Universidad de Murcia
Facultad de Biología
Máster Universitario en Bioinformática

Trabajo Fin de Máster

Aplicación de aprendizaje automático supervisado para el modelado predictivo de patologías con cuerpos de Lewy a partir de perfiles de metilación del ADN

Autor: Carlos C. Ureña Mateo
Tutor académico: Dr. Juan Antonio Botía Blaya
Curso académico: 2024–2025

Índice

Índice de figuras

Índice de tablas

Lista de acrónimos

ADN ácido desoxirribonucleico. 1–6, 19

ARN ácido ribonucleico. 1, 2

CTRL controles. II, 6, 10, 11, 14–16, 18, 19

DLB demencia con cuerpos de Lewy. 3, 6, 10, 11, 14

PD enfermedad de Parkinson. II, 3, 4, 6, 10, 11, 14–16, 18–20

PDD Parkinson con demencia. 3, 6, 10, 11, 14, 15

RF *random forest*. 4, 5, 10, 11, 14–19

SVM máquinas de vectores soporte. II, 4, 5, 10–12, 14–20

1. Abstract

Lewy body diseases, including Parkinson's disease, Parkinson's disease dementia, and dementia with Lewy bodies, are neurodegenerative disorders that may share epigenetic alterations, particularly in DNA methylation profiles. In this study, we designed and implemented a reproducible and modular computational pipeline for the analysis of DNA methylation data, aiming to identify biologically meaningful and discriminative epigenetic patterns through the integration of clinically curated gene panels and supervised machine learning models.

To address the high dimensionality of methylation arrays, we applied a biologically informed filtering strategy based on 16 gene panels, including four related to neurodegeneration and several unrelated control panels. Classification models were trained using three algorithms (SVM with radial and linear kernels, and Random Forest), both at the global level (entire panel) and by functional genomic *loci* (e.g., promoters, TSS regions). Additional models incorporated clinical covariates (age and sex) to assess their potential confounding effects.

While global models did not yield significant results after multiple testing correction, the *locus*-specific analysis revealed statistically significant signals in promoter and TSS200 regions of the *Adult Neurodegenerative Disorders* panel, particularly in the Parkinson's disease vs Control (PD vs CTRL) contrast using radial SVM. In contrast, Random Forest models that included covariates detected significant signals only in negative control panels, suggesting that demographic imbalances, rather than disease-specific methylation differences, were driving those associations.

These results highlight the value of integrating biological knowledge and functional stratification in methylation studies and suggest that promoter regions may harbor epigenetic variation relevant to Parkinson's disease. Further investigation in larger, demographically balanced cohorts will be essential to validate these findings and clarify their clinical implications.

2. Introducción

2.1. Metilación del ADN como mecanismo epigenético

La epigenética es el campo que estudia los cambios heredables en la actividad génica que no implican alteraciones en la secuencia del ácido desoxirribonucleico (ADN) [1, 2]. A diferencia de las mutaciones, estos mecanismos no modifican el código genético en sí, sino que actúan como una capa reguladora adicional que modula la expresión de los genes a través de diversas modificaciones estructurales y químicas.

Entre los principales mecanismos epigenéticos se encuentran la metilación del ADN, las modificaciones postraduccionales de las histonas, la remodelación de la arquitectura nuclear, la pérdida de impronta y la acción de ácido ribonucleico (ARN) no codificante [1, 2]. Estos mecanismos interactúan dinámicamente para establecer patrones de expresión

génica específicos, que pueden verse modificados por factores ambientales y hábitos de vida, como la dieta o la exposición a toxinas, con posibles consecuencias funcionales a largo plazo [1].

La metilación del ADN es la modificación epigenética más ampliamente caracterizada en eucariotas [3], y constituye un eje fundamental en la regulación de la expresión génica. Este proceso implica la transferencia enzimática de un grupo metilo (-CH3) desde la molécula donadora S-adenosilmetionina (SAM) al carbono 5 del anillo de citosina, generando 5-metilcitosina (5-mC), una modificación covalente que ocurre mayoritariamente en dinucleótidos CpG [4, 5].

Las enzimas responsables de esta reacción pertenecen a la familia de las ADN metiltransferasas (DNMTs; DNMT1, DNMT2, DNMT3a y DNMT3b), que participan tanto en la síntesis *de novo* como en el mantenimiento de las marcas metiladas [6]. A este sistema se añaden proteínas lectoras, capaces de reconocer las marcas epigenéticas, así como enzimas demetilasas como TET1-3, que permiten la eliminación activa de 5-mC y otorgan a la metilación un carácter reversible y dinámico [4]. Aunque los patrones de metilación establecidos durante el desarrollo embrionario tienden a mantenerse a lo largo del tiempo, pueden modificarse en respuesta a señales internas y externas, adaptando así la expresión génica a las necesidades contextuales del organismo. Esta plasticidad funcional sitúa a la metilación del ADN como un mecanismo clave tanto en el desarrollo como en procesos fisiopatológicos como el envejecimiento o las enfermedades neurodegenerativas [4].

Funcionalmente, la metilación del ADN desempeña un papel central en la regulación epigenética, asociándose frecuentemente con la represión transcripcional, aunque su efecto depende del contexto genómico en el que se localice [5]. En regiones promotoras, especialmente en islas CpG, la metilación puede bloquear la unión de factores de transcripción o de la ARN polimerasa, impidiendo el inicio de la transcripción y conduciendo al silenciamiento génico [7]. Además, estas marcas pueden ser reconocidas por proteínas con dominios MBD, que reclutan complejos de remodelado de la cromatina y favorecen la compactación del ADN [5].

Estos mecanismos de represión resultan especialmente relevantes para el silenciamiento de elementos genómicos móviles. Cerca del 45 % del genoma de los mamíferos está compuesto por elementos transponibles o secuencias de origen viral, cuya actividad es reprimida por patrones estables de metilación, contribuyendo así a preservar la estabilidad e integridad del genoma [8].

No obstante, la metilación no siempre implica inhibición de la expresión génica. En regiones intragénicas, por ejemplo, puede modular el empalme alternativo (*splicing*) o participar en mecanismos de regulación postranscripcional [9], reflejando en estos casos una función más moduladora que represiva.

2.2. Enfermedades neurodegenerativas y metiloma

La enfermedad del Parkinson (PD) es la segunda enfermedad neurodegenerativa más común en la población anciana, afectando aproximadamente al 1–2 % de las personas mayores de 60 años [3, 10]. Desde el punto de vista clínico, se caracteriza por un inicio progresivo de síntomas motores como bradicinesia, temblor en reposo, rigidez muscular y alteraciones posturales, junto con manifestaciones no motoras frecuentes, como depresión, disfunción autonómica, deterioro cognitivo o trastornos del sueño [11, 12]. Neuropatológicamente, la PD se define por la pérdida progresiva de neuronas dopaminérgicas en la *sustantia nigra pars compacta*, aunque también se han documentado signos de neurodegeneración en la corteza cerebral y otras regiones encefálicas, lo que podría explicar los síntomas no motores en fases avanzadas de la enfermedad [10]. El hallazgo distintivo es la presencia de inclusiones intracelulares conocidas como cuerpos de Lewy, compuestos mayoritariamente por agregados de α-sinucleína, acompañados de neuritas de Lewy con distribución característica en el hipocampo y la *sustantia nigra* [3, 11].

La demencia con cuerpos de Lewy (DLB) y el Parkinson con demencia (PDD) comparten con la PD esta acumulación anómala de α-sinucleína en forma de cuerpos y neuritas de Lewy. Aunque las diferencias clínicas entre estas entidades se centran principalmente en el momento de aparición del deterioro cognitivo respecto a los síntomas motores, todas ellas comparten mecanismos fisiopatológicos convergentes, en particular alteraciones epigenéticas que afectan a la regulación del gen *SNCA*, que codifica la α-sinucleína [4, 13].

