Aplicación de técnicas de aprendizaje supervisado a la selección de biomarcadores para la enfermedad de Parkinson

Jaime Ñíguez Baeza

Tutores:
Juan Antonio Botía Blaya (Universidad de Murcia)
Laura Ibáñez Lladó (Washington Saint Louis University)

Máster de Bioinformática

2023-2024

Índice

Índice de figuras

Índice de tablas

Código disponible

El código fuente de R utilizado y la documentación adicional con los análisis y los resultados obtenidos en detalle se encuentra disponible en el repositorio de GitHub JaimeBioR/circRNA_Parkinson.

Las versiones de software utilizadas en cada apartado aparecen recogidas en cada documento en su apartado final.

1. Resumen

La enfermedad de Parkinson engloba un conjunto de trastornos de origen variado caracterizados por afecciones motoras y no motoras. Aunque constituye la segunda enfermedad neurodegenerativa más común, su completa caracterización plantea un gran reto para la comunidad científica debido a su caracter multifactorial y al elevado número de interacciones existentes entre las vías implicadas en su desarrollo, las cuales pueden ser muy variables entre distintos grupos de pacientes. Esto ha llevado a una diversificación de su campo de estudio que comprende desde el análisis de la influencia de determinados elementos ambientales hasta sus bases ómicas. Sin embargo, el diagnóstico temprano del Parkinson así como encontrar un tratamiento efectivo continúa siendo imposible. En este trabajo se expone el uso de técnicas de aprendizaje supervisado con el fin de seleccionar RNA circulares con potencial para actuar como biomarcadores de la enfermedad de Parkinson. Mediante estas técnicas de Machine Learning se llevó a cabo un entrenamiento de varios modelos predictivos a partir de los valores de expresión de RNA circulares en muestras de plasma de sangre periférica, obteniéndose capacidad predictiva sobre un conjunto de validación independiente a la hora de distinguir entre casos y controles para la enfermedad de Parkinson. Entre los genes utilizados como predictores para los modelos más exitosos se llevó a cabo un análisis de enriquecimiento, encontrándose un conjunto de elementos que presentan concomitancia con vías implicadas en el desarrollo de la enfermedad, a partir de las cuales se seleccionaron RNA circulares candidatos a biomarcadores para la enfermedad. Estos resultados muestran cómo la explotación de RNA circulares puede representar una nueva vía a la hora de estudiar enfermedades de difícil caracterización como el Parkinson. El descubrimiento de nuevos biomarcadores y su uso combinado puede suponer un gran avance para el entendimiento que tenemos sobre estos procesos, permitiendo desarrollar técnicas de prevención, diagnóstico, prognosis y tratamiento que mejoren las expectativas frente al desarrollo de la enfermedad.

2. Introducción

2.1. Parkinson

La enfermedad de Parkinson (PD) está caracterizada por la pérdida de neuronas dopaminérgicas (DN) en la *substantia nigra* del mesencéfalo, lo que ocasiona un cuadro sintomático compuesto por una serie de afecciones motoras denominadas Parkinsonismo. Los principales síntomas son el desarrollo de bradicinesia progresiva, rigidez, temblor en reposo y pérdida del control de la postura, así como otras características no motoras como afecciones del estado de ánimo y el comportamiento, depresión, trastornos cognitivos y la pérdida de autonomía[1 2].

El desarrollo de la enfermedad se ha relacionado tradicionalmente con la presencia de Cuerpos de Lewy, formados principalmente por inclusiones de α-Sinucleina (αSyn). No obstante, la existencia de algunos tipos de Parkinson que carecen de estas estructuras, así como el descubrimiento de otros elementos con un rol importante en su desarrollo, explican la enfermedad como un ente altamente heterogéneo en el que interactúan una gran cantidad de sistemas[3]. Esto no es distinto a nivel genético, existiendo una alta complejidad con cientos de genes, abundante heterogeneidad alélica, penetrancia incompleta y amplio potencial para las interacciones entre genes, así como de estos con el ambiente[3].

El 5-10 % de los pacientes posee un PD de origen monogénico, siendo estos genes clave en las distintas vías moleculares conocidas (SNCA, LRRK2, Parkin, PINK1…)[4]. No obstante, para la mayoría de los casos se desarrolla un Parkinson del tipo esporádico, caracterizado por una predisposición genética y componentes ambientales favorables para su desarrollo como elementos en el estilo de vida, factores de riesgo como la edad, el cual es el más conocido, pero también otros como el sexo o el perfil del microbioma[5].

A lo largo de este apartado se van a describir las principales vías moleculares que han sido relacionadas con la enfermedad, presentándose sus características más importantes, funcionamiento, y los elementos de esta que contribuyan al desarrollo del Parkinson. También se analizarán en algunos casos las interacciones existentes entre las diferentes vías implicadas.

2.1.1. α-Sinucleina como eje central de la enfermedad de Parkinson

La αSyn monomérica se encuentra formada por 3 regiones definidas, su extremo N-terminal comprende los residuos 1-60, presenta una carga neta de 3+ y participa en la unión de la proteína con lípidos de membrana mediante su estructura en α-hélice. La siguiente región es denominada Componente no amiloide (NAC) y supone los residuos 61-95, los cuales son altamente hidrófobos y con tendencia a la agregación. Por último, el extremo C terminal, con residuos 96-140, presenta carga negativa, la cual le permite unirse a cationes y a vesículas sinápticas en presencia de calcio[6].

Su función no está bien definida, aunque se ha demostrado su participación en la liberación y reciclaje de vesículas sinápticas mediante estudios sobre modelos knockout para αSyn[6].

En la enfermedad de Parkinson se ha observado cómo debido a ciertos cambios post traduccionales, la región en α-hélice se ve modificada a una conformación lámina-β altamente insoluble[3]. Este cambio va a suponer la formación de agregados insolubles formados por αSyn patógenas como componente principal en comunión con abundantes lípidos, membranas de orgánulos como mitocondrias y lisosomas y otras proteínas[3]. Estas inclusiones reciben el nombre de Cuerpos de Lewy y suponen tal vez el principal rasgo de la enfermedad de Parkinson a nivel molecular[3].

El proceso de agregación puede desencadenarse por varios factores, como la sobreexpresión del propio gen SNCA de la αSyn[7], por modificaciones postraduccionales como fosforilación de la Ser129[3] o por mutaciones presentes en el gen SNCA, tales como las mutaciones A53T y A30P[7].

Posteriormente, el proceso de agregación se vería amplificado y propagado ya que las moléculas de αSyn patógenas funcionarían como plantilla para el plegamiento patológico de las αSyn endógenas, actuando de una manera similar al mecanismo patológico de los priones, lo que aumentaría su número[8]. Al igual que la αSyn original, las αSyn tóxicas son capaces de secretarse mediante las vesículas extracelulares y ser tomadas por las neuronas vecinas, expandiendo de esta manera sus efectos y acelerando el desarrollo de la enfermedad[29].

Como se verá más adelante, αSyn[7] constituye un elemento desestabilizador en numerosas vías moleculares que se han relacionado con la enfermedad de Parkinson, ya que actúa bloqueando procesos como la autofagia y la degradación vía lisosoma, afectando a la actividad mitocondrial y al aumento de estrés oxidativo[7].

2.1.2. Estrés oxidativo

El uso del oxígeno molecular en el metabolismo aeróbico lleva aparejado un cierto nivel de producción de especies de oxígeno reactivas (ROS), las cuales se caracterizan por poseer un gran poder oxidante sobre múltiples componentes como ácidos nucléicos, proteínas y lípidos, pudiendo llegar a comprometer la homeostasis celular[10]. Frente a esto, se ha propiciado el desarrollo de una serie de estrategias defensivas antioxidantes, las cuales reducen, protegen, previenen y reparan el daño por ROS[10], entre ellas se cuentan herramientas para confinar y neutralizar aquellos elementos degradados que pueden aumentar la producción de ROS, siendo mecanismos clave la autofagia de elementos citosólicos y la apoptosis de células afectadas mediado por respuesta al daño del DNA[11].

En el contexto de la enfermedad de Parkinson, se ha identificado el estrés oxidativo como un elemento transversal al desarrollo de la enfermedad, participando en procesos clave como la degradación de la función mitocondrial y la formación de fibras de αSyn, así como en la degradación vía lisosoma de esta αSyn[12].

De hecho, el proceso de liberación de la dopamina de las vesículas sinápticas se ha determinado como un punto especialmente susceptible al estrés oxidativo, ya que durante todo el proceso se producen ROS y compuestos Dopamina semiquinona[13], fruto de la naturaleza relativamente inestable de la molécula de dopamina, siendo este proceso especialmente destacado en los pacientes de mayor edad[14].

Entre los principales productores de ROS se encuentra la Mitocondria, elemento central de varios tipos de Parkinson, especialmente los complejos I y III de la cadena de transporte de electrones, además de la enzima NADPH-oxidasa (NOX), que también ha sido relacionada con procesos neurotóxicos[12].

El caso de la mitocondria es especial ya que no solo va a generar una gran cantidad de ROS sino que la presencia de estos va a afectar al funcionamiento mitocondrial mediante la peroxidación de lípidos de membrana, lo cual va a modificar la permeabilidad de esta mediante la generación de poros, lo que conduce a daños en el DNA mitocondrial, inducción de la autofagia mitocondrial (mitofagia), apoptosis y neurodegeneración[14].

Por último, en cuanto a la situación de los antioxidantes en la enfermedad de Parkinson, se han observado bajos niveles de expresión de aquellos de origen endógeno como Superóxido dismutasa, Catalasa, Glutatión y Glutatión Peroxidasa[14].

Es interesante la relación que presentan el estrés oxidativo con la presencia de determinados metales como el Hierro (Fe), ya que la transformación de H_2O_2 en el altamente reactivo $OH^{\bullet}$ mediante la reacción de Fenton por oxidación de lípidos y DNA va a producirse en presencia de Fe^{+2}. De esta manera el estrés oxidativo se va a relacionar con la ferroptosis, un tipo de muerte celular dependiente de hierro que se ha relacionado con la degeneración de las neuronas dopaminérgicas en PD[12]. Además de esto, la presencia de altos niveles de metales como Cu, Mn y Fe también va a contribuir al desarrollo de la enfermedad mediante su asociación en la formación de agregados de αSyn en los Cuerpos de Lewy[15].

2.1.3. Neuroinflamación

La microglía supone la primera línea de defensa inmunológica en procesos infecciosos y lesiones del cerebro y la médula espinal, activándose en respuesta a la presencia de agentes infecciosos y elementos patológicos como puedan ser proteínas con plegamiento aberrante y agregados de estas, neurotransmisores alterados y otros elementos reguladores del sistema inmune[16]. Una vez activadas las células de la microglía se va a generar un proceso neuroinflamatorio agudo mediado por citokinas[17].

