Uso de organismos modelo en el descubrimiento de nuevos genes relacionados con el Parkinson Juvenil

Trabajo Fin de Máster realizado por:

Ana Sabater Aguado

Dirigido por:

Juan A. Botía

MÁSTER UNIVERSITARIO EN BIOINFORMÁTICA

Curso 2018-2019

Universidad de Murcia

Murcia, septiembre, 2019

RESUMEN

Para facilitar el descubrimiento y validación de nuevos genes asociados a enfermedades humanas, se ha creado un script en R que permite obtener el número de pares gen-fenotipo relevantes a partir de herramientas de minería de bases de datos genómicas de animales modelo y de humano desarrolladas por InterMine (intermine.org). Los animales modelo son un pilar de la investigación biomédica al proporcionar sistemas más simples que los presentes en humanos y presentar menos problemas éticos. Hemos elaborado un catálogo de bases de datos de animales modelo, especificando las ontologías utilizadas y las herramientas que poseen. Con el objeto de ilustrar la utilidad de nuestra contribución, se han analizado términos de ontologías fenotípicas a partir de un panel de genes con un total de 34 genes conocidos como causantes de Parkinson definidos por Genomics England (Marzo 2017), y se ha analizado la prevalencia de términos relacionados con síntomas conocidos de dicho trastorno. Se extrajeron un total de 9 términos de fenotipo humano (HP) y 5 de fenotipo de mamífero (MP) que estaban asociados a al menos el 15% de estos genes conocidos como causantes de Parkinson. Con estos términos fenotípicos, se procedió a evaluar 52 genes predichos como causantes de Parkinson mediante *machine learning* por Botía y colaboradores (2018) y se testó empíricamente la probabilidad de que dichas predicciones sean significativamente diferentes de 1000 conjuntos de genes humanos tomados al azar. El p-valor empírico demostró que los genes predichos presentaban una mejor caracterización, tanto mediante términos HP como MP relevantes para Parkinson, de lo que se esperaría por azar. Concluimos que este TFM puede ser una contribución importante a la hora de priorizar genes candidatos a su asociación con enfermedades neurológicas, a pesar de las limitaciones inherentes que conlleva el uso de bases de datos de humano y animales modelo que se ha llevado a cabo en el presente trabajo.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

LISTADO DE ACRÓNIMOS

Acrónimo	Descripción
ADN	Ácido desoxirribonucleico
API	Interfaz de programación de aplicaciones
ARN	ARN: Ácido Ribonucleico
CRISPR	Repeticiones palindrómicas cortas agrupadas y regularmente interespaciadas
DO	Disease Ontology
DPO	Drosophila Phenotype Ontology
EP	Enfermedad de Parkinson
FBbt	Fly Anatomy
FBcv	FlyBase Controlled Vocabulary
GO	Gene Ontology
HP	Human Phenotype
HPO	Human Phenotype Ontology
ICD	International Classification of Diseases
ID	Identificación
KEGG	Kyoto Encyclopedia of Genes and Genomes
MeSH	Medical Subject Headings
MGD	Mouse Genome Database
MGI	Mouse Genome Informatics
MIM	Mendelian Inheritance in Man
MOD	Base de datos de Organismos Modelos
MP	Mammalian Phenotype
MPO	Mammalian Phenotype Ontology
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Secuenciación de segunda generación
OMIM	Online Mendelian Inheritance in Man
OWL	Web Ontology Language
QTL	Loci de rasgos cuantitativos
RGD	Rat Genome Database
SNOMED	Systematized Nomenclature of Medicine
SNV	Variaciones de nucleótidos únicos
VTA	Área tegumental ventral
ZFA	Zebrafish Anatomical Ontology
ZFIN	ZebraFish Information Network

LISTA DE ABREVIATURAS

Abreviatura	Descripción
WBPhenotype	WormBase Phenotype

1. INTRODUCCIÓN

En el presente trabajo buscamos evaluar un listado de genes predichos como causantes de la enfermedad de Parkinson Juvenil mediante *machine learning* (Botía, 2018) utilizando bases de datos que almacenan información genética de animales modelo. En primer lugar, se realizó un catálogo de bases de datos genéticas de animales modelo, especificando las ontologías de fenotipo que utilizan y las herramientas de minería de datos que poseen. En segundo lugar, se utilizó un script de R para obtener todos los términos de ontologías fenotípicas asociados a un listado de 34 genes conocidos por ser causantes de este trastorno, obtenido a partir de paneles de genes de Genomics England (https://panelapp.genomicsengland.co.uk/panels/39/gene/IPPK/) y curada por expertos en la materia.

El script correspondiente a la caracterización de genes conocidos por ser causantes de Parkinson está disponible en el siguiente enlace:

https://github.com/a-sabater/UMU-Bioinformatica-TFM/blob/master/TFM-Caracterización.Rmd

Posteriormente, se evaluó la presencia de términos relacionados con síntomas ampliamente conocidos del Parkinson observados en la bibliografía existente. Es decir, se calibró la cantidad de asociaciones gen-fenotipo que pueden encontrarse en dichas bases de datos genéticas de animales modelo para los genes conocidos. Este es un paso imprescindible si, a partir de dichos genes básicos, realizamos predicciones de nuevos genes que deberían considerase como básicos. Ya que, si dichas predicciones son correctas, deberíamos asumir que reproducirán un número de asociaciones gen-fenotipo similar a la obtenida con los genes conocidos. Y si no es similar, al menos debería estar por encima del ruido. Dicho en otras palabras, el número de asociaciones debería ser mejor que el que obtendríamos prediciendo nuevos genes mediante azar.

El script correspondiente con la evaluación la presencia de términos relacionados con síntomas ampliamente conocidos del Parkinson está disponible en el siguiente enlace:

https://github.com/a-sabater/UMU-Bioinformatica-TFM/blob/master/TFM-Evaluación.Rmd

Además, se incluyen los informes generados mediante RStudio de ambos scripts en el repositorio de github facilitado en formato html.

1.1 Enfermedad de Parkinson

La enfermedad de Parkinson (EP) fue descrita por primera vez por James Parkinson en 1817. Se trata de una enfermedad neurodegenerativa crónica y progresiva que afecta al 2% de la población mayor de 60 años y al 4% de la población mayor de edad de 80 (EP de inicio tardío). Sin embargo, el 10% de la enfermedad puede ocurrir en adultos más jóvenes, entre 20 y 50 años de edad (EP juvenil o de inicio temprano). La mayoría de las deficiencias motoras de la EP surgen por la pérdida de neuronas dopaminérgicas en la sustancia nigra pars compacta, lo que conlleva una pérdida de la entrada de dopamina a las neuronas motoras del cerebro anterior. Otro de sus rasgos característicos es el depósito anormal de α-sinucleína, formando los cuerpos de Lewy. La pérdida neuronal producida en la EP lleva a problemas debilitantes con temblor, rigidez muscular y bradicinesia (lentitud de movimientos). No obstante, la EP es un trastorno complejo, que abarca también síntomas no motores entre los que encontramos anomalías sensoriales (deficiencias olfativas), trastornos del sueño, depresión y deterioro cognitivo. Los síntomas no motores pueden preceder a la aparición de signos motores, siendo muy útiles para el diagnóstico temprano e introducción de estrategias neuroprotectoras. Recientemente, se ha estudiado la posibilidad de que las perturbaciones en la red mitocondria-retículo endoplásmico tengan un papel importante en su patogénesis. La EP se clasifica en dos subtipos genéticos: formas monogénicas familiares con herencia mendeliana y formas esporádicas (Gómez-Suaga, 2018; Poewe, 2017).

La EP juvenil o de inicio temprano carece de una definición consistente. Algunos estudios ponen el umbral para el inicio de los síntomas motores antes de los 40 años, sin embargo, otros amplían la definición para incluir el inicio antes de los 50. Los casos de EP de inicio temprano se caracterizan por tener una progresión más benigna y lenta de la enfermedad, un deterioro cognitivo más leve y complicaciones motoras más tempranas, así como un retraso en el inicio de las caídas y una supervivencia más prolongada. Los pacientes con EP juvenil son menos propensos a tener trastornos de la marcha como síntoma de presentación, teniendo una rigidez y bradicinesia más pronunciadas en comparación con la EP de inicio tardío. Sin embargo, suelen experimentar síntomas psiquiátricos y conductuales más significativos, como depresión, ansiedad y trastornos obsesivo-compulsivos, cuyas tasas varían del 5% al 45% (Ferguson, 2017).

Entre los genes cuyas mutaciones causan EP destacamos los loci PARK, que causan parkinsonismo familiar en 5 a 10% de los casos. Las mutaciones en los genes SNCA (α-sinucleína), LRRK2 y VPS35 causan una forma dominante de la EP, mientras que las mutaciones en PARK2 (parkina), PINK1, y PARK7 (también conocida como DJ-1) causan tienen una forma de herencia recesiva y forma juvenil de la EP (Gómez-Suaga, 2018).

1.2 Mecanismos neurológicos relacionados con el Parkinson Juvenil

Los síntomas motores de la EP se deben a una pérdida de neuronas dopaminérgicas. El sistema de dopamina se ha implicado en muchos aspectos diferentes de la función cerebral, incluida la locomoción, el afecto y la cognición (Grace, 2016). A nivel celular, en mamíferos se han identificado cinco subtipos de receptores de dopamina, etiquetados de D1 a D5. Todos ellos funcionan como receptores metabotrópicos, acoplados a proteínas G, lo que significa que ejercen sus efectos a través de un complejo sistema de segundo mensajero. Estos receptores se pueden dividir en dos familias: D1 y D2. El efecto final de la activación tipo D1 (D1 y D5) puede ser la excitación (a través de la apertura de los canales de sodio) o la inhibición (a través de la apertura de los canales de potasio); el efecto final de la activación tipo D2 (D2, D3 y D4) suele ser la inhibición de la neurona objetivo. Los receptores D1 son los receptores de dopamina más numerosos en el sistema nervioso humano; mientras que los receptores D3, D4 y D5 están presentes en niveles significativamente más bajos (Romanelli, 2010). A nivel fisiológico, en los primates se observa que las neuronas dopaminérgicas que se proyectan hacia el cuerpo estriado límbico y cortical o asociativo están localizadas en la sustancia nigra junto con las neuronas relacionadas con el motor; el nivel dorsal de la sustancia nigra está compuesto por neuronas que proyectan al sistema límbico y a la corteza asociativa, y el nivel ventral de la sustancia nigra está compuesto por más neuronas relacionadas con neuronas motoras motor. Las neuronas de dopamina exhiben una conductancia de marcapasos, que es una corriente de membrana de despolarización lenta y espontánea que mantiene su estado de actividad basal. In vivo, el circuito local y las entradas aferentes de GABAérgico cambian el patrón de disparo del marcapasos en un patrón de disparo lento e irregular. Además, las potentes entradas de GABAérgico del pálido ventral son capaces de hiperpolarizar las neuronas del cerebro medio de la dopamina por debajo del umbral para la activación; de hecho, aunque las neuronas de dopamina del mesencéfalo reciben numerosas entradas que afectan su velocidad de disparo, se descubrió que el pálido ventral en particular controla de manera potente la proporción de neuronas de dopamina que se disparan espontáneamente (Grace, 2016).

