Búsqueda de Dianas Genéticas como Polimorfismos (SNPs) en Genes Asociados a la Enfermedad Neurodegenerativa del Alzheimer

María del Pilar Barruz Galián
Máster de Bioinformática y Bioestadística
Biología del desarrollo, cáncer,
biología molecular y farmacología.

Nombre Consultor/a
Dña Ivette Olivares Castiñeira

Nombre Profesor/a responsable de la asignatura
D. Carles Ventura Royo

Madrid 5 de Junio de 2018

A Mi Madre

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Búsqueda de Dianas Genéticas como Polimorfismos (SNPs) en Genes Asociados a la Enfermedad Neurodegenerativa del Alzheimer*
Nombre del autor:	*M.ª del Pilar Barruz Galián*
Nombre del consultor/a:	Ivette Olivares Castiñeira
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster de Bioinformática y Bioestadística*
Área del Trabajo Final:	*Biología del desarrollo, cáncer, biología molecular y farmacología*
Idioma del trabajo:	*Español*
Palabras clave	*Alzheimer, Snps, Microarrays de Expresión*

Resumen del Trabajo:

Resúmen: El Alzheimer es la más común de las demencias. Se caracteriza principalmente por una pérdida progresiva cognitiva y neurodegenerativa. Epidemiológicamente representa un problema ya que a mediados del siglo XXI constituirá un problema socio-sanitario y es un índice de mortalidad, por lo que es esencial poder determinar la etiología de la enfermedad.

Métodos: Para valorar molecularmente la enfermedad del Alzheimer se realiza el siguiente estudio donde se seleccionan, de Gene Expression Omnibus (del dataset GSE84422), seis tejidos, de distintas regiones corticales y subcorticales, relacionados patognomónicamente con el Alzheimer. Las muestras pertenecen a 125 individuos que presentaban distintos grados de afectación según el criterio CERAD. Se determinan qué genes presentan expresión génica diferencial en dichos tejidos cuyo cambio está asociado con los distintos grados de demencia y con las diferentes lesiones neuropatológicas detectadas en los pacientes diagnosticados de Alzheimer. Se clasifican aquellos genes que presentan el 1% de las desviaciones estándar más altas respecto de la intensidad de señal detectada en los microarrys objeto de estudio. Y mediante Métodos de Regresión Penalizada se seleccionan genes asociados a las variables respuesta de interés. De todos los genes detectados se determinan la significación biológica así como las rutas biológicas en las que están enriquecidos (Análisis de Enriquecimiento).

Resultados: De los seis tejidos analizados, cuatro presentan expresión diferencial. En todos los tejidos se ha obtenido el mismo número de genes (al seleccionar el 1% de las desviaciones estándar más altas) excepto en el tejido Amígdala. Se han detectado varios conjuntos de genes asociados a las variables respuesta de interés. En conjunto, la mayoría de los genes detectados, están implicados en procesos de estructuración y comunicación celular, en procesos de transmisión sináptica, en distintas vías de transducción de señal celular y en procesos de destrucción celular (apoptosis).También se han detectado genes implicados en diversas patologías cerebrales y algunos relacionados directamente con el Alzheimer.

Conclusión: Se han detectado genes con los tres métodos de análisis propuestos y, algunos, están asociados a la enfermedad Neurodegenerativa del Alzheimer.

Abstract:

Abstract: Alzheimer's is the most common dementia. Mainly a progressive cognitive and neurodegenerative loss characcteriza Alzheimer´s disease. Epidemiologically, Alzheimer represents a problem since in the middle of the 21st century it will constitute a socio-sanitary problem and it will be an index of mortality, so it is essential to be able to determine the etiology of the disease.

Methods: In order to molecularly assess Alzheimer's disease, the following study was carried out, selecting from Gene Expression Omnibus (from dataset GSE84422) six tissues from different cortical and subcortical regions, pathognomonicly related to Alzheimer's. The samples belong to 125 individuals who presented different degrees of affectation according to the CERAD criteria.The genes that have differential gene expression were determined to those tissues whose change is associated with the different degrees of dementia and with the different neuropathological lesions detected in patients diagnosed with Alzheimer. Those genes with the 1% of the highest standard deviations were classified with respect to the signal intensity detected in the microarray under study; and by means of Penalized Regression Methods, genes associated to the response variables of interest were selected. The biological significance was determined to those genes detected, as well as the biological routes in which they are enriched (Enrichment Analysis).

Results: From the six tissues analyzed, four have differential expression. The same number of genes (1% of the highest standard deviations) had been obtained in all tissues except for the Amygdala tissue. Several sets of genes associated with the variables response of interest had been detected. Overall, most of the genes detected are involved in cell structure and communication processes, in processes of synaptic transmission, in different cellular signal transduction pathways and in processes of cellular destruction (apoptosis). Also, genes involved have been detected in various brain pathologies and some ones directly related to Alzheimer.

Conclusions: Genes have been detected with the three methods of analysis proposed and, some others, are associated with the neurodegenerative disease of Alzheimer.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Tal y como indica la Fundación española de enfermedades neurológicas, un tema de interés social y científico actual son las enfermedades neurodegenerativas conocidas como demencias. Éstas son «un síndrome clínico caracterizado por un deterioro cognitivo persistente y progresivo así como por trastornos conductuales. Afectan a funciones cerebrales superiores como la memoria, el lenguaje, la orientación o percepción espacial entre otras»[8]

«Para poder caracterizar la importancia de estas patologías es necesario conocer cuál es el patrón de aparición de la enfermedad en la población por lo que hay que determinar cuál es la incidencia y la prevalencia de la misma. En cuanto a la incidencia se determina el número de casos nuevos que aparecen en un período de tiempo (por ejemplo rangos de edad) mientras que la prevalencia se refiere al número total (o porcentaje) de casos totales (nuevos y antiguos) detectados en el mismo intervalo de tiempo»[8]

En el estudio realizado por la Fundación española de enfermedades neurológicas [8] «En la demencia se ha visto que tanto la incidencia como la prevalencia se incrementan, de manera exponencial, con la edad en la mayoría de los estudios realizados. Estimaciones de incidencia realizadas en demencia [35] proporcionan las siguientes cifras: se detectan de 5 a 10 nuevos casos por cada mil personas/año en el grupo de edad de 64 a 69 años mientras que para el grupo de edad de 80 a 84 años la incidencia sube hasta 60 nuevos casos por cada mil personas/año. Algunos estudios demuestran incluso que las mujeres por encima de 55 años tienen el doble de riesgo de padecer demencia respecto a los varones por que tienen una mayor expectativa de vida[29]»

«En cuanto a la prevalencia también se observa la misma tendencia. En general, entre 64-69 años se sitúa por debajo del 2% mientras que en el grupo entre 80-84 años la prevalencia se sitúa en el 17% [8] . Esta prevalencia se ha observado, a nivel mundial, en Europa, Sudamérica, Norteamérica y Asia oriental [11] . Y, en España, la prevalencia está, para el grupo de edad de 60-65 años, entre el 5% y el 16% [24] »

Al analizar la prevalencia por tipos de demencia se ve que el Alzheimer es la forma de demencia más frecuente (60-80%) y, a día de hoy, es incurable y terminal [22]. Se da más frecuentemente en personas de edad avanzada (mayores a 65 años) aunque hay casos que se dan a partir de los 40 años e incluso hay casos que desarrollan la enfermedad de manera temprana. Epidemiológicamente la incidencia en el Alzheimer de nuevos casos por cada mil personas aumenta según aumenta la edad, por ejemplo, para mayores de 65 a 69 años la incidencia es de 3 nuevos casos por cada mil personas mientras que para mayores de 80 años la incidencia es de 23 nuevos casos por cada mil personas. También se ha observado una mayor incidencia entre pacientes del sexo femenino sobre todo a partir de los 85 años [2]»

«En segundo lugar aparece como más frecuente la demencia por cuerpos de Lewy (20%). A continuación se encuentra la demencia como consecuencia de tener una patología mixta: vascular y Alzheimer (15%) seguida de la demencia vascular (10%) [23]». «La demencia asociada a Parkinson representa un 4% y por último, entre las demencias menos frecuentes está la demencia frontotemporal y las demencias secundarias (1%) [8]». «Patogénicamente el Alzheimer se caracteriza por pérdida de neuronas y sinapsis en la corteza cerebral así como en ciertas regiones subcorticales. Ello se traduce en una atrofia de las regiones afectadas incluyendo degeneración en el lóbulo temporal y parietal así como partes de la corteza frontal y la circunvolución cingulada [43] »

Relevancia Socio-sanitaria del Alzheimer

«Teniendo en cuenta que la forma de demencia más frecuente es el Alzheimer, se centra este estudio en la misma. Es una enfermedad relevante ya que el progresivo envejecimiento de la población hace prever un incremento subsiguiente de la demencia en general y del Alzheimer en particular. En España se calcula que en el año 2050 uno de cada

tres españoles tendrá más de 65 años y se estima que cerca de un millón de personas padecerán demencia. Incluso se cree que esta estimación está infravalorada ya que hay muchos enfermos se quedan sin diagnosticar [11] »

«En EEUU se espera que la cifra de pacientes con demencia (Alzheimer) esté entre los 11-16 millones de enfermos en el año 2050 [8]. Y a nivel mundial se calcula que habrá, en 2050, 113 millones de personas afectadas de Alzheimer [8]». «Si se cumplen estas previsiones se está ante una auténtica epidemia lo que supone una enorme carga tanto a nivel social como económico como sanitaria. Y afectará tanto al entorno de los individuos enfermos como a la sociedad y a los gobiernos. De hecho la OMS ha instado a los distintos gobiernos a tomar medidas que puedan reducir el impacto socio-sanitario asociado a esta patología.»

«Otro aspecto importante que justifica el esfuerzo a realizar para estudiar el Alzheimer es la comorbilidad asociada a dicha patología. Destaca la prevalencia de factores de riesgo vascular (como la hipertensión y la diabetes mellitus) y los problemas derivados de los mismos a todos los niveles (cerebral, cardiaco y periférico) [37]»

«El deterioro cognitivo y funcional que sufren estos pacientes conlleva riesgos de caídas así como pérdida de movilidad. Y en etapas más avanzadas de la enfermedad los pacientes sufren postración[25]. Un dato relevante es los efectos secundarios de la medicación a la que están expuestos estos pacientes (polifarma) como por ejemplo los síntomas gastrointestinales»

«Además la demencia es uno de los principales predictores de mortalidad situada al nivel de otras enfermedades frecuentes como el cáncer o las enfermedades vasculares [41]. Se estima que el Alzheimer es la responsable de la muerte de aproximadamente el 5% de las personas mayores de 65 años. Este riesgo aumenta hasta un 30% en varones mayores de 85 años y hasta un 50% en mujeres mayores de 85 años [1] »

Por definición son enfermedades crónicas que provocan una mayor dependencia ya que conlleva una pérdida de la capacidad funcional del individuo. La atención a estos pacientes es compleja y requiere el apoyo de la familia, de los médicos y de los servicios sociales. Sería necesario un diagnóstico precoz así como un tratamiento multidisciplinar que incluya al enfermo y a los cuidadores para intentar reducir la carga sanitaria, social y económica de las demencias [8]. Por lo tanto, hay distintos aspectos científicos, éticos, sociales, económicos, sanitarios, etc., que justifica sobradamente el interés en estudiar y poder describir mejor esta patología. E igualmente de importante es poder encontrar fármacos que inhiban o curen la enfermedad.

Resúmen de lo que se sabe desde un punto de vista científico, a día de hoy, del Alzheimer

Etiológicamente las causas del alzheimer no se han descubierto completamente. Pero se han establecido varias hipótesis para describir la enfermedad. Se podría decir que son distintas maneras de explicar la degeneración de las neuronas en estos pacientes. Las hipótesis más recientes son que se deben a trastornos metabólicos y a acumulación de las proteínas beta-amiloide y tau. Hipótesis establecidas tras el análisis realizado en las autopsias del tejido neuronal de estos pacientes[16] .

En cuanto a los trastornos metabólicos se ha asociado con la hiperglucemia y la resistencia a la insulina. Se ha visto que hay receptores de insulina en las células del sistema nervioso central (concretamente en el hipocampo). Cuando la insulina se une a su receptor celular se promueve una cascada de señalización intracelular que produce cambios en la expresión de genes, por un lado relacionados con los procesos de plasticidad sináptica y, por otro lado de enzimas relacionadas con la liberación de la insulina que está unida al receptor y relacionadas con la proteína beta amiloide[4].

Según la segunda hipótesis, la acumulación de proteínas: se han observado acumulaciones de proteínas intracelulares (proteínas tau) que forman los denominados ovillos neurofibrilares. Otros acúmulos que se dan son depósitos extracelulares de proteínas también conocidas como placas seniles. Uno de los componentes de estas placas seniles es el péptido Beta-amiloide que se forma a partir de la degradación de una proteína precursora más grande

(Proteína precursora amieloide, PPA) por parte de enzimas denominadas alfa, beta y gamma secretasa. En concreto la enzima gamma secretasa cataliza la formación del péptido beta-amiloide que es el que tiene mayor significado desde un punto de vista médico. La γ-secretasa depende a su vez de las presenilinas (PSEN) [36] .

Desde un punto de vista genético (desde una perspectiva evolucionista), a día de hoy, se ha relacionado el Alzheimer de inicio temprano con mutaciones en el gen PPA (cromosoma 21) y en los genes PSEN1 y PSEN2 (cromosoma 14 y cromosoma 1, respectivamente)[3]. La proteína PPA (Proteína Precursora de Amiloide) es indispensable para el crecimiento de las neuronas, para su supervivencia y para su reparación si sufren daño. Si esta proteína se divide en fragmentos de menor tamaño por las secretasas (como ya se ha mencionado) se forman los beta-amiloides que se agrupan y depositan fuera de las neuronas constituyendo las placas seniles[28].

Por otra parte, se ha observado que las mutaciones del gen PSEN1 son las responsables de la aparición del Alzheimer de inicio temprano (como a los 23 años) y el Alzheimer de inicio tardío se relaciona con mutaciones en el gen de la apolipoproteína E (APOE). En concreto la mutación en el gen de la APOE4 (asociada con la hiperlipoproteinemia y la hipercolesterolemia familiar) se considera un factor de riesgo ya que tiende a producir acumulación amiloide en el cerebro antes de que aparezcan los primeros síntomas del Alzheimer [30].

Como se ha referido anteriormente, entre otros efectos, se observa un aumento de la concentración de la proteína Beta-amiloide. Este acúmulo parece ser el responsable de la alteración de la homeostasis del ión calcio intracelular induciendo la apoptosis. Se sabe además que la acumulación se realiza de manera selectiva en las mitocondrias de las neuronas y que son capaces de inhibir ciertas funciones enzimáticas [9][45] Aunque la beta-amiloide forma agregados en la sustancia gris del cerebro la pérdida de la función neuronal también se asocia a la formación de ovillos de neurofibrillas intracelulares debido a que se acumulan proteínas tau hiperfosforiladas. Estas proteínas adoptan formas anómalas produciendo la desintegración de los microtúbulos colapsando el sistema de transporte de la neurona. Ello produce una disfunción en la comunicación bioquímica entre neuronas lo que conlleva a la muerte celular.[19]

Se han hecho progresos significativos para intentar comprender la neuropatología, neuroanatomía, neuroquímica y anormalidades de la biología molecular, en los cerebros de los enfermos con Alzheimer [17]; [32]; [38]. De hecho numerosos estudios neurobiológicos han identificado aquellas regiones cerebrales vulnerables en el Alzheimer [7]; [27]; [42].

Línea de investigación a seguir

«Básicamente se han establecido dos hipótesis que pueden explicar la degeneración neuronal en los pacientes con Alzheimer, y que, además, se apoya en términos de biología evolucionista. En síntesis: un primer grupo constituido por aquellas demencias degenerativas que se producen por alteraciones genéticas puntuales (que ocurren al azar en el individuo fundador) y que se transmiten ininterrumpidamente al evitar la selección natural. Y un segundo grupo de causas genéticas complejas como el cambio de expresión o regulación de varios genes. Es decir, existen genes relacionados con el desarrollo del Alzheimer temprano y el Alzheimer tardío» [40]

«Ahora bien es difícil establecer un diagnóstico diferencial con el resto de demencias ya que todas cursan con trastornos cognitivos progresivos así como con trastornos conductuales. De hecho el diagnóstico de estos enfermos se establece en dos niveles: posible o probable. Y sólo se establece confirmación de la enfermedad mediante análisis histológico postmorten. Las herramientas hasta ahora utilizadas para el diagnóstico se basan en determinar ciertas características neurológicas y neuropsicológicas así como en la ausencia de un diagnóstico alternativo. Se apoya en el scaner cerebral y se están desarrollando técnicas basadas en procesamiento de señales electroencefalográficas. Y, aunque se ha mostrado fiabilidad y validez estadística entre los criterios diagnósticos y la confirmación histológica definitiva, se están estableciendo otros objetivos de diagnóstico. Por ejemplo, recientemente se

está realizando análisis de líquido cefalorraquídeo en busca de amiloides beta o proteínas tau. Y, recordemos, la mutación en el gen de la APOE4 es considerado un factor de riesgo para padecer Alzheimer.»[40]

En definitiva se necesitan nuevas propuestas para cambiar y mejorar los criterios diagnósticos de estos pacientes. En otras palabras, es necesario poder disponer de algún "biomarcador" que nos indique si una persona puede o no padecer Alzheimer ya que, como hemos visto, es una enfermedad devastadora y que afecta a todos los ámbitos de la vida de una persona.

El advenimiento de técnicas de análisis de alto rendimiento, como los Microarrays de expresión, han permitido poder estudiar la expresión de varios genes simultáneamente y así definir el Alzheimer asociado a los cambios observados en numerosas rutas biológicas[5]; [10]; [14]; [44].

Es muy importante comprender los cambios de expresión génica que se produce en las diferentes áreas del cerebro en las distintas etapas del Alzheimer; sobre todo porque dichas etapas de la enfermedad se definen según criterios neuropatológicos o por criterios funcionales y cognitivos. Por ello, en este TFM (Trabajo Fin de Máster) se hará:

- Una selección de genes expresados diferencialmente en distintos tipos de tejido nervioso de pacientes afectados (en distinto grado) y se utilizarán aquellas regiones de tejido nervioso que han sido relacionadas patognomónicamente con el Alzheimer.

- Se estudiará cómo se agrupan los genes expresados en los distintos tipos de tejido nervioso de pacientes afectados por Alzheimer.

- Se identificará el significado biológico de los genes seleccionados y/o clasificados.

-Se seleccionarán polimorfismos (snps), si los hay, relacionados con una variable respuesta (clínica o anatomohistopatológica) de interés.

- Se identificará a qué gen pertenece dicho polimorfismo y cuál es su función biológica. Identificar, si es posible, en qué patologías relacionadas con las demencias ha sido descrito dicho gen/es.

- Se establecerá la región genómica a la que pertenecen todos los genes obtenidos.

1.2 Objetivos del Trabajo

1.2.1. OBJETIVOS GENERALES

1.2.1.1 Determinar el nivel de expresión génica en distintos tipos de tejido nervioso de pacientes afectados por Alzheimer

1.2.1.2 Detectar polimorfismos (Snps) asociados al Alzheimer

1.2.1.3 Identificar a qué región genómica pertenecen los genes seleccionados e identificados en los apartados anteriores

1.2.2 OBJETIVOS ESPECÍFICOS

*** Para determinar el nivel de expresión génica en distintos tipos de tejido nervioso**

Selección de genes diferencialmente expresados

1.2.2.1.1 Identificar genes con niveles de RNA diferentes en los grupos experimentales (pacientes con Alzheimer) mediante análisis de datos de microarrays de expresión.

1.2.2.1.2 Anotar los genes seleccionados mediante expresión diferencial y visualizar los perfiles de expresión

1.2.2.1.3 Analizar la significación biológica de los genes seleccionados por expresión diferencial

1.2.2.1.4 Interpretar los resultados: identificar aquellas categorías con significación biológica detectadas, que están significativamente enriquecidas (Enrichment Analysis).

Descubrir clases o grupos de genes en datos de microarrays

1.2.2.1.5 Clasificar genes de datos de microarrays de distintos tipos de tejido nervioso de pacientes con Alzheimer

1.2.2.1.6. Analizar la significación biológica de los genes agrupados

1.2.2.1.7. Interpretar los resultados: identificar si los genes agrupados pertenecen a la misma categoría de significacón biológica y si están significativamente enriquecidas (Enrichment Analysis).

*** Para detectar polimorfismos (Snps) asociados al Alzheimer**

1.2.2.2.1 Seleccionar e identificar los polimorfismos (Snps), si los hay, que están relacionados con la variable clínica respuesta de interés.

1.2.2.2.2. Valorar cuál es el porcentaje de variabilidad (de la variable clínica respuesta de interés) explicada por los polimorfismos (Snps), seleccionados, en su conjunto.

1.2.2.2.3. Determinar si cada uno de los polimorfismos (Snps) seleccionados presentan asociación significativa con la variable clínica respuesta analizada.

1.2.2.2.4. Comprobar en qué genes están situados los polimorfismos (Snps) seleccionados que presentan asociación significativa con la variable clínica respuesta analizada.

1.2.2.2.5. Analizar la categoría de significación biológica de los nuevos genes identificados y analizar si están significativamente enriquecidas (Enrichment Analysis).

*** Identificar la región genómica**

2.2.3.1 Posicionar en el genoma (de referencia) los genes seleccionados, agrupados y detectados.

1.3 Enfoque y método seguido

Hay muchos métodos de análisis implementados en software comerciales (por ejemplo Galaxy) así como herramientas desarrolladas, de código abierto y disponible, para realizar análisis de expresión diferencial. Destacamos el análisis de NGS (RNAseq) y de Microarrays de expresión (tanto de plataformas Affymetrix como plataformas Illumina) utilizando el softwareR. Éste es un entorno de software libre para análisis y gráficos estadísticos.

Si se comparan los pipeline (básicos) de trabajo de RNAseq versus Microarrays de expresión se tiene que en ambos casos hay una parte experimental que difiere una de otra. Al final se obtiene: en el RNAseq los niveles de expresión de los transcritos cuyos valores se dan en forma de “cuenta” (constituye una variable de valores discretos, no continuos) mientras que en los Microarrays, los niveles de expresión de los transcritos, lo valores obtenidos constituyen una variable continua. Y, además, en ambos casos se puede realizar un análisis estadístico que nos permite obtener la expresión diferencial de los transcritos analizados (y, por tanto, del gen correspondiente) así como determinar la significación biológica de los mismos.

El software R incluye también paquetes que permiten descubrir clases o agrupaciones de genes así como implementar métodos (técnicas de Data Mining) que permiten desarrollar modelos matemáticos lineales para detectar polimorfismos (Snps).

Ambas tecnologías experimentales tienen ventajas e inconvenientes desde el punto de vista de análisis de datos. Para este TFM, se ha seleccionado el análisis de Microarrays Affymetrix por varias razones.

En primer lugar porque los datos seleccionados (una buena colección de dataset, unas 2004 muestras, de distintos tipos de tejido nervioso de pacientes con Alzheimer) fueron analizados utilizando Microarrays de Expresión de Affymetrix.

En segundo lugar porque los pipelines de análisis para Microarrays de Affymetrix está muy contrastado y existen muchos paquetes en R que permiten abordar todas las etapas de análisis que se quieren realizar.

En tercer lugar porque los datos seleccionados, de la variable/s clínica/s de interés, se cuantificaron utilizando sistemas de medida objetivos, siendo los valores mayoritariamente continuos (con un rango de valores). Ello permite integrar datos del mismo tipo (variables continuas) en un solo dataset: datos que provienen de la intensidad medida en los microarrays junto a los datos seleccionados de las distintas variables clínicas. Esto facilita mucho el análisis, simplemente por cuestiones técnicas.

Otra razón para seleccionar microarrays de expresión para realizar el análisis es porque el segundo bloque central de análisis que se desea realizar, seleccionar polimorfismos mediante métodos de regresión, requiere que la variable respuesta sea continua.

Una razón de orden práctico es que se usa un solo software para todos los análisis, R, que además es flexible para seleccionar aquello que realmente interesa además de poder obtener resultados gráficos con los que poder también comparar en las distintas etapas del análisis.

R, además, permite obtener el posicionamiento genómico de los Snps obtenidos (genes). Por último, para obtener información de los Snps (Single Nucleotide Polymorphims) seleccionados se pueden consultar bases de datos curadas y contrastadas como GeneCard, OMIM, Decipher, VarScan. Prioritariamente se incluirá la información referida en GeneCard donde se establece la patología relacionada con el Snps seleccionado.

1.4 Planificación del Trabajo

1.4.1 *TAREAS*

* Obtención del conjunto de datos para realizar el análisis (del 19 al 25 Marzo): 24 horas

1.4.1.1 Seleccionar de Genome Omnibus Expression los datos que vamos a usar en el análisis: escoger aquellas partes del tejido nervioso relacionado patogénicamente con el Alzheimer en la literatura.

1.4.1.2 Explorar el diseño experimental que han utilizado para realizar este experimento: básicamente qué tipo de plataforma de análisis han utilizado, si hay réplicas biológicas y/o réplicas técnicas, si presenta pooling (combinación de mARN de diferentes casos en una única muestra), si se trata de arrays de un color o de dos colores.

1.4.1.3. Seleccionar los distintos Scripts en R para llevar a cabo el análisis de los objetivos propuestos. Adaptarlos y/o desarrollar, adecuar, otros nuevos que se irán probando según se ejecuten las distintas partes del análisis.

* Preprocesado general previo de los datos descargados de Genome Omnibus Expression (del 26 al 28 Marzo): 15 horas

1.4.1.4. Lectura (carga) de los datos seleccionados en el programa (R Studio), y con el script correspondiente, con el que haremos el análisis

1.4.1.5. Depurar los datos: Identificar y tratar si hay outliers, si hay valores missings

1.4.1.6. Integrar datos: elaborar la base de datos que integre los valores de los microarrays y los valores de la variable clínica de interés

1.4.1.7. Valorar si hay que transformar los datos: estandarizarlos, ponerlos en la escala adecuada

1.4.1.8. Valorar si hay que reducir los datos: si hay que tener una dimensión más reducida de los mismos

* Seleccionar genes diferencialmente expresados (del 2 al 8 de Abril): 36 horas

1.4.1.9. Exploración y control de calidad de los datos: lo haremos visualizando de manera gráfica los datos. Dependiendo del tipo de array que sea podremos usar, por ejemplo, histogramas y gráficos de densidad, gráfico de componentes principales, gráficos de densidad y “Maplots” .

1.4.1.10. Preprocesado de los datos: Utilizaremos el método robust multi-array average (RMA) de Bioconductor ya que es el método estándar para Affymetrix.Éste se realiza, a su vez, en varias etapas:

a) Filtraje: Eliminar la parte de la señal no atribuible a la expresión (eliminar spots erróneos) llamada background o ruido.

b) Normalizar los datos para hacerlos comparables entre muestras y eliminar posibles sesgos técnicos.

c) Al ser un array de Affymetrix resumiremos la información de un gen en un solo valor.

d) Valorar si es preciso realizar una transformación de los datos de forma que la escala sea razonable y facilite el análisis.

1.4.1.11. Seleccionar genes diferencialmente expresados: análisis basado en modelos lineales. Valorar las hipótesis de contraste que vamos a realizar.

1.4.1.12. Estimación del modelo y selección de los genes:Generar una lista de genes ordenados de más a menos diferencialmente expresados.Representar gráficamente (volcano plot) los genes diferencialmente expresados que han sido seleccionados. Controlar los posibles falsos positivos por distintos métodos: Bonferroni, Benjamini & Hochberg o Benjamini & Yekutieli.

1.4.1.13. Anotación de resultados: Analizar la significación biológica de los genes seleccionados, visualizar los perfiles de expresión (mediante Heatmap) y realizar análisis de enriquecimiento utilizando Bioconductor.

* Clasificar o agrupar genes de datos de microarrays de expresión (del 9 al 15 de Abril): 36 horas

1.4.1.14. Clasificar genes de datos de microarrays de expresión: análisis clúster basado en métodos no jerárquicos.

1.4.1.15. Representar gráficamente los genes clasificados (mediante Heatmap)

1.4.1.16. Anotación de resultados: Analizar la significación biológica de los genes clasificados y realizar análisis de enriquecimiento utilizando Bioconductor.

*Detectar polimorfismos (Snps) asociados al Alzheimer (del 16 al 22 de Abril): 36 horas

1.4.1.17. Seleccionar los polimorfismos (Snps): Modelar la relación entre un conjunto de variables predictoras (Snps) y una variable respuesta continua mediante métodos de regresión penalizados (métodos “shrinkage”).

1.4.1.18. Calcular el porcentaje de variabilidad (de la variable respuesta) explicada por el conjunto de los polimorfismos (Snps) seleccionados mediante métodos de Regresión Lineal Múltiple: Análisis Multivariante

1.4.1.19. Determinar si hay asociación significativa de cada polimorfismo (Snp) seleccionado con la variable respuesta mediante métodos de Regresión Lineal Simple: Análisis Univariante.

1.4.1.20. Comprobar en qué genes están situados los polimorfismos (Snps) seleccionados, que tienen asociación significativa con la variable clínica respuesta analizada mediante métodos de Genómica Computacional. O comprobar si están en zonas intergénicas o zonas de regulación.

1.4.1.21. Analizar la significación biológica de los nuevos genes identificados (ver categorías funcionales) y analizar si están significativamente enriquecidas (Enrichment Analysis) utilizando Bioconductor.