Además del proceso inflamatorio, la activación crónica de la microglía va a activar a la proteína NOX, desencadenando la producción de ROS y aumentando el estrés oxidativo[18].

En el caso del Parkinson, a principios de los años 80 se descubrieron infiltraciones de microglía activada en muestras postmortem de enfermos, habiéndose analizado desde entonces la asociación que la neuroinflamación presenta en el desarrollo de la enfermedad. No obstante, actualmente no está determinado si el proceso neuroinflamatorio contribuye al desarrollo de la enfermedad o si por el contrario es una consecuencia de la misma[16]. Esto ocurre porque en el momento en el que la mayoría de pacientes son diagnosticados ya han perdido un alto porcentaje de neuronas dopaminérgicas, complicando el trazado de la enfermedad en etapas tempranas. Sin embargo, recientemente se ha sugerido que el proceso inflamatorio podría estar sucediento en pacientes en un estadio prodromal de la enfermedad, años antes de su diagnóstico completo[16].

El análisis de muestras postmorten ha demostrado la elevada presencia de citokinas en el cerebro y la médula espinal de los tipos TNF-α, IL-1β, IL-2, IL-4, IL-6, factor de crecimiento de fibroblastos (bFGF) y factor de crecimiento transformante - β 1 (TGF-β 1), siendo algunas de estas moléculas candidatas a biomarcadores de la enfermedad[16] debido a que pueden ser detectadas en muestras de sangre periférica[17].

Por otro lado, genes asociados a varios tipos de Parkinson como puedan ser LRRK2, SNCA, Parkin y PINK1 también se encuentran relacionados con eventos neuroinflamatorios vía activación de la microglía.

La neuroinflamación se conecta con la vía lisosomal del Parkinson ya que las GBA mutadas del lisosoma actúan como activadores de microglía[16].

Otro de los elementos centrales del Parkinson como son las mitocondrias disfuncionales también se relaciona con la neuroinflamación en tanto que la producción de ROS en el orgánulo van a liberar patrones moleculares asociados al daño (DAMPs) que serán reconocidos por los PRRs de la microglía, contribuyendo a su activación y a la perpetuación del proceso neuroinflamatorio[18][10] .

2.1.4. Déficit en la función mitocondrial.

Las neuronas dopaminérgicas de la sustancia negra poseen algunas características que las hacen muy dependientes del flujo continuo de energía desde la mitocondria, ya que tienen un amplio árbol axonal no mielinizado desplegado hacia el cuerpo estriado, así como actividad marcapasos intrínseca, de manera que una alteración de las vías mitocondriales generaría un problema para la homeostasis neuronal[19].

En cuanto a la enfermedad de Parkinson, la función mitocondrial se ha demostrado clave debido a su participación en el proceso de sinapsis, sus interacciones con αSyn y la producción de ROS en la cadena de transporte de electrones mitocondrial[20].

La mitocondria va a participar en procesos esenciales para el funcionamiento del proceso de sinapsis, encargándose de la producción de ATP, almacenamiento de Calcio y metabolismo de lípidos necesarios en el compartimento sináptico[20].

Estos procesos se ven afectados en pacientes de Parkinson de diferentes maneras, por ejemplo, estudios bioquímicos muestran cómo una deficiencia del Complejo I de la cadena respiratoria mitocondrial, afecta al soporte energético de las sinapsis adrenérgicas, significando la pérdida de estas conexiones en la sustancia nigra y un déficit en la producción de dopamina[20]. Además, el transporte mitocondrial anterógrado desde el soma hasta las terminaciones axonales supone un elemento clave para el aporte de energía en esta zona, habiéndose demostrado que los agregados de αSyn son capaces de bloquear proteínas motoras (kinesinas) interfiriendo en este proceso[19].

En cuanto al almacenamiento de Calcio, se ha demostrado su relación con la formación de agregados de αSyn[20] así como con la interacción entre estas y las vesículas sinápticas[6].

La importancia de la mitocondria en el desarrollo de la enfermedad de Parkinson es tal, que una gran parte de los genes tradicionalmente asociados a esta participan en el control de calidad de la mitocondria (PARKIN, PINK1, DJ-1, FBOX7) y en dinámicas de división mitocondrial (OPA1)[20].

En cuanto a la producción de ROS, las αSyn tóxicas participan en este proceso mediante 2 vías, aumentando la absorción de Calcio por las membranas asociadas a mitocondrias, afectando a la producción de ATP, y por otro lado, la acumulación de αSyn va a reducir la actividad de del Complejo I mitocondrial lo cual también conlleva el aumento de la producción de ROS como se expuso anteriormente[20].

2.1.5. Déficit en la función Lisosomal.

Debido a que la degradación de αSyn depende principalmente de la acción lisosomal, es de esperar que la interferencia con los elementos que conforman esta vía sean factores de riesgo para el desarrollo del Parkinson[7].

Por otro lado, el lisosoma participa en el proceso de autofagia mitocondrial (Mitofagia) mediada por chaperonas, creando de esta manera un triángulo de interacciones αSyn-Mitocondoria-Lisosoma de gran importancia para entender la enfermedad de Parkinson[20].

Se han descrito tres maneras en las que una función lisosomal afectada repercute en el desarrollo de la enfermedad, siendo estas el incorrecto reconocimiento de proteínas mal plegadas[21], el bloqueo de enzimas lisosomales[7] y una participación defectuosa en el control de calidad mitocondrial[22].

- **Impidiendo la selección de proteínas mal plegadas**

El reconocimiento de proteínas con plegamiento anormal va a verse afectado en personas que padecen la enfermedad ya que los propios agregados de αSyn son capaces de interactuar con elementos implicados en esta vía, hecho que también se ha relacionado con la destrucción de neuronas dopaminérgicas[21].

Algunos de estos elementos serían la cochaperona BAG5, la cual también participa en autofagia mediante interacción con p62[22], y el propio receptor de autofagia p62, el cual interactúa con las proteínas que van a ser degradadas vía ubiquitina[7].

Esta relación es bidireccional, ya que se ha demostrado como en experimentos de knock-down de BAG5, se promueve la formación de oligómeros de αSyn y se reducen los niveles de p62. Aunque el mecanismo molecular subyacente a esta interacción está aún por dilucidar[22], se ha descubierto como la expresión de BAG5 era notablemente menor en pacientes PD que presentaban la mutación R492X PINK1 (biomarcador para el Parkinson tipo 6) frente al grupo control[21].

- **Impidiendo la acción de enzimas lisosomales**

Por otro lado, mutaciones en genes asociados a enzimas lisosomales como β-glucocerebrosidasa (GBA1) y Galactocerebrosidasa (GALC); catepsinas lisosomales como CTSD y CTSB; proteínas de la membrana lisosomal (SCARB2, TMEM175, LAMP3) y componentes encargados de la acidificación del contenido del lisosoma (ATP13A2, ATP6V0A1), se han mostrado como factores de riesgo para el desarrollo de la enfermedad de Parkinson por bloqueo de esta vía degradativa[7].

Ha sido especialmente bien descrito el papel de GBA1 en el desarrollo de algunos tipos de Parkinson (*GBA1-associated PD*), acelerando la progresión de la enfermedad.

La función de GBA1 es participar en la degradación del esfingolípido lisosomal Glucosilceramida (GluCer), formando glucosa y ceramida en el proceso. Cuando esta vía del metabolismo de ceramida se ve bloqueada, la degradación de αSyn se ve afectada. Además, la acumulación del sustrato GluCer puede acarrear la conversión de αSyn en su versión patológica, potenciando la formación de agregados[7].

- **Afectando el proceso de mitofagia**

La mitofagia supone un proceso vital a la hora de proteger a la célula del estrés oxidativo ya que las mitocondrias disfuncionales constituyen la mayor fuente de producción de ROS, lo que conduce al desarrollo de diversas enfermedades como Alzheimer o Parkinson[23].

El proceso de autofagia mitocondrial es llevado a cabo de manera parecida al reciclaje de otros elementos citoplasmáticos, actuando mediante un proceso vesicular en el que las cochaperonas reconocen proteínas con plegamiento anormal y las conducen hacia su degradación en el lisosoma. El reconocimiento de las mitocondrias que serán degradadas se encuentra mediado por receptores y adaptadores anclados en su membrana externa (OMM) y en ocasiones en la interna (IMM), los cuales van a interactuar con adaptadores citoplasmáticos como la proteína p62, la cual también se encuentra relacionada con el proceso de degradación ubiquitina-proteasoma y conecta ambos[22].

En este proceso las serin-treonin kinasas PINK1, las cuales se acumulan en la OMM como respuesta a la despolarización de la membrana, van a fosforilar tanto a la proteína Parkin, activándola, como a proteínas mitocondriales receptoras[2223]. Tras esto Parkin ubiquitinará los receptores fosforilados creando cadenas de fosfo-ubiquitina, los cuales son reconocidos por el adaptador p62 como una señal de activación de la Autofagia y se producirá la interacción de este con la proteína asociada a microtúbulos LC3-II, la cual conducirá a la fusión con el autofagosoma[24].

De igual manera que con las proteínas citoplasmáticas como αSyn, la autofagia mediada por chaperonas contribuye al control de calidad de proteínas mitocondriales, habiéndose demostrado que la traslocación de la cochaperona BAG5 a la OMM participa en la inducción de la fisión mitocondrial, activación de la vía PINK1/Parkin e inducción de la autofagia, pudiendo significar esto que BAG6 actuaría como un nuevo receptor para p62[22].

Tal y como se mencionó anteriormente en el bloqueo del transporte anterógrado de mitocondrias por medio de αSyn, este bloqueo también va a afectar al transporte retrógrado de mitocondrias defectuosas, impidiendo el procesamiento de estas mediante la vía lisosomal para la autofagia[19].

2.1.6. Modificaciones epigenéticas

Los mecanismos por los que los cambios epigenéticos contribuyen a la enfermedad son complejos, pero en general están relacionados con el aumento o la disminución de la expresión de genes específicos.

En este apartado se van a presentar una serie de mecanismos epigenéticos que se han mostrado relevantes en el desarrollo de enfermedades neurodegenerativas, con especial atención a aquellos procesos relacionados con la enfermedad de Parkinson. La epigenética engloba aquellas vías con capacidad de modular la expresión génica sin afectar a la propia secuencia del DNA, siendo por lo general modificaciones como el silenciamiento de regiones de la cromatina mediante metilación del DNA, diversas modificaciones de las histonas las cuales intervienen en el grado de empaquetamiento del material genético o mediante bloqueo post transcripcional del mRNA resultante, por ejemplo mediante RNAs de interferencia[25].

