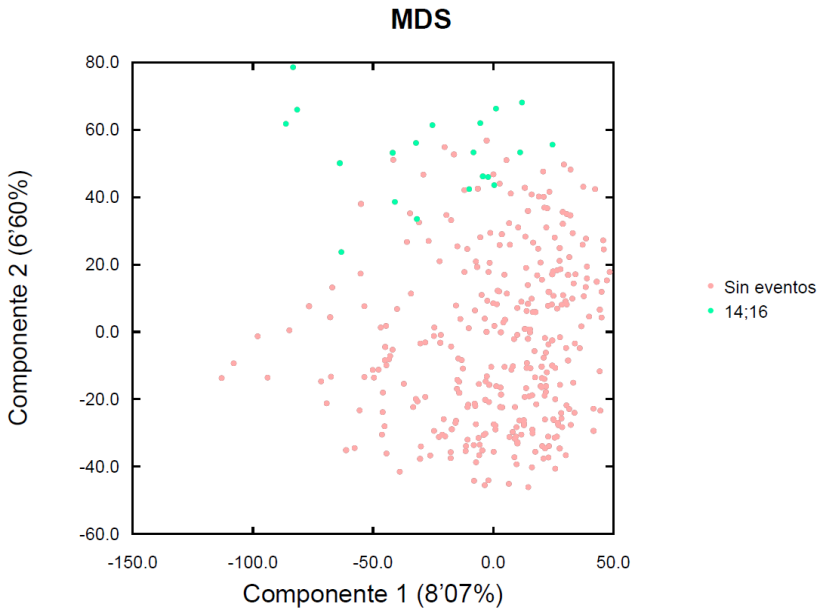


Figura 12. Multidimensional-Scaling para t(14;16)

Para las tres traslocaciones, se puede ver cómo las muestras del grupo que las presentan tienen un comportamiento similar, sin embargo, en el caso de las alteraciones de *TP53* se detectó un patrón diferente tal y como se recoge en la siguiente figura:

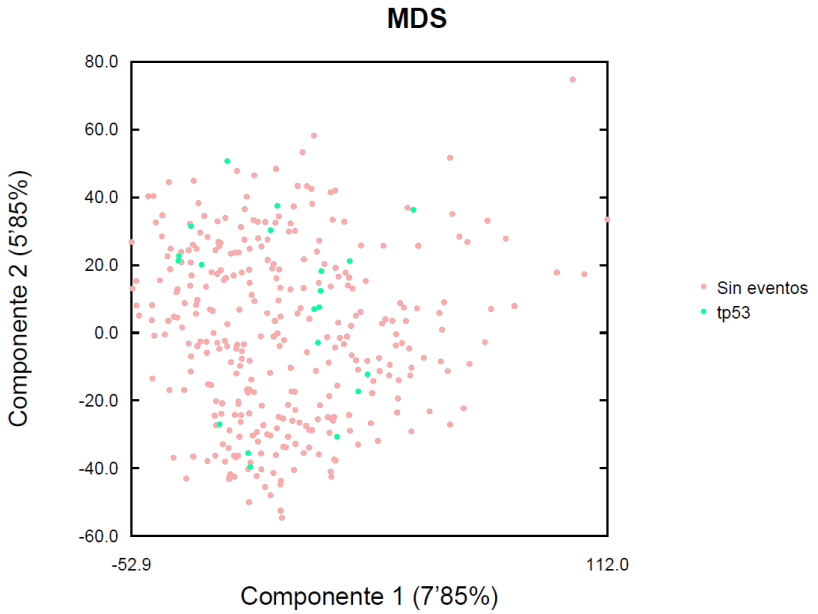


Figura 13. Multidimensional-Scaling para TP53

Así, en el caso de tp53 (deleción y mutación) no hay una diferenciación tan clara en el MDS no supervisado, es decir, utilizando todos los genes de la base de datos. Esto puede deberse a que los pacientes se comportan de manera similar presenten o no presenten TP53, ya que no hay diferencias en la expresión de los genes.

Determinación de las firmas de expresión génica

Tras el estudio de la estructura de los datos se realizó la búsqueda de la firma de expresión génica para cada uno de los grupos. Al realizar la expresión diferencial se seleccionan los genes con un FDR < 0.05 de cada contraste, para finalmente cruzar estas listas de genes significativos mediante un diagrama de Venn y así comprobar cuáles son exclusivos de cada comparación. El diagrama de Venn se realizó con la herramienta online de la Universidad de Gent, disponible en http://bioinformatics.psb.ugent.be/webtools/Venn/.

En la siguiente tabla se muestran los genes significativos para cada contraste, divididos en función del sentido de la expresión.

Tabla 5. Resultados de la expresión diferencial para cada constraste.

Firma	Genes significativos (FDR < 0.05)	Sobreexpresados	Infraexpresados
t(4;14)	8740	4874	3866
t(11;14)	10840	5529	5311
t(14;16)	5888	2309	3579
Alteración *TP53*	1720	374	1346

A continuación se muestra la expresión de los dos genes sobreexpresados (los que presentan un Fold Change > 1) con menor FDR, y con los dos genes infraexpresados (que tienen Fold Change < 1) más significativos para cada comparación, mediante diagramas de caja y bigotes *o Boxplot*.

***TP53* alterado (mutación y deleción) *vs.* Pacientes sin eventos:**

Los dos genes con mayor infraexpresión fueron ENSG00000108523 (*RNF167*) y ENSG00000129245 (*FXR2*), mientras que los dos genes mayor sobreexpresión fueron son ENSG00000146670 (*CDCA5*) y ENSG00000175063 (*UBE2C*).

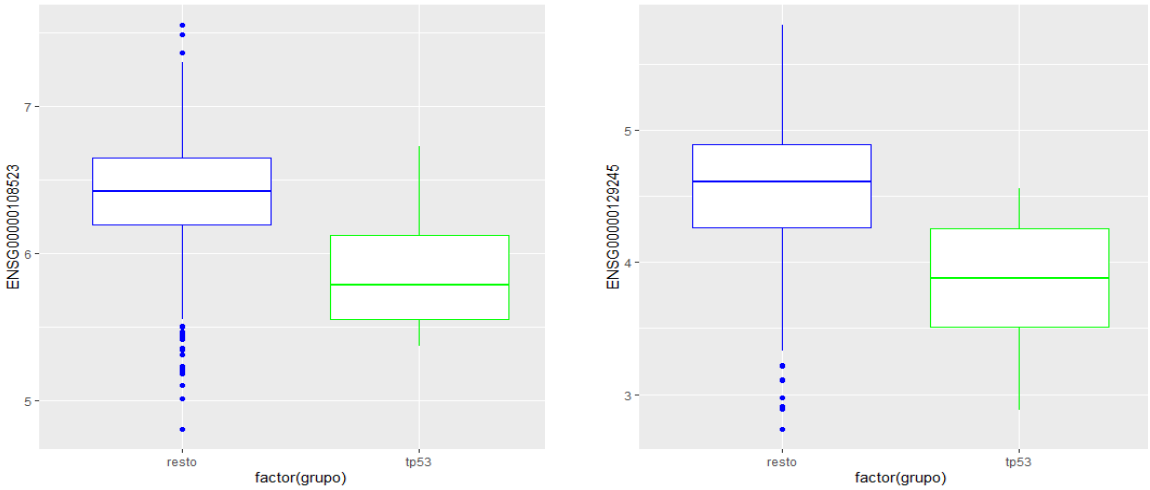


Figura 14. Boxplot para los dos genes más infraexpresados.

Los gráficos anteriores confirman la infraexpresión de estos dos genes en el grupo de pacientes que presentan *TP53* alterado. Curiosamente, la primera fusión génica de *TP53* que se ha descrito en la literatura ha sido con el gen ENSG00000129245 (*FXR2*) (Kanezaki, Toki, Xu, Narayanan, & Ito, 2006), con lo que los datos obtenidos presentan alta concordancia con que la pérdida de función de *TP53* lleve a una infraexpresión de *FXR2*.

Por su parte, el *boxplot* para los dos genes que más se sobreexpresan se recoge en la Figura 15.

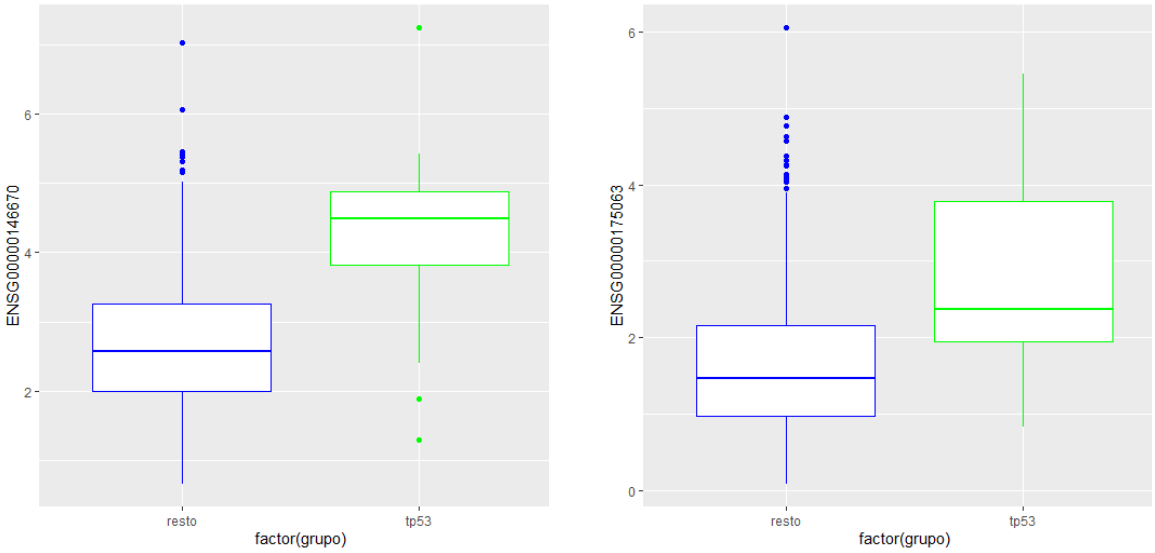


Figura 15. Boxplot para los dos genes más sobreexpresados.