La respuesta rápida y fásica conductual de las neuronas de dopamina se caracteriza por una explosión de disparo, que es impulsada por el tegumento pedunculopontino. Sin embargo, el disparo de estallido solo puede ser impulsado en neuronas de dopamina que ya están disparando espontáneamente. El número de neuronas de dopamina activadas se modula en direcciones opuestas por dos regiones del cerebro: el subículo del hipocampo, que aumenta la capacidad de respuesta de la dopamina en los efectos transitorios. La respuesta primaria observada es una breve inhibición de la activación de la neurona dopaminérgica, sobre todo en las áreas relacionadas con el afecto en el área tegumental ventral (VTA) medial y la sustancia nigra. Por el contrario, el VTA lateral, que se supone que está involucrado en la prominencia, responde al estrés con un aumento transitorio de la excitación que podría ser impulsada por la habénula. Sin embargo, los factores estresantes prolongados aumentan la actividad de la población de las neuronas de dopamina en la extensión medial-lateral del VTA y aumentan el nivel de dopamina en la corteza prefrontal y el núcleo accumbens. Debido a que hay más neuronas de dopamina activadas, se aumenta la respuesta fásica conductualmente relevante (Grace, 2016).

El depósito anormal de α-sinucleína es otro de los rasgos característicos de la EP. Se trata de una proteína pequeña de 140 aminoácidos que se localiza en terminales presinápticas, donde regula el ciclo de las vesículas de neurotransmisores. Además de tener funciones específicas en la modulación de la exocitosis y la endocitosis de las vesículas, en condiciones patológicas los agregados de α-sinucleína en inclusiones somáticas y neuríticas. Esta agregación ocurre en neuronas en la EP, como se ha mencionado con anterioridad. La α-sinucleína se puede encontrar en el núcleo y su presencia está regulada por varios factores, incluido el estrés oxidativo y las modificaciones postraduccionales de la proteína. Aunque cabe destacar que la función potencial de la α-sinucleína nuclear no se conoce tan bien. Se han propuesto muchas actividades de la α-sinucleína nuclear que implican interacciones directas o indirectas con el ADN, incluida la modulación del estado de modificación de histonas o la unión directa de ADN. Se ha demostrado que algunas formas de α-sinucleína agregada tienen actividad de endonucleasa de ADN, además la interacción de la α-sinucleína con el ADN podría en la transcripción. También se ha demostrado que la α-sinucleína nuclear influye en la muerte celular neuronal. También se ha argumentado que la interacción de la α-sinucleína con el ADN regula la función celular normal al influir en la transcripción (Schaser, 2019). Lo que sí ha sido demostrado ampliamente es su implicación en multitud de enfermedades neurodegenerativas (Wong, 2017).

1.3 Animales modelo en neurodesarrollo

Los organismos modelo son organismos que se han utilizado para estudiar procesos biológicos específicos. Debido a la existencia de un ancestro común de todos los seres vivos, muchos de ellos conservan gran parte de su material genético y metabólico, haciéndolos idóneos para trasladar el conocimiento obtenido en ellos a otros organismos más complejos. Han sido muy utilizados en el ámbito de la biomedicina, donde la investigación de enfermedades humanas mediante experimentación directa no es viable o incluso ética. Los animales de experimentación han sido un pilar clave en la obtención de la mayor parte del conocimiento médico actual, en el desarrollo de regímenes de tratamiento para las enfermedades humanas y en el desarrollo de dispositivos médicos (Fields y Johnson, 2005).

Desafortunadamente, las diferencias biológicas entre los organismos modelo hacen que la información biológica obtenida a partir de ellos no siempre sea trasladable a los seres humanos. Se han detectado casos donde la actividad biológica de un medicamento en un organismo modelo no se ha traducido en un efecto real en los seres humano. También se ha detectado la existencia de ciertos sesgos ocultos en la manipulación y uso de modelos animales, como es el caso del aumento en el nivel de estrés sufrido en ratones cuando el científico que los manejaba era varón (Katsnelson, 2014). Aunque estos sesgos no se han estudiado demasiado en la bibliografía, es importante tener en cuenta su posible existencia. Otro de los principales inconvenientes del uso de modelos animales es el respeto cada vez mayor de la sociedad respecto a los derechos de los animales. Actualmente, toda investigación donde se requiera el uso de un animal tiene que ser aprobada por un Comité de Bioética. Además, existe un apoyo internacional para reemplazar, reducir y refinar su uso. Particularmente, la Unión Europea apunta a la sustitución total de los animales para la investigación (Jans, 2018). Sin embargo, las autoridades reguladoras siguen exigiendo que tanto vacunas como otros medicamentos se sometan a una evaluación preclínica en modelos animales antes de ingresar a los ensayos clínicos en humanos (FDA, 2018). Otros problemas a los que se enfrentan los investigadores es la gran carga de administración, el costo monetario, el cuidado de los animales y, en algunos casos, llegan a arriesgar su propia vida debido a la acción de determinados grupos de activistas.

En el ámbito del neurodesarrollo, la necesidad de usar modelos animales viene dada por la alta complejidad de los procesos biológicos en el cerebro humano, los problemas éticos derivados de la experimentación en humanos y la dificultad de obtener tejido cerebral humano en desarrollo y enfermo. A finales de la década de los 70, los expertos en ética y los científicos instauraron la "regla de los 14 días", que limita la investigación en embriones humanos hasta un periodo de quince días después de la fecundación, un momento en que aparecen los primeros indicios del sistema nervioso y el último punto en el que el embrión se puede dividir (Shen, 2018). Además, los investigadores solo tienen acceso a experimentar de forma invasiva en cerebros de pacientes post-mortem, dificultando la investigación de estadíos primarios en enfermedades neurológicas.

Aunque los modelos animales son muy utilizados, cabe destacar que no hay modelos animales totalmente fenocópicas (es decir, no presentan todos los fenotipos asociados con un trastorno) de la EP. Muchos modelos muestran la proteinopatía inicial u otras características patológicas relacionadas con el trastorno humano. Algunos modelos también desarrollan una cascada neurodegenerativa más completa, pero aún no se conoce si la secuencia completa de eventos fisiopatológicos que ocurren en la enfermedad humana se captura totalmente. Sin embargo, es ampliamente conocido que los descubrimientos en modelos animales han conducido a una mayor comprensión de los mecanismos moleculares y celulares que conducen a la disfunción y degeneración de las células cerebrales. Los modelos animales han permitido al campo desarrollar, probar y refinar terapias dirigidas, pero los estudios realizados en modelos con ratones han tenido un poder predictivo deficiente para la eficacia de los fármacos en enfermedades neurodegenerativas humanas. Esta falta de traducción no siempre es atribuible a deficiencias en el modelo animal en sí. A pesar del repertorio cada vez mayor de modelos celulares humanos de enfermedades neurodegenerativas, estos modelos están limitados en términos de maduración y complejidad, incluida la falta de circuitos neuronales complejos, la falta de un complemento completo de complejidad glial y la ausencia de componentes vasculares e inmunológicos (Dawson, 2018).

Los modelos animales tradicionales incluyen el gusano nematodo *Caenorhabditis elegans*, la mosca de la fruta *Drosophila melanogaster*, el pez cebra *Danio rerio*, el ratón *Mus musculus* y la rata *Rattus norvegicus*. A continuación, pasamos a introducir brevemente cada uno de estos organismos y su papel en la investigación en la neurodegeneración, específicamente en la EP.

D. melanogaster proporciona numerosas ventajas para la investigación biomédica. Es un animal fácilmente manipulable gracias a su corta vida útil, cría rápida (9 días de ciclo de vida a 25 y maduración sexual femenina dentro de 8-10 horas después de la eclosión), gran cantidad de progenie (100 huevos al día) y genoma pequeño (cuatro cromosomas). El genoma de la mosca es más pequeño, tiene una redundancia de genes sustancialmente menor que la de los mamíferos y contiene homólogos de aproximadamente el 70% de los genes relacionados con enfermedades humanas. Las herramientas genéticas avanzadas permiten la mutagénesis (aparición de mutaciones), el silenciamiento (pérdida de expresión de un gen) y la sobreexpresión (aumento en la expresión de un gen) del gen de interés de una manera específica y temporalmente controlada de tipo celular. El sistema nervioso de la mosca es más simple, pero comparte muchas características en la estructura, organización y función con sus homólogos de mamíferos. Exhiben varios comportamientos dependientes de dopamina controlados por diferentes subclases de neuronas dopaminérgicas. Estudios recientes han demostrado que al menos un subconjunto de neuronas dopaminérgicas en el grupo protocerebral anterior medial de la mosca tiene roles similares a las neuronas dopaminérgicas en la substancia nigra pars compacta de los mamíferos, lo que hace que las moscas sean un sistema modelo siempre relevante en la investigación de la EP. Los modelos de EP de mosca también se han utilizado para investigar los síntomas no motores de la EP, como el sueño y las disfunciones circadianas, los déficits visuales y las anomalías de la memoria y el aprendizaje de forma eficaz. A pesar de estas ventajas, existen varias limitaciones de este organismo modelo para investigar la EP. Las moscas no son el mejor organismo para los estudios bioquímicos que requieren una gran cantidad de tejido homogéneo y no tienen un sistema inmunitario adaptativo, lo que tiene implicaciones importantes en la fisiopatología de la EP. Otra limitación notable es la ausencia de un homólogo de mosca de SNCA, el gen que codifica la α-

sinucleína. Sin embargo, la formación de inclusiones de tipo cuerpo de Lewy puede ser inducida por la expresión ectópica de α-sinucleína humana en moscas, lo que indica que los mecanismos celulares necesarios para la formación de cuerpos de Lewy y el impacto fisiopatológico de las inclusiones pueden investigarse en moscas (Dabool, 2019; Nagoshi, 2018).

C. elegans posee un pequeño tamaño, un ciclo de vida corto (aproximadamente 3 semanas), una alta tasa de reproducción y un bajo costo de mantenimiento. La transparencia óptica de su cuerpo permite la obtención de imágenes in vivo de células mediante microscopía fluorescente y el hecho de que su genoma está completamente secuenciado son otras de sus ventajas. Sin embargo, su anatomía y comportamiento son, en general, muy diferentes de los mamíferos y los humanos. No obstante, *C. elegans* comparte más del 80% de homología genética con los humanos en términos de genes que están relacionados con la enfermedad de Parkinson (Gaeta, 2019). A pesar de las grandes diferencias anatómicas con respecto a los humanos, el sistema nervioso de *C. elegans* consiste en un anillo nervioso circunfaríngeo y contiene características celulares y moleculares clave de las neuronas de mamíferos, incluidos los sistemas de neurotransmisores conservados (dopamina, GABA, acetilcolina, serotonina, etc.) Moléculas de guía de axones, canales iónicos y características sinápticas. Aunque la α-sinucleína no es endógena a *C. elegans*, la expresión de esta proteína asociada a la EP humana en las neuronas dopaminérgicas de *C. elegans* produce neurodegeneración. Además, la mayoría de los genes de la EP familiar, como PINK1, PARK, DJ-1 y LRRK2 tienen al menos un homólogo de *C. elegans*. Tiene 302 neuronas, de las cuales ocho (ADEL, ADER, CEPDL, CEPDR, CEPVL, CEPVR, PDEL y PDER) son dopaminérgicas, como las implicadas en la EP en humanos. Se han identificado cuatro receptores de dopamina (DOP-1, DOP-2, DOP-3 y DOP-4) en *C. elegans*, incluidos los homólogos de cada una de las dos clases de receptores de dopamina de mamíferos (similares a D1 y D2). Su morfología neuronal puede vincularse a anomalías funcionales para una fácil visualización y cuantificación que permite establecer una correlación entre los comportamientos y las aberraciones en las neuronas objetivo, que son inducidas por mutaciones o exposición a toxinas. En contraste, las limitaciones de un modelo de *C. elegans* incluyen la falta de órganos definidos, incluida la compleja estructura cerebral observada en los humanos y, por lo tanto, la incapacidad de recapitular el mismo conjunto de interacciones complejas que involucran varias células cerebrales y tejidos observados en humanos con EP. Además, la cutícula en su mayoría impermeable y la incapacidad de las células intestinales para absorber algunos tipos de químicos pueden requerir altas dosis de exposición para afectar la fisiología del animal. A pesar de estas limitaciones, *C. elegans* ha demostrado ser útil en la investigación del envejecimiento y en la EP (Maulik, 2017).