Algunos de estos factores epigenéticos van a estar relacionados con estímulos ambientales, generando patrones de expresión génica y desarrollo de patologías como respuesta.

2.1.6.1. Metilación diferencial del DNA en enfermos de Parkinson

La metilación del DNA es el proceso mediante el cual la adición de un grupo metilo es añadido a una de las citosinas presentes en Islas CpG, regiones del genoma en las que se encuentran grandes agrupaciones de Citosina y Guanina, las cuales pueden situarse en regiones promotoras e introns de genes y se encargan de modular su expresión[26].

En el caso de genes con un papel central en el desarrollo de PD como SNCA, LRRK2 y Parkin se han detectado patrones de metilación anormales en su región promotora e intron I en pacientes con PD. En concreto se ha detectado como ciertas variantes del gen SNCA y factores ambientales afectan a los niveles de metilación que presentan estas regiones[26].

Otros estudios sobre metilación de DNA en pacientes de Parkinson han reportado la presencia de islas CpG con metilación diferencial en otros genes asociados con neurodegeneración (SLC12A5, ABCA3, FHIT, FAT1, CPLX2, APBA1, MAGI2, CNTNAP2, ATP8A2 y SMOC2)[27] y patrones de metilación relacionados con el envejecimiento[28].

2.1.6.2. Modificaciones de las histonas y su relación con la enfermedad de Parkinson.

Es especialmente relevante el papel que las modificaciones de la bioquímica de histonas parecen tener sobre la epigenética de trastornos neurodegenerativos tales como las enfermedades de Alzheimer, Huntington, Esclerosis Lateral Amiotrófica o Parkinson[25]. El hecho de que estas modificaciones sean reversibles abre la puerta a su uso terapeutico, bloqueando o estimulando ciertas vías que estén contribuyendo al desarrollo de la enfermedad[29], así como un mayor entendimiento de estas permitiría usarlas como biomarcadores que supongan mejoras en el diagnóstico.

Las histonas comprenden un conjunto de proteínas cuya función se encuentra estrechamente relacionadas con la estructura de la cromatina. En esta, el DNA se encuentra envolviendo agrupaciones octaméricas de histonas formando los denominados nucleosomas, los cuales son la unidad básica de la cromatina. Suponen el principal factor que determina el nivel de condensación de los cromosomas, el cual va a verse modificado a lo largo del ciclo de vida de la célula en función de las necesidades de expresión génica en cada momento.

La manera en la que controlan la expresión de genes se basa en cambios en la afinidad entre estas y el DNA, ya que ligeras modificaciones en la bioquímica de las histonas van a permitir el relajamiento de la estructura de manera que ciertos loci queden expuestos y puedan reclutar factores de unión al DNA que desencadenen la expresión de genes concretos[30]. En general estos cambios van a generar una pérdida de ciertas cargas positivas presentes en las histonas, principalmente los aminoácidos Lisina y Arginina, neutralizando así a la proteína y haciéndola menos afín al DNA.

Existen diversas formas en las que las histonas pueden ser modificadas en su extremo N-terminal, incluyendo acetilación, metilación, fosforilación y SUMOilación[30].

A continuación nos vamos a centrar en aquellas más importante para el Parkinson: Acetilación y Metilación.

- **Acetilación de histonas y su relación con PD**

La acetilación de histonas supone uno de los más importantes cambios que estas presentan y contribuye a un empeoramiento de la afinidad entre la histona y el DNA, generando estructuras más abiertas y por lo tanto activando la expresión génica. El proceso de acetilación se va a producir en residuos de Lisina mediante la enzimas Histona Acetiltransferasas (HATs) y el proceso contrario, la desacetilación por Histona deacetilasas (HDACs), de manera que del equilibrio existente entre estos dos procesos va a depender la homeostasis de la expresión génica y el ciclo celular.

Una de estas enzimas deacetilasas, la Histona Deacetilasa 3 (HDAC3) presenta una actividad neuroprotectora en conjunto con la fosforilasa PINK1, la cual incrementa la actividad deacetilasa de HDAC3 y su unión con p53, desactivándolo y suprimiendo la pérdida de neuronas por la vía de apoptosis dependiente de p53[26].

La fosforilasa PINK1 supone de hecho uno de los elementos centrales de la enfermedad de Parkinson y en este caso, mutaciones en su secuencia van a generar una situación en la que HDAC3 no es capaz de frenar la neurodegeneración causada por la activación de p53.

Al hilo de estos hechos, se han relacionado incrementos en los niveles de acetilación de histonas con el avance de la enfermedad[31] incluyendo regiones hiperacetiladas en más de 20 genes relacionados con el desarrollo de Parkinson familiar y esporádico[26]. No obstante, la función de las desacetilasas va a variar desde aquellas que como HDAC3 presentan una función neuroprotectora, por ejemplo los miembros de la familia de las Sirtuinas SIRT1 y SIRT5, y por el contrario, encontramos desacetilasas con actividad neurotóxica como SIRT2, SIRT3 y SIRT6[25]

- **Metilación de histonas y su relación con PD**

La regulación de la expresión génica mediante metilación y desmetilación de histonas se produce mediante un proceso de monometilación o de dimetilación preferentemente sobre residuos de Lisina (Lys) o Arginina (Arg) presentes en el extremo N-terminal de los monómeros H3 y H4.

Se han descubierto patrones de metilación que se relacionan con la **activación** de la transcripción, por ejemplo la metilación de residuos en H3 como Lys4 (H3K4), Lys14 (H3K14), Lys36 (H3K36), Lys79 (H3K79) o Arg17 (H3R17), mientras que otros se muestran como patrones **represores** de la trancripción, los cuales incluyen tanto metilaciones en H3 en residuos Lys9 (H3K9) o Lys27 (H3K27) como en H4 en Lys20 (H4k20)[26].

Se han identificado patrones de metilación de histonas relacionadas con la región promotora del gen SNCA en pacientes de Parkinson. Estos patrones suponen una triple metilación en la Lys4 de H3 (H3K4me3) y en Lys27 (H3K27me3). En cuanto a su función biológica, se ha determinado que H3K4me3 supone una señal de inicio de la transcripción del gen SNCA, y que esta se encuentra muy representada en pacientes de Parkinson. Por otro lado, metilaciones H3K27me3 se asociaron con la represión de este loci SNCA, contribuyendo al proceso contrario[26].

También se ha analizado el efecto neuroprotector que las histonas demetilasas presentan, restaurando los niveles de metilación no patológicos. El estudio se centró en GSK-J4, un potente inhibidor de las metilaciones en H3K27 y H3K4, el cual frenó la pérdida de neuronas dopaminérgicas tras el tratamiento en ratas con PD inducido mediante la supresión de hierro celular, fuente de estrés oxidativo[32].

2.2. RNA Circulares

El descubrimiento de los RNA no codificantes ha puesto de manifiesto nuevas formas en las que el organismo es capaz de regular diferentes procesos biológicos básicos como la transcripción y la traducción de genes, habiéndose hallado relaciones entre estos RNA y la progresión de distintas enfermedades como el cáncer y diversos procesos neurodegenerativos[33].

Entre las clases de RNA no codificante existe una especialmente desconocida, los RNA circulares (circRNA). Los circRNAs constituyen un tipo de molécula con una característica estructura en forma de anillo completamente cerrada sin extremos 3' o 5' libres. Esto contrasta con las disposiciones observadas en otras moléculas de RNA, las cuales se ajustan al modelo de molécula lineal con extremos libres[34].

Desde su descubrimiento en viroides de plantas hace más de 40 años[35] se ha documentado su presencia en numerosos grupos eukariotas[36] y ha sido durante esta última década cuando su estudio ha aumentado de manera exponencial, habiéndose acumulado evidencia sobre el rol que el circRNA desempeña en procesos biológicos como varios tipos de cáncer[37], enfermedades cardiovasculares[38] y enfermedades neurodegenerativas[39].

Previo a la caracterización de sus funciones biológicas, los circRNA se consideraron como subproductos aberrantes derivado del proceso de splicing canónico, ya que en general los valores de expresión de linearRNA son muy superiores a los de circRNA. Esta suposición también se apoyaba en que ambos procesos se encuentran mediados por la maquinaria molecular del spliceosoma y parten en muchos casos de elementos precursores

comunes, por ejemplo los exones de un mismo pre-mRNA[36]. No obstante, estos procesos divergen pudiendo generarse un circRNA. Para diferenciar estas dos vías, se considerará en este caso de un splicing no canónico o back-splicing[36]. La manera en la que se selecciona cual de estos dos procesos se producirá es aún objeto de estudio, aunque se han caracterizado una serie de factores que se exponen más adelante en este trabajo.

2.2.1. Estructura y biogénesis

La estructura circular característica de estas moléculas va a determinarse en el proceso de selección de exones e intrones conocido como *Splicing* y va a comenzar al igual que con el resto de RNAs, a partir del elemento precursor denominado pre-mRNA, una molécula de RNA recién transcrita y que aún contiene todos los exones e intrones[40].

Este pre-mRNA va a someterse al proceso de maduración en el que se seleccionarán los distintos exones e intrones mediante la maquinaria espliceosomal, generándose distintas versiones del mRNA[40]. Este proceso es común a RNA circulares y lineales, no obstante, lo que ocurre en el caso de los circRNA es que esta misma maquinaria molecular va a generar una unión entre los extremos del mRNA maduro, circularizándolo mediante un enlace fosfodiéster. Este proceso será denominado *Back-Splicing* para diferenciarlo del *Splicing* canónico[3335].

El proceso mediante el cual se lleva a cabo el proceso de circularización no está del todo claro, no obstante, parece estar modulado por la presencia de varios elementos reguladores como secuencias intrónicas complementarias (ICSs) en los extremos de la molécula y proteínas de unión al RNA (RBPs)[36].

2.2.2. Funciones conocidas de los circRNAs

El análisis funcional de los circRNA se encuentra en una etapa relativamente temprana, habiéndose identificado su participación en algunos procesos que regulan el metabolismo celular tales como la transcripción, unión a proteínas y se ha descubierto que presentan potencial para ser traducidos a péptidos[41].

La primera función detectada para los circRNA al comienzo de su estudio fue la de servir como meras esponjas de miRNA. No obstante, con el desarrollo de su estudio se han descrito procesos en los que circRNAs actuaban por si mismos modificando el resultado de rutas metabólicas de formas similares a como se produce el secuestro de miRNA pero con la diferencia de que se realiza sobre proteínas funcionales[41]. Por ejemplo, la interacción del circRNA CDR1a con IGF2BP3 pondría en riesgo la función pro metástasis de este. Del mismo modo, la interacción de CDR1a con p53 bloquearía la acción de la ubiquitina ligasa MDM2 $\cdot$ sobre este último, impidiendo su degradación[41].

