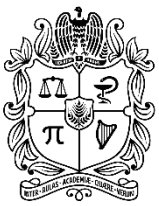


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo de epistasis basado en aprendizaje automático para pacientes con discapacidad intelectual y retraso del neurodesarrollo

Jossie Esteban Murcia Triviño

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogotá, Colombia
2024

Modelo de epistasis basado en aprendizaje automático para pacientes con discapacidad intelectual y retraso del neurodesarrollo

Jossie Esteban Murcia Triviño

Tesis presentada como requisito parcial para optar al título de:

Magister en Bioinformática

Director (a):

Luis Fernando Niño Vásquez, Ph.D.

Codirector (a):

Juan Javier López Rivera, MD. MSc.

Línea de Investigación:

Bioinformática funcional y estructural

Grupo de Investigación:

Laboratorio de investigación en sistemas inteligentes (LISI)

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá, Colombia

2024

Dedicada a mi madre, mis hermanos y mi novia.

As complexity rises, precise statements lose meaning and meaningful statements lose precision.

Lotfi A. Zadeh

Declaración de obra original

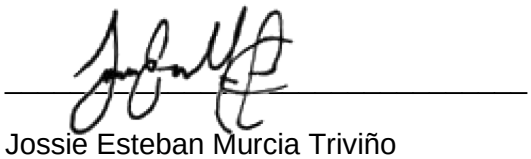
Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.



Jossie Esteban Murcia Triviño

Fecha 29/04/2024

Agradecimientos

A mis directores y mentores en el proceso, el profesor Luis Fernando Niño Vásquez, director del Grupo de Investigación LISI, y el doctor Juan Javier López Rivera, genetista y director del Laboratorio Especializado en Citogenética, quienes me apoyaron en cada una de mis distintas etapas de la investigación a través del trato humano, la transmisión de sus conocimientos y el apoyo emocional. Así mismo, a mis compañeros, tanto del grupo LISI, como del laboratorio que me brindaron importante apoyo, consejos y aportes intelectuales en el marco del desarrollo de este estudio.

A las instituciones. La Universidad Nacional de Colombia por recibirme como uno más de su familia, y brindarme todas las bases teóricas, académicas y éticas para crecer académicamente y lograr completar este proyecto. El grupo Keralty, bajo sus empresas de aseguramiento y prestación de servicio de salud, y en especial a las áreas de Investigación y Laboratorio de patología, por confiar en mi la gestión de los datos y proporcionarme el valioso acompañamiento en el correcto y transparente desarrollo de la investigación.

Resumen

Modelo de epistasis basado en aprendizaje automático para pacientes con discapacidad intelectual y retraso del neurodesarrollo

Los estudios de asociación como epistasis representan un factor importante en la comprensión de la expresión de enfermedades complejas, como lo son los trastornos del neurodesarrollo (TND), que presentan un desafío en el entendimiento de su etiología. Aunque varios estudios han revelado diferentes hallazgos de mutaciones, los efectos de asociación entre polimorfismos de un solo nucleótido (SNP) siguen siendo desconocidos. La reducción de dimensionalidad multifactorial (MDR) es un método de minería de datos por inducción constructiva empleado para detectar interacciones complejas. Este estudio comprendió una cohorte retrospectiva de pacientes pediátricos con prueba de exoma trio por sospecha de alteraciones genéticas para TND. Después de los controles de calidad sobre genotipos, se desarrolló el método MDR bajo la Prueba de desequilibrio de pedigrí (MDR-PDT). Además, se identificaron variantes asociadas individualmente con la enfermad a partir de la prueba de desequilibrio de transmisión (TDT). Se encontró que la variante rs6843524 (SEC24D) significativa por TDT (valor-P=0.003135) evidenció asociaciones con SNP; rs6843524-rs895952 (MDR-PDT valor-P=0.0084) y rs6843524-rs1168666 (MDR-PDT valor-P=0.0079). Aunque las variantes rs1168666 (SETD1B) y rs4974081 (QRICH1) no fueron significativas en MDR, si se identificaron en varios modelos y sus genes destacaron en el análisis de enriquecimiento (FDR 1.11e-05 y 6.55e-05). A pesar de la baja significancia de los modelos MDR-PDT, se lograron validar asociaciones importantes por medio de las otras pruebas y la interpretación biológica. Estos modelos pueden ser muy útiles en el descubrimiento de nuevas variantes, especialmente cuando son desarrollados sobre poblaciones grandes y con un análisis completo desde la secuenciación.