El pez cebra *Danio rerio* es un pez tropical de agua dulce utilizado como organismo modelo animal en la investigación biomédica y neurofarmacológica. El genoma del pez cebra incluye ortólogos del 71% de los genes humanos y un alto grado de conservación en las propiedades funcionales de muchas de las proteínas codificadas (Howe, 2013). El pez cebra se usa cada vez más para estudiar trastornos neurodegenerativos y procesos asociados, como la disfunción mitocondrial y el estrés oxidativo (Pinho, 2019). Se trata de una herramienta potencial para estudiar varias enfermedades cerebrales humanas asociadas con comportamientos sociales anormales (por ejemplo, esquizofrenia y autismo, algunas afecciones neuropsiquiátricas relacionadas con el deterioro del comportamiento social (como depresión y trastornos de ansiedad). Se han llegado a realizar con éxito pruebas de comportamiento más complejas, como laberintos (aprendizaje, memoria y comportamiento exploratorio) y pruebas de sociabilidad con esta especie (Robea, 2018). El sistema dopaminérgico está bien caracterizado tanto en las etapas embrionarias como en la edad adulta del pez cebra. Las neuronas dopaminérgicas se detectaron por primera vez en un grupo de células en el tubérculo posterior del diencéfalo ventral entre las 18 y 19 horas después de la fecundación. La mayoría de las neuronas dopaminérgicas se encuentran en el cerebro del pez cebra en el quinto día después de la fertilización. El pez cebra tiene una barrera hematoencefálica basada en el endotelio funcional en el tercer día postnatal. Se han identificado genes homólogos de proteínas relacionadas con la EP (PARKIN, DJ-1, PINK1 y LRRK2) en el pez cebra. A pesar de la presencia de tres sinucleínas (sinucleína α, β y γ) en humanos, no se ha encontrado homólogos de la α-sinucleína en el pez cebra (Ünal, 2019).

Los pequeños roedores son fáciles de manejar y albergar, tienen un tiempo de gestación más corto y tienen un menor costo de mantenimiento que los modelos de animales más grandes. También existe una base de datos considerable con datos normativos que se han desarrollado y recopilado cuidadosamente durante muchos años (Vink, 2018). Los cerebros de ratones y ratas tienen una neuroanatomía relativamente conservada en comparación con los cerebros humanos, incluida la función y conectividad de los núcleos más afectados en Parkinson. Los modelos de rata parecen tener una mayor validez aparente al reproducir una pérdida de células de la sustancia nigra más robusta y progresiva y una patología de α-sinucleína en comparación con ratones con las mismas mutaciones (Creed, 2018).

Los modelos transgénicos de α-sinucleína de ratón exhiben las características clave de la EP, incluidas las inclusiones α-sinucleína, la disfunción nigrostriatal, los fenotipos motores y las características no motoras, y por lo tanto son válidos, pero con varias limitaciones. La sobreexpresión transgénica de α-sinucleína en ratones no conduce de manera prominente o consistente a la neurodegeneración; además la neurodegeneración en la EP es un proceso relacionado con la edad, y la vida útil de los ratones puede ser demasiado corta para que se produzca la neurodegeneración mediada por α-sinucleína. Unos pocos modelos mostraron pérdida neuronal en regiones extranigral, incluyendo el neocórtex y el hipocampo, que son relevantes para la EP, pero también en la médula espinal, que es menos característica de la EP. Algunos modelos sin pérdida

de células SN exhiben anomalías en el estriado, como la reducción de la tinción con TH, la disminución de los niveles de dopamina y la mayor expresión del transportador de dopamina, lo que sugiere una disfunción nigrostriatal temprana. Estos marcadores, pueden ser medidas de resultados más sólidas en estudios preclínicos. Estos modelos muestran deterioro motor (bradicinesia, rigidez, temblor, deterioro postural y de la marcha), incluidos aquellos sin inclusiones α-sinucleína o disfunción nigrostriatal. Algunos modelos demuestran una hiperactividad que puede reflejar un comportamiento similar a la ansiedad. A menudo, se encuentra que el deterioro motor progresa con la edad, como se espera para un proceso neurodegenerativo. La desactivación de la expresión de α-sinucleína en modelos condicionales reduce la tasa de disminución en el rendimiento de los tests de rotarod, lo que demuestra una correlación de la disfunción motora con la expresión de α-sinucleína. (Koprich, 2017).

En conclusión, los animales modelo han sido, y siguen siendo a día de hoy, un pilar en la investigación biomédica. A pesar de sus diferencias fisiológicas y genéticas con los humanos, han llevado al descubrimiento de numerosos procesos biológicos y constituyen una herramienta necesaria en el estudio de enfermedades y en el desarrollo de medicamentos. Nuestro trabajo busca obtener toda la información relevante sobre nuestros genes de interés en estos modelos animales, con el fin de poder validar su asociación con la enfermedad de Parkinson.

1.4 Ontologías

Varias disciplinas, en particular la medicina, han desarrollado terminologías estandarizadas a lo largo de los años para promover un lenguaje común. Sin embargo, estas terminologías tienen limitaciones respecto a su aplicabilidad computacional y presentan problemas de pérdida de significado cuando existen diferentes terminologías para el mismo fenómeno. Las ontologías surgieron en el siglo XX, usándose de forma extensiva a partir de principios del siglo XXI, para abordar los problemas tanto de la estandarización como de la computabilidad, y se definen como una definición estructurada y semánticamente formalizada. La aplicación de ontologías a la integración de datos tiene la ventaja de permitir la captura de relaciones entre conceptos en un dominio a través de axiomas formales. La capacidad de las ontologías para capturar conocimiento a través de sus axiomas permite además el uso de razonadores automáticos a través de ellos, una propiedad que se ha vuelto cada vez más importante en la forma en que se están utilizando las ontologías actualmente (Gkoutos, 2018). Dado que nuestro trabajo se centra en la valoración sistemática de genes asociados a enfermedades mediante *machine learning* mediante técnicas bioinformáticas, el uso de ontologías es especialmente importante. A la hora de evaluar un listado de genes, el uso de terminologías estandarizadas que permitan comparar los fenotipos en humano de forma eficiente, además de en diferentes animales modelos, es fundamental. En este trabajo realizamos una búsqueda sistemática de términos de ontologías fenotípicas en bases de datos de enfermedades humanas y de información de diversos animales modelo, con la finalidad de validar series de genes predichos como causantes de Parkinson mediante *machine learning* por Botía y colaboradores en 2018.

Human Phenotype Ontology (HPO; https://hpo.jax.org) ha sido creado por *Monarch Initiative*, un consorcio internacional dedicado a la integración semántica de datos biomédicos y de organismos modelo. Su objetivo es mejorar la investigación biomédica, proporcionando recursos bioinformáticos enfocados en el análisis de enfermedades y fenotipos humanos. HPO se ha desarrollado a partir de Orphanet (https://www.orpha.net/consor/cgi-bin/index.php), DECIPHER (decipher.sanger.ac.uk) y *Online Mendelian Inheritance in Man* (OMIM; https://www.omim.org). HPO se ha utilizado para la fenotipificación profunda computacional, medicina de precisión e integración de datos clínicos en investigación traslacional. En contraste con otras terminologías clínicas, HPO fue diseñado específicamente para facilitar diagnósticos diferenciales e investigaciones científicas (destacando las comparaciones entre especies) (Köhler et al, 2017). Además, una característica clave de la HPO es su interoperabilidad lógica con ontologías de investigación básica como *Mammalian Phenotype Ontology* (MP) y Uberon (Köhler, 2018).

En su década inicial (2007-2017) el enfoque de HPO fue sobre enfermedades raras, principalmente mendelianas. La construcción de la versión inicial de la HPO en 2007/2008 se realizó generando una ontología basada en descripciones en la Sinopsis Clínica de OMIM. En 2015 la ontología se amplió mediante la aplicación de un sistema de reconocimiento de conceptos (CR) con reconocimiento de fenotipo a todos los resúmenes disponibles en aquel momento en PubMed para extraer anotaciones fenotípicas para enfermedades comunes (Groza, 2015). Posteriormente, se ha ido ampliando la ontología HPO a otros tipos de enfermedades.

Los términos HPO tienen una ID única con la estructura HP:000X y una etiqueta. La mayoría de los términos tienen definiciones textuales en las que se debe indicar la fuente de la misma; y muchos términos tienen sinónimos dentro de la misma ontología. En cuanto a la estructura de la ontología, se estructura como gráficos acíclicos dirigidos, similares a las jerarquías. Aunque un término más especializado (hijo) puede relacionarse. La HPO proporciona definiciones textuales para facilitar su uso, pero también tiene una representación lógica sólida con definiciones lógicas basadas en OWL. (Köhler et al, 2017). Dentro de la ontología, los términos cuentan con dos campos obligatorios: Referencia, que indica la fuente de la información utilizada para la anotación, y Evidencia, que indica el nivel de evidencia que respalda la anotación. La ontología

HPO permite buscar genes, enfermedades y fenotipos. Cada una de estas modalidades muestra las asociaciones existentes con los otros tipos de datos antes mencionados. Adicionalmente, los genes presentan un visualizador de UniProt (https://www.uniprot.org/). Además, HPO cuenta con diversas herramientas clínicas, genómicas y fenotípicas.

Disease Ontology (DO; disease-ontology.org) se ha desarrollado como una ontología estandarizada de enfermedades humanas, cuyo propósito es proporcionar a la comunidad biomédica descripciones consistentes de características de fenotipo y vocabulario médico relacionado mediante el esfuerzo colaborativo de investigadores de la Universidad Northwestern, el Centro de Medicina Genética, la Escuela de Medicina de la Universidad de Maryland y el Instituto de Ciencias del Genoma. Integra semánticamente enfermedades y vocabularios médicos a través de un extenso mapeo cruzado de términos de DO a términos MeSH, ICD, *thesaurus* del *National Cancer* Institute (NCI), SNOMED y OMIM (Kibbe, 2015).

OMIM es la principal fuente de información genética sobre enfermedades humanas. Se trata de un catálogo de genes humanos, trastornos y rasgos genéticos actualizado de forma continua, destacando la relación molecular entre la variación genética y la expresión fenotípica. OMIM es una continuación de *Mendelian Inheritance in Man* (MIM) del Dr. Victor A. McKusick, que se publicó por última vez en 1998 y está curada en el Instituto de Medicina Genética McKusick-Nathans, en la Escuela de Medicina de la Universidad Johns Hopkins. Todas las entradas en OMIM tienen números MIM únicos y estables. Los genes y los fenotipos se describen en entradas separadas y los prefijos de números MIM ayudan a distinguir el contenido de la entrada: * (asterisco) denota una entrada de gen; + (signo más) denota una entrada que describe un gen y un fenotipo; # (signo de número) denota un fenotipo con una base molecular conocida; % y nulo denotan fenotipos con diferentes niveles de información de apoyo para la ocurrencia familiar. Las relaciones fenotipo-gen, en las que nos centraremos en este trabajo, se tabulan en el Mapa mórbido del genoma humano de OMIM. Las entradas del fenotipo OMIM están vinculadas a Sinopsis clínicas, que son listas tabulares donde las características clínicas de un trastorno están organizadas anatómicamente, y están creadas para ser utilizadas principalmente por los médicos. Las características clínicas en las sinopsis están vinculadas a vocabularios controlados como HPO y DO (como se ha mencionado con anterioridad). En la parte superior de la vista de página completa de Sinopsis Clínica, se proporciona un botón de alternar para ver estas "ID de funciones". Las sinopsis clínicas son específicas para los pacientes que comparten el mismo gen alterado, y sus sinopsis clínicas se pueden ver de lado a lado en la Serie Fenotípica. OMIM utiliza la construcción genómica de referencia GRCh38 y asigna los genes a las ID de Entrez del *National Center for Biotechnology Information* (NCBI) y las ID de Ensembl según las tablas disponibles en NCBI. El mapa genético de OMIM se usa para mostrar las tablas de relación fenotipo-gen/gen-fenotipo, series fenotípicas y vistas *genemap* (Amberger, 2018).