En este caso, se confirma gráficamente que los dos genes tienen una mayor expresión en el grupo tp53 que en el resto de los pacientes. En el caso del gen ENSG00000175063 (*UBE2C*) podemos confirmar a través de una búsqueda en la literatura la regulación por parte de *TP53* mutado de este gen, ya que Bajaj et al en su artículo de 2016 (Bajaj et al., 2016) pudieron comprobar que la mutación de TP53 causa un incremento de este gen, y puede ser la clave en la regulación del punto de control del montaje del huso acromático que realiza *TP53* en la célula.

t(4;14) *vs.* Pacientes sin eventos:

Se repite el proceso que se siguió con tp53 para elegir los 4 genes que van a ser representados. Los dos con mayor sobreexpresión son ENSG00000235034 *(BTG3P1)* y ENSG00000166446 *(CDYL2)* y sus diagramas de cajas se ven representados en la figura 16.

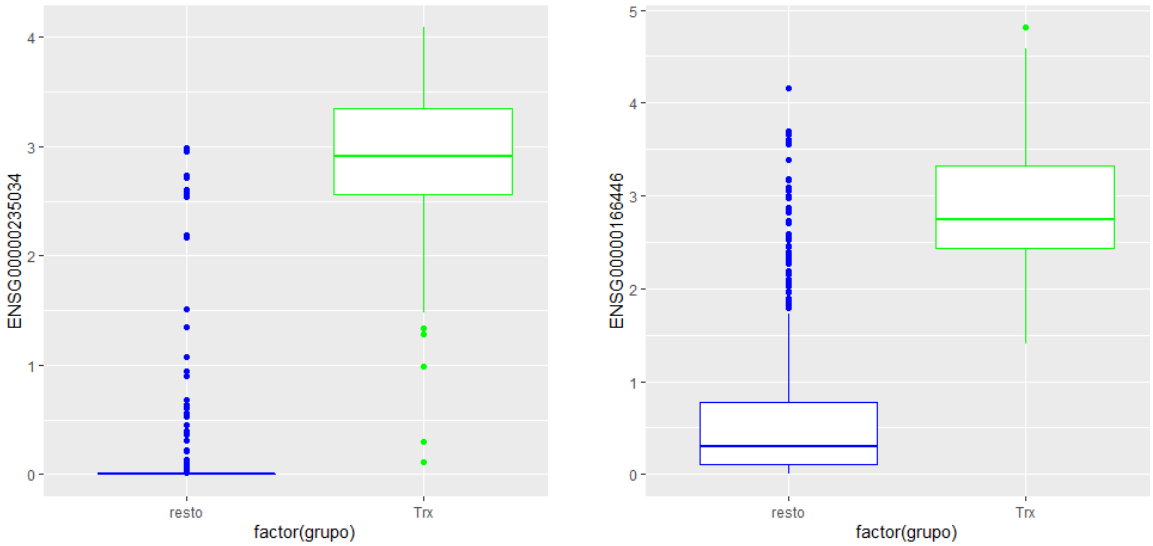


Figura 16. Boxplot para los dos genes que más se sobreexpresan.

Los dos genes que más se infraexpresan están representados en la figura 17 y son ENSG00000165802 *(NSMF)* y ENSG00000197355 *(UAP1L1)*

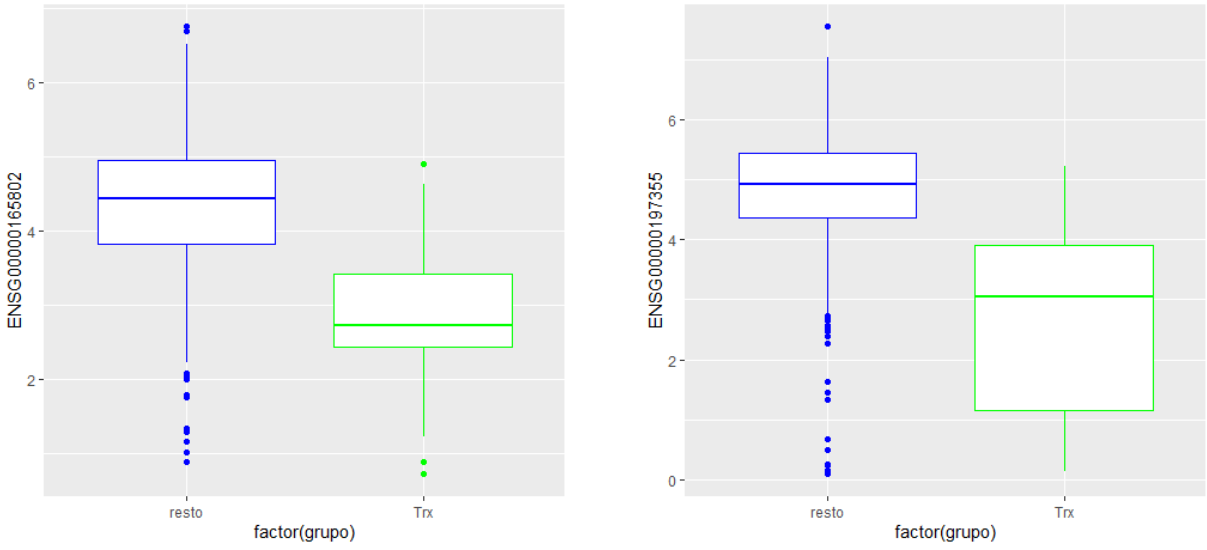


Figura 17. Boxplot para los dos genes infraexpresados.

t(11;14) *vs.* Pacientes sin eventos:

Los dos genes más sobreexpresados para este contraste son ENSG00000197119 *(SLC25A29)* y ENSG00000185100 *(ADSS1)*. Sus diagramas de caja están representados en la figura 18.

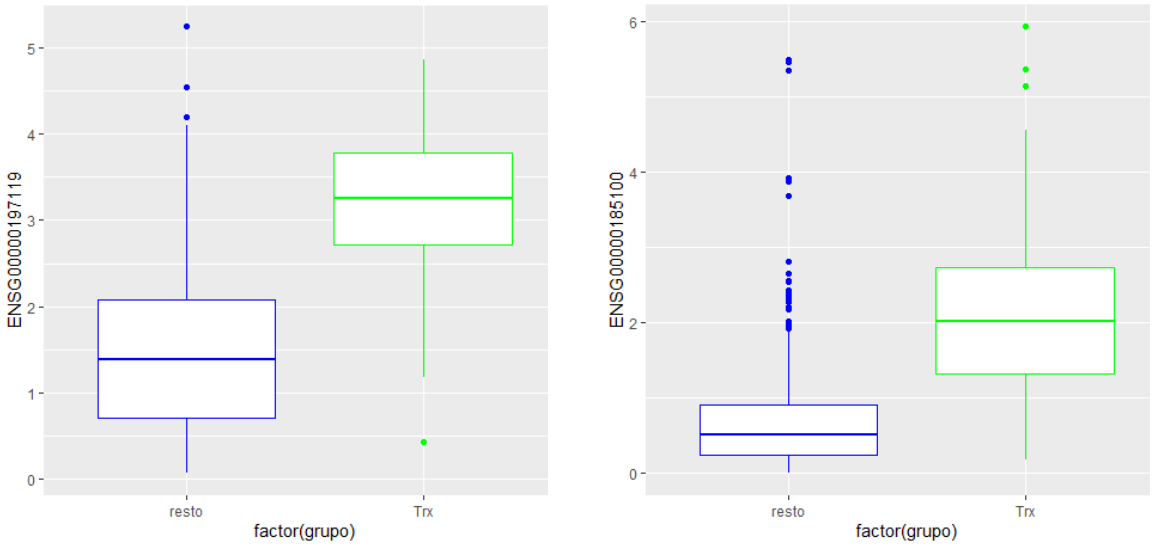


Figura 18. Boxplot para los genes sobreexpresados.