Palabras clave: Epistasis, aprendizaje de máquinas, polimorfismo de un solo nucleótido, trastornos del neurodesarrollo, discapacidad intelectual.

Abstract

Machine learning-based epistasis model for intellectual disability and neurodevelopmental delay

Association studies such as epistasis studies represent an important factor in understanding the expression of complex diseases, such as neurodevelopmental disorders (NDD). These disorders exhibit a challenge around their etiology. Even though certain studies have revealed several mutation findings, the association effects between Single Nucleotide Polymorphisms (SNPs) remain unknown. Multifactor dimensionality reduction (MDR) is a constructive induction data mining approach that can be used to identify those effects. In this work, a retrospective cohort study based on pediatric patients with trio exome analysis due to suspected genetic alterations for NDD was carried out. After developing genotype quality controls, MDR method was performed under Pedigree Imbalance Test (MDR-PDT). In addition, variants individually associated to disease were identified with Transmission Disequilibrium Test (TDT). We found that variant rs6843524 (SEC24D) is TDT significant (P-value=0.003135) and evidenced SNP interactions; rs6843524-rs895952 (MDR-PDT P-value=0.0084) and rs6843524-rs1168666 (MDR-PDT P-value=0.0079). Although variants rs1168666 (SETD1B) and rs4974081 (QRICH1) were not significant by MDR they were identified by several models and their genes were outstanding in enrichment analysis (FDR 1.11e-05 y 6.55e-05). Despite the low significance of MDR-PDT models, important associations were validated through other tests and biological interpretation. These models can be very useful in discovering new variants, especially when they are developed on larger populations and performing a complete analysis beginning from sequencing.

Keywords: Epistasis, machine learning, single nucleotide polymorphism, neurodevelopmental disorders, intellectual disability.

Esta tesis de maestría se sustentó el 22 de abril de 2024 a las 3:00 p.m., y fue evaluada por los siguientes jurados:

Andrés Mauricio Pinzón Velasco, Ph.D.
Profesor, Instituto de Genética
Universidad Nacional de Colombia - Sede Bogotá

Liliana López Kleine, Ph.D.
Profesora, Departamento de Estadística, Facultad de Ciencias
Universidad Nacional de Colombia - Sede Bogotá

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de Símbolos y abreviaturas

Abreviaturas

Abreviatura	Término
SNV	Variante de un solo nucleótido
SNP	Polimorfismo de un solo nucleótido
TND	Trastorno del neurodesarrollo
DD/ID	Retraso de Neurodesarrollo y discapacidad intelectual
MDR	Reducción de dimensionalidad multifactorial
PDT	Prueba de desequilibrio de pedigrí
TDT	Prueba de desequilibrio de transmisión
SEN	Red de epistasis estadística
FDR	Tasa de descubrimiento falso

Introducción

Los trastornos del neurodesarrollo (TND) se han descrito como las afecciones médicas crónicas con mayor prevalencia en la atención primaria pediátrica (American Psychiatric Association, 2013). Se ha detallado que la detección temprana y el tratamiento adecuado de estos trastornos pueden mejorar significativamente la calidad de vida y sintomatología de las personas afectadas. La identificación de variantes o genes causativos de los TND pueden ayudar a mejorar los diagnósticos e implementación de tratamientos que terminan por apoyar en la calidad de vida y la sintomatología de las personas afectadas. Estos trastornos, especialmente aquellos comunes como el retraso de neurodesarrollo (DD) y la discapacidad intelectual (DI), los cuales representan en conjunto un único rasgo (DD/ID), impactan en el funcionamiento cognitivo de los afectados, generando así riesgos serios en el aprendizaje y el desempeño social (Morris-Rosendahl & Crocq, 2020; Thapar et al., 2017). Por estos y otros motivos, los esfuerzos en la detección de variantes genéticas se vuelven imprescindibles para la detección temprana, tratamiento y adecuado manejo sobre poblaciones infantiles con riesgos severos en el desarrollo de comportamientos sociales, aprendizaje y de posible autolesión.