Otro tipo de interacciones circRNA-Proteína se ha detectado en el núcleo, mediante las cuales las proteínas son retenidas por medio de circRNA y reclutadas hacia la cromatina[41].

Su función reguladora comprende también la estabilización de moléculas de mRNA específicas, aumentando su probabilidad de ser traducidas a proteína. Esto se puede observar en como la unión de circPAN3 $\cdot$ con el mRNA para IL-13 α 1 sirve como regulador positivo de la acción de esta sobre las células madre del instestino[41].

Por último, al ser traducidos a péptidos los circRNA van a originar una proteína truncada con funciones similares a la original, aunque en ocasiones como para el circFBXW7, distintas versiones de la proteína pueden agregar funciones de forma independiente y a veces incluso contrarias[41].

También se ha demostrado que un mismo circRNA puede presentar funciones diferentes, combinando su actuación como esponja de miRNA con la interacción y sirviendo como plantilla para proteínas[41].

Por último, el descubrimiento de las interacciones reguladoras de la expresión RNA-RNA y el hecho de que algunos circRNA presenten complementariedad de bases con su contraparte lineal sugieren que los circRNA podrían suponer importantes moduladores de la red de RNA competidores endógenos (ceRNA)[42].

Esta gran capacidad para interactuar con proteínas clave en el metabolismo celular y su gran versatilidad permiten dilucidar más procesos en los que los circRNA se encuentren involucrados en el futuro.

2.3. Características de los circRNA y potencial como biomarcadores

El objetivo principal de este estudio radica en el descubrimiento de circRNA candidatos a biomarcadores para PD partiendo de muestras de sangre periférica. Antes de afrontar el diseño experimental se explicará en qué consiste un biomarcador, qué características son deseables y cuáles de estas podrían estar presentes en moléculas como los circRNA.

Una definición de biomarcador ampliamente aceptada sería la elaborada en 1998 por *the National Institutes of Health*, cuya traducción sería: *"Un indicador mensurable de algún estado o condición biológica que se mide y evalúa objetivamente para examinar procesos biológicos normales, procesos patógenos o respuestas farmacológicas a una intervención terapéutica"*[9], de esta manera, el objetivo de nuestro estudio sería analizar si los distintos valores de expresión de cirRNA específicos constituirían un indicador de procesos biológicos relacionados con la enfermedad de Parkinson.

Para describir aquellas características que deberían poseer los circRNA para ser biomarcadores cualificados vamos a centrarnos en la clasificación expuesta por Zhang, et al.[43], la cual incluye:

- **Estabilidad**: Los biomarcadores moleculares han de poseer estabilidad que les permita permanecer en el tejido el tiempo suficiente para la toma de muestras y su análisis. En cuanto a los circRNA, su estructura circular los hace resistentes a la acción de las exonucleasas RNasas, encontrándose enriquecidos en fluídos corporales como el plasma sanguíneo, el fluído cerebroespinal, la saliva y la orina, así como en células flotantes en estos fluídos como las células sanguíneas y tumorales, y en vesículas circulantes.
- **Sensibilidad**: Es clave para un biomarcador que su medida sea reflejo de aquella condición que esperamos examinar. En el caso de los circRNA, al ser un objeto de estudio relativamente nuevo y no conocerse muchas de sus funciones, su validez debería ser comprobada mediante distintos análisis. Uno de los principales objetivos de este estudio es de hecho el dilucidar su correlación con PD.
- **Especificidad**: La presencia de un biomarcador debe relacionarse con la condición específica y ser distinta de la observada en otras condiciones. A este respecto, los circRNA presentan patrones de expresión específicos de tejido y de estado de desarrollo de distintos procesos, lo que los hace candidatos perfectos a biomarcadores, habiendose asociado algunos de ellos con el desarrollo de varios tipos de cáncer, enfermedades cardiovasculares y neurodegenerativas. No obstante, también se ha comprobado que en ocasiones una expresión diferencial de circRNA en los tejidos no se traslada a una diferencia medible en el suero sanguíneo, siendo esto una limitación de su uso como biomarcadores.
- **Precisión**: La medida de un biomarcador debe ser capaz de relacionarse con exactitud con el estado biológico concreto. En este estudio utilizaremos esta precisión a la hora de seleccionar aquellos circRNA candidatos a ser biomarcadores para PD.
- **Reproducibilidad**: Aquellos biomarcadores relacionados con una condición concreta han de ser capaces de predecir esa misma condición en poblaciones diferentes. En nuestro estudio esto se validará mediante predicciones sobre un subconjunto de pacientes independiente al que se utilizó para la selección de biomarcadores candidatos.

En cuanto a nuestro caso concreto, las enfermedades neurológicas llevan aparejada la limitación en la toma de muestras biológicas, por lo que ser capaces de detectar biomarcadores presentes en el plasma sanguíneo, el cual es mucho más accesible, podría mejorar procedimientos clave como el diagnóstico temprano, la selección de terapias y su seguimiento o la prognosis[43].

Aunque el estudio de los circRNA como biomarcadores se ha centrado hasta el momento en algunos tipos de cáncer, el hecho de que sea el cerebro una de las áreas en las que su presencia se encuentra altamente enriquecida[33] los hace muy interesantes para el estudio de PD y otras enfermedades neurodegenerativas.

Por otro lado, ya se ha demostrado la capacidad de los circRNA para traspasar la barrera hematoencefálica, por lo que sería esperable que aquellos relacionados con el Parkinson pudieran estar presentes en la sangre periférica[43].

La identificación de circRNA como biomarcadores para la enfermedad de Parkinson va a presentar sin embargo una desventaja compartida con muchos estudios basados en datos ómicos, ya que el análisis de miles de

circRNA presentes en sangre va a suponer un problema de alta dimensionalidad, con gran cantidad de datos ruidosos no relacionados con el objeto de estudio. Esto hará completamente necesario efectuar una selección de características para identificar aquellos elementos relevantes.

El hecho de encontrar biomarcadores fiables para la enfermedad de Parkinson, para la cual es muy complicado el diagnóstico prematuro, podría suponer un gran cambio a la hora de afrontar el tratamiento de los afectados por la enfermedad o caracterizar nuevos factores que la desencadenen, lo que permitiría trabajar en su prevención[9].

3. Material y métodos

El estudio está enfocado en la aplicación de técnicas de Machine Learning de aprendizaje supervisado para la selección de biomarcadores de la enfermedad de Parkinson. Este experimento se va a formular como un problema de selección de características, en el que se espera obtener el mejor subconjunto de genes productores de circRNA para separar entre las categorías *Enfermo* y *Control* referentes a la enfermedad de Parkinson.

Este trabajo se enmarca en el proyecto de detección de biomarcadores desarrollado en colaboración con el grupo de investigación de Laura Ibáñez Lladó en *Washington Saint Louis University*. La utilización de Machine Learning pretende servir como complemento al proceso basado en estudios de expresión diferencial llevado a cabo por este grupo. Por esta razón no se incluye este tipo de selección en este trabajo.

El flujo de trabajo (Figura 1) a seguir comprenderá un preprocesamiento de los datos de expresión de circRNAs presentes en sangre, su división en subconjuntos de entrenamiento y de validación, seguido del ajuste de los modelos de Machine Learning y un posterior análisis de su rendimiento. Aquellos genes utilizados como predictores en los modelos más destacados serán utilizados en un análisis de enriquecimiento frente a varias ontologías, cuyos resultados se analizarán en búsqueda de términos enriquecidos relacionados con vías moleculares de importancia para la enfermedad de Parkinson.

Los algoritmos de Machine Learning fueron seleccionados por ser capaces de lidiar con conjuntos de datos de alta dimensionalidad y presentar una alta interpretabilidad, de manera que permitieran la identificación de predictores así como la naturaleza y la intensidad de su relación con la variable respuesta. Se seleccionaron Lasso y Random forest, ambos de amplio uso en el estudio de biomarcadores y selección de genes candidatos[44].

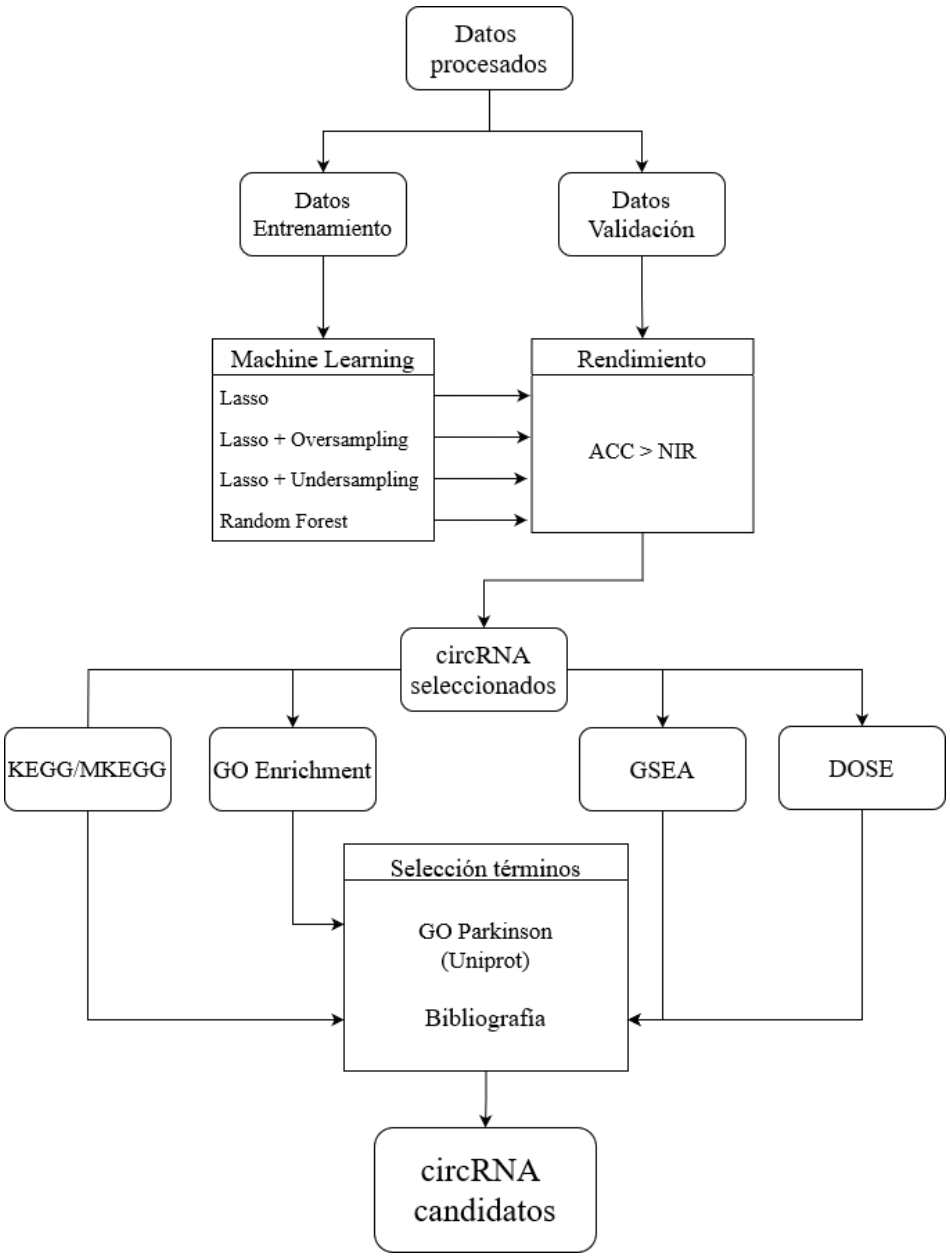

Figura 1: Flujo de trabajo seguido para la selección de biomarcadores

3.1. Contexto clínico

Para la elaboración de este estudio se ha contado con datos procedentes de Parkinson's Progression Markers Initiative (PPMI), un estudio clínico de referencia centrado en la identificación de biomarcadores para la enfermedad de Parkinson, el cual contiene la mayor colección de datos clínicos, imágenes y muestras biológicas en este campo. PPMI se encuentra financiado por The Michael J. Fox Foundation for Parkinson's Research (MJFF) y los datos están a disposición de los investigadores a través de su página web.