A continuación se muestran los dos genes que más se infraexpresan. Estos son ENSG00000135916 *(ITM2C)* y ENSG00000128626 *(MRPS12)*

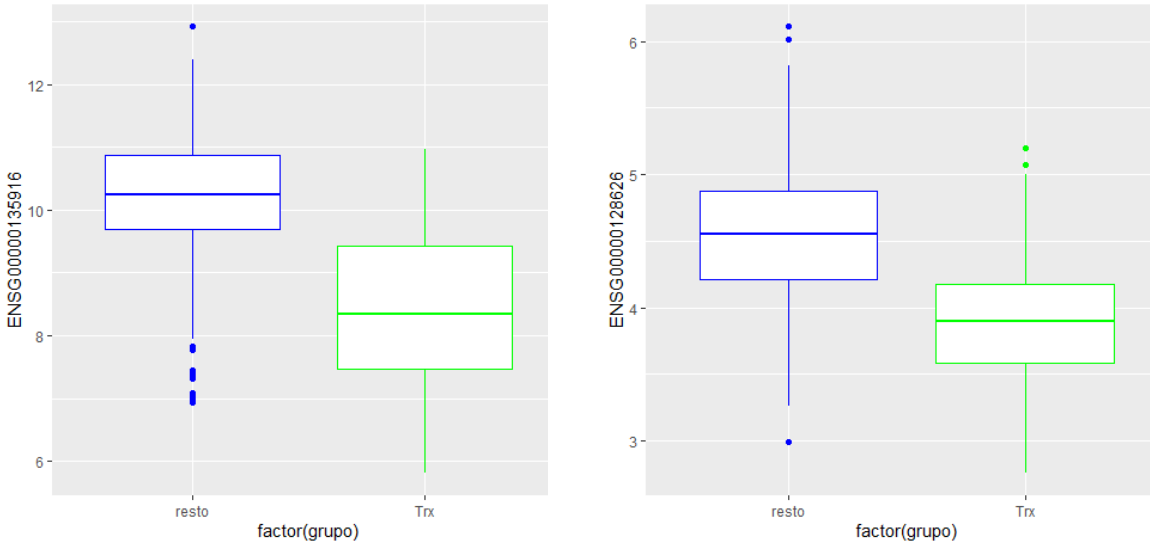


Figura 19. Boxplot para los genes infraexpresados.

t(14;16) *vs.* Pacientes sin eventos:

Los dos genes que más se sobreexpresan en este contraste son ENSG00000159363 *(ATP13A2)* y ENSG00000169242 *(EFNA1)* y sus diagramas de caja están representados en la figura 20.

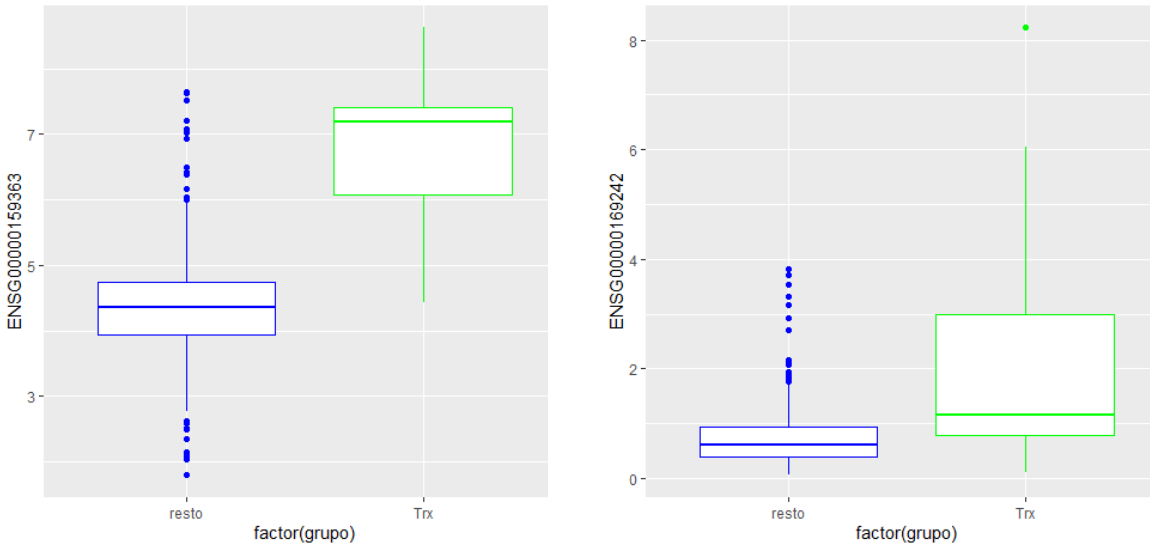


Figura 20. Boxplot para los genes más sobreexpresados.

En la figura 21 se muestran los dos genes más infraexpresados. Éstos son ENSG00000185565 *(LSAMP)* y ENSG0000134369 *(NAV1).*

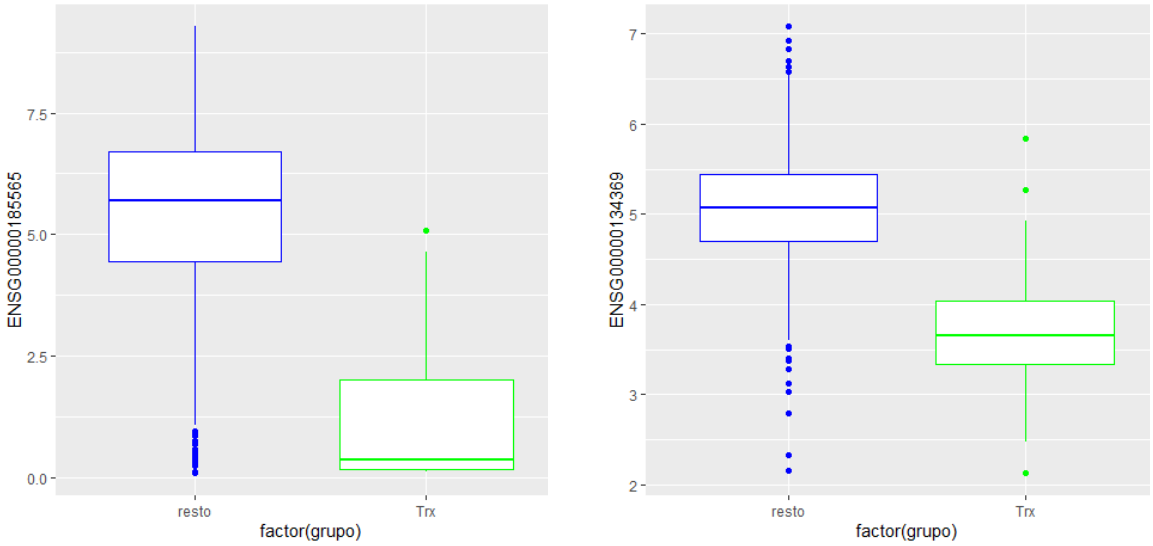


Figura 21. Boxplot para los dos genes más infraexpresados.

Cruce de los genes significativos: diagramas de Venn

El cruce de las listas con los genes significativos de cada firma se llevó a cabo mediante diagramas de Venn, de forma que se obtuvieron los genes que son exclusivos de cada contraste. Como podemos ver a continuación, 2356 genes fueron exclusivos del contraste de la translocación t(4;14), 3685 genes para la translocación t(11;14), 1228 genes para la translocación t(14;16); y finalmente 330 genes fueron exclusivos de la alteración de *TP53* (mutación y deleción):

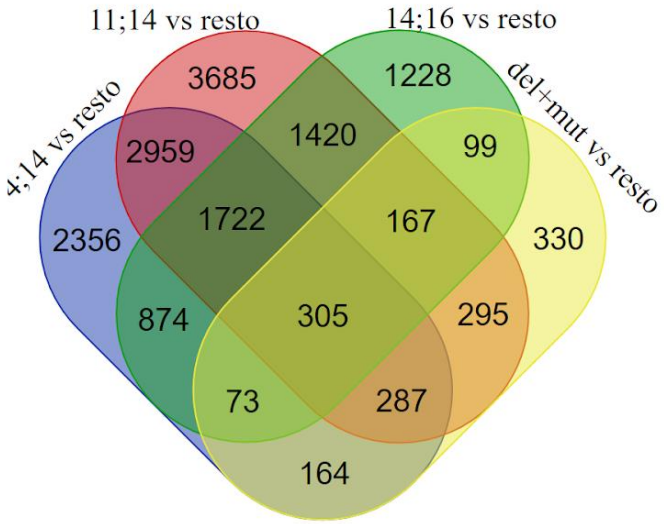


Figura 22. Diagrama de Venn con los genes significativos.

Análisis funcional

Con los genes exclusivos de cada contraste se realizó un análisis de sobrerrepresentación de genes en vías y funciones biológicas (ORA) para determinar las posibles funciones celulares de los genes seleccionados. Esto se realiza en la web *WebGestalt*, donde se extrajeron las 20 rutas y funciones con mayor sobrerrepresentación considerando las bases de datos KEGG, Reactome y Gene Ontology (GO). En todos los casos se empleó *genome protein-coding* como set de referencia.

Los resultados del análisis ORA se muestran en las siguientes Figuras:

***TP53* alterado (mutación y deleción) *vs.* Pacientes sin eventos:**

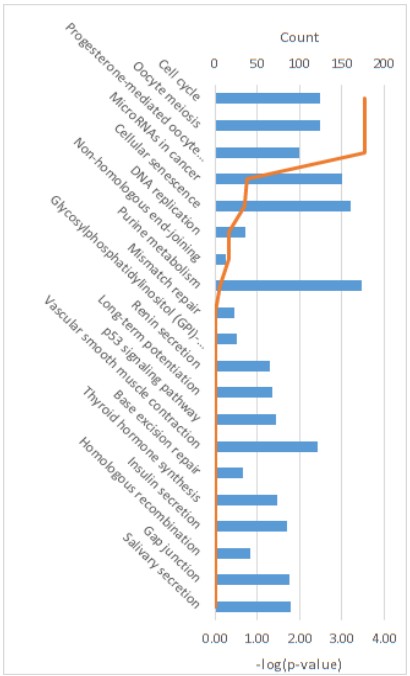


Figura 23. Base KEGG

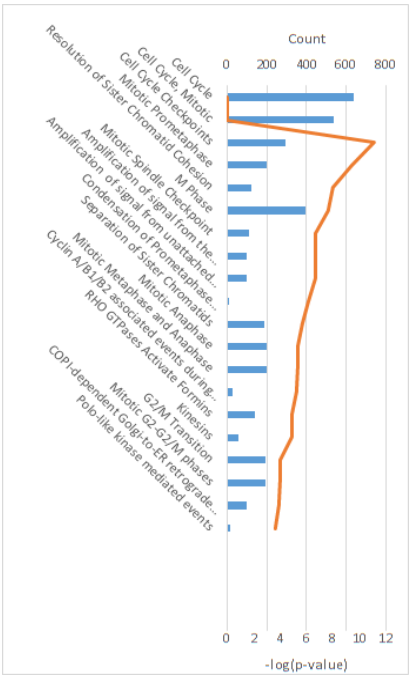


Figura 24. Base Reactome

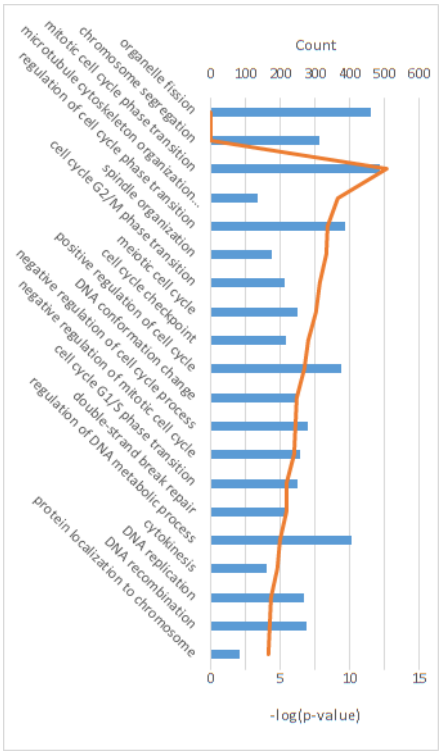


Figura 25. Biological Process no Redundant

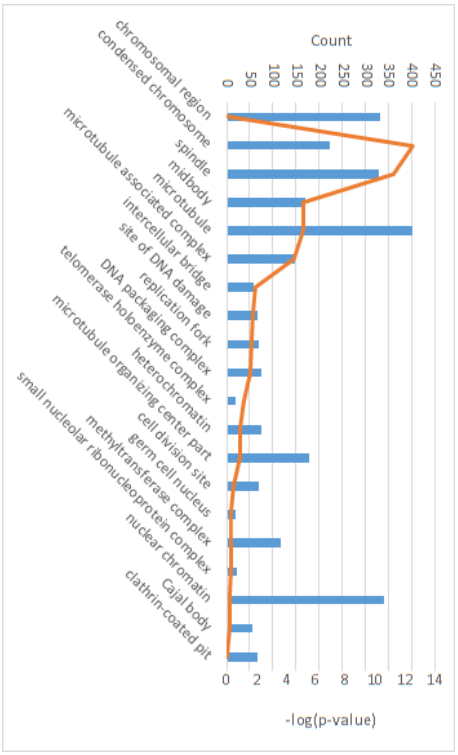


Figura 26. Cellular Component no Redundant

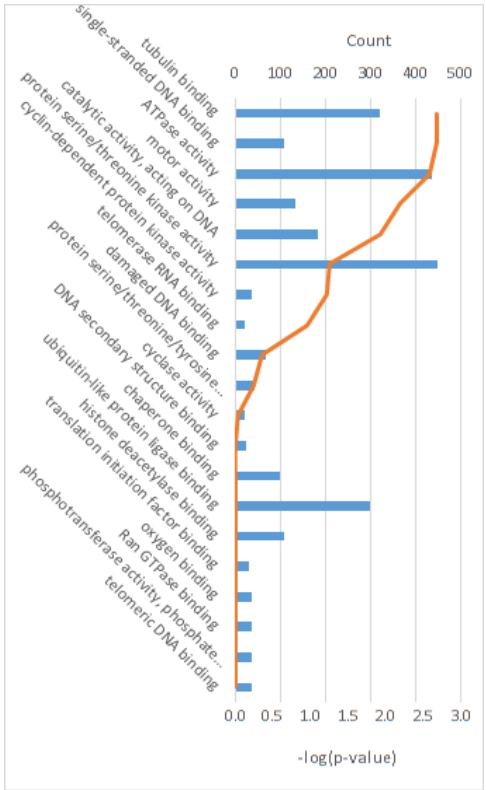


Figura 27. Molecular Function no Redundant

Estos gráficos muestran en uno de los ejes las distintas vías desreguladas, mientras que en el otro se representa el tamaño de cada una de ellas *(Count)*. En el eje secundario se representa el -log10(p-value) que es quien indica si la vía es significativa o no.

De este modo, el resultado del análisis funcional de la firma de la alteración de TP53 reveló una sobrerrepresentación significativa de vías asociadas al ciclo celular en varias vertientes. Así, podemos encontrar genes implicados en los cambios de fase de dicho ciclo, reguladores de la formación del huso acromático o específicos de fase como la prometafase. Así, dichos resultados concuerdan con la posible función del gen TP53 en la regulación del ciclo celular y, por tanto, esta firma detectada podría indicar el mecanismo mediante el cual TP53 actuaría a este nivel.

De manera análoga se realizó el análisis ORA para los genes exclusivos del contraste de t(4;14). En este caso, tal y como muestran las figuras a continuación, no se detectó ninguna vía significativamente sobrerrepresentada:

Para los genes exclusivos del contraste de t(4;14) no hay ninguna vía significativa por lo que no se incluye ningún gráfico en el trabajo.

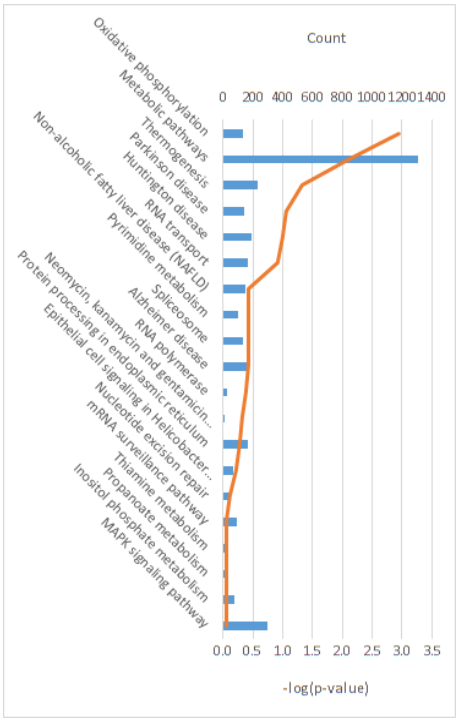


Figura 28. KEGG

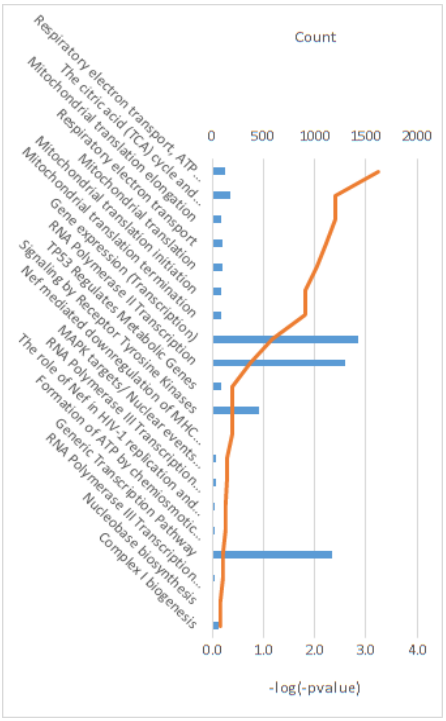


Figura 29. Reactome

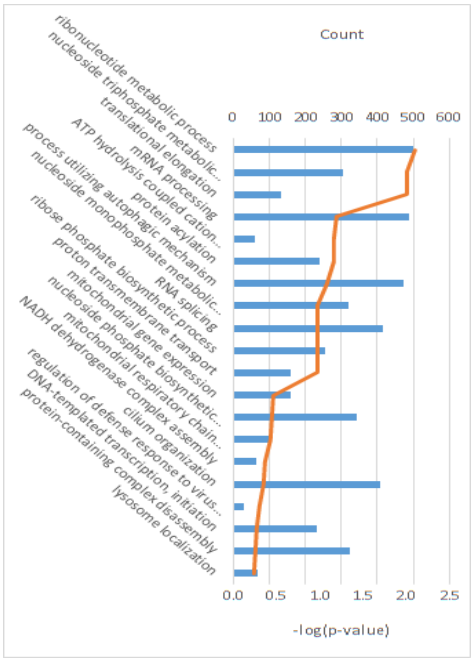


Figura 30. Biological Process no Redundant

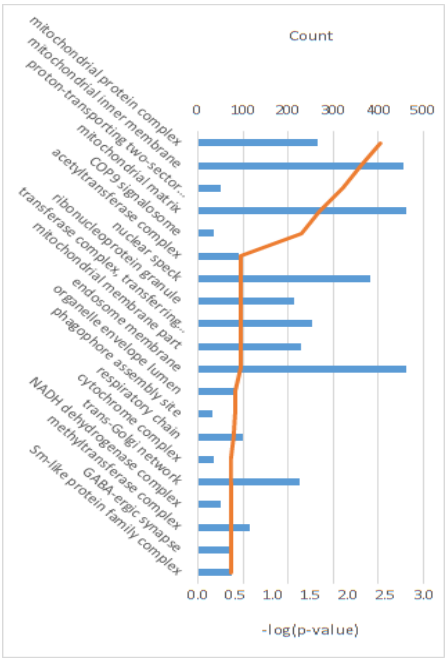


Figura 31. Cellular Component no Redundant

Resulta relevante el hecho de contar con vías en las que encontramos más de 400 genes, como por ejemplo las vías metabólicas utilizando la base de datos Reactome o vías KEGG asociadas con procesos de regulación de la transcripción. Dichos mecanismos podrían tener una función relevante dentro de este subgrupo de MM con lo que sería recomendable indagar en las posibles implicaciones de los genes detectados en el desarrollo y evolución del MM en los pacientes con t(11;14).

