Universitat Oberta de Catalunya

uoc.edu

Estudio integrado de la metilación y expresión génica en sangre y búsqueda de marcas potencialmente asociadas a la enfermedad de Parkinson

Guillermo Paz López

Máster en Bioinformática y Bioestadística (UOC-UB)

Área 1 – Análisis de datos ómicos

Nombre Consultor/a: Helena Brunel Montaner

Nombre Profesor/a responsable de la asignatura: David Merino Arranz

Nombre Tutor del centro colaborador: Andrés González Jiménez

Junio de 2023

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio integrado de la metilación y expresión génica en sangre y búsqueda de marcas potencialmente asociadas a la enfermedad de Parkinson
Nombre del autor:	Guillermo Paz López
Nombre del consultor/a:	Helena Brunel Montaner
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	06/2023
Titulación:	Máster en Bioinformática y Bioestadística (UOC-UB)
Área del Trabajo Final:	Área 1 - Análisis de datos ómicos
Idioma del trabajo:	Español
Número de créditos:	15
Palabras clave	Parkinson, Metilación, RNA-Seq, Integración ómica

Resumen del Trabajo (máximo 250 palabras): Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.

La enfermedad de Parkinson es la segunda enfermedad neurodegenerativa más común, por detrás del Alzheimer, y está caracterizada por una degeneración de las neuronas dopaminérgicas. Para los individuos enfermos su padecimiento está asociado, principalmente, a problemas motrices que dificultan su día a día. Respecto al diagnóstico, la identificación temprana del Parkinson es algo esencial en la lucha contra la enfermedad. Para realizar el diagnóstico precoz se han utilizado multitud de biomarcadores, desde marcadores sintomáticos a neuroimágenes. Otras vías de estudio que también se han contemplado implican el análisis de la expresión génica y de la metilación del ADN. En este estudio se ha optado por realizar un análisis integrado de datos de expresión y metilación en sangre para identificar posibles marcas asociadas a la enfermedad. Para ello se realizaron análisis de expresión y metilación diferencial, además de la integración de éstos. Los diferentes análisis realizados permitieron identificar un

conjunto de genes diferencialmente expresados y/o metilados, procesos biológicos y componentes celulares que previamente ya habían sido vinculados a la enfermedad. Además, se comprobó como las variables de estudio tienen cierto poder de separación entre las muestras control y las muestras de Parkinson, sin llegar a una distinción nítida entre ambos grupos. Estos resultados arrojan luz sobre la capacidad del tejido sanguíneo como predictor del Parkinsonismo.

Abstract (in English, 250 words or less):

Parkinson's disease is the second most common neurodegenerative disease, after Alzheimer's, and it is characterized by a degeneration of dopaminergic neurons. For people with Parkinsonism, their suffering is mainly associated with motor problems that hinder their daily lives. In terms of diagnosis, early identification of Parkinson's is essential in the fight against the disease. To achieve early diagnosis, a multitude of biomarkers have been used, ranging from symptomatic markers to neuroimaging. Other paths of study that have also been considered involve analyzing gene expression and DNA methylation. In this study, an integrated analysis of expression and methylation data in blood was chosen to identify possible markers associated with the disease. For this purpose, expression and differential methylation analyses were performed, in addition to their integration. The different analyses allowed the identification of a set of differentially expressed and/or methylated genes, biological processes, and cellular components that had previously been linked to the pathology. Furthermore, it was observed that the variables studied have some ability to differentiate between control samples and Parkinson's samples, although a clear distinction between the two groups was not reached. These results shed light on the potential of blood tissue as a predictor of Parkinsonism.

Índice

Lista de figuras

Lista de tablas

1 Introducción

1.1 Contexto y justificación del trabajo

La enfermedad de Parkinson es la segunda enfermedad neurodegenerativa más común, tan solo superada por la enfermedad de Alzheimer. Aproximadamente diez millones de personas están afectadas por la enfermedad de Parkinson a nivel mundial, y únicamente en torno al cuatro por cierto es diagnosticada antes de los 50 años (Erkkinen *et* al., 2018; Parkinson's Foundation, s.f.). Esta afecta al sistema nervioso central y se caracteriza por la degeneración progresiva de las neuronas dopaminérgicas presentes en la región pars compacta de la sustancia negra (SNpc, del inglés, *Substantia Nigra pars compacta*) (Gibb & Lees, 1991; Damier *et al.*, 1999), lo que conlleva una reducción en la producción de dopamina (Ehringer & Hornykiewicz, 1960).

La dopamina es un neurotransmisor que ayuda a coordinar el movimiento en el cuerpo (Ungerstedt, 1971), aunque también parece estar involucrada en el control del estado de ánimo, la motivación y el estado de alerta (Marshall *et al*., 1976; Berridge & Robinson, 1989). Como resultado de la degeneración de las células dopaminérgicas, las personas con la enfermedad de Párkinson pueden experimentar temblores, también llamados movimientos parkinsonianos, rigidez muscular, lentitud en los movimientos y dificultad para caminar y mantener el equilibrio. A medida que la enfermedad progresa, también puede haber otros síntomas como depresión, ansiedad, problemas de sueño y problemas cognitivos. Aunque actualmente no hay cura para la enfermedad de Párkinson, existen tratamientos y terapias disponibles que pueden ayudar a controlar los síntomas y mejorar la calidad de vida de las personas que la padecen. Uno de los tratamientos más comunes consiste en la administración de L-DOPA, el precursor metabólico de la dopamina; este es capaz de atravesar la barrera hematoencefálica, a diferencia del neurotransmisor. Aunque su uso no permite recuperar las células perdidas en el proceso neurodegenerativo, sí que fomenta que el resto de las células dopaminérgicas funcionales produzcan una cantidad mayor de dopamina, tratando de mitigar el efecto de la pérdida celular (*The National Collaborating Centre for Chronic Conditions*, 2006).

Frente a la falta de tratamientos efectivos contra la enfermedad surge una demanda imperante por conocer las bases moleculares y bioquímicas de la enfermedad en pos de facilitar su identificación en etapas tempranas previas a su desarrollo mediante la identificación de marcadores moleculares de riesgo.

El estudio de los mecanismos moleculares asociados a una enfermedad puede llevarse a cabo desde múltiples enfoques mediante el uso de las nuevas tecnologías de secuenciación (NGS, del inglés, *Next Generation Sequencing*), por ejemplo a través del estudio de la regulación de la expresión génica. El análisis del transcriptoma, es decir, del ARN derivado de la expresión de los genes, y en concreto del ARNm, puede llevarse a cabo mediante el análisis de datos derivados de la técnica de RNA-Seq. Por otro lado, el análisis del epigenoma puede llevarse a cabo mediante el análisis de la metilación de las posiciones CpG del ADN.

El análisis de expresión diferencial de datos de RNA-Seq es una técnica bioinformática utilizada para comparar los niveles de expresión génica entre dos o más muestras (Wang *et al.,* 2009). Esta técnica se ha convertido en una herramienta esencial para estudiar la regulación génica y ha sido utilizada en una amplia gama de aplicaciones, desde la investigación básica hasta la medicina personalizada (Thin *et al.,* 2021). En esencia, el análisis de expresión diferencial de datos de RNA-Seq implica la comparación del número de lecturas de secuencias de ARN de diferentes muestras y la identificación de diferencias estadísticamente significativas en estos valores. Existen varios paquetes de R utilizados comúnmente para realizar este análisis, entre ellos *DESeq2*, *edgeR* y *limma* (Love *et al.*, 2014; Robinson *et al.*, 2010; Ritchie *et al.,* 2015).

Para llevar a cabo el análisis de expresión diferencial de datos de RNA-Seq, se requiere una serie de pasos. Estos incluyen la alineación de las lecturas de secuencia con un genoma de referencia, la cuantificación de los niveles de expresión génica y la identificación de los genes diferencialmente expresados entre las muestras o DEG (del inglés, *Differentially Expressed Genes*) (Love *et al.*, 2014).

Los análisis de expresión diferencial se han utilizado con éxito en numerosos estudios. En el contexto de la enfermedad de Parkinson, este tipo de análisis ha

permitido identificar multitud de genes diferencialmente expresados entre casos enfermos y casos sanos, como por ejemplo el gen SKP1A (*S-Phase Kinase Associated Protein* 1) (Grünblatt *et al*., 2004), el gen MRPS6 (*Mitochondrial Ribosomal Protein* S6) (Papapetropoulos *et al*., 2006), el gen PDXK (*Pyridoxal Kinase*) (Elstner *et al*., 2009), o el PGC-1α (*Peroxisome proliferatoractivated receptor Gamma Coactivator*-1α) (Zheng *et al*., 2010). Sin embargo, es poco común que se identifiquen DEG comunes entre un estudio y otro, como se comenta en Borrageiro *et al*. (2018). En su lugar, puede ser más interesante para el investigador estudiar las vías y los procesos afectados, entre los que hay más consenso entre estudios.

Otras vías se centran en el estudio de la regulación de la expresión génica y el epigenoma. Entendemos por epigenoma al conjunto de modificaciones que se realizan sobre la molécula de ADN, y que en muchos casos conllevan una regulación de la expresión génica. Una de las modificaciones que se puede realizar sobre el ADN es la metilación de los nucleótidos citosina en posiciones CpG (nucleótidos citosina seguidos de nucleótidos guanina). El conjunto de la información sobre el estado de metilación de todas las citosinas de un genoma constituye el metiloma (Hsu *et al.,* 2018).

Además de los sitios CpG metilados, la metilación de un genoma se estudia a nivel de islas CpG (regiones con un alto número de sitios CpG), CpG *shores* (regiones 2 Kpb corriente arriba y corriente debajo de islas CpG) y CpG *shelves* (regiones 2 Kpb corriente arriba o corriente debajo de las regiones CpG *shores*). En su conjunto, las islas CpG, así como sus áreas colindantes, pueden encontrarse en regiones intragénicas o intergénicas, exónicas o intrónicas, así como regiones no traducibles (UTR, del inglés, *Unstranslated Regions*) (Irizarry *et al.,* 2009).

El estudio de la metilación de sitios y regiones ricas en CpG es de un especial interés, pues la metilación del ADN constituye un proceso natural de la regulación de la expresión génica. La metilación del ADN implica un cambio en su reorganización, provocando que la cromatina sea accesible para las diferentes proteínas (Moylan & Murphy, 2016).

El análisis de metilación diferencial de sitios CpG es una técnica de análisis bioinformático utilizada para identificar cambios en la metilación del ADN en diferentes grupos de muestras (Mikeska *et al.*, 2014). Esto permite comparar la metilación del ADN entre diferentes grupos de muestras y detectar cambios significativos en la metilación de sitios y regiones CpG específicas que pueden estar asociados con diferencias fenotípicas o patológicas (Bock *et al.,* 2010; Laird, 2010). Esta técnica se utiliza en una amplia gama de estudios de investigación, incluyendo estudios de cáncer (Baylin & Ohm, 2006), enfermedades metabólicas (Nitert *et al.,* 2012) y trastornos neurológicos (Irwin *et al.*, 2015). Para llevar a cabo un análisis diferencial de metilación existen una variedad de métodos computacionales, que incluyen desde análisis estadísticos básicos hasta algoritmos más complejos con diferentes tipos de correcciones. Tras el análisis se pueden obtener listas de sitios o posiciones CpG diferencialmente metiladas, o DMP (del inglés, *Differentially Methylated Positions,* posiciones diferencialmente metiladas), regiones diferencialmente metiladas o DMRs (del inglés, *Differentially Methylated Regions,* regiones diferencialmente metiladas) (Neidhart, 2016), y listas de genes diferencialmente metilados.

La identificación de DEG y de DMP y DMRs es crucial para conocer qué diferencias existen entre las poblaciones de estudio a nivel del transcriptoma y del metiloma. De este modo es posible conocer la regulación de qué genes puede tener efecto en la aparición de síntomas de enfermedades, como es la enfermedad de Parkinson. Esta identificación puede hacerse tanto a nivel del tejido de la sustancia negra como de tejido sanguíneo, con un objetivo terapéutico y/o de diagnóstico. El estudio de marcas en el tejido sanguíneo tiene un interés especial sobre todo de cara al diagnóstico temprano de la enfermedad, pues el acceso al tejido sanguíneo es siempre más cómodo respecto al tejido neuronal de las regiones encefálicas.

1.2 Descripción general del proyecto

A lo largo de este proyecto de Trabajo de Fin de Máster se realizará el análisis de datos de RNA-Seq de ARN mensajero y de datos de metilación ADN, tanto de muestras de sangre de pacientes controles sanos como de pacientes con la

enfermedad de Parkinson. El estudio de estos datos se llevará a cabo en dos fases.

Por un lado se aplicará un enfoque aislado, analizando por separado cada ómica. Para ello se utilizarán los protocolos apropiados para manipular y realizar inferencia en cada tipo de dato.

Por otro lado, se llevará a cabo una segunda fase que consistirá en el análisis integrado de los resultados obtenidos a partir de los estudios aislados para obtener una visión más completa sobre las vías afectadas e identificar potenciales biomarcadores asociados a la enfermedad a nivel sanguíneo, algo gran interés para el diagnóstico temprano del Parkinsonismo, como hemos comentado anteriormente. Para esta segunda fase se utilizarán herramientas especializadas en la integración de resultados obtenidos a partir de datos de diferente naturaleza.

1.3 Impacto en sostenibilidad, ético-social y de diversidad

El diagnóstico precoz de la enfermedad de Parkinson es esencial para mitigar los efectos de esta en el paciente, facilitando el acceso temprano a tratamientos y mejorando su calidad de vida. Además, la enfermedad de Parkinson tiene una gran implicación económica.

Según las últimas estimaciones realizadas por el grupo Lewin (*Lewin Group Inc*.), la Fundación Parkinson y la Fundación Michael J. Fox (The Michael J. Fox Foundation for Parkinson’s Research | Parkinson’s Disease , 2019), en EE.UU la combinación directa e indirecta de los costos de la enfermedad de Parkinson a fecha de 2017 asciende a aproximadamente 52 mil millones de dólares anuales. En torno a 25 mil millones de dólares serían de gasto directo, lo que incluye tratamientos, hospitalización, equipamiento médico, y cerca de 27 mil millones de gastos serían indirectos y gastos no médicos, como la contratación de profesionales de asistencia a los enfermos y las adaptaciones que se deben hacer en hogares y medios de transporte. Estos resultados superan a las previsiones realizadas por Huse *et al*. (2005), en las que predijeron que el costo total de la enfermedad de Parkinson en EE.UU sería de aproximadamente 50 mil millones de dólares para el año 2040. Según estudios recientes, la previsión es que para el año 2037 en EE.UU haya más de 1.6 millones de enfermos de

Parkinson, implicando una carga económica que superaría los 79 mil millones de dólares (Yang *et al.*, 2020). Estos datos son indicativos de cómo el aumento en la incidencia de Parkinson, y su impacto asociado, es algo que se ha subestimado. Por tanto, es imperante tomar medidas que permitan reducir esta incidencia, reducir el progreso de la enfermedad y alivianen los síntomas, tanto para mejorar la vida de los pacientes como para reducir el peso económico-social de la enfermedad.

1.4 Objetivos del Trabajo

A continuación, se detallan los objetivos generales y los objetivos específicos. Mientras que los primeros buscan cubrir, *grosso modo*, las aspiraciones del proyecto y el alcance de este, los segundos buscan concretar cada uno de los objetivos generales.

1.4.1 Objetivos generales

Los objetivos generales del proyecto son los siguientes:

- Estudiar las diferencias a nivel de metiloma y transcriptoma de ARNm entre la población sana y la población afectada por la enfermedad de Parkinson.

1.4.2 Objetivos específicos

A continuación, se indican los objetivos específicos, referidos a los objetivos generales indicados anteriormente.

- Estudiar y comparar el nivel de metilación de diferentes posiciones del ADN entre las muestras de población sana y las muestras de población enferma.
- Estudiar y comparar el grado de expresión génica entre las muestras de población sana y las muestras de población enferma.
- Obtener deducciones integradas a partir de ambos tipos de datos e identificar potenciales biomarcadores de la enfermedad de Parkinson.

1.5 Enfoque y método seguido

Para abordar y completar los objetivos planteados en el apartado anterior se procederá a realizar una serie de análisis estadísticos que permitan hacer inferencia sobre datos del metiloma y del transcriptoma obtenidos a partir de diferentes muestras de pacientes sanos y enfermos con Parkinson. Las muestras estudiadas provienen del estudio de Henderson *et al*. (2021).

Por un lado, para el análisis de los datos de metilación se partirá de los archivos IDAT (del inglés, *Intensity Data,* datos de intensidad). Estos contienen información sobre la intensidad de metilación de los diferentes sitios CpG de estudios en cada una de las muestras. El tratamiento de estos datos brutos hasta la obtención de valores-β (valores beta) o valores-M puede hacerse mediante el uso de diversas herramientas, flujos de trabajo y paquetes, como son el paquete *minfi* (Aryee *et al.,* 2014) o el paquete *ChAMP* (Tian *et al*., 2017). También pueden ser manipulados manualmente aplicando las funciones necesarias para procesar y transformar los datos. En el caso de este proyecto, para el análisis de los datos de metilación se utilizará el paquete *ChAMP*. Este permite realizar un tratamiento de los datos de forma integrada, englobando su transformación, filtrado, normalización y análisis diferencial, en muchos casos haciendo uso de otros paquetes de R ampliamente utilizados por la comunidad científica, como es el paquete *limma (*Ritchie *et al.*, 2015). De esta forma se identificarán DMP y DMRs.

Por otro lado, para el análisis de los datos de RNA-Seq de ARNm partiremos de los archivos FASTQ, con información sobre las secuencias y su calidad de secuenciación asociada. Estos datos serán alineados frente al genoma humano de referencia GRChg38.p13. Existen múltiples alineadores para lograr esto: *Bowtie2, BBMAp, HISAT2, Salmon, Kallisto, STAR, TopHat2, Burrows Wheeler Aligner* (*BWA*)*,* etc. Será este último, *STAR* (Dobin *et al.*, 2013), el que utilicemos para realizar el alineamiento de las secuencias frente al genoma de referencia. Tras el alineamiento y el contaje de lecturas por gen y muestra se realizará análisis diferencial para identificar DEG (del inglés, *Differentially Expressed Genes*, genes diferencialmente expresados) entre los dos grupos de estudio. Para este tipo de análisis existen también diversas herramientas y paquetes

estadísticos ampliamente usados, como son *DESeq2* y *EdgeR*. Sin embargo, recientemente ha existido una controversia respecto a la tendencia de estos paquetes a encontrar “falsos positivos”, en este caso, “falsos DEG”, a la hora de realizar análisis diferenciales (Li *et al.,* 2022). Frente a esta situación se optará por utilizar pruebas estadísticas clásicas no paramétricas y ajustes clásicos de los p-valores, como se recomienda en Li *et al.* (2012).

Finalmente se realizará una integración entre los resultados obtenidos en ambos estudios. Existen también múltiples formas de lograr esto. En este caso se realizará, por un lado, un análisis de metilación de rasgos cuantitativos de expresión o eQTM (del inglés, *Expression Quantitative Trait Methylation analysis*). Para ello se utilizarán los datos obtenidos en los pasos anteriores junto al paquete *MatrixEQTL* (Shabalin, 2012). Este paquete cuenta con un enfoque centrado en el uso de datos de metilación y RNA-Seq, idóneo para este proyecto. Por otro lado, se realizará una integración vertical clásica a partir de las variables de ambas ómicas usando el paquete *mixOmics* (Rohart *et al*., 2017) y las funciones que ofrece.

El enriquecimiento funcional se hará en base a la terminología de ontología génica (*Gene Ontology*) y a la base de datos KEGG (*Kyoto Encyclopedia of Genes and Genomes*) usando los paquetes *clusterProfiler* (Wu *et.,* 2021) y *enrichplot* (Yu, 2023).

1.6 Planificación del Trabajo

A continuación, se indican cuáles serán las diferentes tareas que se llevarán a cabo para lograr los objetivos planteados. También se indican las fechas en las que se contempla su realización y una breve mención de las subtareas en las que se dividen.

1.6.1 Tareas

- Fase de control:
    - TAREA 1.1 (22/03 – 02/04): Análisis preliminar, control de calidad y filtrado de los datos de metilación y de RNA-Seq sin procesar.
- 2. Fase de análisis de metilación:

- TAREA 2.1 (03/04 – 16/04): Análisis de los datos (distribución, varianza). Estandarización de los datos e identificación de sitios CpG diferencialmente metilados entre los grupos de estudio.
- TAREA 2.2 (17/04 – 30/04): Estudio de la localización de los sitios CpG e identificación de regiones y genes diferencialmente metilados entre los grupos de estudio.
- TAREA 2.3 (22/05 – 30/05): Preparación de gráficos descriptivos de los resultados obtenidos.

- 3. Fase de análisis de RNA-Seq:
  - TAREA 3.1 (03/04 – 16/04): Alineamiento y obtención de tablas de contaje.
  - TAREA 3.2 (17/04 – 30/04): Cálculo de los niveles de expresión e identificación de genes diferencialmente expresados entre los grupos de estudio.
  - TAREA 3.3 (22/05 – 30/05): Preparación de gráficos descriptivos de los resultados obtenidos.
- 4. Fase de integración:
  - TAREA 4.1 (01/05 – 21/05): Integración de los resultados obtenidos en los estudios anteriores. Pruebas de agrupación no supervisada. Ensayo funcional a partir de las listas de genes seleccionados en base a criterios de significación. Revisión y corrección de posibles errores.
  - TAREA 4.2 (22/05 – 30/05): Preparación de gráficos descriptivos de los resultados obtenidos.
- 5. Fase de redacción de la memoria y preparación de la presentación:
  - TAREA 5.1 (17/04 – 21/05): Redacción de la metodología.
  - TAREA 5.2 (31/05 – 16/06): Repaso y finalización de la redacción de la metodología. Redacción de la discusión y las conclusiones. Redacción de otros apartados (bibliografía, posibles, anexos, etc.).

– TAREA 5.3 (12/’6 – 20/06): Preparación de la presentación.

1.6.2 Calendario

Se muestra a continuación la distribución temporal de las tareas mediante un diagrama de Gantt (Figura 1).

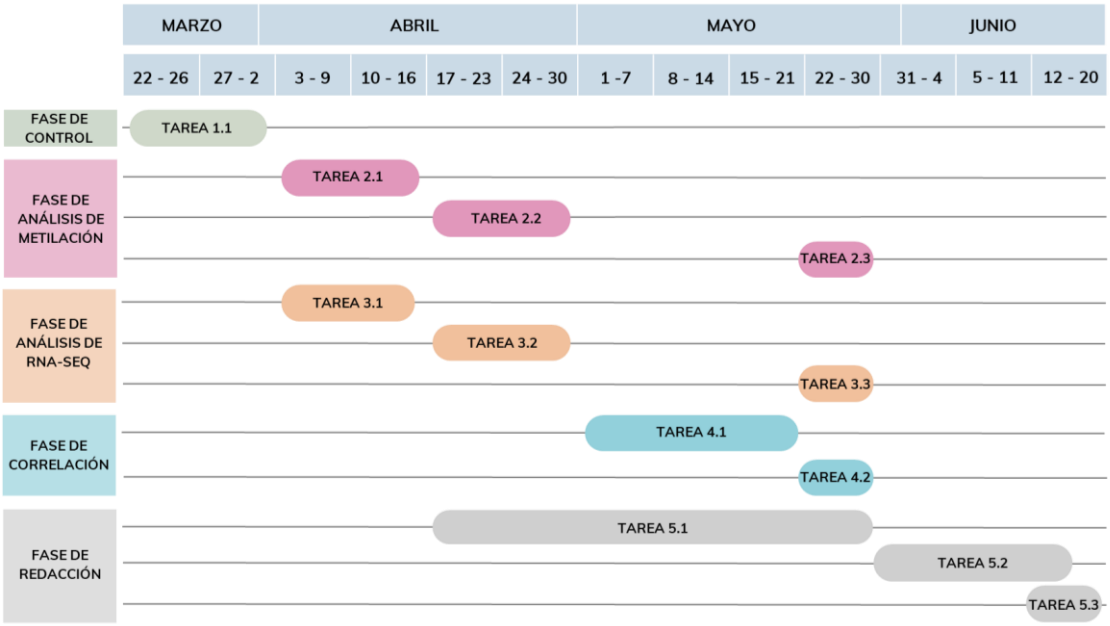

Figura 1. Distribución temporal de las tareas en las que está divida el proyecto.

1.6.3 Hitos

Los hitos planteados en este proyecto servirán como control de cada una de las tareas y de los resultados obtenidos.

- Control sobre la calidad de las secuencias. Se trata de un hito de la etapa inicial. Una vez realizado el control de calidad de las secuencias del análisis de datos RNA-Seq será necesario determinar si es necesario eliminar adaptadores y si es necesario cortar parte del extremo 3’ dado un descenso en la calidad de la secuenciación. También será necesario comprobar la ausencia de burbujas de calidad y, en caso de haberlas, será necesario determinar qué región de la secuencia se utilizará para las fases posteriores del análisis.

- Resultados derivados del análisis de metilación y de RNA-Seq: lista de DMP y DMRs, lista de genes vinculados a las DMP y DMRs, tablas de contaje de expresión y lista de genes diferencialmente expresados Se trata

de hitos de la etapa intermedia. La obtención de estos datos constituye un punto clave en el proyecto, pues son en sí necesarios para la discusión y forman parte de los objetivos. Será necesario comprobar que los resultados obtenidos en cada una de las etapas son lógicos y no se trata de “falsos positivos”. En función de los resultados obtenidos habrá que evaluar la forma de proceder, por ejemplo, a la hora de elegir puntos de corte para determinar significancia (p-valores, *Fold-Change,* etc.).

- Resultados derivados del análisis de integración: listas de genes seleccionados en base a criterios de significación y procesos y rutas funcionales. Se trata de hitos de la etapa final. Estos datos constituirán el punto final del proyecto y son clave para el apartado de discusión. Al igual que en el caso anterior es necesario corroborar que los resultados son lógicos y la ausencia de “falsos positivos”.
- Corrección de errores. Se trata de un hito de la etapa intermedia-final. Aunque no necesariamente tiene por qué ocurrir, en realidad es un evento importante. La detección y corrección de errores de forma eficiente a lo largo y al final de proyecto es algo determinante para facilitar la fluidez de este, así como su transcripción a la memoria.

1.6.4 Análisis de riesgos

Los riesgos generales que se deberán afrontar a lo largo del proyecto son los siguientes.

- Limitación de tiempo. Esta es una de las principales limitaciones. Para poder gestionar el tiempo disponible de la mejor forma posible y mitigar su efecto negativo en el proyecto se han diseñado y distribuido tareas a lo largo de un eje temporal, como se indicó anteriormente. En caso de que, pese a ello, sea notoria la falta de tiempo para poder finalizar todas las tareas planteadas, se optará por eliminar la TAREA 4.1., o subtareas de esta misma.
- Alcance del proyecto. Muy relacionado con la limitación de tiempo, el alcance del proyecto es algo que se ha tenido en cuenta a la hora de seleccionar los objetivos del proyecto y de diseñar las distintas fases en

las que estará dividido. Pese a eso, si el alcance llegase a ser limitante, al igual que en el caso de la limitación de tiempo, se optará por reducir el número de tareas a completar.

- Mala calidad de los datos de partida. Es un problema común en muchos análisis de datos biosanitarios. Para este proyecto se parte de datos preprocesados públicos, por los que se espera que la calidad sea considerablemente alta. Aun así, puede darse el caso de contar con muestras individuales con mala calidad, o con grupos de datos con baja calidad. En el primer caso, y según las circunstancias, se optará por descartar la muestra del análisis. En el segundo caso, y en función del tipo de dato, se optará por intentar tratar de forma que se intente corregir o mitigar el efecto de la baja calidad, o se seleccionarán regiones con una calidad mínimamente aceptable (en el caso de las secuencias de RNA-Seq). En caso de trabajar con datos con una calidad pobre, aunque mínimamente aceptable, esto se tendrá en cuenta en el apartado de la discusión y conclusiones.

Resultados pobres o de bajo interés. A la hora ir completando las diferentes tareas e ir alcanzando los diferentes hitos podrá suceder que los resultados no cumplan las expectativas esperadas. En cada caso, y según las circunstancias propias del análisis realizado, deberán estudiarse los pasos a seguir. En caso de ser estrictamente necesario, por ejemplo, al obtener resultados no concluyentes tras usar datos de baja calidad o fiabilidad para un análisis, se podrán descartar algunos resultados. Estos no se tendrán en cuenta en fases posteriores del análisis, ni se incluirán en los resultados; sin embargo, sí podrán ser mencionados y discutidos, brevemente, en los apartados finales del proyecto.

1.7 Breve resumen de contribuciones y productos obtenidos

A continuación, se indican cuáles son los resultados principales que se obtendrán a lo largo o al finalizar el proyecto.

- Plan de trabajo.
- Memoria.
- Presentación virtual.