Numerosas investigaciones han puesto de manifiesto que las alteraciones en los patrones de metilación del ADN juegan un papel relevante en la fisiopatología de las enfermedades neurodegenerativas. La metilación aberrante de genes implicados en procesos clave como la homeostasis neuronal, la dinámica del citoesqueleto, la respuesta inflamatoria o los ritmos circadianos puede facilitar tanto la aparición como la progresión del daño neurodegenerativo. Como se ilustra en la Figura 1, la metilación de regiones promotoras puede conllevar el silenciamiento transcripcional de genes relevantes para la función neuronal, contribuyendo así a los procesos patológicos que subyacen a estas enfermedades [13].

En el caso de la PD, uno de los hallazgos más consistentes ha sido la hipometilación de una isla CpG situada en el intrón 1 del gen *SNCA*, observada en tejido cerebral postmortem de pacientes, lo cual se asocia con una sobreexpresión del gen y una acumulación tóxica de α-sinucleína. Esta alteración epigenética parece favorecer la formación de agregados proteicos y la posterior disfunción neuronal [14].

Por otro lado, se ha demostrado que la propia α-sinucleína puede interferir en el equilibrio epigenético de la célula al secuestrar la ADN metiltransferasa DNMT1 fuera del núcleo, impidiendo su función en el mantenimiento de la metilación [4]. Esta relocalización anómala provoca una hipometilación global en las neuronas y la sobreexpresión de genes, como *SNCA* y *CYP2E1*. Este último, al participar en rutas de biotransformación, puede generar metabolitos neurotóxicos que contribuyen a la degeneración de las neuronas dopaminérgicas. Este circuito de retroalimentación patológica entre la acumulación de α-sinucleína y la alteración epigenética refuerza la progresión de la enfermedad [4].

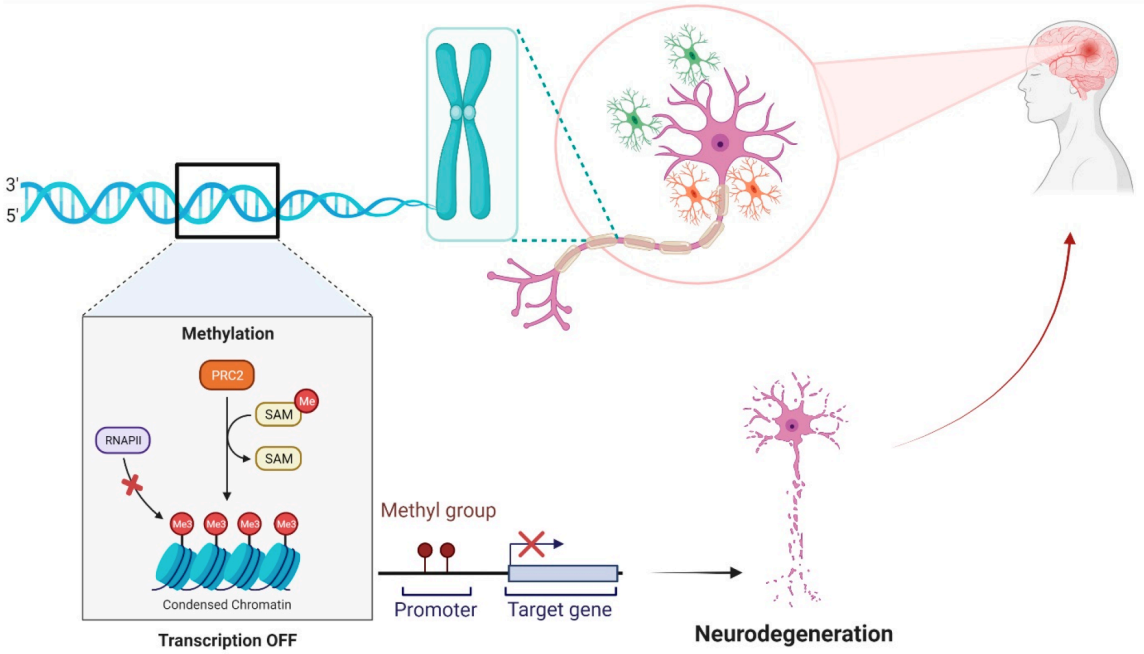

Figura 1: Esquema ilustrativo de cómo la metilación del ADN en regiones promotoras puede suprimir la expresión génica, contribuyendo a la desregulación neuronal y al desarrollo de procesos neurodegenerativos. Imagen adaptada de [13].

Además de *SNCA* y *CYP2E1*, se han identificado cambios de metilación en otros genes relevantes en PD. Por ejemplo, se ha observado hipermetilación en genes como *MAPT*, *MRI1* o *TMEM9*, e hipometilación en genes como *GSST1*, *TUBA3E* o *NOS2*, tanto en tejido cerebral como en sangre periférica [13]. Estos hallazgos respaldan el valor potencial de los perfiles de metilación como biomarcadores epigenéticos en el diagnóstico y seguimiento de la enfermedad.

En conjunto, la evidencia actual sugiere que las alteraciones epigenéticas, en particular, los cambios en la metilación del ADN, no representan un fenómeno secundario, sino un mecanismo central en la patogénesis de las enfermedades neurodegenerativas, entre las que se incluyen las neuropatologías asociadas a cuerpos de Lewy. Comprender estos patrones epigenéticos no solo aporta información sobre los mecanismos moleculares implicados, sino que también abre nuevas oportunidades para el desarrollo de herramientas diagnósticas.

2.3. Aplicaciones del *machine learning* en epigenética

El aprendizaje automático (*machine learning*) se utiliza cada vez más para analizar datos complejos en estudios biomédicos. En el caso de las enfermedades neurodegenerativas, permite detectar patrones en perfiles de metilación del ADN que no son evidentes mediante métodos estadísticos convencionales. En particular, los algoritmos de aprendizaje supervisado (Figura 2), como SVM y *random forest* (RF), aprenden a partir de datos etiquetados y se han aplicado con éxito en la clasificación de pacientes, la identificación de subtipos moleculares y la búsqueda de posibles biomarcadores [15].

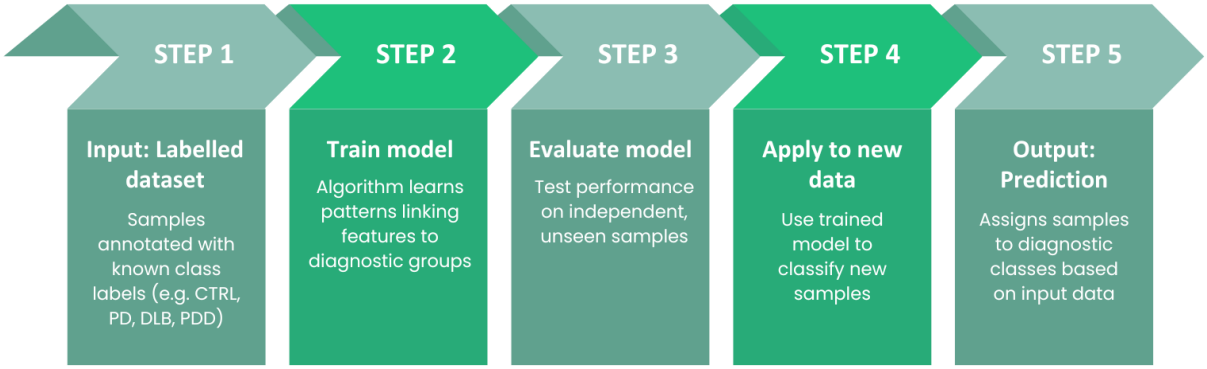

Figura 2: Esquema general del aprendizaje automático supervisado.

En este trabajo se pretende explorar si los perfiles de metilación del ADN contienen información relevante sobre la etiología de las enfermedades estudiadas. Para ello, se entrenan modelos de clasificación supervisada y se evalúa su capacidad para discriminar entre grupos clínicos (casos frente a controles). Si un modelo obtiene un rendimiento superior al esperado por azar, es decir, si clasifica mejor que un modelo aleatorio, se interpreta que existe señal en los datos. Esto sugiere que existen diferencias sistemáticas en los patrones de metilación entre los grupos, lo que podría reflejar alteraciones biológicamente significativas. Partiendo de esta idea se plantean los objetivos del trabajo.