* Identificar a qué región genómica pertenecen los genes seleccionados, agrupados e identificados en los apartados anteriores (del 23 al 29 de Abril): 36 horas

1.4.1.22. Posicionar en el genoma (de referencia) los genes seleccionados, agrupados y detectados mediante métodos de Genómica Computacional: determinar en qué cromosoma están, cuáles son las coordenadas genómicas, etc. Si los Snps seleccionados están en zonas intergénicas o zonas de regulación, establecer su posición genómica, etc.

* Repasar y Organizar (del 3 al 6 de Mayo): 12 horas

1.4.1.23. Recopilar y anotar la bibliografía utilizada

1.4.1.24. Estructurar los resultados obtenidos y escribir las conclusiones

1.4.1.25. Recopilar y anotar los scripts utilizados para el análisis

* Hacer Borrador de la Memoria a presentar (del 7 al 20 de Mayo): 56 horas

1.4.1.26. Redactar Introducción, Materiales y Métodos (del 7 al 13 de Mayo)

1.4.1.27. Redactar Resultados, Discusión y Conclusiones (del 14 al 20 de Mayo)

1.4.1.28. Utilizar este período de tiempo para repetir algún análisis en caso de problemas técnicos o por que se detecten errores de edición en los documentos o para mejorar la redacción.

* Redactar la Memoria definitiva del TFM (del 21 de Mayo al 5 de Junio): 64 horas

1.4.1.29. Completar el resto de apartados de la Memoria (Abstract, Bibliografía, etc.) (del 21 al 27 de Mayo)

1.4.1.30. Revisar la redacción y estructura de la Memoria (del 28 de Mayo al 3 de Junio)

* Hacer la Presentación en Power Point (del 6 al 12 de Junio): 36 horas

* Defensa Pública (del 14 al 25 de Junio)

1.4.2 *CALENDARIO*

Ver diagrama (equivalente) de Gant como documento excel (Material Suplementario 1)

1.4.3. *HITOS*

1.4.3.1 Una de las partes más importantes es tener preparados los dataset que usaremos en los distintos tipos de análisis así como haber comprobado que los scripts de R que usaremos funcionan perfectamente. Por ello a 28 de Marzo de 2018 es esencial haber decidido qué datos usar, haberlos preparado y ver que funcionen los scripts.

1.4.3.2. La primera parte esencial de esta memoria es poder realizar el análisis de selección de genes diferencialmente expresados así como poder clasificar los genes estudiados en los microarrays de expresión. Esto debe estar a 15 de Abril de 2018.

1.4.3.3. La segunda parte esencial es poder detectar los polimorfismos (Snps) asociados a la variable clínica de interés. Esto debería estar a 22 de Abril 2018

1.4.3.4. Es importante en la descripción genética que queremos hacer, posicionar, en el genoma de referencia, los distintos genes obtenidos. Lo cual concluiríamos a 29 de Abril 2018.

1.4.3.5. A 20 de Mayo 2018, e independientemente de los resultados obtenidos, tendremos recopilados todos los datos que se necesita para redactar la memoria. Haremos un borrador previo con todos los apartados.

1.4.3.6. Como fase final es importante tener, a 3 Junio 2018, estructurada toda la Memoria, correctamente redactada y con la discusión y conclusión en función de los resultados obtenidos.

1.4.3.7. A 12 Junio 2018 esencial tener completada la presentación en Power Point

1.4.3.8. A partir del 13 Junio 2018 se preparará la Defensa Pública estableciendo el por qué del TFM, explicar el por qué se han seleccionado las técnicas de análisis elegidas y cuáles han sido los resultados. Establecer las conclusiones obtenidas en este TFM.

1.5 Breve sumario de productos obtenidos

A cada uno de los seis tejidos seleccionados se han realizado tres análisis diferentes para obtener dianas genéticas (Polimorfismos, Snps) relacionadas con el Alzheimer.

En el análisis de expresión diferencial se obtuvieron genes diferencialmente expresados en cuatro tejidos: tres de la corteza cerebral (en LF3 (Prefrontal Cortex), en LP(Superior Parietal Lobule), en LT3 (Superior Temporal Gyrus)) y el otro tejido de la zona subcortical, el SL1 (Amígdala).

En el análisis de descubrimiento de clases se obtienen genes en todos los tejidos ya que se decide seleccionar sólo aquellos genes que se encuentran entre el 1% de las desviaciones estándar más altas. Ello se debe a que en las medidas de intensidad obtenidas en el array de expresión pueden existir valores extremos, altos. Y, al calcular su desviación estándar, se ve que es más grande que la media lo cual indica, probablemente, un sesgo. Como dato relevante indicar que se obtiene el mismo número de genes en todos los tejidos analizados (223) excepto en el tejido Amígdala donde se obtienen 547 genes.

Por último, en el análisis de Regresión penalizada se obtienen distinto número de genes asociados a las dos variables respuestas definidas como variable global clínica y variable global anatomohistopatológica.

En cada tipo de análisis (y por tejido) se ha obtenido un informe (en algunos casos en formato pdf en otros casos en formato .html) y una tabla de anotación funcional (según análisis y por tejido; en formato .html) (Ver Tabla 1).

Tipo Tejido	Prefrontal Cortex	Superior Parietal Lobule	Superior Temporal Gyrus	Temporal Pole	Amígdala	Región Hipocampo
Tipo Análisis	***Informe/Anotación***	***Informe/Anotación***	***Informe/Anotación***	***Informe/Anotación***	***Informe/Anotación***	***Informe/Anotación***
Expresión Diferencial	Si/Si	Si/Si	Si/Si	Si/NO	Si/Si	Si/NO
Clases	Si/Si	Si/Si	Si/Si	Si/Si	Si/Si	Si/Si
Train y Test (clínica)	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO
Train y Test (anatomoh.)	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO
Obtención de Snps clinica	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO
Obtención de Snps anatomoh.	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO	Si/NO
A. Mult y A.Uni Clinica	Si/Si	NO/NO	Si/Si	Si/Si	Si/Si	Si/Si
A. Mult y A.Uni Anatomoh.	Si/Si	NO/NO	Si/Si	NO/NO	Si/Si	Si/Si

Tabla 1. Resumen de los Informes y Tablas de Anotaciones obtenidas en el trabajo

En total son 45 informes (unos en formato .html y otros en pdf) y 19 tablas de anotaciones (todas en formato .html).

1.6 Breve descripción de los otros capítulos de la memoria

Los otros capítulos esenciales de la memoria serán:

1.6.1 Materiales y Métodos donde se explica de dónde se obtienen los datos. Se hace la descripción de la expresión diferencial mediante arrays de expresión. Se explica el diseño experimental que se realiza así como el análisis estadístico de los datos seleccionado en los distintos tipos de análisis. También se hará una breve valoración del porqué de la selección de las variables respuesta utilizadas en la Regresión penalizada.

1.6.2 Resultados Se exponen básicamente, en forma de tablas y gráficos. Se estructura según método de análisis utilizado y según tipo de tejido.

1.6.3 Discusión donde se analizan los resultados obtenidos aunque no muy profusamente

2. Resto de capítulos

2.1 Materiales y Métodos

2.1.1 Materiales

Para obtener dianas genéticas (como polimorfismos (SNPs)) en genes asociados a la Enfermedad Neurodegenerativa del Alzheimer. se utiliza el conjunto de datos de Arrays de Expresión obtenidos con plataformas de Affymetrix publicados el 19 de Agosto de 2016 en GEO (Gene Expression Omnibus) con número GSE84422 bajo el título "Molecular Signatures Underlying Selective Regional Vulnerability to Alzheimer's Disease". Este conjunto de datos, tal y como se indica en la Query dataSet de GSE84422, está constituido por una amplia cohorte de pacientes. En concreto 125 personas que fueron diagnosticadas de Alzheimer en distintos grados.

Estos datos, además, están relacionados con el proyecto Biológico PRJNA329139 ya que se utilizan en la publicación [26]. De este artículo se obtiene el otro conjunto de datos (material suplementario del artículo) relacionados con los pacientes incluidos en GSE84422. Son la tabla S1 (datos demográficos, clínicos y anatomohistopatológicos de los pacientes) y la tabla S2 adaptada (explicación de los datos clínicos y anatomohistopatológicos de la tabla S1 (se adjuntan como Material Suplementario 2).

En dicho artículo [26] se indica que en el diagnóstico se utilizaron dos criterios: la Consortium to Establish a Registry for Alzheimer's Disease (CERAD) y la clinical dementia raiting (CDR) dando lugar a dos variables. En la variable CDR (clinical dementia raiting) los pacientes son diagnosticados y se clasifican en una de las siguientes clases: 0(no demencia), 0.5 (demencia cuestionable), 1(demencia leve), 2(demencia moderada), 3(demencia severa), 4(demencia profunda) y 5(demencia terminal) (se adjunta como Material Suplementario 3 el formulario de la CDR en España). Con la otra variable, la CERAD (Consortium to Establish a Registry for Alzheimer's Disease), se mide la categoría neuropatológica en la que se clasifican los pacientes [13] Así, en el artículo [26] se definen las siguientes categorías neuropatológicas: 1(Normal), 2(Definitiva), 3(Posible) y 4(Probable).

Además, en las tablas S1 y S2 del material suplementario 3, se resumen los distintos hallazgos anatomohistopatológicos encontrados postmorten en los 125 pacientes. Y se resumen en las siguientes variables: braak (neurofibrillas enredadas, le dan un score), PLQ_Mn (promedio de las medidas de densidad de la placa neurítica analizada en 5 regiones cerebrales), NprSum (suma de los scores de calificación de Cerad en distintas áreas cerebrales) y NTrSum (suma de densidad de ovillos neurofibrilares en distintas Áreas).

Los microarrays de expresión (de los que se obtienen los datos con los que se trabaja) se realizaron con el RNA extraído (postmorten) de 19 regiones cerebrales distintas. En este trabajo se seleccionaron en un principio doce tejidos que corresponden a cada una de las regiones cerebrales relacionadas patognómonicamente con el Alzheimer. Los doce tejidos seleccionados (a los que por razones metodológicas se les asignó una abreviatura consignada entre paréntesis) se pueden ver en la siguiente Tabla 2.

Lóbulo Frontal	**Lóbulo Parietal**	**Lóbulo Temporal**	**Sistema Límbico**
Frontal Pole **(LF1)**	Superior Parietal Lobule **(LP)**	Inferior Temporal Gyrus **(LT1)**	Amígdala **(SL1)**
Inferior Frontal Gyrus**(LF2)**		Middle Temporal Gyrus **(LT2)**	Posterior Cingulate Cortex **(SL2)**
Prefrontal Cortex **(LF3)**		Superior Temporal Gyrus **(LT3)**	Anterior Cingulate Cortex **(SL3)**
		Temporal Pole **(LT4)**	Región Hipocampo **(SL4)**

Tabla 2 Abreviatura consignada a los Tejidos seleccionados

Por imponderables espacio-temporales los tejidos seleccionados y analizados se redujeron a seis: Prefrontal Cortex (LF3), Superior Parietal Lobule (LP), Superior Temporal Gyrus (LT3), Temporal Pole (LT4), Amígdala (SL1) y Región Hipocampo (SL4).

2.1.2 Métodos

2.1.2.1 Métodos Generales: plataformas de análisis utilizadas y valoración de la calidad de los datos obtenidos

Tal y como se indica en la Query dataSet de GSE84422, y que está referido en el artículo [26], las muestras de RNA recolectadas fueron trabajadas sobre varias plataformas de Microarrays de Affymetrix para obtener el perfil de expresión de dichas muestras. En concreto se usaron las plataformas de Genoma Humano GPL96 (HG-U133A), GPL97 (HG-U133B) y GPL570 (HG-U133_Plus_2). Se contrastan las plataformas GPL96 y GPL97 donde no se observan grandes diferencias. Dado que hay más tejidos analizados sobre la plataforma GPL96, se selecciona la misma para los tejidos Prefrontal Córtex (LF3), Superior Parietal Lobule (LP), Superior Temporal Gyrus (LT3), Temporal Pole (LT4) y Región Hipocampo (SL4) excepto el tejido Amígdala (SL1) que sólo fue trabajado en la plataforma GPL570.

Los datos seleccionados se obtienen de array de un color de Affymetrix cuyo archivos tienen extensión .CEL (archivo en formato binario que sólo puede ser leído con programas específicos). Los archivos .CEL, por cada microarray chip, contiene los valores de intensidad para cada sonda. A partir de las intensidades de los archivos .CEL se genera una matriz de expresión que contiene una columna por chip con los valores de intensidad absolutos y una fila por grupo de sondas.

Se evalúa la calidad de los datos de intensidad brutos (Rawdata) mediante diferentes tipos de gráficos utilizando el método del paquete “Affy” implementado en R/Bioconductor. Tal y como se indica en [33] se usan los siguientes gráficos: Gráfico de densidad de las intensidades en los distintos arrays (para ver si son similares o diferentes los arrays), boxplot de cómo es la distribución de la intensidad, gráfico de componentes principales de las intensidades de los arrays para ver cómo se agrupan las muestras en función de la intensidad de señal. También se realizan gráficos de agrupación (clúster jerárquico) para ver si hay agrupación de las muestras en función de la intensidad de señal. Y con los gráficos de degradación se valora la calidad de hibridación del ARN a lo largo de los conjuntos de sondas

En la evaluación de la calidad de los datos [33] es importante determinar si el array es problemático o no de cara a posteriores análisis de datos. Para ello se realiza un modelo de análisis de datos no normalizados, utilizando el método del paquete “affyPLM” implementado en R/Bioconductor. Son modelos de análisis de bajo nivel (probe-level-models o PLM) y que «ajustan los valores de intensidad, es decir, a nivel de sondas. Estos modelos dan unos valores estimados que se comparan con los valores reales. De esta forma se obtienen los errores o residuos del ajuste. Al analizar dichos residuos (de manera similar a como se hace con un modelo de regresión) se observa: o que no presentan ningún patrón especial con lo que se puede suponer que el modelo se ajusta relativamente bien o se observan desviaciones con lo que el modelo no explica bien las observaciones. Esto último se atribuirá a la existencia de algún problema con los datos»

El método utilizado en este trabajo,[33], consiste en calcular dos medidas de error:RLE (relative log expression) y NUSE (Normalized Unscaled Standard Errors). RLE es una medida estandarizada de la expresión, «no es de gran utilidad pero debería presentar una distribución similar en todos los arrays» Sin embargo NUSE «representa la distribución de los residuos (valores estimados de los errores en los modelos de bajo nivel, PLM).Si el array es problemático se observa la caja correspondiente en el boxplot desplazada hacia arriba o hacia abajo. Por el contrario, tal y como se indica en [33] si los datos son de calidad, ambos gráficos deben ser centrados y relativamente simétricos”

2.1.2.2 Análisis de Expresión Diferencial

En este trabajo se utiliza un modelo lineal de análisis para identificar genes diferencialmente expresados entre grupos de enfermos (que padecen Alzheimer) establecidos según el criterio CERAD (grado de neuropatología). Para ellos se utiliza el paquete "Limma" [15] implementado en R/Bioconductor.

Un esquema de trabajo para expresión diferencial se puede ver en la Figura 1 basado en la figura de Proceso de Análisis de Microarrays de [33], se da por hecho que la parte experimental ha ido bien y los primeros pasos son comunes al resto de métodos que se van a usar en este trabajo.

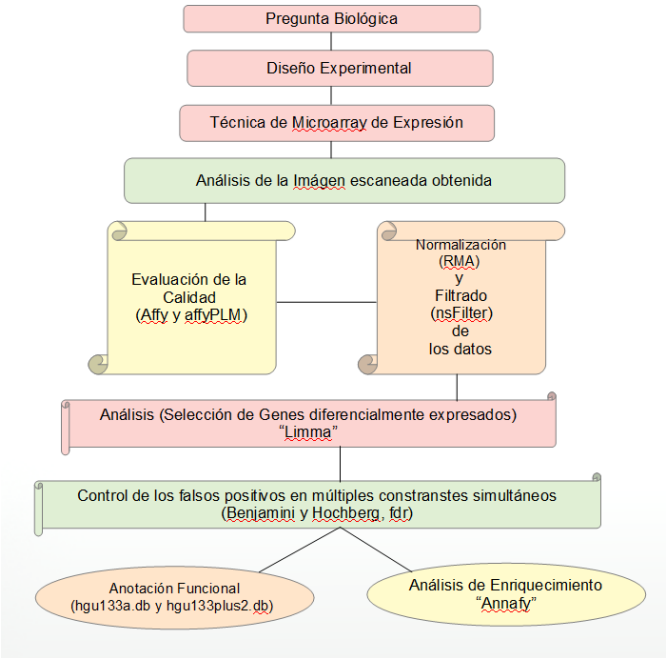


Figura 1 Esquema de trabajo para Expresión Diferencial de Genes

Una vez valorada la calidad de los datos brutos (Rawdata) hay que realizar una etapa denominada Preprocesado donde se Normalizan y Filtran los datos. Es una etapa muy importante ya que se obtiene el conjunto de datos Normalizados y Filtrados con los que se realizarán todos los demás análisis.

«La normalización es un conjunto de técnicas utilizadas para transformar adecuadamente los datos antes de que sean analizados. El objetivo es corregir las diferencias sistemáticas entre muestras (en la misma imagen o entre imágenes). Ello es importante ya que dicha diferencia no es una verdadera variación en la expresión entre muestras biológicas. Es lo que se denomina como diferencias sistemáticas»[33]

«En los Arrays de Affymetrix, tras escanear la imagen, se obtienen una serie de valores de intensidad de cada elemento del chip. En este tipo de arrays se sabe que cada valor no corresponde a la expresión del gen ya que hay múltiples sondas (probes) por cada gen y que originan un conjunto de sondas iguales o probeset. Además, cada grupo de sondas consiste en múltiples pares de sonda, donde cada una puede tener dos elementos: un perfect match que coincide exactamente con el fragmento del gen al que corresponde la sonda (PM) y un mismatch que coincide con el fragmento del gen excepto en el valor central que se ha sustituido por el nucleótido complementario (ello se hace para valorar la hibridación no específica). Actualmente, en las nuevas versiones de microarrays no se usan los mismatch»[33]

Tal y como se estable en [33], para la normalización se puede usar el método RMA(robust multi-array average) del paquete “Affy” implementado en R/Bioconductor. Con ello se convierten las señales individuales en valores de expresión normalizados para cada gen.

El método RMA (robust multi-array average) se fundamenta en la modelización de las intensidades de las sondas lo cual se basa en los distintos valores de la misma sonda entre todos los arrays disponibles.

Los pasos en los que se realiza este método son:[33]

- "Ajuste del ruido de fondo (background) basándose solo en los valores PM (perfect match)"
- "Toma logaritmos en base 2 de cada intensidad ajustada por el background"
- "Realiza una normalización por cuantiles de los valores del paso anterior, sustituye cada valor individual por los promedios de las distribuciones de los valores ordenados de cada array"
- "Estima las intensidades de cada gen separadamente para cada conjunto de sondas. Realiza una técnica similar a una regresión robusta denominada median polish sobre una matriz de datos que tiene los arrays en filas y los grupos de sondas en columnas"

Siguiendo la metodología de análisis de Microarrays estudiada [33] se observa si los datos han sido normalizados o no mediante un gráfico boxplot del objeto raw_data que se genera.

«Pero también es importante realizar un filtraje no específico. Ello es recomendable para eliminar el ruido de fondo y limitar los ajustes posteriores a los necesarios. Con este filtraje se eliminan por un lado los spots cuyas imágenes o señales sean erróneas disminuyendo el ruido de fondo, por otro los spots con señales muy bajas (o que no ha habido hibridación). También se eliminan genes que no presentan variación significativa entre las distintas condiciones experimentales. A veces, se pueden eliminar spots informativos. Por ello se usa la función "nsFilter" que permite eliminar los genes que, o bien varían poco, o bien no se dispone de anotación para ellos. La función nsFilter devuelve los valores filtrados en un objeto expression_Set y un informe de los resultados del filtraje.» [33]

«En la selección diferencial propiamente dicha se buscan aquellos genes que presentan una expresión diferencial entre dos o más condiciones experimentales. En este trabajo se aplicará la aproximación presentada por Smyth [15] y que está basada en la utilización del modelo lineal general, combinada con un método para obtener una estimación mejorada de la varianza.»[33]

Para su desarrollo se ha establecido el siguiente diseño experimental. El objetivo del experimento es encontrar qué genes se expresan diferencialmente en los pacientes que presentan alguno de los cuatro tipos de neuropatología. Teniendo en cuenta que el Microarray de expresión es Affymetrix de un color se tiene que cada muestra ha sido hibridada de manera independiente en cada array. Ello permite hacer comparaciones directas entre muestras y todas las comparaciones son igual de eficientes.

Formalmente, y utilizando la nomenclatura establecida en [33] se tiene, en general, el siguiente modelo lineal:

$$y = X\alpha + \varepsilon,$$

y contrastes:

$$C'\beta$$

«En otras palabras, en los análisis basados en modelos lineales el primer paso para el mismo es crear la matriz de diseño. Y, dado un modelo lineal definido a través de una matriz de diseño, pueden formularse las preguntas de interés como contrastes de hipótesis, es decir, en hacer comparaciones entre los parámetros del modelo»[33]. Formalmente queda establecido como:

Parámetros del modelo

$\alpha_1 = E(\log(M_1/W))$
$\alpha_2 = E(\log(M_2/W))$
$\alpha_3 = E(\log(M_3/W))$
$\alpha_4 = E(\log(M_4/W))$

Ejemplo de Matriz de Diseño

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1010 \\ 0011 \\ 1001 \\ 1100 \\ 0101 \\ 1100 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

La matriz de diseño se elabora en función del tejido que se está analizando. Cada tejido contiene X muestras (una por paciente) y que representa, cada una, la condición neuropatológica establecida en el artículo [26] (Material Suplementario 4, tablas realizadas con distintos datos de los pacientes y que se utilizan para crear la matriz de diseño y el documento target).

Los Contrastes (hipótesis) a realizar en este estudio son: Se valora la expresión diferencial génica entre las cuatro categorías neuropatológicas: A vs B (Normal vs Definitivo), A vs C (Normal vs Posible), A vs D (Normal vs Probable), B vs C (Definitivo vs Posible), B vs D (Definitivo vs Probable) y C vs D (Posible vs Probable). Con lo que se obtienen seis contrastes. Formalmente se puede expresar como:

Matriz de Contraste

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 1010 \\ 0011 \\ 1001 \\ 1100 \\ 0011 \\ 0101 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix}$$

Contrastes realizados

$\beta_1 = \alpha_1 - \alpha_2$
$\beta_2 = \alpha_1 - \alpha_3$
$\beta_3 = \alpha_1 - \alpha_4$
$\beta_4 = \alpha_2 - \alpha_3$
$\beta_5 = \alpha_2 - \alpha_4$
$\beta_6 = \alpha_3 - \alpha_4$

El método de estimación que se usa en este trabajo es Limma que utiliza modelos de Bayes empíricos para combinar la información de toda la matriz de datos y de cada gen individual y así obtener estimaciones de error mejoradas. En otras palabras, se usan métodos de ajuste lineal. Con este análisis se obtienen los estadísticos como Fold-change tmoderados o p-valores ajustados que se utilizan para ordenar los genes de mayor a menor diferencialmente expresados. Lo que hace es “calcular un odds-ratio que viene a ser la razón entre probabilidad de que un gen esté diferencialmente expresado frente a la de que no lo esté y se asocia este valor, denominado estadístico B, con un estadístico t moderado y su p-valor” expresado tal cual en [33]

Además con este método lineal se controlan los falsos positivos que pueden resultar del alto número de contrastes realizados simultáneamente. Para controlar la tasa de falsos positivos los p-valores se ajustan con el método Benjamini y Hochberg (fdr o false discovery rate). Con la función topTable se genera para cada contraste una lista de genes ordenados de más a menos diferencialmente expresados. Los resultados se visualizan utilizando un volcano plot. En abscisas se representan los cambios de expresión en escala logarítmica, y en ordenadas el -log (p-valor) o el estadístico B (datos que se obtienen tras hacer el ajuste lineal). "Cuanto más arriba y más hacia el exterior se encuentre el gen, más posibilidades tiene de estar diferencialmente expresado" [33]

«Sin embargo, tal y como establecen, en [33], cuando se realizan varias comparaciones a la vez puede resultar importante ver qué genes cambian simultáneamente en más de una comparación. Si el número de comparaciones es alto, también puede ser necesario realizar un ajuste de p-valores entre las comparaciones (entre condiciones), distinto del realizado entre genes. Para ello se usa la función "decidetests" (del paquete limma implementado en R/Bioconductor) con la que se seleccionan los genes que cambian en una o más condiciones. Se obtiene así una tabla (res) que contiene para cada gen y para cada comparación un 1 (si el gen está sobreexpresado, up), un 0 (si no hay cambio significativo) o un

-1(si el gen está infraexpresados, down regulado). Los genes así obtenidos son los que interesan en este estudio ya que se obtienen aquellos genes que presentan expresión diferencial en una comparación y/o cambian simultáneamente su nivel de expresión en más de una comparación. Esta expresión diferencial se puede visualizar mediante perfiles de expresión utilizando gráficos como los Heatmaps (se encuentran en el paquete "ggplot" implementado en R/Bioconductor)»[33]

2.1.2.3 Descubrimiento de Clases

En el descubrimiento de clases se incluye un conjunto de métodos y técnicas cuyo objetivo es buscar patrones en los datos de tal forma que se agrupen los que más se parecen entre sí (los más homogéneos). Siguiendo el material de estudio [33], el descubrimiento de clases se basa en dos puntos esenciales: en la medida de distancias para cuantificar la similitud entre observaciones o variables y el algoritmo de agrupación con el que se decide qué datos son similares.

Un esquema de trabajo para agrupamiento de clases se puede ver en la Figura 2 (basado en la figura de Proceso de Análisis de [33]), se da por hecho que la parte experimental ha ido bien y los primeros pasos son comunes al resto de métodos que se van a usar en este trabajo.

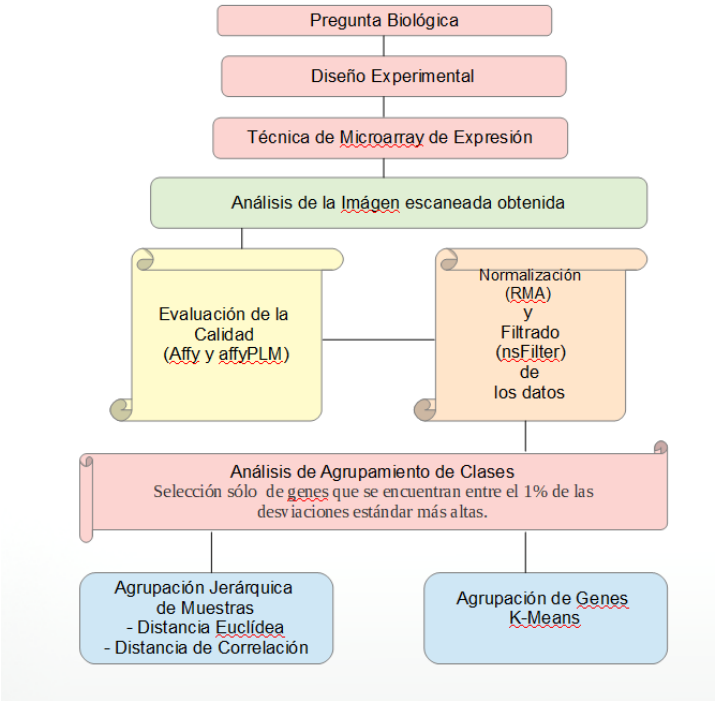


Figura 2 Esquema de trabajo para el Análisis de Descubrimiento de Clases

Para hacer el análisis de agrupamiento de clases se seleccionan sólo aquellos genes que se encuentran entre el 1% de las desviaciones estándar más altas. Básicamente para disminuir la dimensión de los datos ya que por motivos computacionales es difícil hacer este tipo de análisis con datos de alta dimensión. Se escoge el 1% de las desviaciones estándar más altas porque en las medidas de intensidad obtenidas en el array de expresión pueden existir valores extremos, altos. Y al calcular su desviación estándar se ve que es más grande que la media lo cual indica, probablemente, un sesgo.

«En este trabajo se usan dos medidas de distancias: la distancia euclídea y la distancia de correlacción y se aplican dos algoritmos. Se hacen varias agrupaciones para ver cómo se clasifican las muestras y/o los genes según se use un método u otro. Es interesante ver cómo se agrupan las muestras para valorar si hay agrupación según las categorías neuropatológicas establecidas. Para ello se puede realizar Agrupación Jerárquica basada en distancias euclídeas

y Agrupación Jerárquica en función de los coeficientes de correlación.» [33] Concretamente se realiza un clúster jerárquico basado en distancias euclídeas y enlaces promedio (average linkage). Se quiere valorar si hay muestras (pacientes) que tengan niveles de expresión similares.[20]

Sin embargo los perfiles de expresión génica acostumbran a agruparse en función de los coeficientes de correlación. Es decir, se usa el Método Jerárquico teniendo en cuenta las correlaciones entre observaciones (muestras). Se calcula en este caso la distancia media[33]

En la clasificación de los genes es interesante observar si también existe o no agrupación de las muestras según las categorías neuropatológicas. Pero en este caso se puede usar el método de las K-means.