1.8 Breve descripción de los otros capítulos de la memoria

Los siguientes apartados de la memoria contienen información sobre el contexto del estudio de la enfermedad de Parkinson y los procesos y herramientas bioinformáticas existentes, la metodología utilizada, los resultados obtenidos y cómo estos se relacionan con los conocimientos ya existentes de la enfermedad.

En el apartado 2. Estado del arte se describe cuál es la situación actual del estudio del Parkinson, y qué vías de estudio son las más populares. También se indaga en la importancia de desarrollar vías de estudio centrada sen la búsqueda de biomarcadores moleculares, para lo que se pueden utilizar multitud de procesos y herramientas bioinformáticas. En este apartado también se describe en qué punto se encuentra la bioinformática y, en concreto, cuáles son las fases que siguen los análisis de expresión y metilación diferencial, además de las herramientas existentes. Por último, se explica qué es la integración ómica, cuál es su importancia y por qué es interesante plantear un estudio de integración ómica asociado a una patología.

En el apartado 3. Metodología se describen los flujos de trabajo seguidos y las herramientas utilizadas. Se describen las diferentes fases que se han seguido en el proyecto para realizar el análisis de expresión diferencial, el análisis de metilación diferencial, y la integración de los resultados obtenidos.

En el apartado 4. Resultados se describen cuáles han sido los resultados obtenidos tras los diferentes análisis realizados, y en el apartado 5. Discusión estos son comentados en el contexto específico de la enfermedad, buscando los motivos que puede haber detrás de los resultados observados.

Finalmente, en el apartado 6. Conclusiones se establecen cuáles son las conclusiones a las que se ha podido llegar con la realización de este trabajo.

El apartado 7. Glosario contiene información sobre las siglas más utilizados en el trabajo, el 8. Bibliografía contiene las referencias, y el 9. Anexo contiene resultados extra autocontenidos.

2 Estado del arte

Como se ha comentado en apartados anteriores, la enfermedad de Parkinson es una de las enfermedades neurodegenerativas más comunes actualmente, afectando principalmente a la población mayor de 60 años (Erkkinen *et al*., 2018). Múltiples avances respecto a su identificación y caracterización se han hecho a lo largo de la historia. Entre los hitos más importantes destaca la detección de bajos niveles de liberación de dopamina asociada a la enfermedad (Ehringer & Hornykiewicz, 1960), el uso de L-DOPA, precursor directo de la dopamina, como primer tratamiento (Barbeau, 1961), y la identificación de la acumulación alfa-sinucleína como principal causa de los cuerpos de Lewy y primer factor genético asociado a la enfermedad (Spillantini, 1997). Sin embargo, el diagnóstico temprano de la enfermedad, algo que no se ha podido realizar efectivamente aún, sigue siendo crucial para reducir el efecto de esta patología en la población afectada y en la sociedad. Una de las vías para lograr esto es la identificación de biomarcadores asociados a enfermedades (Strimbu *et al*., 2010), que además de ser de gran interés de cara al diagnóstico de una enfermedad (Parnetti *et al*., 2019), también son de interés para evaluar el pronóstico de los pacientes ya diagnosticados (Perlis, 2011).

Existen multitud de tipos de biomarcadores útiles para el diagnóstico de enfermedades (Strimbu *et al*., 2010). Estos pueden ser desde medidas fisiológicas y físicas, hasta marcadores ómicos, como pueden ser los marcadores de expresión génica o de metilación. Sin embargo, hay que añadir que no solo es de interés la identificación de biomarcadores *de novo*, práctica principal a lo largo de los últimos años, sino que poco a poco se hace notoria la necesidad de probar clínicamente los biomarcadores identificados como verdaderamente útiles para el diagnóstico y predictores del padecimiento de la enfermedad (Li & Le, 2017, Chen-Plotkin *et al*., 2018).

Entre los biomarcadores más comunes para el diagnóstico de la enfermedad de Parkinson destacan los síntomas tempranos (previos a la aparición de la deficiencia motriz), como son la disfunción olfatoria o hiposmia (Braak *et al*., 2003) y el trastorno del comportamiento del sueño por movimientos oculares rápidos (RBD, del inglés, *Rapid Eye Movement Sleep Behavior Disorder*) (Iranzo

et al., 2013). También destaca el estudio de neuroimágenes, principalmente del de la vía de la dopamina, por ejemplo a través de la técnica de tomografía computarizada por emisión monofotónica y la tomografía por emisión de positrones con fluorodopa (Suwijn *et al*., 2015). Otra vía común de estudio se centra en la búsqueda de marcadores en biofluidos. Entre los biomarcadores más estudiados destacan la proteína α-sinucleína, como recogen Nalls *et al*. (2014) y Fayyad *et al.* (2019), y las vesículas extracelulares liberadas por células nerviosas, como recogen Tomaso & Furlan (2017). Son de especial interés los biomarcadores identificados en sangre, dada su accesibilidad. Además, estudios más recientes se han centrado en la identificación de biomarcadores a partir del análisis del perfil de expresión génico a nivel de micro-ARNs en sangre (Yang *et al*., 2019; Leggio *et al*., 2017), pero también de ARNm en la barrera hematoencefálica (Correddu *et al*., 2019), y de ARN circulares y su efecto regulador (Fyfe, 2021).

Aunque la búsqueda de biomarcadores asociados a una enfermedad a partir de una única ómica es común, existen otros métodos que se centran en la búsqueda de biomarcadores a partir del estudio integrado de dos o más ómicas.

La integración multiómica es un método que ha tenido un auge considerable en la última década, y sobre todo en los últimos cinco años (2018 – 2023) (Figura 2). Mediante este método es posible obtener deducciones más completas sobre, por ejemplo, una patología. Por ejemplo, integrando datos del transcriptoma y del epigenoma es posible estudiar no solo el nivel de expresión y metilación, sino también el posible efecto de la metilación del ADN sobre la expresión génica. (Shen *et al*., 2012; Mo *et al*., 2013).

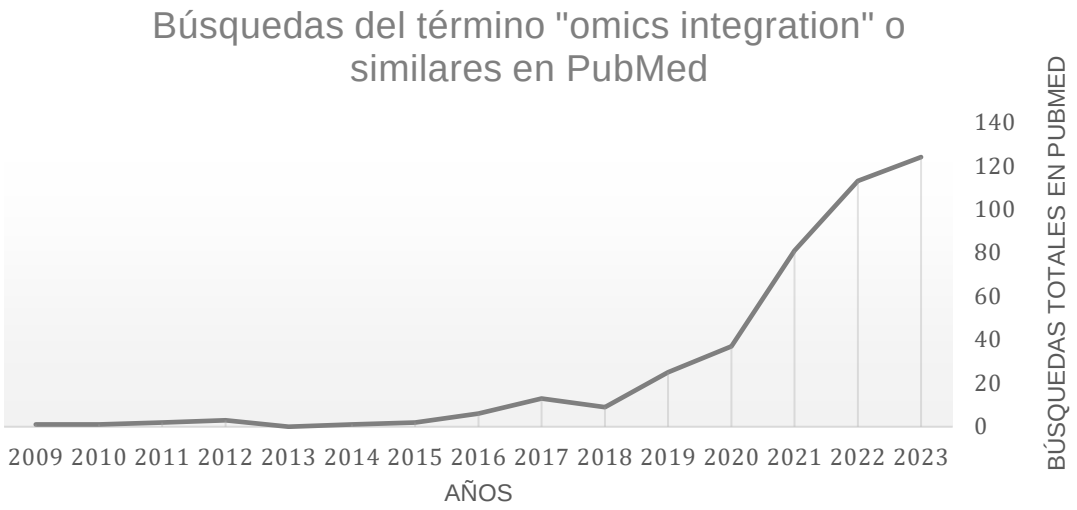

Figura 2. Búsquedas en PubMed de los términos "*omics integration*", "*multiomics integration*", "*omic integration*" y "*multiomic integration*" en conjunto. Fuente: Gráfica de elaboración propia, datos de PubMed (https://pubmed.ncbi.nlm.nih.gov/).

Mediante técnicas de integración la combinación de diferentes biomarcadores, tanto de tipo fisiológico como moleculares, pueden incrementar el poder del diagnóstico temprano (Li & Le, 2020). Sin embargo, los modelos elaborados, pese a su valor predictor, no permiten obtener deducciones integradas sobre las bases moleculares de la enfermedad.

Para poder estudiar los datos procedentes de las diferentes ómicas existen diversas herramientas, metodologías y flujos de trabajo. En este trabajo prestaremos una atención especial al estado actual de los procesos utilizados para manejar dados de transcripción génica, metilación de ADN, y su integración.

El análisis de datos de RNA-Seq requiere la utilización de diversas herramientas con funciones específicas que permitan realizar el alineamiento de las secuencias frente al genoma de referencia, realizar el contaje de lecturas alineadas, transformar los datos obtenidos y ejecutar análisis posteriores. El alineamiento de las lecturas frente al genoma de referencia correspondiente puede hacerse mediante el uso de programas alineadores. Entre los alineadores más comunes utilizados en el ámbito de la bioinformática se encuentran *Bowtie2, BBMAp, HISAT2* y *STAR.* Hay que tener en cuenta que no todos funcionan óptimamente para todos los tipos de ARN. Algunos, por ejemplo, son preferibles para el alineamiento de moléculas cortas (*small RNA-Seq*), como es *STAR*

(Musich *et al.*, 2021). También debe tenerse en cuenta la capacidad del alineador para gestionar las variantes, los tránscritos quiméricos, y el efecto del *splicing* de las moléculas de ARN. El alineador *STAR* (Dobin *et al.*, 2013), uno de los alineadores más utilizados en la comunidad científica, destaca por su precisión y velocidad, requiriendo en muchos casos una capacidad computacional baja, así como por su flexibilidad a la hora de funcionar bien tanto para secuencias cortas (menores de 200 pb) como para secuencias largas. Destaca también por su correcta gestión de eventos como la aparición de variantes y el *splicing* alternativo (Raplee *et al*., 2019; Musich *et al.*, 2021; Brüning *et al*., 2022).

Para el contaje de lecturas alineadas existen múltiples herramientas, como *HTSeq, FeatureCounts o Rsubread*. Sin embargo, el propio alineador *STAR* cuenta con un sistema de contaje desde la versión 2.5. (*--quantMode*) y cuyo resultado es idéntico al obtenido mediante la función *htseq-count* de *HTSeq*, como se comenta en el manual de *STAR* (https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf).

El análisis de expresión diferencial para identificar DEG puede llevarse a cabo en R utilizando multitud de paquetes, como son *DESeq2* y *EdgeR*, ambos muy utilizados por la comunidad científica para este tipo de análisis. Sin embargo, como se comenta en Li *et al.* (2022) y en Schurch *et al*. (2016), ambas herramientas parten de la asunción de normalidad en las distribuciones de los valores de expresión, motivo por el cual la inferencia estadística que llevan a cabo se basa en pruebas paramétricas. Este error lleva a la identificación de falsos DEG. En su lugar, recomiendan el uso de las pruebas estadísticas clásicas pertinentes en función de las asunciones que se pueden hacer sobre la distribución de los datos.

Por otro lado, el análisis de datos de metilación puede llevarse a cabo mediante el uso de diferentes paquetes, como el paquete *missMethyl* (Phipson *et al.,* 2015), *minfi* (Aryee *et al.,* 2014), o, más recientemente, el paquete *ChAMP* (Tian *et al*., 2017). Aunque ambos tienen funciones integradas, es necesario destacar que el paquete *ChAMP* destaca por su uso sencillo y accesible, además de su enfoque centrado en los datos de metilación.

Este paquete permite un análisis completo de datos de metilación de manera sencilla, consecutiva e integrada, desde la transformación de los archivos IDAT hasta la identificación de DMP. Para esta última fase hace uso del paquete *limma* (Ritchie *et al.*, 2015), que también asume normalidad para las distribuciones de valores de las variables, pues la prueba estadística que aplica es una modificación de la prueba T.

Sin embargo, ambos paquetes, como los comentados anteriormente para el análisis de datos de RNA-Seq parten de asunciones sobre la normalidad en la distribución de los datos. No debemos olvidar que es necesario evaluar la distribución de los valores de cada variable para cada grupo y, en base a los resultados, optar por la aplicación de pruebas paramétricas o pruebas no paramétricas.

Tras analizar los dos tipos de datos y una vez se han obtenido los diferentes resultados es posible llevar a cabo la integración de estos.

Existen multitud de técnicas de integración y correlación entre ómicas enfocadas a la caracterización de enfermedades, como se recoge en Hasin *et al*. (2017). Destacan en este trabajo como mediante la aplicación de diferentes ómicas a un estudio es posible obtener una visión general del flujo de la información asociada a una enfermedad. En función del objetivo del estudio de esta se deberán aplicar unas u otras técnicas. En algunos casos, y como se comentó anteriormente, una forma integración se basa en el estudio de la correlación entre variables y la creación de algoritmos de clasificación que utilicen distintos tipos de datos (Cappelli *et al*., 2018). En concreto, en el caso de los datos de RNA-Seq y metilación, un método más integrado y con base molecular es el análisis de metilación de rasgos cuantitativos de expresión aplicado a la metilación, como se comentó anteriormente. Se trata de un tipo de análisis de locus de rasgo cuantitativo o QTL (del inglés, *Quantitative Trait Locus*) (Abiola *et al*., 2003). Mediante esta técnica es posible conocer qué regiones del genoma tienen influencia en la variación fenotípica de un rasgo, normalmente debido a la interacción genética (Powder, 2020), esto se lleva a cabo a través de un proceso denominado como mapeado QTL. El análisis QTL puede aplicarse usando el paquete *MatrixEQTL* (Shabalin, 2012).

Otras vías de integración son menos específicas, y simplemente buscan determinar si existe correlación entre las variables de una ómica y las variables de otra ómica, si distinguir o tener en cuenta especialmente el tipo de ómica. Además, en muchos casos esta metodología abre nuevas formas de identificar biomarcadores asociados a fenotipos, como recoge Subramanian *et al*. (2020).

Existen multitud de herramientas para realizar integración multiómica, y pueden diferenciarse en función distintos criterios, como sus objetivos. Algunas herramientas permiten estudiar la correlación entre distintas ómicas desde una perspectiva contextual, como es el paquete CNAmet de R (Louhimo & Hautaniemi), otros permiten integrar distintas variables con el objetivo de identificar potenciales biomarcadores, como es el paquete *mixOmics* (Rohart *et al*., 2017). Una de las herramientas de integración más potentes es iCLusterPLus (Mo & Shen), con capacidad tanto para identificar biomarcadores como para extraer tener información contextual de la relación de las ómicas. Utilizando esta herramienta ha sido posible identificar subgrupos y patrones moleculares característicos en el cáncer de mama (Shen *et al*., 2009), y del glioblastoma (Shen *et al*., 2012).

En este proyecto se propone el estudio vanguardista de la enfermedad de Parkinson mediante el análisis integrado de los datos de metilación y transcripción del estudio de Henderson *et al.* (2021).

3 Metodología

A continuación, se describen las diferentes fases que han conformado el grueso del análisis estadístico de este proyecto (Figura 3).

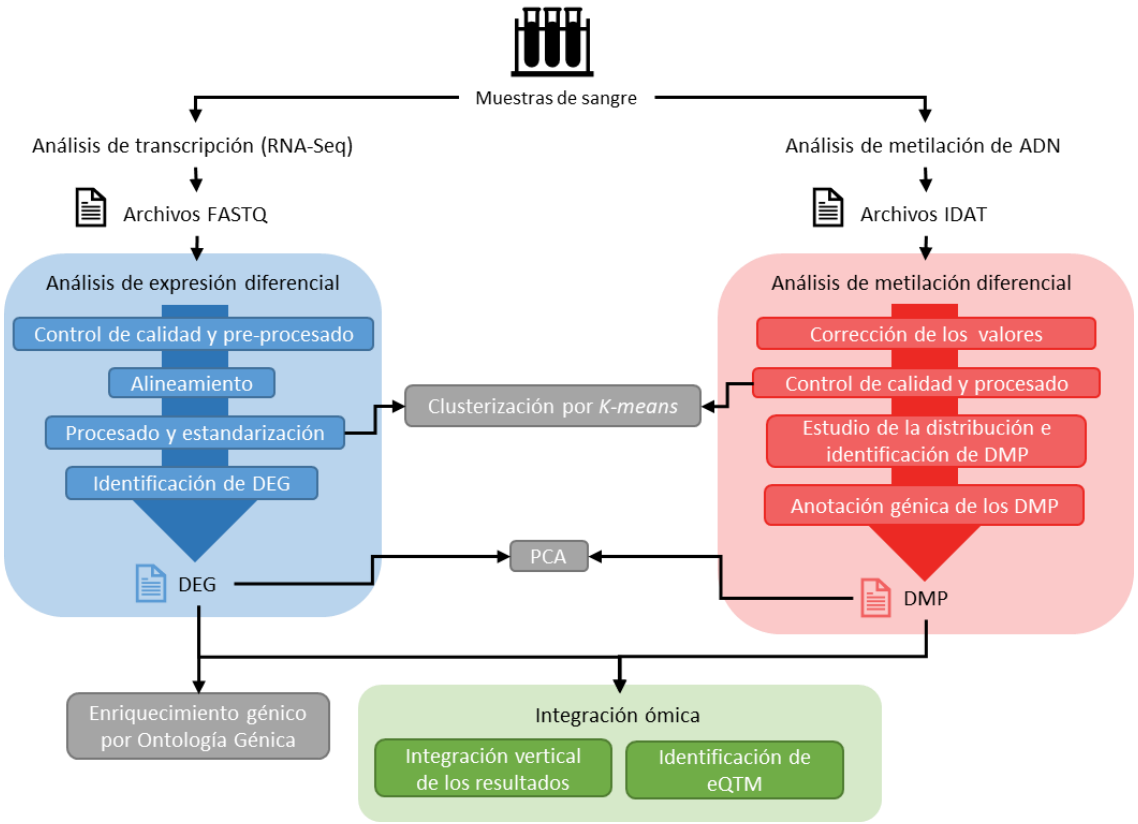

Figura 3. Flujo de trabajo seguido para este proyecto. Se diferencia una vía centrada en el análisis de expresión diferencial y otra centrada en el análisis de metilación diferencial. Los resultados de ambas vías son integrados en una última fase.

3.1 Obtención de los datos brutos

Los datos brutos que se utilizarán en este proyecto provienen del estudio de Henderson *et al.* (2021). Estos se obtuvieron de la base de datos pública *Gene Expression Omnibus* (GEO) (https://www.ncbi.nlm.nih.gov/geo/), y son los asociados al código de serie GSE165083. La serie se divide en un conjunto de datos de metilación (código GSE165081) en formato *array* de *Illumina HumanMethylation450 BeadChip*, y en un conjunto de datos de secuenciación de ARN (código GSE165082) en formato FASTQ.

Los datos de expresión se distribuyen en 26 muestras, mientras que los datos de metilación estaban distribuidos inicialmente en 28 muestras. En ambos casos

las muestras se separaron en un grupo Control (muestras de pacientes sanos en base a la patología de estudio) y en un grupo Parkinson o PD (del inglés, *Parkinson's Disease*, muestras de pacientes diagnosticados con la enfermedad de Parkinson). A los individuos enfermos se les diagnosticó Parkinson en estado temprano según la calificación de Hoehn y Yahr (1.89 ± 0.48) y según la Parte III de la Escala de Evaluación Unificada de la Enfermedad de Parkinson, correspondiente al examen motor (13.9 ± 2.1). Las muestras de sangre extraídas de los pacientes se conservaron a – 80°C. Posteriormente se extrajo el ARN general de los leucocitos sanguíneos, y se conservaron las muestras a – 20 °C.

Antes de comenzar los análisis se omitieron dos muestras del estudio de metilación sin réplicas en el estudio de expresión. Cada muestra tiene asociados un par de archivos FASTQ al tratarse de lecturas pareadas (*paired-end*).

3.2 Métodos para el análisis de expresión diferencial

A continuación, se definen cuáles han sido las diferentes fases seguidas en el análisis de expresión diferencial de los datos de *RNA-Seq*.

3.2.1 Control de calidad y pre-procesado

En primer lugar, se llevó a cabo un control de calidad sobre las secuencias de cDNA con el objetivo de evaluar la calidad de la secuenciación, el número de secuencias, y la existencia de adaptadores derivados del proceso de amplificación y secuenciación con *Illumina HiSeq 2000* (*Illumina, Inc.*).

El control de calidad sobre las secuencias contenidas en los archivos FASTQ se llevó a cabo utilizando las herramientas *FASTQC* (https://www.bioinformatics.babraham.ac.uk/) y *MultiQC* (Ewels *et al.*, 2016). Ningún parámetro extra se utilizó a la hora de utilizar la aplicación *FASTQC.* Los archivos referidos al control de calidad realizado por *FASTQC* se utilizaron como *input* para la aplicación *MultiQC* con el objetivo de obtener un resumen de estos.

Una vez revisados los resultados del control de calidad se procedió a recortar de forma pareada en las secuencias los extremos 5' y 3' por debajo de un valor Phred de calidad de secuenciación de 30. También se eliminaron los adaptadores detectados y las colas de poli(A) para evitar que puedan causar interferencias durante el alineamiento. Para ello se utilizó la herramienta *fastp*

(Chen *et al*., 2018). Los primeros cinco nucleótidos se recortaron de cada secuencia. También se identificaron y recortaron los adaptadores *TruSeq Index* 9 y *TruSeq Index* 12, además de las colas de poli(A).

3.2.2 Alineamiento

Para realizar el alineamiento se utilizó el alineador *STAR* (versión 2.7.9). En primer lugar, y antes de alinear las lecturas, se generó un índice del genoma humano a partir del archivo FASTA del genoma hg38 (versión GRChg38.p13) y el archivo *Gene Transfer Format* (GTF, versión GRCh38.109) de la Universidad Santa Cruz de California (*UCSC*, del inglés, *University of California Santa Cruz*), ambos obtenidos desde la *web* de *Ensembl* (http://www.ensembl.org/Homo_sapiens/Info/Index). El archivo GTF contiene las coordenadas de cada uno de los genes, esencial para poder realizar el contaje de cada una de las lecturas alineadas con cada gen.

El índice del genoma humano se generó utilizando el modo *genomeGenerate* para el parámetro *runMode*. Además, se especificó una longitud para la secuencia donadora/aceptora de 50 pb, la longitud máxima de las secuencias menos uno, siguiendo lo estipulado en el manual de *STAR*. Respecto al archivo GTF, se seleccionaron las características *exon*, gene_*name* y *transcript_name*. De este modo las lecturas se alinearán únicamente frente a los exones del genoma, y se guardará la información del nombre del gen y el transcrito.

Una vez se generó el índice se alinearon las lecturas de forma pareada. Para ello se utilizó el modo *alignReads* para el parámetro *runMode*. Además, se estableció que el archivo output fuese en formato BAM y estuviese ordenado por coordenadas. Para el parámetro *quantMode* se estableció el modo *GeneCounts*, necesario para obtener la tabla de contajes para cada gen. Para el tipo de alineamiento se mantuvo el modo establecido por defecto (*local*). Mediante este no se fuerza el alineamiento de los extremos 5’ y 3’ de las secuencias. Este proceso es denominado como alineamiento con *soft-clipping.* De este modo se evitan los problemas de alineamiento que pueden surgir debido a los restos de adaptadores o la baja calidad de secuenciación de los extremos 3’, principalmente.

A partir de las tablas de contaje y sumando las tres columnas de contajes se comprobó el tipo de librería de *RNA-Seq* generado previamente, del que no se tenía información. Las tres columnas de contaje tienen información sobre las lecturas alineadas sin diferenciar ente *forward* y *reverse* (*unstranded RNA-Seq*), alineadas correctamente según su sentido (las secuencias *forward* alinean con la hebra sentido y las secuencias *reverse* con la hebra antisentido, correspondiente a un *forward stranded RNA-Seq*) y alineadas inversamente según su sentido (las secuencias *forward* alinean con la hebra antisentido y las secuencias *reverse* alinean con la hebra sentido, correspondiente a un *reverse stranded RNA-Seq*). Finalmente se seleccionó la primera columna de contajes, correspondiente a la técnica *unstranded RNA-Seq.* Las primeras columnas de cada una de las tablas de contaje de cada muestra se fusionaron en una sola matriz de contajes.

3.2.3 Procesado y estandarización

Tras su obtención, la matriz con los contajes se cargó en R y se omitieron los contajes asociados a la clasificación *MissingGeneID.* Posteriormente se utilizaron funciones del paquete *edgeR* para llevar a cabo el filtrado de aquellos genes con pocos contajes (menos de 10 contajes en más de un 70 % de las muestras) y la estandarización en función del nivel de expresión general de cada muestra. El método elegido para calcular los factores de estandarización fue el TMM (del inglés, *Trimmed Mean of M-values*), establecido por defecto en la función *calcNormFactors*. La estandarización de los contajes usando los factores de estandarización se realizó utilizando el método de contajes-por-millón (CPM, del inglés, *Counts-Per-Million*), descrito por Chen *et al*. (2016). Para evitar la imposibilidad de calcular el logaritmo en base 0 de los contajes nulos se sumó 2 a cada uno de los contajes de cada gen y muestra antes de estandarizar.

3.2.4 Identificación de DEG

En el caso de los genes, para estudiar la expresión diferencial se utilizó la prueba estadística no paramétrica de la Suma de Rangos de Wilcoxon, también conocida como la prueba U de Mann-Whitney (Mann & Whitney, 1947), asumiendo que los contajes siguen una distribución binomial negativa (Tsonaka & Spitali, 2021). Los p-valores obtenidos se ajustaron mediante el método

desarrollado por Benjamini-Hochberg (1995), también conocido como la corrección de la tasa de falsos descubrimientos o FDR (del inglés, *False Discovery Rate*). Mediante esta prueba es posible reducir la tasa de errores tipo I cometidos a la hora de realizar múltiples comparaciones mediante pruebas estadísticas.

Para obtener el logaritmo en base 2 del *Fold-Change* ($\log_2$(FC) (abreviado como logFC) de cada gen en el grupo PD respecto al grupo Control, entendido como el cambio en el nivel de expresión de cada gen, se realizó un estudio de expresión diferencial mediante funciones del paquete *edgeR*. Del resultado obtenido únicamente se seleccionó la variable con información sobre el logFC de cada gen.

La identificación de DEG se llevó a cabo usando el p-valor obtenido y el logFC. Se aplicó un p-valor umbral de 0.05 y un |logFC| umbral de 1. Todos los genes con un p-valor inferior a 0.05 se consideraron como DEG. Mientras que los genes con un logFC superior a 1 se consideraron como sobreexpresados o *upregulated* en el grupo PD respecto al grupo Control; los genes con un logFC inferior a -1 se consideraron como infraexpresados o *downregulated* en el grupo PD respecto al grupo Control.

Los p-valores y logFC de cada gen se utilizaron para generar *volcano plots* (gráficos de tipo *volcano*). Además, los DEG identificados se utilizaron para generar mapas de calor o *heatmaps* mediante la función *pheatmap* del paquete *pheatmap.* Para la elaboración de los *heatmaps* se estandarizaron los datos y se omitió la clusterización por grupos. Los DEG más sobreexpresados e infraexpresados se representaron mediante diagramas de cajas ordenados en función del logFC.

3.3 Métodos para el análisis de metilación diferencial

El análisis de metilación diferencial se llevó a cabo principalmente utilizando el paquete *ChAMP* (versión 3.16) y el *software* R (versión 4.2.3). En primer lugar, los archivos IDAT brutos, con información sobre los metadatos y los valores de intensidad de metilación de las muestras se cargaron utilizando la función *champ.load*. Al cargar los datos, a partir de los datos IDAT se generaron automáticamente los valores beta necesarios para los siguientes pasos.

En total se obtuvieron 12 muestras PD y 14 Controles. Las muestras 046 y 048 del grupo PD fueron omitidas por no tener réplicas en el estudio de expresión diferencial.

3.3.1 Corrección de los valores de metilación de las sondas tipo II

La matriz de valores beta se estandarizó utilizando la función *champ.norm* del paquete *ChAMP*. Esto se hizo con el objetivo de corregir las diferencias existentes entre los valores de metilación de las sondas tipo I y las sondas tipo II debidas a la química por la cual se calcula el porcentaje de metilación (Dedeurwaerder *et al*., 2011; Teschendorff *et al*., 2013). De este modo, ajustando la distribución de valores de las sondas tipo II para que se ajusten a la distribución de las sondas tipo I, se corrige el efecto *bias* por el cual las sondas tipo II tienen una mayor probabilidad de ser identificadas como diferencialmente metiladas (Teschendorff *et al*., 2013). La estandarización se llevó a cabo mediante el método de normalización cuantitativa de la mezcla de valores beta o BMIQ (del inglés, *Beta-mixture Quantile Normalization*), según se recomienda en Teschendorff *et al*. (2013).

3.3.2 Control de calidad y procesado

Tras la corrección, el control de calidad sobre los valores beta de los diferentes sitios CpG se llevó a cabo mediante la representación de un gráfico de densidad de los valores beta para cada muestra. El objetivo de este gráfico es evaluar si la distribución de los valores beta sigue el perfil esperado, e identificar posibles muestras problemáticas que deban ser omitidas del estudio.