3. Objetivos

Objetivo general

Diseñar e implementar un *pipeline* computacional reproducible y modular para el análisis de datos de metilación del ADN, orientado a la identificación de patrones epigenéticos con valor discriminativo mediante la integración de conocimiento biológico previo (paneles génicos clínicamente relevantes) y el uso complementario de algoritmos de clasificación supervisada (SVM y RF). Este enfoque se aplicará como caso de estudio a un conjunto de datos de metilación en muestras de pacientes con enfermedades neurodegenerativas relacionadas con cuerpos de Lewy y controles sanos.

Objetivos específicos

- Desarrollar una estrategia para la reducción de dimensionalidad basada en la selección informada de sondas de metilación asociadas a genes incluidos en paneles clínicamente relevantes.
- Diseñar un sistema automatizado para la construcción y validación de modelos de clasificación binaria en contextos de alta dimensionalidad y bajo número de muestras, incorporando múltiples algoritmos de aprendizaje supervisado.

- Implementar una partición funcional de las sondas por regiones genómicas específicas (*loci*) con el fin de aumentar la sensibilidad y resolución de los modelos en la detección de señales epigenéticas regionales.
- Integrar covariables clínicas (como sexo y edad) en el entrenamiento de los modelos para evaluar su impacto en la clasificación y controlar posibles factores de confusión.
- Establecer controles negativos basados en paneles no relacionados con las patologías del estudio para identificar y descartar posibles señales espurias derivadas del sesgo de dimensionalidad.

4. Materiales y métodos

4.1. Materiales

El conjunto de datos empleado en este estudio procede de un perfilado epigenético realizado sobre tejido cerebral humano post mortem, concretamente en la materia gris del córtex frontal obtenido a través del Netherlands Brain Bank (NBB). La cohorte está compuesta por 203 individuos distribuidos en cuatro grupos clínicos: 68 CTRL sin patología neurológica, 60 pacientes con PD, 60 con PDD y 15 con DLB. El diseño experimental y las características de la cohorte fueron descritos previamente por los autores del estudio [16].

Los niveles de metilación del ADN fueron cuantificados mediante la plataforma Illumina Infinium HumanMethylationEPIC BeadChip, una tecnología de alta densidad que permite medir más de 850.000 sitios CpG distribuidos a lo largo del genoma humano. Esta técnica proporciona valores beta (β) entre 0 y 1, que representan la proporción relativa de metilación en cada sitio CpG. Las sondas del array se encuentran anotadas funcionalmente en relación con elementos genómicos relevantes —como promotores, exones, regiones génicas, islas CpG, regiones UTR 5' y 3', y zonas intergénicas— siguiendo la anotación oficial del paquete *IlluminaHumanMethylationEPICanno.ilm10b4.hg19* de Bioconductor.

El preprocesamiento de los datos brutos fue realizado por los autores del estudio original, utilizando una combinación de herramientas del ecosistema Bioconductor: *minfi*, *ChAMP* y *wateRmelon*. Se aplicaron filtros de calidad para eliminar sondas poco fiables y muestras con métricas deficientes, además de una normalización de los valores beta mediante el método BMIQ.

Como resultado de este procesamiento, se generaron los archivos utilizados como punto de partida en el presente análisis. En particular, se empleó la matriz *LBDfcFilteredMyNorm.csv*, que contiene los valores beta normalizados para 722.934 sitios CpG en las 203 muestras, y el archivo de metadatos *LBDfcSamplesheet.csv*, que recoge información clínica y técnica asociada a cada muestra, incluyendo el grupo diagnóstico, edad, sexo y variables neuropatológicas adicionales.

Junto con los datos de metilación, el estudio se centró en un conjunto definido de paneles de genes asociados a enfermedades. En total se analizaron 16 paneles génicos,

clasificados en tres categorías: (1) paneles relacionados con patologías neurodegenerativas, que constituyen el núcleo clínico de interés del estudio; (2) paneles asociados a otras áreas biomédicas como el metabolismo, la hematología y la cardiología; y (3) paneles aleatorios utilizados como controles negativos internos.

Los cuatro paneles neurodegenerativos se obtuvieron directamente del recurso PanelApp de Genomics England (https://panelapp.genomicsengland.co.uk/panels/) y están específicamente vinculados con enfermedades del movimiento o del deterioro cognitivo: *Adult Dystonia*, *Adult Neurodegenerative Disorders*, *Early Onset Dementia* y *Parkinson Disease*.

Los nueve paneles biomédicos adicionales fueron seleccionados a partir de versiones curadas en formato CSV de PanelApp. Por último, tres paneles fueron generados aleatoriamente a partir de genes del array EPIC, asegurando que cada uno incluyera genes con al menos una sonda asociada.

Todas las funciones del *pipeline* desarrollado, están disponibles en el repositorio de GitHub asociado a este trabajo: https://github.com/carlos-um/TFM

4.2. Métodos

El análisis de metilación fue llevado a cabo mediante un *pipeline* bioinformático modular, diseñado para garantizar reproducibilidad y adaptabilidad a distintos conjuntos de datos epigenéticos. Su arquitectura se organiza en bloques secuenciales que pueden configurarse de forma flexible en función de la estrategia analítica empleada.

Como se muestra en la Figura 3, el *pipeline* original (Ruta 1), desarrollado previamente, sigue un enfoque exploratorio sin hipótesis explícita, basado en criterios estadísticos internos como la correlación entre sondas o el análisis de expresión diferencial. En el contexto del presente trabajo, se plantea una ruta alternativa (Ruta 2), orientada por hipótesis, cuya principal aportación consiste en incorporar conocimiento biomédico desde las etapas iniciales del análisis. Para ello, se utilizan paneles de genes clínicamente validados como hipótesis a priori sobre regiones genómicas relevantes para cada enfermedad.

Ambas estrategias parten de un problema común: la alta dimensionalidad de los datos de metilación, con 722.934 sitios CpG frente a solo 203 muestras. Además, el estudio incluye cuatro grupos (tres patológicos y uno control), lo que reduce aún más el tamaño de muestra efectivo en cada comparación por pares. Esta desproporción incrementa el riesgo de sobreajuste, dificulta la interpretación de los modelos y supone una carga computacional considerable. Por este motivo, ambas estrategias incorporan una etapa inicial de reducción de dimensionalidad, aunque mediante enfoques distintos.

En la Ruta 2, el análisis sigue una estrategia dividida en cuatro niveles. Primero, se filtran las sondas utilizando paneles de genes clínicamente validados para reducir la dimensionalidad y centrarse en regiones genómicas relevantes. Segundo, se entrenan modelos globales por panel y contraste, para comprobar si el conjunto completo de sondas de cada panel contiene señal. Tercero, las sondas se agrupan por regiones funcionales del gen (*loci*) y se entrenan modelos por separado para cada panel, región y contraste, con el

objetivo de detectar señales más localizadas. Cuarto, se añaden las variables edad y sexo a los modelos por *locus* para evaluar si estas covariables influyen en los resultados.