Este es un método no jerárquico partitivo. El método de las k-means permite estudiar la consistencia de la agrupación a partir del cálculo de las sumas de cuadrados dentro de cada grupo (conocido como within SS). El promedio de estas sumas de cuadrados para todos los grupos creados (variable within SS) es una medida de la variabilidad dentro de los grupos, es decir, como son de parecidos los genes dentro de los grupos. Se obtiene la clasificación de las muestras en función del parecido de los genes. Un inconveniente de la agrupación por k-means es que hace falta escoger el número de clusters (k) antes de realizar la agrupación. [33];[20]

Pero para ello se puede buscar el Número óptimo de clusters de una manera semiautomática, para que la decisión no sea tan subjetiva. En este análisis se hace un Gráfico probando entre k=2 y k=8. Se obtiene el gráfico de SS, donde SS de todos los datos, antes del clustering, es la suma de las varianzas de todas las variables multiplicado por el número de observaciones menos 1 [20]

Los resultados de los métodos divisivos como k-means no pueden visualizarse mediante un dendrograma, por lo que es preciso recurrir a otras representaciones gráficas para obtener el pérfil de expresión (difícil de interpretar) [20]

Hay que recordar que por el método k-means se obtienen qué muestras están incluidas en cada clúster y un inconveniente de este método es que en la mayoría de los casos los clúster no son homogéneos, es decir, están constituidos por muestras que no pertenecen a la misma categoría [33];[20]

Hay que recordar también que en las medidas de intensidad obtenidas en el array de expresión pueden existir valores extremos, altos. Al calcular su desviación estándar se ve que es más grande que la media lo cual indica, probablemente, un sesgo. Por la naturaleza del experimento del que se obtienen los datos (análisis de arrays de expresión) se acepta que a mayor valor de intensidad mayor expresión génica.[33]

2.1.2.4 Método de Regresión Penalizado

El análisis que se aborda tiene como objetivo extraer información de los datos obtenidos de los arrays de expresión y de los datos clínicos y anatomohistológicos que figuran en el artículo [26]. La extracción de la información de los datos se hace mediante una técnica de aprendizaje supervisado: la Regresión (penalizada) que constituye un modelo predictivo con variable respuesta continua donde se intenta valorar si los predictores (Snps) tienen poder predictivo sobre la variable respuesta. [20];[21]

En el estudio de la asignatura de modelos de Regresión Lineal así como en el Curso de Data Mining celebrado en Madrid en 2016, se deja claro que los modelos penalizados, también denominados métodos Shrinkage o métodos de regularización, son métodos lineales que usan una fórmula diferente de estimación de los coeficientes de regresión. Introducen una pequeña cantidad de sesgo en los coeficientes de regresión estimados consiguiendo una reducción sustancial de la varianza. Es por ello que se puede aplicar en problemas de alta dimensionalidad como el que nos ocupa [20];[21]

El término de penalización que se añade en estos modelos viene dado por el parámetro lambda que determina el peso que se le da a la penalización. Si lambda vale cero se obtienen los estimadores de máxima verosimilitud (mínimos cuadrados de la regresión lineal normal). Por el contrario, a mayor valor de lambda tendremos que los coeficientes de regresión beta (subjota) son más pequeños. En otras palabras, cuando lambda aumenta los coeficientes (los beta subjota) se van acercando a cero [20];[21]

En el método utilizado muchos de los coeficientes estimados alcanzan el valor cero y solo unos pocos quedan con valores distintos de cero. Se obtiene un modelo menos complejo porque muchas variables han desaparecido. Por eso es un método que se usa como método de selección de predictores en problemas de alta dimensión[20];[21]

En la elaboración de un modelo predictivo se siguen tres puntos: se construye el modelo predictivo, se evalúa la capacidad predictiva y se evalúa la significación estadística [20]

En este análisis se elabora un modelo predictivo pero (por motivos computacionales) el tercer punto no es posible realizarlo; pero ello no invalida el modelo ya que el objetivo es obtener los polimorfismos (Snps) y se puede evaluar la capacidad de selección de los Snps (lo bien que clasifica el método). Por lo tanto: se construye el modelo para lo que se optimizan los parámetros del modelo (denominados lambda) mediante cross-validation, se construye el modelo con los parámetros óptimos y se evalúa la capacidad predictiva del modelo (su capacidad de clasificación) utilizando técnicas de remuestreo (cross-validation) [20];[21]

Un esquema de trabajo para Método de Regresión Penalizada se puede ver en la Figura 3 (basado en la figura de Proceso de Análisis de Microarrays [33]), se da por hecho que la parte experimental ha ido bien y los primeros pasos son comunes al resto de métodos que se van a usar en este trabajo.

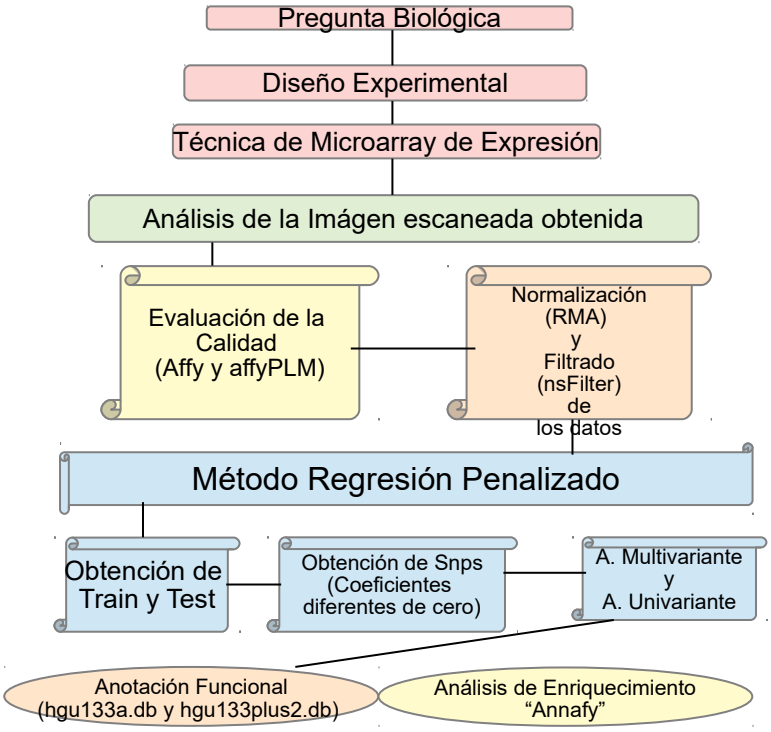


Figura 3 Esquema de trabajo para el Método de Regresión Penalizado

Para realizar este modelo se ejecutan dos partes bien diferenciadas:[20];[21]

1- Se obtienen los dataset, los data train y test, que se usarán en los análisis posteriores, para construir y evaluar el modelo.

2- Se seleccionan los polimorfismos (Snps) mediante técnicas de Data Mining (Métodos Shrinkage)

Además, se valora si la asociación entre los polimorfismos (Snps) seleccionados y la variable respuesta en estudio es significativa.

3- Para ello se realiza Análisis Multivariante que indica el porcentaje de variabilidad que explican los Snps seleccionados en su conjunto respecto a la variable respuesta y un Análisis Univariante para valorar si la asociación, entre los Snps seleccionados (cada uno individualmente) y las variables respuesta definidas, es significativa.

2.1.2.4.1 Obtención de los dataset train y test

El primer paso consiste en obtener el data inicial con el que se trabaja. Se han elaborado las variables respuesta de interés (ver tablas del Material Suplementario 4) a partir de la tabla facilitada en el artículo [26], donde figuran tanto parámetros clínicos como determinaciones anatomohistopatológicas.

Cómo se elaboraron las vgps (valoraciones globales del paciente) tanto clínica (vgpc) como anatomohistopatológica (vgph): Se han obtenido los datos de expresión, de los pacientes correspondientes, de distintos áreas del tejido nervioso. Lo que se quiere valorar es:

1- Por un lado la posible asociación entre los valores de expresión obtenidos en los distintos tejidos respecto a una valoración global clínica de los pacientes. Para ello se suman las dos variables que han valorado en el artículo [26], La primera variable es CDR (clinical dementia raiting) donde los pacientes se clasifican en una de las siguientes clases: 0(no demencia), 0.5 (demencia cuestionable), 1(demencia leve), 2(demencia moderada), 3(demencia severa), 4(demencia profunda) y 5(demencia terminal).

La otra variable que se usa es la CERAD (Consortium to Establish a Registry for Alzheimer's Disease) que mide la categoría neuropatológica en la que se clasifican los pacientes. Así se tienen las siguientes categorías: 1(Normal), 2(Definitiva), 3(Posible) y 4(Probable).

Teniendo en cuenta que la fisiopatología clínica define el estado patológico de los pacientes, se puede definir una variable que se denominara valoración global clínica (vgpc) y que será la suma de CDR y CERAD.

2- Por otro lado se han cuantificado distintos hallazgos anatomohistopatológicos. Por lo tanto, se puede elaborar una variable global histopatológica que se define como la suma de todos los hallazgos realizados en el cerebro del individuo. Se desea valorar si los hallazgos totales anatomohistopatológicos en un cerebro estan asociados significativamente a la mayor o menor expresión génica en determinadas áreas del cerebro. Las variables que se suman serán: braak(neurofibrillas enredadas, le dan un score), PLQ_Mn(promedio de las medidas de densidad de la placa neurítica analizada en 5 regiones cerebrales), NPrSum(suma de los scores de calificación de Cerad en distintas áreas cerebrales) y NTrSum(suma de densidad de ovillos neurofibrilares en distintas Áreas). La variable obtenida es una valoración global histopatológica del tejido nervioso (postmorten) de los pacientes analizados y se denominará vgph.

Tanto en uno como en otro caso se obtiene una variable continua pero que son de escala diferente a las variables continuas de la medida de intensidad de la expresión en los distintos tipos de tejido. Por ello se transforman las variables obtenidas para trabajar en la misma escala. En este caso como en alguna de las variables tiene como valor mínimo cero, no se pueden hacer transformaciones logarítmicas y, por tanto, no se puede comparar que

transformación es la más simétrica para escogerla. Por ello se realiza, para ambas variables, la transformación de raíz cuadrada.[20]

Estas nuevas variables transformadas son las que se integran en el dataset obtenido tras hacer el preprocesado y el filtrado de datos en el análisis de expresión diferencial. Se hace así porque estos datos están normalizados y la diferencia de valores de intensidad se deben realmente a diferencias de expresión y no a cuestiones, por ejemplo, de tipo técnico. En otras palabras, se quiere asociar las variables globales definidas (clínica y anatomohistopatológica) a los valores "reales" de expresión.

Una etapa importante es valorar si existen o no correlaciones entre los datos. Esta etapa se decide no hacer en el caso que nos ocupa ya que interesa más ver si hay relación entre el Snp y la variable respuesta que la correlacción en sí misma. Es decir, se decide no perder información ya que se le da mas peso a la biología que a la ortodoxia estadística [20]

Al dataset de datos normalizados y filtrados, en el que se incorporan las variables respuestas, se le hace la transpuesta ya que interesa que las variables predictoras sean (en este caso) los Snps. Como se ve el data obtenido, de partida para este análisis, es un data donde las filas estan constituidas por el número de la muestra (que pertenece a un paciente) y las columnas por las variables (la vgpc o vgph y los Snps). Sobre este conjunto de datos se elaboran los conjuntos train y test. [20];[21]

La muestra de training se selecciona aleatoriamente con un tamaño n y el resto lo forma la muestra de testing de tamaño N-n.Para obtener el train y el test, se realiza una extracción de los datos aleatoriamente de todas las observaciones, para entrenar al modelo y el resto (test) es para evaluarlo.[20];[21]

2.1.2.4.2 Seleccion de los polimorfismos (Snps) mediante Métodos Shrinkage (Método Lasso)

Una vez obtenidos los subconjuntos train y test en la primera parte de este análisis,el objetivo, en esta segunda, es obtener los polimorfismos (Snps) y evaluar la capacidad de selección de los mismos, se valora cuánto error se comete en la clasificación de los mismos. Por lo tanto: se construye el modelo: para ello se optimizan los parámetros del modelo mediante cross-validation y se construye el modelo con los parámetros óptimos. Y, segundo, se evalúa la capacidad predictiva del modelo utilizando técnicas de remuestreo (cross-validation).Este análisis se realiza de la misma manera para dos variables respuesta distintas: para la variable definida como global clínica y para la variable definida como global anatomohistopatologica [20];[21]

Profundizando en el Método Lasso: Como se dijo en el primer análisis del Método Lasso, los datos de expresión obtenidos en microarrays de expresión se consideran de alta dimensión ya que se obtiene un gran número de predictores (Snps). Por ello se pueden usar métodos de regresión lineal penalizados como el método Lasso para su análisis.[20];[21] Con estos métodos de regularización se controla la magnitud de los coeficientes de regresión, es decir, se buscan estimadores con valores menores. Se introduce una pequeña cantidad de sesgo en los coeficientes de regresión estimados consiguiendo así una reducción de la varianza. En este análisis se utiliza el método Lasso ya que muchos de los coeficientes estimados por el mismo alcanzan el valor de cero y, al final, sólo quedan unos pocos estimadores con valores distintos de cero. Es decir, se obtiene un modelo menos complejo porque muchas variables han desaparecido y, por ello, es un buen método para seleccionar los predictores (en este caso los polimorfismos (Snps)) que están relacionados con la variable respuesta cuantitativa). [20];[21]

Al hacer la selección de los polimorfismos (Snps), se comete un error de clasificación honesto(error de generalización) en el cual la estimación del error se calcula con observaciones que no han participado en la construcción del modelo. El método usado, para obtener una estimación honesta de la capacidad predictiva del modelo, consiste en dividir la muestra original en dos submuestras: Una muestra de entrenamiento (training set) que se utiliza para construir el modelo y una muestra de validación (testing test) donde se realizan las

predicciones y donde se evalúa la capacidad predictiva del modelo. La obtención de train set y testing set es lo que se hizo en la primera parte del análisis [20];[21]

El método Lasso, por lo tanto, es un método predictivo donde los polimorfismos seleccionados (Biomarcadores) están asociados a la variable respuesta cuantitativa de interés y el método de estimación del error usado es la Validación Cruzada. [20];[21]

En el caso de la variable respuesta cuantitativa las medidas usadas para evaluar la capacidad predictiva de un modelo de regresión se basan en las distancias entre los valores observados y los valores predichos por el modelo. En los modelos de regresión se utilizan los residuos, que dan la diferencia entre los valores observados y los valores predichos.[20]

En la Validación Cruzada (utilizada para estimar el error usando train y test) la muestra se divide en K subconjuntos de similar tamaño y ajusta los K modelos dejando cada vez una partición como conjunto de testing y construyendo el modelo con las K-1 restantes. Por lo tanto, se tienen tantos modelos como filas tenga el subconjunto de train. Con ello se reduce la dependencia de los resultados de una única partición aleatoria. Y la estimación del error se calcula como el promedio de los K errores evaluados en las muestras de testing de las K particiones. Es decir, se evalúa el error con una medida que compare los valores observados y los predichos de la variable respuesta exclusivamente en la muestra de testing.[20]

Los K subconjuntos de testing son independientes mientras que los de training son dependientes (comparten una parte de la muestra para construir el modelo). El número de veces que se hace cross-validation (las folds) se elige en función del tamaño de la muestra, normalmente se elige n=10. Y, además, para reducir la dependencia de las particiones aleatorias, se puede repetir varias veces el proceso de validación cruzada. [[20];[21];[33]

Metodológicamente se cargan en el script las funciones que se van a usar. Por un lado la función EvalRegr servirá para evaluar el modelo construido, es una Medida de Evaluación de Modelos de Regresión donde la Variable respuesta es cuantitativa. [20]

En esta función se han incluido las medidas para evaluar la capacidad predictiva de un modelo de regresión. La más importante es el error cuadrático medio (MSE) y su raíz cuadrada que están en las unidades de la variable. El MSE es una medida de la falta de ajuste del modelo a los datos, es decir, cuantifica la diferencia entre las observaciones observadas y las predichas.

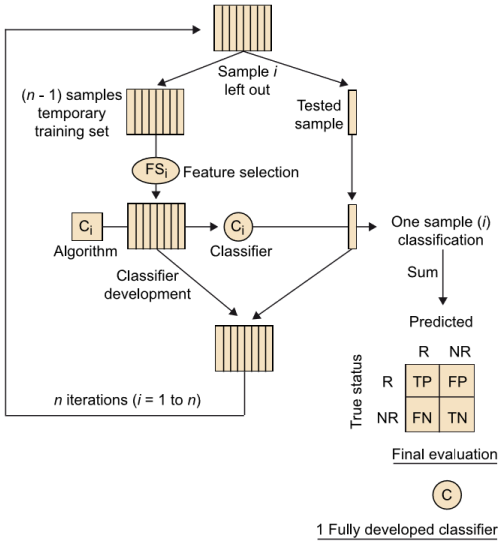


Figura 4 Validación Cruzada con tantas iteracciones como elementos contiene la muestra [33]

Otra medida de evaluación importante es el coeficiente de determinación (R^2) que es una medida del ajuste del modelo y es una medida de la proporción de variabilidad de la

variable respuesta explicada por el modelo ajustado. Si R^2 vale 1 indica un ajuste perfecto y si vale cero indica un ajuste nulo. En otras palabras, es el coeficiente de correlación lineal al cuadrado entre los valores observados y los valores predichos de la variable respuesta. Por lo tanto, no es una medida de precisión del ajuste ya que es una medida relativa; depende de cómo los datos se ajustan a una recta y de la variabilidad de la variable respuesta. [20];[21]

Aunque en la función están incluidas otras medidas de evaluación, en este análisis sólo se evalúa con R2 y RMSE(raíz cuadrada del error cuadrático medio) que es la medida más importante.[20]

En la función creada ProcessLASSO [20] hay dos puntos claros: lo primero es obtener el parámetro lambda óptimo con error mínimo mediante cross-validation. Para ello se usa la función cv.glmnet() y con nfolds =10. En otras palabras, se optimiza el parámetro lambda en la muestra de training mediante 10-fold CV. En este primer punto, además, se obtiene el grafico de los errores frente a los lambdas (en el intervalo de interés). Es el gráfico del Modelo que se selecciona con el MSE mínimo al usar el lambda óptimo y que facilita el número de variables respuesta seleccionadas. Se obtienen tantos modelos como K-1 particiones hay en training. Por motivos computacionales, poca memoria, no se incorpora en este análisis el gráfico de lambdas frente a los errores.[20]

El segundo punto de esta función permite ajustar el modelo mediante el Método Lasso con lambda optimo. Esta función permite obtener las predicciones en otra muestra (testing).Una vez elaboradas todas las funciones y cargadas en el script, se construye el modelo predictivo final con la muestra observada. Para ello se usa la función ProcessLASSO que contiene el proceso completo y se utiliza la funciones EvalRegr para evaluar la capacidad predictiva del modelo realizado[19]

El modelo final es un modelo de regresión Lasso con «x» variables predictoras (según el tejido analizado). Con la función coef() se obtienen los coeficientes distintos de cero y el nombre de las variables (Snps). Se eligen los coeficientes del modelo con el lambda óptimo. Con la función predict() se obtienen las predicciones para el lambda óptimo Los modelos predictivos pueden ser buenos o no sobre la variable respuesta analizada. Para ver si es bueno o no habría que hacer un test de significación estadística del modelo (test de permutaciones). En este caso no se hace por motivos computacionales (se debe establecer el número de permutaciones y debería ser al menos de 1000. Ello supone unas 8 horas de ejecución con un ordenador de 4 núcleos)[20]

2.1.2.4.3 Análisis Multivariante y un Análisis Univariante

En el Análisis Multivariante interesa ver, con los Snps seleccionados, cuanta variabilidad de la variable respuesta explican los snps seleccionados. Y valorar si las variables, en global, todas en conjunto, explican algo de la variable respuesta[20]

Y, por último, se realiza un Análisis Univariante de cada uno de los Snps seleccionados donde se valora si hay relación significativa entre cada Snp y la variable respuesta Este último paso se realiza para obtener aquellos Snps con relación significativa con respecto a la variable respuesta de interés [20]

2.1.2.5 Anotación y Enriquecimiento Funcionalmente

Obtenidos los Snps mediante los tres métodos analizados, lo siguiente es hacer el análisis de significación biológica. Para ello se obtienen las anotaciones de los genes en distintos tipos de bases de datos, principalmente en bases de datos funcionales como Gene Ontology (GO), o la Kyoto Encyclopedia of Genes and Genomes (KEGG). Y en el análisis de Enriquecimiento se establece si una categoría funcional aparece con mayor frecuencia en la lista de genes seleccionados. En este trabajo se usan las siguientes bases de datos para hacer anotación funcional y análisis de Enriquecimiento: hgu133a.db y hgu133Plus2.db. Se usan también bases de datos de organismos, en concreto la base de datos org.Hs.eg.db. Todas ellas implementadas en R/Bioconductor [33]

2.2 Resultados

Las enfermedades Neurodegenerativas (entre ellas el Alzheimer) son un síndrome clínico caracterizado por un deterioro cognitivo persistente y progresivo así como por trastornos conductuales. Afectan a funciones cerebrales superiores y en este trabajo se han seleccionado aquellas muestras de tejido nervioso relacionado, en diversos estudios, con las funciones reseñadas como alteradas en el Alzheimer: principalmente la memoria, lenguaje, orientación o percepción espacial.

Lo primero es analizar la calidad de los datos.[33] Los gráficos obtenidos en los distintos tipos de tejidos son similares pero no iguales. Una vez visto que los datos son de calidad, se analizan los resultatdos de los tres métodos de análisis utilizados (expresión diferencial, descubrimiento de clases y método de regresión penalizada) para obtener dianas genéticas como polimorfismos (SNPs)) en genes asociados a la Enfermedad Neurodegenerativa del Alzheimer. Dada la gran cantidad de resultados obtenidos se presentan los mismos por método de análisis y, dentro de cada método, por tejido analizado.

2.2.1 Control de Calidad de los datos

Para el análisis del área cerebral Lóbulo Frontal se obtienen los datos que corresponden al Tejido Prefrontal Córtex (al que nos referiremos como LF3). Se seleccionaron los 56 pacientes analizados de este área, cuyas muestras van desde el GSM2234357 hasta el GSM2234412 y que presentan distintos grados de afectación neuropatológica.

Para el análisis del área cerebral Lóbulo Parietal se obtienen los datos que corresponden al Tejido Superior Parietal Lobule (al que nos referiremos como LP). Se seleccionaron 50 pacientes (muestras que van desde GSM2234307 hasta GSM2234356) y que presentan distintos grados de afectación neuropatológica.

En el análisis del área cerebral Lóbulo Temporal se obtienen los datos que corresponden al Tejido Superior Temporal Gyrus (al que nos referiremos como LT3). Se seleccionaron 60 pacientes (muestras que van desde GSM2233853 hasta GSM2233912) y que presentan distintos grados de afectación neuropatológica. El otro tejido del Lóbulo Temporal seleccionado es el Temporal Pole. Y para el análisis de este área cerebral se obtienen los datos que corresponden al Tejido Temporal Pole (al que nos referiremos como LT4). Se seleccionaron 58 pacientes (muestras que van desde GSM2234090 hasta GSM2234147) y que presentan distintos grados de afectación neuropatológica.

Para el análisis del cerebral Sistema Límbico se obtienen los datos que corresponden al Tejido Amígdala (al que nos referiremos como SL1). Se seleccionaron 51 pacientes (muestras que van desde GSM2233519 hasta GSM2233569) y que presentan distintos grados de afectación neuropatológica. En este área también se selecciona otro tejido Región Hipocampo. Para el análisis de este área cerebral se obtienen los datos que corresponden al Tejido Región Hipocampo (al que nos referiremos como SL4). Se seleccionaron 55 pacientes (muestras que van desde GSM2234465 hasta GSM2234519) y que presentan distintos grados de afectación neuropatológica.

En todos los tejidos antes de realizar cualquier análisis se evaluó la calidad de los datos obtenidos. La evaluación se lleva a cabo conjuntamente con el informe de Expresión Diferencial para cada Tejido. La calidad se valora con una serie de gráficos. En el ANEXO 1 se observan los histogramas de densidad y el boxplot de la distribución de la señal de intensidad de los arrays (Figura 1 y Figura 2 del ANEXO1). Como se ve, en todos los casos, las diferencias encontradas entre arrays no son reseñables.

En el ANEXO 2 (Figura 3) se observan los gráficos de componentes principales. En ellos se puede ver que las muestras no se agrupan claramente ni en función de la intensidad de la señal ni en función de las categorías neuropatológicas establecidas.

En el ANEXO 3 (Figura 4) se observan los gráficos de degradación de RNA de todos los tejidos analizados. Y se observa como todas las líneas son prácticamente paralelas con lo que el nivel de degradación del RNA es similar en todos los chips.

EN el ANEXO 4 (Figura 5 y Figura 6) se pueden ver los gráficos de los errores no estandarizados y normalizados (NUSE, Normalized Unscaled Standard Errors) para todos los tejidos. Se puede ver que los datos son de calidad ya que presentan una relativa simetría. Es el gráfico de errores o residuos del ajuste del modelo de análisis previo a la normalización de los datos. Es el más informativo ya que representa la distribución de dichos residuos (valores estimados en los modelos de bajo nivel, PLM, antes de ser normalizados). Todos estos gráficos se pueden visualizar mejor en los informes de Expresión Diferencial para cada tejido.

Se puede observar que en el tejido Prefrontal Cortex sólo se observan tres muestras cuyo boxplot está desplazado hacia arriba ligeramente. Se corresponde con las muestras GSM2234365, GSM2234387 y la muestra GSM2234392. En el Tejido Superior Parietal Lobule sólo se observan tres muestras cuyo boxplot está desplazado hacia arriba ligeramente. Se corresponde con las muestras GSM2234309, GSM2234328 y GSM2234334. En el tejido Superior Temporal Gyrus sólo se observan tres muestras cuyo boxplot está desplazado hacia arriba ligeramente. Son las muestras GSM2233867, GSM2233906 y GSM2233910. En el tejido Temporal Pole sólo se observan cuatro muestras cuyo boxplot está desplazado hacia arriba ligeramente. Son las muestras GSM2234103, GSM2234133, GSM2234141 y GSM2234142. En el tejido la Amígdala sólo se observan tres muestras cuyo boxplot está desplazado hacia arriba ligeramente. Son las muestras GSM2233531, GSM2233549 y GSM2233566. Para el tejido Región Hipocampo sólo se observan cuatro muestras cuyo boxplot está desplazado hacia arriba ligeramente. Son las muestras GSM2234468, GSM2234491, GSM2234493 y GSM2234507.

Por lo tanto los datos, antes de ser normalizados para realizar el análisis son de buena calidad. Se puede observar que se ha realizado correctamente la normalización de los datos mediante los boxplot correspondientes a cada tejido (ANEXO 5, Figura 7). Este punto es muy importante ya que el resto de los análisis se inician con el conjunto de datos normalizados y filtrados.

2.2.2 Expresión Diferencial

Lóbulo Frontal

Según el criterio CERAD (Consortium to Establish a Registry for Alzheimer's Disease), se mide la categoria neuropatológica en la que se clasifican los pacientes en las siguientes categorías: 1(Normal), 2(Definitiva), 3(Posible) y 4(Probable). Para poder realizar expresión diferencial y establecer los contrastes de hipótesis se asigna a cada categoría con una letra con lo que quedarían establecidas las categorías como: A(Normal), B(Definitiva), C(Posible) y D(Probable).

Con el método de expresión diferencial se realiza un ajuste lineal donde se estiman los siguientes contrastes:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D)

Y, además, se realizan las pruebas de significación para cada gen y cada comparación. En el tejido Prefrontal Cortex se obtiene para la primera comparación (A vs B), para un p-valor < 0.05, 351 genes diferencialmente expresados. Y para un p-valor < 0.01 NO se obtiene ningún gen diferencialmente expresado.

En la primera comparación, donde se obtienen genes diferencialmente expresados, tenemos los 10 genes más diferencialmente expresados (que tienen un p-valor significativo) (Tabla 3).

Los dos primeros genes VCAN y CALU están implicados en procesos de unión de iones calcio mientras que RGS5 está implicado en la activación del AMPc (Adenosín Monofosfato Cíclico) dependiente de PKA (proteína quinasa A). El gen MINDY2 desempeña un papel regulador en el nivel de recambio de proteínas. El gen EZR, su proteína, sirve como intermediario entre la membrana plasmática y el citoesqueleto de actina. VAMP3 es un gen relacionado con el metabolismo del Calcio y se le relaciona con las neuronas dopaminérgicas. Y el gen MTFR1 juega un importante papel en la respiración aeróbica de la mitocondria. Se pueden observar en el informe adjunto de Expresión Diferencial para Tejido Prefrontal Cortex así como los volcanos plot que se obtienen para cada comparación. En dichos gráficos “cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado” [33]

ID Probe	GEN
211571_s_at	VCAN
200755_s_at	CALU
209070_s_at	RGS5
214691_x_at	MINDY2
211749_s_at	VAMP3
208622_s_at	EZR
201337_s_at	VAM
221848_at	ZGPAT
204620_s_at	VCAN
203207_s_at	MTFR1

Tabla 3 Los 10 genes más diferencialmente expresados en la primera comparación del Tejido Prefrontal Cortex

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación se obtiene que hay 8 genes en la primera comparación (A vs B), 7 de los cuales están down regulados y 1 up regulado. El pérfil de expresión de estos genes viene dado en la Figura 5.

