UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

“Identificación de Biomarcadores asociados a demencia para proponer un sistema biológico neuronal y circulante para su estudio”

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Ciencias

PRESENTA:
Erick Cuevas Fernández

TUTOR PRINCIPAL
Dr. Heriberto Manuel Rivera
Facultad de Nutrición, UAEM

MIEMBROS DEL COMITÉ TUTOR

Dra. Leonor Pérez Martínez
Instituto de Biotecnología, UNAM

Dr. Alejandro Garcíarrubio Granados
Instituto de Biotecnología, UNAM

Cuernavaca, Morelos, México. Marzo 2020.

AGRADECIMIENTOS

“Que el fin del mundo te pille bailando…”

-Joaquín Sabina

Agradezco a todos aquellos que formaron parte intelectual y emocional durante este proceso. A mi padre y madre por enseñarme los valores y actitudes que me han hecho llegar hasta aquí. A mi hermano mayor, un mentor y héroe para mí. Mi hermana menor, ser su ejemplo es una gran motivación. Al tutor principal de esta esta tesis, -lo logramos Doc-, gracias por inculcarme la mentalidad de seguir adelante, gracias por enseñarme a nunca decir “no puedo” ante cualquier adversidad. A mis compañeros de laboratorio, que durante estos dos años se han vuelto parte de mi familia. A mis amigos de la vida que en borracheras, fiestas y partidos me enseñaron el valor de una amistad. A los investigadores implicados en la revisión y evaluación de este trabajo, es hermoso leer y escuchar opiniones objetivas que alimentan la curiosidad y hacen mejorar. Agradezco a los que se fueron y a los que se quedaron.

Resumen:

La demencia es una enfermedad mental, crónica y degenerativa, su principal característica es la neurodegeneración, causando perdida de habilidades cognitivas y facultades mentales. Su incidencia va en aumento a nivel mundial producto de un diagnóstico tardío. Los polimorfismos de un solo nucleótido, variaciones genéticas a nivel poblacional, podrían arrojar pistas respecto a la etiología de la enfermedad para incrementar su entendimiento, y al mismo tiempo, proporcionar información que permita generar herramientas de diagnóstico temprano. Al hacer uso de la "big data" y "machine learning" se encontraron polimorfismos no reportados para las múltiples manifestaciones de la demencia, este trabajo plantea proponer el papel de los SNPs en el contexto de la función de un sistema celular.

Palabras clave

Demencia; SNP: Single Nucleotide Polymorphism; Aprendizaje de máquina; Análisis de textos; Unidad Neurovascular; Enfermedad de Alzheimer; Esclerosis Múltiple; Demencia Vascular; Demencia Frontotemporal; Demencia por cuerpos de Lewy; Parkinson.

Contenido (Índice)

Lista de tablas

Lista de figuras

Abreviaturas y siglas usadas

SNP: Single Nucleotide Polymorphism

AD: Alzheimer´s Disease (Enfermedad de Alzheimer)

MS: Multiple Sclerosis (Esclerosis Múltiple)

LBD: Lewy body Disease (Enfermedad de cuerpos de Lewy)

FTD: Frontotemporal Dementia (Demencia frontotemporal)

VaD: Vascular Dementia (Demencia Vascular)

SNP-Cell CRAD: SNP-Cell Classification and Risk Association of Dementia

GWAS: Genome-wide Association Study

MeSH: Medical Subject Header

Introducción

La demencia es un padecimiento provocado por múltiples factores de naturaleza crónica y progresiva, caracterizado por la pérdida de habilidades cognitivas y emocionales (*Organización Mundial de la Salud CIE-10, 1992*). Existen otras manifestaciones neuropsiquiátricas tales como alteraciones motoras, de la conducta, depresión, ansiedad, alucinaciones y/o delirium como consecuencia de un proceso de degeneración neuronal principalmente asociados a este síndrome (*Gallegos et al., Gutiérrez y Arrieta, 2014*). Como consecuencia de lo anterior, las personas que la padecen, así como su entorno familiar se ven afectados en su calidad de vida y economía. Se estima que el gasto de salud será para el 2030 de 2 trillones de dólares en el mundo (*Alzheimer's Association. World Alzheimer Report 2015*).

En los años 70, se demostró que la demencia senil era indiscernible clínica y neuropatológicamente a casos de demencia en edades menores a los 60 años (*Blessed G et al.1968*). Por lo tanto, fue clara la necesidad de desarrollar nuevas alternativas de diagnóstico diferencial. Actualmente la demencia está catalogada como un trastorno neurocognitivo mayor (*DSM-V. Asociación Americana de Psiquiatría, 2013*). Existen múltiples tipos de demencia que comparten síntomas y que tienen su propia patología (*Zanni & Wick, 2007*). Esta falta de límites entre los diferentes tipos de demencia presenta el inconveniente de dificultar el diagnóstico preciso y como consecuencia un tratamiento ineficaz.

Para hacer frente a esta disyuntiva se han hecho diversos avances en nanotecnología, interconectividad y biotecnología. Estos avances representan un incremento en la cantidad disponible de información generada sin precedentes. Un ejemplo de lo anterior es la información genética y clínica que se genera a través de dispositivos electrónicos (*Kavakiotis et al.,2017*). Para manejar esta inmensa cantidad de datos ("Big Data", BD por sus siglas en inglés) (*Peek, Holmes, & Sun, 2014*), es necesario utilizar herramientas de inteligencia artificial (IA). Esta capacidad computacional puede ser definida en diferentes tipos de aprendizaje de maquina o "machine learning", tal como aprendizaje asistido o supervisado y no supervisado. Aprendizaje de maquina es un método para el análisis de datos que nos permite generar modelos analíticos automáticamente, capaz de realizar predicciones mediante el reconocimiento de patrones con la intención de hacer sentido a la interpretación de un fenómeno biológico, se realiza mediante un proceso de tratamiento de datos múltipasos (selección, preprocesamiento, transformación, minería de datos o data mining, interpretación y evaluación) (*Sanchéz-Mendez,J., et al 2017*). Estos métodos ya han sido puestos a prueba con éxito para el diagnóstico de enfermedades crónico degenerativas, tal como la diabetes mellitus tipo 2 (*Kavakiotis et al., 2017*) y síndrome metabólico (*Sanchéz-Mendez,J., et al. 2017*).

Este trabajo plantea implementar estrategias de aprendizaje de maquina, se utilizan datos de información genética poblacional, específicamente polimorfismos y estudios de asociación de genoma completo (Genome-wide Association Study, GWAS), con la intención de identificar y proponer nuevos biomarcadores para el diagnóstico diferencial y temprano de demencia. Se parte de la idea de que los polimorfismos son moléculas de características medibles y evaluables como un indicador de una condición normal o patológica asociada a un proceso biológico o respuesta farmacológica a una intervención terapéutica que puede ser soluble o no (Atkinson et al. 2001).

Antecedentes

Demencia

La demencia es una enfermedad mental, crónica y progresiva, que esta catalogada como un desorden neurocognitivo mayor (*DMS-V 2013*) y se caracteriza por un proceso gradual de neurodegeneración, con perdida de habilidades cognitivas, perturbación de las facultades mentales y un impacto directo de la autonomía del la persona que la padece (*Alistair Burns et al., 2009*). A nivel molecular se presenta principalmente inflamación, y agregados proteícos.

La demencia se clasifica según sea la causa en: enfermedad de Alzheimer **(AD)**, demencia frontotemporal **(FTD)**, demencia por cuerpos de Lewy o Parkinson **(LBD)**, esclerosis múltiple **(MS)**, demencia vascular **(VaD)**, enfermedad de priones, enfermedad de Huntington y demencia a causa de VIH (*ICD-11 2018*). Las primeras cinco demencias son las que presentan una mayor incidencia a nivel mundial. Se estima que para el 2050 la incidencia de demencia rebasará los 131.5 millones a nivel mundial. Actualmente se calcula, que cada 3 segundos, alguien desarrolla demencia en el mundo (*Alzheimer's Association. World Alzheimer Report 2015*).

El diagnóstico de la demencia inicia con la evaluación de habilidades cognitivas y psicológicas mediante exámenes psicométricos.

Posteriormente para un diagnóstico especifico de los distintos tipos de demencia, se evalúan biomarcadores en líquido cefalorraquídeo para confirmar la acumulación de agregados mieloides y proteínas tau, tomografía por emisión de positrones (PET) para censar el daño cerebral, y resonancia magnética (RM) (*Sperling RA, et al. 2011, Gordon E. et al. 2016*). A nivel fisiológico es difícil encontrar una frontera para discernir entre los distintos tipos de demencia, lo que genera un diagnóstico con poca sensibilidad y además tardío. Esto implica que la mayoría de los casos tengan un diagnóstico *postmortem*.

En el 2011 se calculó que a nivel global 35,600 millones de personas fueron diagnosticadas con demencia. Lo anterior representó un costo estimado de 604,000 millones de dólares para el sector salud. Un año más tarde la demencia fue declarada como prioridad de salud pública por la Organización Mundial de la Salud (OMS) con una estimación del doble de casos cada 20 años (*Dua et al., 2013*). En México se estima una incidencia de 555.53 por cada 100,000 habitantes y se proyecta que para el 2050 la cifra será de al menos el triple, por lo que el impacto de esta enfermedad en los sistemas económico, social y de salud no tendrá precedentes (*Institute for Health Metrics and Evaluation 2017*).

Dadas estas cifras, es de vital importancia entender los mecanismos que desencadenan esta enfermedad para poder tratarla y prevenirla. Desafortunadamente

el estudio de la demencia en humanos es limitado, ya que el estudio del cerebro implica una gran complejidad dada su estructura y la imposibilidad de estudiar mecanismos cerebrales *in-vivo* (*Bassett & Gazzaniga, 2011*). La búsqueda de nuevos biomarcadores posibilitaría el diagnóstico diferencial y oportuno de los tipos de demencia, así como formular hipótesis de mecanismos fisiológicos y tratamientos farmacológicos. Para ello, un biomarcador ideal debe detectar un rasgo patológico fundamental de la enfermedad, debe validarse en cohortes probadas patológicas y debe ser preciso, confiable, económico y detectable a través de un procedimiento no invasivo y simple de realizar (*Bateman et al., 2012; Novelli, Ciccacci, Borgiani, Amati, & Abadie, 2008)*. La detección de biomarcadores en un paciente con un deterioro cognitivo leve podría lograr circunscribir el origen del déficit, o en etapas asintomáticas poder prevenir la incidencia de demencia (*Sperling RA et al. 2011*).

Recientemente, se han contemplado a las enfermedades como entidades que se originan por factores genéticos y factores ambientales (*Baye, Abebe & Wilke, 2011*), dado la complejidad de la demencia y la etiología múltiple que la caracteriza, entender y encontrar los factores genéticos que promueven su aparición podría ayudar a la identificación de biomarcadores genéticos que nos permitan contenderla.

Polimorfismos de un solo nucleótido y estudios de asociación de genoma completo

“La mayoría de las enfermedades comunes, tales como la diabetes, el cáncer, las enfermedades del corazón, los accidentes cerebrovasculares, la depresión y el asma, son afectadas por muchos genes y factores ambientales. Aunque cualquier par de personas no emparentadas tienen alrededor del 99.9 por ciento de las secuencias de su ADN en común, el 0.1 por ciento restante es importante porque contiene las variantes genéticas que influyen en cómo las personas se diferencian entre sí en su riesgo de enfermedades o en la respuesta a medicamentos. El descubrimiento de las variantes en las secuencias de ADN que contribuyen al riesgo de enfermedades, ofrece una excelente oportunidad para entender las causas complejas de muchas enfermedades comunes en los seres humanos” (*Proyecto Internacional HapMap, 2015*).

El genoma humano alberga información acumulada de más de 6 millones de años de evolución en 46 cromosomas, en los cuales se encuentran 3.2 mil millones de pares de bases. Se estima que cada 300 pares de bases se genera un polimorfismo de un solo nucleótido (Single nucleotide polymorphism o **SNP**). Un SNP es un tipo de variabilidad genética generada por una mutación puntual que se establece en al menos el 1% de la población. Existen en regiones codificantes y no codificantes, sinónimos y no sinónimos.

Los SNPs pueden afectar los niveles de expresión genética, modificar la traducción de los RNA mensajeros, alterar el corte y empalme en el “splicing”, cambiar la estabilidad

de los RNA mensajeros y microRNAs; es por ello que se les ha prestado atención en el papel que desempeñan en el desarrollo de patologías de etiología múltiple (*Dooley MA et al., 2003; Kunes J et al., 2009; Andrassi MG et al., 2009; Ramirez-Bello et al., 2011; Su MW et al., 2012*). Durante los últimos 20 años de este siglo, las metodologías analíticas que han permitido la detección de SNPs (*Talenti A et al., 2016; Altshuler D et al., 2000; Griffin T J et al., 2000; Drabovich A et al., 2006; Tahira T et al., 2009*) son las siguientes:

- Polimorfismo en la longitud de fragmentos de restricción (SNP-RFLP).
- Técnicas de secuencias de DNA, como la secuenciación de Sanger o técnicas de siguiente generación.
- Electroforesis capilar.
- Polimorfismo de conformación de cadena simple.
- Cromatografía Liquida Desnaturalizante de Alto Rendimiento (HPLC).
- Espectrometría de masas.

Gracias a estas herramientas, en 2002 inició el proyecto HapMap con el fin de reconocer los distintos haplotipos del genoma. Éste proyecto generó 6 millones de SNPs nuevos en distintas poblaciones que hoy en día se encuentran bases de datos de libre acceso (*International HapMap Project 2015*). Todo este proyecto tuvo un costo aproximado a los $100 millones de dólares. Seguido del proyecto HapMap, en 2008 el proyecto de los "*1000 genomas*", basado en tecnologías de secuenciación de siguiente generación, propuso el objetivo de generar un catalogo con todas las variaciones genéticas humanas con mayor resolución, éste proyecto se estima que tuvo un costo entre $30 millones y $120 millones de dólares (*1000 Genomes Project 2008*).

En 2005 iniciaron los estudios de asociación de genoma completo, para identificar asociaciones entre variaciones genéticas y una muestra poblacional con algún padecimiento (*Peter M., et al., 2017*). El mecanismo mediante el cual la variante genética se asocia con el fenotipo no puede ser obtenido mediante este enfoque, no obstante, se han recabado alrededor de 52,000 asociaciones de variantes genéticas estadísticamente significativas superiores al umbral del valor P de todo el genoma de $5x10^{-8}$. Estos estudios han ayudado a revelar mecanismos que antes se desconocían en enfermedades autoinmunes como la enfermedad de Crohn, en el cual se hallaron dos polimorfismos, rs2241880 y rs1000113 (*Hampe J. et al., 2007; Wellcome Trust Case Control Consortium, 2007*), en el gen ATG16L1 e IRGM respectivamente. Éstos hallazgos mostraron la importancia del rol de la autofagia ya que el rs2241880 promueve la ruptura de la caspasa-3 y disminuye la autofagia durante un estrés celular, alterando la producción de citocinas inflamatorias, por tanto estableciendo inflamación crónica (*Murthy A., et al., 2014*).

Se estima que existen en la base de datos de SNPs (*Sherry ST, et al. 2001*) alrededor de 60 millones de variantes, y 381,000 se localizan en la región del gen. Bases de datos como ClinVar, un archivo público de informes de variantes genéticas relacionadas a un fenotipo cuenta aproximadamente con 900,538 variantes reportadas usando el genoma de referencia GRCh37/hg19, y GWAS catalog con cerca de 52,000 polimorfismos reportados (*Buniello A et al., 2019*); en la *fig.* 1 se resume esta información. Se contempla que aproximadamente 500 SNPs están relacionados a algún tipo de demencia. Actualmente se han asociado variaciones genéticas a distintos tipos de demencia, por ejemplo en Alzheimer se han identificado mutaciones en APOE, PSEN1, PSEN2 y APP como factor de riesgo (*Bertram & Tanzi, 2008*); en demencia frontotemporal se han asociado mutaciones en GRN, c9orf72 y MAPT como agente de riesgo para desarrollar este tipo de demencia (*Meeter, Kaat, Rohrer, & van Swieten, 2017*). También se han ocupado estrategias de la Big Data y meta análisis para identificar biomarcadores en enfermedades psiquiátricas como el trastorno de bipolaridad y esquizofrenia (*Kalia & Costa, 2015*). Wheeler *et al*., han asociado 25 polimorfismos al riesgo de padecer Alzheimer, mediante el análisis de 2.6 millones de SNPs (*Wheeler et al*., *2010*).

Realizar un estudio de asociación de genoma completo requiere una inversión institucional de mas de 100,000 USD y un tamaño de población de al menos el 1% de la población a estudiar, es por ello que se requiere del desarrollo de herramientas que nos permitan utilizar y analizar los datos ya existentes, para entender la complejidad de una enfermedad y dilucidar los mecanismos que desencadenan las variantes genéticas en los padecimientos. La identificación de patrones mediante algoritmos automatizados y de aprendizaje de máquina representan una excelente alternativa que podría ayudar al análisis de grandes volúmenes de datos para poder clasificar la información.

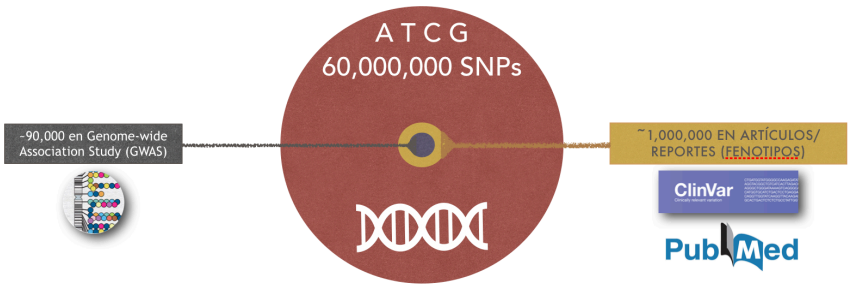


Fig. 1 Esquema de la cantidad de SNPs que existen y en dónde se encuentran; modificado de Amalio Telenti 2018

Aprendizaje de maquina (Machine Learning)

El reconocimiento de patrones consiste en asignar etiquetas (clases) a objetos (instancias) e identificarlos. Los objetos se describen mediante un conjunto de medidas o propiedades denominadas atributos o características (*Kuncheva 2004*), de este modo

cada objeto es un vector de *n* dimensiones, y un conjunto de observaciones o instancias (*m*) es una matriz de *n x m* dimensiones. Existen dos aproximaciones, aprendizaje supervisado y aprendizaje no supervisado, en el primero se especifica la clase a cada objeto, en tanto que esta no se especifica en el segundo (*Gareth Jaimes, "An introduction to statistical learning" 2013*). El "aprendizaje de maquina" es un conjunto de algoritmos que permiten encontrar las fronteras entre múltiples instancias con distintas clases. En la *fig.2* se muestra un ejemplo de dos dimensiones en donde se observa un hiperplano o frontera que divide dos clases, y cada frontera será generada de manera distinta según sea el algoritmo que se utilice. En color rojo se representa una clase y en azul otra clase; los ejes de la *fig. 2* solo muestran como ejemplo las magnitudes que se pueden cuantificar en las distintas clases, el aprendizaje de maquina permite dividir la clase roja y azul automáticamente, y su virtud radica en generar esta frontera (hiperplano) en múltiples dimensiones con miles de datos.

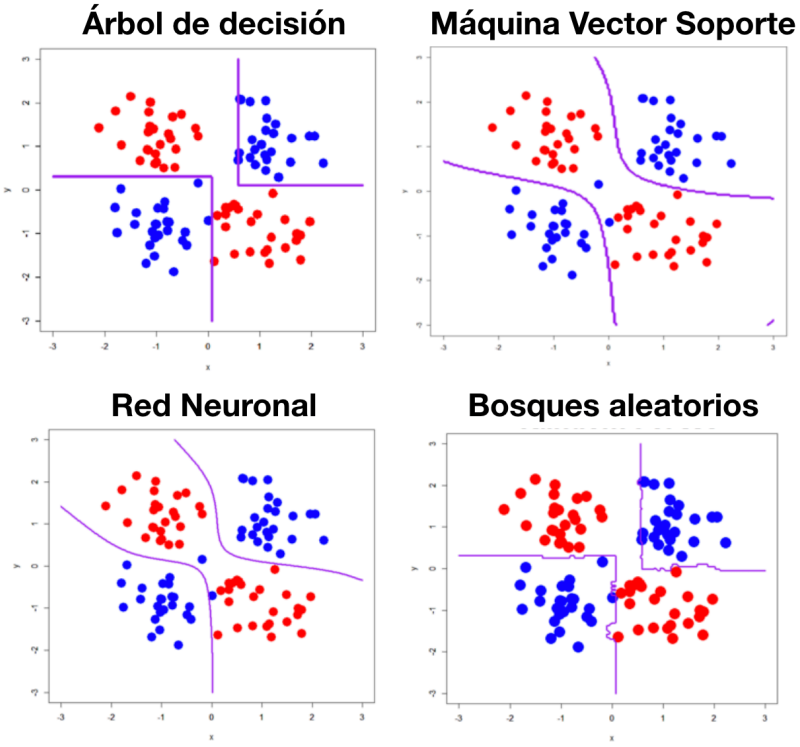


Fig.2 Hiperplano entre dos clases en dos dimensiones utilizando distintos algoritmos: Arboles de decisión, Maquinas Vector Soporte, Redes Neuronales y Bosques Aleatorios. (Modificada de Takashi J 2014; https://tjo-en.hatenablog.com*)*

Al representar una instancia u observación mediante distintas características, se puede generar una representación vectorial multidimensional que permite agrupar al objeto observado con una etiqueta/clase. Con este tipo de metodologías se puede representar a enfermedades o fenómenos como un conjunto de datos distribuidos en patrones.

Existen distintos algoritmos dentro del aprendizaje de máquina supervisado que permiten encontrar el hiperplano para separar nuestras clases. Entre ellos se encuentran los arboles de decisión, los cuales se basan en la toma de decisiones mediante la partición de los atributos. *Bosques Aleatorios* son básicamente la generación de múltiples arboles de decisión y la clasificación se da por una ponderación entre todos los arboles. *AdaBoost* (Adaptive Boosting) es un meta-algoritmo, que permite la interacción con otros clasificadores para optimizar su rendimiento al clasificar. *Maquinas de Vector Soporte* permite generar un hiperplano mediante distintas funciones kernel (núcleo). Se conoce como "modelo", a un algoritmo entrenado con datos, es decir, un algoritmo que ya "aprendió"; en cuanto al aprendizaje, "*se dice que una computadora aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de rendimiento P, si su desempeño en las tareas en T, medido por P, mejora con la experiencia E*" (*Tom Mitchell 2011*). Para el caso de aprendizaje supervisado, una manera de evaluar el rendimiento de cada modelo generado es mediante matrices de confusión, en donde se representa el número de predicciones de cada clase y las instancias en su clase real, esto nos genera la cantidad de verdaderos y falsos negativos, así como verdaderos y falsos positivos.