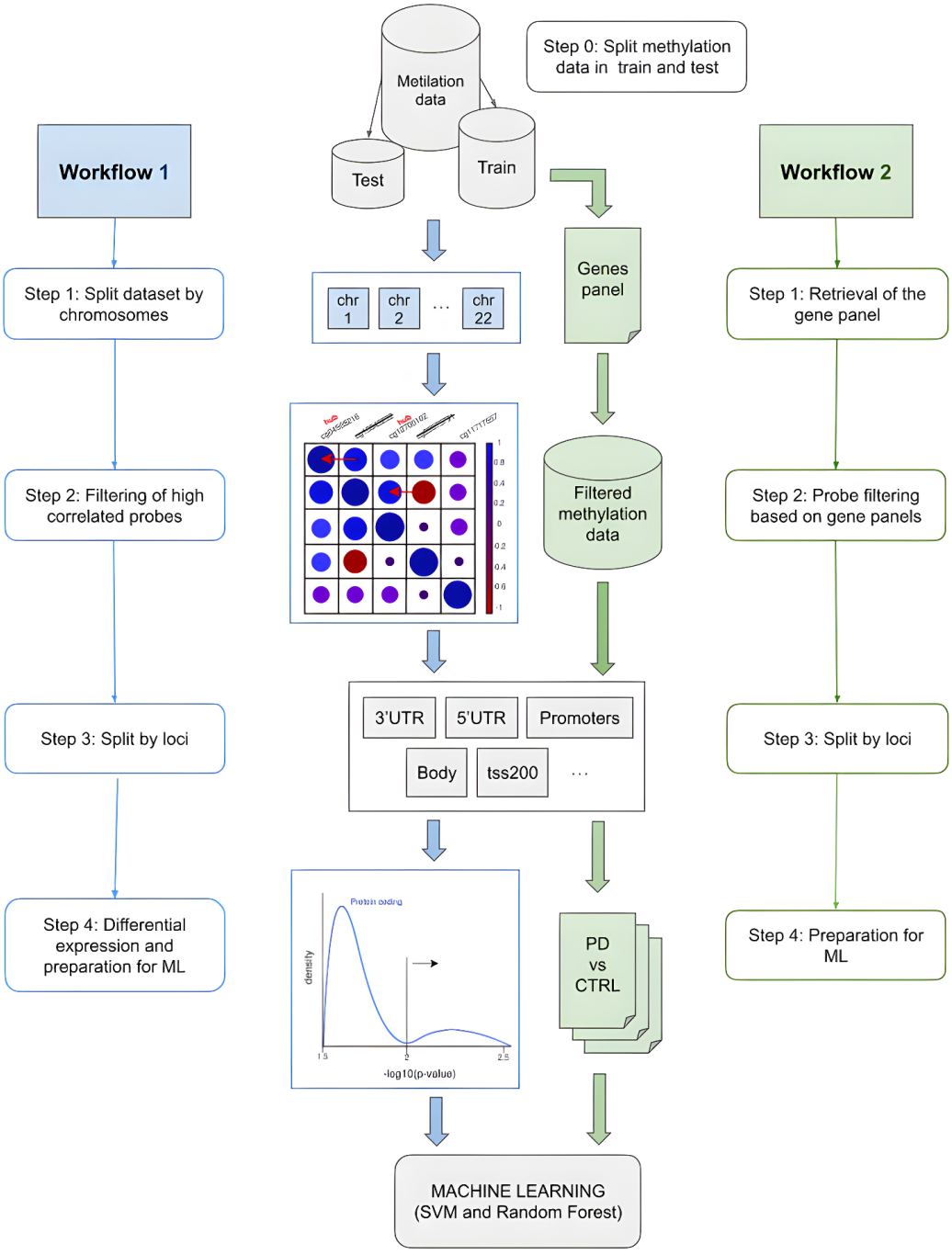

Figura 3: *Pipeline* de análisis de metilación dividido en dos enfoques: uno basado en el genoma completo mediante métodos estadísticos (*Workflow 1*, azul) y otro basado en paneles génicos clínicamente seleccionados (*Workflow 2*, verde). Ambos flujos comparten etapas comunes como la partición por *loci* funcionales y el entrenamiento de modelos de clasificación, pero difieren en las fases iniciales de selección de sondas.

A continuación se describen en detalle los pasos de la estrategia de análisis seguida en el presente trabajo.

4.2.1. Bloque preliminar. Preparación del conjunto de datos y paneles génicos

Antes de aplicar la estrategia en niveles definida para la Ruta 2, se llevó a cabo una fase preliminar que incluyó dos tareas fundamentales: (1) la partición de los datos de metilación en conjuntos de entrenamiento y validación, y (2) la obtención de listas de genes clínicamente relevantes que servirían como base para el filtrado dirigido de sondas.

En primer lugar, el conjunto de datos normalizado de metilación (*LBDfcFilteredMyNorm.csv*) y su correspondiente archivo de metadatos clínicos (*LBDfcSamplesheet.csv*) fueron divididos en dos subconjuntos independientes: uno destinado al entrenamiento de los modelos (*train*) y otro reservado para su evaluación sobre datos no vistos (*test*). Esta partición se realizó en proporción 70/30, aplicando estratificación por grupo diagnóstico para preservar la distribución de clases. Se utilizó una semilla aleatoria fija para asegurar la reproducibilidad del particionado en ejecuciones posteriores.

Paralelamente, se generaron listas de genes con alta evidencia de asociación a enfermedades, que sirvieron como base para el filtrado dirigido de sondas. Estas listas se obtuvieron a partir de dos fuentes complementarias: (1) archivos TSV descargados directamente desde el portal PanelApp de Genomics England, y (2) versiones preprocesadas en formato CSV, correspondientes a paneles seleccionados y curados manualmente. Para procesar cada tipo de archivo se emplearon las funciones *get_green_genes_from_genomics_england()* y *get_high_evidence_genes_from_csv()*, respectivamente, las cuales filtran genes en función de su nivel de evidencia clínica. Se retuvieron únicamente aquellos anotados como *Expert Review Green* o *HighEvidence*, siguiendo la nomenclatura oficial de PanelApp.

El resultado de esta fase fue un conjunto estructurado compuesto por: (1) las matrices de metilación separadas en *train* y *test*, y (2) vectores de símbolos génicos de alta calidad que sirvieron como entrada para el primer nivel de análisis.

4.2.2. Nivel 1: Filtrado dirigido por paneles de genes

Una vez definidos los paneles génicos de interés, se procedió a identificar qué sondas del array de metilación se localizaban dentro de las regiones genómicas correspondientes a los genes seleccionados.

Este procedimiento se implementó mediante la función *filter_methylation_by_genes()*, que toma como entrada un vector de símbolos génicos, junto con los archivos de anotación de sondas y matriz de metilación *train*. En primer lugar, la función obtiene las coordenadas genómicas asociadas a los genes de interés utilizando el paquete *biomaRt* y el recurso Ensembl, asumiendo el ensamblado GRCh37/hg19. Estas coordenadas se transforman al formato BED, estándar para operaciones genómicas.

Simultáneamente, las posiciones de las sondas EPIC se extraen del archivo de anotaciones oficial y se formatean igualmente como un archivo BED. Ambos conjuntos (genes y sondas) se intersectan mediante la herramienta externa *bedtools*, lo que permite identificar

las sondas que se solapan espacialmente con al menos una región génica. A partir de dicha intersección, se extraen los identificadores únicos de las sondas relevantes, que se utilizan para filtrar la matriz de metilación original. El resultado es una nueva matriz con menor número de variables (sondas), que se guarda en disco para su uso en las siguientes etapas del *pipeline*.

Además del archivo de salida, la función proporciona estadísticas resumidas que permiten monitorizar la eficacia del filtrado, como el número de genes introducidos, cuántos se mapearon exitosamente a coordenadas genómicas, y cuántas sondas fueron finalmente retenidas.

4.2.3. Nivel 2. Modelado global por panel génico

Una vez obtenidas las matrices de metilación filtradas por paneles de genes, se procedió al entrenamiento de modelos de clasificación supervisada utilizando todas las sondas asociadas a cada panel, sin distinción por regiones funcionales. El objetivo de este segundo nivel fue evaluar la capacidad discriminativa del conjunto completo de sondas por cada panel clínico por separado, para distinguir entre sujetos control y patologías.

El procedimiento se estructuró en dos etapas consecutivas. En primer lugar, la función *prepare_for_ml_global()* generó matrices de entrenamiento y prueba específicamente diseñadas para cada contraste clínico binario definido: *PD vs CTRL*, *PDD vs CTRL*, *DLB vs CTRL* y *neuro vs CTRL*. Este último contraste agrupa todas las muestras patológicas (PD, PDD y DLB) frente a los controles, y se emplea con el objetivo de evaluar si el incremento en el tamaño muestral permite detectar patrones discriminativos generales que podrían no ser visibles al analizar cada condición por separado. La función toma como entrada la matriz de entrenamiento previamente filtrada por genes, la matriz de test global normalizada y los archivos de anotaciones clínicas, asegurando que las matrices generadas (*train* y *test*) por contraste contengan el mismo conjunto de sondas y que las muestras estén correctamente etiquetadas.