Con base en lo anterior, el desafío más remarcable en estos trastornos y en aras de detectar los diferentes riesgos, persiste el entendimiento de su etiología a partir de la determinación de las interacciones entre factores genéticos, el entorno y los estilos de vida que predisponen el apropiado riesgo (Tărlungeanu & Novarino, 2018). De esta manera, es importante descubrir aquellas interacciones entre variantes genéticas que pueden explicar la enfermedad de manera independiente a partir de interacciones genéticas, enfocándose en el estudio poblacional bajo una aproximación de estudios de interacción, desde la presencia de diferentes tipos de variaciones, como los polimorfismos de un solo nucleótido.

En el área médica, se ha venido gestando un rumbo hacia la comprensión puntual de la biología de las enfermedades humanas y su tratamiento particular, bajo un término acuñado como medicina de precisión (Khoury et al., 2012). El propósito de esta estrategia

es brindarle al paciente, bajo un modelo de atención médica exclusivo, la mejor terapia posible o esa dosis correcta de medicación para su enfermedad a partir de sus rasgos peculiares, es decir, sus características y factores genéticos y ambientales. Por lo tanto, para alcanzar dicho objetivo, es imprescindible el desarrollo de una fase de diagnóstico y, con ella, la identificación de asociaciones de variantes genéticas significativas para una enfermedad de interés. Es de esta manera que los estudios de asociación del genoma completo (GWAS, por su sigla en inglés) han permitido la identificación de aquellos genotipos que pueden explicar rasgos o enfermedades complejas. Sin embargo, para comprender mejor el espectro genético, especialmente en afecciones complejas comunes, es de igual manera importante articular los esfuerzos en la exploración de factores hereditarios alternativos que ayuden a explicar la presencia y severidad de la enfermedad, como es el caso de las interacciones genéticas.

La noción de interacciones genéticas entre un rango de lugares específicos del genoma, denominados *locus* (en plural loci), se conoce como epistasis y se evidencia en estudios de organismos modelo, a partir de la asociación entre dichos locus y un rasgo de interés, donde los genes implicados interactúan entre ellos biológicamente en diferentes niveles ómicos y de manera transversal. De manera que, si estas interacciones son importantes para organismos modelo, son también determinantes para comprender componentes hereditarios diferentes a los que se han identificado a partir de GWAS.

Históricamente, el DD/ID ha sido diagnosticado clínicamente de primera mano bajo análisis de microarreglos cromosómicos (CMA), cuando la causa es desconocida (Moeschler & Shevell, 2014). No obstante, y como se manifestó anteriormente, es elemental, descubrir aquellas interacciones entre variantes genéticas que expliquen el fenómeno de manera independiente a partir de interacciones genéticas, focalizados bajo una aproximación de estudio de interacción de epistasis a partir de polimorfismos de un solo nucleótido (SNP) o variación en el número de copias (CNV), que con ayuda de las pruebas clínicas de CMA se convertirían en la prueba de primera línea en un diagnóstico general (Srivastava et al., 2019).

Las tecnologías disruptivas como la inteligencia artificial y el aprendizaje automático permiten desarrollar estrategias e identificar patrones genéticos en las poblaciones desde la segmentación de sus comportamientos frente a estos trastornos, a partir de sus

características biológicas y estilos de vida, generando así la adecuación de intervenciones sobre el paciente. De esta manera, la identificación de variantes potenciales puede usarse en el desarrollo de dianas terapéuticas, tratamientos personalizados, diagnóstico y seguimiento temprano de pacientes con DD/ID.