Brevemente, el enfoque de aprendizaje de maquina no supervisado permite generar agrupamientos de los datos sin etiqueta dadas las distancias entre ellos en un espacio multidimensional, entre los algoritmos existentes se encuentra *k-means*, *análisis de componentes principales, t-SNE (t-distributed stochastic neighbor embedding) y CLARA (Clustering Large Applications)*, entre otros.

En cuanto a los polimorfismos de un solo nucleótido, el aprendizaje de maquina se ha utilizado para la identificación de SNPs asociados a metilación en el estudio "Genetics of Lipid Lowering Drugs and Diet Network" (GOLDN) (*Mariza de A. et al., 2018*), predecir SNPs involucrados en "Splicing" (*Hui Y. Xiong et al., 2015*) y SNPs asociados con enfermedades mentales mediante su efecto en imágenes cerebrales (*Kristin K. Nicodemus et al., 2010; Xiaoqian Wang et al., 2018*). Existen cerca de 23 algoritmos que emplean aprendizaje de maquina para predecir el efecto de SNPs a nivel de proteína, basado en características estructurales y/o de secuencia (*M. S. Hassan et al., 2019*).

En el sector salud el aprendizaje de maquina ha crecido, permitiendo a través de las variantes genéticas de cada individuo, realizar predicciones sobre el padecimiento que podría sufrir ó predecir las variantes relevantes para la enfermedad (*Daniel S W H et al., 2019*). Como en el caso de la enfermedad coronaria (*Cihan O. et al., 2017*) y cáncer de mama (*Hamid B. et al., 2018*).

Existen cerca de 60 millones de SNPs en nuestro genoma de los cuales se desconoce su asociación con los distintos tipos de demencia. Con la gran cantidad de datos

existentes y qué día a día van en aumento, podemos entrenar una maquina para realizar predicciones en los datos que desconocemos. Si bien existen modelos que nos permiten diagnosticar Alzheimer mediante imágenes de resonancia magnética cerebral (*Lee et al., 2018*), actualmente no existen modelos para predecir SNPs asociados a demencia. El aprendizaje de maquina posibilita clasificar, predecir y explicar conjuntos de datos, pero no permite dar una explicación de causalidad. En el contexto de la demencia es relevante conocer la causalidad, y para ello, entender qué tipo celular es afectado ayudaría a brindar una interpretación de causalidad.

Unidad neurovascular

Antes se conocía que la unidad funcional del cerebro era la neurona (*Edward G. Jones 1999*), pero a lo largo de los años se fue descubriendo el papel relevante que tenían las células de la glía y la barrera hematoencefálica, llevando a postular un modelo como unidad funcional llamado “unidad neurovascular”. Se compone generalmente de epitelio, pericitos, astrocitos, microglía, oligodendrocitos y neuronas (*fig 3*).

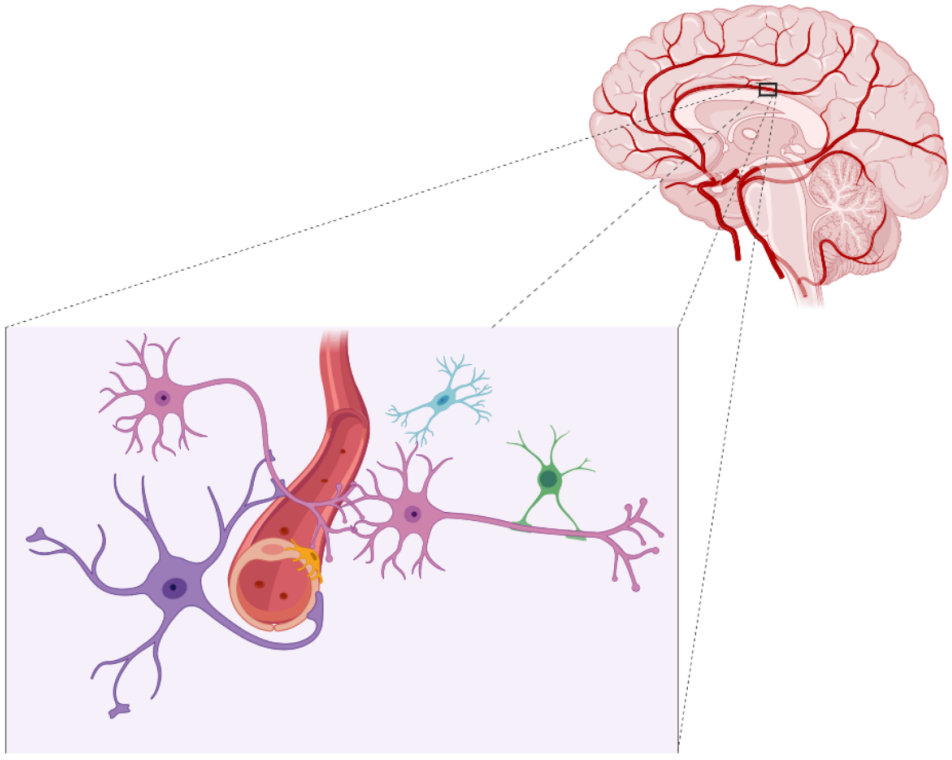


Fig. 3 Esquema de la unidad neurovascular; en morado: astrocitos; amarillo: pericitos; anaranjado: endotelio; rosa: neuronas; azul: microglia; verde: oligodendrocitos.

El cerebro se encuentra en constante comunicación con todo el cuerpo pero parcialmente aislado de metabolitos generados por los demás los sistemas del cuerpo

humano ya que tiene necesidades particulares, esto gracias a la barrera hematoencefálica, compuesta por endotelio vascular que funciona como barrera física, miocitos y pericitos que regulan la dilatación. Ésta barrera regula el flujo de glucosa, monoaminas, y otras moléculas relevantes para el metabolismo del cerebro, también limita la entrada de moléculas dañinas (*Hawkins BT & Davis TP 2005*). Los astrocitos son células gliales que se extienden por muchas regiones del cerebro, tienen contacto directo con la barrera hematoencefálica. Algunas de sus funciones son regular el metabolismo energético de la neuronas así como de neurotransmisores, mantener un pH en el sistema nervioso central y proporcionar soporte físico a la neuronas (*Cai-Yun Liu et al., 2018*). La microglía son células encargadas del sistema inmune del cerebro, que permite la inflamación y la limpieza de agregados proteicos (*Thurgur & Pinteaux, 2019*). Los oligodendrocitos mielinizan y dan soporte a los axónes, por tanto, son relevantes para la comunicación neuronal (*Hamanaka, G et al., 2018*). Las neuronas se interconectan, formando redes grandes y complejas, en las cuales emiten señales, controlando las funciones fisiológicas de todo el cuerpo, se estima hay mas de 90 mil millones neuronas en el cerebro, de tipo excitatorio e inhibitorio (*Herculano-Houzel, 2009*).

En conjunto, cada tipo celular desempeña un papel fundamental para el correcto funcionamiento del sistema nervioso central, y se encuentran vinculados íntimamente en un sistema eficiente de regulación de flujo sanguíneo y metabólico del cerebro (*Armstead & Raghupathi et al., 2011; Abbott & Friedman et al., 2012*). Esta unidad es una estructura vital en la homeostásis del cerebro, cuando uno o más de sus componentes falla, se desencadena un estado patológico (*Bastide et al., 2007; Xing C et al., 2012*). El conocimiento del comportamiento de cada componente de esta unidad, así como sus mecanismos subyacentes, son de relevancia para poder entender padecimientos mentales cómo los tipos de demencia (*Muoio V et al., 2014*).

Análisis de textos de Pudmed

En la tabla 1 se muestra la cantidad de artículos que se descargaron en formato *MEDLINE de la base de datos pubmed*. Se hace una descripción del número de artículos depositados en la base de datos *Pubmed del NCBI* relacionados con los tipos celulares de la unidad neurovascular en abril de 2019, usando como filtro en las búsquedas de “*human*”.

Para contender con información de tipo texto, el aprendizaje de máquina también puede ayudar a hacer una toma de decisiones, si consideramos que una variante genética tiene una localización en el genoma (gen), y este gen tiene información en artículos de *Pubmed.* Si consideramos como clases a los tipos celulares, dada la información del texto se podría hacer una predicción *gen - tipo celular*.

Tabla 1. Número de publicaciones disponibles relacionados a las palabras claves de los tipos celulares.

Palabra clave	Número de artículos
Astrocyte	13,000
Microglia	9,000
Oligodendrocyte	7,000
Perycite	3,000
Brain Endothelial	10,000
Excitatory neuron	6,000
Inhibitory neuron	900

Para usar información tipo texto, se utiliza una transformación de texto a vector, la cual permite desglosar las palabras frecuentes y asignar un valor numérico a cada palabra para ponderar la relevancia de una palabra en un texto dada una colección de textos, de este modo cada texto puede ser expresado en atributos de tipo numérico. Esta aproximación ha sido ampliamente usada para el análisis de textos en redes sociales, para identificar tendencias y preferencias de los usuarios (*Eriko Otsuka et al., 2016*). Junto con aprendizaje de maquina también este enfoque se ha usado en la predicción de enfermedades usando los textos de los expedientes clínicos (*Brown & Kachura, 2019*). En el hospital infantil Rady ubicado en California, EUA, se han ocupado estrategias de aprendizaje de maquina y análisis de texto para la identificación de fenotipos raros y difíciles de diagnosticar, a partir de la información en texto de los expediente clínicos e información del genoma (*Clark et al., 2019*). En la industria farmacéutica se utiliza para la clasificación de potenciales pacientes para hacer pruebas farmacológicas y la venta de fármacos (*Pattisapu N., et al 2019*). Actualmente no existe un modelo que permita asociar un gen a un tipo celular mediante información de texto y que tenga una conexión con la demencia, esto debido a cantidad limitante de datos ómicos a nivel celular asociados a demencia en humanos de la unidad neurovascular. Al relacionar gen y tipo celular mediante los artículos publicados se generó una nueva alternativa de obtención de información relevante para el diseño de experimentos.

Demencia, tipos celulares y SNPs

La base de datos *GWAScatalog* (*Buniello A. et al., 2019*) proporciona los SNPs asociados a una enfermedad o fenotipo. Cada SNP tiene atributos como el valor de P, gen, localización, cromosoma, arquitectura genómica, ubicación en el cromosoma y tipo de variación genética, entre otras. Cada SNP tiene al menos dos frecuencias alélicas, una del alelo ancestral y una “alélica menor”, ambas son distintas a nivel

poblacional. Si se parte de estas características, y se toma en cuenta que existen aproximadamente 592 SNPs asociados a los distintos tipos de demencia con un valor de P menor a $5x10^{-8}$, se puede apreciar claramente la cantidad de SNPs únicos para cada tipo de demencia, lo cual pone de manifiesto que podría existir un patrón, es decir, un hiperplano entre los distintos tipos de demencia (ver *fig. 4)*.

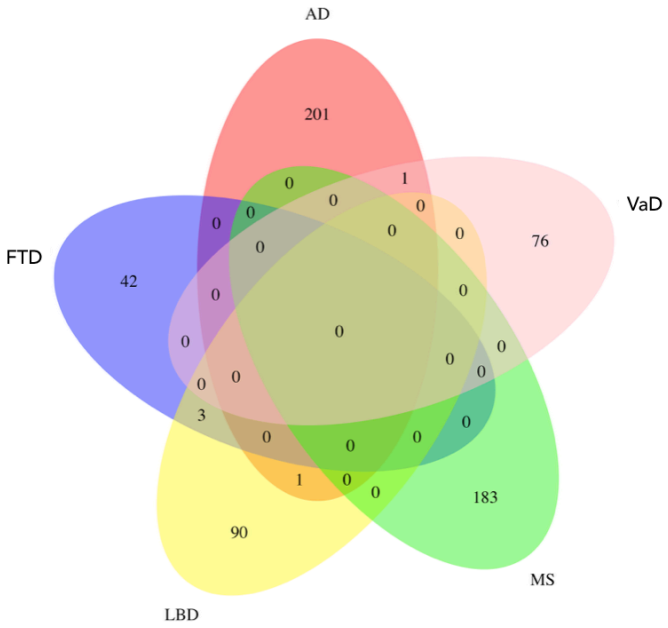


Fig. 4 Diagrama de Venn, cantidad de SNPs que hay entre los diferentes tipos de demencia. AD: Alzheimer's Disease; VaD: Vascular Demencia; MS: Múltiple Sclerosis; LBD: Lewy Body Disease; FTD: Frontotemporal Dementia.

En cuanto a los tipos celulares, si se consideran los resúmenes o "abstracts" de artículos publicados en una búsqueda determinada, y se contemplan las palabras con una frecuencia de al menos 300 apariciones, se sugeriría observar un patrón que resalta cada tipo de célula así como las palabras con mayor frecuencia de aparición (ver *fig. 5)*. Mas aún, si se representan los textos a vector se genera un modelo capaz de clasificar otros textos basado en los tipos celulares. Con base en todo lo anterior se postula la siguiente hipótesis.

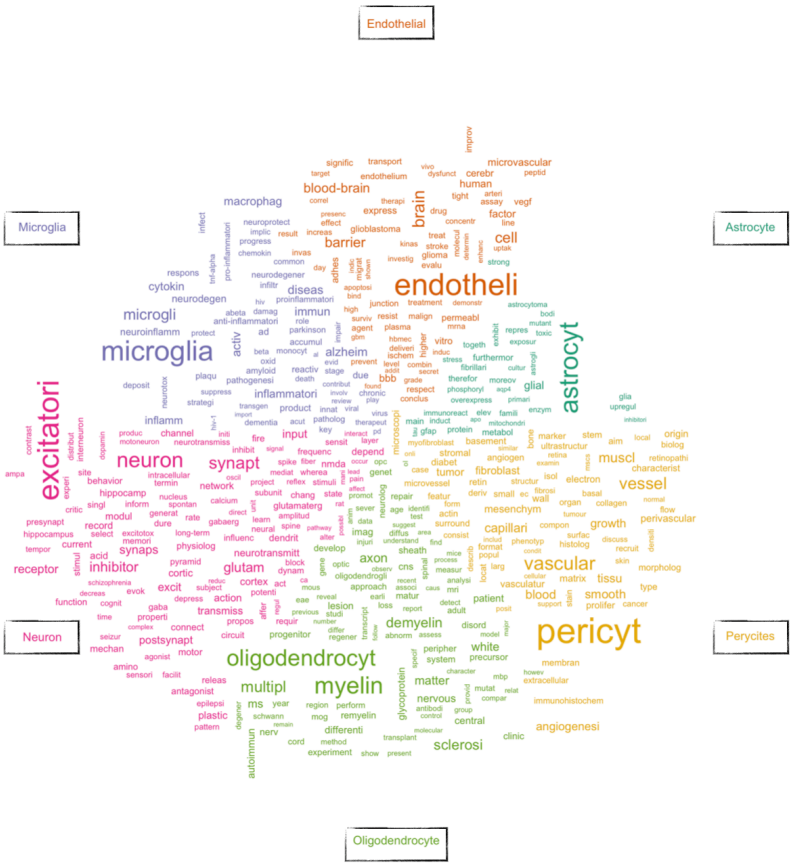


Fig. 5 Nube de palabras con mayor frecuencia en los textos de cada tipo celular. La frecuencia mínima que se tomo fue de 300.

Hipótesis.

Los mecanismos moleculares de las vías metabólicas asociados a demencia, y su relación con los SNPs, desde la perspectiva de un sistema celular, permitirán establecer un modelo biológico para evaluar los biomarcadores asociados a esta enfermedad.

Objetivo general:
Identificar marcadores biológicos (SNPs), involucrados en las vías del metabolismo asociados a demencia, en un sistema biológico neuronal y circulante propuesto para su estudio.

Objetivos particulares:

1. Generar al menos un modelo de aprendizaje de máquina supervisado para asociar SNPs con los diferentes tipos de demencia.
2. Definir los procesos biológicos relacionados con demencia, e interpretarlos en procesos simples asistidos a un lenguaje de programación, o instrucciones en forma de diagramas de flujo.
3. Realizar la selección, pre-procesamiento, transformación, y minería de datos en bases de datos especializadas.
4. Realizar la interpretación y evaluación a partir de criterios de selección de matrices, análisis de textos y predicciones.
5. Proponer un modelo biológico para el estudio de los biomarcadores identificados (SNPs).

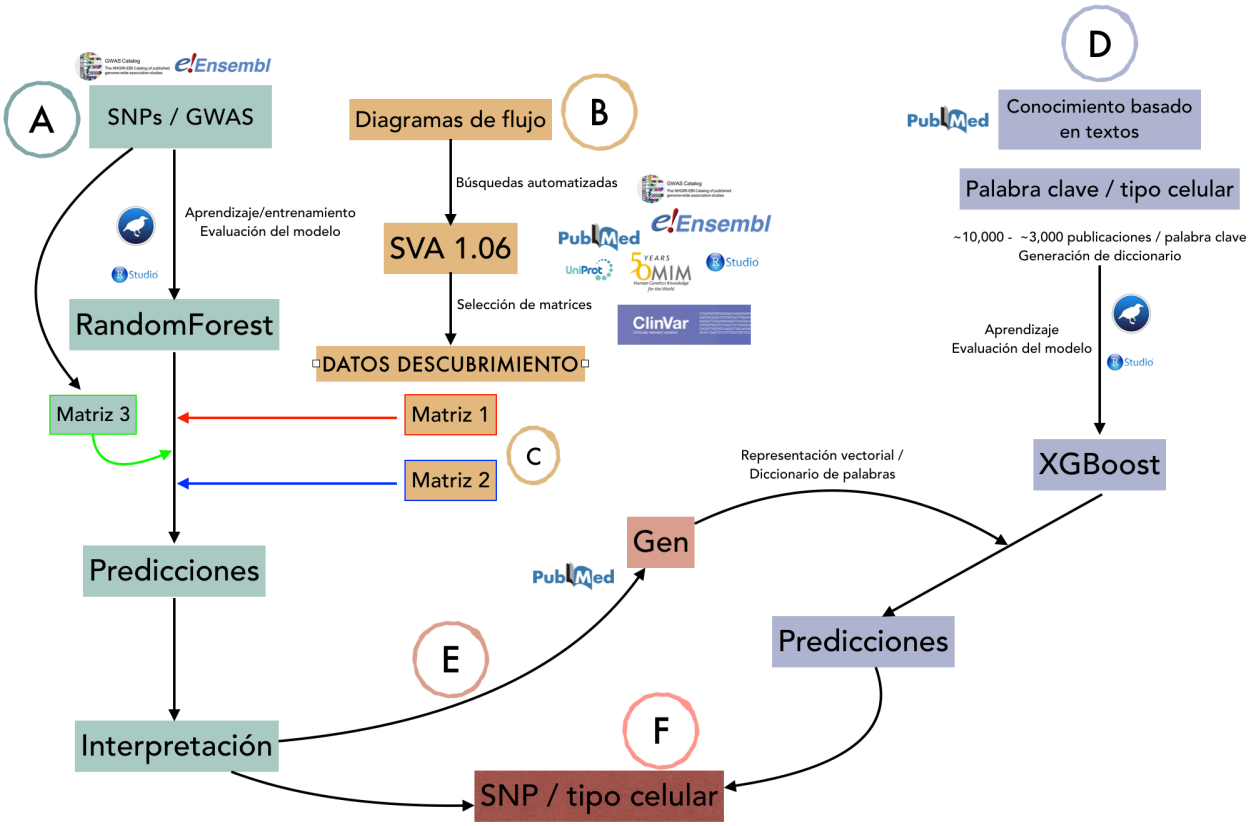


Fig. 6 Esquema de la metodología. A: entrenamiento con datos de SNPs de estudios de asociación de genoma completo con un valor P menor o igual a $5x10^{-8}$. Aquellos SNPs con un valor mayor se agruparon en una matriz 3. El entrenamiento y procesamiento de los datos se llevo a cabo con WEKA y R, el modelo generado con RandomForest sirve para predecir la asociación entre un SNP y un tipo de demencia. B y C: se muestra la obtención de los datos de descubrimiento usando un motor de búsqueda SVA0.11, se obtuvieron dos matrices con datos de SNPs asociados a las palabras clave de cada tipo de demencia. D: la extracción de artículos en formato MEDLINE para cada tipo celular se realizó en pubmed. El procesamiento se realizó en WEKA y R, para generar un modelo de aprendizaje de máquina capaz de predecir mediante textos la relación entre genes y tipos celulares. E: cada SNP se localiza en una región genética, de cada gen se obtuvieron los textos asociados en formato MEDLINE y se usaron como conjunto de descubrimiento para el modelo de XGBoost. F: con ambos modelos (RandomForest y XGBoost) se predice la relación entre SNP y tipo celular.

Metodología

Definición del problema.

Se construyeron representaciones gráficas, es decir, diagramas de flujo, de posibles alternativas de desarrollo para cada tipo de demencia a partir de la identificación e interpretación de procesos biológicos afectados a nivel celular e interpretados a nivel sistémico ya publicados.

Aprendizaje supervisado para SNPs asociados a demencia y textos asociados a tipos celulares.

Datos de SNPs asociados a demencia para entrenar

Los datos se obtuvieron de los estudios de asociación de genoma completo en la base de datos de GWAS Catalog. Se tomaron atributos de esa información y además se agregaron atributos de la base de datos de Ensembl, en la *tabla suplementaria 1* se muestran el tipo y característica de cada atributo. El nombre de cada población se representó con el código poblacional del proyecto de los 1000 genomas. En ella se colectaron las dos frecuencias alélicas (ancestral y menor *"MAF"*) de 26 poblaciones del proyecto de los *1000 genomas*. Se consideró a cada polimorfismo de un solo nucleótido un vector, y a cada elemento *n* del vector una característica del SNP, como se muestra en la *Ec.1*. Por ejemplo, su localización en el genoma, sus frecuencias alélicas en múltiples poblaciones y su rol biológico, y en "*n*+1" su clase, es decir, el tipo de demencia. De esta manera pudimos construir una matriz con todos los SNPs que se han asociado de los distintos tipos de demencia. Esto se esquematiza en la *fig. 6A*.

Ec.1 $SNP = (n_1, n_2, \ldots, n+1)$

Donde: *SNP es Single Nucleotide Polymorphism ;*

n representa a los atributos;

n+1 representa la clase asignada

Datos de texto a vector asociados a los tipos celulares para entrenar.