En cuanto a organismos modelo, *Mammalian Phenotype Ontology* (MPO) es un vocabulario controlado que se ha utilizado en *Mouse Genome Informatics* (MGI; http://www.informatics.jax.org/) (Law y Shaw, 2018) y en *Rat Genome Database* (RGD; https://rgd.mcw.edu/) Shimoyama, 2015) para anotar datos de fenotipos de conjuntos de datos a gran escala, incluidas las pantallas de mutagénesis de ratones y ratas, y de datos descritos en la literatura publicada. *Drosophila Phenotype Ontology* (DPO) (Osumi-Sutherland, 2013) es un vocabulario utilizado en la base de datos *FlyBase* (https://flybase.org) (Thurmond, 2019) que almacena información genética y genómica de drosofila y usa términos de *Gene Ontology* (GO) para tres atributos de los productos genéticos de tipo salvaje: su función molecular, los procesos biológicos en los que desempeñan un papel y su ubicación subcelular. Para respaldar el uso de organismos modelo para promover la investigación en salud humana, los desarrolladores de la MPO han colaborado con el equipo de la HPO para desarrollar definiciones lógicas compatibles, pero estos esfuerzos se limitaron a la comparación de definiciones individuales y dieron como resultado cambios manuales en las respectivas ontologías. El desarrollo basado en patrones ofrece una alternativa más precisa y escalable al desarrollar patrones comunes a los que todas las ontologías fenotípicas (es decir, todos los organismos) pueden referirse y que se pueden aplicar a la vez a una rama completa de una ontología (Köhler, 2018).

El desarrollo de ontologías específicas en el ámbito biológico ha sido clave para la investigación. Dado que nuestro objetivo principal es la validación de genes asociados a la enfermedad de Parkinson, el uso de ontologías estandarizadas es necesario para poder integrar toda la información que extraemos de las bases de datos de animales modelo.

1.5 Paneles de genes NGS

La secuenciación de segunda generación (NGS), también llamada secuenciación masiva paralela, se desarrolló en la última década y permite la secuenciación simultánea de millones de fragmentos de ADN sin el conocimiento previo de la secuencia. La NGS permite la prueba simultánea de múltiples genes de forma más económica y rápida que las plataformas de secuenciación anteriores. Los paneles de biomarcadores de múltiples genes identificados a partir de los datos de expresión génica se pueden usar para diagnosticar enfermedades y o estratificar a los pacientes en diferentes etapas de la enfermedad y se usan para analizar un conjunto de genes asociados con uno o más fenotipos específicos de una enfermedad. Tienen una gran especificidad, alta cobertura y una capacidad probada para caracterizar las variaciones de nucleótidos únicos (SNV) y los pequeños eventos de inserción y deleción de nucleótidos (Lee, 2019).

Con el uso de paneles multigénicos basados en NGS se han producido grandes conjuntos de datos genómicos que suponen una herramienta muy valiosa para ampliar nuestra comprensión de los riesgos que confieren varios genes y mutaciones. También pueden ayudar a identificar nuevas poblaciones de pacientes que pueden beneficiarse de terapias particulares (Shash y Nathalson, 2017). Además, se pueden usar para muchos escenarios de diagnóstico diferentes de manera rentable, lo que permite a los pequeños laboratorios ampliar sus carteras de diagnóstico, así como permitir que los pacientes obtengan diagnósticos personalizados de una manera más rápida y eficiente (Pajusalu, 2018). La selección y prueba de genes en función de la validez clínica y la utilidad de las mutaciones identificadas dentro de los paneles de genes no están reguladas. Cualquier gen, independientemente del peso de la evidencia que asocia una mutación con un fenotipo, puede incluirse en una prueba disponible. De manera similar, no hay un requisito mínimo para los datos que asocian los resultados de las pruebas con una estrategia de gestión clínica que se ha demostrado que mejora los resultados de los pacientes. Además, la clasificación de variantes no es uniforme ni estandarizada entre los laboratorios de ensayos clínicos comerciales y académicos (Shah y Nathalson, 2017). Aunque algunos laboratorios pueden preferir incluir todos los genes posibles que han sido incluso remotamente asociados con el fenotipo de interés con la esperanza de un mejor rendimiento diagnóstico, otros laboratorios adoptan un enfoque más conservador y eligen incluir solo aquellos genes que tienen pruebas sólidas de asociación con el trastorno. Es esencial tener en cuenta que la validación de genes recién descubiertos asociados con un cierto fenotipo requiere análisis funcionales y/o genéticos. Incluir estos genes en un panel sin un análisis adecuado primero puede ser engañoso (Xue, 2015).

Es importante destacar que, aunque los paneles de NGS aumentan el potencial para definir mutaciones patógenas en trastornos neurodegenerativos, también pueden revelar variantes de importancia incierta. La interpretación de los resultados de NGS puede ser difícil y requiere de personal con conocimientos específicos (Williamson, 2018).

1.6 InterMine: herramienta de minería de bases de datos de animales modelos

En cada una de las bases de datos de animales modelo catalogadas en este trabajo, se presenta una parte del software InterMine como herramienta para la minería de información. En la mayoría de los casos, se trata de la única herramienta de minería presente en la misma base de datos. InterMine es una plataforma que reúne datos de los organismos modelo más utilizados en la investigación. Su objetivo es proporcionar interoperabilidad entre tipos de datos relacionados (genes, fenotipo, etc) y brindar a los investigadores una plataforma común única para acceder y analizar los datos. Cada base de datos de Base de datos de Organismos Modelos (MOD) dentro de InterMine proporciona su propio conjunto preferido de ortólogos, facilitando el movimiento entre las bases de datos. Esto tiene la ventaja adicional de que se pueden usar ortólogos curados proporcionados por algunos MOD. El núcleo de las bases de datos MOD - InterMine consiste en datos extraídos de bases de datos funcionales que proporcionan enlaces entre un gen o una proteína y su función, a menudo a través de un vocabulario u ontología controlados, como GO, KEGG y Reactome (Lyne, 2015). InterMine proporciona una interfaz web personalizable en la que los usuarios pueden realizar búsquedas interactivamente utilizando formularios de búsqueda predefinidos (plantillas) o crear su propia consulta utilizando un generador de consultas. Las búsquedas se pueden ejecutar para entidades individuales, como un gen, o conjuntos de entidades, como una lista de genes. Todas las funciones de búsqueda y análisis están disponibles a través de los servicios web (Kalderimis, 2014).

2. OBJETIVOS

2.1 Elaborar un catálogo de bases de datos genéticas disponibles de animales modelo.

Se llevará a cabo una búsqueda y evaluación de bases de datos genómicas para los animales modelo más utilizados en la investigación biomédica, y se ha elaborado un catálogo con las más completas. Dentro de este catálogo identificaremosmos las ontologías de fenotipo que utilizan, así como las herramientas de minería que poseen.

2.2 Crear un mecanismo de uso homogéneo de las bases de datos.

Se creará un método basado en un script en R que utiliza diversas herramientas de InterMine. Este script utilizará un listado de genes de interés y una serie de términos ontológicos relevantes en humanos y en ratón para el trastorno a investigar; y posteriormente devuelverá el número de pares gen-fenotipo relevantes para dicho trastorno.

2.3 Crear un software que utilice bases de datos de animales modelos para evaluar las predicciones de Gen-Fenotipo

Se trata de un objetivo a largo plazo; en este trabajo pretende ser un primer paso en la creación de un software de minería de datos para bases de datos de diferentes modelos animales. La finalidad del mismo, sería la evaluación de predicciones

mediante modelos de *machine learning* para el descubrimiento de nuevos genes implicados en diferentes trastornos humanos.

3. MATERIAL Y MÉTODOS

3.1 Material

En este trabajo utilizamos el lenguaje de programación R, junto con diversos paquetes disponibles para el uso de APIs y la manipulación de datos. Las APIs a las que se accedió mediante el script de R forman parte de InterMine y se corresponden con diferentes bases de datos de modelos animales muy utilizados en la investigación biomédica. Con este script de R se caracterizaron mediante el uso de las APIs de bases de datos de animales modelo 3 tipos de conjuntos de genes diferentes: genes conocidos como causantes de EP, genes predichos como causantes de EP y genes aleatorios. A continuación, se presentan en detalle todos los materiales utilizados en el desarrollo de este trabajo.

3.1.1 Script en R y bases de datos de animales modelo

En este trabajo se ha utilizado el lenguaje de programación R (R Core Team, 2018) para acceder a diferentes servicios web de InterMine y APIs específicas de bases de datos de animales modelo mediante los paquetes curl (Ooms, 2019) y Jsonlite (Ooms, 2014). También se usó el paquete Dplyr (Wickham, 2019) para manipular los datos iniciales. A través de un script de R desarrollado por una alumna para este TFM, se ha accedido a distintas estancias de InterMine que permiten el acceso a bases de datos de los animales modelos más utilizados en la investigación biológica. A continuación, se detallan las bases de datos utilizadas, así como el uso que se le ha dado a dicha herramienta y el tratamiento que han tenido los ortólogos en el caso de bases de datos de especies no humanas.

HumanMine (Smith, 2012) es una base de datos genómicos de Homo sapiens. Una de las plantillas definidas dentro de la API permite obtener a partir de genes humanos los términos HP presentes en la base de datos. Se ha creado un script que devuelve los términos HP de una lista de genes.

MouseMine (Motenko, 2015) proporciona datos integrados de *M. Musculus* desde la base de datos MGI (Law y Shaw, 2018). Esta API nos permite obtener términos DO a partir de genes humanos, a partir de MGI y OMIM. Aunque la herramienta más relevante es la que permite obtener los términos MP a partir de genes humanos. MouseMine obtiene los ortólogos de ratón internamente a través de Homologene (https://www.ncbi.nlm.nih.gov/homologene) y Panther (Mi, 2016; http://www.pantherdb.org/).

RatMine (Smith, 2012) integra muchos tipos de datos de *R. Norvegicus*, *H. Sapiens* y *M. Musculus*. Las estructuras genéticas y otras anotaciones del genoma en RatMine son proporcionadas por una variedad de bases de datos de fuentes curadas que incluyen RGD, National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) y Ensembl (http://www.ensembl.org/Rattus_norvegicus/Info/Index). RatMine también obtiene los ortólogos mediante Panther. En este trabajo hemos utilizado la herramienta que permite obtener los términos MP y a partir de genes humanos.

La base de datos ZFIN contiene notas de fenotipo anotadas manualmente de bibliografía disponible. No es posible acceder a estos fenotipos a través de ZebrafishMine, por lo que se ha utilizado la propia página web de ZFIN.

FlyMine (Lyne, 2007) integra muchos tipos de datos genómicos de *D. melanogaster*. La anotación del genoma de drosofila utiliza como fuente la base de datos FlyBase y para las relaciones de ortología y paralogía entre organismos usa Panther. El primer paso antes de utilizar esta herramienta es obtener los ortólogos de drosofila mediante una plantilla de HumanMine (que también utiliza Panther para las relaciones de ortología). Una vez obtenidos los ortólogos de drosofila, utilizamos la plantilla de FlyMine que devuelve fenotipos de Fly Anatomy (FBbt) y FlyBase Controlled Vocaulary (FBcv) presentes en la base de datos FlyBase.