Una vez generadas las matrices, se entrenaron modelos de clasificación binaria mediante la función *train_models_all()*. En esta etapa se aplicaron tres algoritmos de forma sistemática: SVM con núcleos lineal y radial, y RF sin inclusión de covariables clínicas. Aunque la función permite incorporar variables como edad y sexo, en el presente nivel se optó por evaluar el rendimiento de los paneles sin añadir factores adicionales.

El entrenamiento se realizó aplicando validación cruzada estratificada de cinco particiones sobre el conjunto de entrenamiento, con ajuste automático de hiperparámetros utilizando *caret::train()* y la métrica *ROC* como criterio de optimización. La evaluación de los modelos se llevó a cabo sobre el conjunto de test independiente, y se calcularon métricas clave como la exactitud (*Accuracy*), exactitud balanceada (*Balanced Accuracy*), índice de concordancia (*Kappa*), área bajo la curva ROC (AUC), *p-value*, *p-value* del test de McNemar y tasa de acierto aleatoria (*No-Information Rate*).

Los resultados obtenidos fueron recopilados en un archivo resumen estructurado por panel y contraste clínico.

4.2.4. Nivel 3. Modelado por regiones funcionales del genoma (*loci*)

Con el objetivo de aumentar la resolución funcional del análisis y explorar si determinadas regiones génicas concentran señal epigenética discriminativa, se implementó un tercer nivel basado en el modelado específico por *loci* funcionales. Este enfoque permite identificar patrones de metilación relevantes que podrían quedar enmascarados en el análisis global, al focalizar el modelado en subconjuntos funcionales de los genes.

La función *split_methylation_by_loci()* permitió la segmentación funcional de las sondas filtradas, clasificandolas por panel según las variables *UCSC_RefGene_Group* y *Regulatory_Feature_Group* del paquete de anotación *IlluminaHumanMethylationEPICanno.ilm10b4.hg19*. A partir de esta información, se definieron ocho regiones funcionales: *promoter*, *TSS200*, *TSS1500*, *Body*, *1stExon*, *5'UTR*, *3'UTR* y *ExonBnd*. Para cada categoría se generó un subconjunto independiente de la matriz de metilación, el cual fue almacenado en subdirectorios organizados por región y panel.

Adicionalmente, esta función exporta los identificadores de sondas por región funcional para su uso en análisis posteriores (por ejemplo, diagramas de Venn), y calcula estadísticas clave mediante la llamada interna a la función *save_summary_statistics()*, que cuantifica el número de sondas y genes únicos representados en cada región. Estos datos se guardan en un archivo resumen estandarizado en formato CSV.

Una vez generados los subconjuntos por *locus*, se construyeron las matrices de entrenamiento y prueba para cada región funcional utilizando la función *prepare_for_ml_loci()*. Esta función replica la lógica del nivel anterior, pero adaptada a cada combinación de panel, región funcional y contraste clínico binario (*PD vs CTRL*, *PDD vs CTRL*, *DLB vs CTRL* y *neuro vs CTRL*). Se garantizaron controles de calidad internos para asegurar la coherencia entre las matrices y las etiquetas clínicas, así como la reproducibilidad estructural de los archivos generados.

El entrenamiento de los modelos se llevó a cabo mediante la función *train_models_loci()*, que aplica los mismos algoritmos y esquemas de validación descritos en el nivel anterior. Para cada combinación de panel, contraste y *locus*, se entrenaron modelos independientes utilizando SVM (núcleo lineal y radial) y RF sin inclusión de variables clínicas adicionales. Esta aproximación permitió evaluar el poder discriminativo intrínseco de cada *locus* funcional, identificando patrones localizados de interés biológico.

Los resultados fueron almacenados en una jerarquía estructurada de archivos por panel, *locus* y contraste.

4.2.5. Nivel 4. Incorporación de covariables clínicas: edad y sexo

En este nivel se evaluó el efecto de las variables clínicas edad y sexo sobre el rendimiento de los modelos entrenados por *locus* funcional. El objetivo fue comprobar si las señales obtenidas en los análisis previos en los que no se tuvieron en cuenta covariables, podían explicarse en parte por estos factores, y si era posible mejorar la interpretabilidad de los modelos al tenerlos en cuenta.

Las covariables se incluyeron únicamente en los modelos entrenados con *ranger*, ya que este algoritmo permite controlar explícitamente por variables como edad o sexo mediante el parámetro *always.split.variables*. Este parámetro fuerza que las variables se utilicen en las divisiones de los árboles, garantizando que su efecto esté presente en el modelo, independientemente del número de predictores o la colinealidad. En cambio, SVM trata todas las variables por igual y no dispone de un mecanismo para asegurar el uso de una covariable concreta. En contextos de alta dimensionalidad y colinealidad, el efecto de variables como edad o sexo puede diluirse y no quedar representado en el modelo.

Para entrenar los modelos se utilizaron las matrices de entrenamiento y prueba del Nivel 3 añadiendo las columnas de edad y sexo. El resto del procedimiento se mantuvo idéntico, lo que permitió comparar directamente los modelos con y sin ajuste por covariables. Los resultados se guardaron en archivos separados por panel, contraste, *locus* funcional y configuración de variables.

5. Resultados y discusión

5.1. Filtrado de sondas basado en paneles génicos

La matriz original de metilación contenía 722.934 sondas, un número muy alto en relación con las muestras disponibles. Para poder manejar estos datos, se aplicó un filtrado previo conservando solo las sondas asociadas a genes incluidos en paneles clínicos. Aunque esta estrategia puede hacer que se pierdan algunas señales importantes, permite reducir el ruido y centrarse en un subconjunto más manejable de información, lo que facilita detectar posibles patrones.

La Tabla 1 resume el resultado de este filtrado para cada panel génico, indicando el número total de genes, cuántos se pudieron mapear a coordenadas válidas y cuántas sondas asociadas se retuvieron para el análisis.

Tabla 1: Resumen del número total de genes para cada panel, los genes mapeados y las sondas seleccionadas tras el filtrado.

Panel	**Genes iniciales**	**Genes mapeados**	**Sondas**
Neurodegenerativas			
Parkinson Disease	45	42	1814
Adult Neurodeg. Disorders	126	118	3202
Adult Dystonia	76	71	2463
Early Onset Dementia	26	24	1059
Otras biomédicas			
Combined BT cell Defect	19	17	453
Cytopaenias and Anaemias	89	80	1738
Dilated Cardiomyopathy	32	29	749
Long QT Syndrome	10	10	962
Mitochondrial Disorders	198	146	3008
Mucopolysaccharidosis	17	14	349
Peroxisomal Disorders	34	30	495
Thoracic Aortic Aneurysm	24	22	889
SCID	17	16	356
Aleatorios			
Random Panel-15	15	13	617
Random Panel-30	30	23	442
Random Panel-60	60	50	1670

El número final de sondas por panel osciló entre 349 y 3202, lo que se considera manejable para entrenar modelos de clasificación. Además, desde un punto de vista cualitativo, centrarse en genes con implicación biomédica facilita la interpretación de los resultados obtenidos en etapas posteriores del *pipeline* analítico.

Como análisis complementario, se evaluó el solapamiento entre genes de los distintos paneles utilizando el índice de Jaccard, definido como el cociente entre la intersección y la unión de dos conjuntos. La Figura 4 muestra un mapa de calor jerarquizado con los coeficientes de similitud entre todos los pares de paneles.

Los paneles relacionados con enfermedades neurodegenerativas comparten un número considerable de genes, algo esperable por su proximidad etiológica. En cambio, los paneles de otras categorías clínicas muestran muy poca intersección entre sí y con los neurodegenerativos, lo que indica una buena especificidad en la selección de genes.