Finalmente, se realizó el análisis ORA para la firma de la translocación t(14;16), de modo que a continuación se muestran los gráficos de las vías y funciones más relevantes:

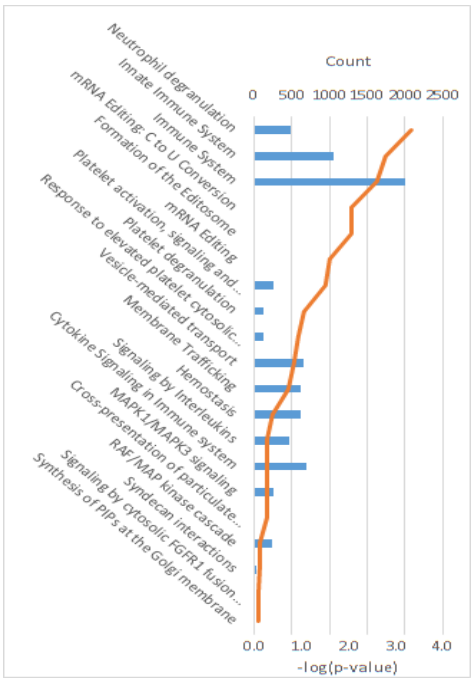


Figura 32. Reactome

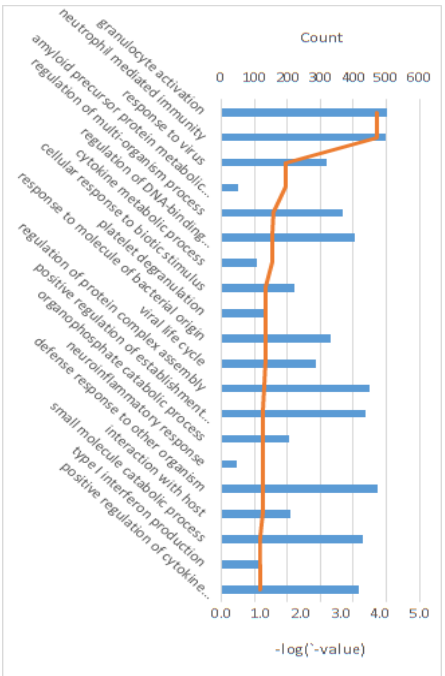


Figura 33. Biological Process no Redundant

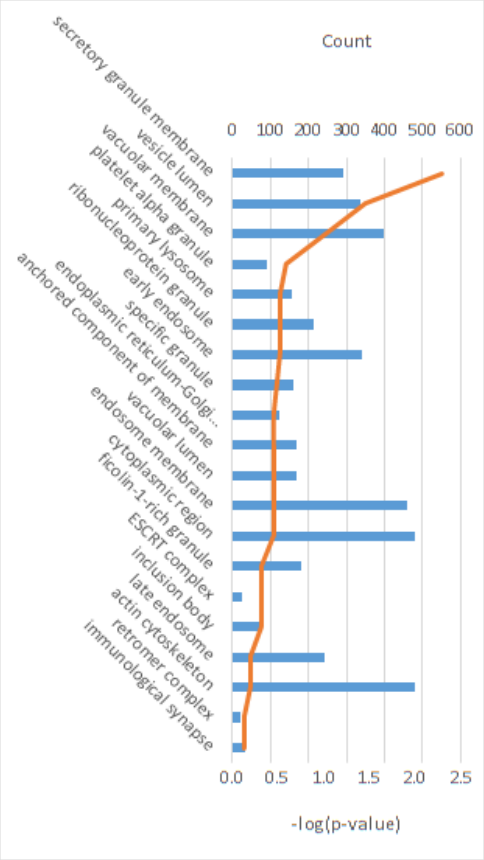


Figura 34. Cellular component no Redundant

En el caso de la t(14;16) encontramos un gran número de vías y funciones significativas asociadas al sistema inmune, lo que podría significar un gran aporte de la desregulación de estos procesos para el desarrollo de la patología en pacientes con MM t(14;16), y desvelar por qué los agentes inmunomoduladores pueden ser una terapia efectiva en este tipo de tumores.

Predicción de la progresión por Machine Learning

Para realizar el análisis de predicción de la progresión del MM mediante *machine learning* se crearon dos conjuntos o matrices de datos de los pacientes, el de entrenamiento y el de validación, de manera que 2/3 de los pacientes fueron seleccionados para el primer conjunto, mientras que el tercio restante pasó a formar parte del conjunto de validación.

Las muestras que forman parte de cada conjunto se eligieron aleatoriamente en el programa SIMFIT, donde se creó una lista de permutaciones aleatorias de 1 hasta n, siendo n en este caso 316 (el número de pacientes que presenta progresión) a cada muestra se le asoció un número de esta progresión, se ordenan de menor a mayor y se eligen las 211 primeras muestras (2/3 de 316). El proceso se repite para los pacientes que no presentan progresión. Al juntar las dos matrices se obtiene la matriz de entrenamiento con 2/3 del total de pacientes (409).

La matriz de validación está formada por las 204 muestras restantes.

Este proceso se realizó para los genes exclusivos de cada contraste. De modo que se obtienen 8 matrices, la de entrenamiento y la de validación de los 4 contrastes.

Predicción con SVM

Para llevar a cabo el estudio predictivo con **SVM** en una primera instancia se deben calcular el Kernel y los parámetros *gamma* y *cost* óptimos. El mejor Kernel será el que tenga un menor error cuadrático medio o *rmse*, que en nuestro caso fue el Kernel radial. En un segundo paso hay que calcular un factor de ponderación para compensar las diferencias muestrales entre los grupos contrastar, para ello se utilizó la siguiente fórmula:

$$\frac{n^o\, total\ de\ muestras}{n^o\, de\ grupos * n^o\, de\ muestras\ en\ el\ grupo\ y}$$

Esta técnica pondera de forma proporcional los pesos de los dos grupos para equipararlos, de modo que los pesos fueron, para el grupo progresión: 409/2*211 = 0.969; y para el grupo no progresión: 409/2*198= 1.033

Posteriormente se obtuvieron los valores óptimos de los parámetros *gamma* y *cost a* través de la ejecución de una función de afinado o tuning en R, siendo en nuestro estudio de la alteración de *TP53* 0.25 y 1 respectivamente.

El Script completo con el que se ha realizado la predicción se encuentra en los anexos.

Este proceso fue repetido para todos los contrastes, donde en el caso de la 4;14 los parámetros óptimos para coste y gamma también fueron 1 y 0.25. mientras que para la t11;14 el kernel elegido fue el radial, que proporcionó unos parámetros óptimos de coste= 1 y gamma= 0.25

Tras el afinado del método se llevó a cabo la predicción sobre la matriz de validación, siendo los resultados obtenidos son los siguientes:

Tabla 6. Resultados de la predicción utilizando las firmas transcripcionales completas.

	Especificidad	sensibilidad	precisión	exactitud
Alteración *TP53*	0	1	0	0.5147
t(4;14)	0	1	0	0.5147
t(11;14)	0	1	0	0.5147
t(14;16)	0	1	0	0.5147

La especificidad obtenida es de cero para todos los casos, es decir, este algoritmo no captura bien a los pacientes que no presentan progresión en la enfermedad. Por el contrario, si captura bien a los que presentan progresión, ya que la sensibilidad es 1. La precisión es cero, lo que indica que la calidad de clasificación es mala. La exactitud es de 0.5147 para todos los contrastes, es aceptable pero no es buena.

El problema que se presenta es que el número de genes con los que se trabaja es muy grande, es decir, hay un número muy elevado de predictores en comparación con el número de muestras, lo que puede producir problemas de sobreajuste. Para solventar este problema se llevó a cabo la selección de los genes con un mayor valor del parámetro "importancia" en un modelo de *Random Forest* con el paquete *Boruta*. Este parámetro mide el grado de asociación de cada una de las variables predictoras sobre la variable respuesta.

Selección de variables con Boruta

El análisis de selección de variables se realizó de forma exclusiva sobre la matriz de entrenamiento.

En el caso de la alteración de *TP53*, se obtuvieron 9 genes determinados como importantes (verde), 20 genes tentativos de ser importantes (amarillo) y 300 genes que no presentaron asociación con la variable dependiente (Figura 33).

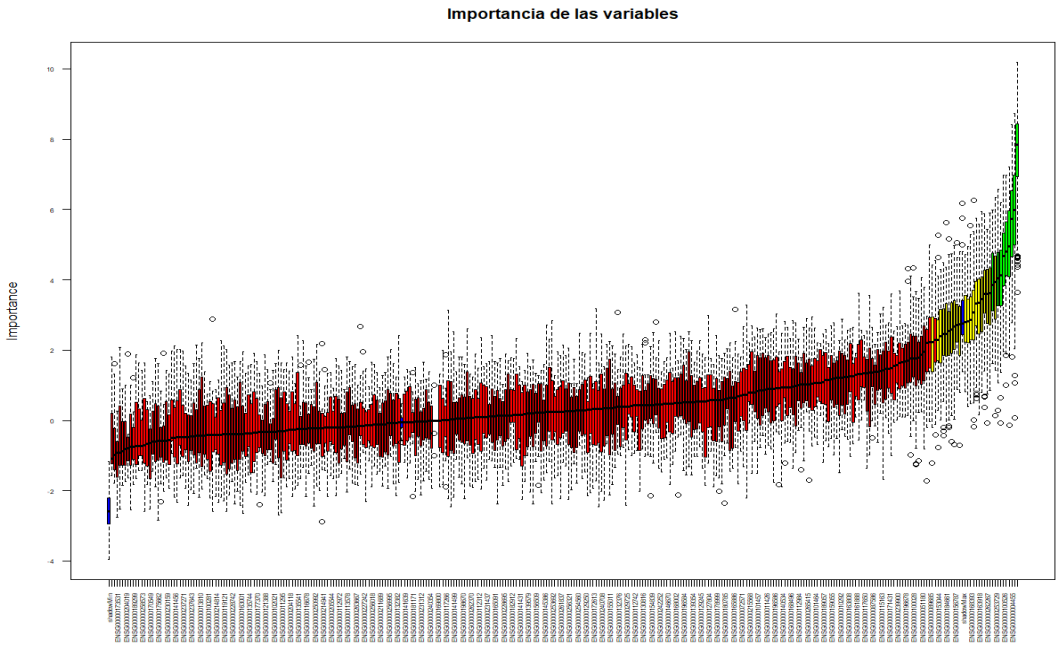


Figura 35. Grafico de importancia de las variables.