Para realizar el entrenamiento se recabaron los artículos en formato *MEDLINE* de los tipos celulares contenidos hasta el 15 de abril de 2019 en la base de datos *Pubmed*. Con los siguientes términos:

1) Astrocyte (Astrocito)
2) Microglia (Microglia)
3) Oligodendrocyte (Oligodendrocito)
4) Pericyte (Pericito)
5) Brain Endotelial (Endotelio)
6) Neuron (Neurona)

Para astrocitos se colectaron 13400 artículos; microglia 9817 artículos; oligodendrocitos 7200 artículos; pericitos 3268 artículos; endotelio 10204 artículos y neuronas (inhibitorias y excitadoras) 7269 artículos. De cada formato *MEDLINE* se seleccionaron los *resúmenes* y los *MH* (Medical Subject Header) en formato *csv* (comma separate value). A cada elemento se añadió su clase, es decir, el tipo celular (búsqueda) del que provenían. Se integró en un solo archivo todos los textos, en formato arff (Attribute-Relation File Format). En WEKA 3.9.3 se uso el filtro StringToWordVector para representar los abstracts como vectores y generar atributos con listas de frecuencias de cada termino (Term Frequency, TF; ver *Ec. 2*) y el inverso de cada articulo que contenga el termino (Inverse Document Frequency, IDF; ver *Ec. 3*) (*T. Joachims 2001*). Para quitar signos de puntuación y palabras vacías como “a”, “an”, “the”, “in”, etc., se utilizaron las siguientes opciones del filtro:

“*weka.filters.unsupervised.attribute.StringToWordVector -R last -W 300 -prune-rate -1.0 -C -T -I -N 1 -L -stemmer "weka.core.stemmers.SnowballStemmer -S english" -stopwords-handler weka.core.stopwords.Rainbow -M 30 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\ \t.,;:\\\'\\\”()?!\”*”

Con esta transformación de datos se genero una matriz con todos los textos convertidos a matriz ver *Ec. 4*. Este proceso se esquematiza en la *fig. 6D*.

Ec. 2

$$TF = \frac{freq(D,p)}{\sum freq(D,p)}$$

Donde TF es la frecuencia de los términos (palabras)
D representa los documentos (abstracts)
p representa las palabras de cada documento

Ec. 3

$$IDF = log\frac{D}{D(p)}$$

Donde IDF es la frecuencia inversa de cada documento
D representa los documentos (abstracts)
p representa las palabras de cada documento

Ec. 4

$$\begin{bmatrix} & P_i & . & . & P_r & Class \\ Doc_i & x_i & . & . & x_r & Cell_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Doc_n & x_i & & & xr & Cell_2 \end{bmatrix}$$

Matriz, donde Doc va desde i hasta n, son el numero de instancias; P son las palabras con mayor frecuencia, y se tomaron como atributos, van desde i hasta r; x son los factores TF-IDF por cada palabra; Class es el tipo celular de la unidad neurovascular.

Pre-procesamiento de datos provenientes de texto asociados a tipo celular

Los datos numéricos se normalizaron según la distribución de cada atributo. Se usó el filtro removeUsless (*Eibe Frank et al., 2016*) para remover aquellos atributos que tuviesen valores constantes y aquellos que excedan la varianza máxima si son nominales. Se eliminaron todos los datos nulos o NA.

Selección de características

Se decidió evaluar el conjunto de variables con métodos de selección y extracción de características. Se usó ganancia de la información (divergencia Kullback-Leibler) y Relief (alivio de inferencia estadística). (PShaltout, 2014, Urbanowicz, 2017).

Entrenamiento para SNPs asociados a demencia

En el caso de los SNPs, dadas las características de nuestros datos (ver *tabla suplementaria 1)*, se optó por el uso de arboles de decisión, como J48 (*Sharma et al., 2013*), LMT (*Sumner M et al., 2005*) y bosques aleatorios (*Breiman L., 2001, ver fig.6A*) para trabajar con atributos continuos y discretos.

Entrenamiento para textos de genes asociados de tipos celulares

Para el caso de textos, se usaron los arboles de decisión y bosques aleatorios, los cuales permiten hacer un modelo a partir de un conjunto de datos mediante la construcción de diagramas con instrucciones lógicas. De este modo se categorizan condiciones para tomar una decisión. Además se utilizó un algoritmo de optimización de gradiente distribuido llamado XGBoost (*Tianqi Chen et al., 2019; ver fig. 6D*) que consiste en la creación de múltiples bosques aleatorios de manera paralela tomando en cuenta una penalización por cada error. El entrenamiento supervisado se realizó en el software WEKA 3.9.3 (*Eibe Frank et al., 2016*) y en R con la paquetería *caret* (*Max Kuhn et al., 2016*).

Evaluación de los modelos de aprendizaje de máquina.

Para medir el desempeño de los algoritmos utilizados se usaron matrices de confusión. Para evaluar el modelo se consideraron parámetros como el porcentaje de instancias correctamente clasificadas (*Kuncheva et al. 2004*), su especificidad (ver *Ec. 5*) y sensibilidad (ver E*c. 6*) en la predicción, y F-score (ver *Ec. 7*), en la que precisión se define como en la *Ec. 8*. También se realizó una validación cruzada de *k* =10.

$$Ec.\ 5 \qquad Especificidad = \frac{VN}{VN + FP}$$

$$Ec.\ 6 \qquad Sensibilidad = \frac{VP}{VP + FN}$$

$$Ec.\ 7 \qquad F - score = 2\frac{(Precision)(Sensibilidad)}{Precision + Sensibilidad}$$

$$Ec.\ 8 \qquad Precision = \frac{VP}{VP + FP}$$

Donde VP es verdaderos positivo; VN verdadero negativo; FP falso positivo; FN falso negativo

Selección, pre-procesamiento, transformación y minería de datos en base de datos especializadas para la generación de datos de descubrimiento de SNPs.

La demencia se clasificó en los tipos de demencia: i) Alzheimer (**AD**), ii) demencia frontotemporal (**FTD**), iii) esclerosis múltiple (**MS**), iv) demencia por cuerpos de Lewy (**LBD**), y v) demencia vascular (**VaD**) (*CIE-11, 2018*). Cada tipo de demencia se representó con un diagrama de flujo con sus elementos alfanuméricos contenidos en el Medical Subject Heading (*MeSH*) de NCBI (*National Center for Biotechnology Information*). Brevemente, el valor de cada elemento alfanúmerico se confirmó con al menos 3 artículos publicados en *Pubmed* como criterio de selección, el SNP esté referido en al menos un estudio GWAS, y que se encuentre vinculado a un factor de riesgo en la base de datos de *ClinVar*. Mediante una plataforma de búsqueda de SNPs automatizada (*SVA v0.11, Loredo M et al., 2018*) se introdujeron los elementos que conforman cada tipo de demencia (ver *fig. Suplementaria 1 A-E*). Las bases de datos de extracción de información relacionadas son: *PubMed*, *GWAScatalog* del EMBL-EBI, *dbSNP*, *ClinVar*, *OMIM*, *Uniprot*, *Swissprot*, *Reactome* y *Ensembl* (*Park Y M et al., 2017; MacArthuer J et al., 2017; Sherry ST et al., 2001; Janet Piñero et al., 2016; Landrum MJ et al., 2018; OMIM® 2018; The UniProt Consortium 2017; Kanehisa et al., 2017*). Todo esto con el fin de generar los conjuntos de descubrimiento para retar nuestro modelo generado con aprendizaje de maquina para asociar SNPs con tipos de demencia. En la fig. 6B se muestra este procedimiento.

Selección y análisis de matrices de descubrimiento.

Los datos obtenidos (ver *fig. 6B*) se agruparon en diferentes matrices para valorizar la información y organizarla. Los criterios de selección se restringieron a la asociación de los elementos que conforman cada matriz con estudios genómicos poblacionales, asociación con fenotipo relacionado con la enfermedad con datos de significancia clínica y al menos 3 estudios publicados. A éste conjunto de datos los agrupamos en la “matriz 1”; a todos aquellos resultados que no cumplieran con alguno de los criterios anteriores se agrupo en una “matriz 2”. Finalmente, se clasificaron los elementos que no tenían un valor de P “p-value” significativo en GWAS asociado algún tipo de demencia en una “matriz 3” (ver *fig. 6C*).

Predicciones de SNPs asociados a demencia

Todas las matrices (1, 2 y 3) se sometieron a la predicción del modelo de clasificación probabilístico para establecer el tipo de demencia al qué pertenecen o podrían estar asociados. Cada predicción arrojará un valor de probabilidad de

pertenecer a la clase basado en la distancia con el centroíde de cada clase en el espacio multidimensional. Como control de predicción negativo se usaron SNPs de estudios de asociación de genoma completo de osteosarcoma y como control positivo se utilizaron SNPs de la base de datos GRASP (*Genome-Wide Repository of Associations Between SNPs and Phenotypes*) distintos a los utilizados para entrenar y con un valor de P menor a $5x10^{-8}$ asociados a Alzheimer y Esclerosis Múltiple. El valor de corte para seleccionar los SNPs con altas probabilidades de pertenecer a una clase fueron fue de 0.5665. Se utilizó este punto de corte dada la distribución de la probabilidad de las predicciones en los controles, ya que se obtenía un 100% de verdaderos positivos y 100% de verdaderos negativos para la enfermedad de Alzheimer (AD; *ver tabla 2 y fig. 8-9*).

Predicciones de genes asociados a tipos celulares.

En el caso de texto se uso también un corte de 0.5. Se utilizaron esos puntos de corte dada la distribución de la probabilidad de las predicciones. De cada gen se recabaron todos los artículos existentes hasta el 15 de abril de 2019, y se les hizo su respectivo tratamiento como en el entrenamiento. En la *fig. 6E* se muestra la parte de este procedimiento.

Lenguaje de programación y software.

El ambiente de software para los análisis estadísticos, cómputo y gráficos se realizó en lenguaje R compilado para plataformas Linux (Fedora Server 29) hospedado en un servidor Thinkserver Lenovo con procesador de Intel Xeon 2.2 Ghz/2400MB con 16 GB, con 128 Gb de RAM y 24 TB de almacenamiento y 32 núcleos con ambiente y MacOS (Mac OS-Sierra v.10.13.4). La generación del modelo analítico con aprendizaje de maquina se realizó con el software WEKA 3.9.3. (*Eibe Frank et al., 2016*) y R.

Representación de grafos

Usando la paquetería "*igraph*" de R-3.5.1 (*Csardi G et al., 2006*), se generaron grafos dirigidos de asociación entre los distintos tipos de demencia, con los SNPs, genes y tipo celular. Se construyeron a partir de una matriz que contenía la lista de vértices (uniones) y los atributos de cada nodo, tomamos como nodo a los tipos de demencia, SNPs, genes y tipo celular; los vértices fueron la asociación, usando nuestro modelo de aprendizaje de maquina, entre SNPs y tipos de demencia; entre SNPs y genes el valor de asociación fue mediante su posición en el genoma; el tipo de célula y el gene fue el valor de predicción del modelo para textos. La edición de cada grafo se hizo en Gephi 0.9.2, se usó una distribución Yifan Hu. En la *fig. 6F* se muestra la parte de este procedimiento.

Función de Bayes

Se usó la función de probabilidad condicional de Bayes para estimar la probabilidad de desarrollar demencia dado un conjunto de SNPs. Donde MAF es la frecuencia alélica menor (Minor Allele Frequency), “Predicción” es el valor de probabilidad de la predicción y p-value, el valor de p de asociación de GWAS (como valor de asociación de 1 - pvalue, tomando en cuenta que el valor p es la probabilidad de no estar asociado). Se tomó en cuenta un factor de 0.5 aleatorio al considerar la contribución de factores externos o ambientales que podrían afectar la probabilidad de desarrollar la enfermedad.

$$P(SNP \mid Demencia) = \frac{P(Demencia \mid SNP)P(SNP)}{\sum P(Demencia \mid SNP)P(SNP)}$$

$$\therefore P\left(SNP \mid Demencia\right) = \frac{\left(1 - pvalue\right)(MAF)}{\sum \left(1 - pvalue\right)(MAF)}(0.5)$$

$$P(Demencia \mid SNP) = 1 - pvalue$$

$$P(SNP) = MAF$$

$$P(Demencia \mid SNP) = Prediccion$$

$$\therefore P\left(SNP \mid Demencia\right) = \frac{\left(Prediccion\right)(MAF)}{\sum \left(Prediccion\right)(MAF)}(0.5)$$

Plataforma web

Con los datos recabados de las predicciones y la función de probabilidad para estimar el riesgo de desarrollar demencia, se construyó “SNP-Cell Classification and Risk Association of Dementia (SNP-Cell CRAD v0.2)” en Shiny R (*Winston Chang et al., 2019*).

Resultados

Modelos para predecir SNPs asociados a demencia

En la *tabla suplementaria 2* se muestran los resultados de la especificidad del modelo para predecir SNPs asociados a demencia usando 69 atributos para el entrenamiento, en los que se tomaron en cuenta las dos frecuencias alélicas, la región del cromosoma, el número de cromosoma, la posición en el genoma, el gen y el tipo de variación genética. En la *tabla suplementaria 3* se muestran la especificidad con 38 atributos para el entrenamiento, con los mismo datos pero solo la frecuencia alélica menor, la región del cromosoma, el número de cromosoma, la posición en el genoma, el gen y el tipo de variación genética. Para ambos conjuntos se utilizaron 602 instancias (*fig. 4*). En ambas tablas se comparan, los valores de sensibilidad entre distintos algoritmos de entrenamiento y selección de características.

En la *tabla suplementaria 4* se reporta la sensibilidad obtenida del modelo entrenado con los 69 atributos mencionados anteriormente; en la *tabla suplementaria 5* se muestra la sensibilidad entrenando con 38 atributos, en ambas tablas se compara el resultado obtenido usando diferentes algoritmos para entrenar y seleccionar características. En la *tabla suplementaria 6* se reporta el F-score para el modelo entrenado con 69 atributos, y en la tabla *suplementaria* 6 con 38 atributos. Como en las tablas anteriores se muestran los resultados para los distintos algoritmos de entrenamiento y selección de características.

El modelo con una mayor sensibilidad, especificidad y F-score fue RandomForest (Bosques aleatorios). En la *fig. 7* se muestran los resultados de sensibilidad, especificidad, F-score y precisión para cada una de las clases. Se utilizaron estas métricas para estimar el error que hay en la predicción de cada una de las clases de nuestro modelo. Las clases LBD, VaD y FTD muestran claramente una perdida en la sensibilidad debido a que las clases se encuentran desbalanceadas respecto a AD y MS. LBD, VaD y FTD sobrerepresentan valores de verdaderos positivos. Para evitar esta sobrerepresentación se utilizó el coeficiente de correlación de Matthews (MCC) como métrica de evolución para las clases desbalanceadas (*ver tabla suplementaria 9*).

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FN})}}$$

Donde TP son verdaderos positivos; TN son verdaderos negativos; FP son falsos positivos; FN son falsos negativos.

Datos de descubrimiento

Se realizaron 5 búsquedas automatizadas en el algoritmo SVA 0.11, una búsqueda por cada tipo de demencia. Para ello se elaboró un diagrama de flujo, en las *fig.*

suplementarias 1 (A-E) se muestran los diagramas de flujo de los tipos de demencia. Los datos se agruparon en matrices (matriz 1 y matriz 2), es importante señalar que los SNPs recabados en esta fase son distintos a los SNPs usados en el entrenamiento. Los SNPs que no tenían un valor P significativo de los GWAS asociados a demencia se agruparon en la matriz 3.

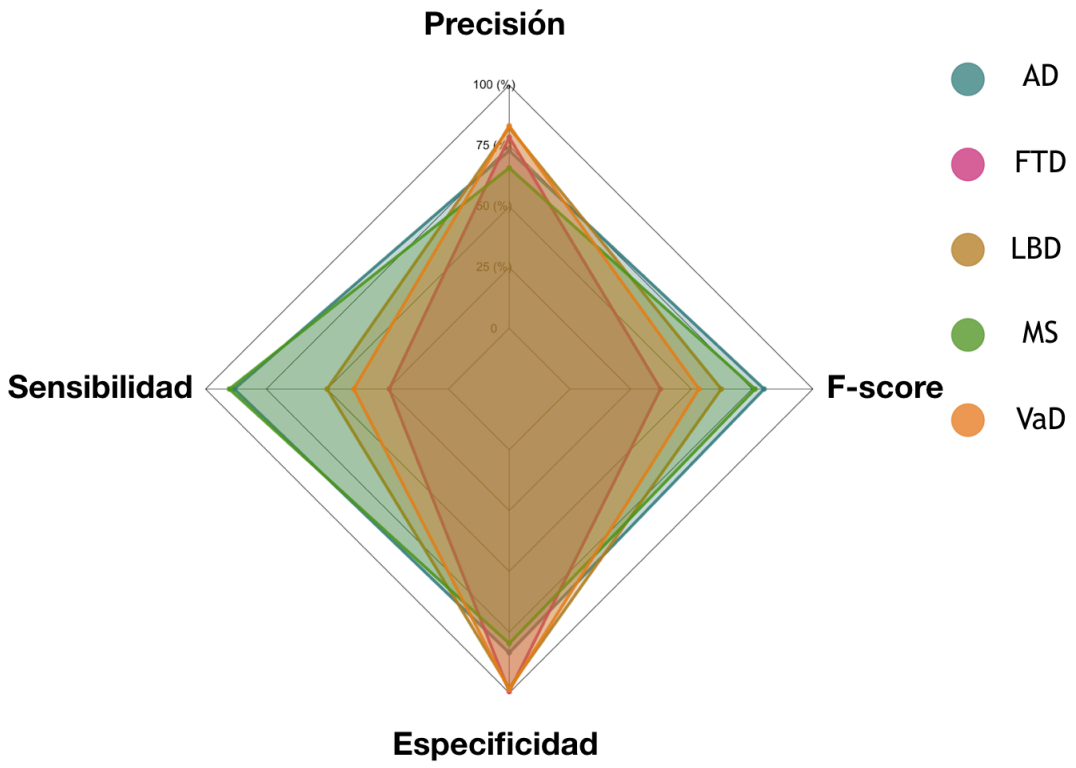


Fig. 7 Radar Chart de los parámetros de validación de nuestro modelo entrando con el algoritmo RandomForest y con 38 atributos. AD, Enfermedad de Alzheimer; FTD, Demencia Frontotemporal; LBD, Enfermedad por cuerpos de Lewy; MS, Esclerosis Múltiple; VaD, Demencia Vascular.

Predicciones de SNPs relacionados a demencia

Como modelo negativo se utilizaron GWAS de osteosarcoma (cáncer de hueso), se partió del supuesto de que estos SNPs no tienen ninguna publicación relacionada con demencia y cómo control positivo se usaron SNPs con un valor P de asociación menor a $5x10^{-8}$ para Alzheimer y Esclerosis Múltiple. En la fig. 8 se muestra la distribución de las predicciones en los controles, así como un claro sesgo de probabilidad. En la *fig. 9* se muestran la predicción de cada SNP de los controles positivos utilizados, de los cuales 11 de MS fueron Falsos Negativos y de AD 0 Falsos Negativos al usar un punto de corte de la media de la distribución de probabilidad (ver fig. 8), por tanto nuestro

modelo es capaz de clasificar correctamente a los SNPs con AD (*ver tabla 2*); los controles negativos tuvieron predicción pero con una probabilidad menor al 0.5665, solo hubo un falso positivo; todas las matrices se sometieron al modelo para generar predicciones de cada uno de los SNPs, se usó un valor de corte de 0.5665, al aplicar este valor de corte, para AD se logró un 100% de sensibilidad y especificidad. En la *figura suplementaria 2A* se observan los SNPs predichos de la matriz 1, en la *figura suplementaria 2B* las predicciones de la matriz 2 y en la *figura suplementaria 2C* las predicciones de la matriz 3, usando este punto de corte se perdió mucha información.

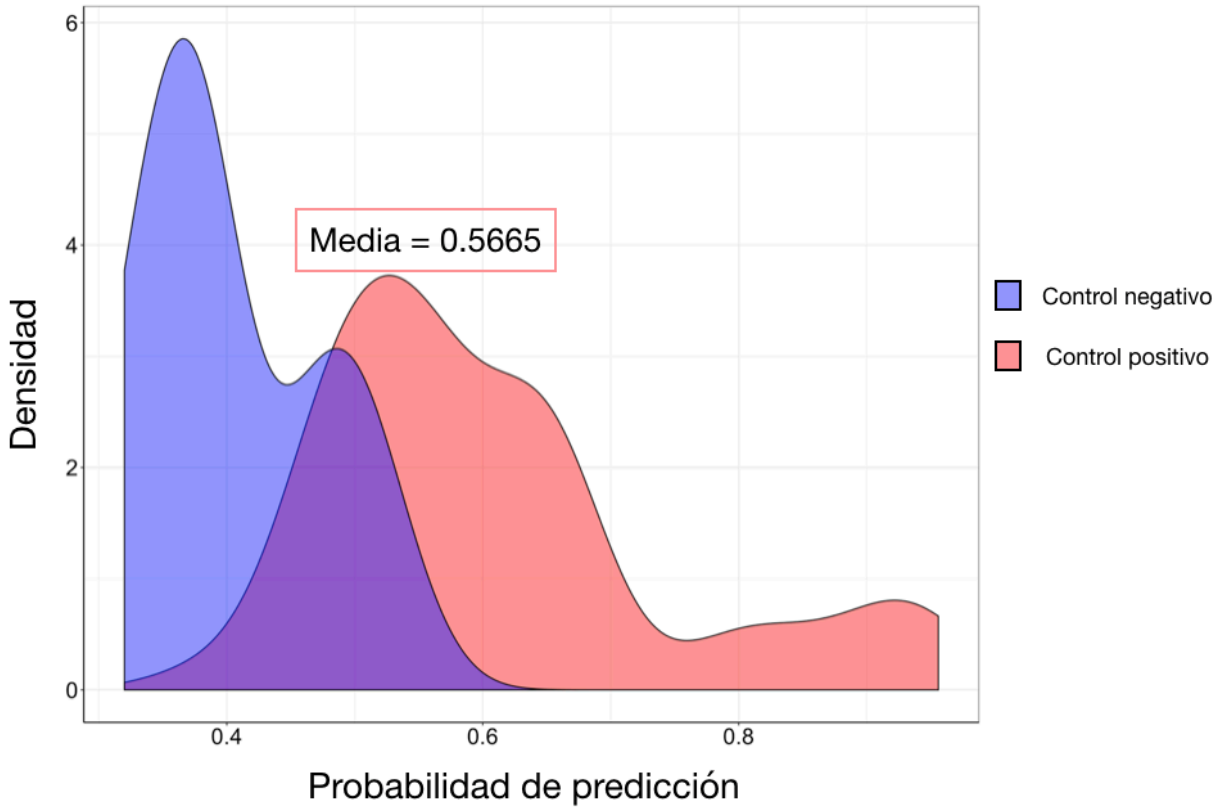


Fig. 8 Distribución de probabilidad de las predicciones. En azul se muestra la densidad del control negativo (osteosarcoma) y en rojo la densidad del control positivo (Alzheimer y Esclerosis Múltiple).