Para identificar el posible efecto *Batch* existente en las muestras se utilizó la función *champ.SVD*. además de un ensayo de clusterización por K-means (Apartado 4.4.). Esta función lleva a cabo un proceso de descomposición en valores singulares (DVS) sobre la matriz de valores beta. Mediante pruebas paramétricas permite identificar el efecto de las covariables sobre los valores. Junto al grupo al que pertenece cada muestra se seleccionaron como covariables la edad y el género de los pacientes de los que se obtuvieron.

Una vez llevada a cabo la DVS de la matriz se representó gráficamente el posible efecto de las covariables conocidas.

3.3.3 Estudio de la normalidad y la homocedasticidad

La normalidad y la homocedasticidad de las distribuciones de valores beta para cada sitio CpG se estudió en función del grupo. Para ello, y puesto que cada grupo está formado por menos de 50 muestras, se utilizó la prueba estadística de Shapiro-Wilks (Mishra *et al*., 2019) para evaluar la normalidad, y la prueba de Levene para evaluar la homocedasticidad, en ambos casos haciendo uso de funciones estadísticas básicas de R.

Una vez aplicadas las pruebas, los p-valores obtenidos se ajustaron mediante el método FDR (Benjamini & Hochberg, 1995).

3.3.4 Identificación de DMP

Para evaluar la existencia de diferencias estadísticamente significativas en los sitios CpG en función de los grupos de estudio se utilizaron las opciones del paquete *limma*. En primer lugar se creó una matriz de diseño comparativa entre ambos grupos. Posteriormente se ajustó un modelo mediante la función *lmFit* y se realizaron los contrastes utilizando las funciones *contrasts.fit* y *eBayes*. La prueba estadisitca paramétrica T de Student, ajustada según los parámetros de *limma*, fue aplicada a las muestras. Los p-valores obtenidos para cada uno de los sitios CpG se ajustaron utilizando el método FDR (Benjamini & Hochberg, 1995). También se calculó para cada sitio CpG la diferencia de la media de valores beta de cada grupo ($\Delta\beta$).

Los criterios utilizados para identificar CpGs como DMP fueron un p-valor obtenido por debajo del umbral de 0.05 y un valor $\Delta\beta$ superior o inferior al $|\Delta\beta|$ umbral de 0.1 y – 0.1, respectivamente. Todos los sitios CpG con un p-valor inferior a 0.05 se consideraron como DMP; aquellos con un valor $\Delta\beta$ superior a 0.1 se consideraron como DMP hipermetilados en el grupo PD respecto al grupo Control, mientras que aquellos con un valor $\Delta\beta$ inferior a -0.1 se consideraron como DMP hipometilados en el grupo PD respecto al grupo Control.

A partir de los p-valores y los $\Delta\beta$ se representó un gráfico de tipo *volcano* en el que quedó reflejado cada sitio CpG y DMP. Los DMP identificados se utilizaron para *heatmaps* a partir de los datos.

3.3.5 Estudio de la distribución de los DMP y anotación génica

Una vez se obtuvieron los DMP se estudió su distribución en función de criterios de proximidad entre CpGs y en función de la región del gen en la que se sitúan. También se realizó la anotación génica, entendida como la identificación de genes con al menos un DMP asociado.

Para ello se utilizó la información contenida en el conjunto *probe.features* del paquete *ChAMPdata*, con información sobre la posición de cada sitio CpG y los genes asociados en caso de no encontrarse en una región intergénica.

Las calificaciones que se tuvieron en cuenta según la proximidad entre CpGs y sus agrupamientos fueron islas o *islands* (gran porcentaje de CpGs, tamaño mayor de 500 pb), *shores* (>2 kpb corriente arriba o abajo de las islas), *shelves* (>4 kpb corriente arriba o abajo de las islas) y el *open sea* (regiones alejadas de las islas); las calificaciones según la región del gen fueron primer exón (*1st exon*), región 3' UTR, región 5' UTR, cuerpo del gen (*body*), región intergénica (*IGR*), región situada 200 pb y 1500 pb corriente arriba del sitio de transcripción (*TSS200* y *TSS1500*, respectivamente).

3.4 Ensayo de clusterización por K-means

A partir de los valores beta corregidos y de los valores de contaje estandarizados se realizaron dos ensayos de clusterización no supervisada por *K*-means, uno para cada ómica, con el objetivo de comprobar si las variables con mayor desviación permiten separar las muestras en los dos grupos esperados, o si pueden identificarse otros grupos.

En primer lugar, se calculó la desviación estándar de cada variable. Puesto que el número de sitios CpG inicialmente estudiados es muy alto, más de 400.000, se seleccionaron aquellos sitios con una desviación superior al 50 % de la media. Para los genes, en mucha menor cantidad, se aplicó un umbral del 10 % de la media.

A partir de las variables seleccionadas se llevó a cabo el ensayo de clusterización utilizando las funciones *fviz_nbclust* (paquete *factoextra)* para determinar el número óptimo de grupos, y *fviz_cluster* (paquete *factoextra*) y *kmeans* (paquete *stats*) para agrupar las muestras.

3.5 Análisis de los Componentes Principales

A partir de los DMP y los DEG identificados se llevaron a cabo múltiples Análisis de los Componentes Principales, o PCA (del inglés, *Principal Component Analysis*) para determinar si las principales fuentes de variabilidad entre las muestras (primer y segundo componente) se corresponden con las diferencias existentes entre los grupos PD y Control.

Para ello se utilizaron las funciones *PCA* (paquete *FactoMineR*) y *fviz_pca_ind* (paquete *factoextra*), además de los datos estandarizados.

En total se realizaron cuatro PCAs a partir de los DMP y DEG identificados, los DMP hiper e hipometilados en PD respecto a Control, y los DEG sobre e infraexpresados en PD respecto a Control.

3.6 Análisis de enriquecimiento génico por ontología génica

El análisis de enriquecimiento génico o enriquecimiento funcional por ontología génica se llevó a cabo haciendo uso de los paquetes *clusterProfiler* y *enrichplot*. Para ello primero los símbolos de los genes se transformaron en identificadores ENTREZ. Estos se utilizaron como *input* de la función *enrichGO* para obtener los procesos biológicos, las funciones moleculares y los componentes celulares significativamente afectados. A partir de los términos obtenidos se generaron gráficos de árbol y redes gen-término.

Los genes utilizados para el enriquecimiento a partir del cual se generaron los gráficos fueron los DEG previamente identificados. También se llevó a cabo el enriquecimiento a partir de aquellos DEG con al menos un DMP asociado, de los que obtuvieron las tablas de términos.

3.7 Integración ómica de los resultados

La integración de los resultados obtenidos a partir del estudio de ambas ómicas por separado se hizo siguiendo dos vías.

Por un lado, se llevó a cabo un análisis eQTM (del inglés, *Expression Quantitative Trait Methylation analysis*) a partir de los DMP y los DEG. Esto se hizo con el objetivo de identificar DMP en regiones corriente arriba (cis) y

corriente abajo (trans) de DEG concretos cuyo nivel de expresión pueda estar correlacionado con el grado de metilación de los DMP.

Por otro lado, se realizó una integración vertical de ambas ómicas, entendiéndose por "integración vertical" como la integración de diferentes ómicas estudiadas en las mismas muestras. Esto se hizo con el objetivo de estudiar la correlación entre ambas ómicas y sus variables, y su capacidad para explicar variabilidad existente entre ambos grupos de estudio.

3.7.1 Análisis de metilación de rasgos cuantitativos de expresión

Para realizar el análisis de metilación de rasgos cuantitativos de expresión o eQTM (del inglés, *expression Quantitative Trait Methylation*) se prepararon las tablas con cada DMP y su localización genómica, las tablas con cada DEG y su localización genómica, además de las matrices de valores beta corregidos y contajes estandarizados. Para identificar la localización de cada gen se utilizaron funciones del paquete *biomaRt*. Mediante la función use*Mart* se seleccionó la base de datos de *Ensembl* con la información del genoma en la versión hg38 (versión GRChg38.p13).

El análisis eQTM se realizó utilizando la función *Matrix_eQTL_main* del paquete *MatrixEQTL.* Se consideraron como eQTM todos aquellos DMP situados en el cuerpo de un DEG, próximos a ellos (eQTM cis) o alejados (eQTM trans), y entre los que exista relación a nivel de metilación-nivel de expresión.

El modelo utilizado para determinar el efecto de las variables genotípicas sobre el fenotipo fue un modelo lineal, y la prueba estadística realizada para determinar significancia fue la prueba T de Student. Los p-valores obtenidos se ajustaron mediante el método FDR.

Además, se establecieron como covariables el género y la edad de los pacientes de los que se obtuvieron las muestras.

3.7.2 Integración vertical mediante DIABLO

Para la integración vertical de ambas ómicas se partió de la matriz de contajes estandarizada y de la matriz de valores beta corregidos.

El método utilizado para llevar a cabo la integración fue el Análisis Discriminante de los Mínimos Cuadrados Parciales en multibloque en su versión *sparse* (*multiblock sPLS-DA*¸ del inglés, *Partial Least Squares*), también conocido como integración tipo DIABLO (del inglés, *Data Integration Analysis for Biomarker Discovery using Latent variable approaches for Omics studies*) (Singh *et al*., 2019), en su versión supervisada. Para poder aplicar este método se utilizaron funciones del paquete de R *mixOmics* (versión 6.24.0) perteneciente al proyecto Bioconductor (https://www.bioconductor.org/).

En primer lugar, se elaboró un modelo PLS basado en la primera componente a partir de ambas ómicas mediante la función *pls*. Seguidamente se extrajo el valor de cada muestra para el primer componente obtenido, por un lado, a partir de la matriz de contaje y, por otro lado, a partir de la matriz de valores beta. Ambos conjuntos se compararon y se obtuvo el valor de correlación entre el primer componente ambas ómicas.

A continuación, se creó la matriz de correlación entre las dos ómicas y se entrenó un modelo PLS-DA basado en las dos matrices de datos ómicos. Se estableció que el modelo valorase la selección de entre 100 y 200 DMP, y entre 50 y 100 DEG. A partir del modelo se realizó la selección del número mínimo posible de variables con la mejor capacidad clasificadora. El modelo se entrenó usando la función *tune.block.splsda*. Se especificó que el modelo incluyese los dos componentes principales derivados de las matrices y que la validación fuese por *cross-validation* con 10 *folds* y 5 repeticiones. El resto de los parámetros se mantuvieron por defecto.

El resultado obtenido, con información sobre el número de DMP y DEG óptimos para la clasificación, así como las matrices de datos, se utilizaron para generar un modelo de análisis discriminante mediante la función *block.splsda*. Los dos primeros componentes principales se incluyeron en el modelo.

A partir del modelo creado se generaron gráficos de correlación entre los componentes principales de cada ómica, las variables predictoras frente a los componentes, las variables más contributivas, y un *heatmap* con clusterización de muestras en base a las variables seleccionadas. Para ello se usaron funciones específicas del paquete *mixOmics*.

4 Resultados

4.1 Control de calidad, normalidad y homocedasticidad de los datos

En primer lugar, se llevaron a cabo una serie de controles sobre los datos para comprobar la calidad de estos. Se analizó tanto la calidad de la secuenciación, como la distribución de los valores beta de metilación de los sitios CpG estudiados. Por un lado, respecto a la secuenciación realizada, se comprobó que esta fue óptima y, por tanto, el riesgo de encontrar nucleótidos erróneos fue mínimo. Además, se observó que el contenido en GC de las secuencias fue de aproximadamente el 50 %. Por otro lado, respecto a los valores beta de metilación, se comprobó que sus distribuciones siguieron el perfil esperado. Respecto al estudio del efecto Batch y de otras covariables en los datos de metilación, no se obtuvieron resultados que indicasen un efecto significativo de covariables como el género o la edad de los pacientes de los que se obtuvieron las muestras. Más información sobre los controles de calidad puede encontrarse en el apartado Anexo I.

Respecto a la normalidad de las distribuciones de los valores de los sitios CpG, se pudo considerar que más del 50 % de sitios CpG siguieron una distribución normal en ambos grupos, y presentaron homocedasticidad entre sí. Por ende, de cara a las fases posteriores, se optó por utilizar pruebas estadísticas paramétricas para estudiar la metilación diferencial.

4.2 Resultados de la clusterización por *K-means*

A partir de las matrices de datos estandarizadas y corregidas se llevaron a cabo ensayos de clusterización por *K-means* para identificar posibles grupos desconocidos o muestras problemáticas. Para ello se identificaron e utilizaron las variables con mayor desviación entre las muestras según los criterios establecidos previamente.

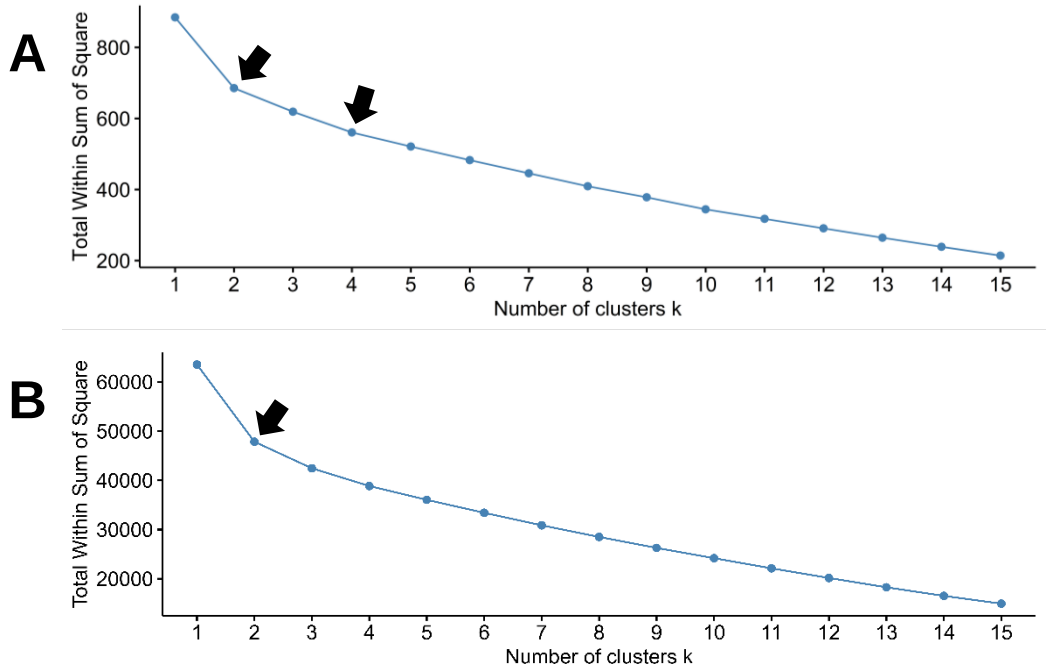

Figura 4. Número óptimo de grupos dependiendo de la suma total de cuadrados dentro del grupo (TWSS), entendida como una medida de la variación entre grupos. Se usaron las variables con mayor dispersión de A) valores beta de metilación corregidos, y B) contajes de expresión estandarizados. Se indica en el eje X el número de grupos entre los que se repartieron las muestras y en el eje Y el valor de TWSS obtenido. Se indican con flechas negras los puntos que pueden considerarse como "codos" en los gráficos, a partir de los cuáles la pendiente gráfica se ve reducida y el valor de TWSS se ve reducido cada vez menos.

La clusterización por *K-means* realizada utilizando la matriz de valores beta corregidos permitió identificar como óptimo el agrupamiento de las muestras en entre dos y cuatro grupos (Figuras 4A). Por otro lado, la clusterización realizada a partir de la matriz de contajes estandarizada permitió identificar como óptimo el agrupamiento en dos grupos (Figura 4B). En ninguno de los gráficos se observa un "codo" evidente, indicativo de que la diferencia entre las muestras es considerablemente elevada, no pudiendo agruparlas en grupos de gran tamaño.

En el caso del agrupamiento en cuatro grupos a partir de los datos de metilación, algunos de estos grupos estuvieron formados solo por una o dos muestras, y los grupos mayores estaban formados por una mezcla de muestras del grupo Control y PD (Figura 5A). Las muestras con mayor diferencia respecto al resto fueron las muestras 051_PD y 021_C, el resto de las muestras se agruparon relativamente próximas en el área bidimensional conformada por los dos primeros componentes. Una situación similar se observó en el agrupamiento de

las muestras en dos grupos a partir de los datos de transcripción (Figura 5B). Las muestras más alejadas en el espacio bidimensional fueron las muestras 012_PD y 028_PD. El resto de las muestras se agruparon relativamente próximas entre sí.

Puesto que no se identificaron nuevos grupos evidentes, los posteriores análisis se desarrollaron teniendo en consideración los dos grupos originales (Control y PD).

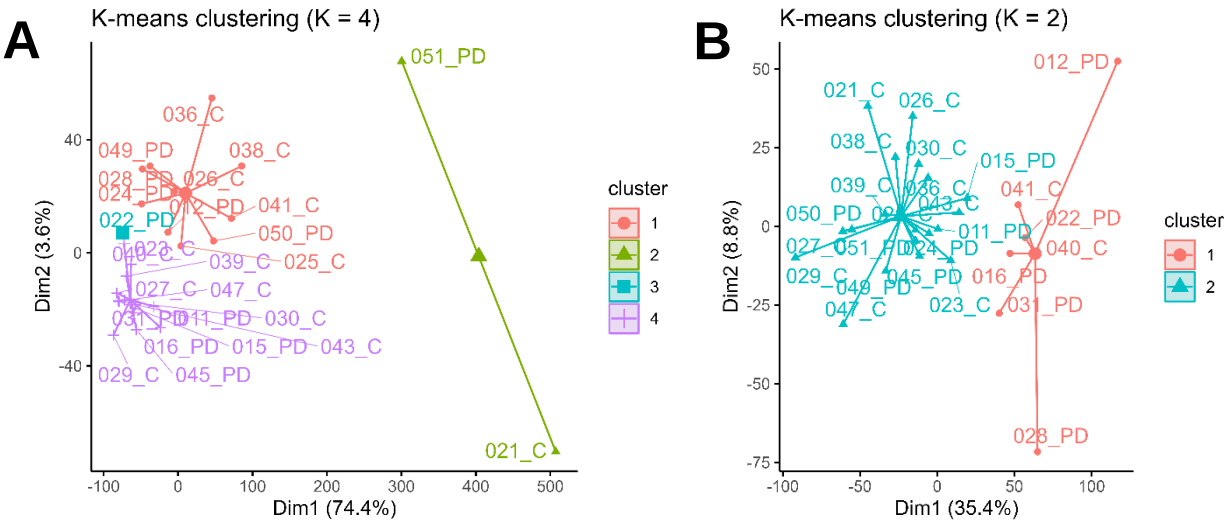

Figura 5. Gráfico bidimensional de los dos primeros componentes de cada matriz de datos. Se indica en A) la descomposición de la matriz de valores beta y el resultado de la clusterización para 4 grupos; y en B) la descomposición de la matriz de contajes y el resultado de la clusterización para 2 grupos.

4.3 Identificación de DMP y DEG

En pos de conocer las diferencias existentes en la expresión génica y la metilación del ADN entre las muestras Control y PD se llevó a cabo la identificación de DEG y DMP. Para ello se realizaron análisis de expresión y metilación diferencial, respectivamente, estableciendo un p-valor umbral de 0.05 en ambos casos, un $|\Delta\beta|$ umbral de 0.1, y un |logFC| umbral de 1.

Dado que el ajuste del p-valor recomendado por algunos autores, como Jafari & Ansari-Pour (2019), produjo una excesiva disminución de los DEG y DMP identificados, se optó por utilizar el p-valor sin ajustar en las fases consecutivas. Más detalles sobre esta decisión se pueden encontrar en el apartado 6. Discusión.

Por un lado, a partir de los datos de metilación se identificaron 14065 DMP, de los cuales 48 estuvieron hipermetilados ($\Delta\beta \geq 0.1$) y 414 hipometilados ($\Delta\beta \leq -0.1$) en el grupo PD respecto al grupo Control (Figura 6A). Además, se comprobó que los DMP generales se distribuían entre 5744 genes, los 48 DMP hipermetilados se distribuían entre 28 genes, y los 414 DMP hipometilados se distribuían entre 174 genes. Por otro lado, de entre los 12264 genes que quedaron tras el filtrado por baja expresión, se identificaron 1678 DEG, 10 sobreexpresados (logFC ≥ 1) y 16 infraexpresados (logFC ≤ – 1) (Figura 6B). Además, de entre los 1678 DEG, 369 tenían asociado al menos 1 CpG con metilación diferencial entre ambos grupos. En el Anexo II se pueden consultar los datos sobre los DEG con al menos 5 DMP asociados, o de interés para el proyecto.

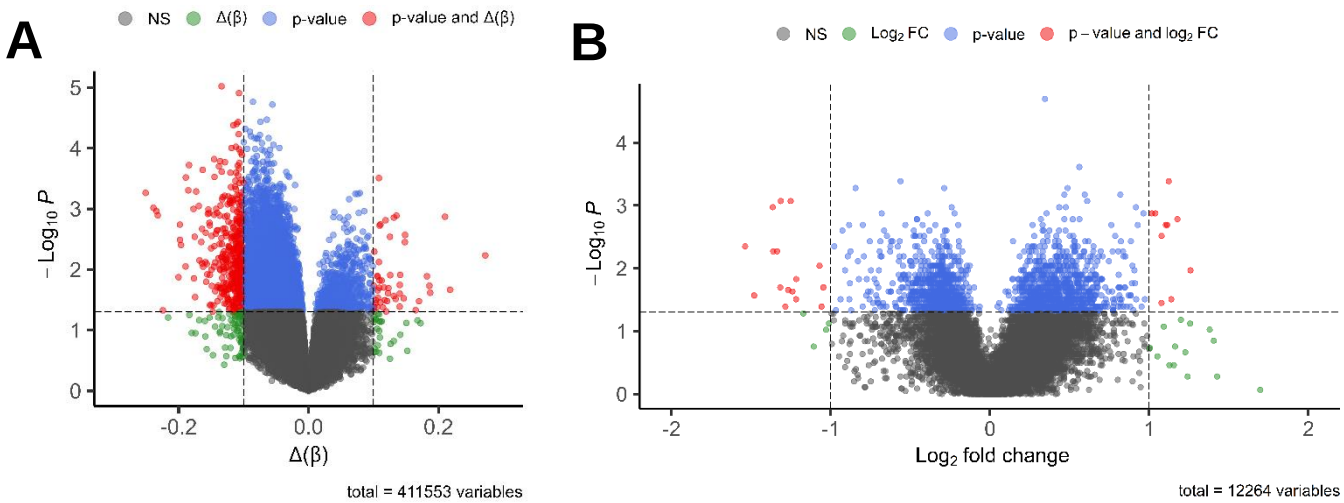

Figura 6. *Volcano plots* de A) los sitios CpG, y B) los genes estudiados. Se indican en el eje X los valores $\Delta\beta$ y $\log_2$(FC), según corresponda, y en el eje Y el -$\log_{10}$(p-valor).

Los DEG y DMP identificados se utilizaron para generar mapas de calor o *heatmaps* (Figuras 7 y 8). En general, en los *heatmaps* generados para ambas ómicas se observaron perfiles diferenciados entre los dos grupos de estudio. La distinción fue algo más notoria en los *heatmaps* de los DMP (Figura 7A), y aún más clara en el caso de los DMP hipo e hipermetilados (Figura 7B).

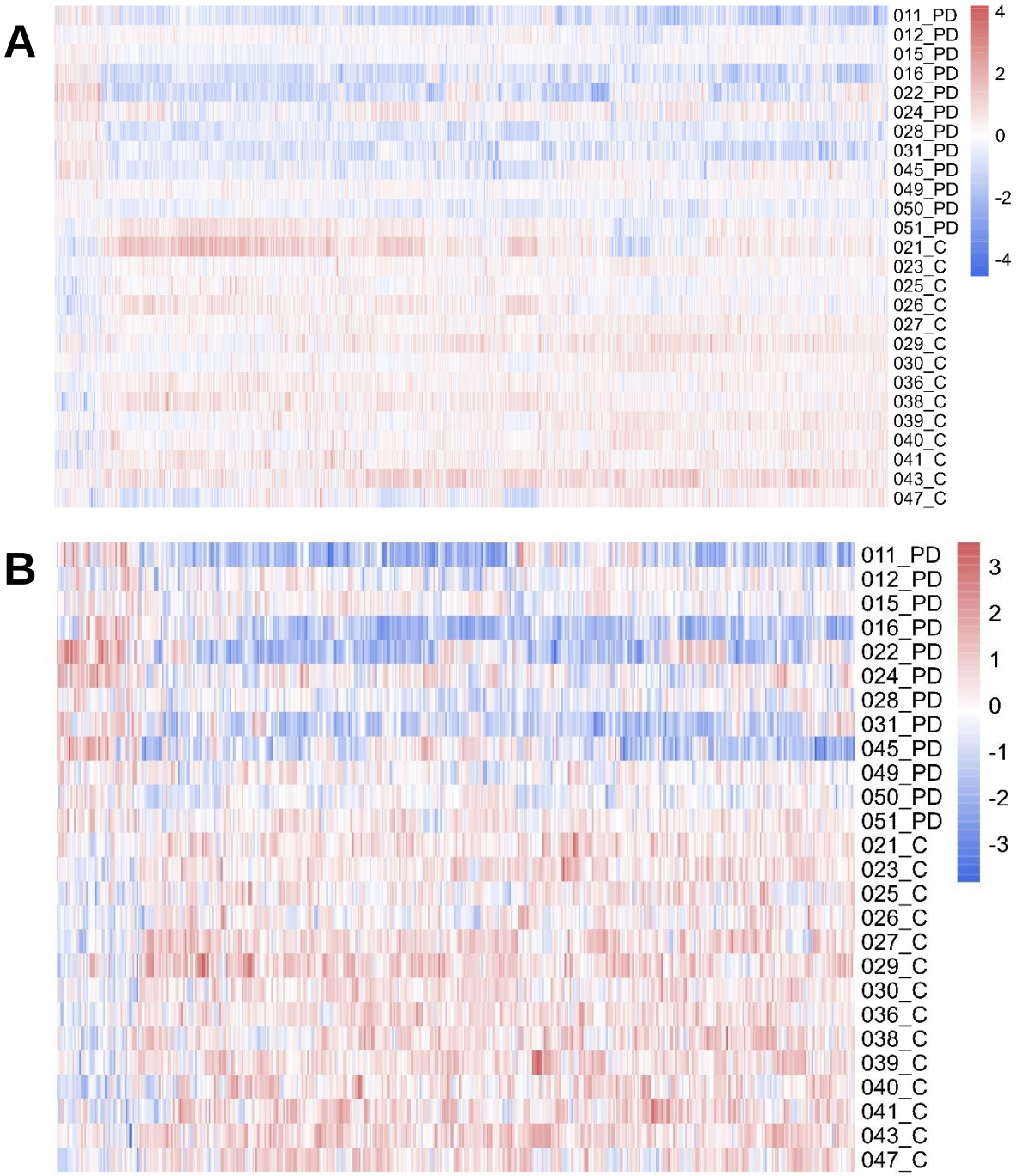

Figura 7. *Heatmaps* obtenidos a partir de los datos de metilación. Para la obtención de cada mapa se utilizaron A) los DMP identificados, y B) los DMP hipermetilados e hipometilados. Las muestras PD pertenecen al grupo Parkinson, mientras que las muestras C pertenecen al grupo Control.

Aunque las muestras de los diferentes grupos mostraron un perfil similar (Figura 7A y 7B), la muestra 021_C mostró perfil significativamente más hipermetilado en comparación al resto de muestras de su mismo grupo (Figura 7A). Además, se observó que la hipometilación de las DMP

hipometiladas en PD respecto a Control fue notoriamente más marcada que la hipermetilación de las DMP hipermetiladas en PD respecto a Control (Figura 7B).

Respecto a los DEG, los valores de expresión fueron menos homogéneos entre los grupos (Figura 8A), existiendo perfiles invertidos entre muestras del mismo grupo. Fue el caso de las muestras 050_PD, 049_PD, 041_C y 040_C. En general, el perfil de expresión fue más similar entre muestras del mismo grupo al utilizar los DEG sobre e infraexpresados (Figura 8B).