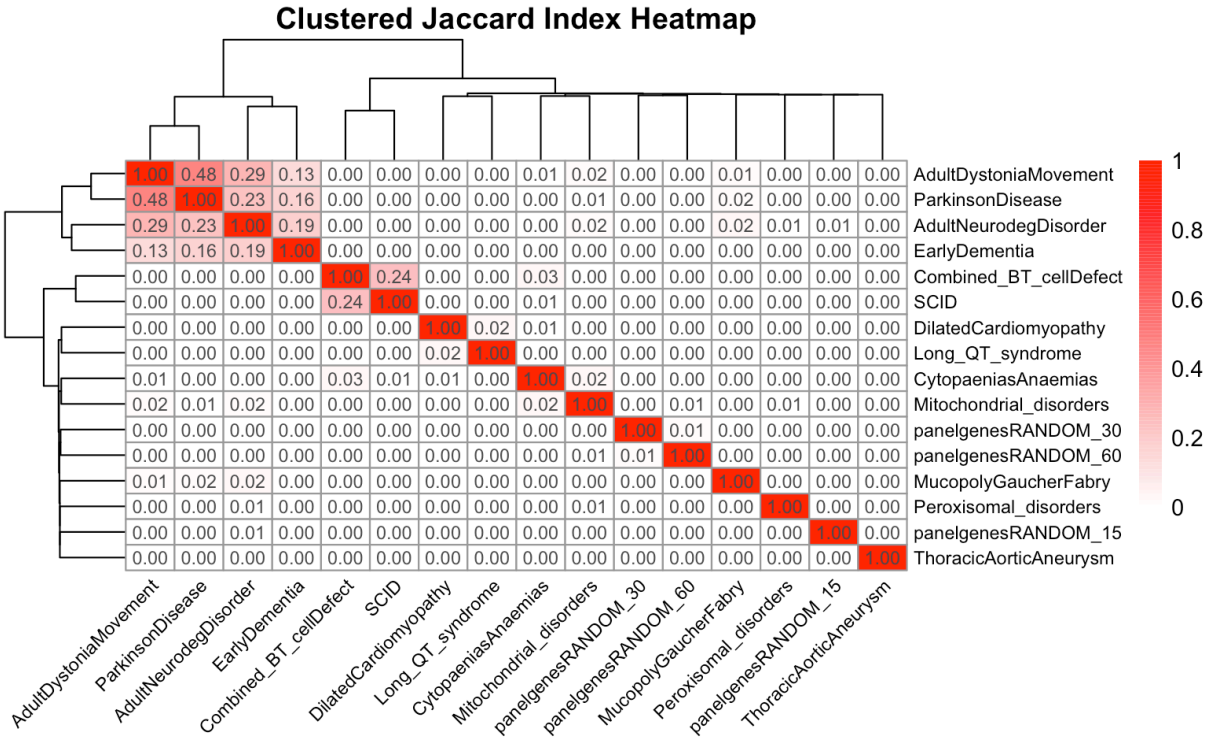

Figura 4: *Heatmap* jerarquizado del índice de Jaccard entre paneles génicos. Se representa el grado de solapamiento entre los conjuntos de genes de cada panel, con una escala de color que va de blanco (sin solapamiento) a rojo (máximo solapamiento).

Estos resultados respaldan la utilidad del enfoque basado en paneles, que permite reducir la dimensionalidad del problema de forma informada. A diferencia de métodos puramente estadísticos, como el filtrado por correlación o el análisis de expresión diferencial, esta estrategia se apoya en conocimiento previo y puede complementar otros enfoques al centrarse en regiones con relevancia biológica para el problema estudiado.

5.2. Evaluación global de señales por panel de genes

Tras aplicar el filtrado por paneles génicos, se evaluó la capacidad discriminativa del conjunto de sondas de cada panel por separado. Para ello, se entrenaron modelos de clasificación para cada combinación de panel y contraste, utilizando tres algoritmos distintos: SVM con kernel radial, SVM lineal y RF. En total, se generaron 192 modelos (16 paneles x 4 contrastes x 3 algoritmos), abarcando los 16 paneles seleccionados en el estudio y los cuatro contrastes clínicos definidos (DLB vs CTRL, PD vs CTRL, PDD vs CTRL, y neuro vs CTRL). Debido a la cantidad de modelos generados en el estudio, las tablas de los resultados pueden consultarse en el repositorio de GitHub (ver en Materiales) con los nombres: Resultados_SVM(RadialyLineal)_All.pdf y Resultados_RF_All.pdf

A los resultados de los modelos se les aplicó una corrección por testado múltiple mediante el método de Bonferroni. Este procedimiento ajusta los *p-values* multiplicándolos por el número de tests realizados dentro de cada tabla de resultados. De este modo,

se obtiene una columna con *p-values* ajustados, que permite interpretar los resultados controlando el riesgo de falsos positivos.

Tras el ajuste, ningún modelo alcanzó significación estadística. No obstante, se observaron algunos *p-values* nominalmente bajos (sin corregir), especialmente en los contrastes que involucran a los grupos PD y PDD. Esta tendencia se repite en distintos paneles, tanto neurológicos como no neurológicos, incluyendo también paneles aleatorios (controles negativos), algo esperable dado el número de pruebas realizadas.

A modo de ejemplo, la Tabla 2 muestra los resultados obtenidos con SVM (kernel radial y lineal) para el panel *EarlyDementia*. Aunque ninguno de los modelos resulta significativo tras la corrección, algunos *p-values* sin ajustar se sitúan por debajo de 0.1, especialmente en el contraste PDD vs CTRL.

Tabla 2: Resultados del modelo SVM (kernel lineal y radial) para el panel *EarlyDementia* entrenado con todo el conjunto de sondas del panel.

Contraste	**Kernel**	***p*-value**	***p*-value ajustado**
PDD vs CTRL	Radial	0.0709	0.5673
PD vs CTRL	Radial	0.0709	0.5673
DLB vs CTRL	Linear	0.6167	1.0000
NEURO vs CTRL	Linear	0.6634	1.0000
PDD vs CTRL	Linear	0.3141	1.0000
PD vs CTRL	Linear	0.3141	1.0000
DLB vs CTRL	Radial	0.6167	1.0000
NEURO vs CTRL	Radial	0.4520	1.0000

La Tabla 3 muestra los resultados de RF para el panel *ParkinsonDisease*. En este caso tampoco se observan resultados significativos, pero el contraste PDD vs CTRL vuelve a mostrar los *p-values* más bajos.

Tabla 3: Resultados del modelo *Random Forest* para el panel *ParkinsonDisease* entrenado con todo el conjunto de sondas del panel.

Contraste	***p*-value**	***p*-value ajustado**
PDD vs CTRL	0.2088	0.8352
DLB vs CTRL	0.6167	1.0000
NEURO vs CTRL	0.5603	1.0000
PD vs CTRL	0.3141	1.0000

A partir de estos resultados, se planteó la hipótesis de que las señales podrían estar diluidas al analizar todas las sondas de un panel como un conjunto, sin tener en cuenta su localización funcional. Por esta razón, se propuso un análisis más detallado mediante la partición de sondas según regiones funcionales del gen (loci), con el objetivo de identificar

posibles focos de señal más localizada y facilitar la interpretación biológica de los patrones detectados.

5.3. Segmentación funcional por *loci*: refinamiento de señales

Para cada combinación de panel, *locus* y contraste se entrenaron modelos binarios utilizando los mismos criterios aplicados en el entrenamiento global. En total, se generaron 1536 modelos (16 paneles x 8 loci x 4 contrastes x 3 algoritmos). Al igual que en el caso anterior, las tablas de los resultados pueden consultarse en el repositorio de GitHub con los nombres: Resultados_RF_loci_noconfounders.pdf, Resultados_SVM_Radial_loci.pdf y Resultados_SVM_Lineal_loci.pdf.

Aunque los modelos RF obtuvieron resultados similares a SVM en los análisis globales y ofrecen mayor interpretabilidad por su capacidad de identificar qué sondas podrían estar afectadas por variables confusoras en fases posteriores, se mantuvieron también modelos SVM para aumentar la sensibilidad en la detección de patrones de metilación diferencial.