Asi, para la nueva predicción de la progresión en el grupo de la alteración de TP53 se seleccionan las 9 variables con mayor importancia sobre la variable dependiente. Las amarillas intermedias son aquellas que son menos importantes y que podrían ser usadas en caso de que no existiera ninguna variable considerada como importante. El nuevo ajuste del modelo SVM con los 9 genes importantes obtuvo como kernel óptimo el lineal y como mejores parámetros cost 0.1 y gamma 0.25

En el caso de la translocación t(4;14) se obtuvo que 8 genes son importantes respecto a la variable dependiente, 22 son menos importantes y 2326 no lo son. Los 8 genes importantes se visualizan en la Figura 36.

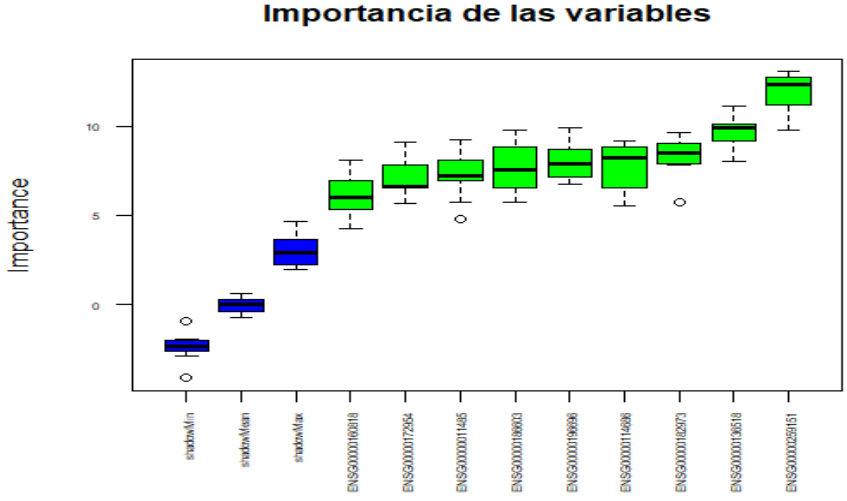


Figura 36. Gráfico de las variables importantes de t(4;14)

El ajuste del modelo SVM con 8 genes importantes determinó que los parámetros óptimos fueron coste 0.01 y gamma 0.25

Para la translocación 11;14, *boruta* devuelve 9 variables importantes reflejadas en el siguiente gráfico, realizándose nuevamente el ajuste del modelo con estas variables y obteniendo como mejores parámetros un coste 10 y gamma 0.25.

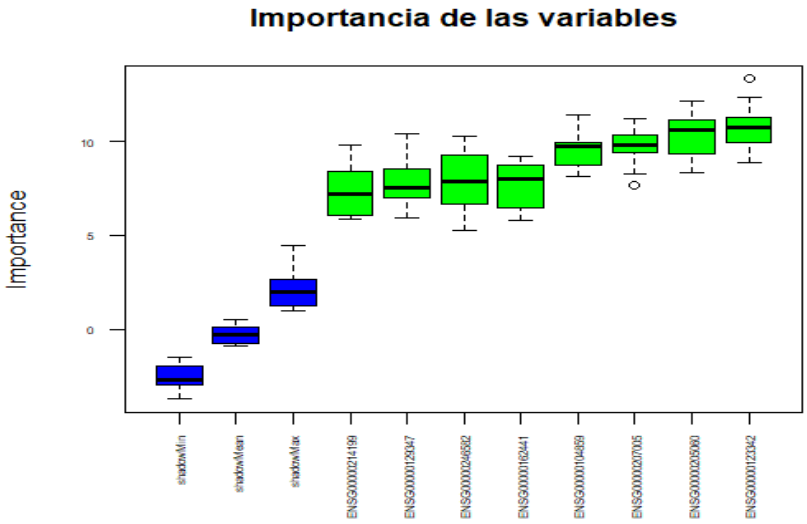


Figura 37. Gráfico para las variables importantes de t(11;14)

Los mejores parámetros son coste 10 y gamma 0.25

Finalmente, en el caso de la translocación t(14;16), de las 1229 variables de partida, 8 resultaron importantes.

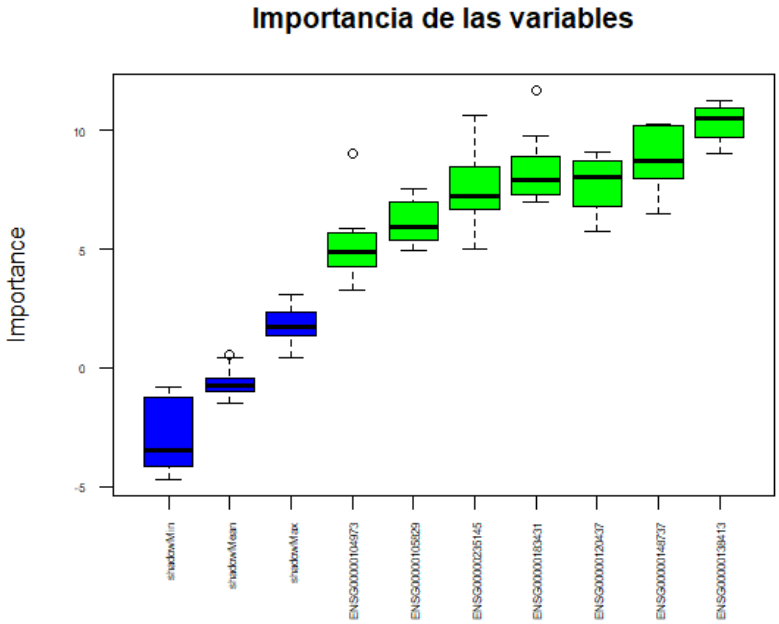


Figura 38. Gráfico de las variables importantes de t(14;16)

A continuación, se muestra la tabla 7 con los resultados de la predicción con las variables importantes devueltas por Boruta.

Tabla 7. Resultado de la predicción con SVM utilizando las variables imortantes.

	especificidad	sensibilidad	VPN	precisión	exactitud
t(4;14)	0.3838	0.7143	0.5588	0.5515	0.5539
t(11;14)	0.6364	0.4952	0.5431	0.5909	0.5637
t(14;16)	0.4747	0.7333	0.6267	0.5969	0.6078
Alteración *TP53*	0.6667	0.4667	0.5410	0.5976	0.5637

Se puede observar una mejora en la predicción, ya que esta vez ni la especificidad ni la precisión son cero, lo que indica que esta vez el modelo si captura a los pacientes sin progresión en la enfermedad. A nivel global, con la exactitud, el mejor resultado es el de la traslocación t(14;16) ya que a partir de 0.6 es considerada moderadamente buena.

El valor predictivo negativo para los cuatro casos se mueve entre 0.54 y 0.62, esto determina que entre un 54 y 62% de los pacientes que no presentan progresión son clasificados correctamente por el modelo. La precisión (o valor predictivo positivo) indica la proporción de pacientes con progresión que están bien clasificados.

KNN (K-Nearest Neighbors)

Para ajustar los modelos de KNN, tanto en las matrices de entrenamiento como en las de validación, se realizó un reajuste la variable grupo ya que esta debe ser una variable de tipo factor, por lo que se recodifican los grupos anteriores “0” y “1” en “g0” y “g1”.

En la siguiente tabla se muestran los resultados de la predicción obtenidos mediante el algoritmo de K-nearest neighbors.

Tabla 8. Resultados de la predicción con KNN

	especificidad	sensibilidad	VPN	precisión	exactitud
t(4;14)	0.7475	0.3429	0.5175	0.5902	0.5392
t(11;14)	0.6162	0.5524	0.5648	0.6042	0.5833
t(14;16)	0.7677	0.3714	0.5352	0.6290	0.5637
Alteración *TP53*	0.4242	0.6190	0.5122	0.5328	0.5245

El algoritmo KNN se muestra más robusto a los pacientes sin progresión ya que tiene una especificidad mayor a la sensibilidad. Según los valores de la precisión, la firma para la cual se clasifican mejor los pacientes con progresión en la enfermedad es t(14;16) (un 62% de ellos).