Modelos para predecir textos asociados a tipos celulares

Para generar los atributos de los textos, se generó un diccionario con las palabras con una frecuencia de aparición de 300 veces. En la *fig. 10A* se observan las palabras con mayor frecuencia al partir de los *abstracts* de los textos, mientras que las palabras con mayor frecuencia derivadas de los *MeSH Terms* se aprecian en la fig. 10B. La clara heterogeneidad de las palabras se pone de manifiesto en los *abstracts y* fue una de las razones por la que se utilizaron en el entrenamiento. Se generó un diccionario de palabras y se aplico una transformación *TF-IDF (ver Ec. 2-3)* para representar

vectorialmente los artículos únicos de cada tipo celular (*ver fig. 11*), de este modo cada clase quedó del siguiente modo:

- Astrocyte: 8925 instancias
- Oligodendrocyte: 5860 instancias
- Microglia: 6910 instancias
- Neuron: 6859 instancias
- Endothelial: 8920 instancias
- Perycites: 3268 instancias

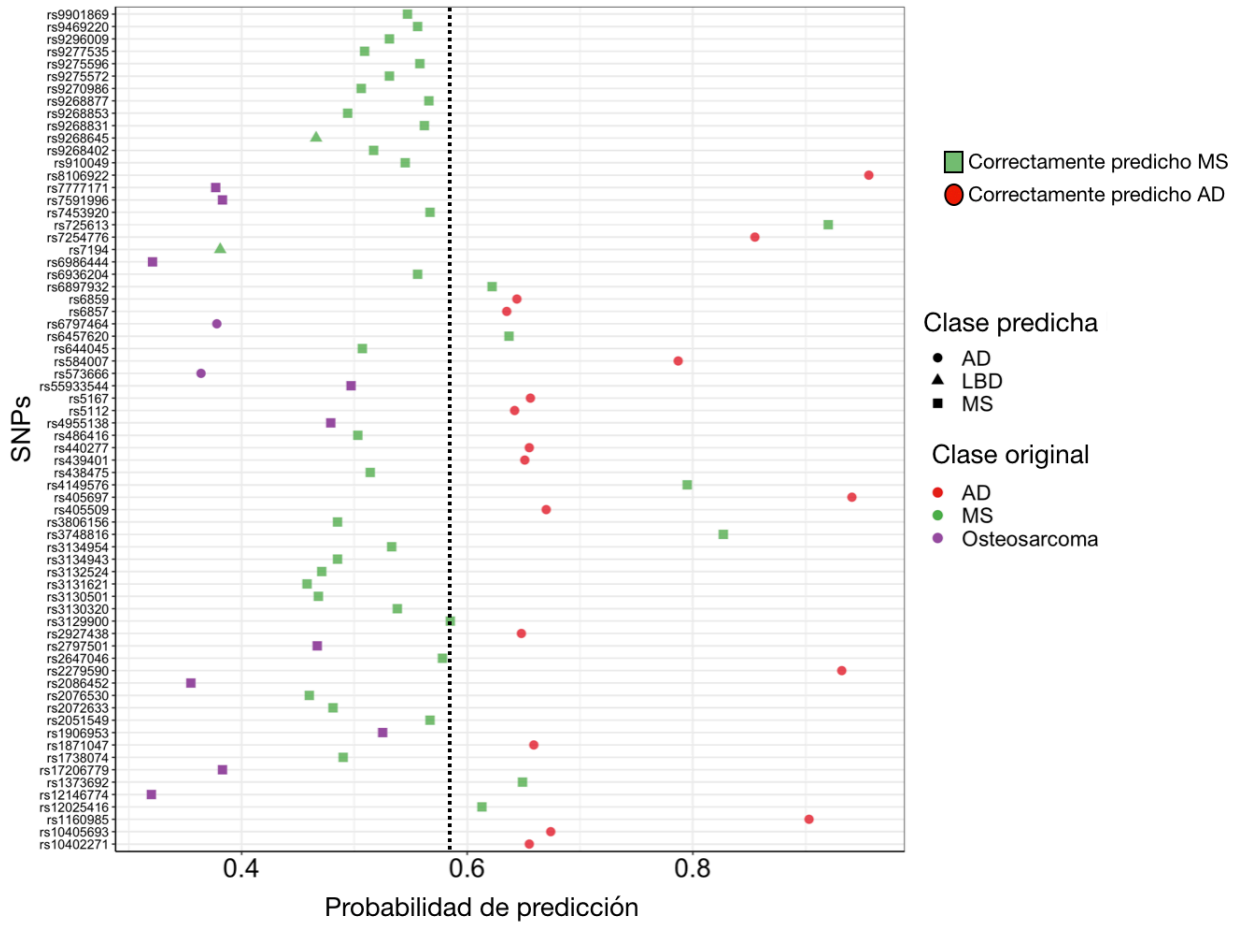


Fig. 9 Probabilidad de Predicción de los SNPs utilizados como control positivo y negativo. En color se muestran las clases originales: rojo, Alzheimer (AD); verde, Esclerosis Múltiple (MS); morado, Osteosarcoma. En figuras se muestra la clase predicha por el modelo generado: circulo, Alzheimer (AD); triangulo, Enfermedad por cuerpos de Lewy (LBD); cuadrado, Esclerosis Múltiple (MS). Cuadro verde y circulo rojo significa por tanto correctamente clasificado

Tabla 2. Matriz de confusión datos de validación, controles positivos y negativos.

A	B	C	D	E	D	Clasificado
39	0	0	0	0	2	**A = AD**
0	0	0	0	0	0	**B = FTD**
0	0	0	0	0	0	**C = LBD**
0	0	2	94	0	10	**D = MS**
0	0	0	0	0	0	**E = VaD**
0	0	0	0	0	0	**D = Osteosarcoma**

Los entrenamientos se llevaron a cabo con XGBoost, y la selección de atributos se llevó a cabo con *infogain y gainratio (ver tabla suplementaria 7),* también se hizo un entrenamiento eliminando las palabras clave de atributo de los nombres de las clases (en las gráficas se encuentra como: *deleted*), como por ejemplo, *astrocyte, endothelial, etc.,* en la *tabla suplementaria 7* se muestran los valores de sensibilidad, especificidad, precisión y F-score. En total se generaron 4 conjuntos de entrenamiento (*all, infogain, gainratio, deleted*), todos los entrenamientos se realizaron con el 85% de los datos y el 25% se uso como prueba. Para validar los modelos se usaron como controles positivos 3328 artículos que tenían intersección entre dos tipos celulares (fig. 11) y cómo controles negativos se usaron 7198 artículos relacionados a *Melanocyte*, *Osteoclast* y *Platelet*. Evaluamos los atributos con mayor relevancia para la clasificación de cada entrenamiento, y observamos que en *infogain* y *gainratio* los atributos con mayor importancia eran distintos *(ver figura suplementaria 3A-C). En la fig. 12C* se muestra que el modelo que tiene menor error al clasificar es el generado con el conjunto de datos de *infogain.*

Fig. 10 Nube de palabras. A) Palabras con mayor frecuencia en los "abstracts". B) Palabras con mayor frecuencia en los "MeSH Terms".

Fig. 11 Diagrama de Venn de los artículos (PMID) para cada tipo celular.

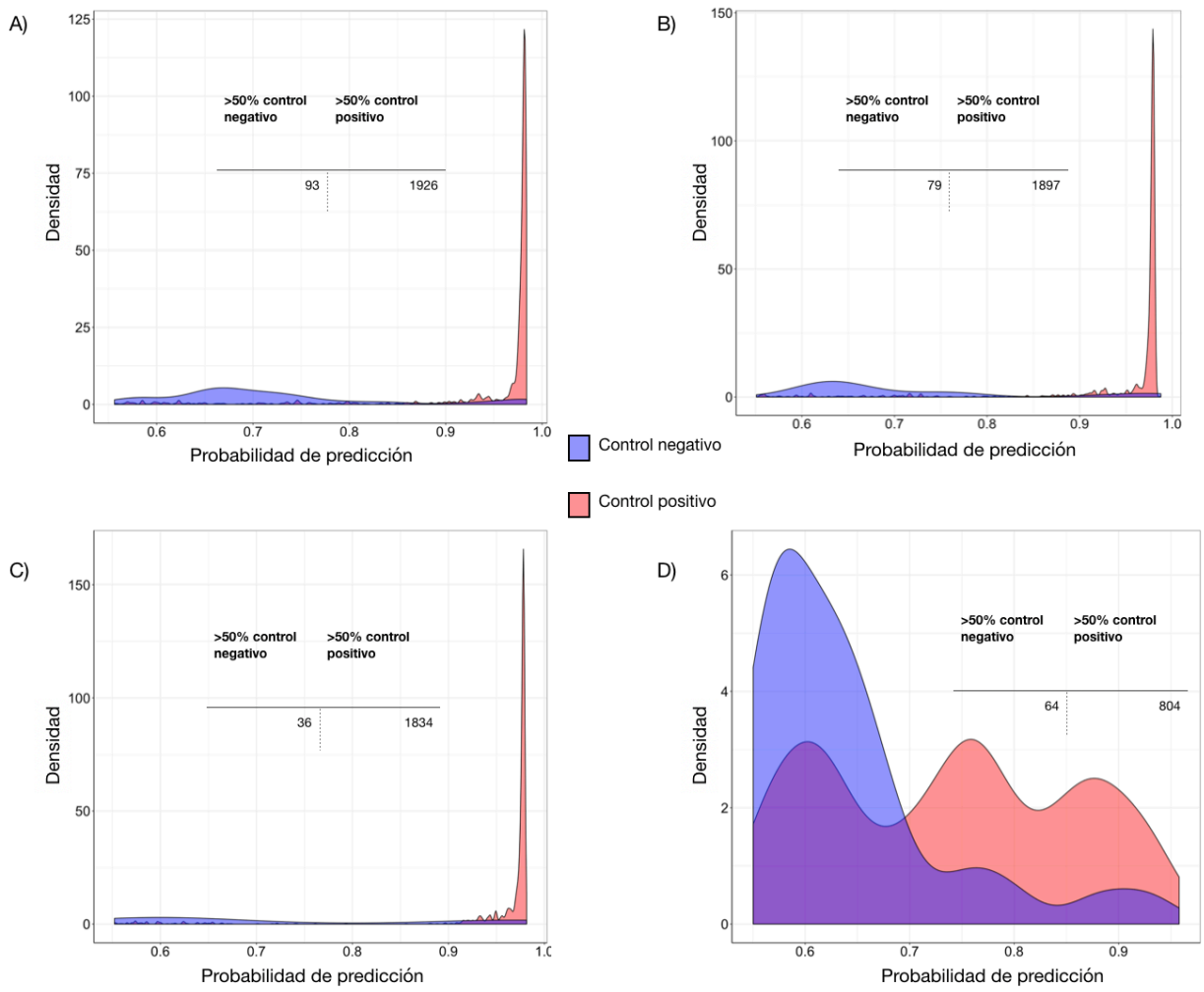


Fig.12.Distribución de probabilidad de predicción de controles positivos y negativos. A) ALL, predicción en conjunto de datos con todos los atributos; B) GainRatio, predicción en conjunto de datos con los atributos seleccionado usando GainRatio; C) InfoGain, predicción en conjunto de datos con los atributos seleccionado usando InfoGain; C) Deleted, predicción en conjunto de datos con ciertos atributos eliminados.

Con el fin de observar la posibilidad de generar el vinculo entre el gen y el tipo celular, se alimento el modelo de predicción de texto con artículos ampliamente reportados para los genes *CD40*, *GABRA1*, *GFAP*, *GLUT1*, *GRM1*, *TJP1*, y como control negativo *Mycobacter* (ver *fig. 13).* Se muestran las predicciones mayores a 0.5 de los artículos (*PMID*) asociados a los genes que se mencionaron anteriormente, como se observa ninguno de los artículos (*abstracts*) que provenía del termino *Mycobacter* fue clasificado con una probabilidad mayor a 0.5 dentro de un tipo celular. En la *figura suplementaria 4* se puede ver la distribución de las predicciones de los artículos.

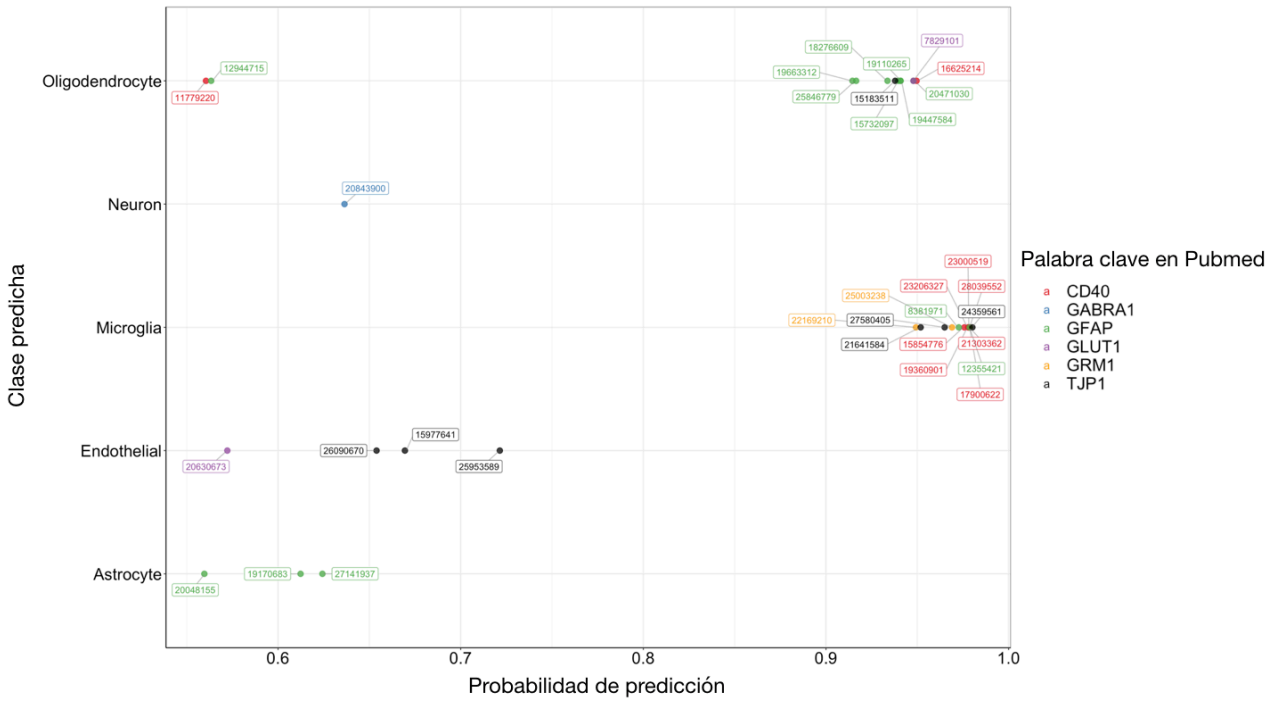


Fig. 13. Predicción/Clasificación de artículos (abstracts) obtenidos de pubmed al buscar cada gen. Se etiquetan con su PubMed ID (PMID).

Tabla 3. Cell-score de los genes CD40, GABRA1, GFAP, GLUT1, GRM1 y TJP1.

Gen	Oligodendrocito	Neurona	Microglia	Endotelio	Astrocito
CD40	0.222	0	0.777	0	0
GABRA1	0	1	0	0	0
GFAP	0.61538	0	0.15385	0	0.2308
GLUT1	0.5	0	0	0.5	0
GRM1	0	0	1	0	0
TJP1	0.1429	0	0.4285	0.4285	0

Con las predicciones/clasificaciones usamos la siguiente fórmula para estimar un valor de asociación entre tipo celular y gen, le llamamos *cell-score*:

$$CellScore = P_{cell}/P_{all}$$

Donde P_{cell} es el número de predicciones al tipo celular en interés, y

P_{all} es el número total de predicciones, ambas refiriendo a un solo gen.

En la *tabla 3* se muestran los *cell-score* de los genes de los cuales se utilizaron sus *abstracts*. Como se esperaba, el marcador GFAP es un marcador de células de la glia (*Xinguang Yang et al., 2019*), TJP1 de la barrera hematoencefálica y GABRA1 un canal muy estudiado en neuronas (*B. T. Hawkins et al., 2006; Shu Yang et al., 2018; Hirose S. 2006*).

Unidad Neurovascular y demencia

Cada SNP tiene una localización en el genoma, y se tiene el reporte de que tipo de variación genética es, y en qué gen se localiza. Tomando en cuenta el gen, podemos asignar una probable asociación con los tipos celulares mediante los artículos que se han publicado a cerca de la función del gen. Para cada gen se descargaron los artículos existentes en formato *MEDLINE* y se hizo la respectiva conversión a vector con los 1034 atributos. Cada gen se sometió al modelo para generar una predicción y un valor de probabilidad de pertenecer a cada clase. Un inconveniente que presenta este enfoque es que no esta diseñado para estudiar los SNPs en zonas no codificantes del genoma, ya que depende exclusivamente de lo que esta publicado en *Pubmed*.

Con grafos dirigidos, se esquematizaron los tipos de demencia, los SNPs asociados, los genes y los tipos celulares de la unidad neurovascular. En cada predicción solo se seleccionaron aquellas con una probabilidad mayor o igual a 0.55, cada vértice toma los valores de predicción de cada modelo que se generó, y el tamaño de los nodos es proporcional al número de vértices de entrada; en las *figs. 14-18* se muestra la red de asociación de los distintos tipos de demencia, el tipo de demencia se encuentra en el centro, de ella salen vértices con los valores de asociación de SNPs, y de los SNPs salen vértices hacia el gen en donde se localizan, y de estos últimos hacia el tipo celular con valores de *cell-escore*. Para los SNPs que se usaron como entrenamiento, se uso la resta de 1 menos su valor de P del estudio GWAS de donde provenía para representar la asociación entre SNP y tipo de demencia.

Desarrollo de un marco de trabajo de integración “SNP-Cell CRAD”

Con la finalidad de integrar la información generada en este trabajo se construyó un marco de trabajo (ver fig. 6 y fig. 20). Brevemente, esta aplicación permite buscar los SNPs, elegir la población con cierto grado de tamizaje por ejemplo, continental y el tipo de demencia; También el marco de trabajo despliega información relativa a la predicción de SNPs, es decir se calcula el valor de probabilidad para cada predicción (“Dementia Score”); también se predice el valor de probabilidad de asociación de texto respecto a los tipos celulares (“Cell-Score”). Esta herramienta permite saber cuál es la probabilidad de desarrollar demencia dada una combinación de SNPs en una población en específico, este cálculo se hace automáticamente usando una función de probabilidad condicional de Bayes. La plataforma permite descargar la tabla de los resultados en formato tsv o csv. También despliega dos gráficos, en uno muestra la variación genética a través de la cantidad de SNPs que hay en los diferentes genes presente en los tipos celulares; mientras que el otro gráfico muestra la probabilidad de desarrollar demencia en distintas poblaciones. Ambos gráficos cambian dinámicamente según los SNPs que el usuario seleccione en la interfaz de la app (*Fig. 19)*.

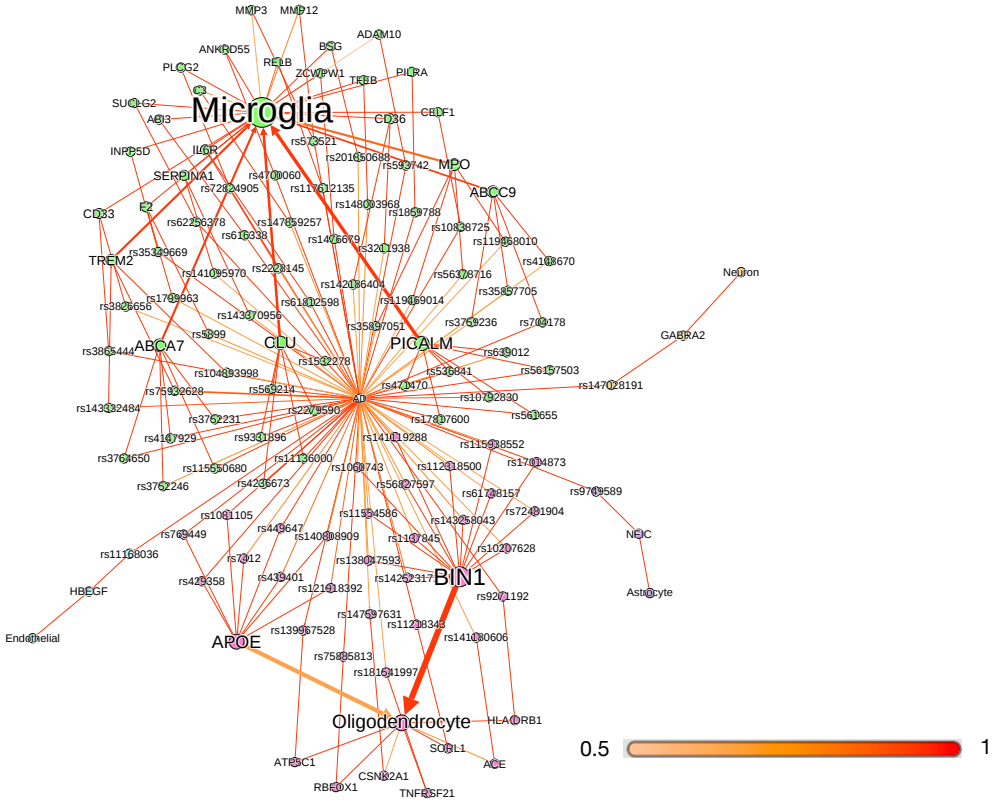


Fig. 14 Red de asociación de la enfermedad de Alzheimer (AD), el tamaño de los nodos es proporcional al grado de entrada, y el color de los vértices es el valor de asociación entre cada nodo.

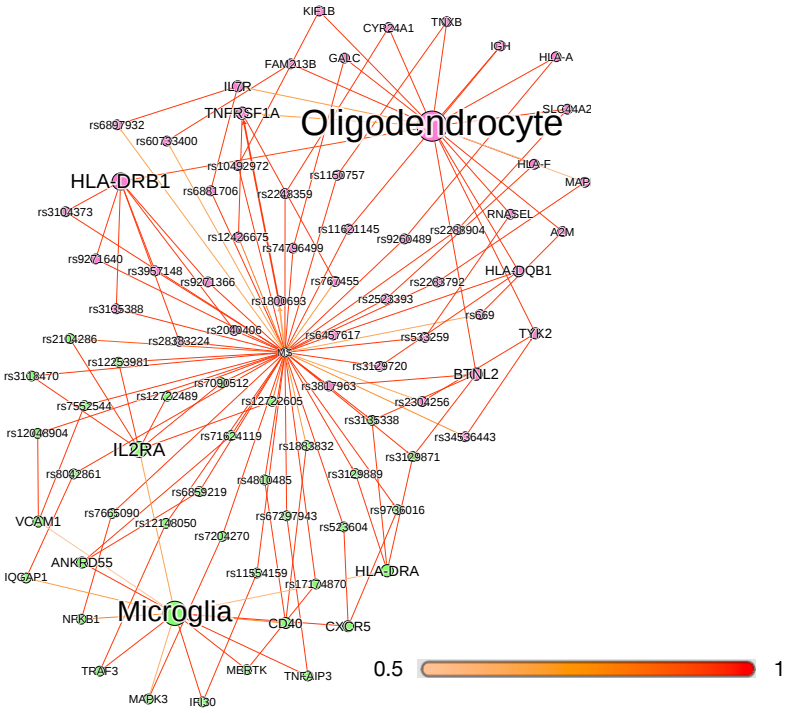


Fig. 15 Red de asociación de Esclerosis Múltiple (MS), el tamaño de los nodos es proporcional al grado de entrada, y el color de los vértices es el valor de asociación entre cada nodo.