Para este estudio se han utilizado datos procedentes de análisis de expresión (RNA-Seq) de genes productores de circRNA procedentes de muestras del suero sanguíneo general de pacientes que presentan la enfermedad y de pacientes control sanos.

Los datos recogidos en PPMI componen un estudio longitudinal de los valores de expresión de cada paciente, ya que la toma de muestras se realiza a lo largo de varias visitas. La primera visita, denominada *baseline*, establece los distintos grupos de pacientes en base a su diagnóstico. A partir de esta toma de datos se llevan a cabo visitas regulares ($V02$, $V04$, $V06$, $V08$...) para un total de 4 visitas el primer año[45] y en adelante de manera bianual. Para este trabajo se partió de un conjunto de valores de expresión de 2176 circRNA para un total de 1530 pacientes enrolados.

PPMI clasifica los pacientes según su diagnóstico en los siguientes grupos:

- **Control**: Participantes que en baseline no presentan trastornos neurológicos significativos, tampoco tienen familiares de primer grado con la enfermedad de Parkinson y presentan transportadores de dopamina (DAT) normales.
- **Prodromal**: Clasifica los individuos que presentan riesgo de desarrollar la enfermedad de Parkinson en el futuro, basándose en características clínicas, variantes genéticas y otros biomarcadores. Algunos de ellos pasarán a ser PD a lo largo de las visitas.
- **SWEDD**: Son aquellos que a pesar de haber sido diagnosticados de Parkinson presentan resultados normales para el escáner de Tomografía computarizada de emisión monofotónica (SPECT por sus siglas en inglés) de DAT, lo que indica que no hay evidencia de déficit dopaminérgico . Algunos de ellos pasarán a ser diagnósticados con Parkinson esporádico (PD) en visitas posteriores a baseline debido al desarrollo de la enfermedad y la aparición de resultados positivos para SPECT DAT.
- **PD**: Agrupa a aquellos pacientes que presentan Parkinson esporádico sin tratar y aquellos con variantes genéticas patogénicas (LRRK2, GBA, SNCA). Todos tienen un diagnóstico clínico de Parkinson y presentan resultados positivos para SPECT de DAT.
- **Genetic Cohort**: Estos pacientes presentan variantes genéticas de riesgo para los genes (LRRK2, GBA, SNCA) y se clasifican de esta manera independientemente de su diagnóstico para la enfermedad de Parkinson.

3.2. Selección de datos

Todo el manejo de datos contenido en el apartado de Material y Métodos, así como la representación de la información en forma de gráficas y tablas se llevó a cabo en el entorno R[46]. Se seleccionaron aquellos pacientes con datos en la primera visita (*Baseline*), los cuales suponen 858 de los 1530 totales, tras lo cual, se escogieron aquellos pertenecientes a los grupos PD y Control, los cuales suponen 538 y 320 pacientes respectivamente. No se utilizaron datos de pacientes con valores nulos para alguno de los valores de expresión, eliminándose al realizar la selección.

3.3. Preprocesamiento

El procesado de datos es un apartado esencial previo al entrenamiento de los modelos predictivos. En nuestro estudio se contó con datos previamente procesados provenientes de RNA-Seq, habiéndose realizado limpieza de datos ausentes o que presentaban una baja variabilidad y normalización de las variables.

También se llevó a cabo un balanceo de clases en algunos de los modelos entrenados con el fin de evitar el sesgo hacia la clase mayoritaria durante la clasificación de muestras. Los métodos utilizados fueron el sobremuestreo (*Oversampling*) de la clase minoritaria hasta que su número representaba el 80 % de la mayoritaria. Así como el submuestreo (*Undersampling*), truncando el número de muestras de la clase mayoritaria hasta que suponían un 25 % más de la minoritaria.

3.4. Modelos predictivos para la selección de circRNA relacionados con Parkinson

La identificación de biomarcadores puede ser afrontada como un problema de selección de características mediante técnicas de Machine Learning. En este estudio se buscará reducir el conjunto de datos completo, el cual presenta probablemente una mayoría de datos ruidosos, para obtener aquel subconjunto de predictores que mejor resuelva el problema de clasificación binaria de pacientes según las categorías *PD* (Casos) y *Healthy Control* (Control).

Entre los algoritmos de selección de características utilizados al tratar datos ómicos se entrenarán modelos de regresión logística con regularización *Least Absolute Shrinkage and Selection Operator* (Lasso)[44]. Por otro lado, también se utilizará un modelo basado en el algoritomo *Random Forest*, con la finalidad de comparar los rendimientos obtenidos y seleccionar aquel o aquellos que nos aporten más información acerca de la relación de nuestros predictores con la enfermedad de Parkinson.

Tanto Lasso como Random Forest presentan una serie de características que los hacen especialmente útiles para la tarea de clasificación sobre nuestro conjunto de datos:

- Capacidad de producir una respuesta binaria.
- Alta interpretabilidad de la respuesta obtenida, lo cual va a permitir trazar una relación más clara entre los predictores y la enfermedad de Parkinson.
- Capacidad de efectuar selección de características y ordenar por importancia los predictores utilizados, de manera que se reduzca la dimensionalidad de los datos a aquellos que presenten una relación más clara con la enfermedad.
- Robustez frente a conjuntos de datos con ruido y con clases desbalanceadas.

3.4.1. Regresión logística con regularización Lasso

Least Absolute Shrinkage and Selection Operator, denominado habitualmente por sus siglas Lasso, constituye un tipo de regularización mediante la cual se contraen los coeficientes de regresión de los predictores mediante un factor de penalización λ, reduciéndolos de manera inversa a su importancia y reduciendo la varianza de los datos. A diferencia de *Ridge Regression*, otro tipo de regularización, Lasso no solo va a reducir el tamaño de los coeficientes, sino que aquellos que no presenten capacidad discriminativa para el proceso predictivo van a reducirse a 0, produciendo de esta manera una selección de características integrada en el proceso del algoritmo de aprendizaje y no durante la etapa de preprocesado[47]. Esta eliminación de características va a generar modelos parsimoniosos, reduciendo el riesgo de sobreajuste, aumentando la intepretabilidad y reduciendo el ruido.

Para el entrenamiento del modelo Lasso se utilizó la librería de R[46] *glmnet*[48], la cual contiene una serie de herramientas para trabajar con modelos de regresión penalizada.

La obtención del valor óptimo de λ se realizó mediante validación cruzada (CV) sobre el conjunto de entrenamiento, seleccionando como parámetro α 1, el cual indica una regularización tipo Lasso. Se utilizó una cuadrícula de valores de λ determinada mediante una semilla de aleatoriedad y se llevó a cabo una CV con 10 pliegues para seleccionar el modelo con mejor rendimiento.

La elección de la métrica se realizó de manera iterativa, determinándose aquella que generaba mejores resultados sobre la capacidad predictiva del modelo. En el caso de Lasso sin balanceo de clases la métrica a minimizar fue el Error Absoluto Medio (*Mean Absolute Error*, *MAE*) . Para los modelos con clases balanceadas mediante sobremuestreo y submuestreo se utilizó como métrica a minimizar la Desviación (*Deviance*).

3.4.2. Random forest

Random forest (RF) es un algoritmo basado en el entrenamiento independiente de un número determinado de árboles de decisión, los cuales actúan en conjunto para elaborar una respuesta. Hace uso de la técnica de muestreo *bootstrap*, la cual selecciona de manera aleatoria un subconjunto de los datos que serán utilizados en el entrenamiento de cada árbol, generando de esta manera árboles poco correlacionados, con resistencia al sobreajuste y con poco sesgo[49]. En el entrenamiento mediante Random Forest se establecen una serie de hiperparámetros que determinan las características del modelo. En este estudio se realizó una selección de aquellos que permitían obtener un modelo más preciso basado en su resultado para la métrica $F - Score$ mediante una validación cruzada con 10 pliegues sobre el conjunto de entrenamiento. $F - Score$ supone la media armónica de la Precisión y la Sensibilidad y es ampliamente utilizada para evaluar el rendimiento de clasificación en modelos de Machine Learning, especialmente cuando los grupos presentan tamaños diferentes[50] como es nuestro caso, con más pacientes diagnosticados con Parkinson que pacientes control.

En este caso, la implementación del modelo Random Forest se realizó utilizando la librería *Ranger*[51] a través del paquete *Caret*[52], de amplio uso en Machine Learning para el entrenamiento y la representación de modelos predictivos.

Los hiperparámetros que se consideraron en el proceso de CV son:

- El número de predictores considerados en cada división del árbol, denominado *mtry*, se estableció en una sucesión de múltiplos de 5 partiendo de 5 hasta la raíz cuadrada del número de genes totales (46,648; truncado a 45).
- Para la regla a seguir al generar los distintos árboles (*Splitrule*) se comparó entre *Gini* y *Hellinger*.
- Para el tamaño mínimo de nodo (*Min.Node.Size*), es decir, el número mínimo de predictores a partir del cual no continuará el crecimiento de los árboles se estableció en la serie 1, 2, 4, 8, 16, 32, 64, 128.

Además se estableció un peso para equilibrar la selección de muestras durante el *bootstrap*, de manera que el modelo obtuviera datos balanceados para las dos variables respuesta.