A continuación, se realiza la predicción con las variables importantes ofrecidas por boruta anteriormente para evitar el sobreajuste:

Tabla 9. Resultados de la predicción con KNN utilizando los genes importantes.

	especificidad	sensibilidad	VPN	precisión	Exactitud
t(4;14)	0.4646	0.6667	0.5679	0.5691	0.5686
t(11;14)	0.5556	0.5333	0.5288	0.5600	0.5441
t(14;16)	0.3939	0.7619	0.6094	0.5714	0.5833
Alteración *TP53*	0.5657	0.5333	0.5333	0.5657	0.5490

No se observa una mejora notoria en la predicción al realizarla con las variables importantes. La exactitud para todas las firmas ronda entre 0.5 y 0.6, lo que se considera aceptable. Para la t(4;14) y la t(14;16) el algoritmo captura mejor a los pacientes con progresión en la enfermedad ya que la sensibilidad es mayor que la especificidad. En el caso de TP53 y t(11;14) se produce lo contrario.

RANDOM FOREST

En la tabla 10 se muestran los resultados para Random Forest realizado con todos los genes de la base de datos.

Tabla 10. Resultados de la predicción con Random Forest para todos los genes.

	Especificidad	sensibilidad	Precisión	VPN	Exactitud
t(4;14)	0.4545	0.6962	0.5844	0.5748	0.5784
t(11;14)	0.4949	0.6571	0.5765	0.5798	0.5784
t(14;16)	0.5253	0.7143	0.6341	0.6148	0.6225
Alteración *TP53*	0.6061	0.6190	0.6000	0.6250	0.6127

Este algoritmo captura mejor a los pacientes con presencia de la progresión, ya que en los cuatro casos la sensibilidad es mayor que la especificidad. La precisión indica que entre un 57% y 63% de los pacientes con progresión se clasifica de forma correcta. Y el valor predictivo negativo (VPN) indica que entre un 57-62% de los pacientes sin progresión se clasifica de forma correcta.

La predicción con todos los genes mediante Random Forest es relativamente buena, esto puede deberse a que como se especificó anteriormente, es una técnica robusta ante conjuntos de datos de gran dimensión.

Tabla 11. Resultado de la predicción con Random Forest para los genes importantes.

	especificidad	sensibilidad	Precisión	VPN	Exactitud
t(4;14)	0.5556	0.5714	0.5500	0.5769	0.5637
t(11;14)	0.5354	0.6190	0.5699	0.5856	0.5784
t(14;16)	0.5152	0.6286	0.5667	0.5789	0.5735

Alteración *TP53*	0.5556	0.5238	0.5238	0.5556	0.5392

Se puede observar en las tablas anteriores que bajo el algoritmo de Random Forest, las predicciones son similares al realizarlas con todos los genes y solo con los importantes, pero en este último caso dicha predicción no está condicionada por el sobreajuste.

Capítulo 5.- Conclusiones

1. Las mutaciones en *TP53* son un factor pronóstico negativo en la progresión del MM con un impacto mayor sobre el tiempo de supervivencia que la deleción de *TP53*.

2. Se logró la identificación de una firma de expresión génica para los grupos de pacientes con las traslocaciones t(4;14), t(11;14) y t(14;16), así como para los pacientes con alteraciones en el gen *TP53*. En este último caso se comprobó que la alteración de *TP53* produce la desregulación de la expresión de genes implicados en el ciclo celular.

3. Los métodos de *machine learning* que obtuvieron mejores resultados fueron el SVM ponderado y Random Forest, con rendimientos cercanos al 60% en la mayor parte de los parámetros de medida del desempeño estudiados.

4. La predicción llevada a cabo con grandes conjuntos de genes no resultó efectiva con los métodos SVM y KNN, posiblemente debido a problemas de sobreajuste. Sin embargo, métodos como Random Forest, diseñados para trabajar con un mayor número de variables, presentaron un mejor desempeño en estos escenarios.

Capítulo 6.- Bibliografía

A, Q., & N, K. (2021). *bedtools: a powerful toolset for genome arithmetic*.

AEAL. (2017). Sintomas y manifestaciones. Retrieved from http://www.aeal.es/mieloma-multiple-espana/3-sintomas-y-manifestaciones/

AECC. (2021). Mieloma Múltiple. Retrieved from https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/mieloma-multiple

Amat Rodrigo, J. (2016). *Comparaciones múltiples: corrección de p-value y FDR*. Retrieved from https://www.cienciadedatos.net/documentos/19b_comparaciones_multiples_correccion_p-value_fdr

Amat Rodrigo, J. (2017). Máquinas de Vector Soporte (Support Vector Machines SVMs). Retrieved from https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Bajaj, S., Alam, S. K., Singha Roy, K., Datta, A., Nath, S., & Roychoudhury, S. (2016). *No TitleE2 Ubiquitin-conjugating Enzyme, UBE2C Gene, Is Reciprocally Regulated by Wild-type and Gain-of-Function Mutant p53*.

Behjati, S., & Tarpey, P. S. (2013). *What is next generation sequencing?* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/

Berzal, F. (2021). *Clustering Jerárquico*.

Biología, A. V. de. (2021). ARN o ácido ribonucleico o RNA. Retrieved from https://www.um.es/molecula/anucl03.htm

Borg, I., & Groenen, P. G. (1997). *Modern Multidimensional Scaling*.

Carrasco, S. (2015). *Técnicas de análisis de expresión diferencial basadas en conceptos para el estudio de datos de RNA-seq usando R y bioconductor*.

Carreño Serra, Á. (2006). *Análisis de supervivencia*.

Chmielecki, J., & Meyersom, M. (2013). *DNA sequencing of cancer: what have we learned?* Retrieved from https://pubmed.ncbi.nlm.nih.gov/24274178/

Clements, J. (2019). *Introduction to Hierarchical Clustering*.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). *Random Forest*.

Delgado, R. (2018). Introducción a la validación cruzada (k-fold Cross Validation) en R. Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/405322_6d94d05e54b24ba99438f49a6f8662a9.html

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1

Geneontology. (2021). THE GENE ONTOLOGY RESOURCE.

Gomez, G., & Cobo, E. (2004). *Hablemos de ... Análisis de supervivencia*.

González, A. (2021). Conceptos básicos de Machine Learning. Retrieved from https://cleverdata.io/conceptos-basicos-machine-learning/

Gupta, N., & Verma, V. K. (2019). *Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality*. Retrieved from https://link.springer.com/chapter/10.1007/978-981-13-8844-6_15

Illumina. (2021). Understanding the genetic code. Retrieved from https://www.illumina.com/techniques/sequencing/dna-sequencing.html

javatpoint. (2021). K-Nearest Neighbor(KNN) Algorithm for Machine Learning. Retrieved from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Kanezaki, R., Toki, T., Xu, G., Narayanan, R., & Ito, E. (2006). *Cloning and characterization of the novel chimeric gene p53/FXR2 in the acute megakaryoblastic leukemia cell line CMK11-5*.

Krzeminski, P., Corchete, L. A., García, J. L., López-Corral, L., Fermiñán, E., García, E. M., … Gutiérrez, N. C. (2016). *Integrative analysis of DNA copy number, DNA methylation and gene expression in multiple myeloma reveals alterations related to relapse*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5348347/

Liao, Y., Shi, Z., & Zhang, B. (2019). WEB-based GEne SeT AnaLysis Toolkit. Retrieved from http://www.webgestalt.org/

Llopis Pérez, J. (2013a). La estadística: una orquesta hecha instrumento. Retrieved from https://jllopisperez.com/2013/01/07/tema-21-analisis-de-supervivencia/

Llopis Pérez, J. (2013b). La estádsitica: una orquesta hecha instrumento. Retrieved from https://jllopisperez.com/2013/01/18/tema-22-regresion-de-cox/

MMRF. (2021). Prognosis. Retrieved from https://themmrf.org/multiple-myeloma/prognosis/

Molina Arias, M. (2015). Hazard ratio: cuando el riesgo varía a lo largo del tiempo. Retrieved from https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322015000300023

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., … Shendure, J. (2009). *Targeted capture and massively parallel sequencing of 12 human exomes*.

Nguyen, B. (2017). Soft Margin Support Vector Machine. Retrieved from https://machinelearningcoban.com/2017/04/13/softmarginsmv/

Noble, W. S. (2006). *What is a support vector machine?*

Peterson, L. E. (2009). *K-nearest neighbor*.

Pita Fernández, S. (2001). Análisis de supervivencia. Retrieved from https://www.fisterra.com/mbe/investiga/supervivencia/supervivencia.asp

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*.

Robinson, M. D., & Oshlack, A. (2010). *A scaling normalization method for differential expression analysis of RNA-seq data*.

Sancho Caparrini, F. (2017). Introducción al aprendizaje automático. Retrieved from http://www.cs.us.es/~fsancho/?e=75

Schott, M. (2019). Random Forest Algorithm for Machine Learning. Retrieved from https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb

Srivastava, D. K., & Bhambhu, L. (2005). *Data classification using support vector machine*.

Storey, J. D. (2010). *False discovery rates*.

Toledano Díaz, M. (2019). Robótica y aplicación de Diagramas de Voronoi. Retrieved from http://fisicotronica.com/robotica-aplicacion-diagramas-voronoi/

Torgerson, W. S. (1958). *Theory and methods of scaling*.

Valencia, U. I. de. (2017). ADN y ARN concepto, diferencias y funciones. Retrieved from https://www.universidadviu.com/es/actualidad/nuestros-expertos/adn-y-arn-concepto-diferencias-y-funciones

Vapnik, V. (1995). *Support-vector networks*.

Wang, Z., Gerstein, M., & Snyder, M. (2009). *RNA-Seq: a revolutionary tool for transcriptomics*.

Yiu, T. (2019). Understanding Random Forest. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zaforas, M. (2017). Machine Learning para dummies. Retrieved from https://www.paradigmadigital.com/techbiz/machine-learning-dummies/

ANEXO

A continuación, se muestran los Scripts de R utilizados a lo largo del trabajo.