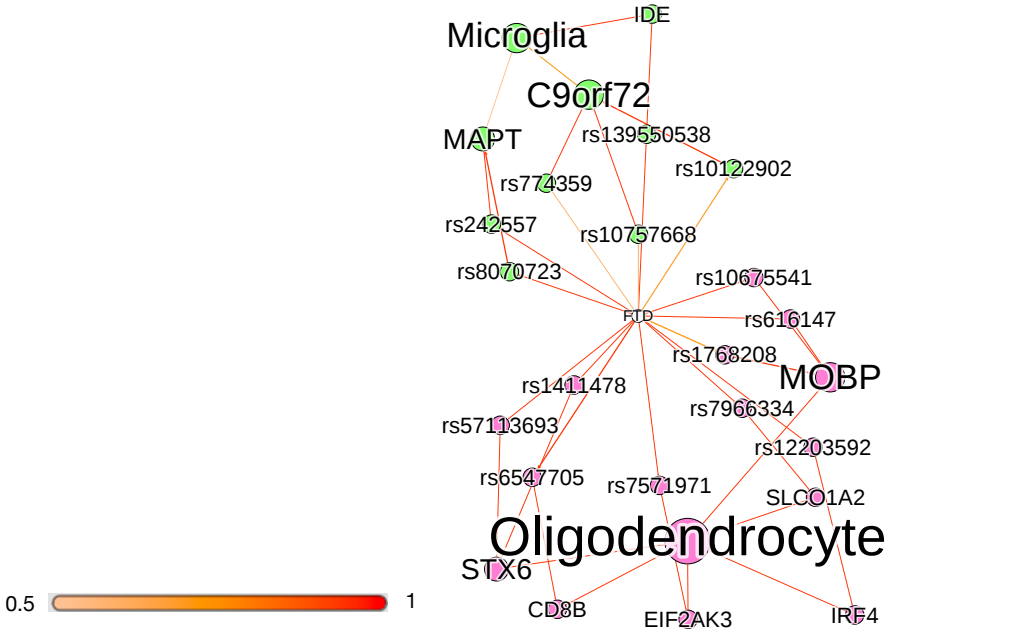


Fig. 16 Red de asociación de Demencia Frontotemporal (FTD), el tamaño de los nodos es proporcional al grado de entrada, y el color de los vértices es el valor de asociación entre cada nodo.

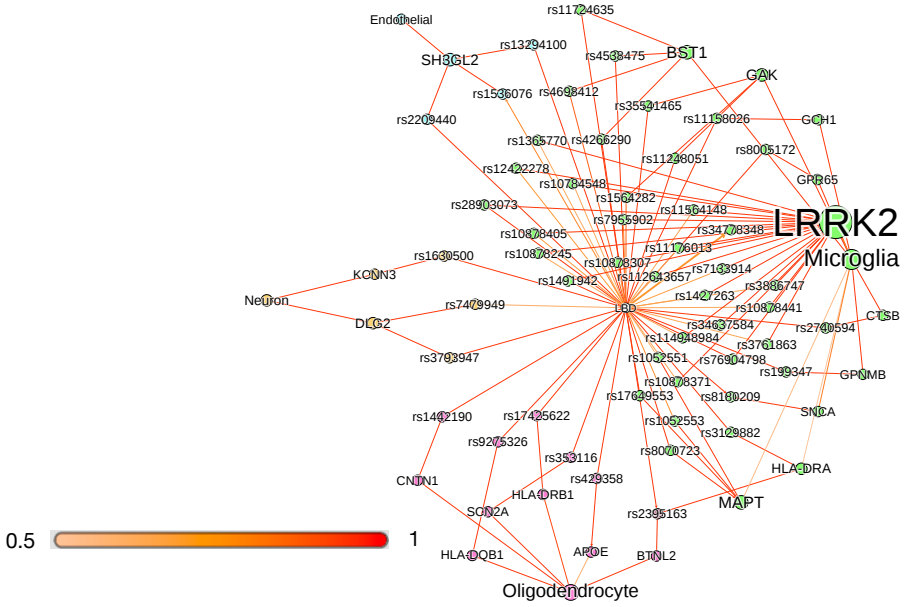


Fig. 17 Red de asociación de la enfermedad por cuerpos de Lewy (LBD), el tamaño de los nodos es proporcional al grado de entrada, y el color de los vértices es el valor de asociación entre cada nodo.

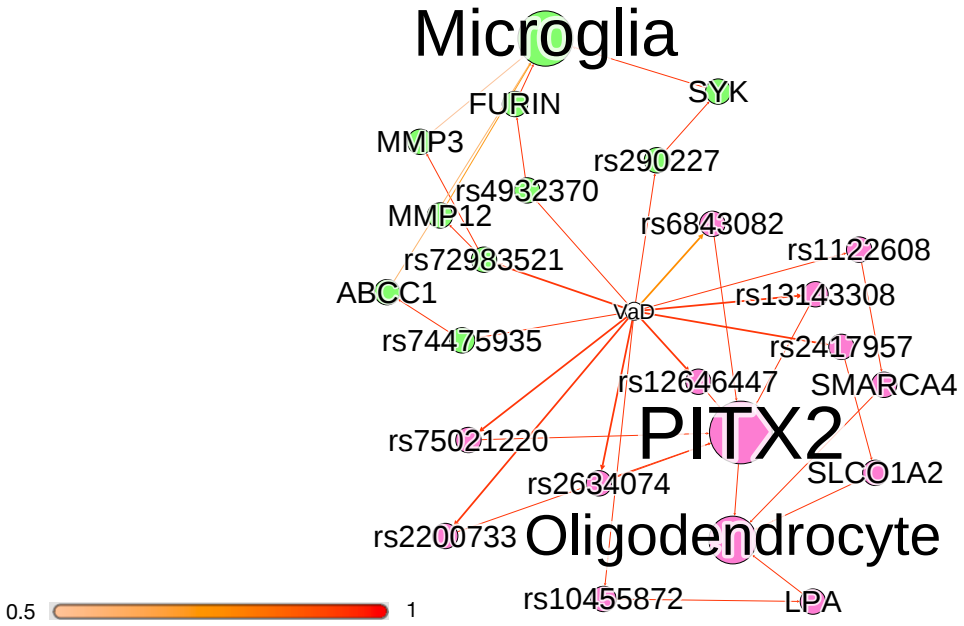


Fig. 18 Red de asociación de Demencia Vascular (VaD), el tamaño de los nodos es proporcional al grado de entrada, y el color de los vértices es el valor de asociación entre cada nodo.

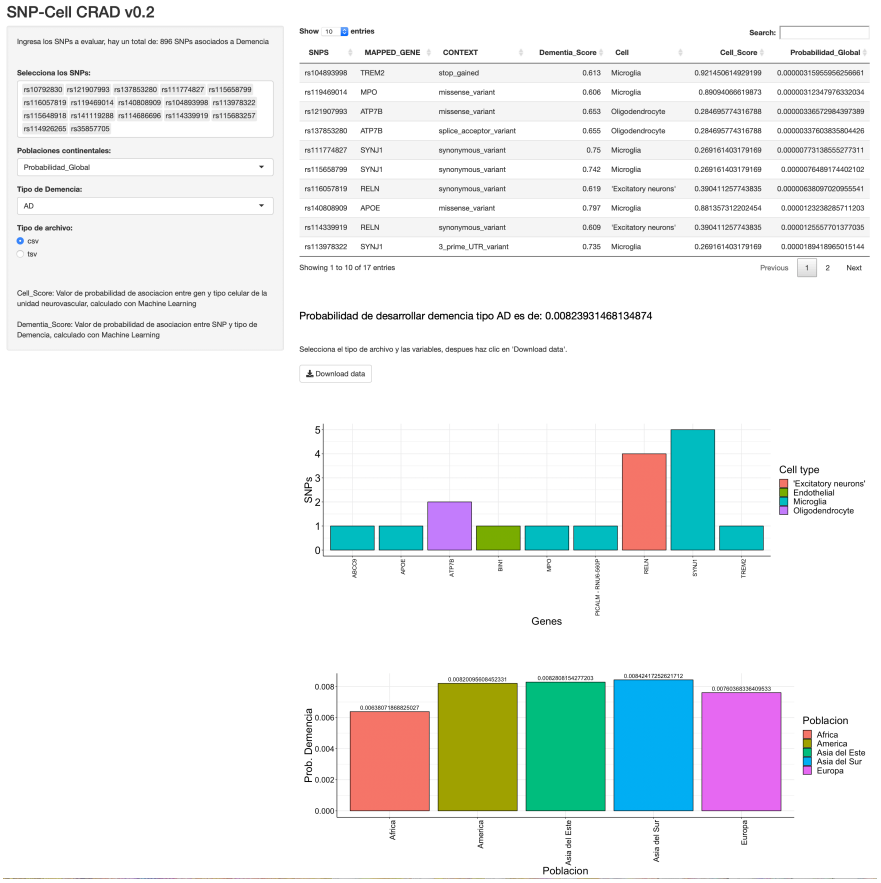


Fig. 19 Interfaz gráfica de la plataforma SNP-Cell Classification and Risk Association of Dementia

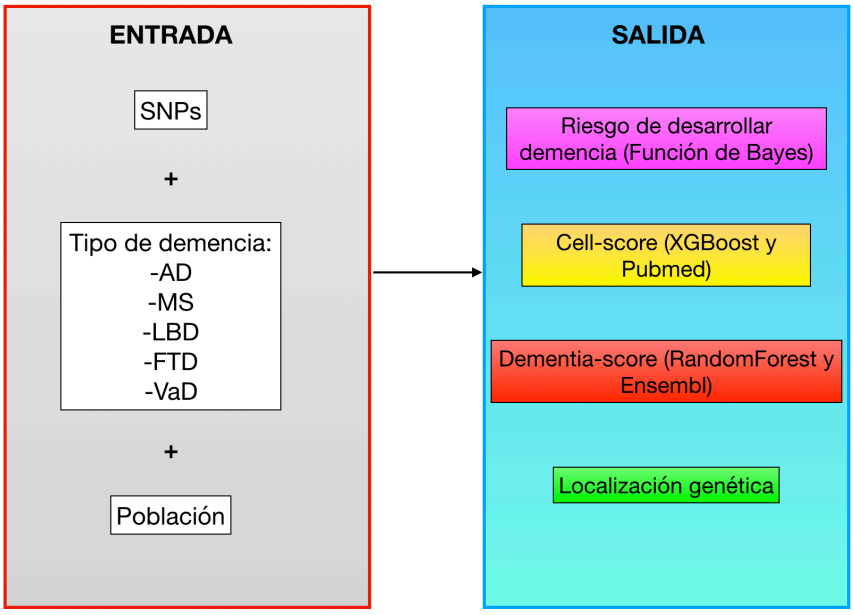


Fig. 20 Pipeline de la Interfaz gráfica de la plataforma SNP-Cell Classification and Risk Association of Dementia. Usa de valores de entrada él ID de los SNPs, el tipo de demencia (considerando solo 5 tipos) y una población a nivel continental. La salida muestra el riesgo de desarrollar la demencia seleccionada en una población en especifico, el tipo celular (cell-score) al que se asocia y su relación con la demencia (dementia-score) en caso de no estar en un estudio GWAS.

Discusión

Aprendizaje de máquina

La selección de características demostró que la naturaleza de los atributos originales son relevantes para realizar la clasificación de nuestro datos (*Tablas suplementarias 2, 4, 6*), se usaron las frecuencias alélicas menores (38 atributos) ya que tuvo un *F-score* de 0.718; nuestro modelo tiene mayores valores de especificidad y sensibilidad para Alzheimer (AD) y esclerosis múltiple (MS), y la clase con menor rendimiento fue demencia frontotemporal (FTD); con una precisión cercana al 0.75 y una especificidad cercana a 1 nos dice que nuestro modelo puede detectar los falsos positivos, por tanto, nos aporta sensibilidad a AD (*fig 7*). Esta diferencia se debe a la cantidad de instancias que se utilizó para entrenar de cada clase, aquellas clases con mayor número de instancias obtuvieron mayores valores de sensibilidad y especificidad; entre más SNPs se obtengan, se incrementa el número de parámetros. Random Forest fue el algoritmo que entrenado, obtuvo el mejor rendimiento, este tipo de algoritmo permite usar variables nominales y de tipo numérico lo que sugiere fuertemente que con los atributos que se integran en bases de datos de libre acceso se puede representar cada SNP para entrenar un robot.

Usamos como control negativo a SNPs de GWAS de osteosarcoma y como control positivo utilizamos SNPs de la base de datos *GRASP* asociados a Alzheimer y Esclerosis Múltiple, esto para establecer un valor corte en la probabilidad de nuestras predicciones (ver *fig. 8,9 y tabla 2*), lo que representa un criterio de selección de SNPs muy riguroso. Usando las tres matrices, se predijeron 294 SNPs con una probabilidad igual o mayor a 0.6 de estar asociados a los diferentes tipos de demencia. Con este enfoque tenemos una manera de predecir asociaciones de SNPs basados en las frecuencias alélicas de todas las poblaciones del mundo y atributos de localización en el genoma; proponemos llamarle a este valor de probabilidad “Dementia Score” (ver fig. 20). Este tipo de análisis nos permite generar una herramienta para predecir asociaciones de SNPs con los tipos de demencia de una manera rápida, económica y cuantificable. Otros autores han utilizado atributos como la frecuencia alélica menor y la posición en el cromosoma para predecir la función de SNPs no codificantes (*Yao Yao, et al., 2019*) y en variaciones genéticas asociadas a autismo (*Jian Zhou et al., 2019*). Aun así para el caso de LBD, VaD y FTD, las predicciones deben tomarse con mucha cautela ya que nuestro modelo tienen valores de especificidad bastante bajos. Desafortunadamente la cantidad de SNPs para cada uno de las clases con bajo rendimiento es limitada, se deben curar aquellos errores de predicción y para futuros trabajos se podrían simular los datos o generar un nuevo modelo solo con las tres clases con menor numero de instancias. Por otro lado hacer una evaluación con el Coeficiente de Correlación de Matthews (MCC) nos hubiese sido mas útil para generar un mejor modelo ya que tenemos clases desbalanceadas, en la *tabla suplementaria 9*

se reportan estos valores de MCC. El modelo generado, usando como datos de prueba los obtenidos en la base de datos de GRASP, nos muestran un 100% de senilidad y especificidad para AD, por tanto, este modelo es tiene excelente rendimiento para la predicción de SNPs asociados a AD. Con este tipo de enfoque podríamos sugerir el diseño de GWAS para poblaciones especificas y la simulación de estos.

Puesto que esta asociación no nos revela información causal de los tipos de demencia, es relevante conocer bajo qué contexto se podrían desenvolver o afectar cada polimorfismo. Para ello se desarrolló un modelo que fuese capaz de aprender de textos relacionados a los tipos celulares de la unidad neurovascular, y predecir los textos de los genes, para así hacer una asociación entre Gen-Célula basándonos en toda la información que se ha publicado en *Pubmed* hasta el 23 de abril de 2019. Este enfoque presenta dos limitantes: i) el entrenamiento depende de los artículos publicados y disponibles en Pubmed, ii) se asume que el contenido del "abstract" presenta información relevante. En la *figura 10* podemos observar que las palabras que describen a nuestros tipos celulares a nivel general provienen de los "*abtracts*" en comparación con los términos MeSH. Derivado de lo anterior se tomaron las palabras con una frecuencia de aparición mayor de 300, para obtener las palabras con mayor presencia en todos los textos, y con la estimación *TF-IDF (ver Ec. 2-3)*, darles un peso por cada clase, de este modo describimos los textos de forma vectorial con las palabras representativas y con la ponderación de relevancia de cada una por cada clase. Para el entrenamiento se optó por XGBoost "gradient boost decisión tree (GBDT)", este algoritmo ha tenido casos de éxito en competencias como "Kaggle" y "KDD cup 2015" (*Tianqi Chen et al., 2016*). XGBoost permite paralelizar la construcción de árboles de decisión penalizando con el error en cada predicción. La ventaja es que nos permite optimizar el manejo masivo de datos, actualizar su aprendizaje, y entrenar nuestro algoritmo en menor tiempo. Al poner a prueba el modelo con conjuntos de datos distintos (selección de características, *tabla suplementaria 7*), el modelo con menor error fue el generado con *InfoGain (fig. 12).* Probablemente por la diferencia en la importancia en la importancia de atributos respecto a GainRatio (*figura suplementaria 3).* Con el modelo generado con *InfoGain* se puede hacer de manera rápida y cuantificable la asociación de un gen y un tipo celular, basándose únicamente en la información publicada, lo que representa una ventaja para poder generar hipótesis en conjunto con un catálogo de estrategias experimentales para demostrarse, ya que se conoce la variación genética y su posible fenotipo, en un contexto celular. Adicionalmente la información y análisis de texto propuesto representa un punto de mejora para el algoritmo SVA0.11 ya que ayudó a optimizar las búsquedas automatizadas que el motor SVA0.11 realiza. Actualmente se sesga la búsqueda con un esquema que depende del manejo del tema por el programador y además plasmarlo en un diagrama simplificado, con el inconveniente de que este proceso depende de información especializada lo que lo hace complejo.

Red de asociación: AD

En la *fig. 14* la mayor variabilidad genética (SNPs) se encuentra en la microglia y en los oligodendrocitos; los genes con mayor numero de SNPs en microglia y oligodendrocito son: PICALM, APOE y CLU.

PICALM se ha reportado estar asociado con la enfermedad de Alzheimer (*Villegas-Llerena et al., 2016, Lo MT et al., 2019*) y estar involucrado en procesos de fagocitosis. APOE tiene un rol importante en la formación de agregados proteicos, ya que en la materia gris esta involucrado en la eliminación de proteínas mieloides mal plegadas, se conoce que los oligodendrocitos, astrocitos y la microglia son las células del cerebro que más expresan este gen (*Zhang Y et al., 2014; C. C. Liu et al., 2013*). En el caso de CLU, también se ha reportado tener mayor expresión en astrocitos y microglia, esta involucrado en la eliminación de fibrillas junto con APOE y en promover la inflamación (*T. Nuutinen et al., 2009*); esta red de asociación nos permite ver que el tipo celular con mayor probabilidad de alteración es la microglia, teniendo sentido que las vías afectadas tengan que ver con la eliminación de agregados proteícos e inflamación como un evento que genera neurodegeneración progresivamente.

En cuanto al tipo variante, es decir, la posición del SNP en el gen, nuestro modelo logró predecir 2 SNPs de tipo “stop gained”, rs104893998 ubicado en el gen TREM2 y rs3211938 en el gen CD36, este tipo de variantes son relevantes ya que el polimorfismo cambia el marco de lectura en la traducción y genera que ésta se detenga, por tanto, produce una proteína con cambios estructurales posibilitando su perdida de función. Se ha estudiado que la perdida de función de TREM2 esta directamente relacionada con la acumulación de proteína beta-mieloide (*Samira Parhizkar et al., 2019*), pero no existe reporte del SNP identificado. En cuando al gen CD36, se ha reportado su papel en la fagocitosis de α-sinucleína y agregados proteicos (*Panicker N et al., 2019*), pero no hay reporte del SNP. Estos dos SNPs podrían ser objetivos de experimentación, ya que tienen un Dementia Score y un Cell Score mayor a 0.6, y además son variantes tipo “stop gained” no publicadas en el contexto de la demencia y el tipo celular.

Red de asociación: FTD

La demencia frontotemporal es caracterizada por una atrofia de los lóbulo temporal y frontal (Nicholas T Olney et al., 2017), en la *fig. 16* se puede observar que los genes con mayor variabilidad genética asociados a FTD son C9orf72, MAPT y MOBP; en el contexto de la microglia y FTD, el gen C9orf72se ha encontrado expresado (*Freischmidt A et al., 2015*) y se sabe que la presencia de GGGGCC en la región promotora podría llevar a una perdida de función celular y una perdida de expresión de este gen, provocando un aumento de citocinas proinflamatorias (*Atanasio A et al., 2016*); bajo el mismo contexto en el gen MAPT, se ha estudiado que la microglia

fagocita y retiene proteínas mal plegadas de tau (Asai H et al., 2015), y se ha demostrado *in-vitro* que tanto microglia de ratón y de humano son capaces de retener y liberar proteínas tau provenientes del espacio extracelular (*Bennett RE et al., 2018*), por otro lado, se ha identificado una relación entre mutaciones intrónicas de MAPT y el aumento de la activación de la microglia (*Lant SB et al., 2014*). En FTD se han asociado mutaciones en GRN, c9orf72 y MAPT al deterioro de los lóbulos frontal y temporal en cerebros *post-mortem* (*Meeter, Kaat, Rohrer, & van Swieten, 2017*).

En el caso de MOBP (*myelin-associated oligodendrocytic basic protein*) asociado a oligodendrocitos, una perdida de función de MOP se ha visto que genera un fenotipo hipomielinizado en ratones, afectando principalmente el sistema nervioso central (*Sperber et al., 2001*), se sabe que se expresa en oligodendrocitos y que esta involucrado en la morfología de estas células a través de la interacción con microtubulos (*Schäfer I et al., 2016*). Los SNPs asociados a FTD son no codificantes, es complicado por ahora conocer el rol que podrían jugar esos SNPs en regiones no codificantes.

Red de asociación: LBD

La enfermedad por cuerpos de Lewy también engloba a la enfermedad de Parkinson, y se caracteriza por generar una acumulación de agregados de alfa-sinucleína, afectando principalmente la sustancia nigra (*Mogi et al., 1994*). Se han asociado 20 SNPs al gen LRRK2 en la microglia (ver fig 17). Este gen ha sido ampliamente estudiado en el contexto del Parkinson, se sabe que la alfa-sinucleína mal plegada genera la activación de la microglia, cuando se depleta a LRRK2 se reducen componentes antioxidantes como SOD2, favoreciendo la desregulación de especies reactivas de oxigeno y promoviendo la inflamación, incitando a pensar que LRKK2 podría estar alterando las vías inflamatorias en la microglia (*Isabella Russo et al., 2019*); se ha reportado su función en la degradación lisosomal y en la modulación de la activación de la microglia (*Schapansky J et al., 2015*). De los SNPs asociados a LRKK2, 7 de ellos son codificantes y podrían ser un blanco de estudio a nivel celular para develar su papel fisiológico si es que afectan la función de LRRK2 en la presencia de alfa sinucleína.