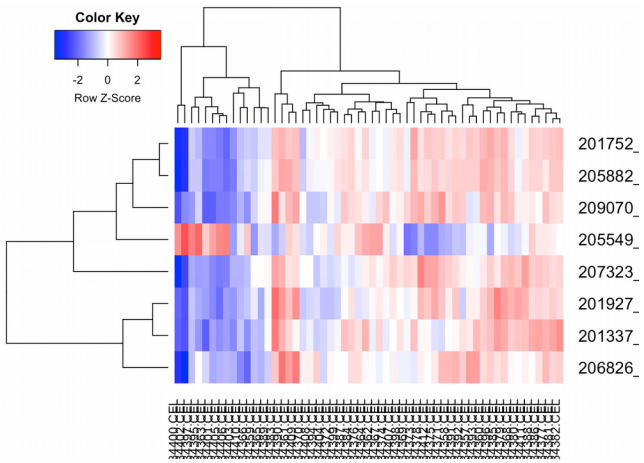


Figura 5 Perfil de Expresión del Tejido Prefrontal Cortex

Los genes corresponden a las sondas que se pueden ver 201752_at, 205882_at, 209070_at, 205549_at, 207323_at, 201927_at,201337_at, 206826_at (se le pone a todas _at porque no se visualiza bien el sufijo que las acompaña). Se puede ver que la sonda 205549_at está sobreexpresada en las muestras GSM2234400, GSM2234402, GSM2234392,

GSM2234395, GSM2234363, GSM2234401, GSM2234405 y GSM2234406 mientras que el resto de genes, en esas mismas muestras, están claramente infraexpresados.

Las características de dichos genes se pueden observar en la siguiente Tabla 4 (la tabla completa se puede ver en carpeta ANEXO 6, Anotación para tejido Prefrontal Cortex en formato .html)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
201337_s_at	VAMP3	Vesicle associated membrane protein 3	1	NM_004781	1p36.23
201752_s_at	ADD3	adducin 3	10	AI763123	10q25.1-q25.2
201927_s_at	PKP4	plakophilin 4	2	BG292559	2q24.1
205549_at	PCP4	Purkinje cell protein 4	21	NM_006198	21q22.2
205882_x_at	ADD3	adducin 3	10	AI818488	10q25.1-q25.2
206826_at	PMP2	peripheral myelin protein 2	8	NM_002677	8q21.13
207323_s_at	MBP	myelin basic protein	18	NM_002385	18q23
209070_s_at	RGS5	regulator of G protein signaling 5	1	AI183997	1q23.3

Tabla 4 Resumen Anotación Funcional del Tejido Prefrontal Cortex para Expresión Diferencial

Al consultar los genes seleccionados en bases de datos constrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 8 genes, cuatro están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

ADD3 está asociado a parálisis cerebral así como a Tetraplejía espástica. Metabólicamente se ha relacionado con rutas de activación de cAMP (Adenosín Monofosfato Cíclico) dependiente de PKA (Proteína quinasa A) y en el transporte de azúcares, sales biliares y ácidos orgánicos, así como de iones metálicos y compuestos aminos. Está incluido como uno de los genes de unión de la calmodulina.

PCP4 es uno de los genes de activación de la calmodulina dependiente de kinasa y constituye una de las proteínas de las células Purkinje,

PMP2 da lugar a una proteína que se localiza en la vaina de mielina de los nervios periféricos. Un defecto en el mismo da lugar a neuropatías desmielinizantes.

MBP es un gen asociado a enfermedades desmielinizantes así como a Esclerosis múltiple. Entre las rutas metabólicas asociadas a este gen se encuentra el proceso de diferenciación de la cresta neural y la diferenciación de las células gliales. Es también, un constituyente de la vaina de mielina.

Se ha visto que la sonda 205549_at que corresponde al gen PCP4 está sobreexpresado en las muestras indicadas y éstas no pertenecen a una sola categoría neuropatológica.

Lóbulo Parietal

Con el método de expresión diferencial aplicado al Tejido Superior Parietal Lobule, se realiza un ajuste lineal donde se estiman los mismos contrastes del tejido anterior:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D)

También se realizan las pruebas de significación para cada gen y cada comparación. Así para la tercera comparación (A vs D), para un p-valor < 0.05, se obtienen 436 genes diferencialmente expresados, para la cuarta comparación (B vs D) se obtienen 32 genes diferencialmente expresados y para la quinta comparación (B vs D) se obtienen 79 genes diferencialmente expresados. Y para un p-valor < 0.01 se obtienen 17 genes diferencialmente expresados en la tercera comparación (A vs D). Los 10 genes más diferencialmente expresados (que tienen un p-valor significativo) obtenidos en cada comparación anterior se pueden observar en el informe de Expresión Diferencial para Tejido Superior Parietal Lobule que se adjunta así como los volcanos plot que se obtienen para cada comparación. En dichos gráficos "cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado" [33]

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación se obtiene en la tercera comparación 36 genes down regulados, en la cuarta comparación 1 gen down regulado y en la quinta comparación 27 genes down regulados. Como genes up regulados tenemos: 14 en la cuarta comparación y 3 en la quinta comparación. El perfil de expresión de estos genes viene dado en la Figura 6. En ella se observan dos clúster claramente diferenciados. Aunque es difícil visualizar, se puede inferir que las sondas que comienzan con AFFX_ se encuentran infraexpresadas en las muestras que forman el clúster de la izquierda mientras que están sobreexpresadas en las muestras que forman el clúster de la derecha.

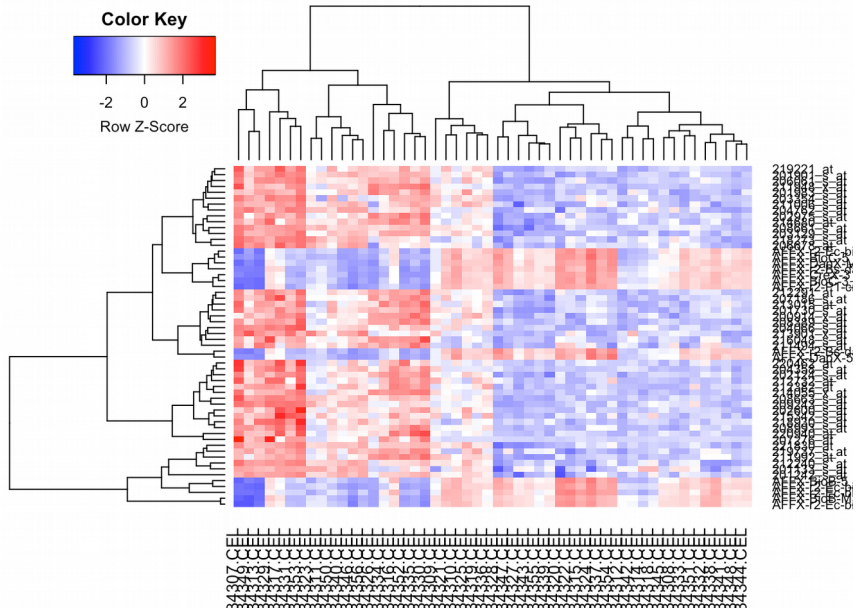


Figura 6 Perfil de Expresión del Tejido Superior Parietal Lobule

En total se seleccionan, en las comparaciones múltiples entre condiciones, 58 genes cuyas características se pueden ver en la Tabla 5 (la tabla completa se puede ver en carpeta ANEXO 6, Anotación para tejido Superior Parietal Lobule en formato .html)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
200914_x_at	KTN1	kinectin 1 (kinesin receptor)	14	BF589024	14q22.1
201133_s_at	PJA2	praja ring finger 2, E3 ubiquitin protein ligase	5	AA142966	5q21.3
201242_s_at	ATP1B1	ATPase, Na+/K+ transporting, beta 1 polypeptide	1	BC000006	1q24
201730_s_at	TPR	translocated promoter region, nuclear basket protein	1	BF110993	1q25

201901_s_at	YY1 YY1 transcription factor	14	Z14077	14q	
201983_s_at	EGFR	epidermal growth factor receptor	7	AW157070	7p12
202124_s_at	TRAK2	trafficking protein, kinesin binding 2	2	AV705253	2q33
202600_s_at	NRIP1	nuclear receptor interacting protein 1	21	AI824012	21q11.2
202975_s_at	RHOBTB3	Rho-related BTB domain containing 3	5	N21138	5q15
203129_s_at	KIF5C	kinesin family member 5C	2	BF059313	2q23.1
203354_s_at	PSD3	pleckstrin and Sec7 domain containing 3	8	AW117368	8p21.3
204066_s_at	AGAP1	ArfGAP with GTPase domain, ankyrin repeat and PH domain 1	2	NM_014914	2q37
204358_s_at				AF169676	
204762_s_at	GNAO1	guanine nucleotide binding protein (G protein), alpha activating activity polypeptide O	16	BE670563	16q13
206061_s_at	DICER1	dicer 1, ribonuclease type III	14	NM_030621	14q32.13
206678_at	GABRA1	gamma-aminobutyric acid (GABA) A receptor, alpha 1	5	NM_000806	5q34
207186_s_at	BPTF	bromodomain PHD finger transcription factor	17	NM_004459	17q24.3
207276_at	CDR1	cerebellar degeneration-related protein 1, 34kDa	X	NM_004065	Xq27.1
208389_s_at	SLC1A2	solute carrier family 1 (glial high affinity glutamate transporter), member 2	11	NM_004171	11p13-p12
208661_s_at				AW510696	
208663_s_at				AI652848	
208993_s_at	PPIG	peptidylprolyl isomerase G (cyclophilin G)	2	AW340788	2q31.1
209243_s_at	PEG3	paternally expressed 3	19	AF208967	19q13.4
211006_s_at	KCNB1	potassium channel, voltage gated Shab related subfamily B, member 1	20	L02840	20q13.2
211494_s_at	SLC4A4	solute carrier family 4 (sodium bicarbonate cotransporter), member 4	4	AF157492	4q21
211948_x_at	PRRC2C	proline-rich coiled-coil 2C	1	BG261071	1q23.3
211992_at	WNK1	WNK lysine deficient protein kinase 1	12	AI445745	12p13.3
212240_s_at	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	5	AI679268	5q13.1
212291_at	HIPK1	homeodomain	1	AI393355	1p13.2

		interacting protein kinase 1			
212382_at	TCF4	transcription factor 4	18	BF433429	18q21.1
212732_at	MEG3	maternally expressed 3 (non-protein coding)	14	AI950273	14q32
213015_at	BBX	bobby sox homolog (Drosophila)	3	BF448315	3q13.1
213901_x_at	RBFOX2	RNA binding protein, fox-1 homolog (C. elegans) 2	22	AW149379	22q13.1
214055_x_at	PRRC2C	proline-rich coiled-coil 2C	1	AW238632	1q23.3
214680_at	NTRK2	neurotrophic tyrosine kinase, receptor, type 2	9	BF674712	9q22.1
215342_s_at	RABGAP1L	RAB GTPase activating protein 1-like	1	AB019490	1q24
216048_s_at	RHOBTB3	Rho-related BTB domain containing 3	5	AK023621	5q15
218273_s_at	PDP1	pyruvate dehyrogenase phosphatase catalytic subunit 1	8	NM_018444	8q22.1
218930_s_at	TMEM106B	transmembrane protein 106B	7	NM_018374	7p21.3
219221_at	ZBTB38	zinc finger and BTB domain containing 38	3	NM_024724	3q23
219737_s_at	PCDH9	protocadherin 9	13	AI524125	13q21.32
220462_at	CSRNP3	cysteine-serine-rich nuclear protein 3	2	NM_024969	2q24.3
220940_at	ANKRD36B	ankyrin repeat domain 36B	2	NM_025190	2q11.2
221830_at	RAP2A	RAP2A, member of RAS oncogene family	13	AI302106	13q34
AFFX-BioB-5_at				AFFX-BIOB-5	
AFFX-BioB-M_at				AFFX-BIOB-M	
AFFX-BioC-3_at				AFFX-BIOC-3	
AFFX-BioC-5_at				AFFX-BIOC-5	
AFFX-CreX-3_at				AFFX-CREX-3	
AFFX-DapX-5_at				AFFX-DAPX-5	
AFFX-DapX-M_at				AFFX-DAPX-M	
AFFX-r2-Bs-dap-5_at				AFFX-R2-BS-DAP-5	
AFFX-r2-Bs-dap-M_at				AFFX-R2-BS-DAP-M	
AFFX-r2-Ec-bioB-5_at				AFFX-R2-EC-BIOB-5	
AFFX-r2-Ec-bioB-M_at				AFFX-R2-EC-BIOB-M	
AFFX-r2-Ec-bioC-3_at				AFFX-R2-EC-BIOC-3	
AFFX-r2-Ec-bioC-5_at				AFFX-R2-EC-BIOC-5	
AFFX-r2-P1-cre-3_at				AFFX-R2-P1-CRE-3	

Tabla 5 Resumen Anotación Funcional del Tejido Superior Parietal Lobule para Expresión Diferencial

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 58 genes, se seleccionan 10 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

PJA2 es esencial para los procesos de memoria a largo plazo mediados por PKA(Proteína quinasa A)

TPR es componente del complejo de poro nuclear (NPC), un complejo requerido para el tráfico a través de la envoltura nuclear.

TRAK2 está relacionado con la Esclerosis Lateral Amiotrófica 2 y la Esclerosis Lateral Amiotrófica Juvenil Entre sus vías relacionadas están el metabolismo y la sinapsis GABAérgica (La síntesis del neurotransmisor ácido gamma aminobutírico (GABA) se inicia por la descarboxilación del Glutamato gracias a la enzima glutamato descarboxilasa (GAD). En el botón sináptico de la neurona presináptica gabaérgica se encuentra almacenado GABA en vesículas. Cuando llega el potencial de acción se produce la entrada del ión Calcio a la neurona presináptica lo que provoca la unión de las vesículas a la membrana presináptica y la liberación del neurotransmisor a la hendidura sináptica. El neurotransmisor GABA se une a receptores ionotrópicos y metabotrópicos de la neurona postsináptica. El neurotransmisor GABA se puede recapturar por las células gliales para ser convertido de nuevo en glutamato. [])

KIF5C La proteína codificada por este gen es una subunidad de cadena pesada de kinesina involucrada en el transporte de moléculas cargadas dentro del sistema nervioso central.

GNAO1 La proteína codificada por este gen representa la subunidad alfa del complejo de transducción de señal de la proteína G heterotrimérica. Los defectos en este gen son una causa de encefalopatía epiléptica de inicio temprano.

GABRA1 Este gen codifica un receptor de ácido gamma-aminobutírico (GABA). GABA es el principal neurotransmisor inhibidor en el cerebro de los mamíferos donde actúa en los receptores ionotrópicos (GABA-A), que son canales de cloruro controlados por ligando. Entre sus vías relacionadas se encuentran la transmisión a través de las sinapsis químicas y la sinapsis GABAérgica.

CDR1 Las enfermedades asociadas con CDR1 incluyen la degeneración cerebelosa y la degeneración cerebelosa paraneoplásica.

SLC1A2 Las enfermedades asociadas con SLC1A2 incluyen Encefalopatía Epiléptica y Encefalopatía de Wernicke. Entre sus vías relacionadas se encuentran la transmisión a través de las sinapsis químicas.

KCNB1 Las enfermedades asociadas con KCNB1 incluyen encefalopatía epiléptica y encefalopatía epiléptica de inicio precoz indeterminado. Entre sus vías relacionadas están la transmisión a través de las sinapsis químicas.

TCF4 son una familia de serina / treonina proteína quinasas relacionadas con el crecimiento, que se activan en respuesta a señales extracelulares y regulan la forma y la motilidad de la célula. Pertenece a los PAK (proteinas activadoras de quinasas) pathways que regulan diversas funciones celulares, incluida la expresión génica, el ensamblaje de actina citoesquelética, las rutas MAPK (proteína quinasa activada por mitógeno), el crecimiento de neuritas, el control del ciclo celular y la apoptosis celular.
(http://saweb2.sabiosciences.com/pathway.php?sn=PAK_Pathway)

Las enfermedades asociadas con ***TMEM106B*** incluyen demencia frontotemporal y afasia progresiva no fluida. Involucrado en la morfogénesis y el mantenimiento de la dendrita regulando el tráfico lisosómico a través de su interacción con MAP6. Puede actuar inhibiendo el transporte retrógrado de lisosomas a lo largo de las dendritas. Requerido para la ramificación dendrítica.

RAP2A Pequeña proteína de unión a GTP que cicla entre una forma inactiva unida al GDP y una activa GTP. En su forma activa, interactúa y regula varios efectores, incluidos MAP4K4, MINK1 y TNIK. Forma parte de un complejo de señalización compuesto por NEDD4, RAP2A y TNIK que regula la extensión de la dendrita neuronal y la arborización durante el desarrollo. De manera más general, es parte de varias cascadas de señalización y puede regular los reordenamientos del citoesqueleto, la migración celular, la adhesión celular y la propagación celular.

Lóbulo Temporal

En este caso se seleccionaron dos Tejidos: el Superior Temporal Gyrus y el Temporal Pole

Superior Temporal Gyrus

Siguiendo el mismo esquema que en los otros dos tejidos, con el método de expresión diferencial aplicado al Tejido Superior Temporal Gyrus, se realiza un ajuste lineal donde se estiman los mismos contrastes de los tejidos anteriores:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D).

También se realizan las pruebas de significación para cada gen y cada comparación. Así para la primera comparación (A vs B), para un p-valor < 0.05, se obtienen 1779 genes diferencialmente expresados. Y para un p-valor < 0.01 NO se obtienen genes diferencialmente expresados. Los 10 genes más diferencialmente expresados (que tienen un p-valor significativo) obtenidos en cada comparación anterior se pueden observar en el informe de Expresión Diferencial para Tejido Superior Temporal Gyrus que se adjunta así como los volcanos plot que se obtienen para cada comparación. En dichos gráficos “cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado” [33]

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación se obtiene en la primera comparación 18 genes up regulados. El perfil de expresión de estos genes viene dado en la Figura 7. En ella se observan dos clúster claramente diferenciados. Los 18 genes seleccionados se encuentran infraexpresados en las muestras que forman el clúster de la izquierda mientras que están sobreexpresados en las muestras que forman el clúster de la derecha.

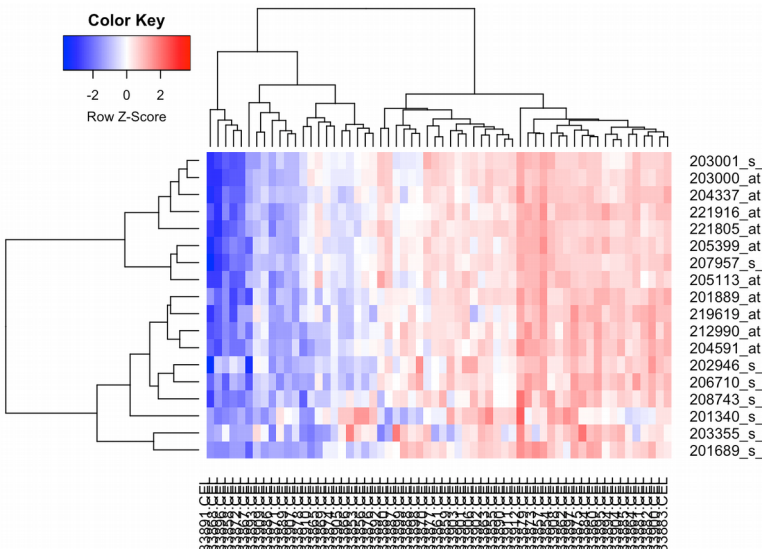


Figura 7 Perfil de Expresión del Tejido Superior Temporal Gyrus

En total se seleccionan, en las comparaciones múltiples entre condiciones, 18 genes cuyas características se pueden ver en la Tabla 6 (la tabla completa se puede ver en carpeta ANEXO 6, Anotación para tejido Superior Temporal Gyrus en formato .html).

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
201340_s_at	ENC1	ectodermal-neural cortex 1 (with BTB domain)	5	AF010314	5q13
201689_s_at	TPD52	tumor protein D52	8	BE974098	8q21.13
201889_at	FAM3C	family with sequence similarity 3, member C	7	NM_014888	7q31
202946_s_at	BTBD3	BTB (POZ) domain containing 3	20	NM_014962	20p12.2
203000_at	STMN2	stathmin 2	8	BF967657	8q21.13
203001_s_at	STMN2	stathmin 2	8	NM_007029	8q21.13
203355_s_at	PSD3	pleckstrin and Sec7 domain containing 3	8	NM_015310	8p21.3
204337_at	RGS4	regulator of G-protein signaling 4	1	AL514445	1q23.3
204591_at	CHL1	cell adhesion molecule L1-like	3	NM_006614	3p26.1
205113_at	NEFM	neurofilament, medium polypeptide	8	NM_005382	8p21
205399_at	DCLK1	doublecortin-like kinase 1	13	NM_004734	13q13
206710_s_at	EPB41L3	erythrocyte membrane protein band 4.1-like 3	18	NM_012307	18p11.32
207957_s_at	PRKCB	protein kinase C, beta	16	NM_002738	16p11.2
208743_s_at	YWHAB	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta	20	BC001359	20q13.1
212990_at	SYNJ1	synaptojanin 1	21	AB020717	21q22.2
219619_at	DIRAS2	DIRAS family, GTP-binding RAS-like 2	9	NM_017594	9q22.2
221805_at	NEFL	neurofilament, light polypeptide	8	AL537457	8p21
221916_at	NEFL	neurofilament, light polypeptide	8	BF055311	8p21

Tabla 6 Resumen Anotación Funcional del Tejido Superior Temporal Gyrus para Expresión Diferencial

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 18 genes, se seleccionan 5 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

BTBD3 Actúa como un regulador clave de la orientación del campo dendrítico durante el desarrollo de la corteza sensorial. También dirige dendritas hacia terminales de axones activos cuando se expresa ectópicamente.

STMN2 Este gen codifica un miembro de la familia Stathmin protein. Las proteínas Stathmin son fosfoproteínas que funcionan en la dinámica de los microtúbulos y en la transducción de señales. La proteína codificada desempeña un papel regulador en el crecimiento neuronal y también se cree que está implicada en la osteogénesis. Las reducciones

en la expresión de este gen se han asociado con el síndrome de Down y la enfermedad de Alzheimer.

NEFM Los neurofilamentos son heteropolímeros de filamentos intermedios tipo IV compuestos de cadenas ligeras, medias y pesadas. Los neurofilamentos comprenden el axoesqueleto y mantienen funcionalmente el calibre neuronal. También pueden desempeñar un papel en el transporte intracelular de axones y dendritas. Este gen codifica la proteína de neurofilamento medio. Esta proteína se usa comúnmente como biomarcador de daño neuronal. Entre sus vías relacionadas se encuentran la esclerosis lateral amiotrófica (ELA) y las vías de diferenciación de células madre neurales.

SYNJ1 Este gen codifica una fosfatasa que regula los niveles de fosfatidilinositol-4,5-bisfosfato de membrana. La expresión de esta enzima puede afectar la transmisión sináptica Las enfermedades asociadas con SYNJ1 incluyen Parkinson y Encefalopatía epiléptica.

NEFL está asociado a distintas neuropatías.

Temporal Pole

Con el método de expresión diferencial aplicado al Tejido Temporal Pole, se realiza un ajuste lineal donde se estiman los mismos contrastes de los tejidos anteriores:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D).

También se realizan las pruebas de significación para cada gen y cada comparación. Así para el tejido Temporal Pole NO se obtienen genes diferencialmente expresados ni para p-valor < 0.05 ni para p-valor < 0.01. En este caso los 10 genes más diferencialmente expresados obtenidos en cada comparación no tienen un p-valor significativo. También se pueden observar los plots volcanos que se obtienen para cada comparación. En dichos gráficos “cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado” [33]

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación NO se obtiene ningún gen diferencialmente expresado.

Sístema Límbico

En este caso se seleccionaron dos Tejidos: Amígdala y Región Hipocampo

Amígdala

Siguiendo el mismo criterio que con los otros tejidos, con el método de expresión diferencial aplicado al Tejido Amígdala, se realiza un ajuste lineal donde se estiman los mismos contrastes de los tejidos anteriores:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D)

También se realizan las pruebas de significación para cada gen y cada comparación. Así para la primera comparación (A vs B), para un p-valor < 0.05, se obtienen 1438 genes diferencialmente expresados y para la comparación tercera (A vs D) se obtienen 18 genes diferencialmente expresados. Y para un p-valor < 0.01 se obtienen para la primera comparación (A vs B) 93 genes diferencialmente expresados y para la tercera comparación (A vs D) se obtienen 8 genes diferencialmente expresados.

Los 10 genes más diferencialmente expresados (que tienen un p-valor significativo) obtenidos en cada comparación anterior se pueden observar en el informe de Expresión Diferencial para Tejido Amígdala que se adjunta así como los volcanos plot que se obtienen para cada comparación. En dichos gráficos "cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado" [33]

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación se obtiene en la primera comparación 17 genes up regulados y 31 genes down regulados. En la tercera comparación se obtienen 9 genes up regulados y 3 genes down regulados. El perfil de expresión de estos genes viene dado en la Figura 8. En ella se observan dos clúster claramente diferenciados. Lo más reseñable de este Heatmap es que hay un grupo de 9 genes que conforman el clúster menor, el de la izquierda (228497_at, 204409_s, 207703_at, 206700_s, 205000_at, 204410_at, 236694_at, 206624_at, 201909_at) que están sobreexpresados en las muestras que finalizan en 519, 529, 526, 525, 528, 522, 530, 533 y 532 mientras que están infraexpresadas en el resto de muestras.

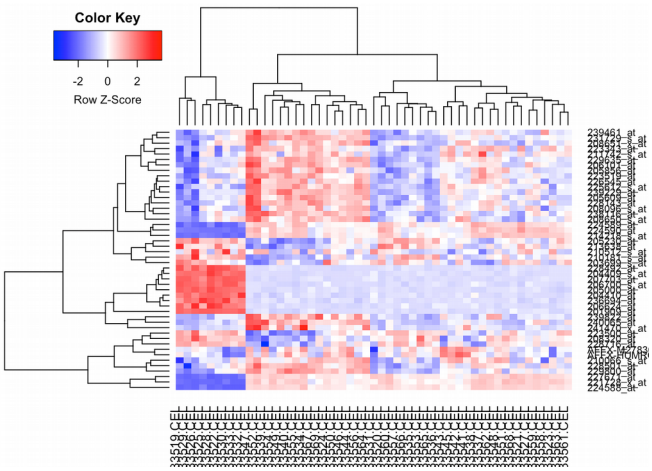


Figura 8 Perfil de Expresión del Tejido Amígdala

En total se seleccionan, en las comparaciones múltiples entre condiciones, 48 genes cuyas características se pueden ver en la Tabla 7(la tabla completa se puede ver en carpeta ANEXO 6, Anotación para tejido Amígdala en formato .html).

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
201909_at	RPS4Y1	ribosomal protein S4, Y-linked 1	Y	NM_001008	Yp11.3
203699_s_at	DIO2	deiodinase, iodothyronine, type II	14	U53506	14q24.2-q24.3
204409_s_at	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	Y	BC005248	Yq11.223
204410_at	EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	Y	NM_004681	Yq11.223
205000_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	Y	NM_004660	Yq11
205230_at	RPH3A	rabphilin 3A	12	NM_014954	12q24.13
205609_at	ANGPT1	angiopoietin 1	8	NM_001146	8q23.1
205856_at	SLC14A1	solute carrier family 14 (urea transporter), member 1 (Kidd blood group)	18	NM_015865	18q11-q12
206101_at	ECM2	extracellular matrix protein 2, female organ and adipocyte specific	9	NM_001393	9q22.3

206624_at	USP9Y	ubiquitin specific peptidase 9, Y-linked	Y	NM_004654	Yq11.2
206700_s_at	KDM5D	lysine (K)-specific demethylase 5D	Y	NM_004653	Yq11
207703_at	NLGN4Y	neuroligin 4, Y-linked	Y	NM_014893	Yq11.221
208096_s_at	COL21A1	collagen, type XXI, alpha 1	6	NM_030820	6p12.3-p11.2
208320_at	CABP1	calcium binding protein 1	12	NM_004276	12q24.31
208650_s_at	CD24	CD24 molecule	6	BG327863	6q21
208651_x_at	CD24	CD24 molecule	6	M58664	6q21
210066_s_at	AQP4	aquaporin 4	18	D63412	18q11.2-q12.1
210181_s_at	CABP1	calcium binding protein 1	12	AF169148	12q24.31
210512_s_at	VEGFA	vascular endothelial growth factor A	6	AF022375	6p12
211742_s_at	EVI2B	ecotropic viral integration site 2B	17	BC005926	17q11.2
213638_at	PHACTR1	phosphatase and actin regulator 1	6	AW054711	6p24.1
214218_s_at	XIST	X inactive specific transcript (non-protein coding)	X	AV699347	Xq13.2
221728_x_at	XIST	X inactive specific transcript (non-protein coding)	X	AA628440	Xq13.2
223343_at	MS4A7	membrane-spanning 4-domains, subfamily A, member 7	11	AI301935	11q12
223500_at	CPLX1	complexin 1	4	BC002471	4p16.3
223519_at	ZAK	sterile alpha motif and leucine zipper containing kinase AZK	2	AW069181	2q24.2
224588_at	XIST	X inactive specific transcript (non-protein coding)	X	AA167449	Xq13.2
224589_at	XIST	X inactive specific transcript (non-protein coding)	X	BF223193	Xq13.2
224590_at	XIST	X inactive specific transcript (non-protein coding)	X	BE644917	Xq13.2
225612_s_at	B3GNT5	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5	3	BE672260	3q28
226545_at	CD109	CD109 molecule	6	AL110152	6q13
227671_at	XIST	X inactive specific transcript (non-protein coding)	X	AV646597	Xq13.2
228143_at	CP	ceruloplasmin (ferroxidase)	3	AI684991	3q23-q25
228492_at	USP9Y	ubiquitin specific peptidase 9, Y-linked	Y	AV681765	Yq11.2
228501_at	GALNT15	polypeptide N-acetylgalactosaminyltransferase 15	3	BF055343	3p25.1
228716_at	THRB	thyroid hormone receptor, beta	3	BG494007	3p24.2
229635_at	LINC01094	long intergenic non-protein coding RNA 1094	4	AW043859	
229800_at	DCLK1	doublecortin-like kinase 1	13	AI129626	13q13
231729_s_at	CAPS	calcyphosine	19	NM_004058	19p13.3

236694_at	TXLNGY	taxilin gamma pseudogene, Y-linked	Y	AW468885	Yq11.222
238116_at	DYNLRB2	dynein, light chain, roadblock-type 2	16	AW959427	16q23.3
239229_at	PHEX	phosphate regulating endopeptidase homolog, X-linked	X	AI342246	Xp22.2-p22.1
239461_at	GALNT15	polypeptide N-acetylgalactosaminyltransferase 15	3	AW205686	3p25.1
239822_at	LINC01354	long intergenic non-protein coding RNA 1354	1	AI869174	1q42.2
240065_at	FAM81B	family with sequence similarity 81, member B	5	AI769413	5q15
241470_x_at	RASSF9	Ras association (RalGDS/AF-6) domain family (N-terminal) member 9	12	R97781	12q21.31
AFFX-HUMRGE/M10098_M_at				AFFX-HUMRGE/M10098_M	
AFFX-M27830_5_at				AFFX-M27830_5	

Tabla 7 Resumen Anotación Funcional del Tejido Amígdala para Expresión Diferencial

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 18 genes, se seleccionan 3 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

NLGN4Y Entre sus vías relacionadas están la transmisión a través de las sinapsis químicas y las interacciones proteína-proteína en las sinapsis.