Por consiguiente, contando con los servicios de atención en salud que se han prestado para este tipo de trastornos, se ha recolectado datos a nivel exómico de los pacientes y sus padres desde el laboratorio clínico. Y es a partir de este gran volumen de información ómica, que se pueden analizar de manera retrospectiva con métodos de aprendizaje automático, para determinar aquellas posibles interacciones genéticas por epistasis que pueden ayudar a comprender la etiología de DD/ID.

El resto de este documento está estructurado como se describe a continuación. En el capítulo 1 se describe el problema relacionado y la justificación para abordarlo. El capítulo 2 consiste en la descripción de los conceptos teóricos fundamentales asociados al estudio, desde la enfermedad hasta los métodos de estudios de asociación. En el capítulo 3 se resaltan los objetivos puntuales del estudio. Luego, en el capítulo 4 se precisa la metodología empleada, comprendiendo los pasos y la descripción de la población. Seguidamente, en el capítulo 5 se describe todo el proceso de análisis y cada uno de los resultados obtenidos durante el proceso. En el capítulos 6 se realiza una discusión de los resultados. Finalmente, en el capítulo 7 Conclusiones se presentan los aspectos destacados derivados del trabajo realizado, mencionando líneas de trabajo futuras.

1. Identificación del problema

1.1 Planteamiento del problema

En el desarrollo de estudios de asociación en un ámbito epistatico es fundamental que alcancen una significancia estadística relevante, a partir de considerar, al mismo tiempo, variantes comunes y raras y cohortes de estudio con heterogeneidad etiológica subyacente (Webber, 2017). De este modo, el objetivo en el descubrimiento de interacciones genotípicas busca estratificar y describir a la población para conocer aquellos factores genéticos que la influencian y la condicionan, especialmente, en contextos de Latinoamérica, ya que la mayoría de los estudios son de composición de ascendencia europea. De acuerdo con lo anterior, se identifica la necesidad de conocer las variantes genéticas que interactúan entre ellas e inciden en DD/ID, sobre la población colombiana, de modo que esta etapa de detección de variantes enriquezca la capacidad de desarrollo de estrategias de diagnóstico y tratamiento de DD/ID.

El abordaje de DD/ID desde la medicina de precisión representa un paso crucial para lograr un diagnóstico e intervención individual de cada paciente con estos trastornos (Tărlungeanu & Novarino, 2018). Uno de los retos es el manejo de los grandes y crecientes volúmenes datos ómicos generados por laboratorios y centros de diagnóstico, gracias a los avances de la secuenciación masiva de próxima generación (NGS, por sus siglas en inglés), de los reportes de expresión génica masiva e incluso de información de neuroimagenología, entre otros. Consecuentemente, es claro que existe la necesidad imperiosa de desarrollar estudios de asociación para el diagnóstico genético-molecular de pacientes con DD/ID, dada la disponibilidad de grandes volúmenes de datos, manteniendo los criterios de significancia estadística y heterogeneidad etiológica, que permitan abordar los datos genéticos de los pacientes y sus núcleos familiares desde una evaluación específica y sustancial.

1.2 Justificación

El DD/ID es un tipo de trastorno del neurodesarrollo; estos trastornos representan una de las afecciones médicas crónicas más prevalentes en la atención primaria pediátrica. La Organización Mundial de la Salud (OMS) estima que 1 de cada 160 niños sufre de alguna forma de trastorno del espectro autista (ASD) (World Health Organization, 2021), 10 de cada 1000 niños sufre de discapacidad intelectual (Cardoso et al., 2019) y, en general, aproximadamente el 17% de la población entre los 3 y los 17 años sufre algún tipo de discapacidad intelectual en Estados Unidos (Savatt & Myers, 2021). En la población latinoamericana se estima que 34 de cada 1000 pacientes pediátricos padecen este tipo de trastornos (Bitta et al., 2018).