La importancia de los predictores se estableció mediante la regla *Permutation*, la cual establece la importancia de cada variable en el modelo final en función de cómo afecte la retirada de esta a su rendimiento. Aquella variable más importante obtiene un valor de 100.00 y el resto se ajustan en función de esta.

El valor de importancia de las variables predictoras se utilizó para seleccionar un subconjunto de genes para el apartado de enriquecimiento. Se consideraron los genes en el 10 % más alto en cuanto a valor de importancia.

3.5. Aplicación de análisis de enriquecimiento y técnicas de anotado a la identificación de procesos biológicos.

Con el fin de obtener una mayor perspectiva acerca del significado biológico de los circRNA seleccionados, se utilizaron una serie de análisis de enriquecimiento frente a bases de datos de ontologías especializadas en términos biológicos.

Se utilizaron las librerías de R[46] *clusterProfiler*[53] y *DOSE*[54] para la obtención de los resultados a partir de las bases de datos de ontologías *Gene Ontology* (GO) (*Biological process*, *Molecular Function* y *Cellular Component*), *Kyoto Encyclopedia of Genes and Genomes* (*KEGG pathway over-representation analysis* y *KEGG module over-representation analysis*), *Disease Ontology Semantic and Enrichment analysis* (DOSE). También se realizó un análisis de enriquecimiento basado en un ranking de genes mediante *Gene Set Enrichment Analysis* (GSEA), técnica que utiliza términos procedentes de distintas ontologías.

Para identificar la relación entre los resultados del enriquecimiento y la enfermedad de Parkinson se llevó a cabo una comparación con aquellos términos pertenecientes a Gene Ontology relacionados con proteínas implicadas en el Parkinson. Este análisis se realizó a través de la librería de R *Uniprot.ws*[55], la cual hace uso del servicio REST API de Uniprot.

De manera complementaria se realizó un análisis de las posibles interacciones entre proteínas implicadas en el Parkinson con predictores relevantes del estudio, para lo cual también se utilizó la librería *Uniprot.ws*.

Para aquellos términos procedentes de ontologías distintas a Gene Ontology su selección se realizó mediante consulta bibliográfica de su relación con PD.

En cuanto al nivel de significancia aceptado, se consideraron aquellos términos bajo el umbral de FDR de 0,05.

4. Resultados.

4.1. Rendimiento de los modelos y selección de características.

El rendimiento de los modelos predictivos se evaluó en base a su capacidad de generalización a nuevas muestras. Se enfrentaron los distintos modelos frente al subconjunto de validación y se calcularon las métricas correspondientes a la matriz de confusión (tabla 1) entre las etiquetas de clase predichas y las reales.

Se consideró un buen rendimiento cuando se obtuvo un valor de *Accuracy* superior al *No Information Rate* (NIR), el cual corresponde a la proporción de la clase más frecuente en el conjunto de datos, siendo en este estudio la clase *Caso* (0.623). El cumplimiento de este hecho permitiría asegurar que los predictores contienen información relevante acerca de la variable respuesta.

Aquellos modelos que cumplían este requisito serían por tanto los basados en Lasso sin balanceo de clases, Lasso con sobremuestreo y Random Forest tal y como se puede observar en la tabla 1, siendo especialmente destacado el rendimiento del modelo Random Forest.

Los predictores que fueron seleccionados para el proceso de enriquecimiento fueron los correspondientes al modelo Lasso (*Acc.* 0.672, *pValue* 0.117), Lasso con sobremuestreo (*Acc.* 0.667, *pValue* 0.152) y RF (*Acc.* 0.729, *pValue* 0.026) por ser aquellos con mejor rendimiento.

En cuanto a los hiperparámetros utilizados en estos 3 modelos, se obtuvo que el modelo de RF que presentaba un mejor rendimiento utilizaba el algoritmo de clasificación *Hellinger*, que presenta un desempeño excelente en problemas de clasificación con desbalanceo entre las clases[56], un número de predictores considerados en cada división del árbol (*mtry*) de 25 y un tamaño mínimo de nodo de 1 predictor, que implica una profundidad alta de los árboles generados. Entre los genes seleccionados del modelo Random Forest se obtuvieron un total de 218 genes que superaron el umbral de importancia establecido.

Para el modelo Lasso, en el proceso de CV se obtuvo un factor de penalización λ de 0,014, el cual implica una reducción en el número de predictores a 238, mientras que para el modelo Lasso con balanceo mediante sobremuestreo el valor óptimo de λ obtenido fue 0,013 y 240 predictores.

Ambos modelos Lasso presentan un rendimiento y un número de predictores similar, no obstante, el hecho de que solo 142 de los predictores sean comunes es motivo suficiente para mantener ambos de cara a los posteriores análisis.

Tabla 1: Rendimiento de los modelos predictivos

	Accuracy	AccuracyNull	AccuracyPValue	N.Pred
Lasso	0.6725146	0.6257310	0.1173751	238
Lasso.OverS	0.6666667	0.6257310	0.1520701	240
Lasso.UnderS	0.5789474	0.6257310	0.9096421	160
RandomForest	0.7294118	0.6235294	0.0265541	2173

4.2. Términos enriquecidos

Tras el filtrado de los resultados basado en su significación estadística y su relación con los términos de relevancia en el Parkinson se obtuvieron un total de 16 términos (tabla 2), entre los cuales se hallan términos redundantes procedentes de distintos análisis de enriquecimiento, que sin embargo se han mantenido por presentar distintos conjuntos de genes. Como se especificó con anterioridad, aquellos términos procedentes de Gene Ontology se seleccionaron automáticamente a través de la base de datos de Uniprot, mientras que aquellos términos enriquecidos en análisis diferentes van acompañados de la referencia bibliográfica que los relaciona con la enfermedad de Parkinson.

Para mayor claridad, los términos enriquecidos se relacionan con las distintas vías de la enfermedad de Parkinson con las que guardan relación en la tabla 3.

Tabla 2. Resultados de enriquecimiento seleccionados

Description	circRNA enriquecidos	FDR
early endosome	ERBB2 MGRN1 GGA3 RAB21 MON2 TPCN1 WASHC2A STX6 TBC1D2B ALS2 GPR107 LDLRAD4 INPP5B MARCHF8 RABGAP1L ZFYVE26 AP1G1 AP3D1	0.0003821
negative regulation of cell cycle	WAPL ATRX CHEK2 RBL1 LATS1 UIMC1 BRCA1 BARD1 PTPN11 RNASEH2B CDK5RAP2 DTL WAC TERF2 MDM2 BUB1	0.0021393
pilocytic astrocytoma[57]	BRAF MBP MAPK1 NF1	0.0059052
early endosome	ALS2 ECPAS EEA1 ERBB2 GGA3 INPP5B LDLRAD4 MAPK1 MGRN1 MON2 RAB21 RABGAP1L TBC1D2B TPCN1 WASHC2A WASHC2C	0.0069800
regulation of chromosome organization	WAPL ATRX XRN1 USP7 CDK5RAP2 MAPK1 TERF2 SETDB2 SLF2 BUB1	0.0225001
nuclear periphery	POLA1 NUP205 SMARCC1 AKAP8L SMARCD1 ATXN1 HNRNPM SMARCA4	0.0230328
forebrain development	KIF14 ATRX ASPM WDR37 ZEB1 PCM1 KDM1A CRKL SCYL2 NCOA1 HERC1 NF1 STIL	0.0267891
nuclear androgen receptor binding	KDM1A KDM4C NSD1 SMARCA4	0.0270850
transcription coactivator activity	JADE1 KAT6B KDM1A KDM4C MED14 MED27 MED4 MMS19 MTDH NCOA6 SMARCA4	0.0333511
N-glycan biosynthesis, complex type ST6GAL2[58]	FUT8	0.0362786
Glycosaminoglycan biosynthesis, heparan sulfate backbone[59]	PFKP	0.0388694

Description	circRNA enriquecidos	FDR
transcription corepressor activity	ATF7IP HDAC9 PHF12 MECP2 MIER2 MED1 NSD1 CASP8AP2 SMARCA4	0.0397259
EGFR tyrosine kinase inhibitor resistance[60]	PRKCB BRAF FOXO3 ERBB2 MAPK1 NF1	0.0435633
Thyroid hormone signaling pathway[61]	PRKCB MED13 NCOA1 SLC16A10 RHEB MAPK1 MDM2	0.0435633
Ubiquitin mediated proteolysis[62]	UBE2K FANCL BRCA1 HERC1 UBE3C UBE2I MDM2	0.0447436
early endosome membrane	GGA3 RAB21 MON2 TPCN1 WASHC2A LDLRAD4 INPP5B MARCHF8	0.0481710

Tabla 3. Relación enriquecimiento con vías asociadas a la enfermedad de Parkinson.

Elementos del Parkinson	Términos enriquecidos
Neurodegeneración	"EGFR tyrosine kinase inhibitor resistance"[60], "negative regulation of cell cycle", "forebrain development", Thyroid hormone signaling pathway
Formación de agregados de αSyn	"Glycosaminoglycan biosynthesis, heparan sulfate backbone"[59], "N-glycan biosynthesis, complex type ST6GAL2"[58]
Control de la transcripción	"regulation of chromosome organization", "nuclear periphery", "transcription corepressor activity", "nuclear androgen receptor binding", "transcription coactivator activity"
Neuroinflamación	"pilocytic astrocytoma"[57], "N-glycan biosynthesis, complex type ST6GAL2"[58]
Vía lisosomal	"early endosome", "Ubiquitin mediated proteolysis"[62], Thyroid hormone signaling pathway
Función mitocondrial	Thyroid hormone signaling pathway

5. Discusión

A continuación se va a explorar la capacidad de nuestro flujo de trabajo de seleccionar circRNA como biomarcadores fiables para la enfermedad de Parkinson a partir de datos pertenecientes al estudio longitudinal de *Parkinson's Progression Markers Initiative*, el cual cuenta con 2176 circRNA cuyo nivel de expresión fue medido en sangre.

Nuestro estudio incorpora 2 fases principalmente, la primera de las cuales supone la selección de predictores relevantes para la enfermedad mediante el uso de algoritmos de Machine Learning lineales (Lasso) y no lineales (Random Forest). El resultado del proceso de selección de características mediante los distintos modelos fue de 511 circRNA en total. En este proceso se buscaba determinar si los RNA circulares presentes en el plasma contienen información acerca del proceso patológico y qué predictores eran los más importantes a ese respecto. Lo cual se confirmó al obtener rendimiento predictivo superior al NIR para 3 de los modelos de Machine Learning entrenados (Random forest, Lasso y Lasso con balanceo de clases). El objetivo concreto de reducción de la dimensionalidad a aquellas variables relevantes a partir de un conjunto original que suponíamos ruidoso fue llevado a cabo de forma exitosa por lo tanto.