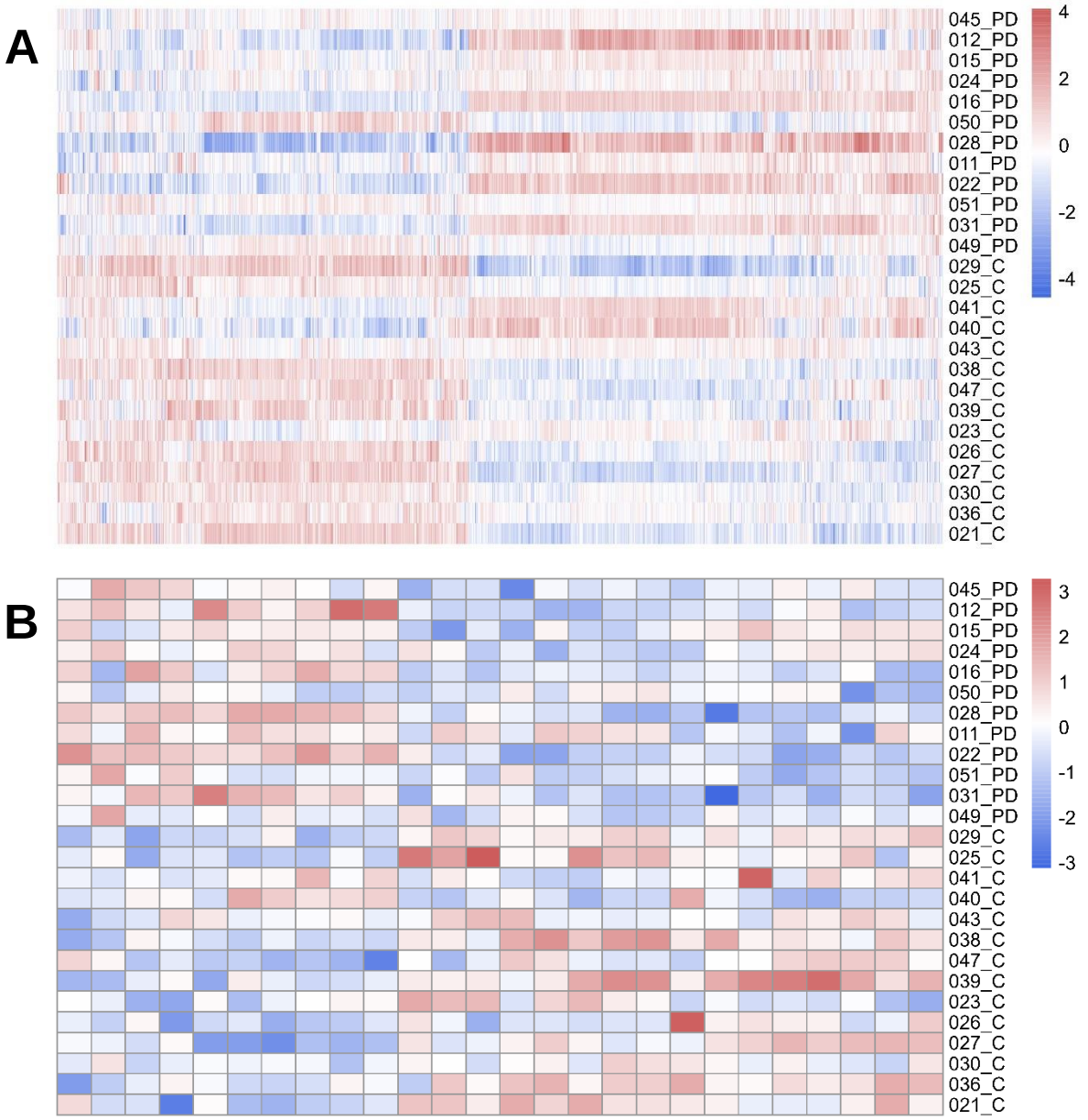

Figura 8. *Heatmaps* obtenidos a partir de los datos de transcripción. Para la obtención de cada mapa se utilizaron A) los DEG identificados, y B) los DEG sobreexpresados e infraexpresados. Las muestras PD pertenecen al grupo Parkinson, mientras que las muestras C pertenecen al grupo Control.

A continuación, se muestran los seis DEG más sobreexpresados (Figura 9A) y los seis DEG más infraexpresados (Figura 9B) en el grupo PD respecto al grupo Control.

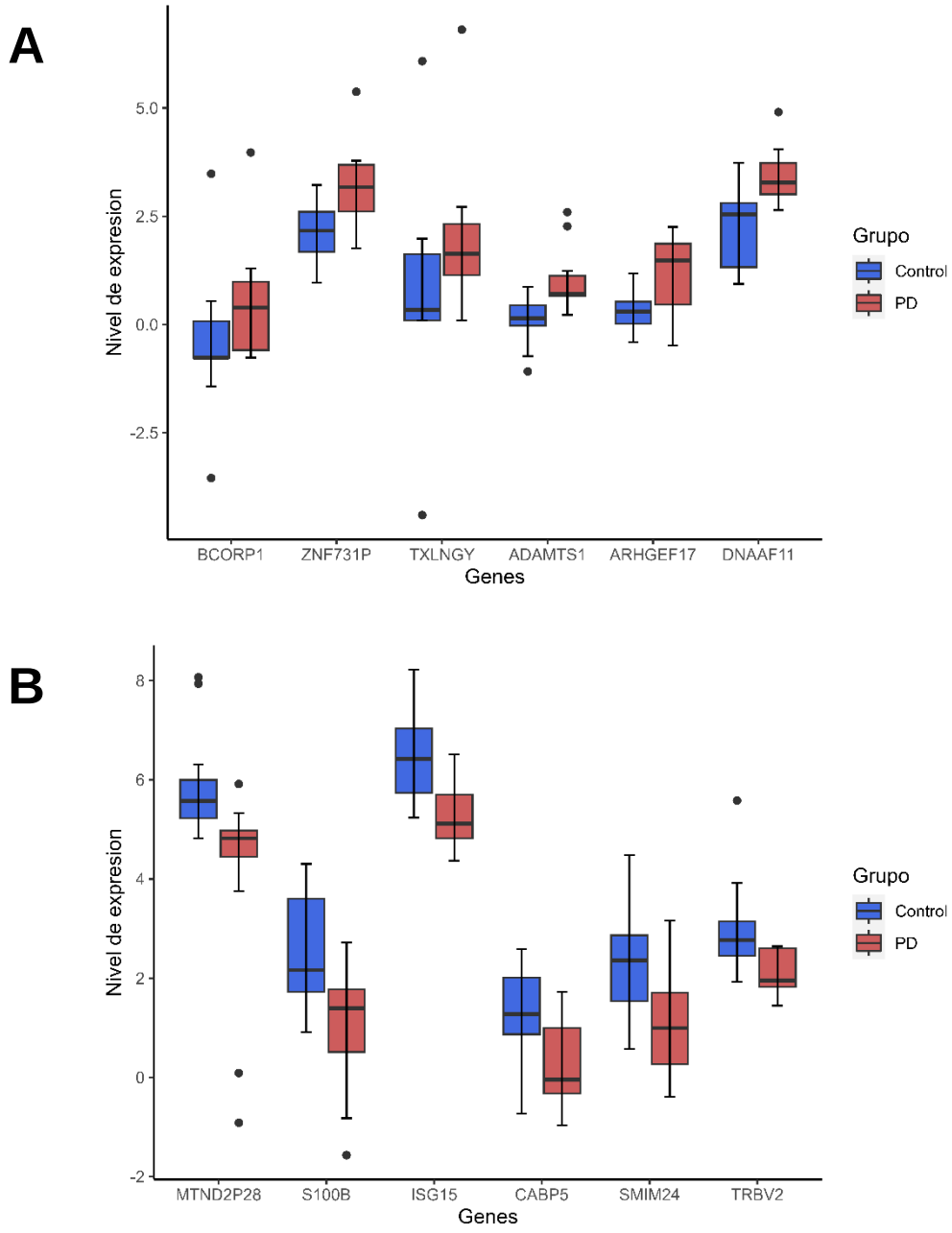

Figura 9. Top 6 DEG A) sobreexpresados y B) infraexpresados en PD respecto a Control. Los genes se encuentran ordenados en función del |logFC| asociado.

4.4 Distribución de DMP

Los DMP identificados se anotaron en función de su localización según la proximidad entre sitios CpG y según la posición respecto al gen asociado (Figuras 10 y 11, respectivamente).

La distribución de los DMP se hizo, en primer lugar, en cuatro regiones definidas según la proximidad entre sitios CpG. Estas fueron las islas o *islands*, los *shores*, los *shelves* y el *open sea* (Figura 10). En segundo lugar, se estudió su distribución en regiones definidas según la posición respecto al gen asociado. Estas fueron la región 5'UTR, el primer exón, el cuerpo génico o *body*, la región 3'UTR, las regiones situadas 200 pb y 1500 pb corriente arriba del sitio de transcripción (*TSS200* y *TSS1500*, respectivamente), y regiones intergénicas (Figura 11).

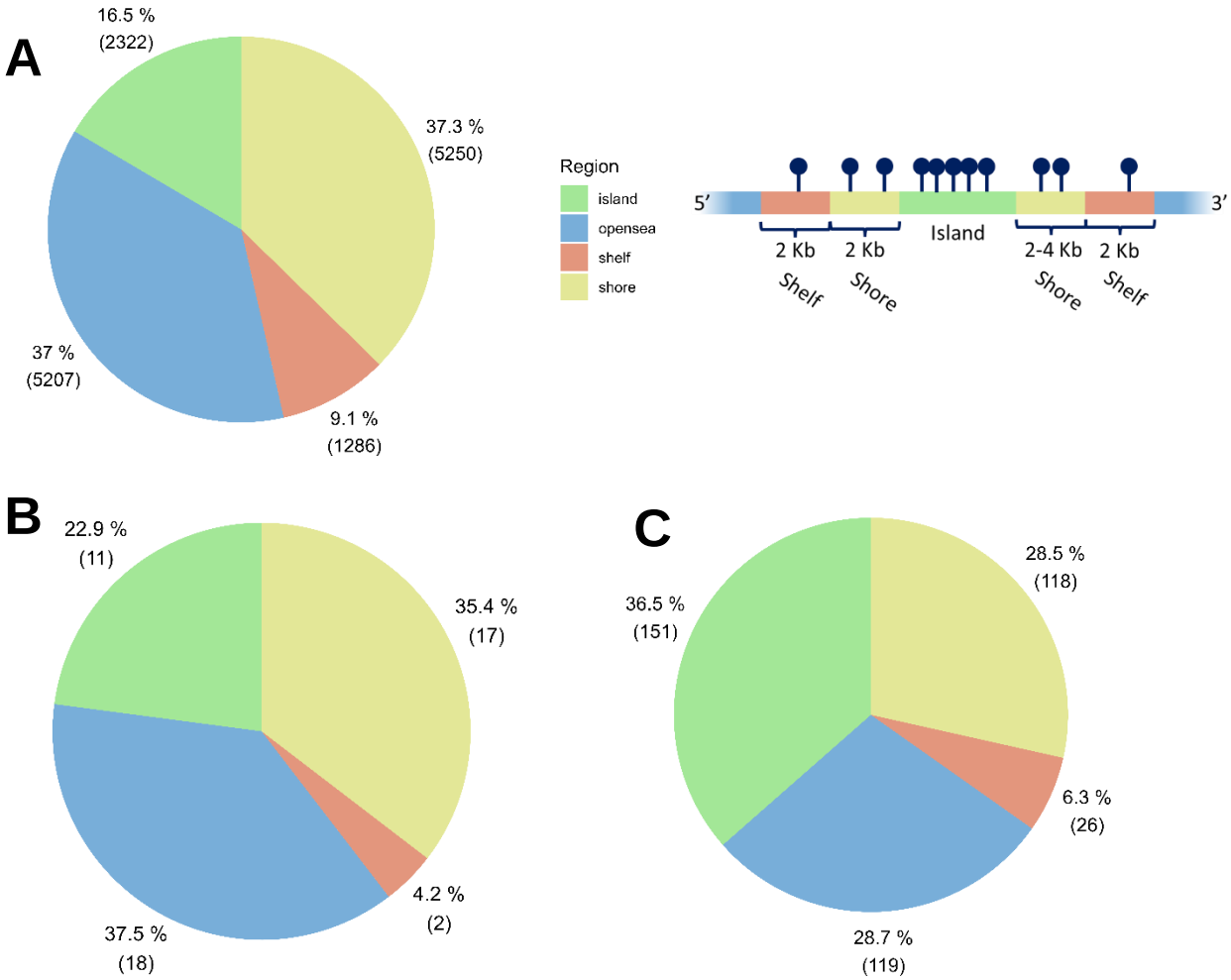

Figura 10. Distribución de los DMP en función de las regiones definidas por proximidad entre sitios CpG. Se muestran en A) la distribución general de los DMP, en B) la distribución de DMP hipermetilados, y en C) la distribución de DMP hipometilados.

Respecto a la primera clasificación, en general la mayoría de los DMP se encontraron en regiones *shores* (37.3 %), relativamente próximas a islas CpG (Figura 10A); el porcentaje de DMP hipermetilados situados en *shores* fue similar (35.4 %), superado por el porcentaje de DMP hipermetilados situados en el *open*

sea (37.5 %). Sin embargo, los DMP hipometilados se situaron principalmente en islas CpG (36.5 %), junto a un porcentaje relativamente alto que se situó en regiones *shore* (28.5 %).

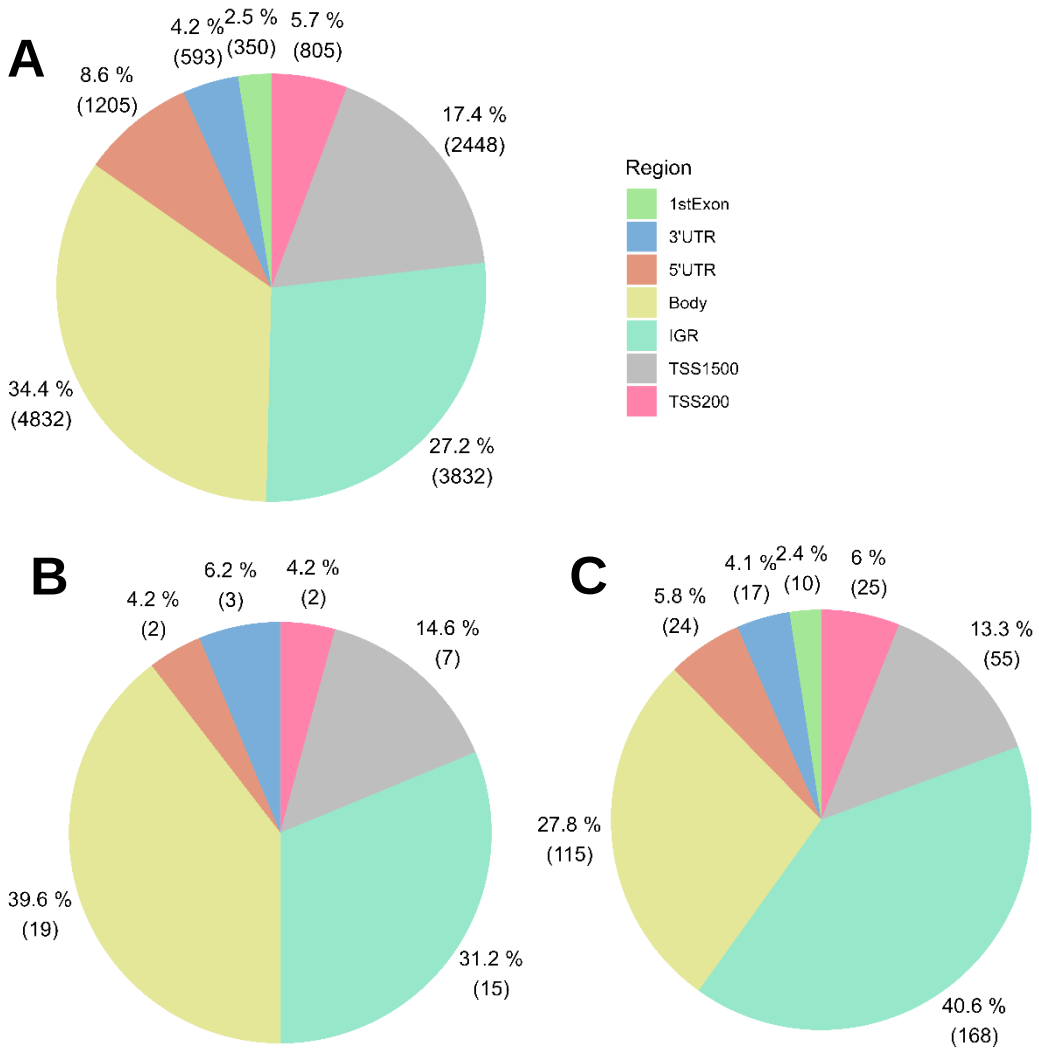

Figura 11. Distribución de los DMP en función de la localización respecto al gen asociado. Se muestran en A) la distribución general de los DMP, en B) la distribución de DMP hipermetilados, y en C) la distribución de DMP hipometilados. Se indican como *1stExon* el primer exón, *Body* el cuerpo del gen, *IGR* las regiones intergénicas, y *TSS1500* y *TSS200* las regiones situadas 1.5 Kpb y 200 pb corriente arriba del sitio de inicio de la transcripción.

Respecto a la segunda clasificación, la mayoría de DMP se encontraron en el cuerpo de genes (34.4 %), junto a un porcentaje relativamente alto situado en regiones TSS1500 y TSS200 (17.4 y 5.7 %, respectivamente. La menor parte de DMP se encontraron en el primer exon, siendo inexistentes en el caso de los DMP hipermetilados (Figura 11B). En el caso de los DMP hipermetilados, el

porcentaje situado en el cuerpo de algún gen fue superior al general (39.9 %), e inferior en el caso de los DMP hipometilados (27.8 %).

4.5 Resultados del Análisis de los Componentes Principales

A partir de los DEG y DMP identificados se llevó a cabo un Análisis de los Componentes Principales (PCA) para cada ómica. Mediante este análisis y a partir de las variables de interés es posible indagar sobre las principales fuentes de variabilidad entre las muestras. El objetivo principal del análisis fue comprobar si los componentes que explican un mayor porcentaje de la variabilidad están relacionados con los fenotipos de estudio (Control y PD), lo que permitiría asumir que la enfermedad es la causa principal la diferencia entre las muestras.

El PCA se llevó a cabo en cuatro bloques: utilizando los 14065 DMP (Figura 12A), utilizando los 462 DMP hiper e hipometilados (Figura 12B), utilizando los 1678 DEG (Figura 13A), y utilizando los 26 DEG sobre e infraexpresados (Figura 13B).

En general se observó cierta separación de las muestras en los dos grupos, Control y PD, a lo largo del primer componente. Sin embargo, esta separación no fue clara en ninguno de los casos, existiendo solapamientos entre los dos grupos a lo largo del primer componente.

Los DMP generales no permitieron agrupar demasiado ninguno de los dos grupos (Figura 12A). Sin embargo, los DMP extremos permitieron un agrupamiento muy homogéneo de las muestras Control, al contrario que el agrupamiento de las muestras PD, mucho más dispersas a lo largo del primer componente (Figura 12B). En ambos casos, el porcentaje de la varianza explicada por el primer componente se encontró en torno al 37 %.

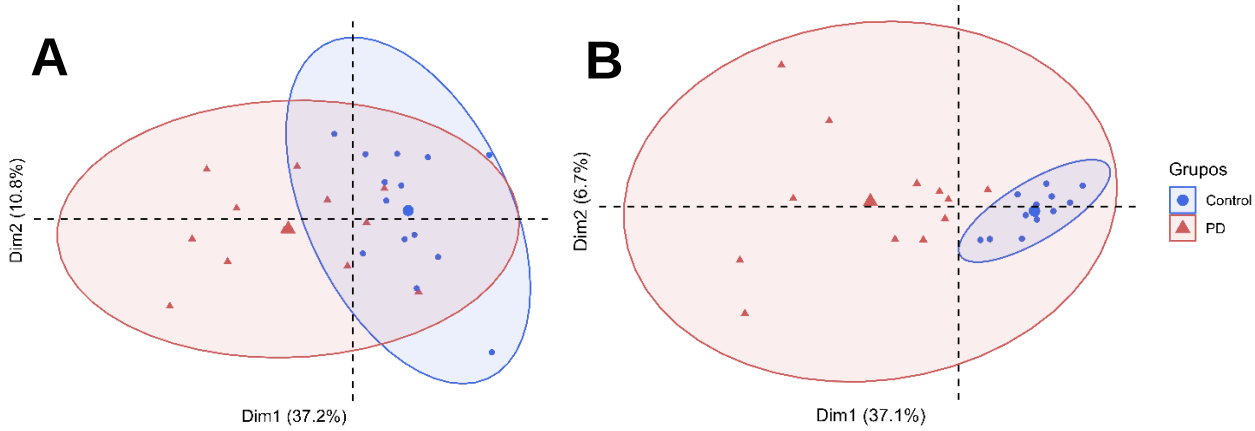

Figura 12. Representación bidimensional de los dos componentes principales obtenidos a partir de los valores beta de A) los DMP identificados y B) los DMP hipermetilados e hipometilados.

El agrupamiento de las muestras en dos clústeres tampoco fue claro al usar los DEG generales (Figura 13A). Sin embargo, la separación fue considerablemente más clara en los gráficos PCA generados a partir de los DEG extremos (Figura 13B). Los DEG extremos permitieron una segregación de las muestras más homogénea, con el menor solapamiento en el gráfico PCA. Además, en general el primer componente de los DEG (~ 62 % y ~ 42 %) explicó un porcentaje de la varianza mayor que el primer componente de los DMP.

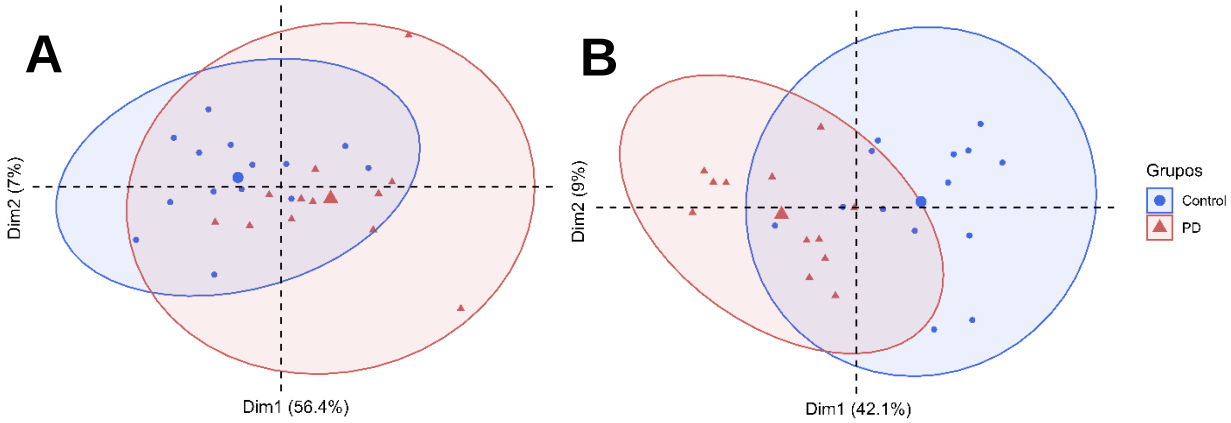

Figura 13. Representación bidimensional de los dos componentes principales obtenidos a partir de los valores beta de A) los DEG identificados y B) los DEG sobreexpresados e infraexpresados.

Respecto al segundo componente obtenido a partir de los datos, este no permitió obtener una separación de las muestras en dos grupos en ningún caso.

4.6 Resultados del enriquecimiento por Ontología génica

A partir de los 1678 DEG identificados se realizó un enriquecimiento por ontología génica para identificar posibles procesos biológicos (Figura 14A y 14B), y componentes celulares (Figura 15A y 15B) potencialmente relacionados con la enfermedad de Parkinson. Aunque las muestras identificadas provinieron de sangre, esta fase permite caracterizar la enfermedad y evaluar la existencia de un paralelismo entre las rutas y estructuras afectadas por la patología en el tejido nervioso y en sangre.

Una vez identificados los diferentes procesos y componentes se generaron redes gen-término para estudiar la posible vinculación entre diferentes términos y los genes comunes (Figura 14B y 15B). En estas redes se destacaron los procesos y componentes de especial interés en relación con la patología de estudio.

Los procesos biológicos identificados estuvieron relacionados principalmente con la modificación postraduccional de las proteínas (Figura 14A), como los clústeres de metilación de lisinas en las histonas y la acetilación de aminoácidos. Entre los distintos procesos se identificaron varios relacionados con la ubiquitinización, destacado en la Figura 14B. También hubo un conjunto de procesos biológicos relacionados con la modificación postranscripcional del ARN, y en concreto del ARN mensajero (Figura 14A).

Respecto a los componentes celulares, estos fueron considerablemente más diversos, agrupados en diferentes clústeres (Figura 15A). Se identificaron componentes relacionados con la mitocondria, destacados en la Figura 15B, el huso mitótico, las ribonucleoproteínas citoplasmáticas, la vacuola lisosomal lítica vacuolar, y el completo histónico proteína acetiltransferasa.

El número de procesos biológicos y componentes celulares significativos fue superior a cien en ambos casos. También se llevó a cabo el enriquecimiento a partir de los 369 DEG con al menos un DMP asociado, aunque no se obtuvieron términos significativamente afectados. Más detalles sobre los procesos biológicos y los componentes celulares identificados pueden consultarse en el Anexo III.

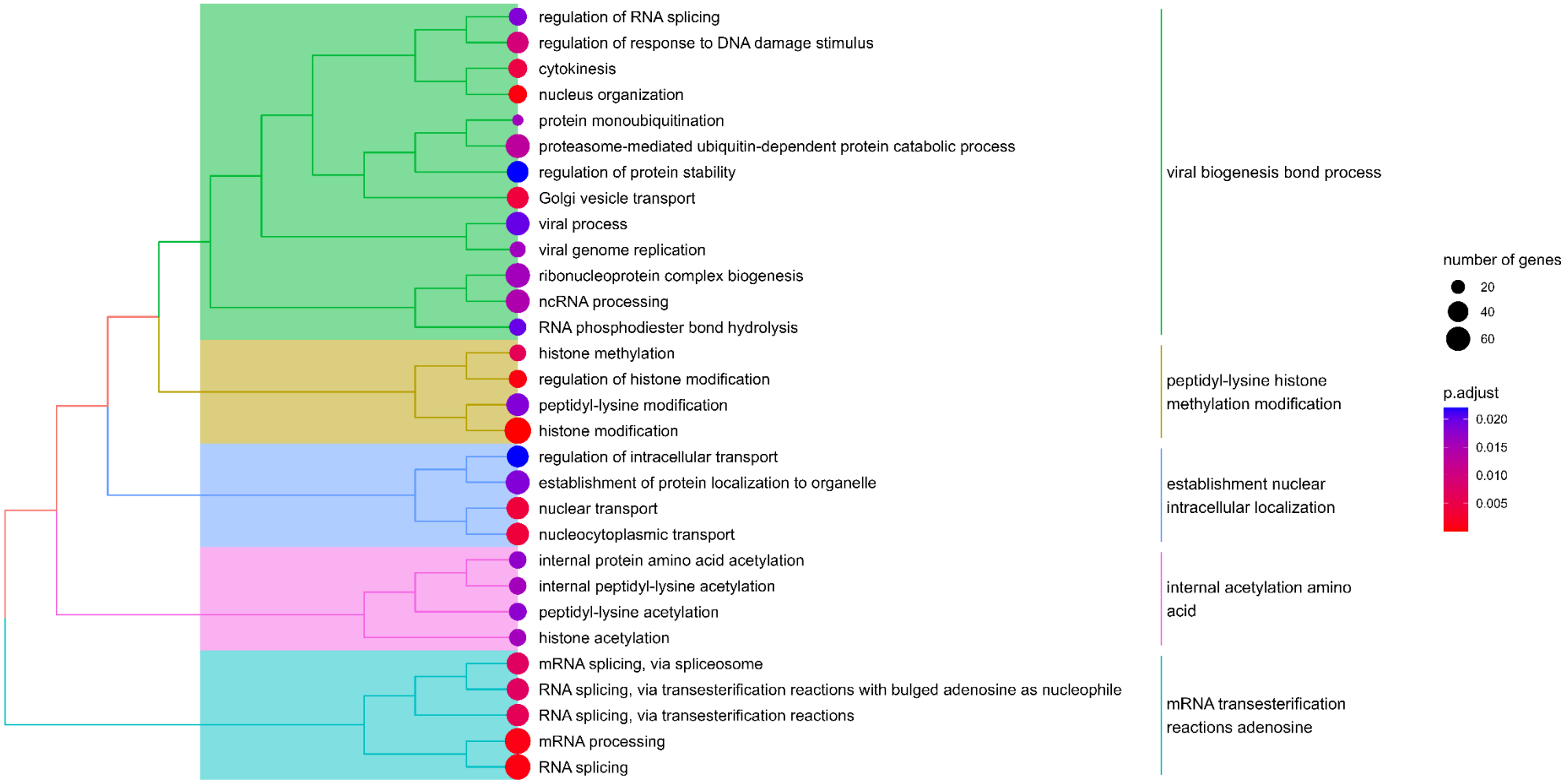

Figura 14A. Gráfico de árbol de los procesos biológicos más significativos y su agrupación en categorías superiores. Se indica el número de genes que contribuyen a cada proceso biológico y el FDR asociado a cada proceso. La agrupación de los procesos se hizo en cinco grupos.