Red de asociación: MS

La esclerosis múltiple (MS) es una enfermedad autoinmune que se caracteriza por la progresiva desmielinización de los axónes, inflamación e infiltración de linfocitos T; los genes con mayor variabilidad en nuestra red de asociación (ver *fig. 15*) son HLA-DRB1, IL2RA y TNFRSF1A, todos con variantes no codificantes. TNFRSF1A codifica al receptor de TNF miembro de la superfamilia 1A. En modelos de ratón, cuando hay

ausencia de TNFR1, los linfocitos T viajan hacia las zonas perivasculares y no invaden la parenquima (*Gimenez et al., 2004*), la vía de señalización de TNFα permite que los linfocitos lleguen a la materia blanca y que recluten células presentadoras de antígeno, lo que ocasiona que las células T, una vez que reconoció al antígeno de la mielina, reclute monocitos y se inicie una respuesta autoinmune (*Angela S. Archambault et al., 2006*). En conjunto ILR2A y HLA-DRB1, se podría pensar que la barrera hematoencefálica en MS es propensa a generar respuestas inflamatorias y el trafico de linfocitos T y monocitos.

Red de asociación: VaD

En el caso de la demencia vascular (VaD), el gen con mayor variabilidad es PITX2 (ver *fig. 18*). Todas las variantes asociadas a VaD son no codificantes. PITX2 se ha asociado como factor de riesgo genético asociado a fibrilación atrial y demencia (*Rollo J et al., 2015*), en modelo murino la depleción de Pitx2 genera una reducción de músculo liso en vasos sanguíneos cerebrales y un aumento en la densidad de estos (*Curtis R. French et al., 2014*).

Universo de biomarcadores y riesgo

En la *fig suplementaria. 4* se muestra todo el universo de SNPs que aun falta por explorar, y en colores tenues de aristas se pueden observar todos aquellos genes que son poco probable de estar asociados a un tipo celular de la unidad neurovascular, por tanto quedan los demás SNPs por estudiar y evaluar; se ha reportado que AD es una enfermedad que para su desarrollo muchos sistemas celulares ajenos no pertenecientes al sistema nervioso central podrían estar contribuyendo a su desarrollo patológico (*Wang J et al., 2017*) por ello es importante recalcar que este enfoque solo es para la unidad neurovascular, hay sistemas celulares que no se tomaron en consideración pero que podría ser relevante que se tomen en cuenta para siguientes estudios, contemplando el cuerpo humano como un todo funcionando en coordinación.

Se ha reportado que aproximadamente la tasa de mutación en una célula humana es de $2.66x10^{-9}$ por par de base (bp) en cada mitosis, por tanto, a mayor número de mitosis mayor tasa de mutaciones somáticas. De estas mutaciones se sabe que en las regiones no codificantes el 50.89% son intergénicas y 35.71% intrónicas (*Brandon Milholland et al., 2016*); la mayoría de las demencias se originan a edades adultas, hace sentido que la acumulación de mutaciones somáticas sea un factor de riesgo que desencadene la demencia, y que si la persona presenta desde el nacimiento variaciones (SNPs), entonces sea mayor la probabilidad y la predisposición a desarrollar algún tipo de demencia, ya que se verían afectados progresivamente los sistemas celulares. Es cierto que poco se conoce de las variantes no codificantes, y

queda mucho aun por estudiar su papel en las demencias, pero con este tipo de enfoques podríamos evaluar esos efectos en un sistema controlado como lo son las líneas celulares.

Tabla 4. Variantes “missense” de Alzheimer.

Gen	SNP	Variante	Class	Dementia-score	Cell-score	Cell	Path name	P-value “reactome.org"	General pathway
ABCA7	rs3752246	missense_variant	AD	0.723	1	Microglia	ABC transporters in lipid homeostasis	0.002035373385738337	Transport of small molecules
APOE	rs429358	missense_variant	AD	1.00	0.61	Oligodendrocyte	Nuclear signaling by ERBB4	1.0881478366253639e-5	Signal Transduction
APOE	rs7412	missense_variant	AD	0.736	0.61	Oligodendrocyte	Nuclear signaling by ERBB4	1.0881478366253639e-5	Signal Transduction
APOE	rs140808909	missense_variant	AD	0.797	0.61	Oligodendrocyte	Nuclear signaling by ERBB4	1.0881478366253639e-5	Signal Transduction
APOE	rs121918392	missense_variant	AD	0.858	0.61	Oligodendrocyte	Nuclear signaling by ERBB4	1.0881478366253639e-5	Signal Transduction
ATP5C1	rs139967528	missense_variant	AD	1.00	1	Oligodendrocyte	Formation of ATP by chemiosmotic coupling	0.0016142616507579532	Metabolism
BIN1	rs112318500	missense_variant	AD	0.731	1	Oligodendrocyte	Clathrin-mediated endocytosis	0.011299831555306006	Vesicle-mediated transport
BIN1	rs138047593	missense_variant	AD	0.684	1	Oligodendrocyte	Clathrin-mediated endocytosis	0.011299831555306006	Vesicle-mediated transport
BSG	rs201850688	missense_variant	AD	0.692	1	Microglia	Defective SLC16A1 causes symptomatic deficiency in lactate transport (SDLT)	4.912970241437442e-4	Disease
C3	rs147859257	missense_variant	AD	0.706	0.688	Microglia	Alternative complement activation	4.211117349803839e-4	Immune System
CD36	rs142186404	missense_variant	AD	0.639	1	Microglia	Transcriptional regulation of white adipocyte differentiation	5.745657024158746e-5	Developmental Biology
IL6R	rs2228145	missense_variant	AD	0.669	1	Microglia	MAPK1 (ERK2) activation	2.1268262738738386e-6	Signal Transduction
MPO	rs119469014	missense_variant	AD	0.606	0.851	Microglia	Events associated with phagocytolytic activity of PMN cells	0.001333520494104401	Immune System
MPO	rs119468010	missense_variant	AD	0.608	0.851	Microglia	Events associated with phagocytolytic activity of PMN cells	0.001333520494104401	Immune System
MPO	rs56378716	missense_variant	AD	0.609	0.851	Microglia	Events associated with phagocytolytic activity of PMN cells	0.001333520494104401	Immune System
PILRA	rs1859788	missense_variant	AD	1	1	Microglia	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	0.022178551375631628	Immune System
PLCG2	rs72824905	missense_variant	AD	1.00	1	Microglia	Erythropoietin activates Phospholipase C gamma (PLCG)	9.124087591241281e-4	Signal Transduction
SERPINA1	rs143370956	missense_variant	AD	0.612	1	Microglia	Cargo concentration in the ER	0.0025968556990454417	Vesicle-mediated transport
TREM2	rs143332484	missense_variant	AD	1.00	1	Microglia	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	1.0909366408906607e-5	Immune System
TREM2	rs75932628	missense_variant	AD	1.00	1	Microglia	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	1.0909366408906607e-5	Immune System

Generamos con la información de las predicciones y con una función de probabilidad la app “SNP-Cell CRAD” v0.2, con ella podemos predecir la asociación entre SNPs y tipos de demencia, generar un contexto celular para el estudio de SNPs, y esto podría funcionar como un catalogo inteligente para proponer experimentos y nuevas hipótesis.

Esta plataforma también nos permite conocer una probabilidad a nivel poblacional de desarrollar demencia dada una combinación de SNPs. Para futuras versiones de la plataforma se considerarán valores epidemiológicos para tomar en cuenta los factores externos que podrían estar involucrados en el desarrollo de estas patologías. En la tabla 4 se muestra como ejemplo los SNPs tipo "*missense*" que probablemente estén afectando la estructura-función de nivel proteína, la plataforma nos permite obtener este tipo de tablas en la que se muestra la asociación a un tipo de demencia, el tipo celular y la función reportada del gen al que pertenece el SNP.

Conclusiones

A. Se exploraron con éxito diversos protocolos de inteligencia artificial y se desarrolló una herramienta con aprendizaje de máquina para la predicción de SNPs asociados a cinco tipos de demencia, y proporcionarles un valor probabilistico, óptimo para la enfermedad de Alzheimer (AD).

B. Se desarrolló y se validó una herramienta basada en aprendizaje maquina mediante XGboost para asociar genes a tipos celulares de la unidad neurovascular. basado en los artículos reportados (textos) mediante XGboost y proporcionar un valor probabilístico.

C. Se implementó un proceso de paralización automática para "Cell Score", con la intención de optimizar los recursos de memoria y procesadores.

D. Se utilizó y adaptó la función de probabilidad de Bayes para estimar la probabilidad de desarrollar alguno de los cinco tipos de demencia, y calcular la probabilidad acumulada de múltiples SNPs en poblaciones continentales específicas.

E. Con toda la información se creo un marco de trabajo con interface web "SNP-Cell CRAD v0.2", con el objeto de crear un catálogo de posibles experimentos y estimar el riego de desarrollar demencia.

F. Se encontraron 235 biomarcadores genéticos posibles en poblaciones especificas a nivel continental.

Perspectivas

- Se requiere establecer una función de estimación de riesgo tomando en cuenta otros factores ambientales y epidemiológicos así cómo estudiar los cinco tipos de demencia.
- Adaptar estas estrategias a modelos experimentales de interés como el caso del modelo murino.
- Explorar aspectos farmacogenéticos del producto de la expresión de los genes y sus variantes o polimorfismos.

- Implementar protocolos de última generación como el aprendizaje federado con estrategias de aprendizaje profundo como proceso de automatización de los modelos creados en este trabajo.
- Explorar los otros tipos de SNPs en otros tipos celulares.
- Estudiar los SNPs no codificantes.
- Mejorar el modelo para VaD, LBD y FTD.

Bibliografía

Abbott, N. J. and Friedman, A. (2012), Overview and introduction: The blood–brain barrier in health and disease. Epilepsia, 53: 1-6. doi:10.1111/j.1528-1167.2012.03696.x

Alistar Burns; Steve Iliffe. (14 February 2009). Dementia. BMJ, 338, 405-409.

Altshuler, D; Pollara, V J; Cowles, C R; Van Etten, W J; Baldwin, J; Linton, L; Lander, E S (2000). «An SNP map of the human genome generated by reduced representation shotgun sequencing». *Nature* **407** (6803): 513-6.

Alzheimer's Association . (2018). 2018 Alzheimer's disease facts and figures.. Alzheimer's & Dementia. ELSEVIER, 14, 367-429.

Andrassi MG. Metabolic syndrome, diabetes and atherosclerosis: influ- ence of gene-enviroment interaction. Mutat Res. 2009;667:35-43.

Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. (2016). caret: Classification and Regression Training. R package version 6.0-71.https://CRAN.R-project.org/package=caret

Anzela Niraula, Kristina G. Witcher, John F. Sheridan, Jonathan P. Godbout, Interleukin-6 Induced by Social Stress Promotes a Unique Transcriptional Signature in the Monocytes That Facilitate Anxiety, Biological Psychiatry, 2018, ISSN 0006-3223.

Atkinson, A. J., Colburn, W. A., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J. and Zeger, S. L. (2001), Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics, 69: 89-9.

Bagui, S.C. (2005). Combining Pattern Classifiers: Methods and Algorithms. Technometrics (Vol. 47). https://doi.org/10.1198/tech.2005.s320

Bastide, M., Ouk, T., Plaisier, F., Pétrault, O., Stolc, S., & Bordet, R. (2007). Neurogliovascular unit after cerebral ischemia: Is the vascular wall a pharmacological target. *Psychoneuroendocrinology*, *32*(SUPPL 1), 36–39. https://doi.org/10.1016/j.psyneuen.2007.03.015

Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., Morris, J. C. (2012). Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. *New England Journal of* Medicine, 367(9), 795–804. https://doi.org/10.1056/NEJMoa1202753

Baye, T. M., Abebe, T., & Wilke, R. A. (2011). Genotype-environment interactions and their translational implications. *Personalized Medicine*, *8*(1), 59–70. https://doi.org/10.2217/pme.10.7

Bertram, L., & Tanzi, R. E. (2008). Thirty years of Alzheimer's disease genetics: the implications of systematic meta analyses. *Nature Reviews Neuroscience*, *9*(10), 768–778. https://doi.org/10.1038/nrn2494

Bezzini D., Battaglia M.A. (2017) Multiple Sclerosis Epidemiology in Europe. In: Asea A., Geraci F., Kaur P. (eds) Multiple Sclerosis: Bench to Bedside. Advances in Experimental Medicine and Biology, vol 958. Springer, Cham.

Blessed G, Tomlinson BE, Roth M. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. Br J Psychiatry 1968;114: 797-811.

Bo Wang, Armin Pourshafeie, Marinka Zitnik, Junjie Zhu, Carlos D. Bustamante, Serafim Batzoglou & Jure Leskovec. (2018). Network enhancement as a general method to denoise weighted biological networks. Nature communications, 9, 1-8.

Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324

Brown, A. D., & Kachura, J. R. (2019). Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization. *Journal of the American College of Radiology*, 1–5. https://doi.org/10.1016/j.jacr.2018.12.004

Changhong Xing, Kazuhide Hayakawa, Josephine Lok, Ken Arai & Eng H Lo (2012) Injury and repair in the neurovascular unit, Neurological Research, 34:4, 325-330, DOI: 10.1179/1743132812Y.0000000019

Clark, M. M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., Kingsmore, S. F. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science Translational Medicine*, *11*(489), eaat6177. https://doi.org/10.1126/scitranslmed.aat6177

Courtney Humphries. (2015). Can We Identify Every Kind of Cell in the Body?. USA (May 18, 2015). MIT Technology Review.

Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org.

David B. Hogan, Kirsten M. Fiest, Jodie I. Roberts, Colleen J. Maxwell, Jonathan Dykeman, Tamara Pringsheim, Thomas Steeves, Eric E. Smith, Dawn Pearson, Nathalie Jetté. (2016). The Prevalence and Incidence os Dementia with Lewy bodies: a Systematic Review. The Canadian Journal of Neurological Sciences Inc., 43, S83-95.

David B. Hogan, Nathalie Jetté, Kirsten M. Fiest, Jodie I. Roberts, Dawn Pearson, Eric E. Smith, Pamela Roach, Andrew Kirk, Tamara Pringsheim, Collen J. Maxwell. (2016). The Prevalence and Incidence of Frontotemporal Dementia: a Systematic Review. The Canadian Journal of Neurological Sciences Inc., 43, S96-S109.

Dooley MA, Hogan SL. Environmental epidemiology and risk factors for autoimmune disease. Curr Opin Rheumatol. 2003;15:99-103.

Drabovich, A.P.; Krylov, S.N. (2006). Identification of base pairs in single-nucleotide polymorphisms by MutS protein-mediated capillary electrophoresis. *Analytical chemistry* **78** (6): 2035-8.

Edgar Correa, Víctor Paredes, Braulio Martínez. (2016). Prevalence of multiple sclerosis in Latin America and its relationship with European migration . Multiple Sclerosis Journal - Experimental, Translational and Clinical, 2, 2055217316666407.

Edward G. Jones (1999) Golgi, Cajal and the Neuron Doctrine, Journal of the History of the Neurosciences, 8:2, 170-178, DOI: 10.1076/jhin.8.2.170.1838

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

EMSP. (2015). Multiple sclerosis in Europe. 2018, de EMSP Sitio web: http://www.emsp.org/wp-content/uploads/2015/08/MS-in-EU-access.pdf

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. (2018 Jan 4). The Reactome Pathway Knowledgebase. Nucleic Acids Res, 46, D649-D655.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer.

Georgina M. Aldridge; Allison Birnschein; Natalie L. Denburg; Nandakumar S. Narayanan. (12 March 2018). Parkinson's Disease Dementia and Dementia with Lewy Bodies Have Similar Neuropsychological Profiles. Frontiers in Neurology, 9, 1-8.

Gordon, E., Rohrer, J. D. & Fox, N. C. Advances in neuroimaging in frontotemporal dementia. J. Neurochem. 138, 193–210 (2016).

Gregorio Alanis-Lobato. (23 September 2015). Mining protein interactomes to improve their reliability and support the advancement of network medicine. Frontiers in Genetics, 6, 1-8.

Griffin, T J; Smith, L M (2000). Genetic identification by mass spectrometric analysis of single-nucleotide polymorphisms: ternary encoding of genotypes. *Analytical chemistry* **72** (14): 3298-302.

Hamanaka, G., Ohtomo, R., Takase, H., Lok, J., & Arai, K. (2018). Role of oligodendrocyte-neurovascular unit in white matter repair. *Neuroscience Letters*, *684*(June), 175–180. https://doi.org/10.1016/j.neulet.2018.07.016.

Hawkins, B.T., Lundeen, T.F., Norwood, K.M. et al. Diabetologia (2007) 50: 202. https://doi.org/10.1007/s00125-006-0485-z.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, *3*(November), 1–11. https://doi.org/10.3389/neuro.09.031.2009

Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, Quaid Morris, Yoseph Barash, Adrian R. Krainer, Nebojsa Jojic, Stephen W. Scherer, Benjamin J. Blencowe, Brendan J. Frey. (9 January 2015). The human splicing code reveals new insights into the genetic determinants of disease. Science, 347, 1-8.

Institute for Health Metrics and Evaluation (IHME). GBD Compare. Seattle, WA: IHME, University of Washington, 2015. Available from http://vizhub.healthdata.org/gbd-compare. (Accessed [02-25-2020])

Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucl. Acids Res. (2016).

Kalia, M., & Costa, J. (2015). and future prospects. *Metabolism*, *64*(3), S11–S15. https://doi.org/10.1016/j.metabol.2014.10.026.

Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353-D361 (2017).

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, *15*, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005.

Kristin K. Nicodemus, Amanda J. Law, Eugenia Radulescu, Augustin Luna, Bhaskar Kolachana, Radhakrishna Vakkalanka, Dan Rujescu, Ina Giegling, Richard E. Straub, Kate McGee, Bert Gold, Michael Dean, Pierandrea Muglia, Joseph H. Callicott, Hao-Yang Tan, Daniel R. Weinberger . (2010). Biological Validation of Increased Schizophrenia Risk With NRG1, ERBB4, and AKT1 Epistasis via Functional Neuroimaging in Healthy Controls. Arch gen Psychiatry, 67, 991-1001.

Kunes J, Zicha J. The interaction of genetic and enviromental factors in the etiology of hypertension. Physiol Res. 2009;58 Suppl 2:33-41.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4.

Lee, J. S., Kim, C., Shin, J. H., Cho, H., Shin, D. S., Kim, N., … Seong, J. K. (2018). Machine Learning-based Individual Assessment of Cortical Atrophy Pattern in Alzheimer's Disease Spectrum: Development of the Classifier and Longitudinal Evaluation. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-22277-x.

Liara Rizzi, Idiane Rosset, Matheus Roriz-Cruz. (2014). Global Epidemiology of Dementia: Alzheimer's and Vascular Types. BioMed Research International, 2014, 8.

Liu, C.-Y., Yang, Y., Ju, W.-N., Wang, X., & Zhang, H.-L. (2018). Emerging Roles of Astrocytes in Neuro-Vascular Unit and the Tripartite Synapse With Emphasis on Reactive Gliosis in the Context of Alzheimer's Disease. *Frontiers in Cellular Neuroscience*, *12*(July), 1–12. https://doi.org/10.3389/fncel.2018.00193.

Loredo M, Sánchez-Méndez, J. Beltran N, Cuevas E, Paniagua D, Rivera HM*. 2018. Registro en trámite. Selección de variantes o SNPs Automatizada asociada a demencia (SVA v1.06). Indautor. México.

Ludmila I. Kuncheva. (2004). Combining Pattern Classifiers. Methods and Algorithms. New Yersey, USA: Wiley-Interscience.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington Z, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, and Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, 2017, Vol. 45 (Database issue): D896-D901.

Mariza de Andrade, E. Warwick Daw, Aldi T. Kraja, Virginia Fisher, Lan Wang, Ke Hu, Jing Li, Razvan Romanescu, Jenna Veenstra, Rui Sun, Haoyi Weng and Wenda Zhou. (2018). The challenge of detecting genotype-by- methylation interaction: GAW20. BMC Genetics, 19, 119-125.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams,Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Thorsten Joachims. (2001). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.

Meeter, L. H., Kaat, L. D., Rohrer, J. D., & van Swieten, J. C. (2017). Imaging and fluid biomarkers in frontotemporal dementia. *Nature Reviews Neurology*, *13*(7), 406–419. https://doi.org/10.1038/nrneurol.2017.75.

Melissa A. Powell, Raiford T. Black, Terry L. Smith, Thomas M. Reeves and Linda L. Phillips. (2018). Matrix metalloproteinase 9 and osteopontin interact to support synaptogenesis in the olfactory bulb following mild traumatic brain injury. Journal of Neurotrauma, DOI: 10.1089/neu.2018.5994, 1-52.

Mieko Ogino, Shuichi Okamoto, Hiroyuki Ohta, Mariko Sakamoto, Yusuke Nakamura, Kosuke Iwasaki, Manami Yoshida, Shinzo Hiroi, Izumi Kawachi. (2017 Nov). Prevalence, treatments and medical cost of multiple sclerosis in Japan based on analysis of health insurance claims database. Clinical & Experimental Neuroimmunology, 8, 318-326.

Muoio, V. , Persson, P. B. and Sendeski, M. M. (2014), The neurovascular unit concept review. Acta Physiol, 210: 790-798. doi:10.1111/apha.12250.

Novelli, G., Ciccacci, C., Borgiani, P., Amati, M. P., & Abadie, E. (2008). Genetic tests and genomic biomarkers: Regulation, qualification and validation. *Clinical Cases in Mineral and Bone Metabolism*.

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2018. World Wide Web URL: https://omim.org/

Organización Mundial de la Salud. Clasificación Internacional de Enfermedades, décimo primera edición (CIE-11). Geneva: WHO; 2018.

Otsuka, E., Wallace, S. A., & Chiu, D. (2016). A hashtag recommendation system for twitter data streams. *Computational Social Networks*, *3*(1). https://doi.org/10.1186/s40649-016-0028-9

Park Y.M., Squizzato S., Buso N., Gur T., Lopez R. (2017)The EBI search engine: EBI search as a service making biological data accessible for all Nucleic Acids Research, May 2, 2017.

Pattisapu, N., Gupta, M., Kumaraguru, P., & Varma, V. (2019). A Distant Supervision Based Approach to Medical Persona Classification. *Journal of Biomedical Informatics*, *94*(September 2018), 103205. https://doi.org/10.1016/j.jbi.2019.103205.

Peek, N., Holmes, J. H., & Sun, J. (2014). Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics. *IMIA Yearbook*, *9*(1), 42–47. https://doi.org/10.15265/IY-2014-0018.

Ramírez-Bello J, Pérez-Méndez O, Ramírez-Fuentes S, Carrillo-Sánchez S, Vargas-Alarcón G, Fragoso JM. Genetic and genomic studies in hy- pertension: an actualization of the genomic studies. Arch Cardiol Mex. 2011;81:240-50.