CD24 Las enfermedades asociadas con CD24 incluyen la esclerosis múltiple.

CPLX1 Entre sus vías relacionadas están la transmisión a través de las sinapsis químicas y el ciclo de liberación de neurotransmisores.

Región Hipocampo

Con el método de expresión diferencial aplicado al Tejido Temporal Pole, se realiza un ajuste lineal donde se estiman los mismos contrastes de los tejidos anteriores:

-Normal versus Definitivo (A vs B)
-Normal versus Posible (A vs C)
-Normal versus Probable (A vs D)
-Definitivo versus Posible (B vs C)
-Definitivo versus Probable (B vs D) y
-Posible versus Probable (C vs D)

También se realizan las pruebas de significación para cada gen y cada comparación. Así para el tejido Región Hipocampo NO se obtienen genes diferencialmente expresados ni para p-valor < 0.05 ni para p-valor < 0.01. En este caso los 10 genes más diferencialmente expresados obtenidos en cada comparación no tienen un p-valor significativo. También se pueden observar los plots volcanos que se obtienen para cada comparación. En dichos gráficos "cuánto más hacia arriba y hacia afuera se encuentra un gen más fuerte es la evidencia de que el gen está diferencialmente expresado" [33]

Por otro lado en cuanto a los genes que cambian simultáneamente en más de una comparación NO se obtiene ningún gen diferencialmente expresado.

2.2.3 Descubrimiento de Clases

Se puede observar en los informes que se adjuntan de Descubrimiento de Clases (uno por cada tejido analizado bien en formato pdf o bien en formato .html) que para hacer este análisis se seleccionan sólo aquellos genes que se encuentran entre el 1% de las desviaciones estándar más altas.

Se seleccionan en todos los casos 223 genes excepto para el Tejido Amígdala que se seleccionan 547 genes. En todos los casos al realizar la agrupación jerárquica (basado tanto en las distancias euclídeas como la basada en función de los coeficientes de correlación) de las muestras se tiene en cuenta sólo a estos 223 genes (a los 547 en el caso del tejido Amígdala) y, en ningún caso, se observa agrupación homogénea entre muestras que pudiera indicar, asociar, un nivel de neuropatología a los 223 genes seleccionados (547 en el caso del tejido Amígdala).

En el agrupamiento de genes mediante k-means (datos que se pueden observar en los informes de Descubrimiento de clases que adjuntan, unos en formato pdf y otros en formato .html) se obtienen el siguiente número de clúster según el tejido:

- Prefrontal Cortex: cinco clúster (óptimo) en donde se observa: que en el primer clúster se agrupan 20 muestras, en el clúster 2 se agrupan 7 muestras, en el clúster 3 se agrupan 8 muestras, en el clúster 4 se agrupan 14 muestras y en el clúster 5 se agrupan 7 muestras.

- Superior Parietal Lobule: seis clúster (óptimo) en donde se tiene: 3 muestras en el clúster 1, 6 muestras en el clúster 2, doce muestras que se agrupan en el clúster 3, doce muestras que se agrupan en el clúster 4, 10 muestras que se agrupan en el clúster 5 y siete muestras que se agrupan en el clúster 6.

- Superior Temporal Gyrus: cinco clúster (óptimo) en donde se tiene 11 muestras en el clúster 1, 19 muestras en el clúster 2, ocho muestras en el clúster 3, 10 muestras en el clúster 4 y 12 muestras en el clúster 5.

- Temporal Pole: tres clúster (óptimo) en donde se tiene 21 muestras en el clúster 1, 27 muestras en el clúster 2 y 10 muestras en el clúster 3.

- Amígdala: cuatro clúster (óptimo) en donde se tiene 11 muestras en el clúster 1, 15 muestras en el clúster 2, 13 muestras en el clúster 3 y 12 muestras en el clúster 4.
- Región Hipocampo: cuatro clúster (óptimo) en donde se tiene 3 muestras en el clúster 1, 12 muestras en clúster 2, 16 muestras en clúster 3 y 24 muestras en clúster 4.

Al igual que en el caso anterior, en el agrupamiento de genes por k-means no se obtiene ningún grupo homogéneo en relación al nivel de neuropatología. Además, tal y como se observa en la Figura 9 que corresponde al perfil de expresión del primer clúster de cada tejido, es muy difícil de interpretar.

Por ello, por motivos computacionales y por escasez de tiempo, se hace la anotación funcional y el análisis de enriquecimiento sobre los genes seleccionados que se encuentran entre el 1% de las desviaciones estándar más altas (la tabla completa se puede ver en carpeta ANEXO 7, Anotación para distintos tejidos en formato .html). Se podría entender, por la naturaleza del experimento que se realiza para obtener los niveles de intensidad de señal, que dichos genes se corresponden con aquellos que presentan sobreexpresión. Es decir, teniendo en cuenta que el nivel de intensidad de señal se corresponde con el nivel de expresión de un gen (realmente se analizan n veces la intensidad de un gen) cuando la señal es más alta, se desvía (en este caso en el 1%), es porque hay mayor expresión.

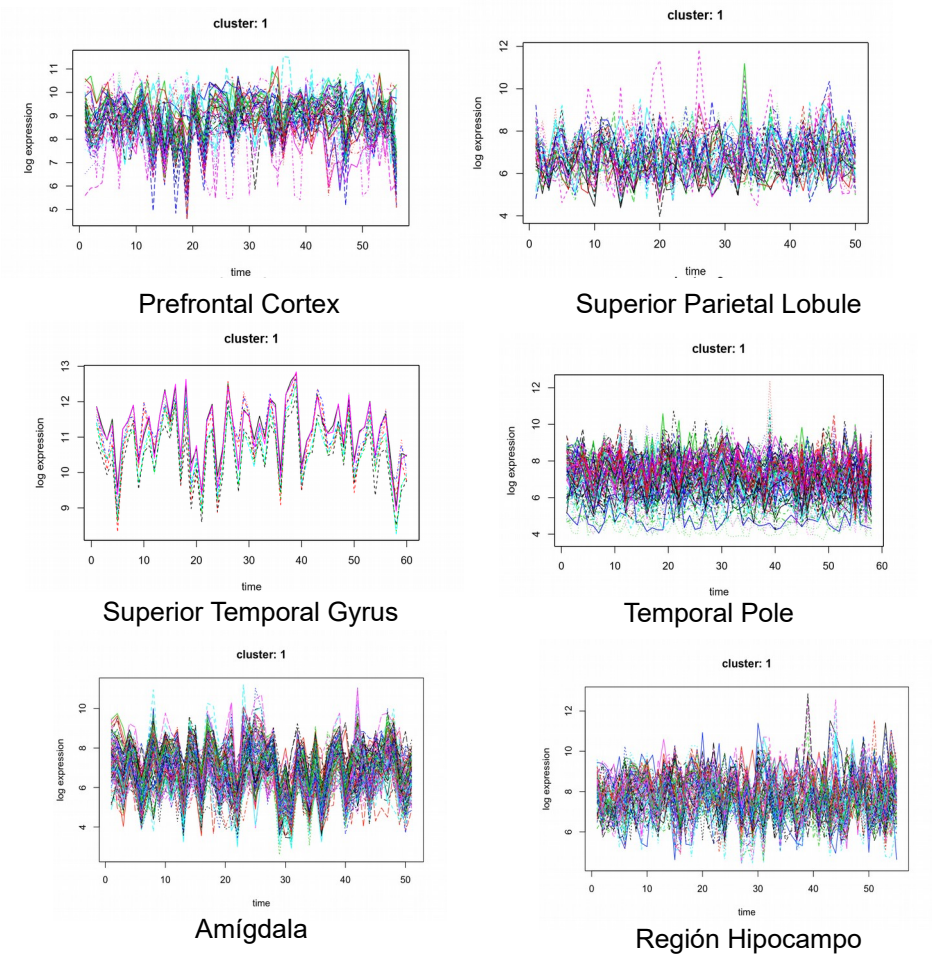


Figura 9 Perfil deExpresión de los clúster 1 (en Descubrimiento de clases) de todos los tejidos

Dado el número elevado de genes que se obtienen en cada tejido mediante el método de descubrimiento de clases, se muestran, en orden alfabético, en la Tabla 8 los 223 genes detectados en cada uno de los tejidos (excepto en el Amígdala que se detecta 547 genes).

Prefrontal Cortex	**Superior Parietal Lobule**	**Superior Temporal Gyrus**	**Temporal Pole**	**Región Hipocampo**
ACOT7	ACOT7	ABCA8	ACOT7	ACOT7
ACTB	ACTB	ACTB	ACTA2	ACTA2
ADD3	ADD3	ADAM23	AMPH	ACTB
ADM	ADD3	APOLD1	APOLD1	ANXA1
APOLD1	ADM	AQP4	AQP4	APOE
AQP4	ANKRD36B	AQP4	AQP4	AQP4
AQP4	AQP4	AQP4	AQP4	AQP4
AQP4	AQP4	ARL6IP5	ARPP19	ARPP21
ARPP21	AQP4	ARPP19	ATP1A3	ATP1A2
ATP1A3	AQP4	ARPP21	ATP2B1	ATP1A3
ATP1B1	ATP1A3	ATP1A3	ATP6V1A	ATP6V1A
ATP2B1	ATP1B1	ATP1B1	ATP6V1B2	ATP6V1G2
ATP2B1	ATP2B1	ATP2B1	ATP6V1G2	ATXN1
ATP6V1A	ATP6AP2	ATP6AP2	ATRNL1	C1QB
ATRNL1	ATP6V1A	ATP6V1A	ATXN1	CA2
ATXN1	ATP6V1B2	ATP6V1B2	BAG3	CABP1
BBX	ATRNL1	ATXN1	C10orf10	CABP1
C10orf10	ATXN1	C10orf10	C1QB	CABP1
C1QB	BEX4	C1QB	CABP1	CACNG3
CABP1	C10orf10	CABP1	CALY	CALY
CALY	C1QB	CALY	CCK	CAMK2B
CAMK2B	CABP1	CAMK2B	CD14	CAMK2B
CAMTA1	CAMK2B	CAP2	CD163	CAMTA1
CAPN3	CCK	CAPN3	CD163	CAPN3
CCK	CD44	CCK	CD44	CCK
CD14	CDK5R1	CCL2	CDK5R1	CD14
CD163	CDR1			

CD163	CEBPD	CD14	CEBPD	CD163
CD44	CHGB	CD163	CHGB	CD163
CDK5R1	CHI3L1	CD163	CHI3L1	CD24
CEBPD	CHI3L1	CD44	CHI3L1	CD24
CHGB	CHL1	CDK5R1	CHL1	CD44
CHI3L1	CLDND1	CEBPD	CLASP2	CDK5R1
CHI3L1	CNN3	CHGB	CNN3	CHGB
CHL1	CXCL14	CHI3L1	CNR1	CHI3L1
CIRBP	DCLK1	CHI3L1	CX3CR1	CHI3L1
CLASP2	DDX17	CHL1	CYP1B1	CHL1
CNN3	DDX17	CLASP2	DCLK1	CHN1
CNR1	DDX3Y	CNN3	DCLK1	CNN3
DCLK1	DICER1	CTBP1	DDX17	CNR1
DDX3Y	DIRAS2	DCLK1	DDX17	CX3CR1
DICER1	DTNA	DDX3Y	DIRAS2	DCLK1
DICER1	DTNA	DIRAS2	DMXL2	DDX17
DIRAS2	EEF1A2	DSTN	DOCK3	DDX3Y
DTNA	EGFR	DTNA	DSP	DIRAS2
DTNA	EGR4	DTNA	DTNA	DNM1
DTNA	ENC1	DTNA	DYNC1I1	EEF1A2
EEF1A2	ENO1	EEF1A2	EEF1A2	EGR4
EGFR	EPB41L3	EGR4	EIF5A	ENC1
EGR4	EPB41L3	EIF5A	ENC1	ENC1
EIF5A	FAM3C	ENC1	ENC1	ENO1
ENC1	FBXO21	ENC1	ENO2	EPB41L3
ENC1	FOS	ENO1	EPB41L3	F5
ENO1	G3BP2	ENPP2	ERBB2IP	FAM3C
ETNPPL	GABBR2	EPB41L3	ETNPPL	FGF13
F3	GABRA1	EPB41L3	FAM3C	FOLR1
FAM3C	GABRA2	EVI2A	G3BP2	FOS
G3BP2	GABRG2	FAM3C	G3BP2	FOXG1
GABBR2	GAD1	FBXO21	GABBR2	FRRS1L
GABBR2	GADD45B	FGF13	GABBR2	GABBR2
GABBR2	GADD45B	GABBR2	GABBR2	GABBR2
GABRA1	GAP43	GABBR2	GABRA1	GABBR2
GABRA2	GLRB	GABBR2	GABRA2	GABRA5
GABRA5	GLRB	GABRA1	GABRA5	GAD1
GABRG2	GLUL	GABRA2	GABRG2	GADD45B
GAD1	GLUL	GABRA5	GAD1	GAP43
GADD45B	GNAO1	GABRG2	GADD45B	GAPDH
GAP43	GOT1	GAD1	GAP43	GBP1
GLRB	HBB	GADD45B	GBP1	GLUL
GLRB	HBB	GAP43	GJA1	GLUL
GLUL	HBB	GAS7	GLRB	GPX3
GLUL	HILPDA	GBP1	GLRB	GPX3
GOT1	HPRT1	GLRB	GLS	HBB
HAMP	HSP90AA1	GLRB	GLS	HBB
HBB	HSP90AB1	GLS	GLS	HBB
HBB	IDS	GLUL	GNAL	HLA-DPA1
HBB	IER5	GNAO1	GOT1	HLA-DPA1
HILPDA	IL1RL1	GOT1	GUCY1B3	HLA-DRA
HPRT1	ISG15	GRIA2	HAMP	HLA-DRA
HSP90AA1	ITPR1	GUCY1B3	HBB	HPRT1
ID4	KCNB1	HAMP	HBB	HSP90AB1
IER5	KDM5D	HBB	HBB	HTR2C
IL1RL1	KIF5C	HBB	HLA-DPA1	ID4
INPP5F	KLC1	HBB	HLA-DRA	IFI44L
ISG15	KLC1	HLA-DRA	HLA-DRA	IGFBP7
ITPR1	KRAS	HLA-DRA	HPCAL1	IL1RL1
KDM5D	KTN1	HLF	HPRT1	INA
KIAA1107	MAFF	HPRT1	HSPB1	INPP5F
KIF5C	MAP1B	HSP90AA1	ID4	ISG15
KLC1	MBNL2	HSP90AB1	IDS	KCNQ2
KLC1	MBP	HSPB1	IFI44L	KIF5C
KTN1	MBP	IDS	ISG15	KLC1
MAFF	MCTP1	IL1RL1	ITFG1	KLC1
MBNL2	MDH1	ITPR1	KCTD12	LDB2
MBP	MEF2C	KIF5C	KLC1	LPPR4
MBP	MEG3	KLC1	KLC1	LY6H
MCTP1	MOAP1	KLC1	KTN1	MAFF
MDH1	MSMO1	LY6H	LY6H	MAG
MLLT11	MYT1L	MAFF	MAFF	MBP
MOAP1	NAP1L2	MBP	MAP2	MBP
MT1M	NAP1L3	MCTP1	MDH1	MCTP1
MYT1L	NBEA	MDH1	MEF2C	MEF2C

NAP1L2	NCALD	MEF2C	MEF2C	MGP
NAP1L3	NDRG2	MYT1L	MOAP1	MKL2
NCALD	NEFH	NAP1L2	MT1M	MLLT11
NDRG2	NEFL	NAP1L3	MYT1L	MOAP1
NEFH	NEFM	NBEA	NAP1L2	MYH11
NEFL	NELL2	NCALD	NAP1L3	MYL9
NEFL	NPTX2	NDRG2	NBEA	MYT1L
NEFM	NRN1	NEFH	NCALD	NAP1L2
NELL2	NRXN1	NEFL	NEFH	NAP1L3
NMNAT2	NSF	NEFL	NEFL	NBEA
NPTX2	NTRK2	NEFM	NEFL	NDRG2
NRN1	NTRK2	NELL2	NEFM	NEFH
NSF	OPCML	NMNAT2	NELL2	NEFH
NTRK2	OXR1	NPTN	NMNAT2	NEFL
NTRK2	PCDH8	NPTX2	NPTX2	NEFL
OPCML	PCDH9	NRN1	NRN1	NEFM
OSBPL8	PCMT1	NSF	NRXN1	NELL2
PCDH8	PCMT1	NTRK2	NSF	NMNAT2
PCDH9	PDP1	NTRK2	OPCML	NPTX2
PCMT1	PEG3	OLFM1	OXR1	NPTXR
PCP4	PI4KA	PCMT1	PCDH8	NRGN
PDP1	PJA2	PCP4	PCMT1	NSF
PEG3	PLK2	PEG3	PEG3	NTRK2
PLK2	PMP2	PIK3R1	PI4KA	OAZ1
PMP2	PNMAL1	PJA2	PLK2	OLFM1
PNMAL1	PPAP2B	PLCB1	PNMA1	OLFM1
PPAP2B	PPAP2B	PLK2	PNMA2	OPCML
PPAP2B	PREPL	PMP2	PNMAL1	PCDH8
PPP3CA	PREPL	PNMA2	PPP3CA	PCP4
PREPL	PRKACB	PNMAL1	PPP3CB	PEG3
PREPL	PRKAR1A	PPAP2B	PREPL	PI4KA
PRKACB	PRKCB	PPAP2B	PREPL	PLD3
PRKCB	PRKCI	PPP3CA	PRKACB	PLK2
PSD3	PSD3	PREPL	PRKCB	PNMA2
PSD3	PSD3	PREPL	PSD3	PNMAL1
RAP2A	RAB2A	PRKACB	PSD3	POLR2E
RCAN2	RAN	PRKAR1A	PSD3	PPAP2B
REEP1	RAP2A	PRKCB	REEP1	PPP3CA
RGS1	RGS1	PSD3	RGS1	PPP3CA
RGS4	RGS4	PSD3	RGS4	PREPL
RGS4	RGS4	RAP2A	RGS4	PRKCB
RHOBTB3	RGS4	RBFOX1	RGS4	PROS1
RHOBTB3	RGS5	RGS1	RGS7	PSD3
RNF6	RHOBTB3	RGS4	RPS4Y1	PSD3
RPS4Y1	RHOBTB3	RGS4	RTN1	PVALB
RTN1	RNF6	RGS4	S100A8	RBFOX1
S100A8	RPS4Y1	RGS7	SCD5	RCAN2
S100A9	RTN1	RHOBTB3	SCG2	REEP1
SCD5	S100A8	RNF6	SCG5	RGS1
SCG2	SCD5	RPS4Y1	SCN3B	RGS1
SCG5	SCG2	RTN1	SERPINA3	RGS4
SERPINA3	SCG5	S100A8	SERPINI1	RGS4
SERPINI1	SCN8A	SCG2	SH3GL2	RGS4
SH3GL2	SEC23A	SCG5	SLC17A7	RGS7
SLC17A7	SERPINA3	SCN3B	SLC26A2	RPS4Y1
SLC1A2	SERPINI1	SERPINA3	SLC47A1	RTN1
SNAP25	SH3GL2	SERPINI1	SLCO4A1	S100A8
SNAP25	SLC17A7	SH3GL2	SNAP25	SCD5
SNAP91	SLC1A2	SLC17A7	SNAP25	SCG2
SNCA	SNAP25	SNAP25	SNAP91	SCN3B
SOX9	SOX9	SNAP25	SNCA	SCN3B
SRGN	SPP1	SNAP91	SNCA	SERPINA3
SRGN	SRGN	SNCA	SNCB	SERPINF1
SRRM2	SRGN	SOX9	SPP1	SERPINI1
STMN2	SRRM2	SPP1	SRGN	SFRP1
STMN2	SST	SRGN	SRGN	SH3GL2
SYN2	STMN2	SRGN	SST	SLC13A4
SYNGR3	STMN2	STMN2	STMN2	SLC17A6
SYNJ1	SUCLA2	STMN2	STMN2	SLC17A7
SYT1	SYNJ1	SYN2	SUB1	SLC17A7
SYT1	SYT1	SYNJ1	SYN2	SNAP25
TAGLN	SYT1	SYT1	SYNJ1	SNAP25
THY1	THY1	SYT1	SYT1	SNAP91
TNPO1	TNPO1	TAGLN	SYT1	SNCB
TPD52	TPD52	TF	TAGLN	SRGN

TPR	TPD52	TF	THY1	STMN2
TUBA4A	TPR	THY1	TOMM20	STMN2
TXNIP	TUBA4A	TIMP1	TPD52	SV2C
UCHL1	TXNIP	TPD52	TPD52	SYNGR3
VAMP3	UCHL1	TPR	TUBA4A	SYNJ1
VEGFA	VDAC1	TUBA4A	TXNIP	SYT1
VSNL1	VEGFA	TXNIP	UCHL1	SYT1
VSNL1	**VSNL1**	UCHL1	VDAC1	TAGLN
WAC	WAC	VAMP1	**VSNL1**	TCF7L2
XIST	XIST	VDAC1	**VSNL1**	TCF7L2
XIST	XIST	**VSNL1**	WIF1	THY1
YBX3	YBX3	**VSNL1**	XIST	TMEM144
YWHAB	YWHAB	WIF1	XIST	TPD52
YWHAB	YWHAB	XIST	YBX3	TPH1
YWHAE	YWHAE	XIST	YWHAB	TPM2
YWHAH	YWHAH	YBX3	YWHAB	TTR
YWHAZ	YWHAZ	YWHAB	YWHAH	UCHL1
YWHAZ	YWHAZ	YWHAB	YWHAZ	VAMP1
	YY1	YWHAE	YWHAZ	**VSNL1**
	ZBTB38	YWHAH	ZBTB20	**VSNL1**
	ZC3H15	YWHAZ	ZIC1	XIST
		YWHAZ		XIST
				YBX3
				YWHAB
				YWHAB
				YWHAH
				YWHAZ
				ZIC1

Tabla 8 Genes seleccionados, para todos los tejidos excepto el tejido Amígdala, por el método de agrupamiento de clases

Se observa en esta comparación que en los cinco tejidos se seleccionan mayoritariamente los mismos genes. La mayoría de los genes implicados se encuentran en el cromosoma 1. El gen señalado (VSNL1,Visinin-like protein 1) está relacionado con el Alzheimer y es un gen que ya se había obtenido en otros estudios. Este gen da lugar a proteínas neuronales de unión al calcio. La proteína codificada se expresa fuertemente en las membranas de células granulares del cerebelo, al ser dependiente del ión Calcio modula las vías de señalización intracelular del sistema nervioso central al regular directa o indirectamente la actividad de la adenilil ciclasa. Las enfermedades asociadas con VSNL1 incluyen encefalopatía (aguda con convulsiones bifásicas y difusión tardía reducida) y enfermedad de Alzheimer (https://www.genecards.org/cgi-bin/carddisp.pl?gene=VSNL1&keywords=VSNL1)

Como se ve este gen se sitúa en el cromosoma 2. La gran mayoría de genes, en lo poco que se ha podido analizar, están implicados en procesos estructurales celulares (adhesinas, tubulinas, etc) , en procesos de ubiquitinización y en procesos dependientes de calcio que modulan las vías de señalización intracelular. Este gen, el VSNL1 también es seleccionado en el Tejido Amígdala así como otros detectados en los otros tejidos. Además, algunos de los genes seleccionados por agrupamiento de clases también fueron seleccionados por expresión diferencial.

2. 2.4 Método de Regresión Penalizado

Con este método de análisis de Data Mining se obtienen modelos predictivos donde se seleccionan las variables predictoras (en este caso los SNPs) asociadas a variables respuestas cuantitativas de interés. En el artículo [26] (en su material suplementario) se facilitan datos clínicos y anatomohistopatológicos (postmorten) de los 125 pacientes de esta cohorte (ver Material Suplementario 2, Tabla S1 y Tabla S2). Con dichos datos se definen dos variable cuantitativas: una global clínica y una global anatomohistopatológica. Con ello se describe el fenotipo fisiopatológico del individuo y se pretende intentar asociarlo a aquellos polimorfismos (SNPs), que según su nivel de expresión, puedan ser los responsables de dicho fenotipo. El resultado de asociar las variables elaboradas (clínica y anatomohistopatológica) se observa en los modelos que se obtienen para los distintos tejidos para cada una de las variables (ver Informes de obtención de Snps para variable clínica y variable anatomohistopatológica para cada tejido en formato pdf que se adjunta).

Se seleccionan los modelos que tienen un buen coeficiente de determinación (R^2). Esta es una medida del ajuste del modelo y es la proporción de variabilidad de la variable dependiente explicada por el modelo de regresión. Ahora bien, este coeficiente se considera que sobrevalora la variabilidad explicada ya que al construirse el modelo con las muestras de train, el modelo estará sobreajustado. Y aumenta según se incorporan variables predictoras (en este caso Snps)[20]

Por ello se usa, además, para evaluar la capacidad predictiva del modelo la raíz cuadrada del error cuadrático medio (RMSE). Esta medida es una de las más importante que se usan y es una medida de la falta de ajuste del modelo a los datos (cuantifica la diferencia entre las observaciones observadas y las predichas). Es una medida calculada que varía por muestra[20]

Lóbulo Frontal

En el Tejido Prefrontal Cortex, para la *variable respuesta global clínica* el modelo seleccionado es: (nota: la X que aparece delante de cada sonda se genera al ejecutar el script y se quita, manualmente con un editor de texto, al hacer la anotación y el enriquecimiento funcional)

X202730_s_at X202756_s_at X203696_s_at X203830_at X204555_s_at X204919_at X205145_s_at X205761_s_at X206653_at X208005_at X208102_s_at X209665_at X209790_s_at X210473_s_at X210630_s_at X212188_at X216091_s_at X218739_at X220409_at X220666_at
X220726_at X220932_at

$R2
[1] 0.9128502

$RMSEtrain
[1] 0.2736989

$RMSEtest
[1] 0.4377576

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos (ver Informe de Obtención de Snps para la variable global clínica para el Tejido Prefrontal Cortex en formato pdf) se ve que la mayoría tienen un R^2 variable, por lo que el elegido, 0.91, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,27, uno de los más bajos, mientras el RMSE de train del resto de los modelos es también variable. Y, además, la muestra testing tiene uno de los valores menores (0,44) en relación al resto de la mayoría de los modelos pero no se diferencia mucho. Por lo tanto, el mejor modelo es el seleccionado: tiene un alto coeficiente de determinación (0,91) junto con un RMSE para train más bajo respecto a los demás (0,27; son los datos que se utilizan para construir el modelo) y, además, es el que tiene el RMSE (diferencia entre observadas y predichas) para testing que no se diferencia mucho del resto de los modelos y es ligeramente más bajo que los demás.[20]

En este modelo seleccionado tenemos 22 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Prefrontal Cortex para variable clínica) el test de ANOVA nos indica que alguno de los 22 coeficientes (snps) de regresión es distinto de cero (con un p=3.981e-09). En los distintos test de la t de Student se ve que dichos coeficientes son X202730_s_at (p=0.00294), X206653_at (p=0.02736), X209790_s_at (p=0.00584), X210473_s_at (p=0.00581) y X220932_at (p=0.01568). Este modelo, es decir, los 22 Snps en conjunto, explican un 87% de la variabilidad de la variable vgpc (un 78% si usamos el valor ajustado).

Por último se realiza un Análisis Univariante de los 22 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Prefrontal Cortex para variable clínica), de los 22 modelos estimados 19 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). De hecho aquellos que tienen un p-value muy bajo explican en torno al 25% de dicha variabilidad. Los 3 modelos estimados restantes NO explican nada de la variabilidad global de la variable vgpc (test F del ANOVA, $p > 0.05$).