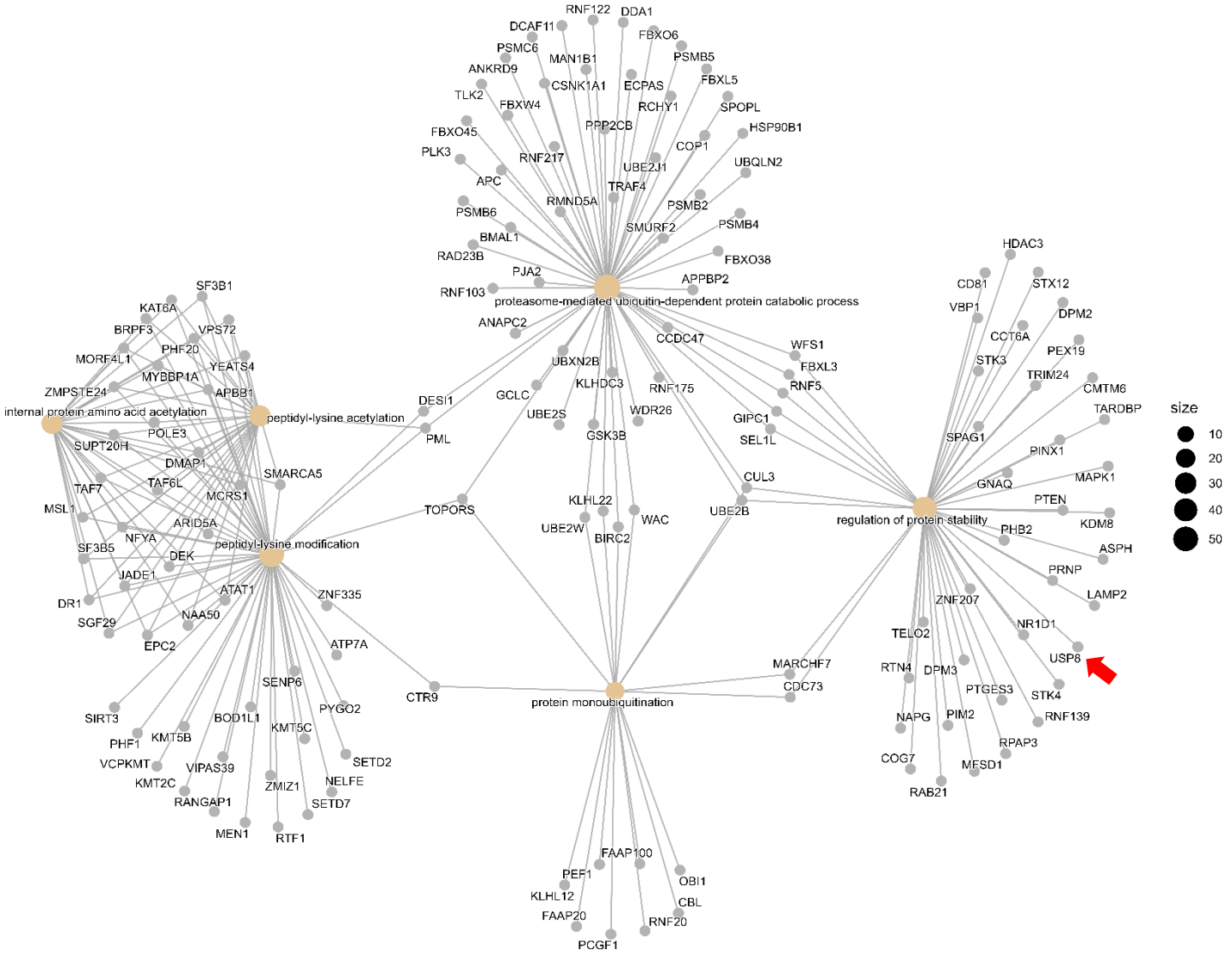

Figura 14B. Gráfico red de genes-procesos biológicos. Se muestran términos específicos relacionados con la ubiquitinización y términos generales relacionados con la modificación proteica. Se indica en rojo el gen USP8, relacionado con el Parkinson (Alexopoulou et al., 2016).

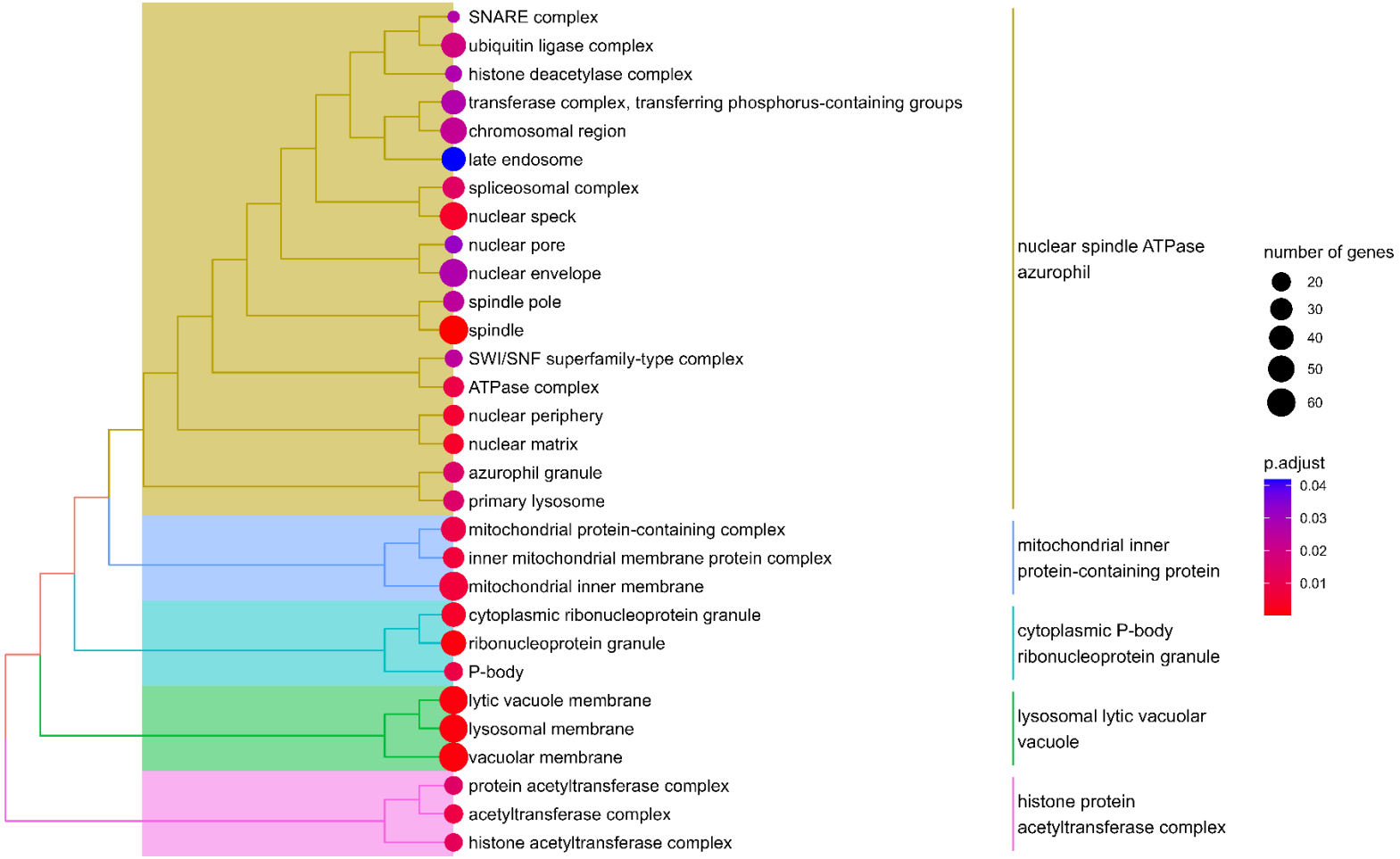

Figura 15A. Gráfico de árbol de los componentes celulares más significativos y su agrupación en categorías superiores. Se indica el número de genes que están asociados a cada componente celular y el FDR asociado a cada proceso. La agrupación de los procesos se hizo en cinco grupos.

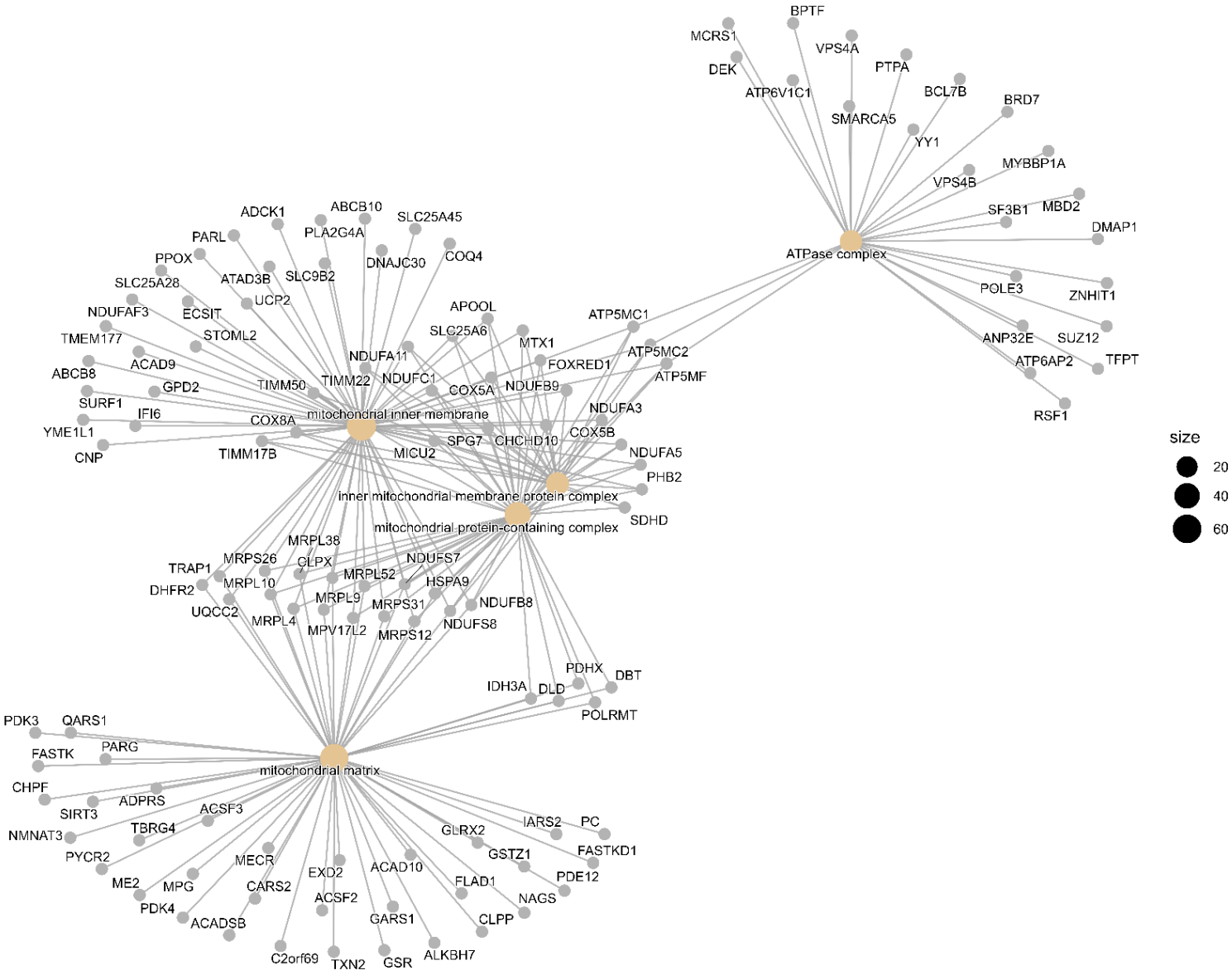

Figura 15B. Gráfico red de genes-componentes celulares. Se muestran términos relacionados con la mitocondria.

4.7 Resultados de la integración ómica

A partir de los DEG y DMP identificados se llevó a cabo la integración ómica con el objetivo de estudiar la interacción entre ambas ómicas y evaluar la capacidad de las variables seleccionadas de actuar como posibles biomarcadores de la enfermedad.

La integración de ambas ómicas se realizó mediante dos vías. Por un lado, se estudió la relación entre DEG y DMP en función de la correlación entre sus valores y la posición de los DMP respecto a los DEG. Se consideraron como eQTM aquellos DMP con un FDR ≤ 0.05. Estos se distribuyeron en dos grupos en función de su localización respecto al gen: cis (cercanos al gen) o trans (alejados del gen). Se identificaron 16 eQTM cis (Tabla 1), cada uno asociado a un gen diferente, y 20921 eQTM trans, distribuidos entre 1294 genes.

Tabla 1. Información sobre los 16 eQTM identificados y los genes asociados.

CpGs - eQTM	Gen	Estadístico T	p-valor	FDR
cg26724450	C8orf33	5.12	~ 0	0.019
cg02406092	FCGBP	4.65	~ 0	0.031
cg05098432	ENTPD1	4.34	~ 0	0.041
cg06034548	MCTP1	-4.20	~ 0	0.041
cg00394658	PTPRJ	-4.17	~ 0	0.041
cg13782866	DISC1	-4.10	~ 0	0.041
cg13728834	TMEM8B	4.02	~ 0	0.041
cg12361262	GGNBP2	-3.99	~ 0	0.041
cg15138109	NCOA1	-3.88	~ 0	0.045
cg11697111	C7orf50	3.80	~ 0	0.045
cg02843500	RUSF1	3.77	~ 0	0.045
cg20483374	CBL	-3.77	~ 0	0.045
cg25417405	USP8	-3.74	0.001	0.045
cg13692972	RTN3	-3.71	0.001	0.045
cg15225325	UIMC1	-3.71	0.001	0.045
cg20212624	MSL1	-3.66	0.001	0.047

Una vez se obtuvieron los DEG, DMP, y eQTM se llevó a cabo una búsqueda bibliográfica con el objetivo de identificar aquellos genes potencialmente relacionados con la enfermedad (Tabla 2). Para ello se tuvieron en cuenta los

DEG, los genes asociados a DMP, los DEG asociados a DMP, y los DEG asociados a eQTM cis.

Tabla 2. Información sobre los 14 genes de interés identificados en los diferentes análisis. Estos están relacionados con la enfermedad de Parkinson o con otras enfermedades neurodegenerativas, el correcto neurodesarrollo o la capacidad motriz.

Categoría	Gen	Nº DMP o Nombre eQTM	p-valor Wilcoxon	FDR Wilcoxon	logFC	p-valor eQTM	FDR eQTM
Genes con mayor número de DMP	PTPRN2[1]	30	-	-	-	-	-
	MAD1L1[1]	25	-	-	-	-	-
Genes con mayor número de DMP hipometilados	DDR1[1]	7	-	-	-	-	-
	C1orf65[2]	6	-	-	-	-	-
Genes con mayor número de DMP hipermetilados	MDGA1[2]	3	-	-	-	-	-
DEG infraexpresados	S100B[1]	-	0.027	0.339	- 1.48	-	-
	ISG15[1]	-	0.001	0.339	- 1.364	-	-
DEG con al menos un DMP	CYB561[1]	2	0.002	0.339	- 0.46	-	-
	CTBP2[1]	6	0.001	0.339	0.49	-	-
	CACNB4[2]	4	0.002	0.339	1.10	-	-
DEG con eQTM cis	FCGBP[1]	cg02406092	0.003	0.339	- 0.647	0.001	0.339
	ENTPD1[2]	cg05098432	0.031	0.339	0.249	0.002	0.339
	MCTP1[2]	cg06034548	0.036	0.339	0.442	0.002	0.339
	USP8[1]	cg25417405	0.015	0.339	0.432	0.001	0.339

[1] Genes directamente vinculados a la enfermedad de Parkinson.

[2] Genes relacionados con otras enfermedades neurodegenerativas, el correcto neurodesarrollo o la capacidad motriz. Las columnas p-valor Wilcoxon, FDR Wilcoxon, logFC se obtuvieron en la fase de análisis de expresión diferencial. Las columnas p-valor eQTM y FDR eQTM se obtuvieron en la fase de análisis de eQTM.

Por otro lado, se llevó a cabo una integración vertical entre ambas ómicas mediante el método DIABLO. Mediante esta técnica es posible estudiar la correlación entre las distintas variables de cada ómica, en este caso genes diferencialmente expresados y CpG diferencialmente metiladas, así como su correlación con el primer y el segundo componente de cada ómica.

Previo al análisis, se estudió la correlación entre ambas ómicas a partir de sus primeros componentes principales obtenidos tras un análisis PLS. El valor de correlación obtenido fue de 0.69, considerablemente alto. A continuación, se estudió la selección del número de variables óptimas para generar el modelo.

Esto se hizo mediante el entrenamiento de un modelo de selección de variables (Figura 16).

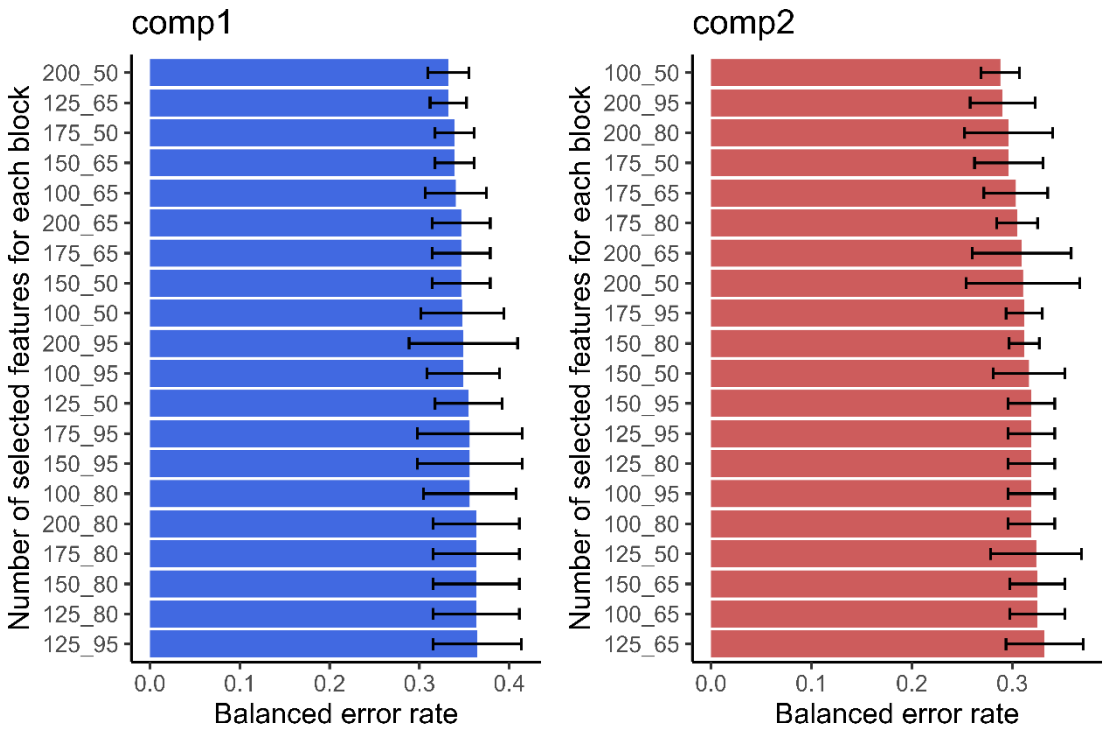

Figura 16. Tasa de error asociada a cada componente a la hora de clasificar cada una de las 26 muestras en función del número de variables seleccionadas. Se indica en el eje X la tasa de error y en el eje Y el número de variables utilizadas, el primer número se corresponde con el número de DMP seleccionado, y el segundo número se corresponde con el número de DEG seleccionado.

Se puede ver cómo el modelo clasificador que cometió un menor número de errores fue aquel que seleccionó 200 DMP y 50 DEG para el primer componente principal de cada ómica, y 100 DMP y 50 DEG para el segundo componente principal de cada ómica (Figura 16).

Una vez establecido el número de variables para cada componente se llevó a cabo el análisis PLS-DA por bloques. A partir del análisis realizado se generaron gráficos con datos de correlación entre los componentes principales de las ómicas analizadas (Figura 17A y 17B) y con la distribución de las variables a lo largo de los dos componentes principales (Figura 18).

En general se observó una correlación muy alta entre los primeros y segundos componentes de cada ómica, con valores de 0.92 y 0.89, respectivamente (Figura 17). Además, se pudo apreciar que la separación de las muestras a lo largo del primer componente de ambas ómicas permitió distinguir mejor los grupos de estudio que la separación de las muestras a lo largo del segundo componente.

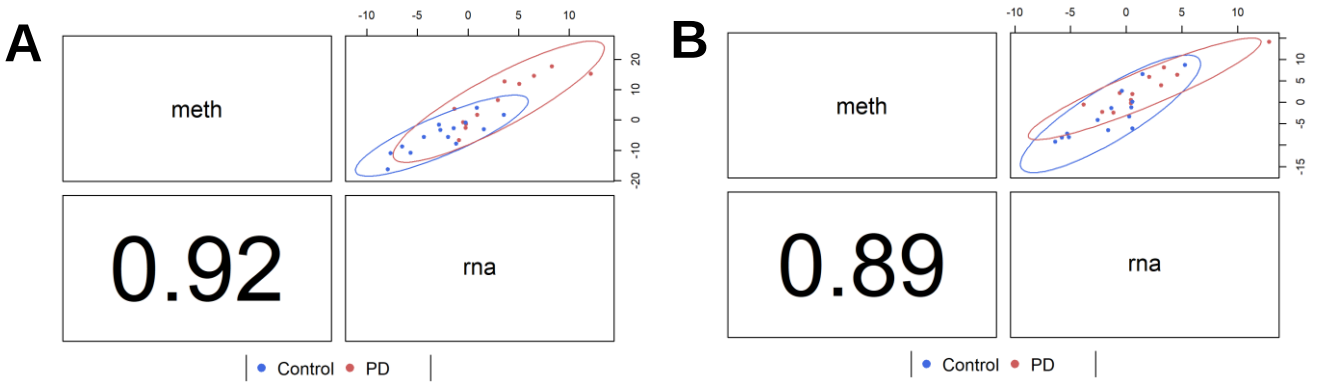

Figura 17. Multigráfico con información sobre la correlación entre los componentes principales de cada ómica. En el marco superior-derecho se muestra el gráfico de los valores de lo componentes enfrentados, en el marco inferior-izquierdo se indica la correlación entre los componentes de ambas ómicas. *Meth* = datos de metilación (DMP), *rna* = datos de transcripción (DEG). Se indica en A) correlación entre el primer componente principal de ambas ómicas, y B) correlación entre el segundo componente principal de ambas ómicas.

Respecto a la relación de las variables con cada uno de los componentes y entre sí, se observaron varios clústeres de correlación situados en la corona del gráfico de correlación (Figura 18). Estas variables se pueden considerar como altamente correlacionadas con uno o con los dos componentes. Se distinguió un gran clúster DEG y DMP correlacionados negativamente con el primer componente, y un segundo gran clúster de, principalmente, DMP correlacionados positivamente con el primer componente. También se observó un tercer clúster de DEG y DMP correlacionados negativamente con el segundo componente.

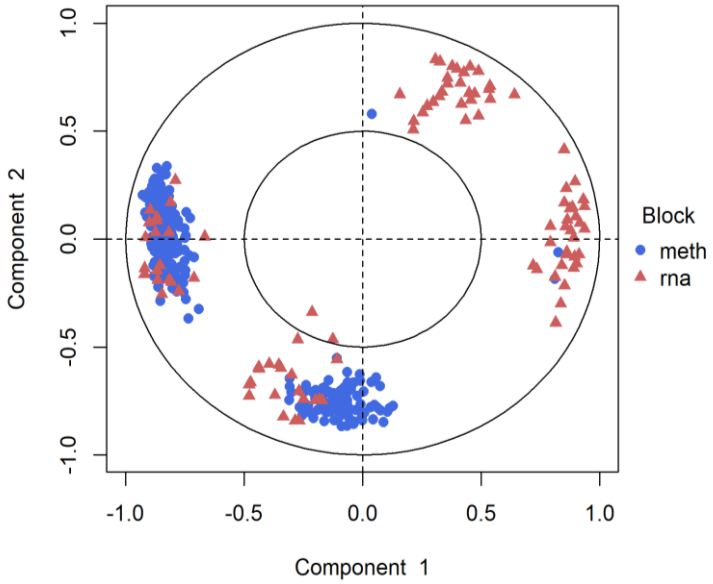

Figura 18. Distribución de las variables de ambas ómicas según su correlación con el primer y el segundo componente. Las coordenadas de cada variable se obtuvieron a parir de la correlación con cada uno de los dos componentes. *Meth* = datos de metilación (DMP), *rna* = datos de transcripción (DEG).

A partir del modelo generado mediante integración DIABLO se comprobó cuáles fueron las variables que más contribuyeron a separar ambos grupos a lo largo del primer componente (Figura 19). Por un lado, los veinte DMP más contributivos para el primer componente tuvieron asociados valores de contribución negativos, indicativo de una mejor discriminación del grupo Control. Por otro lado, los veinte DEG más contributivos no mostraron una discriminación dirigida hacia uno de los grupos; en su lugar, algunos DEG permitieron discriminar mejor el grupo Control, con valores de contribución negativos para el primer componente, mientras que otros lograron una discriminación mejor del grupo PD, con valores de contribución positivos.

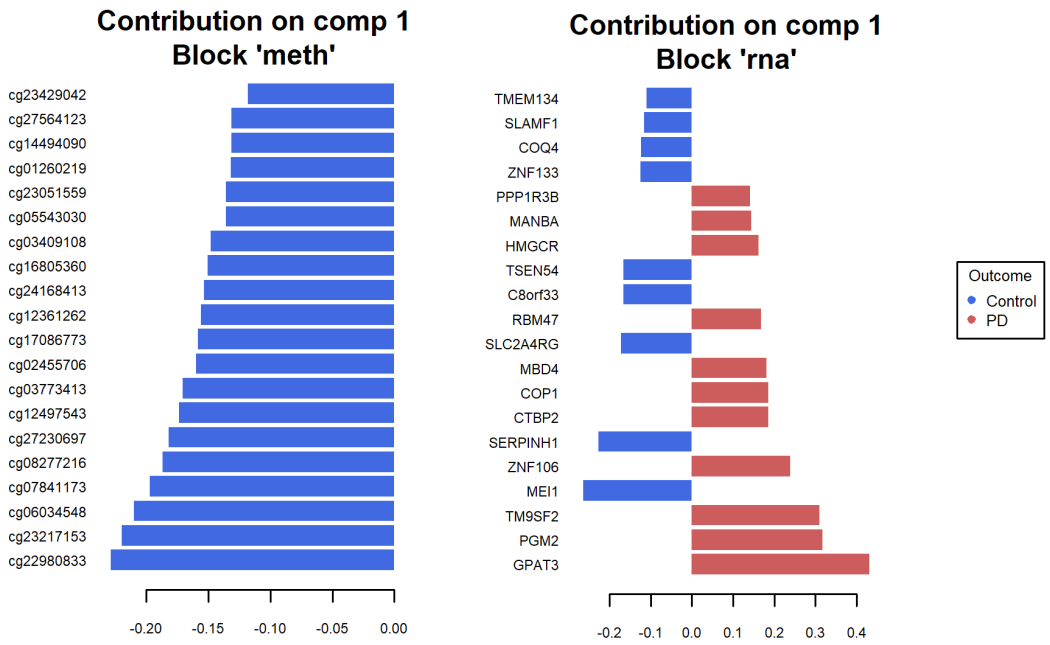

Figura 19. Valores de contribución para el primer componente de las veinte variables más contributivas. Se diferencia la contribución de los DMP y de los DEG . *Meth* = datos de metilación (DMP), *rna* = datos de transcripción (DEG).

Por último, a partir de las variables seleccionadas en el modelo de integración se generó un *heatmap* con información sobre las variables, las muestras, además de una clusterización no supervisada de las muestras (Figura 19). Se distinguió un grupo principalmente conformado por muestras del grupo PD y otro grupo principalmente conformado por muestras Control. Dentro del grupo mayoritariamente PD se identificó un clúster PD con un perfil muy similar. A su vez, dentro del grupo mayoritariamente Control se identificó un clúster Control con un perfil bastante similar. Al margen de los subclústeres bien definidos, en ambos grupos se observaron varias muestras erróneamente agrupadas. Esta son las muestras 051_PD, 012_PD, 049_PD, agrupadas junto a muestras Control, y las muestras 040_C y 047_C, agrupadas junto a muestras PD. Destaca además la muestra 021_C, con un perfil de metilación acorde al grupo Control pero marcadamente extremo.