Sanchéz-Mendez,J., *et al*. 2017. Libro. Edén del Conocimiento. Cap libro. Búsqueda y selección de marcadores moleculares asociados a enfermedades crónico degenerativas *in silico*: un análisis de interacciones complejas. ISBN: 978-607-8093-86-1

Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision_Tree_Analysis_on_J48_Algorithm. *International Journal of Advanced Research InComputer Science and Software Engineering*, *3*(6), 1114–1119. Retrieved from https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of geneticvariation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11)

Shinichi Hirose, A new paradigm of channelopathy in epilepsy syndromes: Intracellular trafficking abnormality of channel molecules, Epilepsy Research, Volume 70, Supplement, 2006, Pages 206-217, ISSN 0920-1211, https://doi.org/10.1016/j.eplepsyres.2005.12.007.

Shu Yang, Hong Jin & Zhigang Zhao (2018) An ECV304 monoculture model for permeability assessment of blood–brain barrier, Neurological Research, 40:2, 117-121, DOI: 10.1080/01616412.2017.1398882

Sperlin RA, Aisen PS, Beckett DA, Craft S, Fagan AM, et el. Towards defining the preclinical stages of Alzheimer´s disease: recommendations from the National Institute of Aging and the Alzheimer´s Association workgroups on diagnostic guidelines for Alzheimer´s disease. Alzheimers Dement 2011; 7: 280-92.

Su MW, Tung KY, Liang PH, Tsai CH, Kuo NW, Lee YL. Gene-gene and gene-environmental interaction of childhood asthma: a multifactor di- mension reduction approach. PLoS One. 2012;7:e30694.

Tahira, T.; Kukita, Y.; Higasa, K.; Okazaki, Y.; Yoshinaga, A.; Hayashi, K. (2009). Estimation of SNP allele frequencies by SSCP analysis of pooled DNA. *Methods Mol Biol*. Methods in Molecular Biology **578**: 193-207.

Telenti, A., et al. (2016). «Deep sequencing of 10,000 human genomes». *Proceedings of the National Academy of Sciences, 201613365*.

The Ronald and Nancy Reagan Research Institute of the Alzheimer's Association and the National Institute on Aging Working Group. Consensus report of the Working Group on: "molecular and biochemical markers of Alzheimer's disease". Neurobiol. Aging 19, 109–116 (1998).

The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45: D158-D169 (2017).

Thurgur, H., & Pinteaux, E. (2019). Microglia in the Neurovascular Unit: Blood–Brain Barrier–microglia Interactions After Central Nervous System Disorders. *Neuroscience*, *405*, 55–67. https://doi.org/10.1016/j.neuroscience.2018.06.046

Visscher PM; Wray NR, Zhang Q; Sklar P; McCarthy MI; Brown MA; Yang J. (2017 Jul 6). 10 Years of GWAS Discovery: Biology, Function, and Translation.. CellPress, 101, 5-22.

Wheeler, M. T., Dewey, F. E., Knowles, J. W., Pavlovic, A. B., Quake, S. R., Pushkarev, D., Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *Www.Thelancet.Com Lancet*, *375*, 1525–35. https://doi.org/10.1016/S0140-6736(10)60599-5

William Montgomery, Kaname Ueda, Margaret Jorgensen, Shari Stathis, Yuanyuan Cheng, Tomomi Nakamura. (2018). Epidemiology, associated burden, and current clinical practice for the diagnosis and management of Alzheimer's disease in Japan . ClinicoEconomics and Outcomes Research, 10, 13-28.

William M Armstead & Ramesh Raghupathi (2011) Endothelin and the neurovascular unit in pediatric traumatic brain injury, Neurological Research, 33:2, 127-132, DOI: 10.1179/016164111X12881719352138

World Alzheimer Report 2015 Alzheimer´s Association. Alzheimers Dement. 13, 325-373 (2017)

Wu, Y.-T-, Brayne, C., & Matthews, F. E. (2015). Prevalence of dementia in East Asia: a synthetic review of time trends. International Journal of Geriatric Psychiatry, 30(8), 793-801.

Xiaoqian Wang, Jingwen Yan, Xiaohui Yao, Sungeun Kim, Kwangsik Nho, Shannon L. Risacher, Andrew J. Saykin Li Shen And Heng Huang. (2018). Longitudinal Genotype–Phenotype Association Study through Temporal Structure Auto-Learning Predictive Model. Journal of Computational Biology, 25, 809-824.

Xinguang Yang, Qingmei Huang, Huacai Yang, Si Liu, Baikeng Chen, Tianni Liu, Jie Yang, Haiyan Yao, Shaopeng Lin, Xiaohui Chen, Honghua Zhuang, Youming long, Cong Gao, Astrocytic damage in glial fibrillary acidic protein astrocytopathy during initial attack, Multiple Sclerosis and Related Disorders, Volume 29, 2019, Pages 94-99, ISSN 2211-0348, https://doi.org/10.1016/j.msard.2019.01.036.

Zanni, G. R., & Wick, J. Y. (2007). Differentiating dementias in long-term care patients. *Consultant Pharmacist*, *22*(1), 14–28.

Material Suplementario

Tabla Suplementaria 1. Tipo de atributos usados en el modelo para relacionar SNPs con demencia

Atributo	Base de datos	Tipo de dato
ALL	ENSEMBL	Número
ALL_MAF	ENSEMBL	Número
AFR	ENSEMBL	Número
AFR_MAF	ENSEMBL	Número
AMR	ENSEMBL	Número
AMR_MAF	ENSEMBL	Número
EAS	ENSEMBL	Número
EAS_MAF	ENSEMBL	Número
EUR	ENSEMBL	Número
EUR_MAF	ENSEMBL	Número
SAS	ENSEMBL	Número
SAS_MAF	ENSEMBL	Número
CHB	ENSEMBL	Número
CHB_MAF	ENSEMBL	Número
JPT	ENSEMBL	Número
JPT_MAF	ENSEMBL	Número
CHS	ENSEMBL	Número
CHS_MAF	ENSEMBL	Número
CDX	ENSEMBL	Número
CDX_MAF	ENSEMBL	Número
KHV	ENSEMBL	Número
KHV_MAF	ENSEMBL	Número
CEU	ENSEMBL	Número
CEU_MAF	ENSEMBL	Número
TSI	ENSEMBL	Número
TSI_MAF	ENSEMBL	Número
FIN	ENSEMBL	Número
FIN_MAF	ENSEMBL	Número
GBR	ENSEMBL	Número
GBR_MAF	ENSEMBL	Número
IBS	ENSEMBL	Número
IBS_MAF	ENSEMBL	Número
YRI	ENSEMBL	Número
YRI_MAF	ENSEMBL	Número
ITU	ENSEMBL	Número

Atributo	Base de datos	Tipo de dato
LWK	ENSEMBL	Número
LWK_MAF	ENSEMBL	Número
GWD	ENSEMBL	Número
GWD_MAF	ENSEMBL	Número
MSL	ENSEMBL	Número
MSL_MAF	ENSEMBL	Número
ESN	ENSEMBL	Número
ESN_MAF	ENSEMBL	Número
ASW	ENSEMBL	Número
ASW_MAF	ENSEMBL	Número
ACB	ENSEMBL	Número
ACB_MAF	ENSEMBL	Número
MXL	ENSEMBL	Número
MXL_MAF	ENSEMBL	Número
PUR	ENSEMBL	Número
PUR_MAF	ENSEMBL	Número
CLM	ENSEMBL	Número
CLM_MAF	ENSEMBL	Número
PEL	ENSEMBL	Número
PEL_MAF	ENSEMBL	Número
GIH	ENSEMBL	Número
GIH_MAF	ENSEMBL	Número
PJL	ENSEMBL	Número
PJL_MAF	ENSEMBL	Número
BEB	ENSEMBL	Número
BEB_MAF	ENSEMBL	Número
STU	ENSEMBL	Número
STU_MAF	ENSEMBL	Número
ITU_MAF	ENSEMBL	Número
REGION	GWASCatalog	Nominal
CHR_ID	GWASCatalog	Número
CHR_POS	GWASCatalog	Número
MAPPED_GENE	GWASCatalog	Nominal
CONTEXT	dbSNP	Nominal
Clase	GWASCatalog	Nominal

Tabla Suplementaria 2. Especificidad del modelo usando como conjunto de entrenamiento 69 atributos

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.83	0.876	0.868	0.844	0.859	0.882
Relief	0.83	0.876	0.868	0.844	0.859	0.882
PCA	0.847	0.86	0.865	0.845	0.866	0.856
GainRatio	0.83	0.876	0.868	0.844	0.859	0.882
S/selección	0.83	0.876	0.868	0.844	0.859	0.882

Tabla Suplementaria 3. Especificidad del modelo usando como conjunto de entrenamiento 38 atributos.

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.83	0.877	0.872	0.878	0.864	0.878
Relief	0.83	0.877	0.872	0.878	0.864	0.878
PCA	0.862	0.86	0.865	0.864	0.842	0.87
GainRatio	0.83	0.877	0.872	0.878	0.864	0.878
S/selección	0.83	0.877	0.872	0.878	0.864	0.878

Tabla Suplementaria 4. Sensibilidad del modelo usando como conjunto de entrenamiento 69 atributos.

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.61	0.694	0.703	0.693	0.651	0.708
Relief	0.61	0.694	0.703	0.693	0.651	0.708
PCA	0.545	0.585	0.674	0.673	0.654	0.63
GainRatio	0.61	0.694	0.703	0.693	0.651	0.708
S/selección	0.61	0.694	0.703	0.693	0.651	0.708

Tabla Suplementaria 5. Sensibilidad del modelo usando como conjunto de entrenamiento 38 atributos

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.61	0.698	0.704	0.718	0.669	0.703
Relief	0.61	0.698	0.704	0.718	0.669	0.703
PCA	0.54	0.598	0.673	0.671	0.661	0.64
GainRatio	0.61	0.698	0.704	0.718	0.669	0.708
S/selección	0.61	0.698	0.704	0.718	0.669	0.703

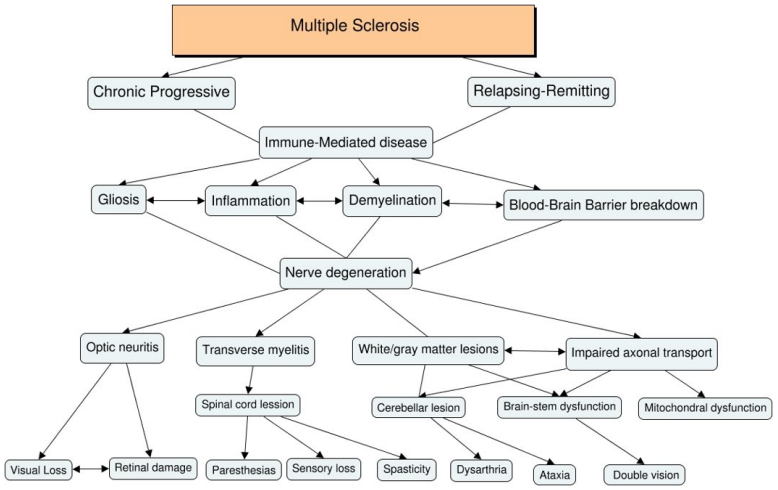
Tabla Suplementaria 6. F-score del modelo usando como conjunto de entrenamiento 69 atributos.

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.594	0.679	0.675	0.667	0.624	0.691
Relief	0.594	0.679	0.675	0.667	0.624	0.691
PCA	0.54	0.577	0.655	0.653	0.64	0.615
GainRatio	0.594	0.679	0.675	0.667	0.624	0.691
S/selección	0.594	0.679	0.675	0.667	0.624	0.691

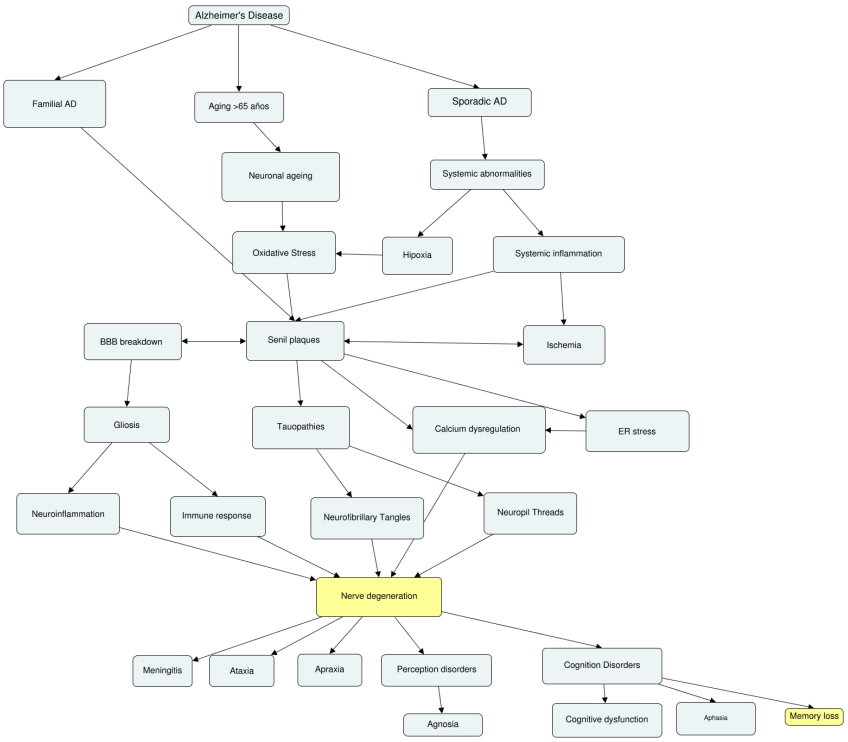
Tabla Suplementaria 7. F-score del modelo usando como conjunto de entrenamiento 38 atributos.

Selección de características / Algoritmo	J48	LMT	RandomForest 1000I	RandomForest	Vote (5)	Vote (forest)
InfoGain	0.594	0.682	0.681	0.694	0.646	0.686
Relief	0.594	0.682	0.681	0.694	0.646	0.686
PCA	0.541	0.588	0.651	0.65	0.64	0.625
GainRatio	0.594	0.682	0.681	0.694	0.646	0.686
S/selección	0.594	0.682	0.681	0.694	0.646	0.686

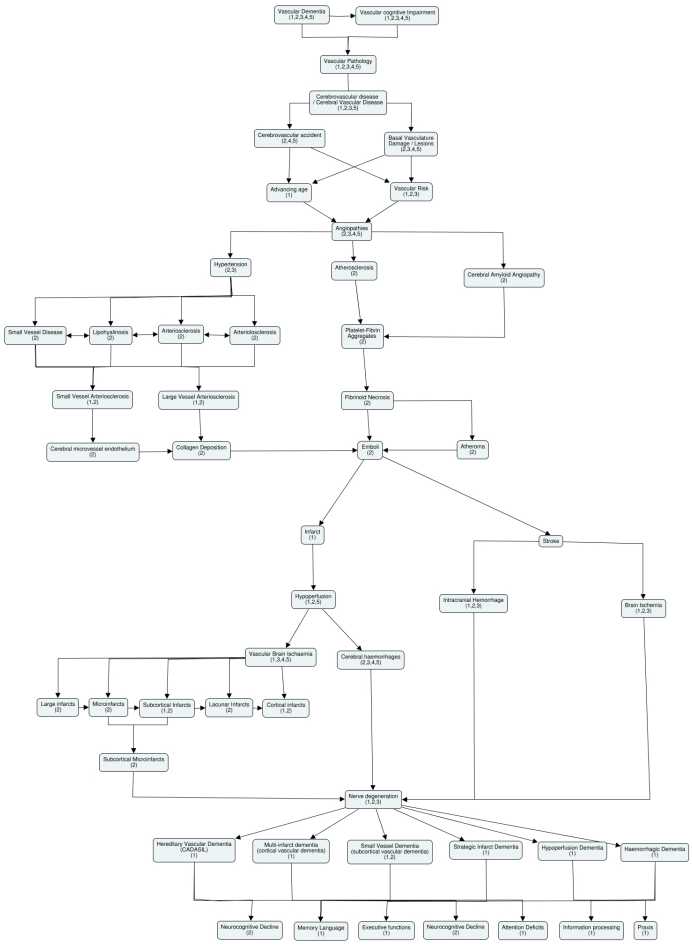
A)


Multiple Sclerosis
Chronic Progressive
Relapsing-Remitting
Immune-Mediated disease
Gliosis
Inflammation
Demyelination
Blood-Brain Barrier breakdown
Nerve degeneration
Optic neuritis
Transverse myelitis
White/gray matter lesions
Impaired axonal transport
Spinal cord lession
Cerebellar lesion
Brain-stem dysfunction
Mitochondral dysfunction
Visual Loss
Retinal damage
Paresthesias
Sensory loss
Spasticity
Dysarthria
Ataxia
Double vision

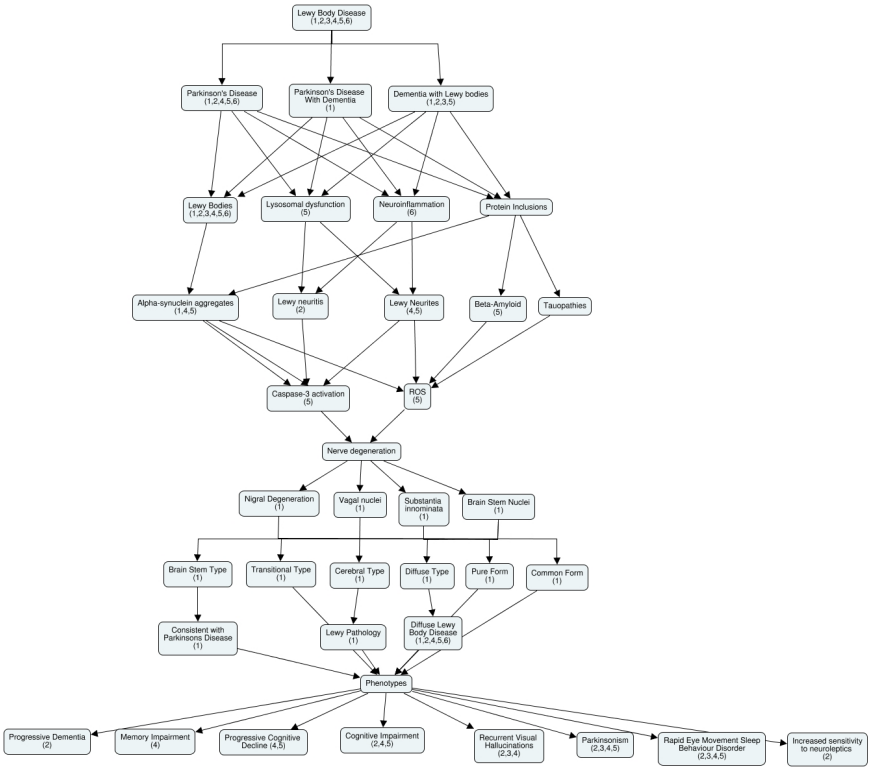
B)


Alzheimer's Disease
Familial AD
Aging >65 años
Sporadic AD
Neuronal ageing
Systemic abnormalities
Oxidative Stress
Hipoxia
Systemic inflammation
BBB breakdown
Senil plaques
Ischemia
Gliosis
Tauopathies
Calcium dysregulation
ER stress
Neuroinflammation
Immune response
Neurofibrillary Tangles
Neuropil Threads
Nerve degeneration
Meningitis
Ataxia
Apraxia
Perception disorders
Cognition Disorders
Agnosia
Cognitive dysfunction
Aphasia
Memory loss

C)


Vascular Dementia (1,2,3,4,5)
Vascular cognitive Impairment (1,2,3,4,5)
Vascular Pathology (1,2,3,4,5)
Cerebrovascular disease Cerebral Vascular Disease (1,2,3,5)
Cerebrovascular accident (2,4,5)
Basal Vasculature Damage / Lesions (2,3,4,5)
Advancing age (1)
Vascular Risk (1,2,3)
Angiopathies (2,3,4,5)
Hypertension (2,3)
Atherosclerosis (2)
Cerebral Amyloid Angiopathy (2)
Small Vessel Disease (2)
Lipohyalinosis (2)
Arteriosclerosis (2)
Arteriolosclerosis (2)
Platelet-Fibrin Aggregates (2)
Small Vessel Arteriosclerosis (1,2)
Large Vessel Arteriosclerosis (1,2)
Fibrinoid Necrosis (2)
Cerebral microvessel endothelium (2)
Collagen Deposition (2)
Emboli (2)
Atheroma (2)
Infarct (1)
Stroke
Hypoperfusion (1,2,5)
Intracranial Hemorrhage (1,2,3)
Brain Ischemia (1,2,3)
Vascular Brain Ischaemia (1,2,4,5)
Cerebral haemorrhages (2,3,4,5)
Large infarcts (2)
Microinfarcts (2)
Subcortical Infarcts (1,2)
Lacunar Infarcts (2)
Cortical infarcts (1,2)
Subcortical Microinfarcts (2)
Nerve degeneration (1,2,3)
Hereditary Vascular Dementia (CADASIL) (1)
Multi-infarct dementia (cortical vascular dementia) (1)
Small Vessel Dementia (subcortical vascular dementia) (1,2)
Strategic Infarct Dementia (1)
Hypoperfusion Dementia (1)
Haemorrhagic Dementia (1)
Neurocognitive Decline (2)
Memory Language (1)
Executive functions (1)
Neurocognitive Decline (2)
Attention Deficits (1)
Information processing (1)
Ptosis (1)

D)


Lewy Body Disease (1,2,3,4,5,6)
Parkinson's Disease (1,2,4,5,6)
Parkinson's Disease With Dementia (1)
Dementia with Lewy bodies (1,2,3,5)
Lewy Bodies (1,2,3,4,5,6)
Lysosomal dysfunction (5)
Neuroinflammation (6)
Protein Inclusions
Alpha-synuclein aggregates (1,4,5)
Lewy neuritis (2)
Lewy Neurites (4,5)
Beta-Amyloid (5)
Tauopathies
Caspase-3 activation (5)
ROS (5)
Nerve degeneration
Nigral Degeneration (1)
Vagal nuclei (1)
Substantia innominata (1)
Brain Stem Nuclei (1)
Brain Stem Type (1)
Transitional Type (1)
Cerebral Type (1)
Diffuse Type (1)
Pure Form (1)
Common Form (1)
Consistent with Parkinsons Disease (1)
Lewy Pathology (1)
Diffuse Lewy Body Disease (1,2,4,5,6)
Phenotypes
Progressive Dementia (2)
Memory Impairment (4)
Progressive Cognitive Decline (4,5)
Cognitive Impairment (2,4,5)
Recurrent Visual Hallucinations (2,3,4)
Parkinsonism (2,3,4,5)
Rapid Eye Movement Sleep Behaviour Disorder (2,3,4,5)
Increased sensitivity to neuroleptics (2)