La anotación funcional y análisis de enriquecimiento de los 19 Snps con relación significativa con la variable respuesta global clínica se puede observar en la Tabla 9 (la tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Prefrontal Cortex para variable clínica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
202730_s_at				NM_014456	
202756_s_at	GPC1	glypican 1	2	NM_002081	2q35-q37
203696_s_at	RFC2	replication factor C (activator 1) 2, 40kDa	7	NM_002914	7q11.23
203830_at	C17orf75	chromosome 17 open reading frame 75	17	NM_022344	17q11.2
204555_s_at	PPP1R3D	protein phosphatase 1, regulatory subunit 3D	20	NM_006242	20q13.3
204919_at				NM_007244	
205145_s_at	MYL5	myosin, light chain 5, regulatory	4	NM_002477	4p16.3
206653_at	POLR3G	polymerase (RNA) III (DNA directed) polypeptide G (32kD)	5	BF062139	5q14.3
208005_at	NTN1	netrin 1	17	NM_004822	17p13.1
208102_s_at	PSD	pleckstrin and Sec7 domain containing	10	NM_002779	10q24
209790_s_at	CASP6	caspase 6, apoptosis-related cysteine peptidase	4	BC000305	4q25
210473_s_at	ADGRA3	adhesion G protein-coupled receptor A3	4	M37712	4p15.2
210630_s_at	RAD52	RAD52 homolog (S. cerevisiae)	12	AF125949	12p13-p12.2
212188_at	KCTD12	potassium channel tetramerization domain containing 12	13	AA551075	13q22.3
216091_s_at	BTRC	beta-transducin repeat containing E3 ubiquitin protein ligase	10	AF101784	10q24.32
218739_at	ABHD5	abhydrolase domain containing 5	3	NM_016006	3p21

220409_at	CAMSAP1	calmodulin regulated spectrin-associated protein 1	9	NM_018627	9q34.3
220666_at				NM_018611	
220726_at				NM_025100	

Tabla 9 Resumen Anotación Funcional del Tejido Prefrontal Cortex para Método Regresión Penalizado para la variable global clínica

Al consultar los genes seleccionados en bases de datos constrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 19 genes, se seleccionan 4 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

PPP1R3D La fosforilación de residuos de serina y treonina en proteínas es un paso crucial en la regulación de muchas funciones celulares que van desde la regulación hormonal hasta la división celular e incluso la memoria a corto plazo. Es un gen codificador de proteínas Entre sus vías relacionadas están la señalización beta-adrenérgica y la activación de PKA dependiente de AMPc

MYL5 implicado en la PAK pathway

CASP6 Este gen codifica un miembro de la familia de enzimas cisteína-ácido aspártico (caspasa). La activación secuencial de las caspasas juega un papel central en la fase de ejecución de la apoptosis celular

KCTD12 relacionado con la activación de PKA dependiente de AMPc

En el Tejido Prefrontal Cortex, para la *variable respuesta global anatomohistopatológica* el modelo seleccionado es:

X201035_s_at X202756_s_at X203013_at X203604_at X203736_s_at X203785_s_at X204919_at X205223_at X205314_x_at X205742_at X205756_s_at X206088_at X206234_s_at X206459_s_at X206653_at X207723_s_at X208110_x_at X208371_s_at X208903_at X210505_at X213477_x_at X214014_at X214333_x_at X214371_at X214951_at X215389_s_at X216109_at X216626_at X217333_at X218267_at X218512_at X219294_at X219654_at X219670_at X220318_at X220703_at X220762_s_at X220795_s_at X220807_at X222066_at X32032_at

$R2
[1] 0.9995861

$RMSEtrain
[1] 0.07127321

$RMSEtest
[1] 1.525097

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 bastante alto, por lo que el elegido, 0.99, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,07, el más bajo. Y, además, la muestra testing tiene uno de los valores menores (1,52) en relación al resto de la mayoría de los modelos pero no se diferencia mucho del valor del resto de los modelos. Por lo tanto, el mejor modelo es el seleccionado: tiene un alto coeficiente de determinación (0,99) junto con un RMSE para train más bajo respecto a los demás (0,07; son los datos que se utilizan para construir el modelo) y, además, es el que tiene el RMSE (diferencia entre observadas y predichas) para testing más bajo y que no se diferencia mucho del resto de los modelos.

En este modelo seleccionado tenemos 41 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjuntado (A.Multivariante y A. Univariante para Tejido Prefrontal Cortex para variable anatomohistopatológica) el test de ANOVA nos indica que alguno de los 41 coeficientes (snps) de regresión es distinto de cero (p= 0.0001688). En los distintos test de la t de Student se ve que dichos coeficientes son X203736_s_at (p=0.019885), X214333_x_at (p= 0.041415), X219654_at (p=0.000418) y X220807_at (p= 0.04660). Este modelo, es decir, los 41 Snps en conjunto, explican un 95% de la variabilidad de la variable vgph (un 82% si usamos el valor ajustado).

Por último se realiza un Análisis Univariante de los 41 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Prefrontal Cortex), de los 41 modelos estimados 36 explican una parte significativa de la variabilidad global de la variable vgph (test F del ANOVA, $p < 0.05$). De hecho aquellos que tienen un p-value muy bajo explican en torno al 25% de dicha variabilidad. Los 5 modelos estimados restantes NO explican nada de la variabilidad global de la variable vgph (test F del ANOVA, $p > 0.05$).

La anotación funcional y análisis de enriquecimiento de los 36 Snps con relación significativa con la variable respuesta global anatomohistopatológica se puede observar en la Tabla 10 (la tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Prefrontal Cortex para variable anatomohistopatológica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
201035_s_at	HADH	hydroxyacyl-CoA dehydrogenase	4	BC000306	4q22-q26
202756_s_at	GPC1	glypican 1	2	NM_002081	2q35-q37
203013_at	ECD	ecdysoneless homolog (Drosophila)	10	NM_007265	10q22.3
203604_at	ZNF516	zinc finger protein 516	18	N38750	18q23
203736_s_at	PPFIBP1	PTPRF interacting protein, binding protein 1 (liprin beta 1)	12	NM_003622	12p12.1
203785_s_at	DDX28	DEAD (Asp-Glu-Ala-Asp) box polypeptide 28	16	NM_018380	16q22.1
204919_at				NM_007244	
205223_at	DEPDC5	DEP domain containing 5	22	NM_014662	22q12.3
205742_at	TNNI3	troponin I type 3 (cardiac)	19	NM_000363	19q13.4
205756_s_at	F8	coagulation factor VIII, procoagulant component	X	NM_000132	Xq28
206234_s_at	MMP17	matrix metallopeptidase 17 (membrane-inserted)	12	NM_016155	12q24.3
206459_s_at	WNT2B	wingless-type MMTV integration site family, member 2B	1	AB045117	1p13
206653_at	POLR3G	polymerase (RNA) III (DNA directed) polypeptide G (32kD)	5	BF062139	5q14.3
208903_at	RPS28	ribosomal protein S28	19	BF431363	19p13.2

210505_at	ADH7	alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide	4	U07821	4q23-q24
214014_at	CDC42EP2	CDC42 effector protein (Rho GTPase binding) 2	11	W81196	11q13
214371_at	TSSK2	testis-specific serine kinase 2	22	AI652441	22q11.21
215389_s_at	TNNT2	troponin T type 2 (cardiac)	1	X79857	1q32
216109_at	MED13L	mediator complex subunit 13-like	12	AK025348	12q24.21
216626_at				AL050026	
217333_at				AL031903	
218267_at	CINP	cyclin-dependent kinase 2 interacting protein	14	NM_016550	14q32.31
218512_at	WDR12	WD repeat domain 12	2	NM_018256	2q33.2
219294_at	CENPQ	centromere protein Q	6	NM_018132	6p12.3
219654_at	HACD1	3-hydroxyacyl-CoA dehydratase 1	10	NM_014241	10p12.33
219670_at	BEND5	BEN domain containing 5	1	NM_024603	1p33
220318_at	EPN3	epsin 3	17	NM_017957	17q21.33
220703_at	IDI2-AS1	IDI2 antisense RNA 1	10	NM_018470	10p15.3
220795_s_at	BEGAIN	brain-enriched guanylate kinase-associated	14	NM_020836	14q32.2
220807_at	HBQ1	hemoglobin, theta 1	16	NM_005331	16p13.3
222066_at	EPB41L1	erythrocyte membrane protein band 4.1-like 1	20	AA573523	20q11.2-q12
32032_at				L77566	

Tabla 10 Resumen Anotación Funcional del Tejido Prefrontal Cortex para Método Regresión Penalizado para la variable global anatomohistopatológica

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 36 genes, se seleccionan 2 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

MMP17 Este gen codifica una peptidasa que está involucrada en la descomposición de la matriz extracelular en procesos fisiológicos normales
EPN3 es un gen codificador de proteína. Entre sus vías relacionadas están PAK (Proteína activadora de quinasa) Pathway

Lóbulo Parietal

En el Tejido Superior Parietal Lobule tanto para la *variable respuesta global clínica* como para la *variable respuesta global anatomohistopatológica* NO se selecciona ningún modelo ya que en ambos casos los parámetros de evaluación de los modelos tienen todos el mismo valor (Ver informes de Selección de Polimorfismos para variable global clínica y variable global anatomohistopatológica para el tejido Superior Parietal Lobule).

Lóbulo Temporal

Tejido Superior Temporal Gyrus

En el Tejido Superior Temporal Gyrus, para la *variable respuesta global clínica* el modelo seleccionado es:

X202489_s_at X202776_at X203650_at X203815_at X203934_at X204055_s_at X204100_at X204135_at X205248_at X207765_s_at X208146_s_at X210524_x_at X210913_at X211887_x_at X212056_at X213430_at X213566_at X214006_s_at X214112_s_at X214322_at X214696_at X216930_at X217893_s_at X218383_at X218655_s_at X219364_at X219410_at X220539_at X221025_x_at X221967_at

$R2
[1] 0.9397115

$RMSEtrain
[1] 0.2413876

$RMSEtest
[1] 0.6036565

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 en torno a 0,55 y algunos en torno a 0,30, por lo que el elegido, 0.94, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,24, de los más bajos. Si es verdad que la muestra testing tiene uno de los valores mayores (0,60) en relación al resto de la mayoría de los modelos. Por lo tanto, el mejor modelo es el seleccionado: tiene un alto coeficiente de determinación (0,94) junto con un RMSE para train bajo respecto a los demás (0,24; son los datos que se utilizan para construir el modelo).

En este modelo seleccionado tenemos 30 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjuntado(A.Multivariante y A. Univariante para Tejido Superior Temporal Gyrus para variable global clínica) . El test de ANOVA nos indica que algunos de los 30 coeficientes (Snps) de regresión es distinto de cero (p=1.746e05). En los distintos test de la t de Student se ve que dichos coeficientes son X204135_at (p= 0.01988), X210913_at(p=0.02101) y X214696_at(p=0.00589). Este modelo explica un 84% de la variabilidad de la variable vgpc (un 67% si usamos el valor ajustado)

Por último se realiza un Análisis Univariante de los 30 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Superior Temporal Gyrus), de los 30 modelos estimados 24 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). Los 6 modelos estimados restantes NO explican nada de la variabilidad global de la variable vgpc (test F del ANOVA, $p > 0.05$).

La anotación funcional y análisis de enriquecimiento de los 24 Snps con relación significativa con la variable respuesta global clínica se puede observar en la Tabla 11 (tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Superior Temporal Gyrus para variable clínica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
202489_s_at	FXYD3	FXYD domain containing ion transport regulator 3	19	BC005238	19q13.12
202776_at	DNTTIP2	deoxynucleotidyltransferase, terminal, interacting protein 2	1	NM_014597	1p22.1
203815_at	GSTT1	glutathione S-transferase theta 1	22	NM_000853	22q11.23
204135_at	FILIP1L	filamin A interacting protein 1-like	3	NM_014890	3q12.1
205248_at	DOPEY2	dopey family member 2	21	NM_005128	21q22.2
207765_s_at	FAM214B	family with sequence similarity 214, member B	9	NM_025182	9p13.3
208146_s_at	CPVL	carboxypeptidase, vitellogenic-like	7	NM_031311	7p15.1
210913_at	CDH20	cadherin 20, type 2	18	AF217289	18q21.33
212056_at	GSE1	Gse1 coiled-coil protein	16	D80004	16q24.1
213430_at	RUFY3	RUN and FYVE domain containing 3	4	BF224071	4q13.3
213566_at	RNASE6	ribonuclease, RNase A family, k6	14	NM_005615	14q11.2
214112_s_at				AA543076	
214322_at	CAMK2G	calcium/calmodulin-dependent protein kinase II gamma	10	AA284757	10q22
214696_at				AF070569	
217893_s_at	AKIRIN1	akirin 1	1	NM_024595	1p34.3
218383_at				NM_017815	
218655_s_at	CWC25	CWC25 spliceosome-associated protein homolog (S. cerevisiae)	17	NM_017748	17q12
219364_at	DHX58	DEXH (Asp-Glu-X-His) box polypeptide 58	17	NM_024119	17q21.2
219410_at	TMEM45A	transmembrane protein 45A	3	NM_018004	3q12.2
220539_at	CFAP46	cilia and flagella associated protein 46	10	NM_017609	10q26.3
221967_at	NXPH4	neurexophilin 4	12	AI933199	12q13.3

Tabla 11 Resumen Anotación Funcional del Tejido Superior Temporal Gyrus para Método Regresión Penalizado para la variable global clínica

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 24 genes, se seleccionan 1 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

Las enfermedades asociadas con **GSE1** incluyen neoplasias neuroectodérmicas primitivas del sistema nervioso central.

En el Tejido Superior Temporal Gyrus, para la *variable respuesta global anatomohistopatológica* el modelo seleccionado es:

X201035_s_at X201069_at X202822_at X203650_at X204938_s_at X205453_at X205609_at X206091_at X215492_x_at X217893_s_at X219985_at X222325_at X34868_at X58916_at

$R2
[1] 0.8151795

$RMSEtrain
[1] 1.341967

$RMSEtest
[1] 1.94277

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 por debajo a 0,80, por lo que el elegido, 0.81, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 1,34 mientras el RMSE de train del resto de los modelos está de media por encima de 1,40. Si es verdad que la muestra testing tiene uno de los valores mayores (1,94) en relación al resto de la mayoría de los modelos aunque la RMSE de testing en el resto de modelos no es tan diferente al elegido

Por lo tanto, el mejor modelo es el seleccionado: tiene un alto coeficiente de determinación (0,81) junto con un RMSE para train, el más bajo (1,34; son los datos que se utilizan para construir el modelo) y, además, es el que tiene el RMSE (diferencia entre observadas y predichas) para testing que no se diferencia mcho del resto de los modelos.

En este modelo seleccionado tenemos 14 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjuntado(A.Multivariante y A. Univariante para Tejido Superior Temporal Gyrus para variable global anatomohistopatológica) . El test de ANOVA nos indica que alguno de los 14 coeficientes (snps) de regresión es distinto de cero (p=3.246e-06). En los distintos test de la t de Student se ve que dichos coeficientes son X222325_at (p= 0.048383) y X34868_at(p=0.000729).Este modelo explica un 64%% de la variabilidad de la variable vgph (un 53% si usamos el valor ajustado)

Por último se realiza un Análisis Univariante de los 14 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Superior Temporal Gyrus), los 14 modelos estimados explican una parte significativa de la variabilidad global de la variable vgph (test F del ANOVA, $p < 0.05$).

La anotación funcional y análisis de enriquecimiento de los 14 Snps con relación significativa con la variable respuesta global anatomohistopatológica se puede observar en la Tabla 12 (tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Superior Temporal Gyrus para variable anatomohistopatológica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
201035_s_at	HADH	hydroxyacyl-CoA dehydrogenase	4	BC000306	4q22-q26
201069_at	MMP2	matrix metallopeptidase 2	16	NM_004530	16q12.2
202822_at	LPP	LIM domain containing preferred translocation partner in lipoma	3	BF221852	3q28

203650_at	PROCR	protein C receptor, endothelial	20	NM_006404	20q11.2
204938_s_at	PLN	phospholamban	6	M60411	6q22.1
205453_at	HOXB2	homeobox B2	17	NM_002145	17q21.32
205609_at	ANGPT1	angiopoietin 1	8	NM_001146	8q23.1
206091_at	MATN3	matrilin 3	2	NM_002381	2p24-p23
217893_s_at	AKIRIN1	akirin 1	1	NM_024595	1p34.3
219985_at	HS3ST3A1	heparan sulfate (glucosamine) 3-O-sulfotransferase 3A1	17	NM_006042	17p12
222325_at				AW974812	
34868_at	SMG5	SMG5 nonsense mediated mRNA decay factor	1	AB029012	1q21.2
58916_at				AI672101	

Tabla 12 Resumen Anotación Funcional del Tejido Superior Temporal Gyrus para Método Regresión Penalizado para la variable global anatomohistopatológica

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 14 genes, no se seleccionan ninguno ya que no existe en esta selección ningún gen que esté implicado en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular.

Tejido Temporal Pole

En el Tejido Temporal Pole, para la *variable respuesta global clínica* el modelo seleccionado es:

X200713_s_at X201897_s_at X202081_at X204302_s_at X205086_s_at X206019_at X208415_x_at X208561_at X209998_at X213119_at X213493_at X216004_s_at X218722_s_at X219864_s_at X220127_s_at X37802_r_at X81811_at

$R2
[1] 0.889319

$RMSEtrain
[1] 0.3203463

$RMSEtest
[1] 0.5840187

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 medio, por lo que el elegido, 0.89, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,32, un valor más bajo que en el resto de modelos. Además la muestra testing tiene un valor ligeramente inferior al resto de modelos (0,58) pero no se diferencia mucho. Por lo tanto, el mejor modelo es el seleccionado: tiene un aceptable coeficiente de determinación (0,89) junto con un RMSE para train más bajo respecto a los demás (0,32; son los datos que se utilizan para construir el modelo) y, además, es el que tiene un RMSE para testing aceptable.

En este modelo seleccionado tenemos 17 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable

respuesta global clínica. Tal y como se observa en el Informe adjuntado(A.Multivariante y A. Univariante para Tejido Temporal Pole para variable global clínica) . El test de ANOVA nos indica que alguno de los 17 coeficientes (snps) de regresión es distinto de cero (p=9.564e-07). En los distintos test de la t de Student se ve que dichos coeficientes son X208561_at (p= 0.0269) y X37802_r_at(p=0.0087).Este modelo explica un 73% de la variabilidad de la variable vgpc (un 61% si usamos el valor ajustado)

Por último se realiza un Análisis Univariante de los 17 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Temporal Pole), de los 17 modelos estimados 12 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). Los 5 modelos restantes NO explican nada de la variabilidad global de la variable vgpc(test F del ANOVA, $p > 0.05$)

La anotación funcional y análisis de enriquecimiento de los 12 Snps con relación significativa con la variable respuesta global clínica se puede observar en la Tabla 13 (tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Temporal Pole para variable clínica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
200713_s_at	MAPRE1	microtubule-associated protein, RP/EB family, member 1	20	NM_012325	20q11.1-q11.23
204302_s_at	CTIF	CBP80/20-dependent translation initiation factor	18	U55962	18q21.1
205086_s_at	NCAPH2	non-SMC condensin II complex, subunit H2	22	NM_014551	22q13.33
206019_at	RBM19	RNA binding motif protein 19	12	NM_014852	12q24.21
208561_at	ABCC9	ATP-binding cassette, sub-family C (CFTR/MRP), member 9	12	NM_020297	12p12.1
213493_at	SNED1	sushi, nidogen and EGF-like domains 1	2	BF509657	2q37.3
218722_s_at	CCDC51	coiled-coil domain containing 51	3	NM_024661	3p21.31
219864_s_at	RCAN3	RCAN family member 3	1	NM_013441	1p36.11
220127_s_at	FBXL12	F-box and leucine-rich repeat protein 12	19	NM_017703	19p13.2
37802_r_at	FAM63B	family with sequence similarity 63, member B	15	AL049226	15q21.3
81811_at				AI744451	

Tabla 13 Resumen Anotación Funcional del Tejido Temporal Pole para Método Regresión Penalizado para la variable global clinica

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 12 genes, se seleccionan 1 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

MAPRE1 Esta proteína se localiza en los microtúbulos, especialmente en los extremos crecientes, en las células interfásicas. Entre sus vías relacionadas se encuentran los eventos de señalización de N-cadherina y PAK Pathway.

En el Tejido Temporal Pole, para la *variable respuesta global anatomohistopatológica* NO se selecciona ningún modelo ya que los coeficientes de determinación están en torno a 0,47. Y la RMSE de train presenta valores mayores que los de test. Por lo tanto, ningún modelo es aceptable y no se selecciona ninguno

Sistema Límbico

Tejido Amígdala

En el Tejido Amígdala, para la *variable respuesta global clínica* el modelo seleccionado es:

X1558793_at X1558900_at X1562919_at X203243_s_at X203504_s_at X206700_s_at X209600_s_at X222738_at X227945_at X235424_at

$R2
[1] 0.866815

$RMSEtrain
[1] 0.3237223

$RMSEtest
[1] 0.538844

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 en torno a 0,83, por lo que el elegido, 0.87, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,32 mientras el RMSE de train del resto de los modelos está de media en 0,36. La muestra testing tiene un valor de 0.53 que no se diferencia del que se observa en el resto de modelos. Por lo tanto, el mejor modelo es el seleccionado: tiene un aceptable coeficiente de determinación (0,87) junto con un RMSE para train más bajo respecto a los demás (0,32; son los datos que se utilizan para construir el modelo) y, además, es el que tiene el RMSE (diferencia entre observadas y predichas) para testing que no se diferencia mucho del resto de los modelos.

En este modelo seleccionado tenemos 10 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjuntado(A.Multivariante y A. Univariante para Tejido Amígdala para variable global clínica). El test de ANOVA nos indica que alguno de los 10 coeficientes (snps) de regresión es distinto de cero (p=6.46e08). En los distintos test de la t de Student se ve que dichos coeficientes son X206700_s_at (p=0.000908) y X235424_at (p=0.002646) Este modelo explica un 71% de la variabilidad de la variable vgpc (un 63% si usamos el valor ajustado)

Por último se realiza un Análisis Univariante de los 10 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Amígdala), de los 10 modelos estimados 8 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). De hecho aquellos que tienen un p-value muy bajo explican en torno al 25% de dicha variabilidad. Los 2 modelos estimados restantes NO explican nada de la variabilidad global de la variable vgpc (test F del ANOCVA, $p > 0.05$).

La anotación funcional y análisis de enriquecimiento de los 8 Snps con relación significativa con la variable respuesta global clínica se puede observar en la Tabla 14 (tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Amígdala para variable clínica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
206700_s_at	KDM5D	lysine (K)-specific demethylase 5D	Y	NM_004653	Yq11

Tabla 14 Resumen Anotación Funcional del Tejido Amígdala para Método Regresión Penalizado para la variable global clínica

Como se observa con el método de anotación funcional utilizado sólo se obtiene un gen.Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que este gen, no está implicado en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular.

En el Tejido Amígdala, para la *variable respuesta global anatomohistopatológica* el modelo seleccionado es:

X1555896_a_at X1556849_at X1561136_at X1561882_at X1563188_at X1568616_a_at X200646_s_at X200649_at X201439_at X204919_at X205001_s_at X207036_x_at X207473_at X212892_at X214691_x_at X216027_at X216931_at X217263_x_at X218500_at X220677_s_at X221479_s_at X221728_x_at X221851_at X222205_x_at X222679_s_at X223161_at X225346_at X225412_at X228449_at X230921_s_at X230958_s_at X235303_at X237597_at X238154_at X243206_at X243755_at

$R2
[1] 0.9996077

$RMSEtrain
[1] 0.08068911

$RMSEtest
[1] 1.691882

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos se ve que la mayoría tienen un R^2 variable alto, por lo que el elegido, 0.99, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Pero, además, tiene un RMSE en la muestra de train de 0,08, uno de los más bajo. Además la muestra testing tiene un valor de 1,69, algo más bajo que en los otros modelos con un coeficiente de determinación y RMSE de train similar al escogido. Por lo tanto, el mejor modelo es el seleccionado: tiene un alto coeficiente de determinación (0,99) junto con un RMSE para train bajo (0,08; son los datos que se utilizan para construir el modelo)

En este modelo seleccionado tenemos 37 coeficientes (Snps) con los que se realiza un Análisis Multivariante para ver cuánto explican, en conjunto, de la variabilidad de la variable respuesta global clínica. Tal y como se observa en el Informe adjuntado(A.Multivariante y A. Univariante para Tejido Amígdala para variable global anatomohistopatológico). El test de ANOVA nos indica que alguno de los 37 coeficientes (snps) de regresión es distinto de cero (p=0.0003884). En los distintos test de la t de Student se ve que dichos coeficientes son X1561882_at (p=0.0141). Este modelo explica un 94% de la variabilidad de la variable vgph (un 80% si usamos el valor ajustado)

Por último se realiza un Análisis Univariante de los 37 Snps seleccionados. Tal y como se ve en el Informe adjunto (A.Multivariante y A. Univariante para Tejido Amígdala), De los 37 modelos estimados 35 explican una parte significativa de la variabilidad global de la variable vgph (test F del ANOVA, $p < 0.05$). De hecho aquellos que tienen un p-value muy bajo explican en torno al 25% de dicha variabilidad e incluso en algunos casos están en torno al 45%. Los 2 modelos estimados restantes NO explican nada de la variabilidad global de la variable vgph (test F del ANOCVA, $p >0.05$).

La anotación funcional y análisis de enriquecimiento de los 35 Snps con relación significativa con la variable respuesta global anatomohistopatológica se puede observar en la Tabla 15 (tabla completa se adjunta en formato .html Snps obtenidos para el Tejido Amígdala para variable anatomohistopatológica)

Probe	Symbol	Description	Chromosome	GenBank	Cytoband
1555896_a_at	ADAM15	ADAM metallopeptidase domain 15	1	BM973999	1q21.3
1556849_at				AU146310	
1561136_at	GYPE	glycophorin E (MNS blood group)	4	AF085899	4q31.1
1561882_at	SYTL3	synaptotagmin-like 3	6	BC042966	6q25.3
1563188_at	LOC102723448	uncharacterized LOC102723448	3	BC039672	
200646_s_at	NUCB1	nucleobindin 1	19	NM_006184	19q13.33
200649_at	NUCB1	nucleobindin 1	19	BC002356	19q13.33
201439_at	GBF1	golgi brefeldin A resistant guanine nucleotide exchange factor 1	10	NM_004193	10q24
204919_at				NM_007244	
205001_s_at	DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	Y	AF000985	Yq11
207473_at	MLN	motilin	6	NM_002418	6p21.3
212892_at	ZNF282	zinc finger protein 282	7	AW130128	7q36.1
216027_at	TMX4	thioredoxin-related transmembrane protein 4	20	AI005473	20p12
216931_at				L23852	
218500_at	THEM6	thioesterase superfamily member 6	8	NM_016647	8q24.3
220677_s_at	ADAMTS8	ADAM metallopeptidase with thrombospondin type 1 motif, 8	11	NM_007037	11q25
221479_s_at	BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like	8	AF060922	8p21
221851_at	DCAF15	DDB1 and CUL4 associated factor 15	19	AI073983	19p13.12
222679_s_at	DCUN1D1	DCN1, defective in cullin neddylation 1, domain containing 1	3	AW468880	3q26.3
223161_at	KIAA1147	KIAA1147	7	AA029331	7q34
225346_at	MTERF2	mitochondrial transcription termination factor 2	12	NM_025198	12q24.1

225412_at	TMEM87B	transmembrane protein 87B	2	AA761169	2q13
228449_at				BG260069	
230921_s_at				BE467612	
230958_s_at				BE670797	
235303_at	TRMT10B	tRNA methyltransferase 10 homolog B (S. cerevisiae)	9	AV728846	9p13.2
235424_at	FAM122A	family with sequence similarity 122A	9	N66727	9q21.11
237597_at				AI655887	
238154_at	CEP70	centrosomal protein 70kDa	3	AI285884	3q22.3
243755_at	PRLR	prolactin receptor	5	AI628734	5p13.2

Tabla 15 Resumen Anotación Funcional del Tejido Amígdala para Método Regresión Penalizado para la variable global anatomohistopatológica

Al consultar los genes seleccionados en bases de datos contrastadas (como Genecards, OMIM, DECIPHER, principalmente) así como buscar información mediante Genómica Computacional (básicamente en Ensembl, USCS Browser) se observa que de los 35 genes, se seleccionan 1 por que están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general así como con procesos de destrucción celular. Los resúmenes de la información de los genes se han obtenido, casi literalmente, de GeneCards (https://www.genecards.org/Search/)

KIAA1147 Puede desempeñar un papel en la neuritogénesis, así como en la recuperación y / o reestructuración neuronal en el hipocampo después de la isquemia cerebral leve.

Tejido Región Hipocampo

En el Tejido Región Hipocampo, para la *variable respuesta global clínica* el modelo seleccionado es:

Se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos (adjunto) se ve que la mayoría tienen un R^2 muy bajo, es decir el ajuste del modelo no es bueno. El que presenta mayor coeficiente de determinación, 0.78, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Sin embargo se observa que el valor de RMSE para train y para test son muy similares. En otras palabras no se obtiene ningún modelo satisfactorio.

En el Tejido Región , para la *variable respuesta global anatomohistopatológica* el modelo seleccionado es:

Igual que en el caso anterior se debe escoger aquel modelo que tenga el mayor R^2 con el menor RMSE en el grupo de testing. Si se observan los datos de los modelos obtenidos (adjunto) se ve que la mayoría tienen un R^2 muy bajo, es decir el ajuste del modelo no es bueno. El que presenta mayor coeficiente de determinación, 0.72, es el que mejor valor tiene (indica un buen ajuste, su valor oscila entre 0 y 1). Sin embargo se observa que el valor de RMSE para train y para test son muy similares. En otras palabras no se obtiene ningún modelo satisfactorio.

2.3 Discusión

Desde distintas áreas científicas se están realizando muchos esfuerzos para poder caracterizar la etiología de la enfermedad del Alzheimer. Y a ello contribuye, además, el avance tecnológico que se ha producido en el campo de la genética con el advenimiento de tecnologías como la NGS y los Microarrays (expresión) que permiten analizar simultáneamente una gran cantidad de datos

.