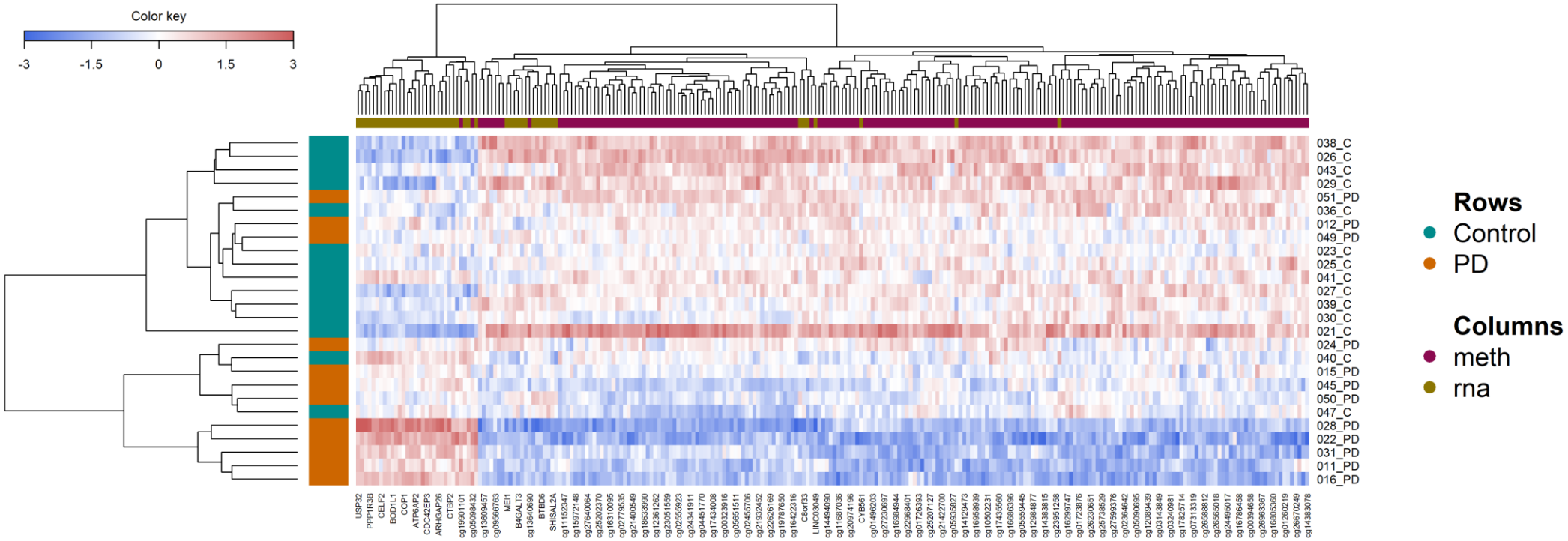

Figura 20. *Heatmap* generado con las variables seleccionadas por el método sPLS. Las muestras se clusterizaron de forma no supervisada a partir de las variables. *Meth* = datos de metilación (DMP), *rna* = datos de transcripción (DEG).

5 Discusión

A lo largo de las diferentes fases de este proyecto se han llevado a cabo diversos análisis y estudios con el objetivo de determinar si existen diferencias significativas a nivel de expresión y metilación del genoma en sangre entre pacientes con la enfermedad de Parkinson y pacientes sanos. La búsqueda de estas diferencias se hizo, por un lado, con el propósito principal de identificar posibles biomarcadores de la enfermedad en sangre y que faciliten su diagnóstico, y, por otro lado, con el segundo propósito de caracterizar la enfermedad a nivel de procesos y estructuras celulares afectadas, y ver si los procesos involucrados afectados en sangre pueden considerarse un reflejo de los procesos involucrados en el tejido nervioso.

Los estudios de la expresión y metilación diferencial entre los grupos se realizaron utilizando diferentes análisis estadísticos mediante los cuales identificar DEG y DMP. En general, se pudo observar que las diferencias entre ambos grupos no fueron lo suficientemente marcadas como para poder considerarse estadísticamente significativas a un FDR umbral de 0.05. Por ello, la selección de los DEG y DMP se hizo teniendo en cuenta un p-valor umbral no ajustado de 0.05. De este modo, lo que se consideró fue la tendencia de los genes a estar diferencialmente expresados y de las posiciones CpG a estar diferencialmente metiladas (p-valor ≤ 0.05, FDR > 0.05).

La ausencia de diferencias marcadas puede observarse en los mapas de calor (Figuras 7 y 8). En ambos casos se observaron perfiles de color ligeramente diferenciados pero tenues; estas diferencias se ven acentuadas en el caso de los DEG y DMP extremos, en los que los perfiles son algo más evidentes. Además, en algunos casos se observaron perfiles de expresión invertidos entre muestras del mismo grupo (Figura 8A). La explicación más lógica para la ausencia de marcas diferenciadas entre los dos grupos es que el tejido sanguíneo no es un buen predictor del tejido nervioso, de modo que las marcas que podrían observarse en las células corticales no tienen por qué poder observarse en sangre. Estudios previos han reportado la baja precisión de la sangre a la hora de reflejar el estado del tejido nervioso, tanto en pacientes sanos

como en pacientes con esquizofrenia (Hannon *et al*., 2015; Walton *et al*., 2016). Aunque parte de las marcas observadas en tejido nervioso pueden prevalecer en sangre, en general estas se ven considerablemente reducidas. Sin embargo, no por ello la sangre pierde valor a la hora de servir para identificar enfermedades que afectan al tejido nervioso. De hecho, una vía de estudio de interés consiste en identificar biomarcadores correlacionados entre el tejido sanguíneo y el tejido nervioso, para posteriormente estudiar su capacidad como predictor de la enfermedad (Hannon *et al*., 2015).

A la hora de realizar el análisis de los gráficos PCA, inicialmente se esperaría una buena clusterización de las muestras utilizando variables previamente seleccionadas en base a diferentes criterios estadísticos de interés, como el caso de los DEG y DMP. Sin embargo, las muestras no se separaron homogéneamente en los dos grupos esperados a lo largo del eje del primer componente (Figuras 12 y 13), aunque sí es posible apreciar una tendencia de separación entre los dos grupos. Al comparar los gráficos PCA de ambas ómicas se observó una mejor distribución de las muestras según los DEG extremos, lo que sería indicativo de que éstos explican un mayor porcentaje de la varianza entre los dos grupos que los DMP extremos.

Respecto a la distribución de los DMP identificados, estos se situaron principalmente en el cuerpo de genes y en islas y regiones *shore*. Casi el 40 % se encontró en cuerpos génicos, según la localización respecto al gen asociado (Figura 11). Es sabido que la metilación en los cuerpos de los genes está en muchos casos asociada a un incremento de la expresión del gen (Wolf *et al*., 1984), algo que se ha podido corroborar en estudios llevados a cabo tanto en el cromosoma X humano (Hellman & Chess, 2007), como en modelos animales y vegetales (Feng *et al*., 2010). Sin embargo, ninguno de los genes sobreexpresados mostró DMP hipermetilados asociados. El porcentaje de hipermetilación en los cuerpos génicos podría estar relacionado con modificaciones en la regulación de la expresión de los genes, sin que esto lleve a una sobreexpresión del gen detectable, aunque no por ello inexistente. De hecho, el número de DEG sobreexpresados, independientemente del p-valor asociado, fue de 899, algo superior al número de DEG infraexpresados, 779.

La potencial sobreexpresión de los genes inducida por la hipermetilación de los cuerpos génicos es algo que, además, se ve apoyado por el porcentaje de DMP situados en regiones de islas CpG y *shores* (Figura 10), considerablemente elevado. Por un lado, es interesante observar cómo el porcentaje de DMP hipometilados situados en islas es cercano al 37 %, superior al porcentaje de DMP hipermetilados situados en estas regiones, del 22 %. Las islas CpG son áreas con una gran densidad de CpGs y que típicamente se encuentran en promotores, o en áreas cercanas al inicio de la transcripción en vertebrados (Saxonov *et al*., 2006). Destaca el hecho de que la hipometilación de los promotores de los genes es algo tradicionalmente asociado a un incremento de la expresión de los genes (Bird, 2002; Irizarry *et al*., 2009). Por otro lado, se ha podido comprobar que la mayoría de las diferencias de metilación entre tejidos y entre células normales y cancerosas se producen en CpGs situados en las regiones *shore* (Irizarry *et al*., 2009). De un modo similar, la tendencia a una metilación diferencial observada en las regiones *shore* podría estar relacionada con el padecimiento de la enfermedad de Parkinson, pudiendo tener un efecto en la regulación de la expresión de los genes.

Por tanto, es posible discernir lo que parece ser una diferencia en la metilación de regiones reguladoras entre pacientes Control y PD. Esta regulación, aunque sutil, tiende a dirigirse hacia la hipometilación de islas y *shores* de CpGs y la hipermetilación de los cuerpos génicos, aunque sin llegar a implicar una sobreexpresión génica significativa, como cabría esperar.

Una vez se identificaron los DEG, DMP , eQTL y sus genes asociados se indagó par aaveriguar qué genes guardan relación con el Parkinsonismo, con otras enfermedades neurodegenerativas, con el neurodesarrollo o con la capacidad motriz. En este proyecto, a la hora de determinar qué genes, o CpGs, sirven como potenciales biomarcadores de una enfermedad, un criterio de interés que se tuvo en cuenta fue la vinculación con la enfermedad. La idea detrás de esta búsqueda es que un gen diferencialmente expresado entre el grupo Control y el grupo PD que además esté relacionado con un proceso nervioso en las neuronas es más fiable como biomarcador de una patología que un gen diferencialmente expresado pero sin vinculación con la enfermedad.

Se identificaron 14 genes de interés relacionados directamente con la enfermedad de Parkinson, con el correcto neurodesarrollo y el desarrollo motriz, o con el Alzheimer, otra enfermedad de carácter neurodegenerativo (Tabla 2). La mayoría de estos genes desempeñan su función en el tejido neuronal, y no en la sangre. Sin embargo, el haberlos identificado en este biofluido podría implicar que los cambios que se producen en el tejido neuronal debidos al padecimiento de Parkinsonismo se ven reflejados en el tejido sanguíneo.

Los 14 genes identificados fueron DEG, genes con un gran número de DMP asociados, y DEG con un gran número de DMP o eQTL asociados. En base al número de DMP asociados se identificaron varios genes directamente relacionados con la enfermedad de Parkinson: DDR1, entre los genes con un mayor número de DMP hipometilados, y PTPRN2 y MAD1L1, entre los genes con un mayor número de DMP (hiper e hipometilados) (Tabla 2).

Lo sobreexpresión del gen DDR1 en pacientes diagnosticados con la enfermedad de Alzheimer y Parkinsonismo ha sido previamente estudiado por Zhu *et al*. (2015), aunque la forma en la que se relaciona con el proceso de neurodegeneración es mayoritariamente desconocida. Entre los procesos y funciones vinculados a DDR1 destaca la modulación de la actividad de la microglía y de la matriz de metaloproteasas, un conjunto de enzimas endopeptidasas con función degradativa de proteínas de la matriz extracelular. Se ha podido comprobar como en algunas situaciones la sobreexpresión de DDR1 induce la secreción de la metaloproteinasa MMP-9, conllevando una degradación de la matriz extracelular y dañando la barrera hematoencefálica (Zhu *et al*., 2015). Además, recientemente se ha podido comprobar que los genes de la familia DDR (DDR1 y DDR2) se encuentran sobreexpresados en la vía nicroestriada, región del sistema nervioso central que alberga en torno al 80 % de la dopamina cerebral (Flavio, 2017), y en el hipocampo de individuos diagnosticados con la enfermedad de Parkinson (Hebron *et al*., 2017). En este trabajo se detectó que el gen DDR1 tuvo 7 DMP hipometilados y 22 DMP generales asociados, siendo uno de los genes con más DMP; sin embargo, en no se detectó una expresión diferencial entre ambos grupos. Dado su carácter asociado a procesos degradativos de proteínas en células nerviosas, y por ende su vinculación directa con la enfermedad de Parkinson, estudiar la metilación del

gen DDR1 y su consecuente regulación es de especial interés. Además, los resultados aquí observados parecen indicar que el gen DDR1 podría estar también bastante regulado en sangre, lo que lo convierte en un potencial gen de interés como biomarcador, con múltiples posiciones CpG que pueden estudiarse.

Respecto al gen PTPRN2, codificante del receptor proteico tipo N2 de la tirosina fosfatasa, este mostró 30 DMP asociados, siendo uno de ellos un DMP extremamente hipometilado. Estudios recientes lo reportaron como un gen con grandes diferencias de metilación, no solo en función del padecimiento de Parkinsonismo, sino también del género del individuo (Kochmanski *et al*., 2022). El estudio de Kochmanski *et al*. (2022) reportó que los varones mostraron signos de hipermetilación en PTPRN2 en el cerebro respecto a los controles, mientras que las mujeres mostraron signos de hipometilación. Estudios anteriores ya lo identificaron como un marcador genético diferencialmente metilado asociado al deterioro cognitivo y al deterioro de la función motora en pacientes PD (Chuang *et al*, 2019). Además, la deleción de PTPRN2 en ratones se ha visto asociada a una reducción en la secreción de diversos neurotransmisores, entre ellos la dopamina, directamente relacionada con el Parkinsonismo (Nishimura *et al*, 2009). Estos resultados coinciden con lo observado en este proyecto, y son indicativos de que el grado de metilación de este gen puede funcionar como biomarcador del padecimiento de la enfermedad de Parkinson, con diferencias en función del género.

Respecto al gen MAD1L1, con 25 DMP asociados, éste codifica la proteína MAD1, con función de control del ensamblaje del huso mitótico. Aunque el gen ha sido previamente relacionado con el padecimiento de Parkinsonismo según variantes identificadas (Guo *et al*., 2018), se desconocen las vías por las cuáles el gen puede vincularse con la enfermedad. Sin embargo, la importancia de los microtúbulos en el padecimiento de Parkinson es algo que sí ha sido estudiado previamente, y su relación con la alfa-sinucleína. Los cúmulos de esta proteína pueden interaccionar con los microtúbulos y estructuras derivadas, afectando a su estabilidad (Prots *et al*., 2013).

También se identificaron otros genes de interés, aunque no directamente relacionados con Parkinson. Éstos fueron el gen C1orf65, con un gran número de DMP hipometilados, y el gen MDGA1, con un gran número de DMP hipermetilados. Por un lado, estudios recientes han identificado regiones diferencialmente metiladas en el promotor del gen C1orf65 entre muestras de sangre de individuos sangos y enfermos de Alzheimer (Silva *et al*., 2022). Por otro lado, el gen MDGA1 cumple una función de control en el hipocampo, regulando negativamente la inhibición de la sinapsis mediada por la proteína precursora amiloide (Kim *et al*., 2022). En ambos casos, aunque no exista una relación directa con el Parkinson, sí existe un paralelismo. Aunque los DMP de ambos genes puedan servir como marcadores de metilación, sería necesario realizar más estudios de metilación diferencial asociada explícitamente a Parkinson.

Por último, se identificó el gen PRKN, también llamado PARK2, con un único DMP asociado. El producto del gen PRKN cumple con funciones reguladoras en el proceso de ubiquitinización, y fue uno de los primeros genes en ser etiquetado como causante del Parkinsonismo (Kitada *et al*., 1968). En ninguno de estos casos los genes estuvieron diferencialmente expresados entre los grupos.

Respecto a los DEG, se identificaron dos genes altamente infraexpresados en PD respecto a Control y relacionados directamente con la enfermedad de Parkinson. Estos fueron S100B e ISG15 (Figura 9B).

El gen S100B, expresado principalmente en los astrocitos, codifica para una proteína de unión a calcio. Entre las funciones que desempeña se encuentra la homeostasis de calcio, el metabolismo energético y la regulación y movilidad del citoesqueleto (Donato, 2001). Estudios *post-mortem* realizados en el encéfalo han revelado un incremento en los niveles de S100B en la sustancia negra de los pacientes PD, entre otras áreas (Sathe *et al*., 20012). La contradicción entre este resultado y los resultados obtenidos en este trabajo parecerían indicar una inversión en la expresión del gen en el tejido nervioso frente al tejido sanguíneo. El otro DEG infraexpresado en sangre fue ISG15. Este gen actúa en las células de la sustancia negra mediante un proceso similar a la ubiquitinización, denominado comúnmente como *ISGylation* (Haas *et al*., 1987). Su producto

peptídico es capaz de modificar la estructura de las proteínas a las que se asocia de forma covalente, lo que las lleva a ser degradadas o a ver alterada su función. Precisamente una de las proteínas diana del producto de ISG15 es la proteína parkina, producto del gen PRKN. Mediante la unión con la parkina, ISG15 es capaz de inducir su actividad ubiquitinizadora (Im *et al*., 2016), lo que desemboca en la degradación asociada a ubiquitinización de la proteína TRAF6 (Henn *et al*., 2007), un factor de necrosis tumoral. Una reducción en la expresión de ISG15 podría conllevar una reducción de la actividad de la parkina, provocando en definitiva una acumulación de sus sustratos y pudiendo conllevar, en definitiva, la muerte celular en el tejido nervioso. Aunque no se puede asumir que toda esta concatenación de efectos suceda en sangre, es plausible pensar que, de nuevo, lo observado en sangre es un reflejo de las rutas génicas afectadas en las células nerviosas.

También se identificaron dos DEG con al menos un DMP asociado y relacionados directamente con el Parkinson, CYB561 y CTBP2, y un DEG con al menos un DMP e implicado en el correcto neurodesarrollo y la aparición de problemas motrices, CACNB4. El gen CYB561, codificante del citocromo b561 y con función homeostática, ha sido previamente señalado por su significancia en estudios de redes de co-expresión a partir de muestras de sangre de PD, destacando su carácter como un posible biomarcador (Chatterjee *et al.*, 2017). Sin embargo, se desconoce su posible rol en la enfermedad. Respecto al gen CTBP2, los genes de la familia CTBP codifican proteínas represoras de la transcripción. Mediante la represión de la expresión de genes pro-apoptóticos estas proteínas son capaces de promover la supervivencia de las neuronas y los astrocitos, entre otras. Sin embargo, la sobreexpresión de CTBP2 en las células corticales provoca alteraciones en el correcto neurodesarrollo (Wang *et al*., 2019). Estudios recientes con modelos de ratones comprobaron que el gen CTBP2 está sobreexpresado en la sustancia negra en ratones ancianos modelo de PD respecto a los ratones jóvenes (Saraiva *et al*, 2023). Estos resultados coinciden con lo observado, una sobreexpresión de CTBP2 en muestras de sangre de pacientes PD, lo que podría indicar un reflejo parcial en sangre de las vías génicas afectadas en las células corticales. Por último, recientemente se comprobó que el gen CACNB4, codificante de la subunidad β4 de los canales de

calcio tipo P/Q, juega un papel importante en el neurodesarrollo; variantes sin sentido en este gen implican afecciones graves y alteraciones de las funciones desempeñadas por la proteína (Coste de Bagneaux *et al*., 2020).

Además, mediante el análisis de eQTM se identificaron 16 genes con al menos un DMP asociado como eQTM (Tabla 1). Estos 16 genes están vinculados con el correcto neurodesarrollo, con enfermedades neurodegenerativas, o con procesos o áreas específicas del encéfalo, como es el caso de los genes ENTPD1 (Calame *et al*., 2022; Mamelona *et al*., 2019) y MCTP1, siendo este último codificador de un circRNA relacionado con el desarrollo de MSA (del inglés, *Multiple System Atrophy*) (Chen *et al*., 2016), una afección degenerativa neurológica rara que afecta a las funciones motrices normales del cuerpo, provocando movimientos involuntarios, y en muchos casos afectando a funciones internas, como la presión sanguínea. De entre los DEG con eQTM fue de especial interés el gen USP8 dada su relación con el proceso de ubiquitinización (Figura 15B). Su producto cumple una función desubiquitinizadora de la proteína alfa-sinucleína, evitando su degradación (Alexopoulou *et al*., 2016). Las estructuras derivadas de la acumulación anormal de la proteína alfa-sinucleína en las neuronas de la sustancia negra se denominan cuerpos de Lewy, comúnmente asociados a la enfermedad de Parkinson (Polymeropoulos *et al*., 1996; Polymeropoulos *et al*., 1997). En este trabajo se observó una sobreexpresión de USP8 en muestras de sangre PD respecto a Control. De ser un reflejo de la regulación en las células neuronales, este resultado podría ser indicativo de un exceso de desubiquitinización de la proteína alfa-sinucleína, lo que conllevaría su acumulación en las neuronas y la consecuente aparición de cuerpos de Lewy.

Mediante la identificación de diferentes genes relacionados con la enfermedad de Parkinson se comprobó que la regulación de la ubiquitinización proteica en las neuronas y otras células neuronales es un proceso clave en la aparición del Parkinsonismo. Este mecanismo de marcaje también se vio reflejado en tres términos obtenidos tras el enriquecimiento por ontología génica. Éstos fueron el "procesado catabólico de proteínas mediado por proteasoma dependiente de ubiquitina" (*proteasome-mediated ubiquitin-dependent protein catabolic process,* GO:0043161), la monoubiquitinización de proteínas (GO:0006513) (Figura 14A),

y el complejo ubiquitina ligasa (GO:0000151). Además, también destacaron otros procesos relacionados con la modificación y la regulación de la estabilidad proteica (Figura 14B), con varios genes en común con los procesos de ubiquitinización. Estos resultados son indicativos de que una parte de los genes diferencialmente expresados están relacionados con vías de regulación y degradación proteica, algo que, como se ha comentado anteriormente, está vinculado a la enfermedad de Parkinson.

Respecto a los componentes celulares, destacó un bloque conformado por estructuras mitocondriales: complejo proteico de la membrana interna mitocondrial (GO:0098800), membrana interna mitocondrial (GO:0005743), complejo mitocondrial proteico (GO:0098798), matriz mitocondrial (GO:0005759) y complejo ATPasa (GO:1904949). Sin embargo, hay que tener en cuenta que estos componentes, salvo por el complejo ATPasa, comparten múltiples de los genes contribuyentes (Figura 16B).

Aunque la teoría del incorrecto funcionamiento de la ubiquitinización en las células neuronales y la consecuente acumulación de proteínas es una de las más plausibles y más probada a la hora de caracterizar la enfermedad, existen otras teorías que explican la maquinaria molecular que hay tras la enfermedad. Entre ellas se encuentra la teoría del estrés oxidativo. Un error en el funcionamiento de la cadena de transferencia electrónica mitocondrial puede llevar a una reducción en la concentración de ATP. Puesto que la ubiquitinización y la sucesiva degradación proteica mediada por proteasoma es un proceso que requiere de un gran aporte energético, una falta de ATP puede llevar a un mal funcionamiento del proceso, con la consecuente acumulación anormal de proteínas (Huang *et al*., 2013). Además, se ha comprobado que la exposición al estrés oxidativo en el fluido cerebroespinal puede conllevar el malfuncionamiento de los procesos celulares y un desbalance energético. Éstos pueden acabar por derivar en daño oxidativo mitocondrial y del ADN asociados con la enfermedad de Parkinson (Isobe *et al*., 2010).

En general, los resultados obtenidos parecen indicar que existe una alteración de parte de la maquinaria celular consecuencia de la tendencia a una expresión y metilación diferencial de los genes y sitios CpG. Sin embargo, a la hora de

determinar la utilidad de estos genes y CpGs como potenciales biomarcadores del Parkinsonismo es necesario recurrir a otras técnicas.

En este trabajo se optó por una vía de estudio integrada entre ambas ómicas a través de dos fases de selección. En primer lugar, y como se ha ido comentando en apartados anteriores, se seleccionaron variables con una tendencia a mostrar valores diferentes entre ambos grupos, aunque sin poder considerarlas como estadísticamente diferentes. Posteriormente, se realizó una segunda selección de los DMP y DEG que permiten una mejor clasificación de las muestras mediante algoritmos de clasificación (Figura 16). Seguidamente, mediante la integración de ambas ómicas por el método DIABLO se estudió la correlación entre las variables y los componentes principales.

La alta correlación observada entre los primeros componentes de ambas ómicas sería indicativa de que las variables seleccionadas de ambas ómicas logran una distribución muy similar de las muestras a lo largo del primer componente. Además, y como se comentó previamente, el primer componente parece explicar las principales diferencias debidas a la enfermedad de Parkinson. Respecto a las variables predictoras, los resultados observados tras la integración fueron indicativos de que las variables que mejor parecen separar ambos grupos (clústeres de variables correlacionados con el primer componente en la Figura 18) no logran una separación nítida; en su lugar lo que se observa es una “tendencia a una separación”. En general, se pudo comprobar que los DMP más contributivos para el primer componente lograron distinguir mejor las muestras del grupo Control que las del grupo PD (Figura 19), algo que ya se comprobó mediante el gráfico PCA realizado previamente utilizando los DMP más extremos (Figura 12B). Una teoría posible sobre la viabilidad de estos DMP como potenciales biomarcadores es que podrían servir para detectar específicamente casos Control, pero no casos PD. Por tanto, estos DMP podrían servir para descartar la existencia de Parkinson, pero no para confirmar su existencia. Respecto a los DEG más contributivos identificados, hubo genes que permitieron detectar más específicamente las muestras del grupo Control, mientras que otros permitieron detectar más específicamente las muestras PD. En general, se puede deducir que el primer componente parece discriminar algo mejor al grupo Control que al grupo PD. Sin embargo, al combinar las mejores

variables predictoras de cada ómica en un modelo de clusterización jerárquica no supervisada se logró una clusterización relativamente buena de cada grupo (Figura 20). Estos resultados permiten catalogar a estas variables como potenciales biomarcadores de la enfermedad de Parkinson, aunque sería necesario realizar análisis de clasificación para poder confirmar su efectividad como biomarcadores.

A través de los diferentes análisis se han identificado una serie de genes diferencialmente expresados y CpGs diferencialmente metilados entre muestras Control y PD que, además, han mostrado tener relación con la patología de estudio, con otras alteraciones neurodegenerativas, o con procesos neuronales. Sin embargo, en muchos casos hubo muestras con patrones poco marcados o invertidos respecto a su grupo. Este resultado sería indicativo de que los procesos que pueden estar afectados en las distintas regiones del sistema nervioso no son los mismos procesos que pueden estar afectados en sangre. Esto es lógico, pues son dos tejidos altamente diferenciados y que cumplen funciones muy distintas. Sin embargo, sí puede hablarse de lo que parece ser cierto efecto de la enfermedad en sangre. Aunque no es posible hablar de biomarcadores claros, a través de este trabajo se han podido identificar genes de interés que parecen estar afectados en sangre a nivel de expresión y metilación y que guardan relación conocida con la enfermedad.

6 Conclusiones

6.1 Conclusiones

Los análisis realizados a lo largo de este proyecto han servido para estudiar y caracterizar los cambios asociados a la enfermedad de Parkinson en el tejido sanguíneo a partir de datos de expresión génica y metilación del ADN. En general, los objetivos del proyecto se han cumplido, habiendo estudiado tanto los datos de expresión, los datos de metilación, como su integración para identificar potenciales biomarcadores de la enfermedad en sangre. Sin embargo, al obtener resultados no estadísticamente significativos entre los dos grupos ha surgido una limitación de cara a los análisis que se pueden realizar y a las deducciones que se pueden hacer.

A través de este proyecto se ha logrado lo siguiente:

- Se han identificado un grupo de marcadores de expresión y metilación en sangre que logran cierta separación entre muestras Control y muestras PD, y que podrían actuar como potenciales biomarcadores de la patología.
- Se han identificado genes, procesos (ubiquitinización y degradación proteica), y componentes mitocondriales que se han reportado como vinculados al Parkinsonismo y que en este trabajo se ha comprobado que están afectados en sangre.

6.2 Líneas de futuro

Por un lado, una vía interesante sería el estudio de las regiones diferencialmente metiladas, que aquí, debido a la falta de tiempo, no han podido trabajarse. El estudio de las regiones diferencialmente metiladas puede hacerse en muchos casos partiendo de CpGs sin diferencias estadísticamente significativas de metilación, pues lo que se comprueba es una tendencia general a la hiper o hipometilación por grupos de CpGs.

Por otro lado, para poder determinar con certeza la capacidad predictora de los potenciales biomarcadores identificados sería necesario elaborar modelos de clasificación mediante algoritmos de *machine learning*, como la clasificación por k-NN (del inglés, k-*Nearest Neighboors*) o la clasificación por el algoritmo de

Random Forest. Sin embargo, esto sería inviable en este caso dado el pequeño tamaño muestral. En su lugar, sería interesante plantear diversos análisis de clusterización jerárquica no supervisada a partir de combinaciones de marcadores de expresión y de metilación seleccionados en base a diversos criterios. De este modo se podría evaluar qué marcadores logran una mejor separación de los datos, y si la integración de estos consigue una mejoría. También podría valorarse la realización de modelos de clasificación con validación cruzada. Estos no permiten obtener una puntuación sobre la capacidad del modelo para clasificar casos nuevos, pero sí permite obtener estimación.

Además, mediante un estudio de validaición con un tamaño muestral mayor sería posible comprobar si los resultados aquí observados se mantienen, o si por el contrario los resultados obtenidos son debido a un tamaño muestral pequeño.