WormMine (Lee, 2018) integra datos genómicos de *C. elegans* pertenecientes a la base de datos WormBase. Con esta herramienta utilizamos el mismo esquema que con FlyMine. Tras obtener los ortólogos mediante HumanMine utilizamos la plantilla que devuelve el fenotipo de la ontología WormBase Phenotype(WBPhenotype).

3.1.2 Conjuntos de genes y términos de ontologías fenotipos relevantes para la EP

El método desarrollado en este TFM requiere el uso de una lista de símbolos de genes humanos. Para la comprobación del método se utilizaron 3 tipos de conjuntos de genes:

1) Genes conocidos por ser causantes de PD

Estos genes provienen de un panel de genes generado por Genomics England (https://panelapp.genomicsengland.co.uk/panels/39/gene/IPPK/; 31 de marzo de 2017) para "Neurología y trastornos del desarrollo neurológico". Estos paneles de genes están altamente curados por más de 45 clínicos y científicos. Se trata de un listado de 34 símbolos de genes humanos.

2) Genes predichos mediante *machine learning* como causantes de PD

En este trabajo se busca caracterizar los genes predichos como causantes de EP mediante *machine learning* por Botía y colaboradores (Botía, 2018). El listado consta de un total de 52 símbolos de genes humanos.

3) Genes aleatorios

Los subconjuntos de genes aleatorios con los que se pretende comprobar la probabilidad de que los genes predichos como causantes de EP sean realmente causantes de dicho trastorno sea significativamente mayor que las probabilidades de que se escojan genes al azar. Se han creado subconjuntos aleatorios de 52 genes humanos a partir de un fichero con casi 18000 genes codificantes de proteínas en humanos. Este fichero de genes codificantes es el mismo que se utilizó para desarrollar el algoritmo de *machine learning* en el trabajo de Botía y colaboradores (Botía, 2018).

En cuanto a los términos fenotípicos relevantes para la EP, se detectaron 9 fenotipos HP y 5 fenotipos MP que se asociaban a más del 15% de los genes conocidos como causantes de EP.

Términos HP relevantes para la EP: Parkinsonismo (HP:0001300), Inestabilidad postural (HP:0002172), Bradicinesia (HP:0002067), Tremor (HP:0001337), Distonia (HP:0001332), Deterioro cognitivo (HP:0100543), Demencia (HP:0000726), Depresión (HP:0000716) y Atrofia óptica (HP:0000648).

Términos MP relevantes para la EP: Pérdida de neuronas dopaminérgicas (MP:0003244), Tremores (MP:0000745), Ataxia (MP:0001393), Trastornos de la marcha (MP:0001406) y Aprendizaje espacial anormal (MP:0001463).

3.2 Métodos

El método utilizado consiste en el uso de un script de R que se divide en dos pasos diferenciados, que se traducen en dos códigos generados disponibles en el siguiente repositorio https://github.com/a-sabater/UMU-Bioinformatica-TFM, que también incluye los informes en formato html generados mediante RStudio y los ficheros con los conjuntos de genes utilizados en la realización de la prueba con generes relacionados con la EP.

El primer paso en la creación del script de R consiste en la extracción mediante APIs que forman parte del software InterMine de términos fenotípicos para diferentes animales modelo y para humano a partir de un listado de genes conocidos como causantes de EP. Estos genes se obtuvieron mediante un panel de genes de Genomics England, como se ha mencionado en el apartado de materiales. Una vez obtenidos todos los términos fenotípicos, se procedió a evaluar la presencia de términos fenotípicos asociados a síntomas ampliamente estudiados en la EP, y su prevalencia en el listado de genes conocidos como causantes de EP. Este primer paso se lleva a cabo en el script TFM-Caracterización.Rmd.

Los siguientes pasos detallados a continuación se llevan a cabo en el script TFM-Evaluación.Rmd, disponible en el repositorio mencionado con anterioridad.

El segundo paso es la caracterización de genes predichos como causantes de EP mediante *machine learning*. Se decidió utilizar las bases de datos de HumanMine y MouseMine porque contenían un mayor número de genes conocidos (más del 90%) y devolvían más términos relevantes para la EP (prevalencia mayor del 15%, es decir, estaban presentes en al menos 5 genes conocidos). Para esta parte, se crearon dos funciones que utilizan tres parámetros (lista de símbolos de genes humanos, función que utiliza una API específica de InterMine y una lista de términos relevantes) y devuelve el número de pares gen-fenotipo relevantes dentro de esa lista de genes usada. Ambas funciones tienen dos partes diferenciadas:

1) Itera sobre los símbolos de la lista de genes (primer parámetro), utilizándolos para completar la url del servicio web de la API especificada por el usuario (segundo parámetro), que permite obtener términos de ontologías de fenotipo a partir de símbolos de genes. Filtra los términos fenotípicos de cada uno de los genes, y los añade a una lista creando una lista de listas. Posteriormente, estas listas son nombradas con cada uno de los símbolos de genes contenidos en el vector con la lista de genes humanos.
2) Se filtra la lista obtenida en el paso de caracterización previo utilizando el vector de términos fenotípicos relevantes (tercer parámetro), obteniendo los genes que presentan cada uno de esos términos y calculando el ratio del número total de fenotipos relevantes en el listado de genes.

El tercer paso es la evaluación de los ratios del número de pares gen-fenotipo de cada conjunto de genes especificado en los materiales. Se calcula un p-valor empírico para comprobar si existe una diferencia significativa entre los ratios de los genes predichos como causantes de EP y los genes tomados al azar. La fórmula utilizada es la siguiente:

$$p - valor = \frac{(1 + \sum(ratios\ de\ genes\ al\ azar\ mayores\ de\ los\ ratios\ de\ genes\ predichos))}{(1 + total\ de\ ratios\ de\ genes\ al\ azar)}$$

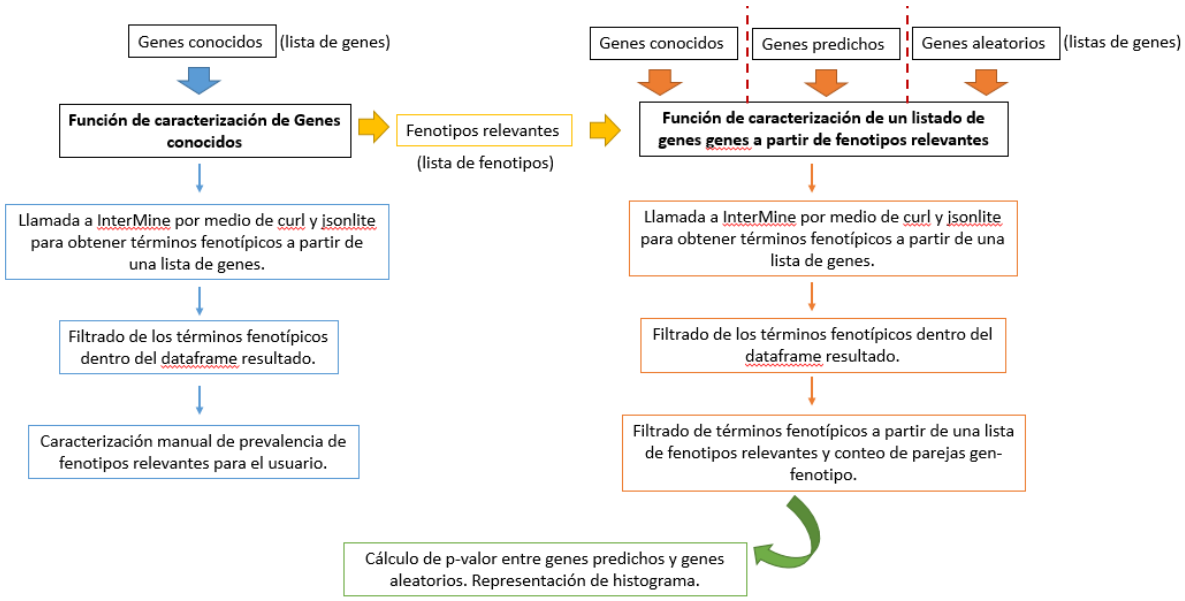

Figura 1- Diagrama estructural del método utilizado en el presente trabajo.

En la figura 1 se muestra el diagrama estructura del método desarrollado, cuyo objetivo es crear un método para la minería de bases de datos que permita la exploración de un listado de símbolos de genes humanos y una lista de términos fenotípicos relevantes para cualquier fenómeno que se quiera investigar. Más específicamente, el método en su conjunto se basa en la evaluación empírica mediante bases de datos genómicas y ontologías de fenotipo de la probabilidad de que el número de fenotipos relevantes presentes en un conjunto de predicciones sea significativamente superior a las que se esperaría por azar.

Esto nos permitirá arrojar algo de luz sobre nuevas predicciones de asociaciones de genes con fenotipos con un enfoque *in silico*. Evidentemente, esto es solo el primer paso para la verificación posterior de las asociaciones en el laboratorio. Pero pretendemos arrojar nuevas evidencias sobre posibles experimentos de laboratorio que aumenten la probabilidad de que dichos experimentos sean positivos.

4. RESULTADOS

El primer paso que se llevó a cabo en el presente trabajo fue crear un catálogo de bases de datos de diferentes animales modelo muy utilizados en la investigación biomédica. Una vez identificadas las bases de datos de interés para nuestro trabajo, se procedió a caracterizar un panel de 34 genes de Genomics England con una asociación conocida a la EP. Tras analizar los términos ontológicos que mejor caracterizaban a esos genes conocidos, se procedió a caracterizar las predicciones realizadas mediante *machine learning* por Botía y colaboradores (Botía, 2018) y 1000 subconjuntos de genes humanos aleatorios con la finalidad de evaluar si dichas predicciones tienen significativamente un mayor número de pares gen-fenotipo relevante para la EP que un conjunto de genes tomados al azar.

4.1 Catálogo de bases de datos genómicos de animales modelo

Tabla 1- Tabla resumen de las bases de datos genómicos de animales modelo.

Organismo modelo	Base de datos	Ontología de fenotipo	APIs de acceso
M. musculus	MGI (Law y Shaw, 2018)	MPO (Smith y Eppig, 2015)	MouseMine
R. norvegicus	RGD (Shimoyama, 2015)	MPO (Smith y Eppig, 2015)	RatMine
D. renio	ZFIN (Howe, 2016)	ZFA (Van Slyke, 2014)	ZebrafishMine
C. elegans	Wormbase (Lee, 2018)	Worm Phenotype Ontology (Schindelman, 2011)	WormMine, WormBase API
D. melanogaster	FlyBase (Thurmond, 2019)	FBbt y FBcv (Osumi-Sutherland, 2013)	FlyMine

En la Tabla 1 recopilamos las bases de datos genéticas de animales modelo que se han utilizado en la caracterización del panel de genes de Genomics England, especificando el organismo modelo cuya información almacenan, las ontologías de fenotipo que utilizan y las APIs de minería de información que poseen. Destacamos que en la columna de APIs se puede observar que la mayoría de ellas presenta como herramienta exclusiva de minería el software InterMine, que es la que se ha utilizado en este trabajo tanto para la caracterización de un panel de genes conocidos y un listado de genes predichos como causantes de EP.

4.1.1 Mouse Genome Informatics (http://www.informatics.jax.org)

MGI es la base de datos internacional para ratones de laboratorio, que proporciona datos genéticos, genómicos y biológicos integrados para facilitar el estudio de la salud y las enfermedades humanas.