El cerebro es el soporte anatomo-funcional de la actividad psíquica humana: cognitiva, afectiva y motivacional y sigue un patrón de maduración propio de la especie humana. Distintos estudios realizados han determinado cuáles son las principales áreas funcionales de la corteza cerebral: Lóbulo frontal, Lóbulo Parietal, Lóbulo Temporal y Lóbulo Occipital (todos ellos relacionados con el área ósea en la que están ubicados). Desde finales del siglo XX y principios del siglo XXI se ha establecido un área más, subcortical, denominada Sistema Límbico.

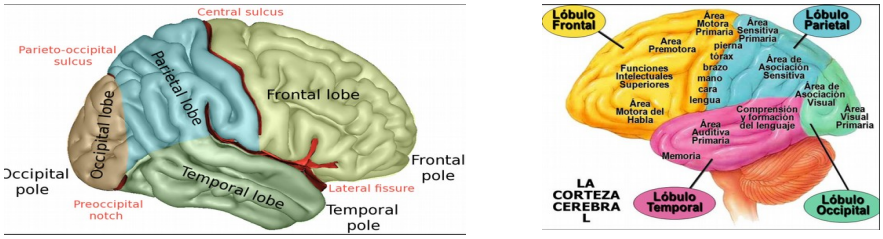


Figura 10 Áreas Funcionales cerebrales

Para este trabajo se ha seleccionado aquella área del tejido nervioso relacionado con las funciones reseñadas como alteradas en el Alzheimer: memoria, lenguaje, orientación o percepción espacial entre otras.

Lóbulo Frontal | Lóbulo Parietal | Lóbulo Temporal | Sistema Límbico

Figura 11 Áreas de Tejido Seleccionadas

Así del Lóbulo Frontal se selecciona el Tejido Prefrontal Cortex situado en el área prefrontal que ocupa la mitad anterior del lóbulo frontal. Funcionalmente este lóbulo está relacionado con el pensamiento, el juicio, el intelecto, la atención, el comportamiento y el pensamiento abstracto. En otras palabras en lo relacionado con lo cognitivo.

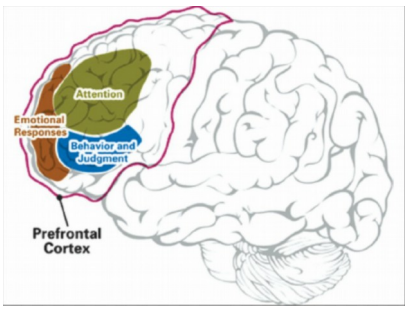


Figura 12 Área Cerebral Prefrontal Cortex

Del Lóbulo Parietal se selecciona el Tejido Superior Parietal Lobule. El lóbulo parietal está situado en la zona posteriosuperior de la corteza cerebral, detrás del surco central y por encima del lóbulo occipital. Funcionalmente este lóbulo interpreta la información sobre el tacto y los receptores de extensión de los músculos y articulaciones. Está relacionado con las áreas encefálicas que controlan el movimiento, con el cálculo, con la orientación y ciertos tipos de reconocimiento así como la memoria.

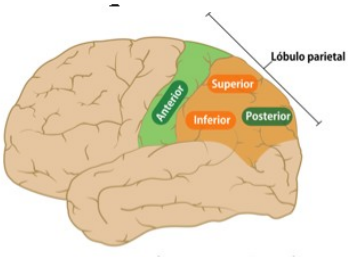


Figura 13 Área Cerebral Lobule Parietal: Superior Parietal Lobule

Del Lóbulo Temporal se seleccionan los Tejidos Superior Temporal Gyrus y el Temporal Pole. Este lóbulo comprende todo el tejido situado por debajo de la cisura de Silvio y por delante de la corteza occipital. Es el receptor principal de la información auditiva y se considera esencial para el lenguaje hablado. Funcionalmente está relacionado con la memoria visual, auditiva y comprensión del habla. Y, además, incluye el área de lenguaje (área de Wernicke).

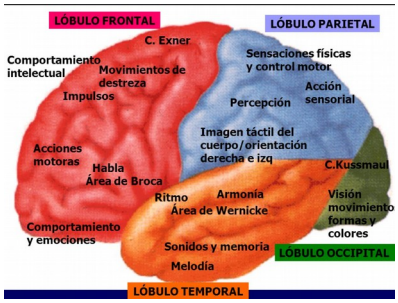


Figura 14 Área Cerebral Lobule Temporal: Superior Temporal Gyrus y Temporal Pole

Y del Sistema Límbico se selecciona el Tejido Amígdala y el Tejido Región Hipocampo, ambos relacionados, principalmente, en la literatura con la memoria, una de las características que define el Alzheimer. El sistema límbico está constituido por un conjunto de estructuras cuya función está relacionada con el aprendizaje, la memoria y con las respuestas emocionales. Es el que configura nuestra personalidad, nuestros recuerdos.

Es el que nos hace ser como somos. Las estructuras que lo constituyen son: la amígdala, el tálamo, hipotálamo, hipófisis, hipocampo, el área septal (fórnix, cuerpo calloso y fibras de asociación), corteza orbitofrontal y la circunvolución del cíngulo. Funcionalmente está relacionado con la memoria, la atención, los instintos sexuales, las emociones (como el miedo, el placer y la agresividad).

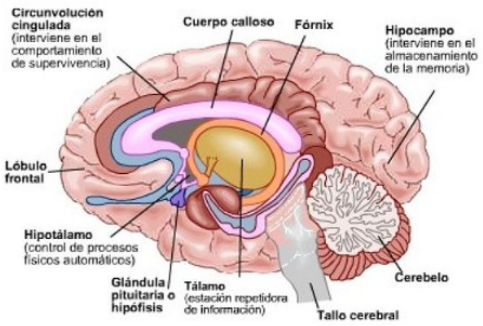


Figura 15 Área Cerebral Sistema Límbico: Amígdala y Región Hipocampo

Se puede decir que son los tejidos que están principalmente relacionados con las funciones alteradas que definen el Alzheimer. Con estos tejidos se realizan diversos Microarrays de expresión (Affymetrix) y los datos obtenidos se analizan, en este trabajo, con tres métodos diferentes.

En el análisis de expresión diferencial se observa que no se obtienen ningún gen diferencialmente expresado en el Tejido Temporal Pole y el Tejido Región Hipocampo. Pero en los otros cuatro tejidos sí se obtienen genes diferencialmente expresados:

- En el tejido Prefrontal Cortex : se obtienen 8 genes en la primera comparación (A vs B), 7 de los cuales están down regulados y 1 up regulado.

- En el Superior Parietal Lobule: se obtienen en la tercera comparación (A vs D) 36 genes down regulados, en la cuarta comparación (B vs C) 1 gen down regulado y en la quinta comparación (B vs D) 27 genes down regulados. Como genes up regulados tenemos: 14 en la cuarta comparación (B vs C) y 3 en la quinta comparación (B vs D).

- En el tejido Superior Temporal Gyrus se obtiene en la primera comparación (A vs B) 18 genes up regulados.

- En el tejido Amígdala: se obtiene en la primera comparación(A vs B) 17 genes up regulados y 31 genes down regulados, en la tercera comparación (A vs C) se obtienen 9 genes up regulados y 3 genes down regulados.

Se observa, por tanto, que el tejido que presenta mayor número de genes diferencialmente expresados es el Superior Parietal Lobule que está relacionado funcionalmente con el cálculo, con la orientación y ciertos tipos de reconocimiento así como la memoria.

En general el perfil de expresión es down regulado y donde se obtiene la mayor diferencia de expresión en todos los tejidos, excepto en el Superior Parietal Lobule, es cuando hacemos el contraste Normal vs Definitivo. En el tejido Superior Parietal el mayor número de genes down regulados se obtienen en las comparaciones Normal vs Probable y Definitivo vs Probable.

Además se puede ver que del Tejido Prefrontal Cortex se obtiene la anotación funcional para 8 genes de los cuales 4 están implicados en procesos patológicos relacionados con patologías cerebrales y con la Demencia en general. Estos genes son *ADD3*, *PCP4, PMP2, MBP.*

En el tejido Superior Parietal Lobule se obtiene la anotación funcional para 58 genes, de los cuales 10 están implicados en procesos patológicos relacionados con patologías cerebrales y con la demencia en general. Son: PJA2, TPR ,TRAK2, KIF5C, GNAO1, GABRA1,CDR1, SLC1A2, KCNB1, TCF4, TMEM106B (relacionado con demencia frontotemporal), RAP2A.

En el tejido Superior Temporal Gyrus se obtiene la anotación funcional para 18 genes de los cuales 5 están implicados en procesos patológicos relacionados con patología cerebrales y con la demencia en general. Estos genes son: BTBD3, STMN2 (relacionado con enfermedad de Alzheimer) NEFM, SYNJ1, NEFL

En el tejido de la Amígdala se obtiene la anotación funcional para 48 genes de los cuales 3 están implicados en procesos patológicos relacionados con patología cerebral y con la demencia en general son: NLGN4Y, CD24, CPLX1

Por lo tanto, por expresión diferencial se obtienen dos genes claramente identificados, relacionados, con las demencias: el TMEM106B relacionado con la demencia frontotemporal y el STMN2 relacionado con la enfermedad de Alzheimer. Y el tejido Superior Parietal Lobule es el que presenta mayor número de genes down regulados seguido del tejido Amígdala.

En el análisis de Descubrimiento de clases no se observa ningún agrupamiento homogéneo de muestras con ninguno de los algoritmos utilizados. Sin embargo al hacer la selección de los genes que se encuentran entre el 1% de las desviaciones estándar más altas,

se obtiene el mismo número de genes en todos los tejidos (223) excepto en la Amígdala que se obtienen 547. Al hacer la anotación funcional y análisis de enriquecimiento se observa que la mayoría de los genes están presentes en los 5 tejidos mencionados e incluso en el tejido Amígdala (Ver Tabla 8).

Se observa que la gran mayoría de los genes analizados, están implicados en procesos estructurales celulares, en procesos de ubiquitinización y en procesos dependientes de calcio que modulan las vías de señalización intracelular. Destaca el gen VSNL1 (relacionado con el Alzheimer) y que se encuentra en todos los tejidos, incluida la Amígdala. Algunos de los genes que fueron seleccionados por agrupamiento de clases también fueron seleccionados por expresión diferencial.

En el análisis de Regresión Penalizada en relación a la variable respuesta global clínica se observa

- En el tejido Prefrontal Cortex se obtienen 22 genes que en conjunto, explican un 87% de la variabilidad de la variable vgpc (un 78% si usamos el valor ajustado). De los 22 modelos estimados por A. Univariante, 19 explican (individualmente) una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 19 genes se seleccionan 4 por que están implicados en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular. Son: PPP1R3D, MYL5 CASP6, KCTD12

- En el Tejido Superior Temporal Gyrus se obtienen 30 genes, que en conjunto, explican un 84% de la variabilidad de la variable vgpc (un 67% si usamos el valor ajustado). De los 30 modelos estimados por A. Univariante, 24 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 24 genes se seleciconan 1 por que está implicado en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular. Es GSE1

- En el Tejido Temporal Pole se obtienen 17 genes, que en conjunto, explican explica un 73% de la variabilidad de la variable vgpc (un 61% si usamos el valor ajustado) . De los 17 modelos estimados por A. Univariante, 12 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 12 genes se seleccionan 1 por que está implicado en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular. Es MAPRE1

- En el Tejido Amígdala se obtienen 10 genes, que en conjunto, explica un 71% de la variabilidad de la variable vgpc (un 63% si usamos el valor ajustado). De los 10 modelos estimados por A. Univariante, 8 explican una parte significativa de la variabilidad global de la variable vgpc (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 8 genes no se selecciona ninguno por que en las bases consultadas, ninguno está implicado en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular.

En el análisis de Regresión Penalizada en relación a la variable respuesta global anatomohistopatológica:

- En el tejido Prefrontal Cortex se obtienen 41 genes que en conjunto explican un 95% de la variabilidad de la variable vgph (un 82% si usamos el valor ajustado). De los 41 modelos estimados por A. Univariante 36 explican una parte significativa de la variabilidad global de la variable vgph (test F del ANOVA, $p < 0.05$) En la anotación funcional de esos 36 genes se seleccionan 2 por que están implicados en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular. Son MMP17, EPN3

- En el Tejido Superior Temporal Gyrus se obtienen 14 genes, que en conjunto, explica un 64% de la variabilidad de la variable vgph (un 53% si usamos el valor ajustado) Los 14 modelos estimados por A. Univariante explican una parte significativa de la variabilidad global de la

variable vgph (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 14 genes no se selecciona ninguno porque en las bases consultadas, ninguno está implicado en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular.

- En el Tejido Temporal Pole para esta variable NO se selecciona ningún modelo ya que los coeficientes de determinación están en torno a 0,47. Y la RMSE de train presenta valores mayores que los de test. Por lo tanto, ningún modelo es aceptable y no se selecciona ninguno. Con este método no se puede seleccionar ningún gen en el tejido Temporal Pole.

- En el Tejido Amígdala se obtienen 37 genes, que en conjunto, explica un 94% de la variabilidad de la variable vgph (un 80% si usamos el valor ajustado) De los 37 modelos estimados por A. Univariante 35 explican una parte significativa de la variabilidad global de la variable vgph (test F del ANOVA, $p < 0.05$). En la anotación funcional de esos 35 genes se selecciona 1 gen que está implicado en procesos patológicos relacionados con patología cerebral y con Demencia en general así como en procesos de destrucción celular. Es KIAA1147

En definitiva, se ha podido establecer asociación significativa de la variable respuesta global clínica y los genes seleccionados en los cuatro tejidos: Prefrontal Cortex, Superior Temporal Gyrus, Temporal Pole y Amígdala.

Y asociación significativa de la variable respuesta global anatomohistopatológica y los genes seleccionados en los tres tejidos: Prefrontal Cortex, Superior Temporal Gyrus y Amígdala.

Se puede ver como en los Tejidos Superior Parietal Lobule y Región Hipocampo no se obtiene ningún modelo bueno ni para la variable global clínica ni para la variable global anatomohistopatológica. Para esta última variable tampoco se obtiene un buen modelo para el tejido Temporal Pole.

Se observa, además, que los modelos con una mejor asociación entre los snps seleccionados y la variable respuesta global clínica se da en los Tejidos Prefrontal Cortex y el Tejido Superior Temporal Gyrus. Mientras que en relación a la variable respuesta anatomohistopatológica la mejor asociación está en el Prefrontal Cortex y el Tejido Amígdala.

Si se valoran los tres métodos conjuntamente es una constante encontrar genes relacionados con la activación de PKA dependiente de AMPc, son genes que están relacionados con procesos patológicos cerebrales y con Demencia en general así como en procesos de destrucción celular. También se observa un gran número de genes que forman parte de la PAK pathway.

3. Conclusiones

Con los tres métodos de análisis utilizados se han obtenido distintas Dianas genéticas (SNPs) relacionadas con los procesos que caracterizan a la enfermedad del Alzheimer. A nivel molecular, básicamente, genes implicados en procesos de destrucción celular.

Algunos genes están claramente identificados y relacionados con las demencias. Tenemos por ejemplo TMEM106B (relacionado con demencia frontotemporal), STMN2 y VSNL1, ambos, relacionados con la enfermedad de Alzheimer.

Otros genes seleccionados (por cualquiera de los tres métodos) están relacionados con la activación de PKA dependiente de AMPc. Por ejemplo: ADD3, PJA2, TCF4. Las proteínas quinasas dependientes de AMPc juegan un papel muy importante en diversos procesos celulares.

Es muy relevante que en los 5 tejidos corticales se obtenga el mismo número de genes al utilizar el método de Agrupamiento de clases. Concretamente cuando se seleccionan sólo aquellos genes que se encuentran entre el 1% de las desviaciones estándar más altas. Hay que tener en cuenta que el número de muestras no es el mismo en todos los tejidos. Y, es significativo, que la mayoría de los 223 genes seleccionados en los 5 tejidos sean iguales (y algunos estén entre los genes seleccionados en el tejido Amígdala)

En este trabajo he aprendido a organizar el tiempo. Al utilizar tres métodos diferentes he podido valorar los pros y los contras de cada uno. Si bien, se necesita más experiencia para poder determinar cuál es el más adecuado desde un punto de vista ortodoxo. Ha sido gratificante ver, que aunque los fundamentos usados en cada método eran diferentes, y los genes obtenidos son distintos, se puede concluir que la mayoría pertenecen a la misma categoría genética.

Todos los objetivos planteados se han logrado. Sin embargo, he de decir, que el trabajo ha sido exhaustivo. Tras finalizarlo, si tuviese que volver a planificarlo, decididamente me quedaría sólo con un método y, quizás, con aquellos tejidos más relevantes en la enfermedad del Alzheimer. Ahora bien, no me arrepiento, ha sido gratificante y sirve de experiencia. Pero es tal la cantidad de datos obtenida que se necesita más tiempo para hacer un análisis más profundo.

Me ha gustado la metodología y la planificación. Si bien he de decir que han ido surgiendo imponderables, no planificados, que han hecho variar brevemente la planificación inicial. Sin embargo, he de criticar el aspecto formal. Es decir, no se considera como tarea planificada y son muchos los detalles a tener en cuenta para que la memoria tenga una más que aceptable presentación. En mi caso creo habrá muchos detalles formales a los que, por falta de tiempo, no he podido prestarles la atención debida.

A lo largo de la memoria he reiterado, hasta la saciedad, el nombre de algunos documentos que he utilizado como base para desarrollar esta memoria. Ha sido largo, en algunos casos difícil, pero ha merecido la pena. Con lo leído en ciertos artículos y con las conclusiones obtenidas, queda mucho por profundizar en los resultados de expresión de tejidos de pacientes con la enferemedad de Alzheimer. En concreto llama la atención lo de los 223 genes y hace que me plantee si no forman parte de alguna red génica. De hecho algunas investigaciones van en esa línea.

4. Glosario

AMPc	Adenosín Monofosfato cíclico
APOE4	Apolipoproteína
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
CDR	clinical dementia raiting
GABA	Ácido gamma aminobutírico
GAD	Glutamato Descarboxilasa
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
NGS	Next Generation Sequencing
NUSE	Normalized Unscaled Standard Errors
OMS	Organización Mundial de la Salud
PAK	Proteína activadora de quinasa
PKA	Proteína quinasa A
PPA	Proteína precursora amieloide
PSEN	presenilinas
RLE	relative log expression
RMA	robust multi-array average
RMSE	raíz cuadrada del error cuadrático medio
Snps	Single Nucleotide Polymorphims
TFM	Trabajo Fín de Máster
TRAIN	Entrenamiento
TEST	Evaluación
Vgpc	Variable global clínica
Vgph	Variable global anatomohistopatológica
within SS	cuadrados dentro de cada grupo

5. Bibliografía

[1] Aevarsson O, Svanborg A, Skoog I. Seven-year survival rate after age 85 years: relation to Alzheimer disease and vascular dementia. Arch Neurol. 1998 Sep;55(9):1226-32.

[2] Bermejo-Pareja F, Benito-Leon J, Vega S, Medrano MJ, Roman GC. Incidence and subtypes of dementia in three elderly populations of central Spain. J Neurol Sci. 2008 Jan 15;264(1-2):63-72.

[3] Bekris L, Yu C, Bird T, Tsuang D. (2010). «Genetics of Alzheimer Disease.». *J Geri Psyc Neur.* **23**: 213-227.

[4] Biessels GJ, Kappelle LJ; Utrecht Diabetic Encephalopathy Study Group. [Increased risk of Alzheimer's disease in Type II diabetes: insulin resistance of the brain or insulin-induced amyloid pathology?] (en inglés). Biochem Soc Trans. 2005 Nov;33(Pt 5):1041-4.

[5] Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A 2004;101:2173–2178. [PubMed: 14769913]

[6] Brodaty H, Breteler MM, Dekosky ST, Dorenlot P, Fratiglioni L, Hock C, et al. The world of dementia beyond 2020. J Am Geriatr Soc. 2011 May;59(5):923-7

[7] Bussiere T, Gold G, Kovari E, Giannakopoulos P, Bouras C, Perl DP, Morrison JH, Hof PR. Stereologic analysis of neurofibrillary tangle formation in prefrontal cortex area 9 in aging and Alzheimer's disease. Neuroscience 2003;117:577–592. [PubMed: 12617964]

[8] Cristina Prieto Jurczynska, Miriam Eimil Ortiz, Carlos López de Silanes de Miguel, Marcos Llanero Luque IMPACTO SOCIAL DE LA ENFERMEDAD DE ALZHEIMER Y OTRAS DEMENCIAS 2011 (fundación española de enfermedades neurológicas)

[9] Chen X, Yan SD (diciembre de 2006). «Mitochondrial Abeta: a potential cause of metabolic dysfunction in Alzheimer's disease». *IUBMB Life* **58** (12): 686-94. PMID17424907

[10] Emilsson L, Saetre P, Jazin E. Alzheimer's disease: mRNA expression profiles of multiple patients show alterations of genes involved with calcium signaling. Neurobiol Dis 2006;21:618–625. [PubMed: 16257224]

[11] Ferri CP, Prince M, Brayne C, Brodaty H, Fratiglioni L, Ganguli M, et al. Global prevalence of dementia: a Delphi consensus study. Lancet. 2005 Dec 17;366(9503):2112_7

[12] Förstl H, Maelicke A, Weichel C. Prólogo: Epidemiología. In: Förstl H, Maelicke A, Weichel C, editors. Demencia: J & C Ediciones Médicas; 2007. p. 2-7

[13] Gerda G. Fillenbaum, PhD1,2,*, Gerald van Belle, PhD3, John C. Morris, MD4, Richard C. Mohs, PhD5, Suzanne S. Mirra, MD6, Patricia C. Davis, MD7, Pierre N. Tariot, MD8, Jeremy M. Silverman, PhD9, Christopher M. Clark, MD10, Kathleen A. Welsh-Bohmer, PhD11, and Albert Heyman, MD12 CERAD (Consortium to Establish a Registry for Alzheimer's Disease) The first 20 years Alzheimers Dement. 2008 March ; 4(2): 96–109

[14] Ginsberg SD, Che S, Counts SE, Mufson EJ. Single cell gene expression profiling in Alzheimer's disease. NeuroRx 2006;3:302–318. [PubMed: 16815214]

[15] G. K. Smyth (febrero, 2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments". Statistical Applications in Genetics and Molecular Biology (vol. 3, núm. 1). http://www.degruyter.com/view/j/sagmb.2004.3.1/sagmb.2004.3.1.1027/sagmb.2004.3.1.1027.xml.

[16] Hardy J, Allsop D (octubre de 1991). «Amyloid deposition as the central event in the aetiology of Alzheimer's disease». *Trends Pharmacol. Sci.* **12** (10): 383-88.

[17] Hardy J. A hundred years of Alzheimer's disease research. Neuron 2006;52:3–13. [PubMed: 17015223]

[18] Heidi C. Rossetti, M.S.1, C Munro Cullum, PhD1,3, Linda S. Hynan, PhD2,1, and Laura Lacritz, PhD1 The CERAD Neuropsychological Battery Total Score and the Progression of Alzheimer's Disease Alzheimer Dis Assoc Disord . 2010 ; 24(2): 138–142.

[19] Hernández F, Avila J (septiembre de 2007). «Tauopathies». *Cell. Mol. Life Sci.* **64** (17): 2219-33. PMID17604998

[20] Herranz Valera J. Técnicas Estadísticas de Data Mining con R (2016)

[21] J. J. Faraway (2004). Linear Models with R (2.ª ed.). Chapman and Hall/CRC.

[22] Jellinger KA, Attems J. Prevalence of dementia disorders in the oldest-old: an autopsy study. Acta Nuropathol. 2010 Apr;119(4):421-33.

[23] Kalaria RN, Maestre GE, Arizaga R, Friedland RP, Galasko D, Hall K, et al. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. Lancet Neurol. 2008 Sep;7(9):812-26.

[24] López-Pousa S, Garre-Olmo J. La demencia: concepto y epidemiología. In: Alberca R, López-Pousa S, editors. Enfermedad de Alzheimer y otras demencias. Madrid: Editorial Panamericana; 2010. p. 29-38

[25] Malone DC, McLaughlin TP, Wahl PM, Leibman C, Arrighi HM, Cziraky MJ, et al. Burden of Alzheimer's disease and association with negative health outcomes. Am J Manag Care. 2009 Aug;15(8):481-8.

[26] Minghui Wang, Panos Roussos, Andrew McKenzie, Xianxiao Zhou, Yuji Kajiwara, Kristen J. Brennand, Gabriele C. De Luca, John F. Crary, Patrizia Casaccia, Joseph D. Buxbaum, Michelle Ehrlich, Sam Gandy, Alison Goate, Pavel Katsel, Eric Schadt1, Vahram Haroutunian and Bin Zhang. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer´s disease Genome Medicine 2016. p. 8:104

[27] Morrison, JH.; Hof, PR.; Rapp, PR. Neuropathology of normal aging in cerebral cortex.. In: Beal, MF.; Lang, AE.; Ludolph, A., editors. Neurodegenerative Diseases. Cambridge University Press; Cambridge: 2005. p. 396-406

[28] Ohnishi S, Takano K (marzo de 2004). «Amyloid fibrils from the viewpoint of protein folding». *Cell. Mol. Life Sci.* **61** (5): 511-24

[29] Ott A, Breteler MM, van Harskamp F, Stijnen T, Hofman A. Incidence and risk of dementia. The Rotterdam Study. Am J Epidemiol. 1998 Mar 15;147(6):574-80

[30] Perez N, Menendez S, Rodriguez J. (2002). «inas, Apo E y enfermedad de Alzheimer.». *Rev Cuba Inve Biom.* **21**: 262-269.

[31] R. A. Irizarry; B. Hobbs; F. Collin; Y. D. Beazer-Barclay; K. J. Antonellis; U. Scherf; T. P. Speed (2003). “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”. Biostatistics (núm. 4, págs. 249-264).

[32] Roberson ED, Mucke L. 100 years and counting: prospects for defeating Alzheimer's disease. Science 2006;314:781–784. [PubMed: 17082448]

[33] Ruiz de Villa M C, Sanchez-Pla, A, PID_00192743. UOC (2017)

[34] Sandoval-Salazar C., Ramírez Emiliano J., Solís Ortiz S. Inhibición GABAérgica en la ingesta alimentaria Rev Mex Neurici Sept-Oct 2013; 14(5):262-271

[35] Savva GM, Brayne C. Epidemiología y repercusión de la demencia. In: Weiner MF, Lipton AM, editors. Manual de Enfermedad de Alzheimer y otras demencias. Madrid: Editorial Panamericana; 2010. p. 17-21

[36] Small D, Klaver D, Foa L. (2010). «Presenilins and the γ-secretase: still a complex problem.». *Mol Brain.*: 1-6.

[37] Stephan BC, Brayne C, Savva GM, Matthews FE. Occurrence of medical co-morbidity in mild cognitive impairment: implications for generalisation of MCI research. Age Ageing. 2011 Jul;40(4):501-7

[38] Terry RD. Alzheimer's disease and the aging brain. J Geriatr Psychiatry Neurol 2006;19:125–128. [PubMed: 16880353]

[39] Thies W, Bleiler L. 2011 Alzheimer's disease facts and figures. Alzheimers Dement. 2011 Mar;7(2):208-44.

[40] Trujillo Tiebas M. J, Gómez Pérez J.L, Daschner A Medicina Evolucionista Vol III 2017. p:67-76

[41]Tschanz JT, Corcoran C, Skoog I, Khachaturian AS, Herrick J, Hayden KM, et al. Dementia: the leading predictor of death in a defined elderly population: the Cache County Study. Neurology. 2004 Apr 13;62(7):1156-62.