6.3 Seguimiento de la planificación

La planificación general planteada originalmente se ha logrado mantener a lo largo del proyecto, únicamente con pequeñas modificaciones temporales que se han ido haciendo sobre la marcha en función de las necesidades y los problemas que han ido surgiendo. Estos problemas consistieron principalmente en la obtención de resultados poco satisfactorios. A la hora de solventar estos problemas se optó por plantear vías alternativas para realizar los análisis, aunque los resultados obtenidos fueron muy similares. Por tanto, se puede considerar que la metodología planteada ha sido adecuada y ha servido para realizar los análisis esperados.

7 Glosario

A continuación, se muestran en orden alfabético las abreviaturas más comunes utilizadas a lo largo del trabajo y su significado.

- DIABLO: *Data Integration Analysis for Biomarker Discovery using Latent variable approaches for Omics studies.* Método de integración vertical también llamado sPLS-DA. Del paquete de R *mixOmics.*
- DEG: Gen/es diferencialmente expresado/s (*Differentially Expressed Genes).*
- DMP: Posición/es diferencialmente metilada/s (*Differentially methylated Position*).
- eQTM: Metilación de rasgos cuantitativos de expresión (*expression Quantitative Trait Methylation*)
- PCA: Análisis de los componentes principales (*Principal Component Analysis*).

8 Bibliografía

Abiola, O., Angel, J. M., Avner, P., Bachmanov, A. A., Belknap, J. K., Bennett, B., Blankenhorn, E. P., Blizard, D. A., Bolivar, V., Brockmann, G. A., Buck, K. J., Bureau, J. F., Casley, W. L., Chesler, E. J., Cheverud, J. M., Churchill, G. A., Cook, M., Crabbe, J. C., Crusio, W. E., Darvasi, A., *et al*. Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. *Nature reviews. Genetics*, 4(11), 911–916.

Alexopoulou, Z., Lang, J., Perrett, R. M., Elschami, M., Hurry, M. E., Kim, H. T., Mazaraki, D., Szabo, A., Kessler, B. M., Goldberg, A. L., Ansorge, O., Fulga, T. A., & Tofaris, G. K. (2016). Deubiquitinase Usp8 regulates α-synuclein clearance and modifies its toxicity in Lewy body disease. *Proceedings of the National Academy of Sciences of the United States of America*, 113(32), E4688–E4697.

Aryee M. J., Jaffe A. E., Corrada-Bravo H., Ladd-Acosta C., Feinberg A. P., Hansen K. D., Irizarry R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics*, 30(10), 1363–1369.

Barbeau A. (1961). Biochemistry of Parkinson's disease. *Proceedings of the seventh international congress of neurology, Rome, Sept, Societa Grafica Romana, Rome*, 2: 925.

Baylin S. B., Ohm J. E. (2006) Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction?. *Nat Rev Cancer*, 6(2), 107–116.

Benjamini Y., Hochberg Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Jo*urnal of the Royal Statistical Society*, Series B. 57 (1): 289–300.

Berridge, K. C., Venier, I. L., & Robinson, T. E. (1989). Taste reactivity analysis of 6-hydroxydopamine-induced aphagia: implications for arousal and anhedonia hypotheses of dopamine function. *Behavioral neuroscience*, 103(1), 36–45.

Bird A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1), 6–21.

Bock C., Tomazou E. M., Brinkman A. B., Müller F., Simmer F., Gu H., Jäger N., Gnirke A., Stunnenberg H. G., Meissner A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol*, 28(10), 1106–14.

Borrageiro, G., Haylett, W., Seedat, S., Kuivaniemi, H., & Bardien, S. (2018). A review of genome-wide transcriptomics studies in Parkinson's disease. *The European journal of neuroscience*, 47(1), 1–16.

Braak, H., Del Tredici, K., Rüb, U., de Vos, R. A., Jansen Steur, E. N., & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of aging*, 24(2), 197–211.

Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., & John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *GigaScience*, 11, giac001.

Calame, D. G., Herman, I., Maroofian, R., Marshall, A. E., Donis, K. C., Fatih, J. M., Mitani, T., Du, H., Grochowski, C. M., Sousa, S. B., Gijavanekar, C., Bakhtiari, S., Ito, Y. A., Rocca, C., Hunter, J. V., Sutton, V. R., Emrick, L. T., Boycott, K. M., Lossos, A., Fellig, Y., … Lupski, J. R. (2022). Biallelic Variants in the Ectonucleotidase ENTPD1 Cause a Complex Neurodevelopmental Disorder with Intellectual Disability, Distinct White Matter Abnormalities, and Spastic Paraplegia. *Annals of neurology*, 92(2), 304–321.

Cappelli, E., Felici, G., & Weitschek, E. (2018). Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. *BioData mining*, 11, 22.

Chen, B. J., Mills, J. D., Takenaka, K., Bliim, N., Halliday, G. M., & Janitz, M. (2016). Characterization of circular RNAs landscape in multiple system atrophy brain. *Journal of neurochemistry*, 139(3), 485–496.

Chen-Plotkin, A. S., Albin, R., Alcalay, R., Babcock, D., Bajaj, V., Bowman, D., Buko, A., Cedarbaum, J., Chelsky, D., Cookson, M. R., Dawson, T. M.,

Dewey, R., Foroud, T., Frasier, M., German, D., Gwinn, K., Huang, X., Kopil, C., Kremer, T., Lasch, S., et al., Zhang, J. (2018). Finding useful biomarkers for Parkinson's disease. *Science translational medicine*, 10(454), eaam6003.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, 34(17), i884–i890.

Chen Y., Lun A. T. L., Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research,* 5, 1438.

Coste de Bagneaux, P., von Elsner, L., Bierhals, T., Campiglio, M., Johannsen, J., Obermair, G. J., Hempel, M., Flucher, B. E., & Kutsche, K. (2020). A homozygous missense variant in CACNB4 encoding the auxiliary calcium channel beta4 subunit causes a severe neurodevelopmental disorder and impairs channel and non-channel functions. *PLoS genetics*, 16(3), e1008625.

Correddu, D., & Leung, I. K. H. (2019). Targeting mRNA translation in Parkinson's disease. *Drug discovery today*, 24(6), 1295–1303.

Chatterjee, P., Roy, D., Bhattacharyya, M., & Bandyopadhyay, S. (2017). Biological networks in Parkinson's disease: an insight into the epigenetic mechanisms associated with this disease. *BMC genomics*, 18(1), 721.

Chuang, Y. H., Lu, A. T., Paul, K. C., Folle, A. D., Bronstein, J. M., Bordelon, Y., Horvath, S., & Ritz, B. (2019). Longitudinal Epigenome-Wide Methylation Study of Cognitive Decline and Motor Progression in Parkinson's Disease. *Journal of Parkinson's disease*, 9(2), 389–400.

Croese, T., & Furlan, R. (2018). Extracellular vesicles in neurodegenerative diseases. *Molecular aspects of medicine*, 60, 52–61.

Damier, P., Hirsch, E. C., Agid, Y., & Graybiel, A. M. (1999). The substantia nigra of the human brain. II. Patterns of loss of dopamine-containing neurons in Parkinson's disease. *Brain: a journal of neurology*, 122 (8), 1437–1448.

Dawson, T. M., & Dawson, V. L. (2003). Rare genetic mutations shed light on the pathogenesis of Parkinson disease. *The Journal of clinical investigation*, 111(2), 145–151.

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, 3(6), 771–784.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England),* 29(1), 15–21.

Donato R. (2001). S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *The international journal of biochemistry & cell biol*ogy, 33(7), 637–668.

Ehringer, H., & Hornykiewicz, O. (1960). *Klinische Wochenschrift*, 38, 1236–1239.

Elstner, M., Morris, C. M., Heim, K., Lichtner, P., Bender, A., Mehta, D., Schulte, C., Sharma, M., Hudson, G., Goldwurm, S., Giovanetti, A., Zeviani, M., Burn, D. J., McKeith, I. G., Perry, R. H., Jaros, E., Krüger, R., Wichmann, H. E., Schreiber, S., Campbell, H., … Turnbull, D. M. (2009). Single-cell expression profiling of dopaminergic neurons combined with association analysis identifies pyridoxal kinase as Parkinson's disease gene. *Annals of neurology*, 66(6), 792–798.

Erkkinen, M. G., Kim, M. O., & Geschwind, M. D. (2018). Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases. *Cold Spring Harbor perspectives in biology*, 10(4), a033118.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England),* 32(19), 3047–3048.

Fayyad, M., Salim, S., Majbour, N., Erskine, D., Stoops, E., Mollenhauer, B., & El-Agnaf, O. M. A. (2019). Parkinson's disease biomarkers based on α-synuclein. *Journal of neurochemistry*, 150(5), 626–636.

Feng, S., Cokus, S. J., Zhang, X., Chen, P. Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., & Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19), 8689–8694.

Fyfe I. (2021). RNA biomarkers of Parkinson disease. *Nature reviews. Neurology*, 17(3), 132.

Gibb, W. R., & Lees, A. J. (1991). Anatomy, pigmentation, ventral and dorsal subpopulations of the substantia nigra, and differential cell death in Parkinson's disease. *Journal of neurology, neurosurgery, and psychiatry*, 54(5), 388–396.

Guo, J. F., Zhang, L., Li, K., Mei, J. P., Xue, J., Chen, J., Tang, X., Shen, L., Jiang, H., Chen, C., Guo, H., Wu, X. L., Sun, S. L., Xu, Q., Sun, Q. Y., Chan, P., Shang, H. F., Wang, T., Zhao, G. H., Liu, J. Y., … Tang, B. S. (2018). Coding mutations in NUS1 contribute to Parkinson's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 115(45), 11567–11572.

Guzmán, F. (2017). Vías dopaminérgicas y antipsicóticos. Programa de actualización en psicofarmacología 2018. Instituto de Psicofarmacología. Consultado el 10 de junio de 2023 en https://psicofarmacologia.com/antipsicoticos/vias-dopaminergicas-y-antipsicoticos

Grünblatt, E., Mandel, S., Jacob-Hirsch, J., Zeligson, S., Amariglo, N., Rechavi, G., Li, J., Ravid, R., Roggendorf, W., Riederer, P., & Youdim, M. B. (2004). Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell adhesion/cellular matrix and vesicle trafficking genes. *Journal of neural transmission (Vienna, Austria : 1996)*, 111(12), 1543–1573.

Hannon, E., Lunnon, K., Schalkwyk, L., & Mill, J. (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, 10(11), 1024–1032.

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome biology, 18(1), 83.

Haas, A. L., Ahrens, P., Bright, P. M., & Ankel, H. (1987). Interferon induces a 15-kilodalton protein exhibiting marked homology to ubiquitin. *The Journal of biological chemistry*, 262(23), 11315–11323.

H. B. Mann. D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." *Ann. Math. Statist*. 18 (1) 50 - 60, March, 1947.

Hebron, M., Peyton, M., Liu, X., Gao, X., Wang, R., Lonskaya, I., & Moussa, C. E. (2017). Discoidin domain receptor inhibition reduces neuropathology and attenuates inflammation in neurodegeneration models. *Journal of neuroimmunology,* 311, 1–9.

Hellman, A., & Chess, A. (2007). Gene body-specific methylation on the active X chromosome. *Science (New York, N.Y.),* 315(5815), 1141–1143.

Henderson, A. R., Wang, Q., Meechoovet, B., Siniard, A. L., Naymik, M., De Both, M., Huentelman, M. J., Caselli, R. J., Driver-Dunckley, E., Dunckley, T. (2021). DNA Methylation and Expression Profiles of Whole Blood in Parkinson's Disease. *Frontiers in genetics*, 12, 640266.

Henn, I. H., Bouman, L., Schlehe, J. S., Schlierf, A., Schramm, J. E., Wegener, E., Nakaso, K., Culmsee, C., Berninger, B., Krappmann, D., Tatzelt, J., & Winklhofer, K. F. (2007). Parkin mediates neuroprotection through activation of IkappaB kinase/nuclear factor-kappaB signaling. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(8), 1868–1878.

Huang, Q., Wang, H., Perry, S. W., & Figueiredo-Pereira, M. E. (2013). Negative regulation of 26S proteasome stability via calpain-mediated cleavage of

Rpn10 subunit upon mitochondrial dysfunction in neurons. *The Journal of biological chemistry*, 288(17), 12161–12174.

Huse, D. M., Schulman, K., Orsini, L., Castelli-Haley, J., Kennedy, S., & Lenhart, G. (2005). Burden of illness in Parkinson's disease. *Movement disorders : official journal of the Movement Disorder Society*, 20(11), 1449–1454.

Hsu, F. M., Gohain, M., Chang, P., Lu, J. H., Chen, P. Y. (2018). Chapter 4 - Bioinformatics of Epigenomic Data Generated From Next-Generation Sequencing. *Epigenetics in Human Disease (Second Edition)*, 65–106.

Im, E., Yoo, L., Hyun, M., Shin, W. H., & Chung, K. C. (2016). Covalent ISG15 conjugation positively regulates the ubiquitin E3 ligase activity of parkin. *Open biology*, 6(8), 160193.

Iranzo, A., Tolosa, E., Gelpi, E., Molinuevo, J. L., Valldeoriola, F., Serradell, M., Sanchez-Valle, R., Vilaseca, I., Lomeña, F., Vilas, D., Lladó, A., Gaig, C., & Santamaria, J. (2013). Neurodegenerative disease status and post-mortem pathology in idiopathic rapid-eye-movement sleep behaviour disorder: an observational cohort study. *The Lancet. Neurology*, 12(5), 443–453.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S., & Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2), 178–186.

Irwin, D. J., Cairns, N. J., Grossman, M., McMillan, C. T., Lee, E. B., Van Deerlin, V. M., Lee, V. M., Trojanowski, J. Q. (2015). Frontotemporal lobar degeneration: defining phenotypic diversity through personalized medicine. *Acta neuropathologica*, 129(4), 469–491.

Isobe, C., Abe, T., & Terayama, Y. (2010). Levels of reduced and oxidized coenzyme Q-10 and 8-hydroxy-2'-deoxyguanosine in the cerebrospinal fluid of patients with living Parkinson's disease demonstrate that mitochondrial oxidative damage and/or oxidative DNA damage contributes to the neurodegenerative process. *Neuroscience letters,* 469(1), 159–163.

Jafari, M., & Ansari-Pour, N. (2019). Why, When and How to Adjust Your P Values?. *Cell journal*, 20(4), 604–607.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, 8(1), 118–127.

Kochmanski, J., Kuhn, N. C., & Bernstein, A. I. (2022). Parkinson's disease-associated, sex-specific changes in DNA methylation at PARK7 (DJ-1), SLC17A6 (VGLUT2), PTPRN2 (IA-2β), and NR4A2 (NURR1) in cortical neurons. *NPJ Parkinson's disease*, 8(1), 120.

Laird P. W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*, 11(3), 191–203.

Leggio, L., Vivarelli, S., L'Episcopo, F., Tirolo, C., Caniglia, S., Testa, N., Marchetti, B., & Iraci, N. (2017). microRNAs in Parkinson's Disease: From Pathogenesis to Novel Diagnostic and Therapeutic Approaches. *International journal of molecular sciences*, 18(12), 2698.

Li, S., & Le, W. (2017). Biomarker Discovery in Parkinson's Disease: Present Challenges and Future Opportunities. *Neuroscience bulletin*, 33(5), 481–482.

Li, T., & Le, W. (2020). Biomarkers for Parkinson's Disease: How Good Are They?. *Neuroscience bulletin*, 36(2), 183–194.

Li, Y., Ge, X., Peng, F., Li, W., Li, J. J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome biology*, 23(1), 79.

Louhimo , R. & Hautaniemi, S. (2011). CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, 27(6), 887–888.

Love, M. I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 1–21.

Marshall, J. F., Levitan, D., & Stricker, E. M. (1976). Activation-induced restoration of sensorimotor functions in rats with dopamine-depleting brain lesions. *Journal of comparative and physiological psychology*, 90(6), 536–546.

Mikeska T., Craig J. M. (2014) DNA methylation biomarkers: cancer and beyond. *Genes (Basel),* 5(3), 821–864.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, 22(1), 67–72.

Mo, Q. & Shen, R. (2023). *iClusterPlus: Integrative clustering of multi-type genomic data*. R package version 1.36.1.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11), 4245–4250.

Moylan, C., Murphy, S. (2016). Chapter 2 - DNA Methylation: Basic Principles. *Medical Epigenetics*, 11-31.

Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Frontiers in plant science, 12, 657240.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., Kara, E., Bras, J., Sharma, M., Schulte, C., Keller, M. F., Arepalli, S., Letson, C., Edsall, C., Stefansson, H., Liu, X., Pliner, H., Lee, J. H., Cheng, R., et al. Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature genetics*, 46(9), 989–993.

Neidhart, M. (2016). DNA Methylation – Introduction. DNA Methylation and Complex Human Disease, 1ª edición, 1–8.

Nitert, M. D., Dayeh, T., Volkov, P., Elgzyri, T., Hall, E., Nilsson, E., Yang, B. T., Lang, S., Parikh, H., Wessman, Y., Weishaupt, H., Attema, J., Abels, M., Wierup, N., Almgren, P., Jansson, P. A., Rönn, T., Hansson, O., Eriksson, K. F., Groop, L., … Ling, C. (2012). Impact of an exercise intervention on DNA methylation in skeletal muscle from first-degree relatives of patients with type 2 diabetes. *Diabetes*, 61(12), 3322–3332.

Nishimura, T., Kubosaki, A., Ito, Y., & Notkins, A. L. (2009). Disturbances in the secretion of neurotransmitters in IA-2/IA-2beta null mice: changes in behavior, learning and lifespan. *Neuroscience*, 159(2), 427–437.

Papapetropoulos, S., Ffrench-Mullen, J., McCorquodale, D., Qin, Y., Pablo, J., & Mash, D. C. (2006). Multiregional gene expression profiling identifies MRPS6 as a possible candidate gene for Parkinson's disease. *Gene expression*, 13(3), 205–215.

Parnetti, L., Gaetani, L., Eusebi, P., Paciotti, S., Hansson, O., El-Agnaf, O., Mollenhauer, B., Blennow, K., & Calabresi, P. (2019). CSF and blood biomarkers for Parkinson's disease. *The Lancet. Neurology*, 18(6), 573–586.

Parkinson’s Foundation. (s. f.). Statistics. Get informed about Parkinson's disease with these key numbers. Consultado el 1 de junio de 2023 en https://www.parkinson.org/understanding-parkinsons/statistics

Perlis R. H. (2011). Translating biomarkers to clinical practice. *Molecular psyc*hiatry, 16(11), 1076–1087.

Polymeropoulos, M. H., Higgins, J. J., Golbe, L. I., Johnson, W. G., Ide, S. E., Di Iorio, G., Sanges, G., Stenroos, E. S., Pho, L. T., Schaffer, A. A., Lazzarini, A. M., Nussbaum, R. L., & Duvoisin, R. C. (1996). Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science (New York, N.Y.),* 274(5290), 1197–1199.

Polymeropoulos, M. H., Lavedan, C., Leroy, E., Ide, S. E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., Stenroos, E. S., Chandrasekharappa, S., Athanassiadou, A., Papapetropoulos, T., Johnson, W. G., Lazzarini, A. M., Duvoisin, R. C., Di Iorio, G., Golbe, L. I.,

& Nussbaum, R. L. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science (New York, N.Y.),* 276(5321), 2045–2047.

Powder K. E. (2020). Quantitative Trait Loci (QTL) Mapping. *Methods in molecular biology (Clifton, N.J.),* 2082, 211–229.

Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Böhm, K. J., & Winner, B. (2013). α-Synuclein oligomers impair neuronal microtubule-kinesin interplay. *The Journal of biological chemistry*, 288(30), 21742–21754.

Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta physiologica Scandinavica. Supplementum*, 367, 95–122.

Raplee, I. D., Evsikov, A. V., & Marín de Evsikova, C. (2019). Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. *Journal of personalized medicine*, 9(2), 18.

Ritchie M. E., Phipson B., Wu D., Hu Y., Law C. W., Shi W., Smyth G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47.

Robinson, M. D., McCarthy, D. J., Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11), e1005752.

Saraiva, C., Lopes-Nunes, J., Esteves, M., Santos, T., Vale, A., Cristóvão, A. C., Ferreira, R., & Bernardino, L. (2023). CtBP Neuroprotective Role in Toxin-Based Parkinson's Disease Models: From Expression Pattern to Dopaminergic Survival. *Molecular neurobiology*.

Sathe, K., Maetzler, W., Lang, J. D., Mounsey, R. B., Fleckenstein, C., Martin, H. L., Schulte, C., Mustafa, S., Synofzik, M., Vukovic, Z., Itohara, S., Berg, D., & Teismann, P. (2012). S100B is increased in Parkinson's disease and ablation protects against MPTP-induced toxicity through the RAGE and TNF-α pathway. *Brain : a journal of neurology*, 135(Pt 11), 3336–3347.

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5), 1412–1417.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA (New York, N.Y.),* 22(6), 839–851.

Shabalin, A. A. (2012) Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10), 1353–1358.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics (Oxford, England), 25(22), 2906–2912.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., & Sander, C. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PloS one*, 7(4), e35236.

Silva, T., Young, J. I., Zhang, L., Gomez, L., Schmidt, M. A., Varma, A., Chen, X. S., Martin, E. R., & Wang, L. (2022). Cross-tissue analysis of blood and brain epigenome-wide association studies in Alzheimer's disease. *Nature communications*, 13(1), 4852.

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., & Lê Cao, K. A. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics (Oxford, England)*, 35(17), 3055–3062.

Spillantini, M. G., Schmidt, M. L., Lee, V. M., Trojanowski, J. Q., Jakes, R., & Goedert, M. (1997). Alpha-synuclein in Lewy bodies. *Nature*, 388(6645), 839–840.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers?. *Current opinion in HIV and AIDS*, 5(6), 463–466.

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and biology insights*, 14, 1177932219899051.

Suwijn, S. R., van Boheemen, C. J., de Haan, R. J., Tissingh, G., Booij, J., & de Bie, R. M. (2015). The diagnostic accuracy of dopamine transporter SPECT imaging to detect nigrostriatal cell loss in patients with Parkinson's disease or clinically uncertain parkinsonism: a systematic review. *EJNMMI research*, 5, 12.

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., & Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics (Oxford, England),* 29(2), 189–196.

The Michael J. Fox Foundation for Parkinson's Research | Parkinson's Disease (2019). *Parkinson's Disease Economic Burden on Patients, Families and the Federal Government Is $52 Billion, Doubling Previous Estimates*. Consultado el 14 de junio en https://www.michaeljfox.org/publication/parkinsons-disease-economic-burden-patients-families-and-federal-government-52-billion

The National Collaborating Centre for Chronic Conditions (2006). Symptomatic pharmacological therapy in Parkinson's disease. *Parkinson's Disease. London: Royal College of Physicians,* 59–100.

Thind, A. S., Monga, I., Thakur, P. K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M., & Ashford, B. (2021). Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology. *Briefings in bioinformatics*, 22(6), bbab259.

Tian Y., Morris T. J., Webster A. P., Yang Z., Beck S., Andrew F., Teschendorff A. E. (2017). ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*, btx513.

Tsonaka, R., & Spitali, P. (2021). Negative Binomial mixed models estimated with the maximum likelihood method can be used for longitudinal RNAseq data. *Briefings in bioinformatics*, 22(4), bbaa264.

Walton, E., Hass, J., Liu, J., Roffman, J. L., Bernardoni, F., Roessner, V., Kirsch, M., Schackert, G., Calhoun, V., & Ehrlich, S. (2016). Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophrenia bulletin*, 42(2), 406–414.

Wang, H., Xiao, Z., Zheng, J., Wu, J., Hu, X. L., Yang, X., & Shen, Q. (2019). ZEB1 Represses Neural Differentiation and Cooperates with CTBP2 to Dynamically Regulate Cell Migration during Neocortex Development. *Cell reports*, 27(8), 2335–2353.e6.

Wang Z., Gerstein M., Snyder M. (January 2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63.

Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T., & Migeon, B. R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proceedings of the National Academy of Sciences of the United States of America*, 81(9), 2806–2810.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Cambridge (Mass.))*, 2(3), 100141.

Yang, W., Hamilton, J. L., Kopil, C., Beck, J. C., Tanner, C. M., Albin, R. L., Ray Dorsey, E., Dahodwala, N., Cintina, I., Hogan, P., & Thompson, T. (2020). Current and projected future economic burden of Parkinson's disease in the U.S. *NPJ Parkinson's disease*, 6, 15.

Yang, Z., Li, T., Li, S., Wei, M., Qi, H., Shen, B., Chang, R. C., Le, W., & Piao, F. (2019). Altered Expression Levels of MicroRNA-132 and Nurr1 in

Peripheral Blood of Parkinson's Disease: Potential Disease Biomarkers. *ACS chemical neuroscience*, 10(5), 2243–2249.

Yu, G. (2023). enrichplot: Visualization of Functional Enrichment Result. R package version 1.20.0, https://yulab-smu.top/biomedical-knowledge-mining-book/.

Zheng, B., Liao, Z., Locascio, J.J., Lesniak, K.A., Roderick, S.S., Watt, M.L., Eklund, A.C., Zhang-James, Y., Kim, P.D., Hauser, M.A., Grünblatt, E., Moran, L.B., Mandel, S.A., Riederer, P., Miller, R.M., Federoff, H.J., Wüllner, U., Papapetropoulos, S., Youdim, M.B., Cantuti-Castelvetri, I., Young, A.B., Vance, J.M., Davis, R.L., Hedreen, J.C., Adler, C.H., Beach, T.G., Graeber, M.B., Middleton, F.A., Rochet, J.-C., Scherzer, C.R., & Global PD Gene Expression (GPEX) Consortium (2010). PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. *Sci. Transl. Med*., 2, 52ra73.

Zhu, M., Xing, D., Lu, Z., Fan, Y., Hou, W., Dong, H., Xiong, L., & Dong, H. (2015). DDR1 may play a key role in destruction of the blood-brain barrier after cerebral ischemia-reperfusion. *Neuroscience research*, 96, 14–19.

9 Anexo

Anexo I: Resultados del control de calidad y del estudio de la distribución de los valores de las variables

Los archivos FASTQ con las secuencias de cDNA, derivadas la secuenciación de ARN, se sometieron a un control de calidad para identificar regiones con baja calidad de secuenciación y posibles adaptadores contaminantes en los extremos 3’ de las secuencias.

En primer lugar, se muestra calidad media de secuenciación para cada base y archivo FASTQ (Figura A1). Se puede observar cómo, en general, la calidad media de las secuencias oscila entre la puntuación *Phred* de 30 y 40, con una media general de en torno a 38. Esto es indicativo de que la secuenciación se ha llevado a cabo correctamente, y hay una alta probabilidad de que cada base secuenciada se corresponda correctamente con la base identificada. Se pudo observar una pequeña caída al inicio y al final de las secuencias, en los extremos 5’ y 3’, respectivamente. En ambos casos se asume que es un artefacto debido a la propia química por la cual se lleva a cabo la secuenciación, reduciendo la calidad de la secuenciación normalmente en las primeras y últimas bases secuenciadas.

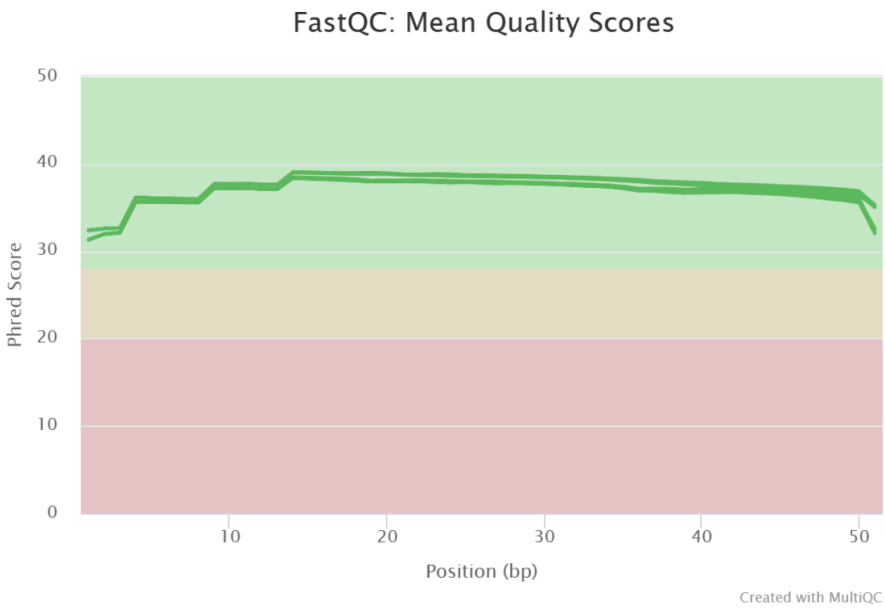

Figura A1. Calidad media de la secuenciación de cada base y archivo FASTQ. En general, la puntuación *Phred* media es alta, indicativo de una secuenciación correcta.

Algo similar podemos observar en la Figura A2, donde se muestra que el mayor volumen de lecturas se encuentra en torno a la calidad media de secuenciación de 38.