MGI aloja múltiples bases de datos y recursos de datos, incluyendo *Mouse Genome Database* (MGD), *Gene Expression Database*, *Mouse Models of Human Cancer database* y GO. proporciona información sobre: gen del ratón, alelo y nomenclatura de cepas; características del genoma genoma de referencia C57BL / 6J; anotaciones fenotípicas; anotaciones funcionales, expresión génica del desarrollo; y modelos de ratón de enfermedades humanas. Esta base de datos tiene acceso a términos de ontologías fenotípicas MP, HPO y DO. También cuenta con diversas herramientas que buscan integrar información de ratón y de humano (Bult, 2019).

MouseMine aparece como parte de los proyectos que contribuyen a este recurso, siendo la API oficial para realizar análisis genómicos fuera de la versión de navegador de MGI.

Como se ha mencionado con anterioridad, el ratón es el principal modelo animal utilizado en biomedicina. Por la facilidad de manejo y bajo coste de mantenimiento de este animal, se ha obtenido mucha información genética a través de la curación de artículos científicos que se ha añadido a la base de datos MGI. Más específicamente, es el modelo animal más utilizado para estudiar enfermedades neurodegenerativas, entre ellas la EP.

4.1.2 Rat Genome Database (https://rgd.mcw.edu/)

RGD es el principal recurso para los datos genéticos, genómicos y fenotípicos de la rata desde su creación en 1999. La rata es un importante modelo animal para farmacología, toxicología, fisiología y patología.

RGD también incorpora genes humanos y de ratón, loci de rasgos cuantitativos (QTL) y polimorfismos de longitud de secuencia simple (SSLP) tanto en informes como en herramientas genómicas. RGD ofrece múltiples herramientas y recursos de software, incluidos múltiples navegadores de genoma, portales de enfermedades, buscador de ontologías, PhenoMiner y el portal de rutas *Pathway*. Esta base de datos tiene acceso a los términos de la ontología fenotípica MP. (Shimoyama, 2015).

RatMine se presenta como una de las herramientas de análisis y visualización de información genómica de esta base de datos.

La rata comparte muchas de las ventajas del ratón, pero ha sido menos utilizado en la investigación biomédica. No obstante, y como se ha explicado con anterioridad, los modelos de rata parecen una pérdida de neuronas dopaminérgicas y una patología de α-sinucleína muy similar a la presente en humanos con EP.

4.1.3 ZebraFish Information Network (https://zfin.org/)

ZFIN es el principal recurso para datos genéticos, genómicos, fenotípicos y de desarrollo del pez cebra (*D. rerio*).

La información de esta base de datos pasa por una curación manual y por la integración de datos completos que incluyen genes de pez cebra, mutantes, líneas y construcciones transgénicas, fenotipos, genotipos, expresiones genéticas, morfolinos, CRISPR, anticuerpos, estructuras anatómicas, modelos de enfermedades humanas y publicaciones. Entre los organismos modelo de vertebrados, el pez cebra destaca por la generación rápida de líneas mutantes dirigidas por secuencia, la caracterización de fenotipos, incluyendo patrones de expresión génica, y la generación de modelos de enfermedades humanas. Entre las mejoras recientes de ZFIN se destacan: el uso de la ontología de las condiciones experimentales del pez cebra, el soporte para fenotipos de expresión génica, modelos de enfermedades humanas, detalles de mutación en los niveles de ADN, ARN y proteínas (Howe, 2016).

ZebrafishMine es una herramienta genómica que permite realizar minería de datos en la base de datos ZFIN. No obstante, no permite acceder a las notas de fenotipo contenidas en la base de datos ZFIN. Por esto se ha utilizado la propia página web de ZFIN para la caracterización de los genes.

Su relevancia como modelo en numerosas investigaciones biomédicas viene dado por el hecho de que el genoma del pez cebra incluye ortólogos del 71% de los genes humanos y un alto grado de conservación en las propiedades funcionales de muchas de las proteínas codificadas. En cuanto a la EP, el sistema dopaminérgico está bien caracterizado tanto en las etapas embrionarias como en la edad adulta del pez cebra y se han identificado genes homólogos de proteínas relacionadas con la EP (PARKIN, DJ-1, PINK1 y LRRK2) en el pez cebra.

4.1.4 WormBase (https://www.wormbase.org)

WormBase es un consorcio internacional de biólogos e informáticos dedicados a proporcionar a la comunidad de investigación información precisa y actualizada sobre la genética, la genómica y la biología de *C. elegans* y nematodos relacionados.

WormBase cura modelos de enfermedad humana de *C. elegans*. Los genes se marcan como 'Modelos experimentales' basados en la curación manual de la literatura de o como 'Modelos potenciales' basados en la ortología a los genes humanos desde Ensembl a través de la conexión Ensembl-Compara y el mapeo posterior de los genes implicados en OMIM. Estos genes que son curados como modelos potenciales y/o experimentales están asociados con enfermedades humanas a través de términos de DO (Lee, 2017).

El servicio de minería de datos principal de WormBase se basa en la plataforma de software de almacenamiento de datos de fuente abierta InterMine.

C. elegans ha permitido el descubrimiento de muchos procesos genéticos, y el hecho de que su genoma está completamente secuenciado es una gran ventaja a la hora de su uso en investigación biomédica. En cuanto a su uso para investigar la EP, *C. elegans* comparte más del 80% de homología genética con los humanos en términos de genes relacionados con dicha enfermedad, y la mayoría de los genes de la EP familiar, como PINK1, PARK, DJ-1 y LRRK2 tienen al menos un homólogo en este modelo animal.

4.1.5 FlyBase (https://flybase.org/)

FlyBase es el principal repositorio y portal web para datos genéticos relacionados con *D. melanogaster*, la mosca de la fruta.

Durante varios años, FlyBase ha alojado datos y desarrollado herramientas para identificar ortólogos de genes de moscas en múltiples organismos. Esto incluye datos de ortología de OrthoDB (https://www.orthodb.org) y meta-análisis de DIOPT (https://www.flyrnai.org/cgi-bin/DRSC_orthologs.pl). (Thurmond, 2019). Proporciona información fenotípica basada en la *Zebrafish Anatomical Ontology* (ZFA).

FlyMine permite acceder a la base de datos FlyBase y realizar minería de datos.

La mosca de la fruta ha sido ampliamente utilizada en investigación genética, debido a que su genoma es más pequeño, tiene una redundancia de genes sustancialmente menor que la de los mamíferos y contiene homólogos de aproximadamente el 70% de los genes relacionados con enfermedades humanas. En cuanto a la EP, se conoce que exhiben varios comportamientos dependientes de dopamina controlados por diferentes subclases de neuronas dopaminérgicas.

4.2 Análisis de términos fenotípicos asociados a Parkinson

Mediante un script de R, se llama a las diferentes API de minería de InterMine mencionadas con anterioridad. A partir de un listado de 34 genes conocidos como causantes de EP, se extraen los términos de ontologías fenotípicas usadas en cada una de las bases de datos y se procede a evaluar la presencia de términos relacionados con los síntomas estudiados de la EP. La finalidad de este apartado es caracterizar los fenotipos asociados a la EP y analizar su prevalencia en nuestro listado de genes conocidos por causar EP. El script está disponible en https://github.com/a-sabater/UMU-Bioinformatica-TFM/blob/master/TFM-Caracterización.Rmd.

4.2.1 Fenotipos motores

La EP es principalmente conocida por sus síntomas motores, como es el caso de la inestabilidad postural, la bradicinesia o la distonia. Estos fenotipos han sido muy estudiados, y nuestro objetivo en este apartado es analizar la presencia de términos de ontologías fenotípicas relacionados con los mismos en un listado de 34 genes asociados a la EP.

4.2.1.1 HumanMine

Encontramos el término Parkinsonismo (HP:0001300) en el 64,71% de los genes, el término Inestabilidad Postural (HP:0002172) está asociado con 11 genes (el 32,4%), el término Bradicinesia (HP:0002067) en 21 genes (61,76%), el término Tremor (HP:0001337) en 21 genes (61,76%) y el término Distonia (HP:0001332) en 22 genes (64,71%). De los 34 genes estudiados, dos no estaban relacionados con ninguno de estos términos HP: SPG11 y RAB39B.

4.2.1.2 MouseMine

Los fenotipos que encontramos relevantes para la EP los dividiremos en varios grupos:

- Fenotipo relacionado con neuronas dopaminérgicas

Tabla 2- Tabla resumen con los términos MP asociados al fenotipo de pérdida de neuronas dopaminérgicas.

Nombre del término	Id del término
Número reducido de neuronas dopaminérgicas	MP:0011448
Pérdida de neuronas dopaminérgicas	MP:0003244
Morfología anormal de la neurona dopaminérgica	MP:0003243
Morfología anormal de sustanca nigra pars compacta	MP:0013219
Disminución del tamaño de la sustancia negra	MP:0020543
Niveles disminuidos de dopamina	MP:0005643
Liberación anormal de dopamina sináptica	MP:0010149

Se observa al menos uno de los términos listados en la tabla 3 en 9 genes (26,5% del total).

- Trastornos del movimiento

Tabla 3- Tabla resumen con los términos MP asociados a fenotipos de trastornos del movimiento.

Nombre del término	Id del término
Comportamiento locomotor anormal	MP:0001392
Activación locomotora anormal	MP:0003313
Trastornos de la marcha	MP:0001406
Tremores	MP:0000745
Ataxia	MP:0001393
Distonia	MP:0005323
Bradikinesia	MP:0005156

Los Trastornos de la marcha (MP:0001406) han sido asociados a Parkinson en ratón en varios artículos (Mak, 2017). 19 genes (el 55,88%) están asociados a al menos uno de los términos listados en la Tabla 4.

- Cuerpos de inclusión de α-sinucleína

Aunque no todos los modelos de ratón presentan este fenotipo, es uno de los síntomas más estudiados de la EP. El término Cuerpo de inclusión de α-sinucleína (MP:0008493) está presente en 3 genes (8,82% del total).

4.2.1.3 RatMine

En cuanto a los términos MP, RatMine devuelve información de 5 genes (14,71% del total). Esto se debe a que los ortólogos en rata de la mayoría de nuestros genes figuran como RefSeq provisional (siguiendo la guía del NCBI) en RGD y están sujetos a validación para corregir errores de anotación y proporcionar anotación en un formato más consistente. Este hecho explica por qué se han obtenido términos DO, dado que se infieren a partir de su ortología.

PARK7 presenta los términos: Pérdida de neuronas dopaminérgicas (MP:0011448) y Coordinación / equilibrio motor anormal (MP:0001516). PINK1 está asociado con los términos: Pérdida de neuronas dopaminérgicas (MP:0011448), Trastornos de la marcha (MP:0001406) y Coordinación / equilibrio motor anormal (MP:0001516).

4.2.1.4 FlyMine

Esta API nos permite obtener los fenotipos a partir de los genes ortólogos en *D. melanogaster*.

De un total de 34 genes, 6 (17,65%) no devolvieron ningún ortólogo a partir de la herramienta HumanMine. En cuanto a los fenotipos de FlyBase, 18 genes (64,29%) presentaron el fenotipo Comportamiento locomotor defectuoso (FBcv:0000414) y 6 genes (21,34%) se localizaron en Neuronas dopaminérgicas (FBbt:00005131).

4.2.1.5 WormMine

Esta herramienta nos permite obtener los términos WBPhenotype a partir de los ortólogos de *C. elegans*.

De un total de 34 genes, 16 (47,06%) no devolvieron ningún ortólogo a partir de la herramienta HumanMine. De los 16 restantes, 3 (18,75%) presentan un fenotipo relacionado con el sistema locomotor y 2 (12,5%) presentan un fenotipo relacionado con la dopamina.

4.2.1.6 ZFIN

Se ha utilizado la propia página web de ZFIN para acceder notas de fenotipo anotadas manualmente de bibliografía disponible.

Se observó un fenotipo relacionado con la locomoción en 6 genes (26,09%), y otros 6 genes fueron encontrados en neuronas dopaminérgicas (26,09%).