A diferencia del análisis global, esta aproximación permitió identificar señales estadísticamente significativas tras corrección por testado múltiple, concretamente en el panel *AdultNeurodegenerativeDisorder*, en el contraste PD vs CTRL, empleando SVM con *kernel* radial. La Tabla 4 recoge los resultados más destacados, que corresponden a dos regiones funcionales con papel en la regulación de la transcripción: *promoter* y *TSS200*.

Tabla 4: Resultados más destacados de los modelos SVM radial por regiones funcionales en el panel *AdultNeurodegenerativeDisorder*. Se destaca el contraste PD vs CTRL en regiones promotoras con significancia ajustada.

Contraste	**Región funcional**	***p*-value**	***p*-value ajustado**
PD vs CTRL	promoter	0.0000465	0.0015
PD vs CTRL	TSS200	0.0002007	0.0064

Estos resultados refuerzan la hipótesis de que las señales relevantes estaban diluidas al considerar conjuntamente todas las sondas del panel, y se manifiestan al focalizar el análisis en regiones funcionales específicas. En particular, SVM radial parece capturar estructuras no lineales en los datos que no son detectadas por SVM lineal ni por RF, lo que justifica su uso como herramienta especialmente sensible en este tipo de análisis.

5.4. Evaluación del efecto de variables confusoras en los modelos *Random Forest*

En este punto del análisis nos cuestionamos si las señales obtenidas en el panel *AdultNeurodegenerativeDisorder* con SVM radial podrían estar afectadas por variables de confusión. Aunque SVM radial no incorpora explícitamente covariables, en el caso de que exista un desequilibrio en la distribución de edad o sexo entre los grupos, el modelo puede

estar captando indirectamente esa señal reflejada en la metilación (por ejemplo, si ciertas sondas están asociadas a edad o sexo).

Por esta razón, se amplió el análisis mediante la inclusión de sexo y edad como variables confusoras en los modelos RF entrenados por *locus*. Aunque RF sin *confounders* no alcanzó significación estadística, en este contexto su inclusión permite explorar el efecto de covariables clínicas, ya que, como se comenta en el apartado de la Metodología, SVM trata todas las variables por igual y no dispone de un mecanismo para asegurar el uso de una covariable concreta. Se generaron un total de 1024 modelos (16 paneles x 8 loci x 4 contrastes x 2 configuraciones con covariable). Los resultados se encuentran en el repositorio de GitHub con los nombres: *Resultados_RF_loci_Age.pdf* y *Resultados_RF_loci_Sex.pdf*.

En una primera fase, los modelos RF sin variables confusoras no detectaron señales estadísticamente significativas tras corrección por testado múltiple. Sin embargo, al introducir sexo o edad como covariables, emergieron señales significativas en varios paneles.

Es llamativo que estas señales aparezcan únicamente en paneles considerados controles negativos, es decir, conjuntos de genes sin asociación conocida con patologías neurodegenerativas. Por ejemplo, en el panel *Combined_BT_cellDefect*, al incluír sexo como covariable se obtuvo un *p*-value ajustado significativo para el contraste PD vs CTRL en la región UTR-5' (Tabla 5). Mientras que en el panel *Peroxisomal_disorders*, al incorporar la edad se identificó una señal significativa en el contraste PDD vs CTRL para la región TSS1500 (Tabla 6).

Tabla 5: Ejemplo de resultados más destacados de *Random Forest* por *loci* para el panel *Combined_BT_cellDefect* incluyendo sexo como variable confusora.

Contraste	**Región**	***p*-value**	***p*-value ajustado**
PD vs CTRL	utr5	0.0002	0.0064
PDD vs CTRL	promoter	0.0007	0.0236

Tabla 6: Ejemplo de resultados más destacados de *Random Forest* por *loci* para el panel *Peroxisomal_disorders* incluyendo edad como variable confusora.

Contraste	**Región**	***p*-value**	***p*-value ajustado**
PDD vs CTRL	tss1500	0.0007	0.0236

Esto sugiere que la inclusión de las variables demográficas en los modelos permite identificar diferencias que no reflejan alteraciones genuinas en la metilación asociadas al fenotipo clínico, sino que probablemente corresponden a distribuciones desbalanceadas de edad o sexo entre los grupos comparados.

Para evaluar esta posibilidad, se analizaron estadísticamente las diferencias de edad y sexo entre casos y controles en los tres contrastes clínicos principales (PD, PDD y DLB vs CTRL). Dado que se trata de variables de naturaleza distinta, se aplicaron pruebas

estadísticas específicas: un test de Wilcoxon para comparar las edades (variable continua no paramétrica), y un test exacto de Fisher para evaluar las proporciones de sexo (variable categórica binaria).

Los resultados, recogidos en las Tablas 7 y 8, muestran diferencias significativas en ambas variables. En concreto, los grupos de pacientes presentan medianas de edad más bajas y una proporción sustancialmente mayor de varones en comparación con los controles.

Tabla 7: Evaluación del sesgo por edad entre grupos clínicos mediante test de Wilcoxon.

Contraste	**Edad mediana (casos)**	**Edad mediana (controles)**	***p*-value**
PD vs CTRL	77.0	83.5	0.0063
PDD vs CTRL	78.0	83.5	0.0120
DLB vs CTRL	74.5	83.5	0.0577

Tabla 8: Evaluación del sesgo por sexo entre grupos clínicos mediante test de Fisher.

Contraste	**% Hombres (casos)**	**% Hombres (controles)**	***p*-value**
PD vs CTRL	55.9	27.9	0.0020
PDD vs CTRL	71.7	27.9	1.2e-06
DLB vs CTRL	81.2	27.9	1.3e-04

Por lo tanto, se confirma la existencia de un desequilibrio significativo en la distribución de edad y sexo entre los grupos clínicos, lo que apoya la hipótesis de que las señales obtenidas en los modelos al incluir estas variables podrían reflejar artefactos derivados de variables de confusión, más que asociaciones epigenéticas genuinas.

Por otro lado, recordemos que en el panel *AdultNeurodegenerativeDisorder* se identificó una señal significativa en el contraste PD vs CTRL al emplear SVM con *kernel* radial en dos regiones funcionales: *promoter* y *TSS200*. Sin embargo, los modelos RF, tanto con como sin la inclusión de covariables (edad y sexo), no mostraron un rendimiento significativo en este panel. Esto impide evaluar la importancia relativa de estas covariables y no permite determinar si la señal observada con SVM radial podría estar influida por factores de confusión. Además, dado que RF y SVM radial son algoritmos con características diferentes (uno basado en árboles de decisión y el otro en funciones kernel), la falta de señal en RF no puede considerarse una validación ni una refutación directa del resultado obtenido con SVM. Por tanto, la señal detectada en el contraste PD vs CTRL queda abierta la posibilidad de que refleje patrones reales de metilación diferencial que podrían ser objeto de estudio en trabajos futuros.

A partir de estos resultados, se proponen varias líneas de mejora. Como siguiente paso, convendría revisar las regiones funcionales utilizadas (como *promoter* o *TSS*) y considerar la inclusión de otras, como *enhancers* o regiones intergénicas, especialmente en contextos reguladores complejos. También sería útil comparar el rendimiento del enfoque basado

en aprendizaje automático con métodos clásicos de análisis de expresión diferencial, para entender qué aporta cada uno. Por otro lado, sería iteresante explorar la posibilidad de adaptar modelos SVM que permitan incluir covariables como sexo o edad, de forma similar a lo que ya hace RF. Estas mejoras permitirían evaluar mejor las señales detectadas y aumentar la interpretabilidad del análisis.

6. Conclusiones

Este trabajo tenía como objetivo general el diseño y aplicación de un *pipeline* modular y reproducible para el análisis de datos de metilación del ADN, orientado a la identificación de patrones epigenéticos con valor discriminativo en muestras de pacientes con enfermedades neurodegenerativas relacionadas con cuerpos de Lewy. Para alcanzar este objetivo, se plantearon una serie de objetivos específicos cuya evaluación permite extraer las siguientes conclusiones:

1. **Reducción de dimensionalidad mediante paneles de genes.** Se logró la reducción de la dimensionalidad seleccionando únicamente las sondas asociadas a genes incluidos en paneles clínicamente relevantes. Este filtrado, basado en conocimiento previo, permitió disminuir el número de predictores sin comprometer la relevancia biológica de los datos analizados. En comparación con enfoques puramente estadísticos, esta estrategia ofrece una alternativa informada que mejora tanto la eficiencia computacional como la interpretabilidad de los modelos.