[42] Von Gunten A, Kovari E, Rivara CB, Bouras C, Hof PR, Giannakopoulos P. Stereologic analysis of hippocampal Alzheimer's disease pathology in the oldest-old: Evidence for sparing of the entorhinal cortex and CA1 field. Experimental Neurology 2005;193:198–206. [PubMed: 15817278

[43] Wenk GL (2003). «Neuropathologic changes in Alzheimer's disease». *J Clin Psychiatry*. 64 Suppl 9: 7-10

[44] Xu PT, Li YJ, Qin XJ, Scherzer CR, Xu H, Schmechel DE, Hulette CM, Ervin J, Gullans SR, Haines J, Pericak-Vance MA, Gilbert JR. Differences in apolipo-protein E3/3 and E4/4 allele-specific gene expression in hippocampus in Alzheimer disease. Neurobiol Dis 2006;21:256–275. [PubMed: 16198584]

[45] Yankner BA, Duffy LK, Kirschner DA (octubre de 1990). «Neurotrophic and neurotoxic effects of amyloid beta protein: reversal by tachykinin neuropeptides». *Science (journal).* **250** (4978): 279-82. PMID221853

(https://www.genecards.org/Search/)

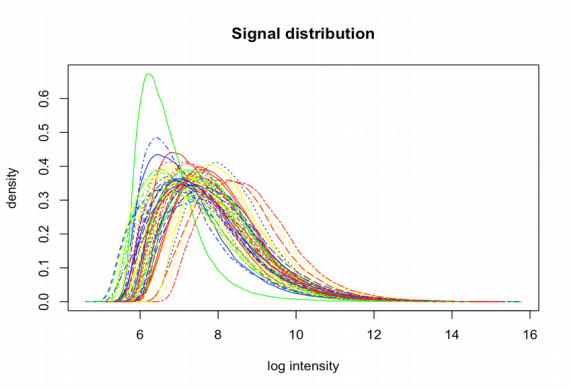
(http://saweb2.sabiosciences.com/pathway.php?sn=PAK_Pathway)

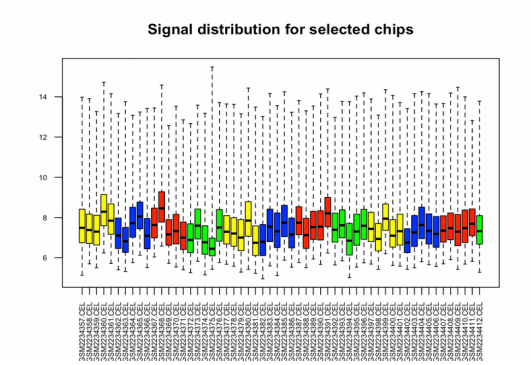
(https://www.genecards.org/cgi-bin/carddisp.pl?gene=VSNL1&keywords=VSNL1)

6. Anexos

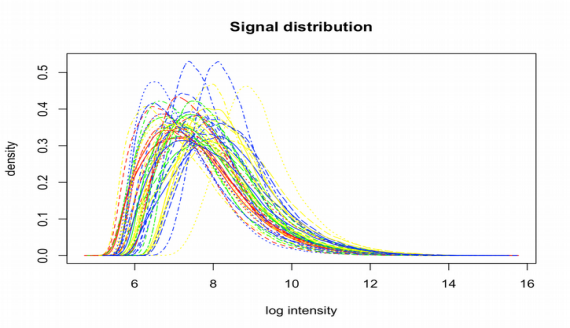
ANEXO 1

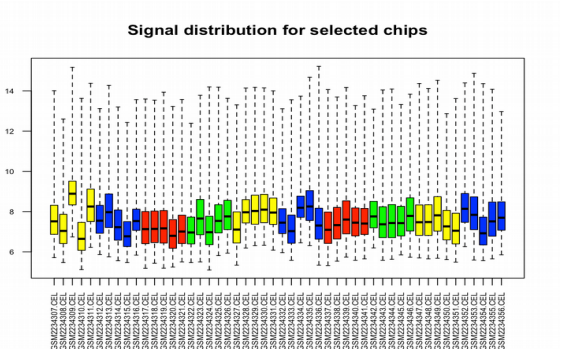
A la izquierda se observan los gráficos de densidad (distribución de los arrays) y a la derecha se observan los boxplot de los arrays de los distintos tejidos. No hay diferencias reseñables entre muestras en ningún tejido



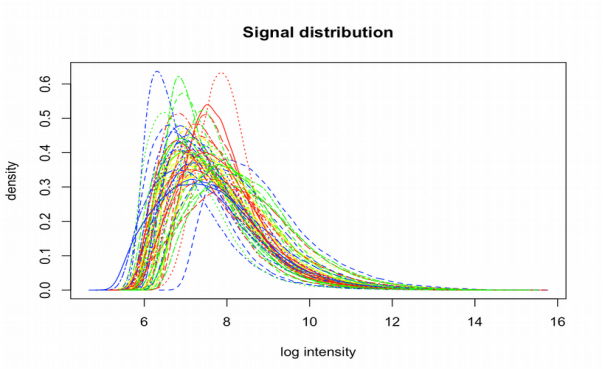


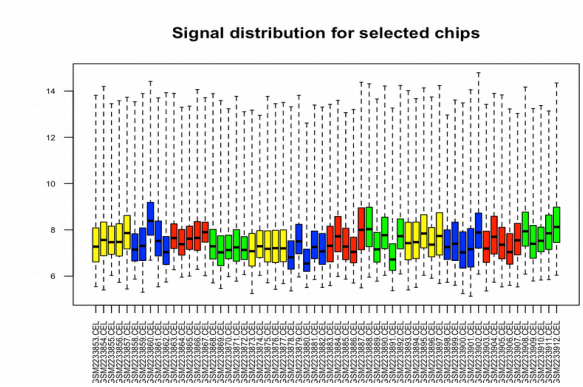
Tejido Prefrontal Cortex





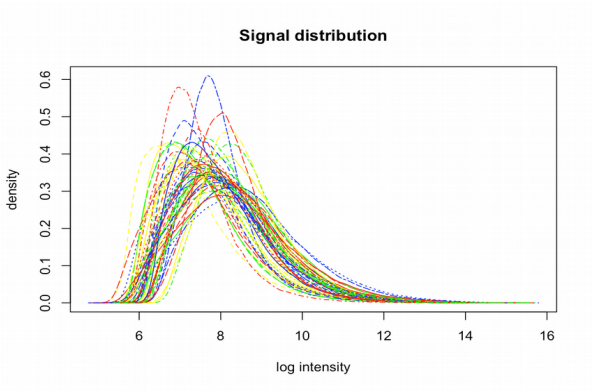
Tejido Superior Parietal Lobule

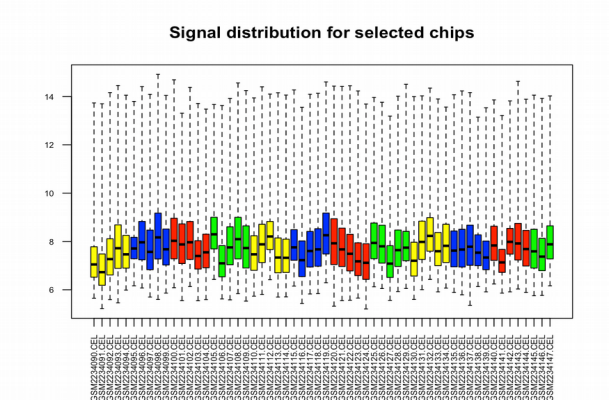




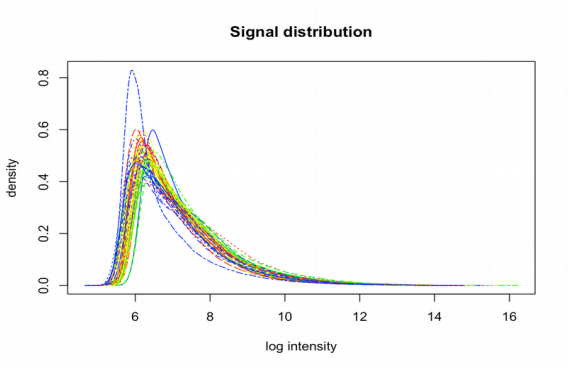
Tejido Superior Temporal Gyrus

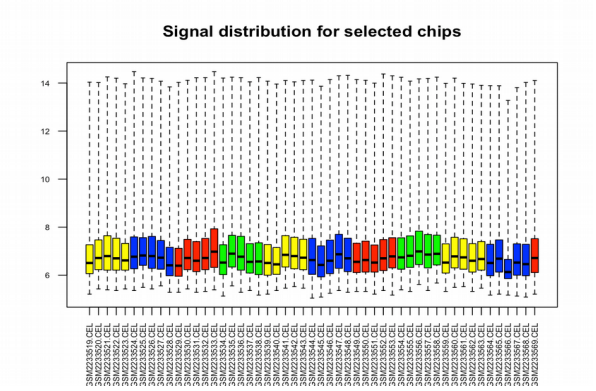
Figura1 Gráfico Densidad y Boxplot de los Tejidos Prefrontal Cortex, Superior Parietal Lobule y Superior Temporal Gyrus



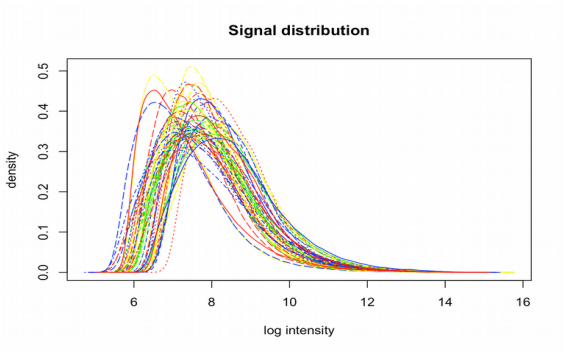


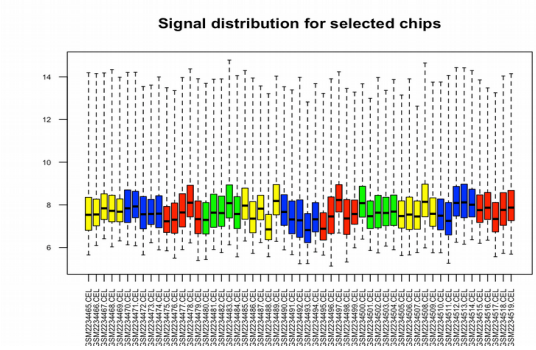
Temporal Pole





Tejido Amígdala



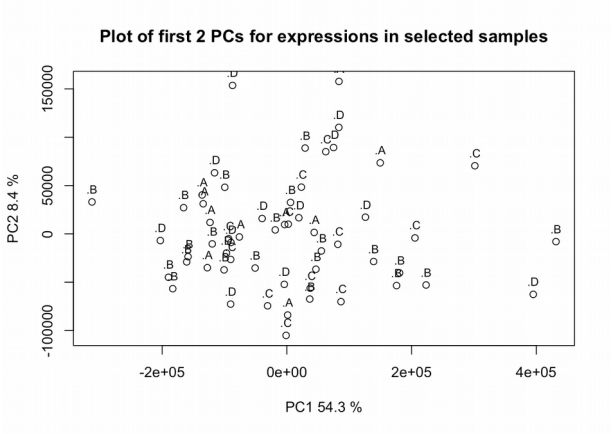


Tejido Región Hipocampo

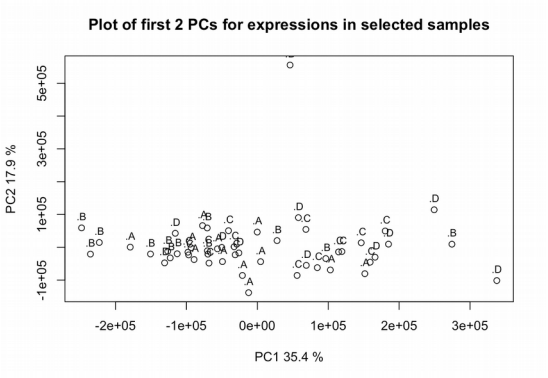
Figura 2 Gráfico Densidad y Boxplot de los Tejidos Temporal Pole, Amígdala y Región Hipocampo

ANEXO 2

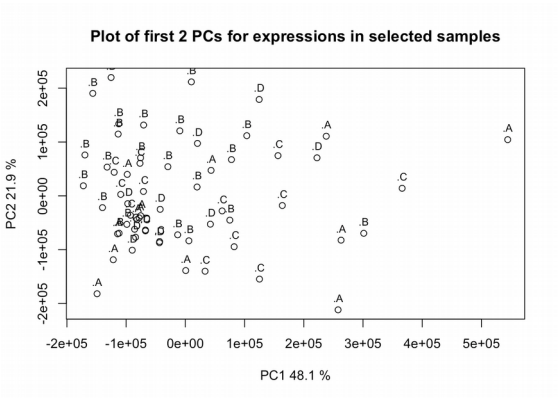
Gráfico de Componentes Principales para los distintos tejidos analizados. Excepto en el Tejido Superior Parietal Lobule donde las muestras presentan intensidad de señal baja, el resto presentan intensidad de señal baja, media y alta. No hay agrupación clara entre muestras en ningún tejido.



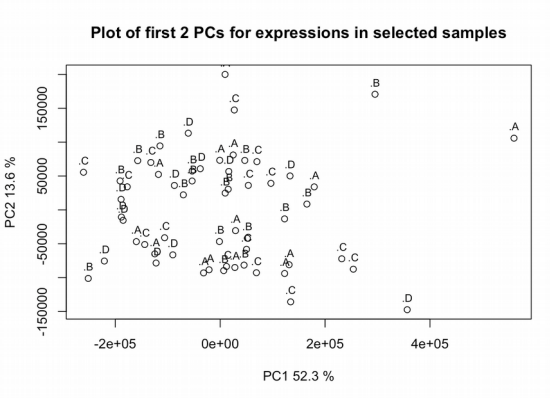
Tejido Superior Parietal Lobule



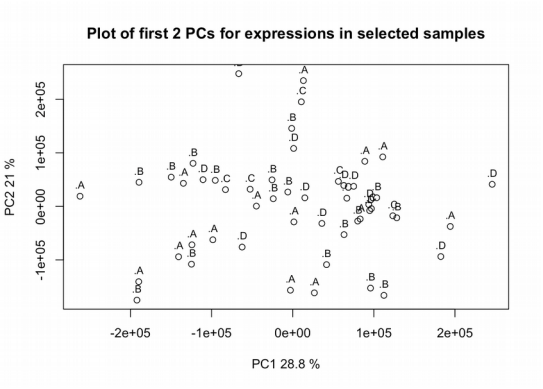
Tejido Prefrontal Cortex



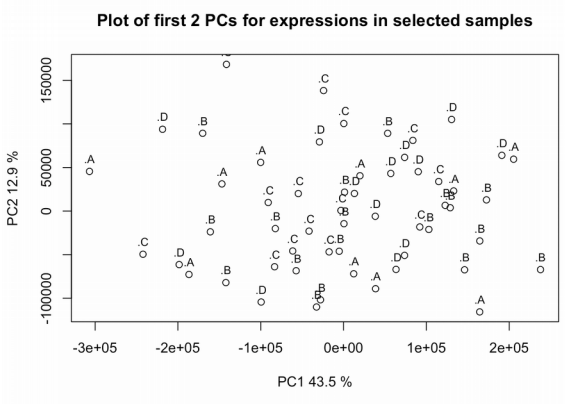
Tejido Superior Temporal Gyrus



Tejido Temporal Pole



Tejido Amígdala

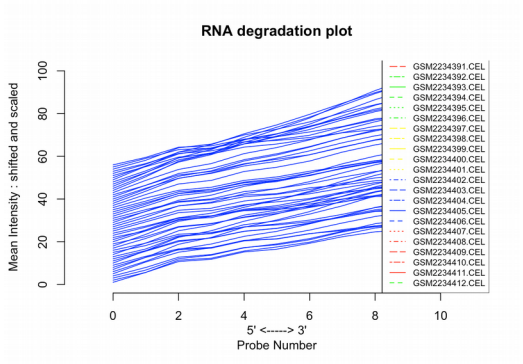


Tejido Región Hipocampo

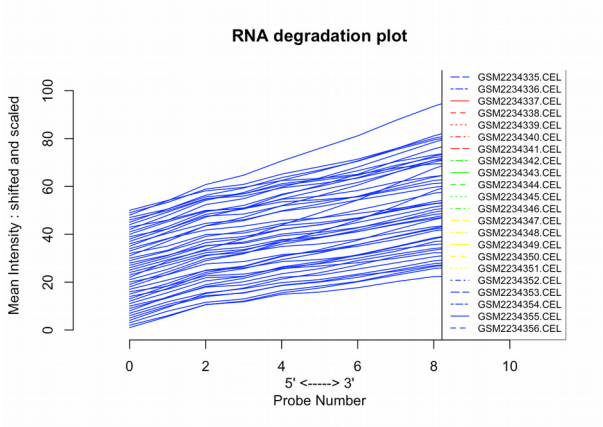
Figura 3 Gráfico Componentes Principales para todos los tejidos

ANEXO 3

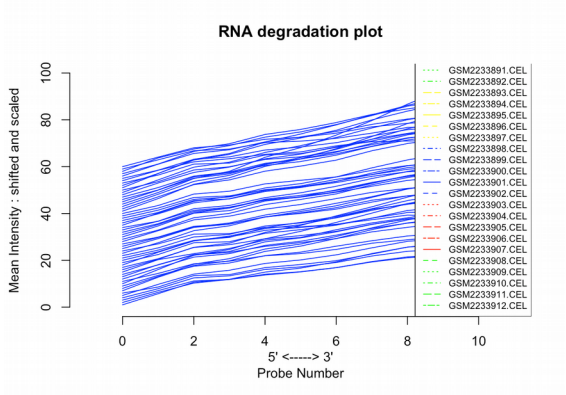
Gráfico de degradación del RNA. En todos los tejidos se observa que las líneas son casi paralelas. Con lo que el nivel de degradación del RNA es similar en todos los chips



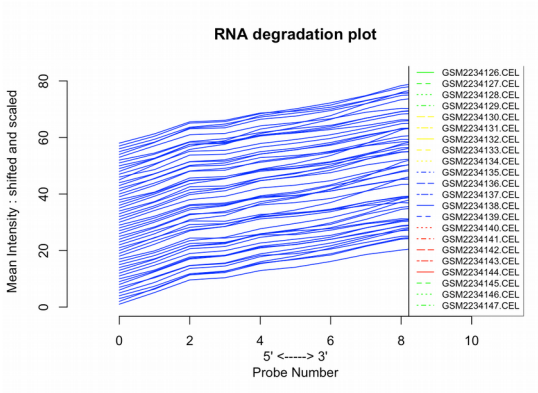
Tejido Prefrontal Cortex



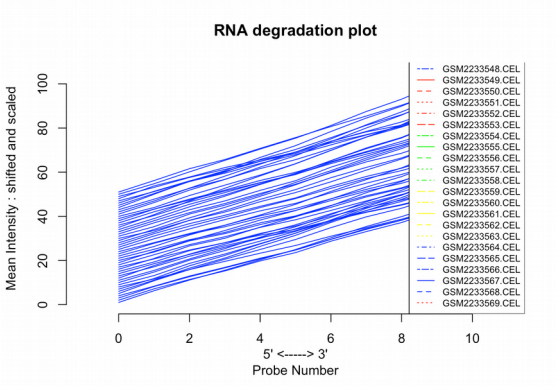
Tejido Superior Parietal Lobule



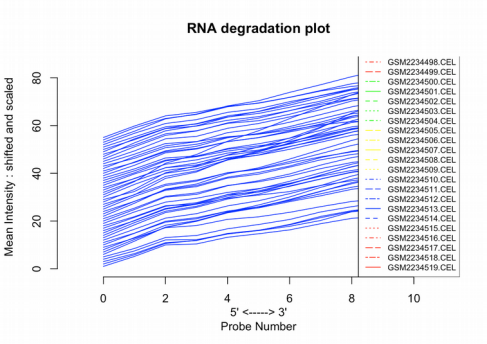
Tejido Superior Temporal Gyrus



Tejido Temporal Pole



Tejido Amígdala

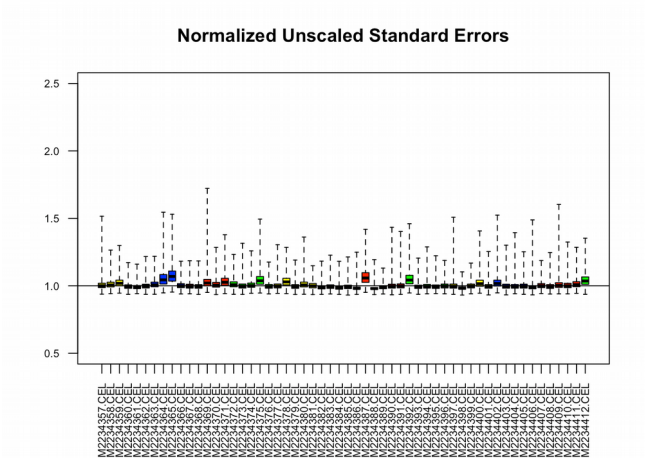


Tejido Región Hipocampo

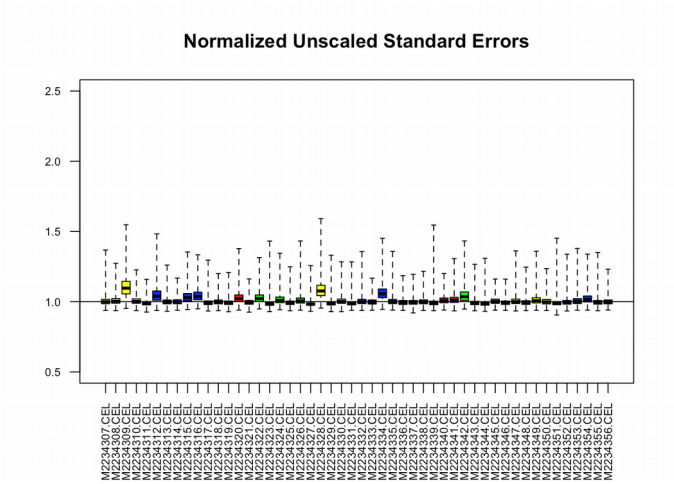
Figura 4 Gráfico Degradación del RNA en todos los Tejidos

ANEXO 4

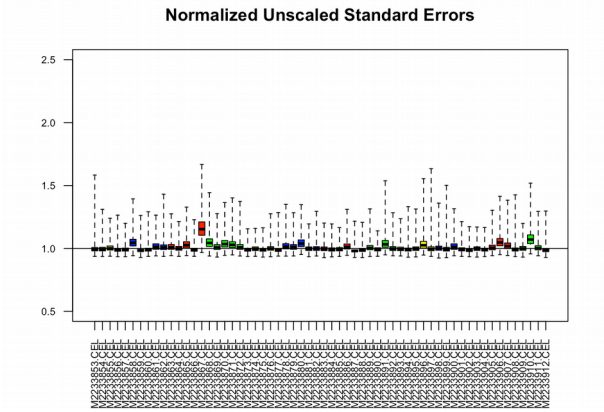
Gráficos Normalized Unscaled Standard Errors (NUSE) para todos los tejidos (que se corresponde con el error no estandarizado y normalizado). Se observa que los datos son de calidad, presentan una relativa simetría. Es el gráfico de errores o residuos del ajuste del modelo de análisis previo a la normalización de los datos



Tejido Prefrontal Cortex

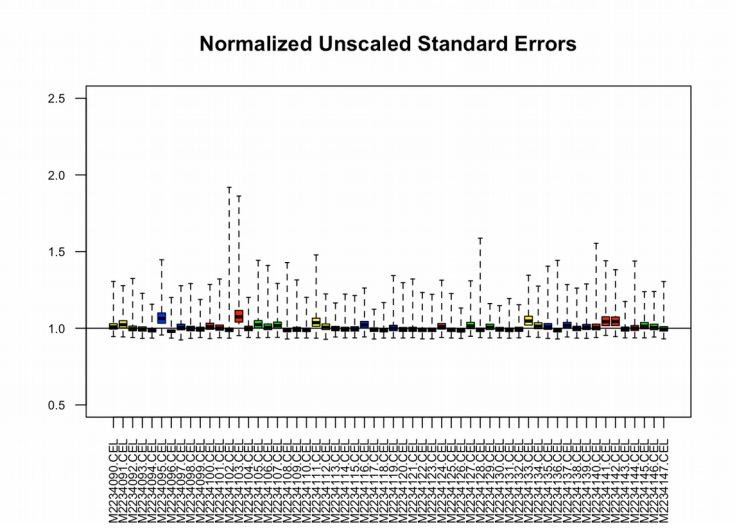


Tejido Superior Parietal Lobule

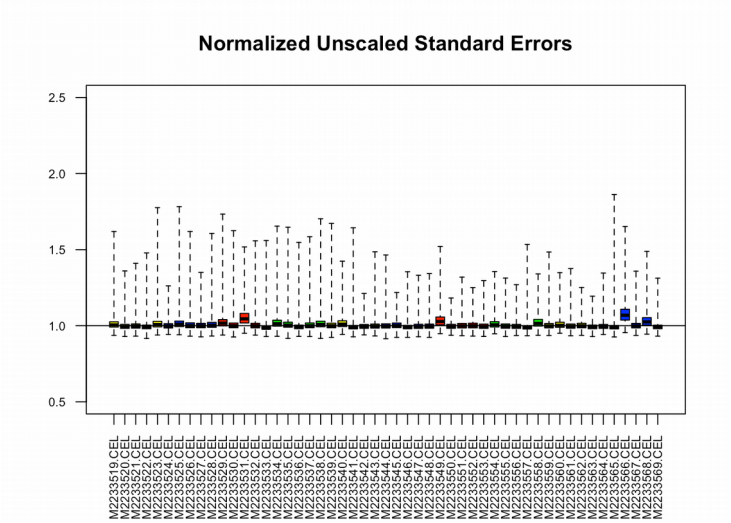


Tejido Superior Temporal Gyrus

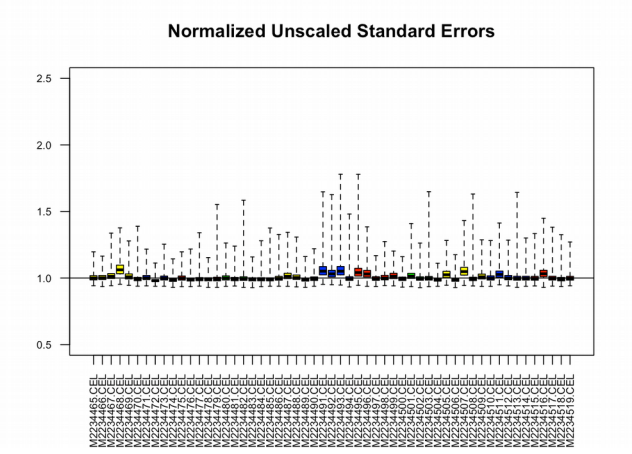
Figura 5 Gráfico de Normalized Unscaled Standard Errors (NUSE) para los Tejidos Prefrontal Cortex, Superior Parietal Lobule y Superior Temporal Gyrus.



Tejido Temporal Pole



Tejido Amígdala

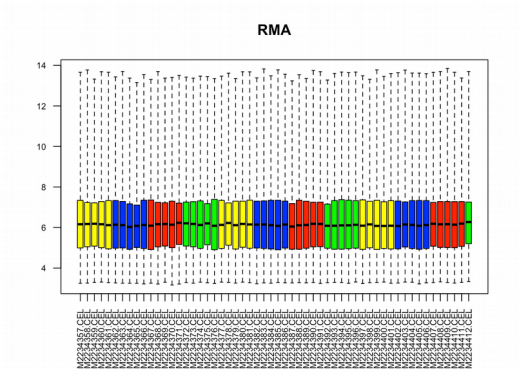


Tejido Región Hipocampo

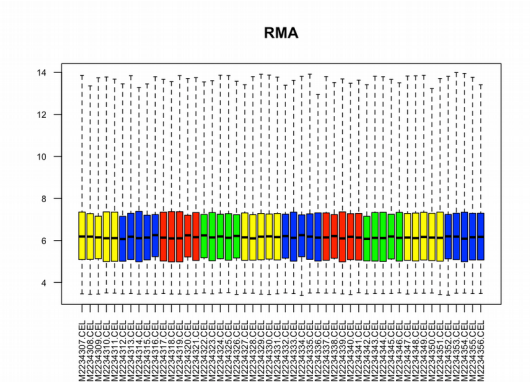
Figura 6 Gráfico de Normalized Unscaled Standard Errors (NUSE) para los Tejidos Temporal Pole, Amígdala y Región Hipocampo

ANEXO 5

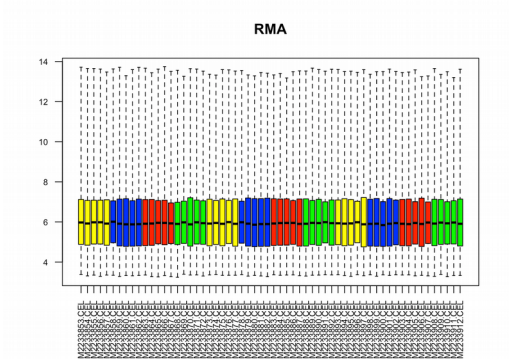
Boxplot de los datos Normalizados para todos los Tejidos.



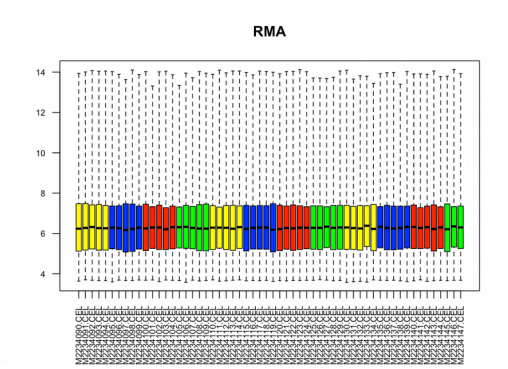
Tejido Prefrontal Cortex



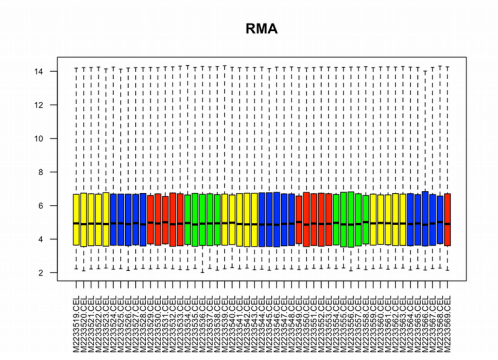
Tejido Superior Parietal Lobule



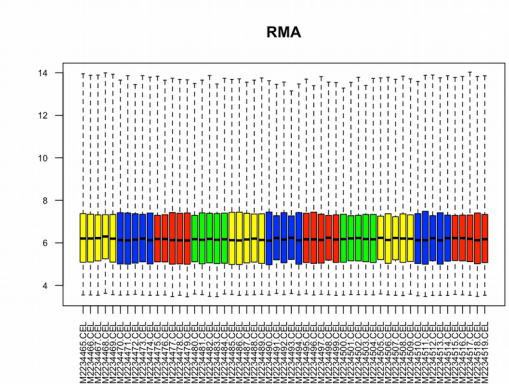
Tejido Superior Temporal Gyrus



Tejido Temporal Pole



Tejido Amígdala



Tejido Región Hipocampo

Figura 7 Gráficos de Boxplot de los datos normalizados para todos los tejidos