En la segunda fase de nuestro estudio se realizó un análisis de enriquecimiento utilizando los predictores seleccionados en la fase 1 mediante técnicas de sobrerrepresentación de genes (GO, KEGG, MKEGG, DOSE) y de conjuntos de genes (GSEA). Los términos del enriquecimiento obtenidos se filtraron en base a su relación con vías moleculares implicadas en PD, obteniéndose resultados significativos para un total de 16 términos enriquecidos entre los que se encontraban 93 de los predictores.

Se obtuvieron vías moleculares relacionadas con la formación de agregados de α-Sinucleina, función mitocondrial y lisosomal, control de expresión génica, neuroinflamación y procesos apoptóticos neurodegenerativos.

La contribución de los modelos predictivos a la selección de los 93 biomarcadores candidatos se encuentra repartida tal y como se expone en la figura 2, lo que justifica la inclusión de todos ellos en el proceso de enriquecimiento, ya que aportan información no redundante acerca de la enfermedad. Por otro lado, esta distribución también confirma los resultados de rendimiento de los distintos modelos, ya que se puede observar cómo el número de contribuciones únicas (predictores aportados solo por 1 modelo) es directamente proporcional al rendimiento evaluado.

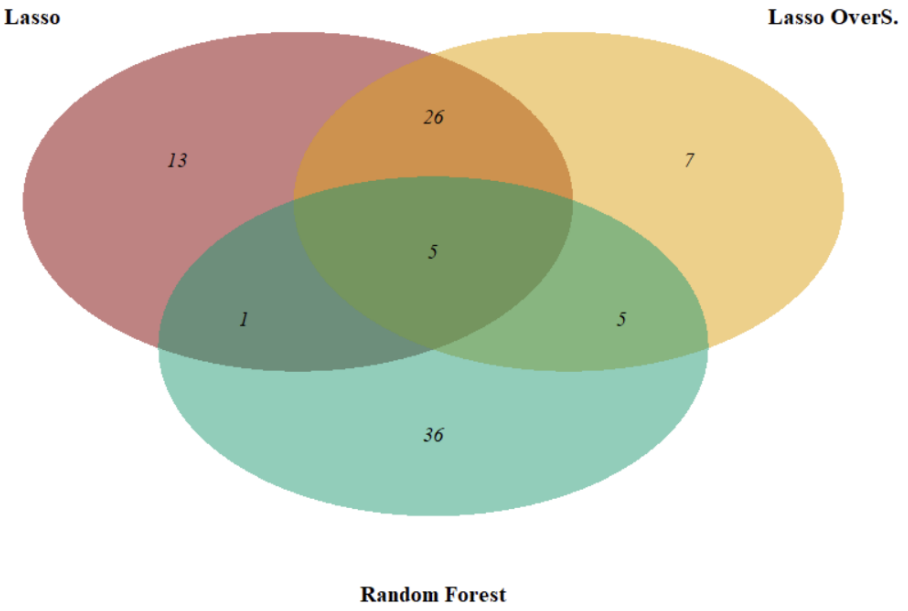

Figura 2: Contribución de los distintos modelos a los biomarcadores seleccionados

En la figura 2 se puede apreciar cómo un total de 5 biomarcadores candidatos se encontraban presentes en los tres modelos como predictores. Estos son **circERBB2**, **circRNASEH2B**, **circKDM1A**, **circUBE2I**

y **circRABGAP1L** y consideramos que representan aquellos biomarcadores candidatos más relevantes.

La caracterización de las funciones realizadas por estos circRNA es variable, por ejemplo, no hay referencias acerca de la acción del RNA circular procdente de KDM1A, aunque su contraparte lineal, la desmetilasa de histonas KDM1a, participa en el proceso de desmetilación de H3K4me, el cual se observó que estaba sobrerepresentado en pacientes de Parkinson[26]. Esto podría sugerir algún tipo de interacción con circKDM1A, fomentando de esta manera el patrón de metilaciones relacionado con la enfermedad de Parkinson.

De circERBB2 también se ha estudiado su acción como circRNA, la cual se encuentra relacionada con el control de la transcripción, promoviendo el desarrollo de cáncer de vesícula[63]. En nuestro estudio se encuentra enriquecido en los términos *early endosome* y *EGFR tyrosine kinase inhibitor resistance* de igual forma que circMAPK1, lo cual puede indicar una especial relevancia en el control de estas vías por medio de circRNA.

Por otro lado, RABGAP1L se encuentra enriquecido en el término *early endosome* relacionado con la vía lisosomal de gran importancia en el desarrollo de Parkinson y ya fue enriquecido en el término relacionado con el Parkinson *Rab GTPase binding*(GO:0017137) en estudios anteriores[64].

circRNASEH2B por su parte ha sido relacionado por su posible implicación en la enfermedad de Alzheimer[65], mienstras que su contraparte lineal se ha relacionado con otras neuropatías por su función antineuroinflamatoria y en la neurogénesis[66], en consonancia con esto, en nuestro análisis de enriquecimiento se encuentra representado en la vía *negative regulation of cell cycle*.

La acción del RNA circular circUBE2I se ha relacionado con la inducción de ferroptosis en queratinocitos en un contexto de estrés oxidativo por H_2O_2 o radiación UVA, mediante la represión de la actividad antioxidante[67]. Este tipo de apoptosis dependiente de hierro se ha relacionado con al neurodegeneración de las neuronas dopaminérgicas en la enfermedad de Parkinson[12].

Como puede observarse, aquellos predictores preferentemente seleccionados mediante técnicas de Machine Learning presentan relaciones coherentes con una posible implicación en PD.

Otros biomarcadores candidatos a tener en cuenta serían aquellos que se encontraban representados en varios términos en el enriquecimiento, destacando circMAPK1 por su presencia en 4 de ellos (*early endosome*, *regulation of chromosome organization*, *EGFR tyrosine kinase inhibitor resistance* y *Thyroid hormone signaling pathway*), así como la existencia de estudios acerca de su acción antitumoral observada en cáncer de estómago mediante bloqueo de la vía MAPK, implicada en la proliferación celular, inflamación y apoptosis[68], procesos que se han incluído en este trabajo por su relación con la enfermedad de Parkinson.

Y por otro lado SMARCA4 (*nuclear periphery*, *nuclear androgen receptor binding*, *transcription coactivator activity* y *transcription corepressor activity*), del cual no hay información disponible acerca de su rol como circRNA pero su contraparte lineal codifica la proteína BRG1, implicada en modificaciones de la cromatina en el contexto del control de expresión génica[69].

6. Conclusiones

En este estudio se han aplicado de forma exitosa técnicas de Machine Learning basadas en Lasso y Random Forest en el proceso de identificación de biomarcadores para la enfermedad de Parkinson a partir de un conjunto de valores de expresión de RNA circulares.

Los modelos predictivos resultantes han sido capaces de afrontar la clasificación binaria entre pacientes enfermos y control, demostrando haber incorporado información relevante sobre la enfermedad de Parkinson.

Muchos estudios[37383839] se están refiriendo al papel regulador de los circRNA y su influencia en el desarrollo de enfermedades ya que sus características los convierten en serios candidatos a biomarcadores. No obstante, pocos estudios han arrojado luz sobre las funciones específicas que estos RNA realizan en este contexto. Aquí se han detectado mediante análisis *in silico* algunos de los circRNA que podrían participar en vías concretas, las cuales podrán ser validadas en el futuro mediante análisis *in vivo*.

Tal y como se ha expuesto a lo largo de este trabajo, la enfermedad de Parkinson se desarrolla sobre muchos y muy variados procesos, la mayoría de los cuales no han sido del todo caracterizados, por lo que la aparición de nuevos biomarcadores podrían generar un mejor entendimiento de las bases que la componen. Igualmente, la inmensa mayoría de las funciones y las repercusiones de los circRNA en el metabolismo son desconocidas, por lo que su entendimiento puede dar respuesta a muchas preguntas acerca de la naturaleza de determinados procesos que, como sucede con el Parkinson, no la han encontrado.

Como resultado de este estudio se han propuesto una serie de biomarcadores candidatos y se ha demostrado que el uso de flujos de trabajo basados en Machine Learning para la selección de genes y biomarcadores constituye una herramienta efectiva a la hora de darle significado a los datos generados en la era ómica.