4.2.2 Fenotipos no motores

Los principales fenotipos no motores de la EP son el deterioro cognitivo, los trastornos psiquiátricos, las anomalías sensoriales y los trastornos del sueño. Estos fenotipos han sido ampliamente estudiados, y nuestro objetivo en este apartado es analizar la presencia de términos de ontologías de fenotipo relacionados con los mismos en un listado de 34 genes asociados a la EP.

4.2.2.1 HumanMine

Los fenotipos no motores de la EP se han estudiado ampliamente en pacientes, confirmándose una correlación con diferentes formas de esta enfermedad en varios genes muy conocidos. En este subapartado presentamos términos HP asociados con estas cuatro categorías de síntomas no motores.

- Deterioro cognitivo

12 genes (35,29%) se asociaron con algún término relacionado con el deterioro cognitivo. Destaca el término Deterioro mental (HP:0100543) que estuvo presente en 7 genes (20,59%).

- Trastornos psiquiátricos

Un total de 27 genes (79,41%) poseían al menos un término relacionado con un trastorno psiquiátrico. De estos, los más prevalentes fueron Demencia (HP:0000726) con presencia en 11 genes (32,35%), Depresión (HP:0000716) en 9 genes (26,47%) y Alucinaciones (HP:0000738) en 6 genes (17,65%).

- Anomalías sensoriales

19 genes (55,88%) presentaron al menos un término relacionado con una anomalía sensorial. Los más prevalentes estaban relacionados con la visión. Destacamos el término Atrofia óptica (HP:0000648), presente en 7 genes (20,59%).

- Transtornos del sueño

Un total de 5 genes (14,71%) se asociaron con el término Alteración del sueño (HP:0002360). El gen DCTN1 presentaba además los términos Insomnio (HP:0100785) y Apnea del sueño (HP:0010535).

4.2.2.2 MouseMine

En este apartado pretendemos analizar los términos MP relacionados con los síntomas no motores asociados a la EP.

- Deterioro cognitivo

El deterioro cognitivo se investiga en roedores mediante varios protocolos que buscan evaluar la capacidad de aprendizaje y la memoria a corto y largo plazo de los individuos. 8 genes (25%) presentaron al menos un término relacionado con el aprendizaje o la memoria. Destacamos el término Aprendizaje espacial anormal (MP:0001463), presente en 6 genes (18,75%).

- Trastornos psiquiátricos

En modelos de ratón, este tipo de fenotipos se investigan causando estrés a los individuos y monitoreando su ansiedad. Un total de 3 genes (9,38%) se asociaron con un aumento de la ansiedad. Destacamos los términos Mayor respuesta relacionada con la ansiedad (MP:0001363) y Aumento de la tigmotaxis (MP:0002797), presentes en 2 genes (6,25%).

- Anomalías sensoriales

Un total de 10 genes (31,25%) se asociaron con al menos un término relacionado con anomalías sensoriales. De nuevo, los términos más prevalentes estaban relacionados con la visión, apareciendo en un total de 6 genes (15,63%).

- Trastornos del sueño

Se encontró tan solo 1 gen (3,13%) con un fenotipo aso ciado a trastornos del sueño.

- Trastornos de la autofagia

Tabla 4- Tabla resumen con los términos MP asociados a fenotipos de trastornos de la autofagia.

Nombre del término	Id del término
Autofagia anormal	MP:0008260
Mitofagia anormal	MP:0014046
Autofagia deteriorada	MP:0030940

Los trastornos en la autofagia han sido recientemente asociados a la EP (Gómez-Suaga, 2018; Hansen, 2018). Se observa al menos uno de los términos asociados a trastornos de autofagia presentados en la tabla 5 en 4 genes (11,76% del total).

4.2.2.3 RatMine

De los 5 genes que devuelve la herramienta RatMine, ninguno devuelve términos relacionados con fenotipos no motores.

4.2.2.4 FlyMine

De los 28 genes que devolvieron ortólogos de *D. melanogaster*, 14 genes (50%) tenían fenotipos que se manifestaban en una estructura del sistema visual. Destacamos el término Ojo (FBbt:00004508) presente en 12 genes (42,86%). 5 genes (17,86%) presentaban un fenotipo asociado a trastornos del sueño, 3 genes (10,71%) con el término Sueño defectuoso (FBcv:0000705) y 2 genes (7,14%) con el término Comportamiento circadiano defectuoso (FBcv:0000679). 3 genes (10,71%) se asociaron con el término Memoria defectuosa (FBcv:0000398).

4.2.2.5 WormMine

De los 16 genes que devolvían ortólogos en *C. elegans*, 1 gen (6,25%) presentaba el fenotipo Variante de respuesta al estrés del organismo (WBPhenotype:0000067), 3 genes (18,75%) tenían al menos un término hijo de Variante de alimentación (WBPhenotype:0002056), 2 genes (12,50%) se asociaron a un término hijo de Variante de comportamiento de alimentación (WBPhenotype:0000662), y 1 gen (6,25%) un fenotipo relacionado con la autofagia (WBPhenotype:0000736).

4.2.2.6 ZFIN

De los 34 genes analizados en este trabajo, tan solo 1 (2,94%) no se encontraba en la base de datos ZFIN. Un total de 10 genes (30,30%) se asocian con al menos un término ZFA relacionado con anomalías sensoriales. Destacamos el término Ojo (ZFA:0000107), presente en 8 genes (24,24%). En cuanto a los trastornos psiquiátricos, 4 genes (12,12%) están asociados al término Tigmotaxis (GO:0001966).

4.3 Evaluación de genes asociados al Parkinson mediante *Machine Learning*

Para la evaluación de genes asociados al Parkinson se han usado las herramientas HumanMine y MouseMine, como se explica en el apartado de materiales. Se han utilizado 9 términos HP y 5 términos MP que estaban presentes en más del 15% de genes conocidos, para que los resultados sean robustos.

El script está disponible en https://github.com/a-sabater/UMU-Bioinformatica-TFM/blob/master/TFM-Evaluación.Rmd.

Tabla 5- Tabla que muestra los ratios de pares gen-fenotipo HP y MP obtenidos para el conjunto de genes conocidos y predichos.

Conjunto de genes	Ratio de pares gen-fenotipo HP	Ratio de pares gen-fenotipo MP
Conocidos	4,0294	0,9706
Predichos	0,1538	0,4038

En la tabla 6, presentamos los ratios del número de pares gen-fenotipo para los términos HP y los términos MP, a partir del conjunto de genes conocidos y predichos como causantes de EP. En cuanto a los 1000 conjuntos de genes aleatorios, el p-valor obtenido para los términos HP fue de 0,041, y para los términos MP de 0,0009. Con un p-valor inferior a 0,5, concluimos que el ratio del número de pares gen-fenotipo HP y MP del conjunto de genes predichos es significativamente superior al que esperaríamos por simple azar.

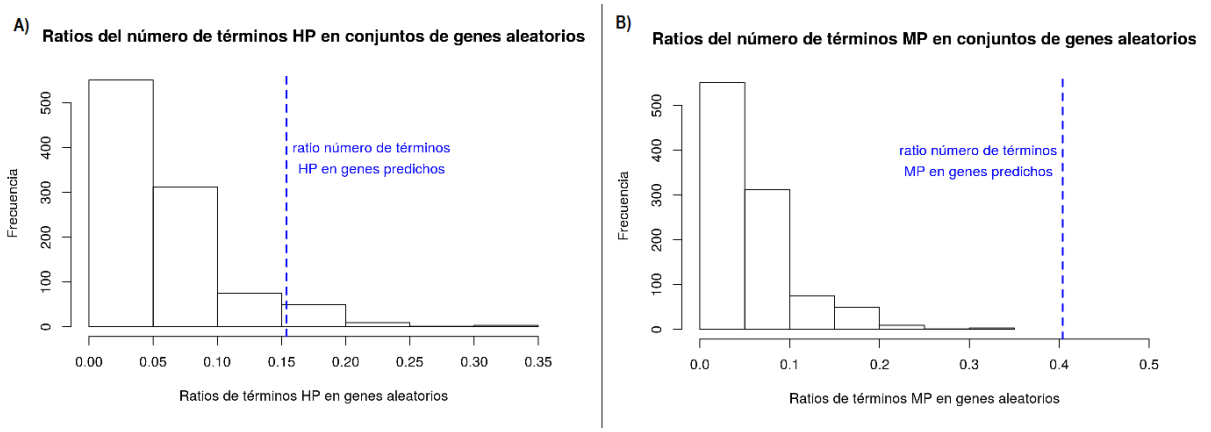

Figura 2- Histogramas correspondiente a: A) los ratios de pares gen-fenotipo HP de los conjuntos de genes tomados al azar, la línea discontinua azul corresponde con el ratio del conjunto de genes predichos; B) los ratios de pares gen-fenotipo MP de los conjuntos de genes tomados al aza, la línea discontinua azul representa el ratio del conjunto de genes predichos.

En la figura 2 presentamos los histogramas correspondientes a los conjuntos de genes generados aleatoriamente y la posición del ratio de pares gen-fenotipo en el conjunto de genes predichos marcada con una línea discontinua azul. Nos permite ver visualmente que, en ambos casos, el ratio en el conjunto de genes predichos como causantes de EP es mayor de lo esperado por azar.

5. DISCUSIÓN

El método propuesto en el presente trabajo, basado en recabar términos fenotípicos a partir de bases de datos de animales modelo, pretende dotar de robustez a la búsqueda de genes asociados a enfermedades. Dado que, como se ha dejado constancia con anterioridad, el sistema nervioso de humanos es muy complejo y la ética en la investigación en humanos no permite la experimentación profunda y sistemática en humanos; el uso de animales modelos ha sido una necesidad en la investigación biomédica. La capacidad de utilizar la investigación en diversas especies animales permite obtener más información que aquella investigación que solo explora a la especia humana, y, además, analizar la evolución de los sistemas biológicos.

No obstante, cabe destacar que presenta una serie de sesgos y limitaciones. Las bases de datos enumeradas en nuestro catálogo actualizan su información mediante la curación de artículos publicados, por lo que la presencia de determinados genes en dichas bases de datos se ve sesgada por los ámbitos que despiertan mayor interés para los investigadores. A su vez, la información asociada a estos genes depende de los métodos utilizados en cada especie animal y a las tendencias en la ciencia actual. También es importante constatar que las diferencias fisiológicas entre diferentes especies animales, como se han mencionado en más detalle en el caso de la EP con anterioridad.

Por mostrar términos prevalentes en al menos un 50% de genes, se decidió utilizar las bases de datos HumanMine y MouseMine en la evaluación de genes predichos como causantes de EP. Estas bases de datos fueron las únicas en mostrar más de un término fenotípico que estuviera presente en al menos el 15% de los genes conocidos, por lo que se consideró que eran las únicas que caracterizaban de forma adecuada la EP. En la caracterización de paneles de Genomics England observamos la presencia de un mayor número de términos fenotípicos HP asociados a Parkinson con respecto a los términos MP. También son más prevalentes los términos relacionados con fenotipos motores a los no motores y un total de 9 términos HP y 5 términos MP se encontraban en más del 15% de los genes conocidos. Dado que se trata de genes conocidos como causantes de Parkinson, han sido muy estudiados en humanos con el fin de buscar dianas terapéuticas, la caracterización de las bases fisiológicas de la enfermedad y otros muchos ámbitos.