Normalización con edgeR:

```
countsTable <-
read.table("RNAseq.txt",header=TRUE,sep="\t",row.names = 1)
library(edgeR)
grupos <- as.factor(c(rep(1,300),rep(0,343)))
dgeObj <- DGEList(countsTable, group=factor(grupos))
dgeObj <- calcNormFactors(dgeObj)
d1 <- estimateCommonDisp(dgeObj)
d1 <- estimateGLMTrendedDisp(d1)
d1 <- estimateTagwiseDisp(d1)
plotBCV(d1)
??calcNormfactors
cp<- cpm(dgeObj)
options("max.print"=1E9)
options("width"=10000)
sink("matriz_norm_edge.txt")
print(cp)
sink()
```

Análisis de supervivencia:

```
library(readxl)
a<- read_excel("RNAseq.xlsx", sheet=5)
library(survival)
library(survminer)
library(tidyr)



tp53 <- a%>%
  mutate(del=ifelse(deleciones=="si","del",NA)) %>%
  mutate(mut= ifelse(mutaciones=="si","mut",NA)) %>%
  mutate(no=ifelse(mutaciones=="no"&deleciones=="no", "sin
tp53",NA))%>%
  unite(del,mut,no,col="grupo",sep="+", na.rm=T)
```

```
  t414 <- a%>%
    mutate(Trx=ifelse(t4_14=="si","t(4;14)",NA)) %>%
    mutate(no=ifelse(t4_14=="no", "resto",NA)) %>%
    unite(Trx,no,col="grupo",sep="+",na.rm=T)

  t1114 <- a%>%
    mutate(Trx=ifelse(t11_14=="si","t(11;14)",NA)) %>%
    mutate(no=ifelse(t11_14=="no", "resto",NA)) %>%
    unite(Trx,no,col="grupo",sep="+",na.rm=T)

  t1416 <- a%>%
    mutate(Trx=ifelse(t14_16=="si","t(14;16)",NA)) %>%
    mutate(no=ifelse(t14_16=="no", "resto",NA)) %>%
    unite(Trx,no,col="grupo",sep="+",na.rm=T)

  ### kaplan meier


  survObject <- Surv(tp53$ttpfs,tp53$censpfs)
  fit <- survfit(survObject~grupo,data=tp53)
  ggsurvplot(fit,data=tp53,xlab="tiempo de supervivencia", conf.int
= FALSE, title="tp 53")


  survObject <- Surv(t414$ttpfs,t414$censpfs)
  fit <- survfit(survObject~grupo,data=t414)
  ggsurvplot(fit,data=tp53,xlab="tiempo de supervivencia", conf.int
= FALSE, title="t(4;14)")


  survObject <- Surv(t1114$ttpfs,t1114$censpfs)
  fit <- survfit(survObject~grupo,data=t1114)
  ggsurvplot(fit,data=tp53,xlab="tiempo de supervivencia", conf.int
= FALSE, title="t(11;14)")


  survObject <- Surv(t1416$ttpfs,t1416$censpfs)
  fit <- survfit(survObject~grupo,data=t1416)
  ggsurvplot(fit,data=tp53,xlab="tiempo de supervivencia", conf.int
= FALSE, title="t(14;16)")
```

Análisis de expresión diferecial:

```
c<- exactTest(d1, pair=c(1,2))
c1<-topTags(c,n=length(countsTable[,1]),adjust.method =
"BH",sort.by = "PValue",p.value = 1)
options("max.print"=1E9)
options("width"=10000)
sink("exp_diferencial_edgeR")
print(c1)
sink()  ## hacer expresion diferencial

de1<- decideTestsDGE(c,adjust.method = "BH",p.value = 0.05)
summary(de1) ## ver genes sobre, bajo expresados
table(c1$table$FDR<0.05) #genes significativos y no significativos
table(c1$table$FDR<0.05/nrow(c1$table))
hist(c1$table$FDR, breaks=40, main="Histograma de FDR")
abline(v=0.05,col="red", lwd=3)
```

Boxplot de los genes sobre e infraexpresados:

```
a<- read.table("boxplot.txt", sep="\t", header=TRUE)
library(ggplot2)
bp<-ggplot(a, aes(x=factor(grupo),y=ENSG00000197355))
bp+geom_boxplot(col=c("blue","green"))
```

Machine Learning

SVM:

```
a <- read.table("entrenamiento.txt", header=TRUE,sep="\t",
row.names = 1)
tabla <- as.data.frame(t(a))
x <- subset(tabla,select=-Grupo)
y <- tabla$Grupo
y <- as.factor(y)
##Determinar el mejor kernel (elegir mejor success_rate)
set.seed(1234)
library(OptimClassifier)
modelFit <- Optim.SVM(Grupo~., data=tabla, p = 0.7, seed=2018)
modelFit
#determinar parámetros óptimos SVM
```

```
  ##PARA CALCULAR LOS PESOS DE CADA GRUPO SE HACE INVERSAMENTE
PROPORCIONAL A LA FRECUENCIA: n_samples / (n_classes *
np.bincount(y))

  ##es decir: N° total de muestras / (N° de grupos * Muestras en
grupo y)

  library(e1071)

  svm_tune <- tune(svm,train.x=x,
train.y=y,kernel="radial",tunecontrol = tune.control(cross
=length(y)),class.weights=
c("0"=1.033,"1"=0.969),ranges=list(cost=10^(-3:2),gamma=2^(-2:4)))


  plot(svm_tune)

  ##The choice of k is usually 5 or 10, but there is no formal rule.
As k gets larger, the difference in size between the training set
and the resampling subsets gets smaller. As this difference
decreases, the bias of the technique becomes smaller

  ##cross =lenght(y) = LOOCV, approximately unbiased

  #svm_tune <- tune(svm,train.x=x,
train.y=y,kernel="sigmoid",class.weights= c(),ranges=list(cost=10^(-
5:15),gamma=10^(-15:3)))

  print(svm_tune)

  ##Con los mejores "cost" y "gamma" del modelo anterior ajustar el
modelo de SVM:

  svm_model_after_tune <-
svm(x,y,kernel="radial",cost=1,gamma=0.25,class.weights=c("0"=1.033,
"1"=0.969),cross =length(y))

  summary(svm_model_after_tune)

  pred <- predict(svm_model_after_tune,x)

  table(pred,y)

  sink("g_SVM_prediccion_LOO_train.txt")

  print(table(pred,y))

  sink()


  #Prediccion en test

  z <- read.table("validacion.txt",header=TRUE,sep="\t",row.names=1)

  testab <- as.data.frame(t(z))

  xtest <- subset(testab,select=-Grupo)

  pred2 <- predict(svm_model_after_tune,xtest)

  ytest <- testab$Grupo

  ytest <- as.factor(ytest)

  table(pred2,ytest)

  sink("h_prediccion_LOO_test.txt")
```

```
print(table(pred2,ytest))
sink()

library(caret)
confusionMatrix(table(pred2,ytest))
```

KNN:

```
tabla <-
read.table("entrenamiento_var_importantes.txt",header=TRUE,sep="\t",
row.names=1)
testab <-
read.table("validacion_var_importantes.txt",header=TRUE,sep="\t",row
.names=1)
x <- subset(tabla,select=-Grupo)
y <- tabla$Grupo
y <- as.factor(y)
testab2 <- subset(testab,select=-Grupo)
testabR <- testab$Grupo

library(caret)
x = trainControl(method ="cv",
                 classProbs = TRUE,
                 summaryFunction = multiClassSummary)
set.seed(1234)
modelKNN <- train(Grupo ~ ., data=tabla, method = "knn",
                  preProcess = c("center", "scale"),
                  trControl = x,
                  metric = "ROC",
                  tuneLength = 10)
sink("s1_KNN_Model.txt")
print(modelKNN)
sink()
# Validation
pred_KNN <- predict(modelKNN,testab2)
confusionMatrix(pred_KNN,as.factor(testabR))
sink("m_KNN_Results.txt")
print(pred_KNN)
sink()
sink("n_Confusion_KNN.txt")
```

Random Forest:

```
modelRF <- train(as.factor(Grupo) ~ ., data=tabla, method = "rf",
                 preProcess = c("center", "scale"),
                 trControl = x,
                 metric = "AUC",
                 tuneLength = 10)
sink("t1_RF_Model.txt")
print(modelRF)
sink()
# Validation
pred_RF <- predict(modelRF,testab2)
confusionMatrix(pred_RF,as.factor(testabR))
sink("o_RF_Results.txt")
print(pred_RF)
sink()
sink("p_Confusion_RF.txt")
print(confusionMatrix(pred_RF,as.factor(testabR)))
sink()
```

Selección de variables importantes con Boruta:

```
tabla <-
read.table("entrenamiento_var_importantes.txt",header=TRUE,sep="\t",
row.names=1)
tabla <- as.data.frame(t(tabla))
names(tabla)
##Convertir variables categoricas en factor
tabla$Grupo = as.factor(tabla$Grupo)
##Resument
summary(tabla)
##Lanzar boruta
library(Boruta)
set.seed(123)
boruta <- Boruta(Grupo~.,data=tabla,doTrace=2)
print(boruta)
plot(boruta,cex.axis=0.5,las=2,xlab="",main="Importancia de las
variables",srt=45)
options("width"=10)
sink("d_Recod_Variables_Sel_ZBS_2018_Vdec.txt")
```

```
print(boruta$finalDecision)
sink()
##Plot_boruta history
#plotImpHistory(boruta)
#Importance of variables
boruta.df <- attStats(boruta)
options("width"=10000)
options("max.print"=1E9)
sink("b3_Importance_of_Scaled_variables_ALL_samples.txt")
print(boruta.df)
sink()
```