7. Bibliografía

1. Reich SG, Savitt JM. Parkinson's disease. *Medical Clinics of North America* 2019; 103: 337–350.
2. Müller-Nedebock AC, Dekker MCJ, Farrer MJ, et al. Different pieces of the same puzzle: A multifaceted perspective on the complex biological basis of parkinson's disease. *npj Parkinson's Disease*; 9. Epub ahead of print July 2023. DOI: 10.1038/s41531-023-00535-8.
3. Ye H, Robak LA, Yu M, et al. Genetics and pathogenesis of parkinson's syndrome. *Annual Review of Pathology: Mechanisms of Disease* 2023; 18: 95–121.
4. Cherian A, K.P D, Vijayaraghavan A. Parkinson's disease – genetic cause. *Current Opinion in Neurology* 2023; 36: 292–301.
5. Costa HN, Esteves AR, Empadinhas N, et al. Parkinson's disease: A multisystem disorder. *Neuroscience Bulletin* 2022; 39: 113–124.
6. Stephens AD, Villegas AF, Chung CW, et al. A-synuclein fibril and synaptic vesicle interactions lead to vesicle destruction and increased lipid-associated fibril uptake into iPSC-derived neurons. *Communications Biology*; 6. Epub ahead of print May 2023. DOI: 10.1038/s42003-023-04884-1.
7. Zunke F, Winner B, Richter F, et al. Editorial: Intracellular mechanisms of alpha-synuclein processing. *Frontiers in Cell and Developmental Biology*; 9. Epub ahead of print September 2021. DOI: 10.3389/fcell.2021.752378.
8. Horsager J, Andersen KB, Knudsen K, et al. Brain-first versus body-first parkinson's disease: A multimodal imaging case-control study. *Brain* 2020; 143: 3077–3088.
9. Li T, Le W. Biomarkers for parkinson's disease: How good are they? *Neuroscience Bulletin* 2019; 36: 183–194.
10. Li YR, Trush M. Defining ROS in biology and medicine. *Reactive Oxygen Species*; 1. Epub ahead of print January 2016. DOI: 10.20455/ros.2016.803.
11. Srinivas US, Tan BWQ, Vellayappan BA, et al. ROS and the DNA damage response in cancer. *Redox Biology* 2019; 25: 101084.
12. Dong-Chen X, Yong C, Yang X, et al. Signaling pathways in parkinson's disease: Molecular mechanisms and therapeutic interventions. *Signal Transduction and Targeted Therapy*; 8. Epub ahead of print February 2023. DOI: 10.1038/s41392-023-01353-3.
13. Dionísio PA, Amaral JD, Rodrigues CMP. Oxidative stress and regulated cell death in parkinson's disease. *Ageing Research Reviews* 2021; 67: 101263.
14. Islam MdT. Oxidative stress and mitochondrial dysfunction-linked neurodegenerative disorders. *Neurological Research* 2016; 39: 73–82.
15. Mezzaroba L, Alfieri DF, Colado Simão AN, et al. The role of zinc, copper, manganese and iron in neurodegenerative diseases. *NeuroToxicology* 2019; 74: 230–241.
16. Araújo B, Caridade-Silva R, Soares-Guedes C, et al. Neuroinflammation and parkinson's disease—from neurodegeneration to therapeutic opportunities. *Cells* 2022; 11: 2908.
17. Liu T-W, Chen C-M, Chang K-H. Biomarker of neuroinflammation in parkinson's disease. *International Journal of Molecular Sciences* 2022; 23: 4148.
18. Chakrabarti S, Bisaglia M. Oxidative stress and neuroinflammation in parkinson's disease: The role of dopamine oxidation products. *Antioxidants* 2023; 12: 955.
19. Jain R, Begum N, Tryphena KP, et al. Inter and intracellular mitochondrial transfer: Future of mitochondrial transplant therapy in parkinson's disease. *Biomedicine & Pharmacotherapy* 2023; 159: 114268.
20. Minakaki G, Krainc D, Burbulla LF. The convergence of alpha-synuclein, mitochondrial, and lysosomal pathways in vulnerability of midbrain dopaminergic neurons in parkinson's disease. *Frontiers in Cell and Developmental Biology*; 8. Epub ahead of print December 2020. DOI: 10.3389/fcell.2020.580634.
21. Fu Y, Chen Y, Tian H, et al. Identification of BAG5 as a potential biomarker for parkinson's disease patients with R492X PINK1 mutation. *Frontiers in Neuroscience*; 16. Epub ahead of print July 2022. DOI: 10.3389/fnins.2022.903958.
22. Pattingre S, Turtoi A. BAG family members as mitophagy regulators in mammals. *Cells* 2022; 11: 681.
23. Gupta MK, Randhawa PK, Masternak MM. Role of BAG5 in protein quality control: Double-edged sword? *Frontiers in Aging*; 3. Epub ahead of print March 2022. DOI: 10.3389/fragi.2022.844168.
24. Tanida I, Ueno T, Kominami E. LC3 and autophagy. In: *Methods in molecular biology™*. Humana Press, pp. 77–88.
25. Li Y, Gu Z, Lin S, et al. Histone deacetylases as epigenetic targets for treating parkinson's disease. *Brain Sciences* 2022; 12: 672.
26. Song H, Chen J, Huang J, et al. Epigenetic modification in parkinson's disease. *Frontiers in Cell and Developmental Biology*; 11. Epub ahead of print June 2023. DOI: 10.3389/fcell.2023.1123621.
27. Masliah E, Dumaop W, Galasko D, et al. Distinctive patterns of DNA methylation associated with parkinson disease: Identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* 2013; 8: 1030–1038.

28. Noroozi R, Ghafouri-Fard S, Pisarek A, et al. DNA methylation-based age clocks: From age prediction to age reversion. *Ageing Research Reviews* 2021; 68: 101314.
29. Perri F, Longo F, Giuliano M, et al. Epigenetic control of gene expression: Potential implications for cancer treatment. *Critical Reviews in Oncology/Hematology* 2017; 111: 166–172.
30. Basavarajappa BS, Subbanna S. Histone methylation regulation in neurodegenerative disorders. *International Journal of Molecular Sciences* 2021; 22: 4654.
31. Harrison IF, Smith AD, Dexter DT. Pathological histone acetylation in parkinson's disease: Neuroprotection and inhibition of microglial activation through SIRT 2 inhibition. *Neuroscience Letters* 2018; 666: 48–57.
32. Mu M-D, Qian Z-M, Yang S-X, et al. Therapeutic effect of a histone demethylase inhibitor in parkinson's disease. *Cell Death & Disease*; 11. Epub ahead of print October 2020. DOI: 10.1038/s41419-020-03105-5.
33. Mehta SL, Dempsey RJ, Vemuganti R. Role of circular RNAs in brain development and CNS diseases. *Progress in Neurobiology* 2020; 186: 101746.
34. Chen L-L, Yang L. Regulation of circRNA biogenesis. *RNA Biology* 2015; 12: 381–388.
35. Sanger HL, Klotz G, Riesner D, et al. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proceedings of the National Academy of Sciences* 1976; 73: 3852–3856.
36. Li X, Yang L, Chen L-L. The biogenesis, functions, and challenges of circular RNAs. *Molecular Cell* 2018; 71: 428–442.
37. Tang X, Ren H, Guo M, et al. Review on circular RNAs and new insights into their roles in cancer. *Computational and Structural Biotechnology Journal* 2021; 19: 910–928.
38. Wang H, Yang J, Yang J, et al. Circular RNAs: Novel rising stars in cardiovascular disease research. *International Journal of Cardiology* 2016; 202: 726–727.
39. Wang Q, Qu L, Chen X, et al. Progress in understanding the relationship between circular RNAs and neurological disorders. *Journal of Molecular Neuroscience* 2018; 65: 546–556.
40. Chen L, Wang C, Sun H, et al. The bioinformatics toolbox for circRNA discovery and analysis. *Briefings in Bioinformatics* 2020; 22: 1706–1728.
41. Zhou W-Y, Cai Z-R, Liu J, et al. Circular RNA: Metabolism, functions and interactions with proteins. *Molecular Cancer*; 19. Epub ahead of print December 2020. DOI: 10.1186/s12943-020-01286-3.
42. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature* 2014; 505: 344–352.
43. Zhang Z, Yang T, Xiao J. Circular RNAs: Promising biomarkers for human diseases. *EBioMedicine* 2018; 34: 267–274.
44. Climente-González H, Azencott C-A, Kaski S, et al. Block HSIC lasso: Model-free biomarker detection for ultra-high dimensional data. *Bioinformatics* 2019; 35: i427–i435.
45. Parkinson's progression markers initiative. Help and resources, http://dx.doi.org/10.1080/15548627.2015.1017192 (accessed January 17, 2024).
46. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/ (2023).
47. Albaradei S, Thafar M, Alsaedi A, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. *Computational and Structural Biotechnology Journal* 2021; 19: 5008–5018.
48. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*; 33. Epub ahead of print 2010. DOI: 10.18637/jss.v033.i01.
49. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*; 7. Epub ahead of print January 2006. DOI: 10.1186/1471-2105-7-3.
50. Parkinson E, Liberatore F, Watkins WJ, et al. Gene filtering strategies for machine learning guided biomarker discovery using neonatal sepsis RNA-seq data. *Frontiers in Genetics*; 14. Epub ahead of print April 2023. DOI: 10.3389/fgene.2023.1158352.
51. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 2017; 77: 1–17.
52. Kuhn M. Building predictive models in r using the caret package. *Journal of Statistical Software*; 28. Epub ahead of print 2008. DOI: 10.18637/jss.v028.i05.
53. Yu G, Wang L-G, Han Y, et al. clusterProfiler: An r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 2012; 16: 284–287.
54. Yu G, Wang L-G, Yan G-R, et al. DOSE: An r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2014; 31: 608–609.
55. Carlson M, Ramos M. *UniProt.ws: R interface to UniProt web services*. Epub ahead of print 2023. DOI: 10.18129/B9.bioc.UniProt.ws.
56. Aler R, Valls JM, Boström H. Study of hellinger distance as a splitting metric for random forests in balanced and imbalanced classification datasets. *Expert Systems with Applications* 2020; 149: 113264.

57. Mencke P, Hanss Z, Boussaad I, et al. Bidirectional relation between parkinson's disease and glioblastoma multiforme. *Frontiers in Neurology*; 11. Epub ahead of print August 2020. DOI: 10.3389/fneur.2020.00898.

58. Conroy LR, Hawkinson TR, Young LEA, et al. Emerging roles of n-linked glycosylation in brain physiology and disorders. *Trends in Endocrinology & Metabolism* 2021; 32: 980–993.

59. Mehra S, Ghosh D, Kumar R, et al. Glycosaminoglycans have variable effects on -synuclein aggregation and differentially affect the activities of the resulting amyloid fibrils. *Journal of Biological Chemistry* 2018; 293: 12975–12991.

60. Jin J, Xue L, Bai X, et al. Association between epidermal growth factor receptor gene polymorphisms and susceptibility to parkinson's disease. *Neuroscience Letters* 2020; 736: 135273.

61. Xu J, Zhao C, Liu Y, et al. Genetic correlation between thyroid hormones and parkinson's disease. *Clinical and Experimental Immunology* 2022; 208: 372–379.

62. Lehtonen Š, Sonninen T-M, Wojciechowski S, et al. Dysfunction of cellular proteostasis in parkinson's disease. *Frontiers in Neuroscience*; 13. Epub ahead of print May 2019. DOI: 10.3389/fnins.2019.00457.

63. Huang X, He M, Huang S, et al. Circular RNA circERBB2 promotes gallbladder cancer progression by regulating PA2G4-dependent rDNA transcription. *Molecular Cancer*; 18. Epub ahead of print November 2019. DOI: 10.1186/s12943-019-1098-8.

64. Yin X, Wang M, Wang W, et al. Identification of potential miRNA-mRNA regulatory network contributing to parkinson's disease. *Parkinson's Disease* 2022; 2022: 1–12.

65. Dube U, Del-Aguila JL, Li Z, et al. An atlas of cortical circular RNA expression in alzheimer disease brains demonstrates clinical and pathological associations. *Nature Neuroscience* 2019; 22: 1903–1912.

66. Aditi, Downing SM, Schreiner PA, et al. Genome instability independent of type i interferon signaling drives neuropathology caused by impaired ribonucleotide excision repair. *Neuron* 2021; 109: 3962–3979.e6.

67. Yi P, Huang Y, Zhao X, et al. A novel UVA-associated circUBE2I mediates ferroptosis in HaCaT cells. *Photochemistry and Photobiology*. Epub ahead of print November 2023. DOI: 10.1111/php.13885.

68. Jiang T, Xia Y, Lv J, et al. A novel protein encoded by circMAPK1 inhibits progression of gastric cancer by suppressing activation of MAPK signaling. *Molecular Cancer*; 20. Epub ahead of print April 2021. DOI: 10.1186/s12943-021-01358-y.

69. Trotter KW, Archer TK. The BRG1 transcriptional coregulator. *Nuclear Receptor Signaling* 2008; 6: nrs.06004.