2. **Implementación de modelos supervisados en entornos de alta dimensionalidad.** Se desarrolló y ejecutó un sistema automatizado para el entrenamiento y validación de modelos de clasificación binaria, empleando múltiples algoritmos: RF, SVM lineal y radial. Aunque los análisis globales no revelaron señales significativas, el sistema permitió una evaluación sistemática y comparativa de cada panel y contraste clínico.

3. **Segmentación funcional por loci.** La incorporación de una partición funcional de las sondas por regiones genómicas específicas (loci) permitió aumentar la resolución del análisis y detectar señales localizadas. En particular, se identificaron señales estadísticamente significativas en regiones *promoter* y *TSS200* del panel *AdultNeurodegenerativeDisorder* mediante SVM con kernel radial para el contraste PD vs CTRL, lo que sugiere una naturaleza no lineal de los patrones diferenciales de metilación y confirma la utilizad de focalizar el análisis en regiones funcionales del gen.

4. **Evaluación del impacto de covariables clínicas.** La inclusión de sexo y edad como covariables en RF evidenció su efecto en la clasificación. Aunque no se observó señal en el panel *AdultNeurodegenerativeDisorder*, donde SVM radial sí la había detectado, aparecieron asociaciones significativas en paneles de control negativo debido al desequilibrio en la distribucion de edad y sexo entre los grupos clínicos. Esto indica

que el desequilibrio demográfico entre grupos puede generar señales espurias si no se ajusta correctamente.

5. **Utilidad de los controles negativos.** Los paneles sin relación conocida con las enfermedades estudiadas permitieron detectar señales atribuibles al sesgo de dimensionalidad y al efecto de covariables clínicas. Su inclusión fue clave para identificar patrones inconsistentes y descartar asociaciones no específicas.

En conjunto, los objetivos planteados al inicio del estudio han sido alcanzados satisfactoriamente. Los resultados obtenidos muestran que el análisis por *loci* funcionales, combinando con modelos no lineales como SVM radial, permite detectar señales epigenéticas con capacidad discriminativa que no emergen en enfoques globales. Sin embargo, la influencia de variables clínicas como la edad y el sexo no puede descartarse completamente en las señales obtenidas en regiones *promoter* y *TSS200* del panel *AdultNeurodegenerativeDisorder* para PD, por lo tanto, estos resultados deberían investigarse en futuros trabajos.

Bibliografía

[1] James P. Hamilton. "Epigenetics: Principles and Practice". En: *Digestive Diseases* 29.2 (2011), págs. 130-135. ISSN: 1421-9875. DOI: 10.1159/000323874. URL: http://dx.doi.org/10.1159/000323874.

[2] Rajan Jain y Jonathan A. Epstein. "Epigenetics". En: *Congenital Heart Diseases: The Broken Heart*. Springer International Publishing, 2024, págs. 341-364. ISBN: 9783031440878. DOI: 10.1007/978-3-031-44087-8_18. URL: http://dx.doi.org/10.1007/978-3-031-44087-8_18.

[3] Ya Feng, Joseph Jankovic y Yun-Cheng Wu. "Epigenetic mechanisms in Parkinson's disease". En: *Journal of the Neurological Sciences* 349.1–2 (feb. de 2015), págs. 3-9. ISSN: 0022-510X. DOI: 10.1016/j.jns.2014.12.017. URL: http://dx.doi.org/10.1016/j.jns.2014.12.017.

[4] Samareh Younesian et al. "The DNA Methylation in Neurological Diseases". En: *Cells* 11.21 (oct. de 2022), pág. 3439. ISSN: 2073-4409. DOI: 10.3390/cells11213439. URL: http://dx.doi.org/10.3390/cells11213439.

[5] Roy Lardenoije et al. "The epigenetics of aging and neurodegeneration". En: *Progress in Neurobiology* 131 (ago. de 2015), págs. 21-64. ISSN: 0301-0082. DOI: 10.1016/j.pneurobio.2015.05.002. URL: http://dx.doi.org/10.1016/j.pneurobio.2015.05.002.

[6] Andrii Rudenko y Li-Huei Tsai. "Epigenetic modifications in the nervous system and their impact upon cognitive impairments". En: *Neuropharmacology* 80 (mayo de 2014), págs. 70-82. ISSN: 0028-3908. DOI: 10.1016/j.neuropharm.2014.01.043. URL: http://dx.doi.org/10.1016/j.neuropharm.2014.01.043.

[7] Ullrich Wüllner et al. "<scp>DNA</scp>methylation in Parkinson's disease". En: *Journal of Neurochemistry* 139.S1 (jun. de 2016), págs. 108-120. ISSN: 1471-4159. DOI: 10.1111/jnc.13646. URL: http://dx.doi.org/10.1111/jnc.13646.

[8] Lisa D Moore, Thuc Le y Guoping Fan. "DNA Methylation and Its Basic Function". En: *Neuropsychopharmacology* 38.1 (jul. de 2012), págs. 23-38. ISSN: 1740-634X. DOI: 10.1038/npp.2012.112. URL: http://dx.doi.org/10.1038/npp.2012.112.

[9] Alika K Maunakea et al. "Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition". En: *Cell Research* 23.11 (ago. de 2013), págs. 1256-1269. ISSN: 1748-7838. DOI: 10.1038/cr.2013.110. URL: http://dx.doi.org/10.1038/cr.2013.110.

[10] Eliezer Masliah et al. "Distinctive patterns of DNA methylation associated with Parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes". En: *Epigenetics* 8.10 (oct. de 2013), págs. 1030-1038. ISSN: 1559-2308. DOI: 10.4161/epi.25865. URL: http://dx.doi.org/10.4161/epi.25865.

[11] Mehmet Ozansoy y A. Nazli Başak. "The Central Theme of Parkinson's Disease: -Synuclein". En: *Molecular Neurobiology* 47.2 (nov. de 2012), págs. 460-465. ISSN: 1559-1182. DOI: 10.1007/s12035-012-8369-3. URL: http://dx.doi.org/10.1007/s12035-012-8369-3.

[12] D. W. Dickson. "Parkinson's Disease and Parkinsonism: Neuropathology". En: *Cold Spring Harbor Perspectives in Medicine* 2.8 (jun. de 2012), a009258-a009258. ISSN: 2157-1422. DOI: 10.1101/cshperspect.a009258. URL: http://dx.doi.org/10.1101/cshperspect.a009258.

[13] Gagandeep Kaur et al. "DNA Methylation: A Promising Approach in Management of Alzheimer's Disease and Other Neurodegenerative Disorders". En: *Biology* 11.1 (ene. de 2022), pág. 90. ISSN: 2079-7737. DOI: 10.3390/biology11010090. URL: http://dx.doi.org/10.3390/biology11010090.

[14] Lasse Pihlstrøm et al. "Parkinson's disease correlates with promoter methylation in the -synuclein gene". En: *Movement Disorders* 30.4 (dic. de 2014), págs. 577-580. ISSN: 1531-8257. DOI: 10.1002/mds.26073. URL: http://dx.doi.org/10.1002/mds.26073.

[15] Monika A. Myszczynska et al. "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases". En: *Nature Reviews Neurology* 16.8 (jul. de 2020), págs. 440-456. ISSN: 1759-4766. DOI: 10.1038/s41582-020-0377-8. URL: http://dx.doi.org/10.1038/s41582-020-0377-8.

[16] Megha Murthy et al. "DNA methylation patterns in the frontal lobe white matter of multiple system atrophy, Parkinson's disease, and progressive supranuclear palsy: a cross-comparative investigation". En: *Acta Neuropathologica* 148.1 (jul. de 2024). ISSN: 1432-0533. DOI: 10.1007/s00401-024-02764-4. URL: http://dx.doi.org/10.1007/s00401-024-02764-4.