E)

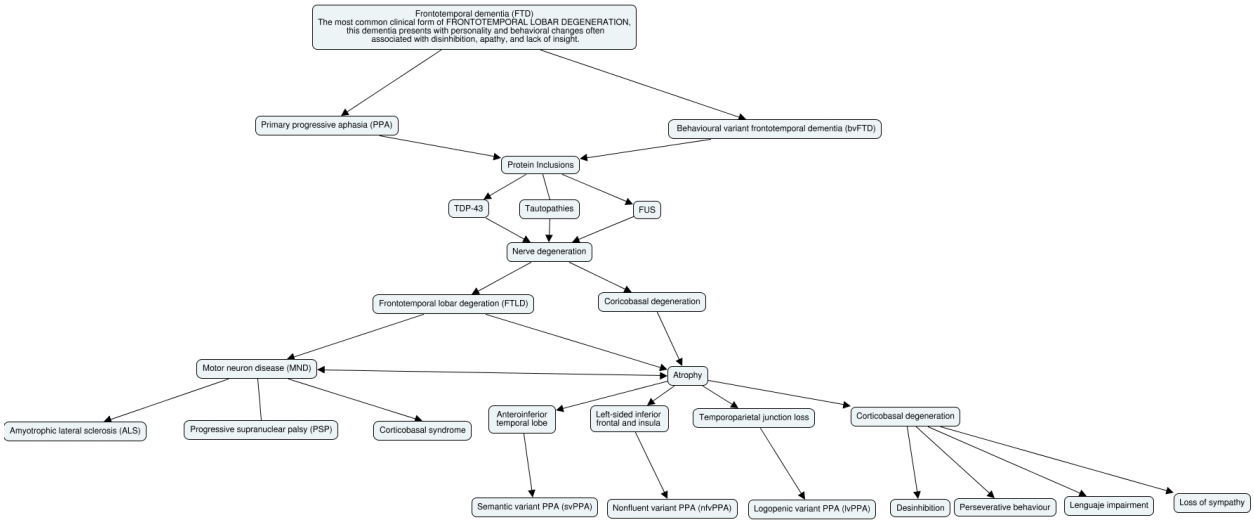
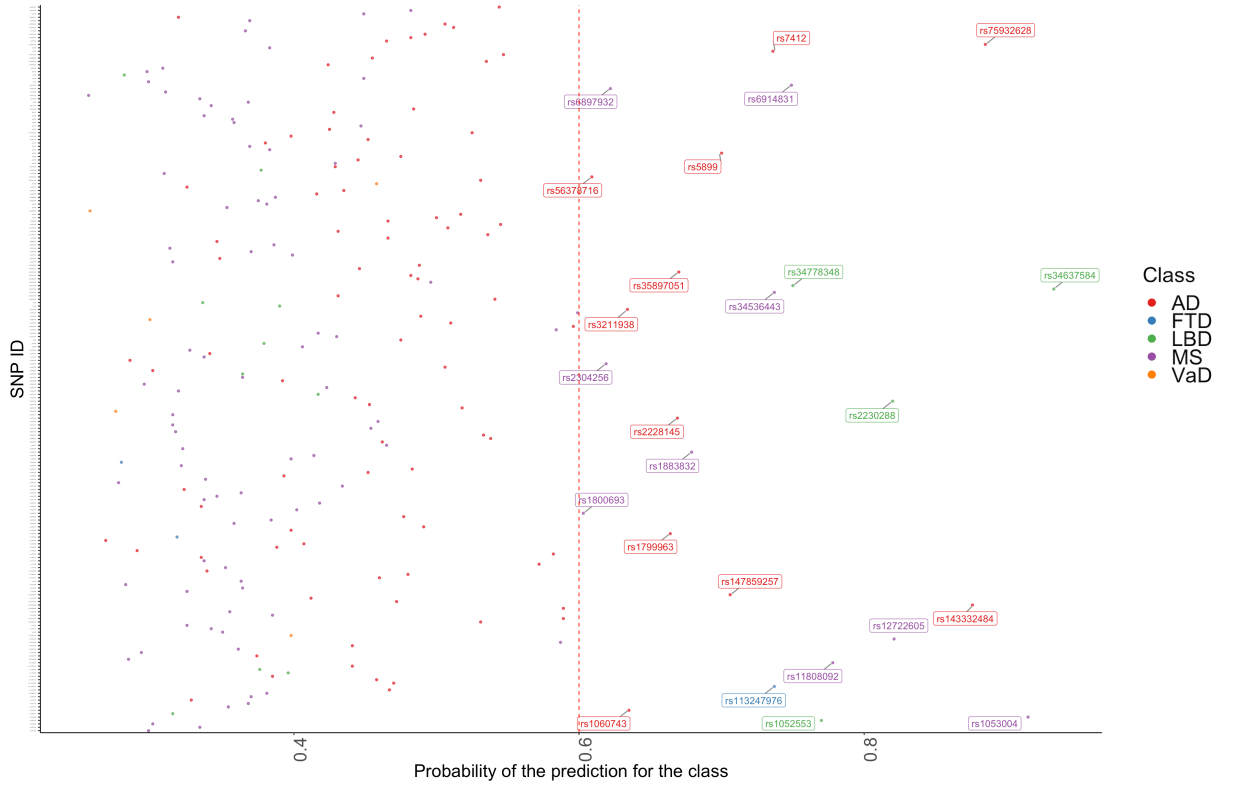
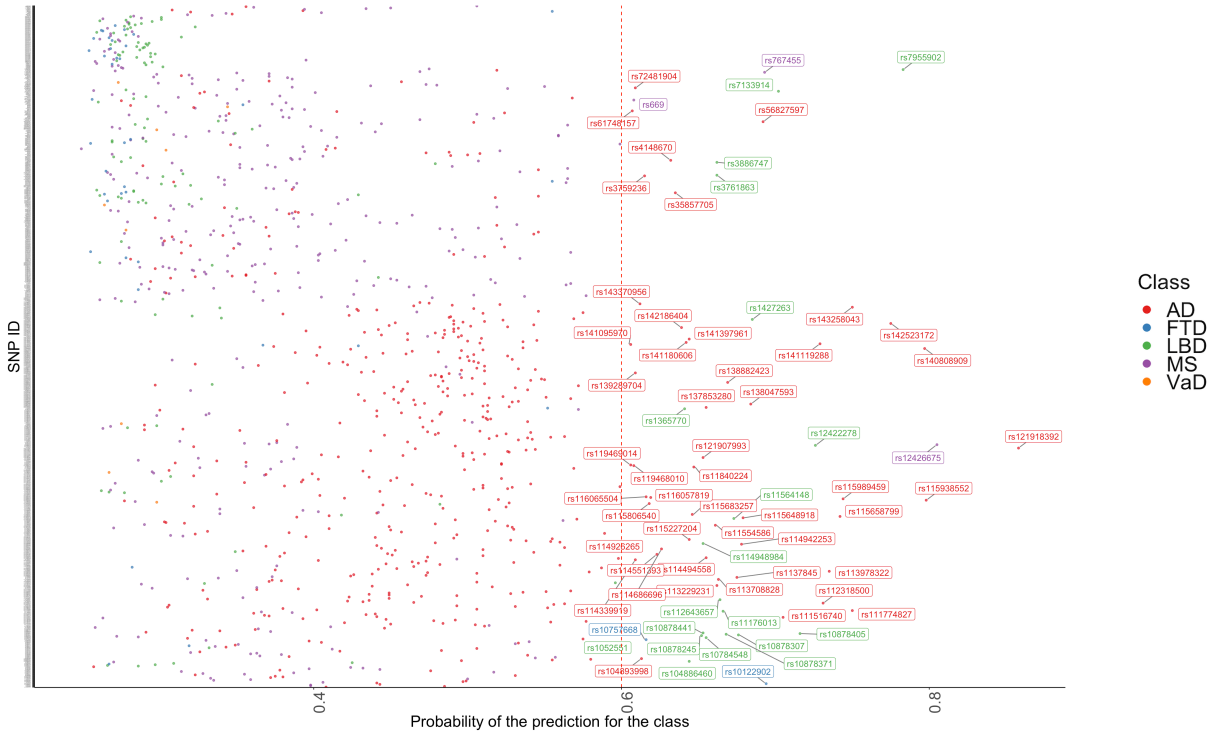


Fig. Suplementaria 1 Diagramas de flujo para búsquedas. A) MS; B) AD; C) VaD; D) LBD; E) FTD.

A)



B)



C)

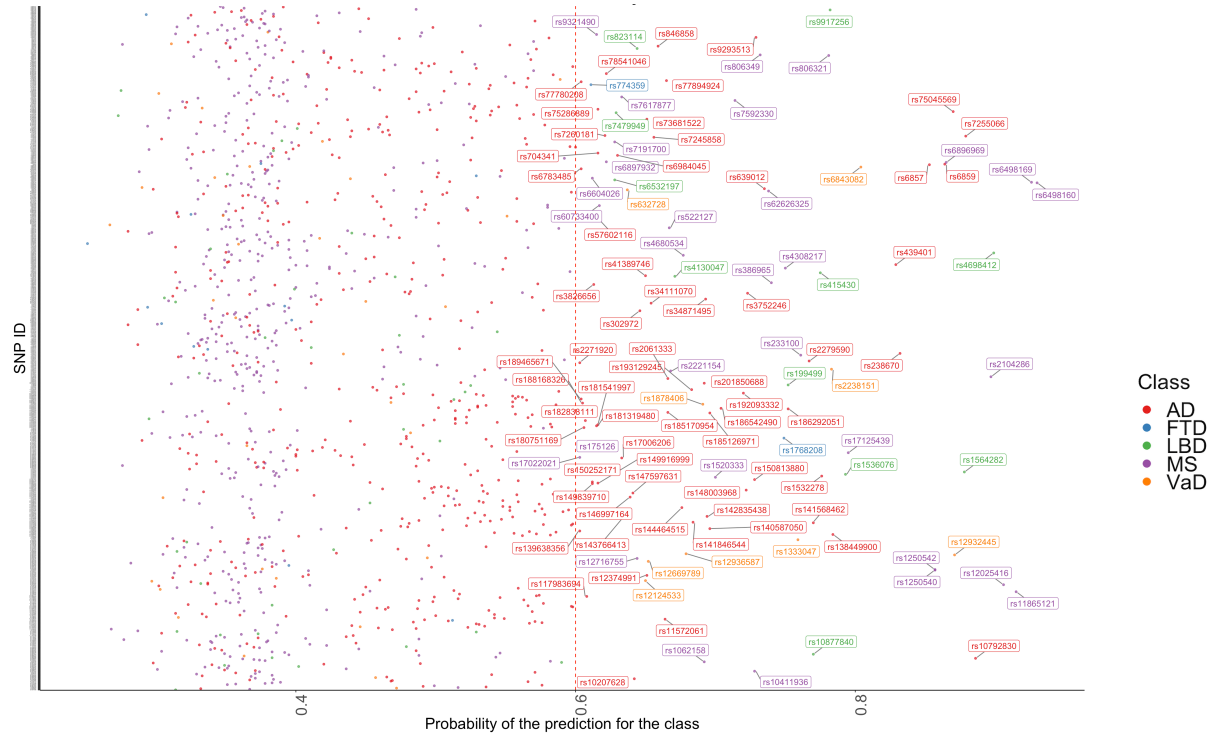


Fig. Suplementaria 2. Predicción en datos de descubrimiento. AD, Alzheimer Disease; FTD, Frontotemporal Dementia; LBD, Lewy Body Disease; MS, Multiple Sclerosis; VaD, Vascular Dementia. A) Matriz 1; B) Matriz 2; C) Matriz 3.

Tabla suplementaria 8. Resultados de entrenamiento con XGBoost con distintos conjuntos de entrenamiento.

Model	Class	Sensitivity	Specificity	Precision	F1
all	Astrocyte	0.912556053811659	0.980092204526404	0.927811550151976	0.920120572720422
all	Endothelial	0.919282511210762	0.984911986588432	0.944700460829493	0.931818181818182
all	Microglia	0.918918918918919	0.993102089081592	0.964539007092199	0.941176470588235
all	Neuron	0.965014577259475	0.98366463294627	0.922862453531598	0.943467933491686
all	Oligodendrocyte	0.903299203640501	0.975912827375263	0.86304347826087	0.882712618121178
all	Perycite	0.902040816326531	0.988078291814947	0.868369351669941	0.884884884884885
gainratio	Astrocyte	0.939461883408072	0.958507963118189	0.863917525773196	0.900107411385607
gainratio	Endothelial	0.917040358744395	0.985750209555742	0.947490347490347	0.932016710976073
gainratio	Microglia	0.91988416988417	0.992707922743398	0.962626262626263	0.940769990128332
gainratio	Neuron	0.948493683187561	0.991143475693761	0.955925563173359	0.95219512195122
gainratio	Oligodendrocyte	0.886234357224118	0.987382909577519	0.92189349112426	0.903712296983759
gainratio	Perycite	0.9	0.988790035587189	0.875	0.887323943661972
infogain	Astrocyte	0.926756352765321	0.970871751886002	0.899202320522117	0.912771439087229
infogain	Endothelial	0.917040358744395	0.984702430846605	0.943846153846154	0.930250189537528
infogain	Microglia	0.912162162162162	0.993102089081592	0.964285714285714	0.9375
infogain	Neuron	0.965014577259475	0.982483763038772	0.917744916820702	0.940786357176693
infogain	Oligodendrocyte	0.890784982935154	0.983559548843433	0.901035673187572	0.895881006864989
infogain	Perycite	0.893877551020408	0.988967971530249	0.876	0.884848484848485
deleted	Astrocyte	0.724962630792227	0.847862531433361	0.571933962264151	0.639419907712591
deleted	Endothelial	0.786248131539611	0.924140821458508	0.743988684582744	0.76453488372093
deleted	Microglia	0.683397683397683	0.966101694915254	0.804545454545455	0.739039665970772
deleted	Neuron	0.861030126336249	0.970675063963787	0.856038647342995	0.858527131782946
deleted	Oligodendrocyte	0.725824800910125	0.982794876696616	0.876373626373626	0.794026135656503
deleted	Perycite	0.538775510204082	0.983451957295374	0.739495798319328	0.623376623376623

Tabla suplementaria 9. Valores de MCC del RandomForest del modelo para relacionar SNPs con demencia.

Clase	MCC
AD	0.692
FTD	0.375
LBD	0.592
MS	0.628
VaD	0.533
Promedio	**0.613**

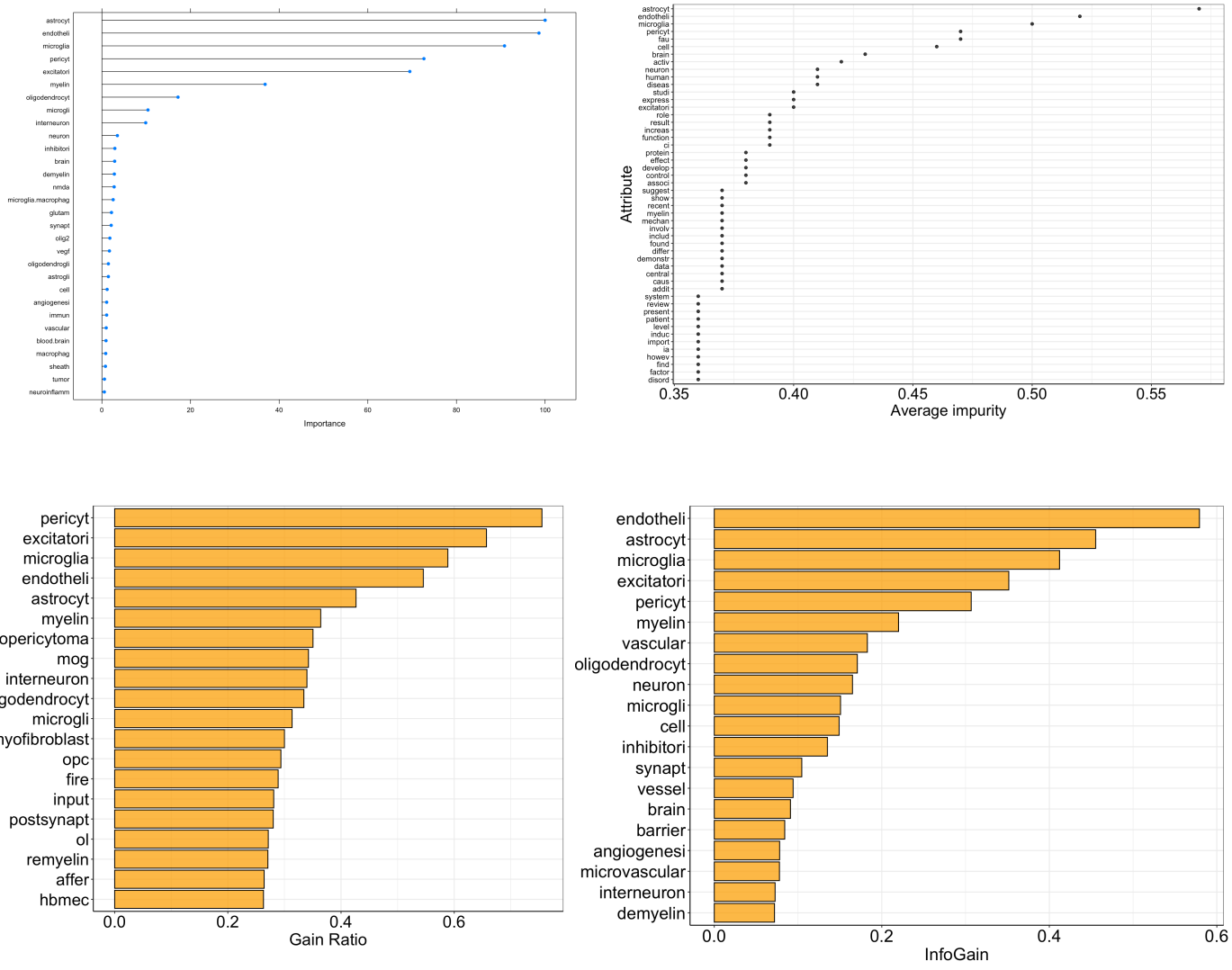


Figura suplementaria 3. Importancia de los atributos en cada conjunto de datos que se utilizó para entrenar. A. Todos los atributos

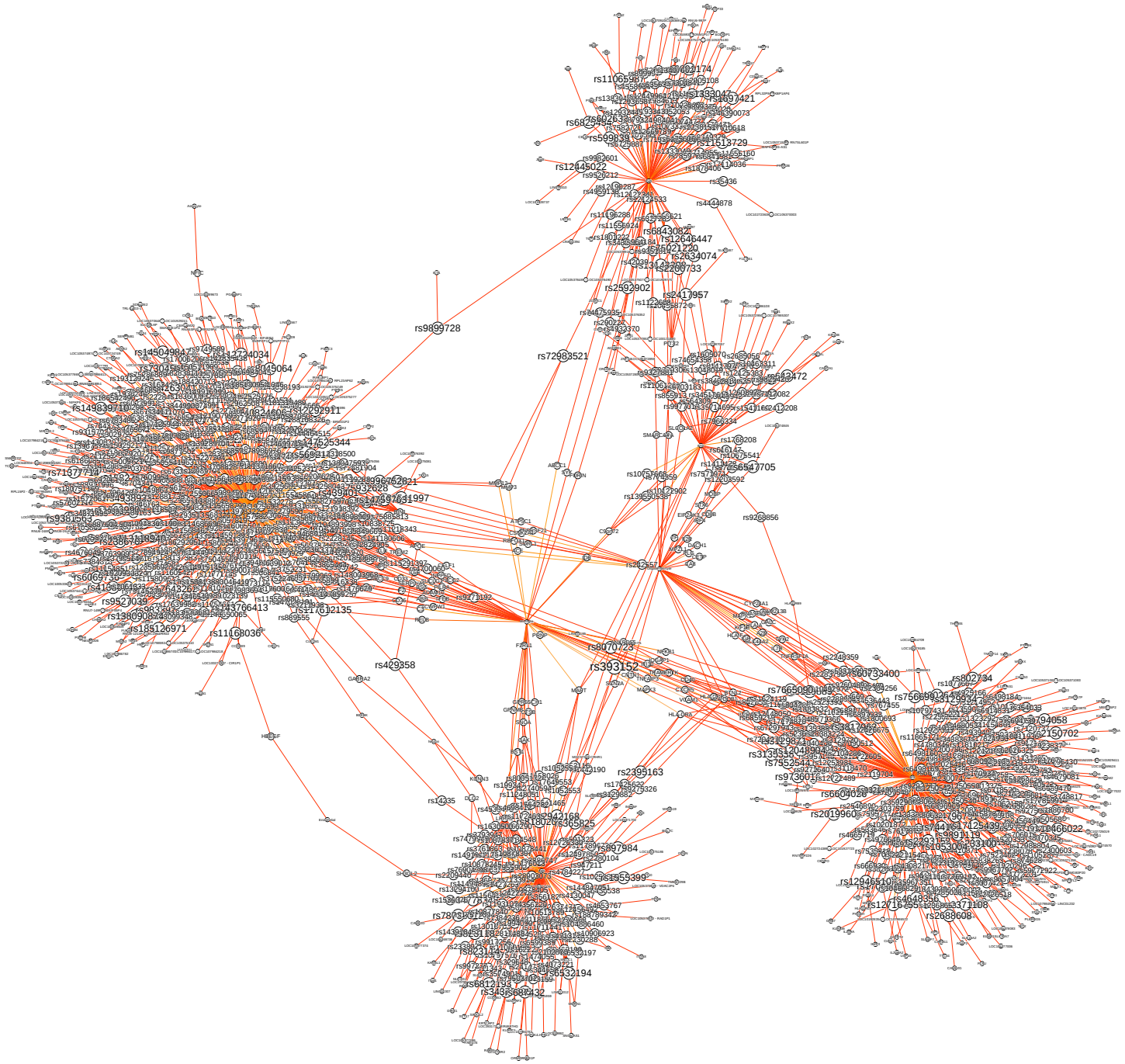


Figura suplementaria 4. Red de asociación del universo de SNPs asociados a los cinco tipos de demencia.

SOCIEDAD MEXICANA DE BIOQUIMICA A.C.

Otorga la presente

CONSTANCIA a:

Erick Cuevas Fernández

Quien asistió y presentó el trabajo:

Identification and Classification of Single Nucleotide Polymorphisms as Biomarkers associated with Dementia through Data Mining and Machine Learning

Por:
Erick Cuevas Fernández, Carlos Moncada Vázquez, Joel Sánchez Mendez, Montserrat Loredo Guillen, Heriberto Manuel Rivera

En la modalidad de presentación oral durante el
XXXII Congreso Nacional de Bioquímica
4 - 9 de noviembre de 2018, Ixtapa Zihuatanejo, Gro.

Atentamente

Dra. Irene B. Castaño Navarro
Presidente