En el conjunto de genes conocidos como causantes de EP se observa un ratio de pares gen-fenotipo de 4,0294 (es decir, por cada gen conocido se observan 4 términos fenotípicos asociados al mismo, de media para términos HP), cuatro veces superior al 0,9706 observado en los términos MP. En los genes predichos observamos el fenómeno opuesto, donde el ratio de términos MP de 0,4038 es superior al 0,1538 de los términos HP. Los genes predichos por Botía y colaboradores no están asociados a EP, por lo que no se han estudiado en humanos con relación a este trastorno. Esto explica perfectamente el ratio de 0,1538, dado que era esperado que no estuvieran asociados a sus términos HP asociados a este trastorno. No obstante, las pruebas en ratones permiten obtener información del comportamiento y del estado cognitivo en la mayoría de investigaciones biomédicas. Esto nos es muy útil a la hora de estudiar enfermedades neurológicas.

En cuanto al método utilizado, observamos que los genes predichos tienen significativamente más términos de ontologías fenotípicas de lo que cabría esperar por azar tanto en la base de datos HumanMine como en MouseMine. Este hecho nos

hace pensar que el algoritmo creado por Botía y colaboradores en 2018 sería adecuado para la búsqueda de nuevos genes implicados en enfermedades neurológicas. No obstante, destacamos que solo se ha evaluado una enfermedad neurológica y que se deberían evaluar un mayor número de trastornos para testar la robustez tanto del método empleado en este trabajo como del algoritmo de *machine learning* presentado por los autores originales.

6. CONCLUSIONES

El objetivo de este trabajo es definir un mecanismo genérico de evaluación de predicciones de asociaciones (gen, enfermedad) para enfermedades genéticas monogénicas. Y hemos demostrado su utilidad mediante su aplicación a la evaluación de predicciones de nuevas asociaciones de genes con variantes monogénicas del Parkinson. La metodología de predicción está ilustrada en el trabajo de Botía y colaboradores (Botía, 2018).

La caracterización previa de genes conocidos para el trastorno de interés es un requisito indispensable a la hora de usar este método. El método de minería de bases de datos genómicas de humano y de ratón realizado en este trabajo ha permitido la obtención de términos que definen de forma bastante precisa varios síntomas importantes de dicho trastorno. No obstante, han de tenerse en cuenta los sesgos inherentes al uso de animales modelo y de bases de datos genómicas. Así como destacamos que en este trabajo solo se ha caracterizado una enfermedad neurológica, por lo que sería necesario caracterizar más trastornos neurológicos con el fin de evaluar la robustez del método de minería de datos presentado, así como del algoritmo de *machine learning* creado por Botía.

BIBLIOGRAFÍA

- Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2018). OMIM. org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research*, 47(D1), D1038-D1043.
- Botia, J. A., Guelfi, S., Zhang, D., D'Sa, K., Reinolds, R., Onah, D., ... & Houlden, H. (2018). G2P: Using machine learning to understand and predict genes causing rare neurological disorders. *bioRxiv*, 288845.
- Bult C.J., Blake J. A., Smith C. L., Kadin JA, Richardson J. E., the Mouse Genome Database Group. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 8;47 (D1): D801–D806.
- Creed, R. B., & Goldberg, M. S. (2018). New Developments in Genetic rat models of Parkinson's Disease. *Movement Disorders*, *33*(5), 717-729.
- Dabool, L., Juravlev, L., Hakim-Mishnaevski, K., & Kurant, E. (2019). Modeling Parkinson's disease in adult Drosophila. *Journal of neuroscience methods*, *311*, 89-94.
- Ferguson, L. W., Rajput, A. H., & Rajput, A. (2016). Early-onset vs. Late-onset Parkinson's disease: A Clinical-pathological Study. *Canadian Journal of Neurological Sciences*, *43*(1), 113-119.
- Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2017). The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, *19*(5), 1008-1021.
- Gómez-Suaga, P., Bravo-San Pedro, J. M., González-Polo, R. A., Fuentes, J. M., & Niso-Santano, M. (2018). ER–mitochondria signaling in Parkinson's disease. *Cell death & disease*, 9(3), 337.
- Grace, A. A. (2016). Dysregulation of the dopamine system in the pathophysiology of schizophrenia and depression. *Nature Reviews Neuroscience*, 17(8), 524–532.
- Groza, T., Köhler, S., Moldenhauer, D., Vasilevsky, N., Baynam, G., Zemojtel, T., ... & Vasant, D. (2015). The human phenotype ontology: semantic unification of common and rare disease. *The American Journal of Human Genetics*, *97*(1), 111-124.
- Hansen, M., Rubinsztein, D. C., & Walker, D. W. (2018). Autophagy as a promoter of longevity: insights from model organisms. *Nature Reviews Molecular Cell Biology*.
- Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., ... & McLaren, S. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498.
- Howe, D. G., Bradford, Y. M., Eagle, A., Fashena, D., Frazer, K., Kalita, P., ... & Pich, C. (2016). The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching. *Nucleic acids research*, *45*(D1), D758-D768.
- Jans, V., Dondorp, W., Goossens, E., Mertes, H., Pennings, G., & de Wert, G. (2018). Balancing animal welfare and assisted reproduction: ethics of preclinical animal research for testing new reproductive technologies. *Medicine, Health Care and Philosophy*, *21*(4), 537-545.
- Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., ... & Micklem, G. (2014). InterMine: extensive web services for modern biology. *Nucleic acids research*, *42*(W1), W468-W472.
- Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., … Schriml, L. M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, *43*(Database issue), D1071–D1078. doi:10.1093/nar/gku1011.
- Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., … Robinson, P. N. (2017). The Human Phenotype Ontology in 2017. *Nucleic acids research*, *45*(D1), D865–D876. doi:10.1093/nar/gkw1039.
- Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J. P., ... & Osumi-Sutherland, D. (2018). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research*, *47*(D1), D1018-D1027.
- Law, M., & Shaw, D. R. (2018). Mouse Genome Informatics (MGI) Is the International Resource for Information on the Laboratory Mouse. In *Eukaryotic Genomic Databases* (pp. 141-161). Humana Press, New York, NY.
- Lee, R., Howe, K. L., Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., … Sternberg, P. W. (2018). WormBase 2017: molting into a new stage. *Nucleic acids research*, *46*(D1), D869–D874. doi:10.1093/nar/gkx998.
- Lee, M. Y., Kim, T. K., Walters, K. A., & Wang, K. (2019). A biological function based biomarker panel optimization process. *Scientific reports*, *9*(1), 7365.
- Lei, P., Ayton, S., Finkelstein, D. I., Spoerri, L., Ciccotosto, G. D., Wright, D. K., ... & Roberts, B. R. (2012). Tau deficiency induces parkinsonism with dementia by impairing APP-mediated iron export. *Nature medicine*, *18*(2), 291.
- Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., ... & Rana, D. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome biology*, *8*(7), R129.
- Lyne, R., Sullivan, J., Butano, D., Contrino, S., Heimbach, J., Hu, F., ... & Balakrishnan, R. (2015). Cross-organism analysis using InterMine. *genesis*, *53*(8), 547-560.
- Mak, M. K., Wong-Yu, I. S., Shen, X., & Chung, C. L. (2017). Long-term effects of exercise and physical therapy in people with Parkinson disease. *Nature Reviews Neurology, 13(11), 689–703.*

- Maulik, M., Mitra, S., Bult-Ito, A., Taylor, B. E., & Vayndorf, E. M. (2017). Behavioral Phenotyping and Pathological Indicators of Parkinson's Disease in *C. elegans* Models. *Frontiers in genetics*, *8*, 77. doi:10.3389/fgene.2017.00077
- Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2016). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, *45*(D1), D183-D189.
- Motenko H., Neuhauser S., O'Keefe M., Richardson J. (2015). MouseMine: a new data warehouse for MGI. *Mammalian Genome*. 26(7-8): 325:30.
- Nagoshi E. (2018). *Drosophila* Models of Sporadic Parkinson's Disease. *International journal of molecular sciences*, *19*(11), 3343.
- Ooms J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.
- Ooms J. (2019). curl: A Modern and Flexible Web Client for R. R package version 4.0. https://CRAN.R-project.org/package=curl
- Pajusalu, S., Kahre, T., Roomere, H., Murumets, Ü., Roht, L., Simenson, K., ... & Õunap, K. (2018). Large gene panel sequencing in clinical diagnostics—results from 501 consecutive cases. *Clinical genetics*, *93*(1), 78-83.
- Pinho, B. R., Reis, S. D., Hartley, R. C., Murphy, M. P., & Oliveira, J. M. (2019). Mitochondrial superoxide generation induces a parkinsonian phenotype in zebrafish and huntingtin aggregation in human cells. *Free Radical Biology and Medicine*, *130*, 318-327.
- Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., ... & Lang, A. E. (2017). Parkinson disease. *Nature reviews Disease primers*, 3, 17013.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Robea, M. A., Strungaru, Ș. A., Lenzi, C., Nicoara, M., & Ciobica, A. (2018). The Importance of Rotenone in Generating Neurological and Psychiatric Features in Zebrafish-Relevance for a Parkinson's Disease Model. *Academy of Romanian Scientists Annals-Series on Biological Sciences*, *7*(1), 59-67.
- Romanelli, R. J., Williams, J. T., & Neve, K. A. (2010). Dopamine receptor signaling: intracellular pathways to behavior. In *The dopamine receptors* (pp. 137-173). Humana Press, Totowa, NJ.
- Schindelman, G., Fernandes, J. S., Bastiani, C. A., Yook, K., & Sternberg, P. W. (2011). Worm Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community. *BMC bioinformatics*, *12*(1), 32.
- Schulte, E. C., Fukumori, A., Mollenhauer, B., Hor, H., Arzberger, T., Perneczky, R., ... & Eckstein, G. (2015). Rare variants in β-Amyloid precursor protein (APP) and Parkinson's disease. *European Journal of Human Genetics*, *23*(10), 1328.
- Shah, P. D., & Nathanson, K. L. (2017). Application of panel-based tests for inherited risk of cancer. *Annual review of genomics and human genetics*, *18*, 201-227.
- Shen, H. (2018). The labs growing human embryos for longer than ever before. Nature, 559, 19-22.
- Shimoyama, M., De Pons, J., Hayman, G. T., Laulederkind, S. J., Liu, W., Nigam, R., … Jacob, H. (2015). The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic acids research*, *43*(Database issue), D743–D750. doi:10.1093/nar/gku1026.
- Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., ... & Stepan, R. (2012). InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, *28*(23), 3163-3165.
- Smith, C. L., & Eppig, J. T. (2015). Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *Journal of biomedical semantics*, *6*(1), 11.
- Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., ... & Kaufman, T. C. (2018). FlyBase 2.0: the next generation. *Nucleic acids research*, *47*(D1), D759-D765.
- Ünal, İ., & Emekli-Alturfan, E. (2019). Fishing for Parkinson's Disease: A review of the literature. *Journal of Clinical Neuroscience.* doi:10.1016/j.jocn.2019.01.015
- Van Slyke, C. E., Bradford, Y. M., Westerfield, M., & Haendel, M. A. (2014). The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *Journal of biomedical semantics*, 5(1), 12.
- Vink, R. (2018). Large animal models of traumatic brain injury. *Journal of neuroscience research*, *96*(4), 527-535.
- Wickham H., François R., Henry L. and Kirill Müller K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr
- Williamson, J. C., Bonello, M., & Larner, A. J. (2018). Genetic investigation in dementia: new interpretive challenges. *Progress in Neurology and Psychiatry*, *22*(4), 6-8.
- Wong, Y. C., & Krainc, D. (2017). α-synuclein toxicity in neurodegeneration: mechanism and therapeutic strategies. *Nature Medicine*, 23(2), 1–13.

- Xue, Y., Ankala, A., Wilcox, W. R., & Hegde, M. R. (2015). Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genetics in Medicine*, *17*(6), 444.

Recursos informáticos

- FDA. (2018) Why are animals used for testing medical products? Retrieved from: https://www.fda.gov/about-fda/fda-basics/why-are-animals-used-testing-medical-products