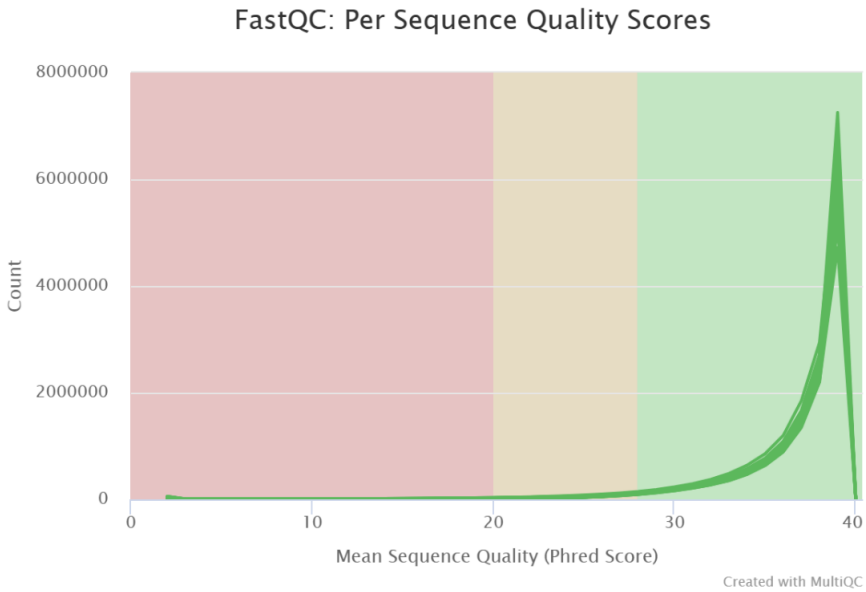

Figura A2. Número de lecturas frente a la calidad media de cada secuencia.

Respecto al contenido en G+C de las lecturas, este fue, como se esperaba, de aproximadamente el 50 % (Figura A3), con una mayor concentración de lecturas con un contenido en G+C cerca del 60 %.

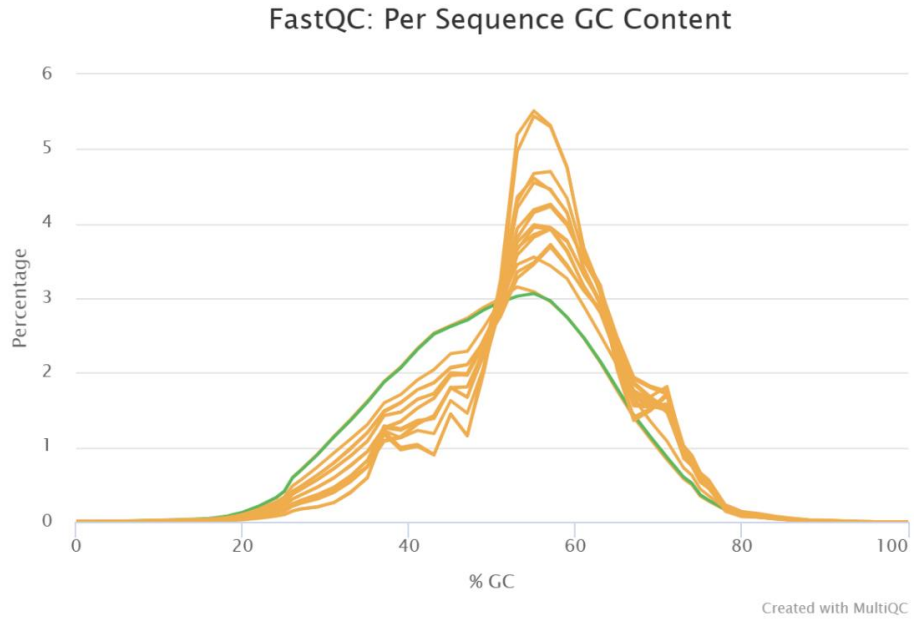

Figura A3. Porcentaje de contenido G+C de los archivos FASTQ. El grueso de las lecturas tiene un porcentaje de G+C cerca del 60 %.

Por último, se evaluó el contenido de secuencias sobrerrepresentadas (Figura A4). Se pudo observar que este contenido es, en general, ínfimo. Estas secuencias eran, principalmente, colas de poli(A) y restos de los adaptadores adaptadores *TruSeq Index* 9 y *TruSeq Index* 12 de Illumina.

Ninguna de las secuencias repetitivas detectadas superó el 1 % de presencia; aún así, se procedió a una fase de procesado de las secuencias en las que se eliminaron los adaptadores detectados y las colas de poli(A) existentes.

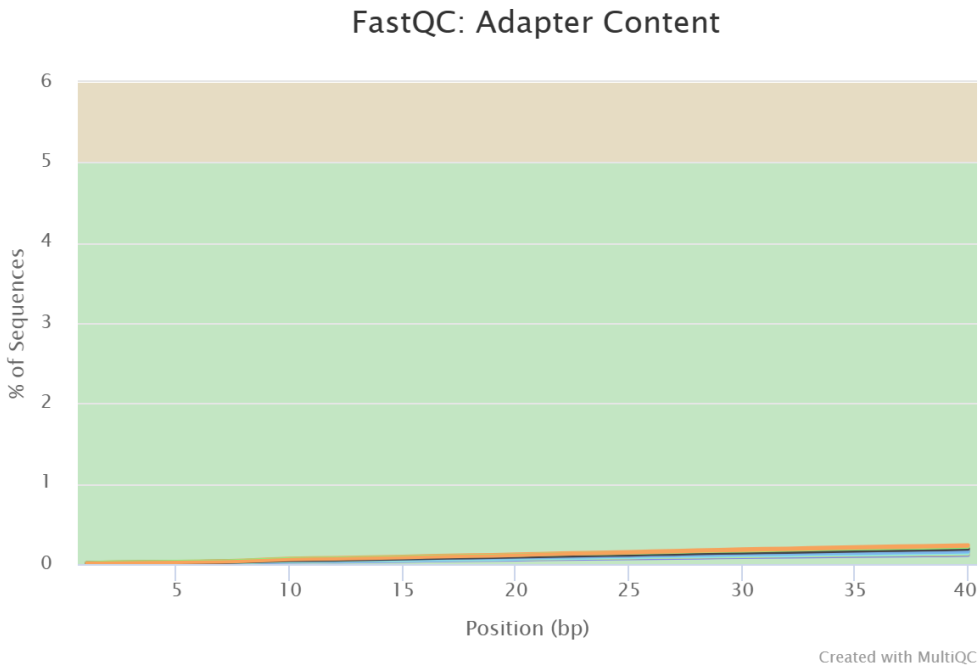

Figura A4. Porcentaje de lecturas con secuencias sobrerrepresentadas detectadas. El porcentaje de presencia de las secuencias repetitivas fue inferior al 1 % en todos los casos.

Además, se observó que la longitud de todas las lecturas fue de 51 pares de bases.

Respecto a los valores beta de metilación de las diferentes sondas, el control de calidad consistió en la observación de la distribución de estos para cada una de las 26 muestras (Figura A5) y la detección y corrección del efecto Batch (Figuras A6 y A7).

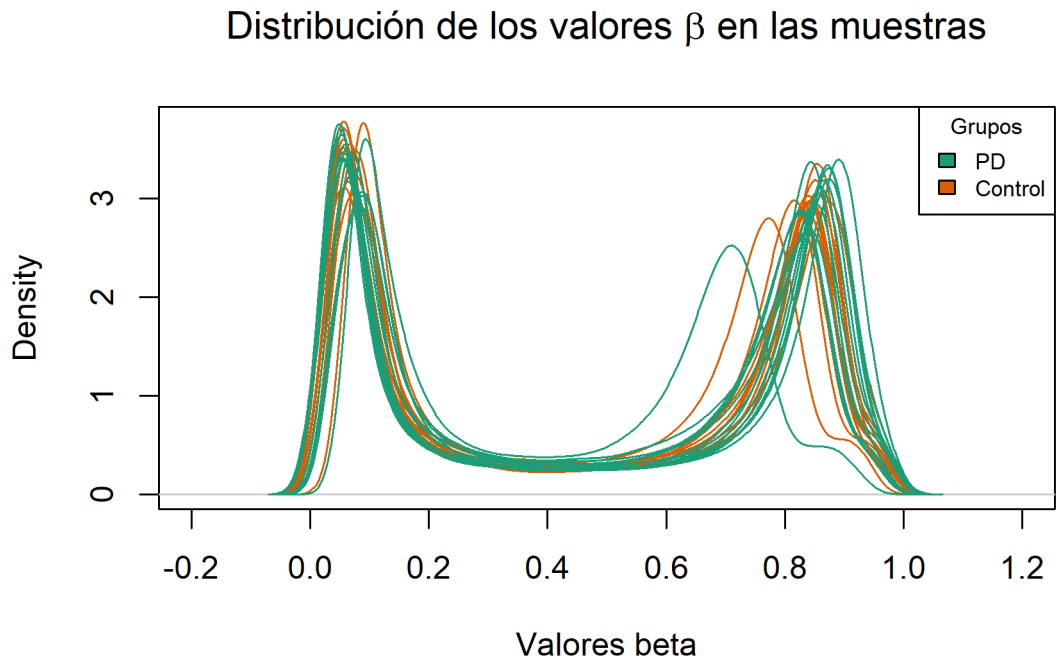

Figura A5. Distribución de los valores beta para cada una de las muestras del grupo PD (enfermos de Parkinson, verde) y Control (naranja).

Se puede observar cómo la distribución de los valores beta sigue una distribución aproximadamente bimodal para cada una de las muestras, con una mayor concentración de valores beta en torno al 0.1 y en torno al 0.8. Estos valores se corresponden con proporciones de metilación, siendo los valores menores marcas de hipometilación, y los valores mayores marcas de hipermetilación. Solo dos muestras mostraron un pico de hipermetilación algo más alejado de la media general. Pese a eso, la distribución general de valores beta fue la esperada.

Una vez se normalizaron los valores beta para equiparar las sondas tipo I y tipo II se procedió a estudiar el posible efecto Batch de las covariables Edad y Género sobre la metilación. Esto se hizo con el objetivo de identificar posibles diferencias significativas en función a estas covariables en pos de mitigar su efecto. Para ello se aplicó una DVS sobre la matriz de valores beta. Luego se observó el efecto de la variable Grupo y de las covariables sobre los principales componentes encargados de explicar la variabilidad de la matriz (Figura A6).

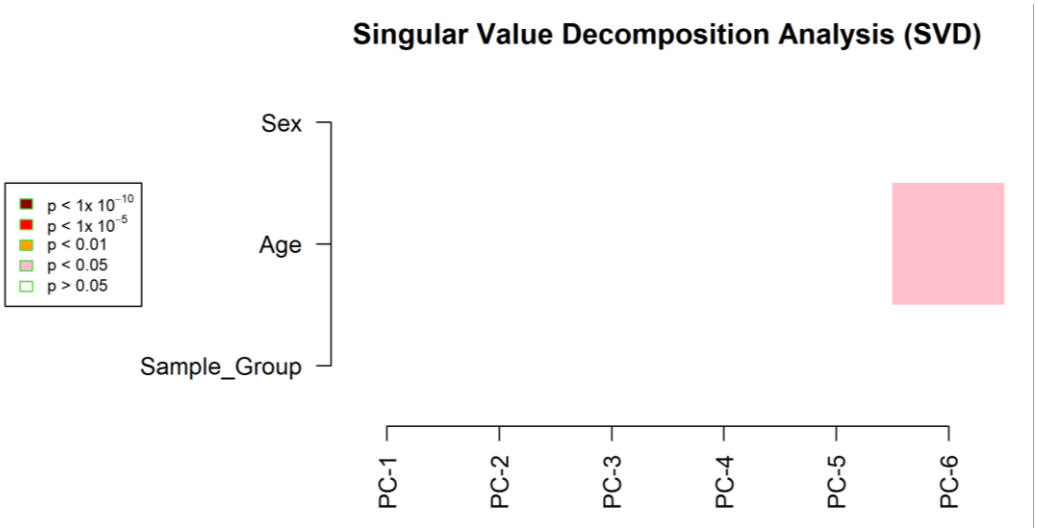

Figura A6. Mapa de calor con el efecto de la variable Grupo y las covariables Edad (*Age*) y Género (*Sex*) en los diferentes componentes de la variabilidad de la matriz de valores beta. Se indican los componentes en el eje X y las variables en el eje Y.

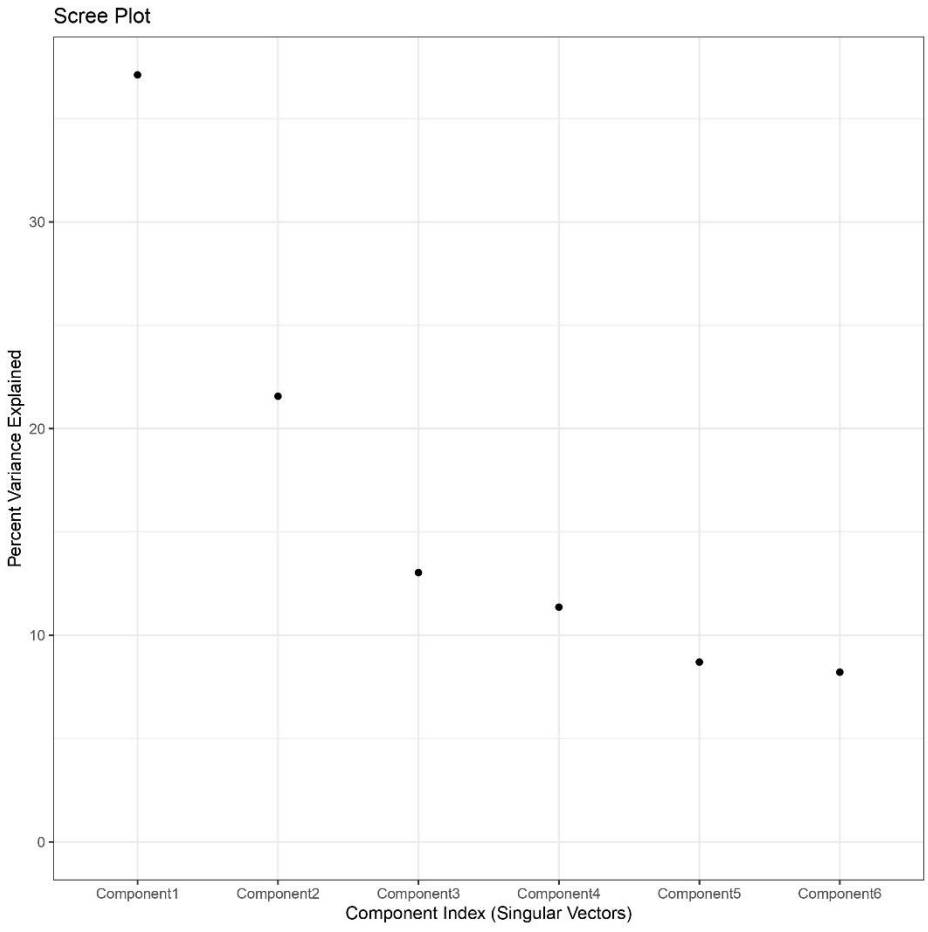

Figura A7. Porcentaje de variabilidad explicada por cada componente principal derivado de la matriz de valores beta corregidos. Se indican los componentes en el eje X y el porcentaje en el eje Y.

Podemos observar cómo se ha detectado correlación entre el sexto componente deconvolucionado y la covariable Edad (Figura A6). Sin embargo, no consideramos este resultado como efecto Batch puesto que el nivel de correlación es pequeño (≤ 0.05) y los 5 componentes superiores explican más de un 80 % de la varianza del conjunto (Figura A7). Por ende, por no se realizó la corrección del efecto.

Respecto a la distribución de los valores de metilación para cada una de las variables de estudio, más del 50 % de las variables pudieron considerarse como normalmente distribuidas y homocedásticas entre los grupos en base a los resultados obtenidos tras las pruebas de Shapiro-Wilks y Levene respectivamente aplicadas.

Anexo II: Genes identificados como DEG y con al menos un DMP

Tabla A1. Genes identificados como DEG con al menos 5 DMP asociado o de interés en el trabajo.

Gen	p-valor	FDR	Estadístico W	logFC	Nº DMP
CTBP2[1]	0.001	0.339	20	0.49	6
CACNB4[2]	0.002	0.339	26	1.10	4
CYB561[1]	0.002	0.339	143	-0.46	2
C7orf50	0.006	0.339	136	-0.57	14
ZMIZ1	0.031	0.339	42	0.53	13
RECQL5	0.036	0.339	125	-0.30	10
LPP	0.020	0.339	39	0.33	6
NFYA	0.046	0.339	45	0.31	6
CTBP2	0.001	0.339	20	0.49	6
PRDM2	0.005	0.339	31	0.29	5
RBM47	0.013	0.339	36	0.66	5
GPSM1	0.013	0.339	132	-0.57	5
SORL1	0.031	0.339	42	0.36	5
ENTPD1	0.031	0.339	42	0.25	5
CACNA2D4	0.036	0.339	125	-0.28	5
ATP8B4	0.006	0.339	32	0.84	5

[1] Genes directamente vinculados a la enfermedad de Parkinson.

[2] Genes relacionados con el correcto neurodesarrollo y la capacidad motriz.

Las columnas p-valor, FDR, Estadístico W y logFC se obtuvieron en la fase de análisis de expresión diferencial.

Anexo III: Resultados del enriquecimiento

Tabla A2. Procesos biológicos identificados como significativos. Se indican en la zona superior los procesos biológicos de interés, separados entre aquellos relacionados con la ubiquitinización (arriba) o con procesos de modificación proteica (abajo), y en la zona inferior los procesos biológicos más significativos en base a FDR e ilustrados en la Figura 14A.

Identificador	Descripción del término	Ratio de genes	p-valor	FDR
GO:0043161	*proteasome-mediated ubiquitin-dependent protein catabolic process*	59/1465	0.000	0.013
GO:0006513	*protein monoubiquitination*	18/1465	0.000	0.016
GO:0006475	*internal protein amino acid acetylation*	28/1465	0.000	0.017
GO:0018394	*peptidyl-lysine acetylation*	29/1465	0.000	0.018
GO:0018205	*peptidyl-lysine modification*	51/1465	0.000	0.018
GO:0031647	*regulation of protein stability*	45/1465	0.000	0.022
GO:0016570	*histone modification*	76/1465	0.000	0.000
GO:0031056	*regulation of histone modification*	30/1465	0.000	0.001
GO:0008380	*RNA splicing*	68/1465	0.000	0.001
GO:0006397	*mRNA processing*	71/1465	0.000	0.001
GO:0006913	*nucleocytoplasmic transport*	50/1465	0.000	0.004
GO:0051169	*nuclear transport*	50/1465	0.000	0.004
GO:0048193	*Golgi vesicle transport*	46/1465	0.000	0.004
GO:0000910	*cytokinesis*	33/1465	0.000	0.004
GO:0000375	*RNA splicing, via transesterification reactions*	49/1465	0.000	0.006
GO:0016571	*histone methylation*	26/1465	0.000	0.006
GO:0000377	*RNA splicing, via transesterification reactions with bulged adenosine as nucleophile*	48/1465	0.000	0.007
GO:0000398	*mRNA splicing, via spliceosome*	48/1465	0.000	0.007
GO:2001020	*regulation of response to DNA damage stimulus*	46/1465	0.000	0.009
GO:0034470	*ncRNA processing*	58/1465	0.000	0.014
GO:0022613	*ribonucleoprotein complex biogenesis*	62/1465	0.000	0.016
GO:0018393	*internal peptidyl-lysine acetylation*	28/1465	0.000	0.016
GO:0016573	*histone acetylation*	27/1465	0.000	0.016
GO:0019079	*viral genome replication*	24/1465	0.000	0.016
GO:0043484	*regulation of RNA splicing*	30/1465	0.000	0.018
GO:0072594	*establishment of protein localization to organelle*	59/1465	0.000	0.018
GO:0016032	*viral process*	56/1465	0.000	0.020
GO:0090501	*RNA phosphodiester bond hydrolysis*	27/1465	0.000	0.020
GO:0032386	*regulation of intracellular transport*	46/1465	0.000	0.022
GO:0032386	*regulation of intracellular transport*	46/1465	0.000	0.022

Tabla A3. Componentes celulares identificados como significativos. Se indican en la zona superior los componentes celulares de interés, separados entre aquellos relacionados con la mitocondria (arriba) y relacionado con la ubiquitinización (abajo), y en la zona inferior los procesos biológicos más significativos en base a FDR e ilustrados en la Figura 15A.

Identificador	Descripción del término	Ratio de genes	p-valor	FDR
GO:0098800	*inner mitochondrial membrane protein complex*	27/1535	0.000	0.007
GO:0005743	*mitochondrial inner membrane*	63/1535	0.000	0.007
GO:0098798	*mitochondrial protein-containing complex*	42/1535	0.000	0.009
GO:1904949	*ATPase complex*	25/1535	0.000	0.009
GO:0005759	*mitochondrial matrix*	56/1535	0.002	0.047
GO:0000151	*ubiquitin ligase complex*	41/1535	0.001	0.019
GO:0005819	*spindle*	64/1535	0.000	0.000
GO:0005774	*vacuolar membrane*	65/1535	0.000	0.001
GO:0005765	*lysosomal membrane*	59/1535	0.000	0.001
GO:0098852	*lytic vacuole membrane*	59/1535	0.000	0.001
GO:0035770	*ribonucleoprotein granule*	43/1535	0.000	0.001
GO:0036464	*cytoplasmic ribonucleoprotein granule*	39/1535	0.000	0.004
GO:0016363	*nuclear matrix*	24/1535	0.000	0.004
GO:0016607	*nuclear speck*	56/1535	0.000	0.004
GO:0034399	*nuclear periphery*	26/1535	0.000	0.006
GO:1902493	*acetyltransferase complex*	20/1535	0.000	0.009
GO:0000932	*P-body*	19/1535	0.000	0.009
GO:0000123	*histone acetyltransferase complex*	18/1535	0.000	0.012
GO:0005681	*spliceosomal complex*	30/1535	0.000	0.014
GO:0031248	*protein acetyltransferase complex*	19/1535	0.000	0.014
GO:0005766	*primary lysosome*	25/1535	0.000	0.015
GO:0042582	*azurophil granule*	25/1535	0.000	0.015
GO:0098687	*chromosomal region*	49/1535	0.001	0.023
GO:0070603	*SWI/SNF superfamily-type complex*	17/1535	0.001	0.024
GO:0000922	*spindle pole*	26/1535	0.001	0.024
GO:0031201	*SNARE complex*	11/1535	0.001	0.027
GO:0061695	*transferase complex, transferring phosphorus-containing groups*	40/1535	0.001	0.027
GO:0000118	*histone deacetylase complex*	15/1535	0.001	0.027
GO:0005635	*nuclear envelope*	58/1535	0.001	0.027
GO:0005643	*nuclear pore*	17/1535	0.001	0.032
GO:0005770	*late endosome*	38/1535	0.002	0.042

A partir de los DEG identificados se llevó un enriquecimiento basado en la Ontología Génica para identificar procesos biológicos (Figura A8) y componentes celulares (Figura A9) significativamente afectados.

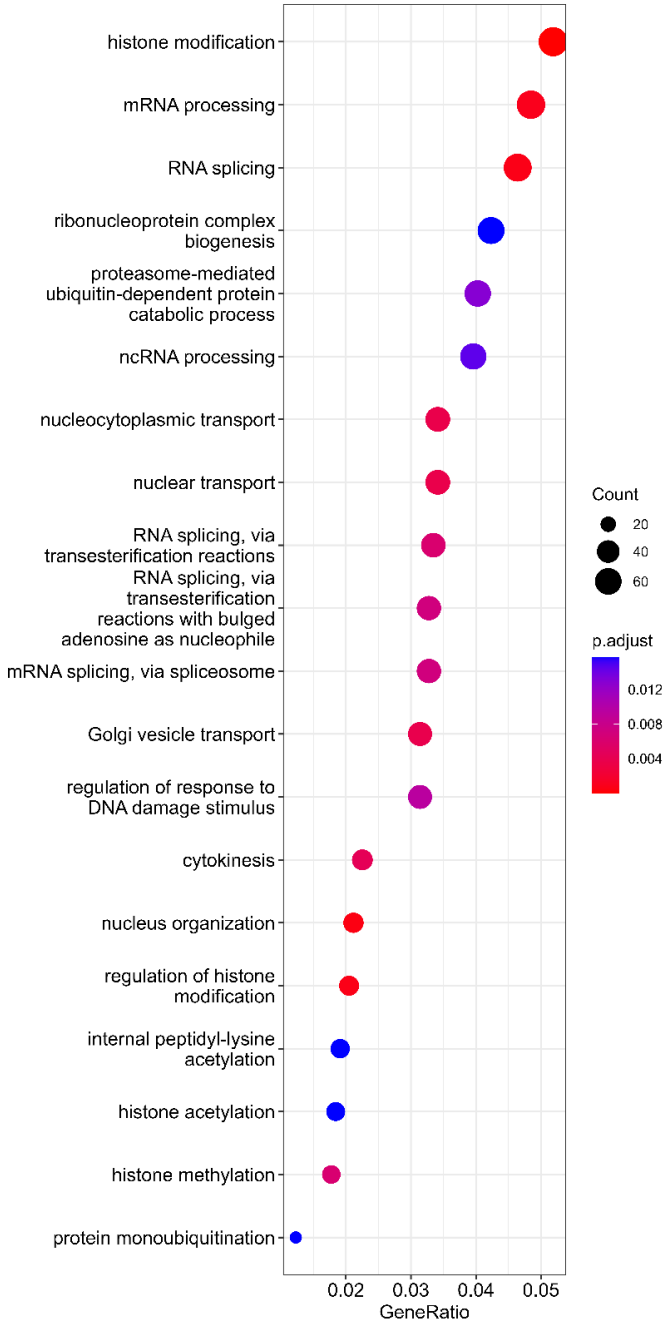

Figura A8. Top procesos biológicos significativos obtenidos a partir de los DEG identificados. Se indica el número de genes que contribuyen a cada término, el p-valor ajustado obtenido y la proporción de DEG que contribuyen al término entre los DEG totales.

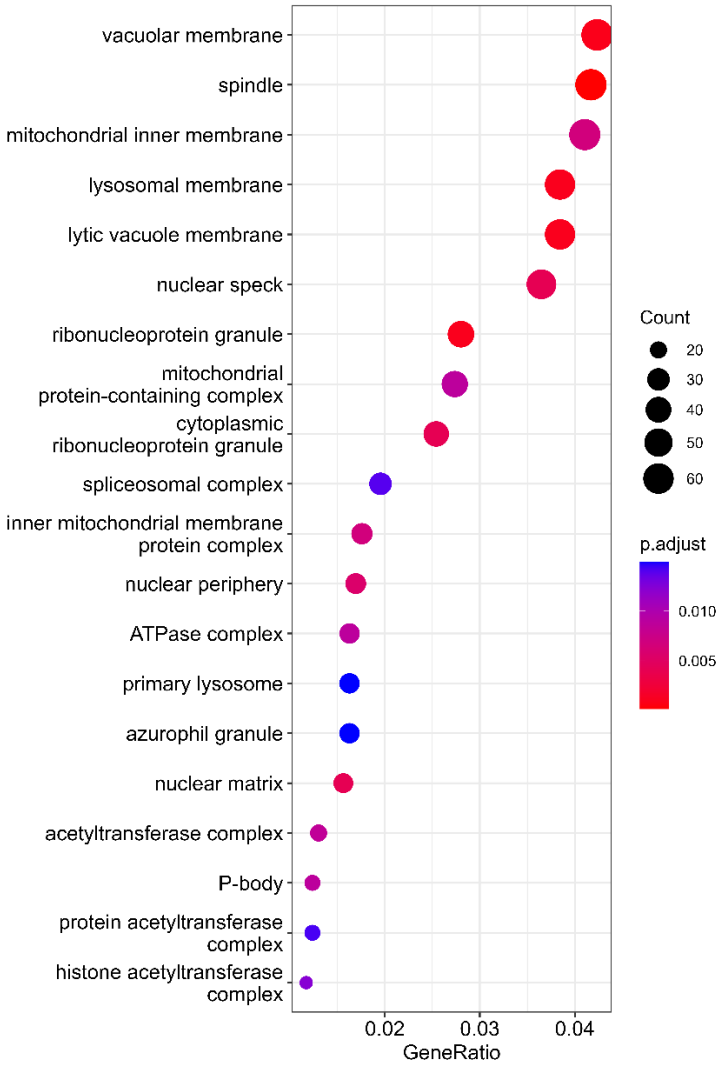

Figura A9. Top componentes celulares significativos obtenidos a partir de los DEG identificados. Se indica el número de genes que contribuyen a cada término, el p-valor ajustado obtenido y la proporción de DEG que contribuyen al término entre los DEG totales.

Anexo IV: Código utilizado

El código utilizado para este proyecto puede consultarse en el siguiente repositorio público de *GitHub*: https://github.com/guillepl/parkinson_tfm. Se distinguen los archivos con códigos *bash* (.sh) y los archivos con código de R (.R), así como el orden en el que deben utilizarse.