En la última década, los estudios genéticos han permitido conocer de manera trascendental las causas y factores inmersos de DD/ID, particularmente, en virtud de la secuenciación del exoma se pueden identificar causas recesivas (Vissers et al., 2016) y, por ende, a partir de CMA y, en concordancia con los hallazgos encontrados por estudios genéticos, ha sido posible el diagnóstico molecular preventivo en pacientes con DD/ID. Evidentemente, varios genes y variantes se han relacionado contundentemente con DD/ID en un carácter aislado y de manera no sindrómica. Sin embargo, quedan en el camino muchos genotipos por descubrir, pronosticando un número superior a 1000 en la próxima década. También, es esencial su estudio a nivel de heterogeneidad genética y no sindrómica, en función de enfoques asociados a la secuenciación del genoma (Vissers et al., 2016), donde su ocurrencia esta probablemente causada por una interacción amplia e intensa entre interacciones genéticas y con el medio ambiente, de manera que varias variantes genéticas modulan la actividad a nivel proteína, en lugar de una única mutación (Morgan et al., 2015).

En virtud de establecer este tipo de estudios de interacciones genéticas epistáticas, muchos de los estudios sobre trastornos del neurodesarrollo, parten de pruebas estadísticas confiables en estos tipos de evaluaciones (X. Zhang et al., 2010). En la definición de estas pruebas, esencialmente la epistasis, se lleva a cabo un espacio de búsqueda exhaustivo cuando se necesitan asociar solamente dos variantes de la forma $PN(N-1)/2$ donde N es el número de variantes a ser estudiadas y P el número de pacientes o casos (Vanderweele, 2010; X. Zhang et al., 2010). El desarrollo de dichas

pruebas de permutación permite alcanzar dicha confiabilidad, pero también implica que el espacio de búsqueda sea extenso y su definición sería realmente $KPN(N-1)/2$ donde K es el número de permutaciones, y en consecuencia los tiempos de solución en el análisis de las interacciones serían muy amplios (X. Zhang et al., 2010).

Por tanto, el uso de estas pruebas estadísticas presenta cierta debilidad en el desarrollo eficiente de los estudios de epistasis en enfermedades comunes complejas (Cole et al., 2017; Manavalan & Priya, 2020), pero así mismo, comprende un gran dominio sobre los controles en la minimización de tasas de error, que permiten comprobar que la captura de las interacciones se mantenga bajo una estructura correlacional del genotipo, contando con el desarrollo de pruebas de permutación que distribuyan el fenotipo en la población y se cuente con cierta heterogeneidad genética (Ohi et al., 2013).

En este contexto, es imprescindible que la comunidad científica se enfoque en desarrollar algoritmos que identifiquen interacciones genéticas para DD/ID y si es factible, la incidencia del medio ambiente y los estilos de vida, con hallazgos incidentales evaluados, que permitan desarrollar estrategias terapéuticas. Así mismo, esos estudios de interacción deben estar direccionados a evaluar las pruebas de permutación o similares, desempeñándose en tasas de error bajas y concluyendo con cohortes de estudio suficientes a nivel de heterogeneidad genética. Por lo tanto, y considerando lo anterior, sería ideal hacer uso de la infraestructura informática actual, incluyendo cálculos paralelos, capturando de manera efectiva las interacciones, para el desarrollo estratégico poblacional en la identificación puntual de estas asociaciones y su tratamiento.

En consecuencia, es fundamental que el desarrollo de estudios de asociación en un ámbito epistatico alcance una significancia estadística relevante, a partir de considerar, al mismo tiempo variantes comunes y raras y cohortes de estudio con heterogeneidad etiológica subyacente (Webber, 2017). De este modo, el objetivo en el descubrimiento de interacciones genotípicas debería buscar estratificar y describir a la población para conocer aquellos factores genéticos que la influencian y la condicionan, especialmente, en contextos de Latinoamérica, ya que la mayoría de los estudios son de composición de ascendencia europea. De acuerdo con lo anterior y debido a que este tipo de estudios no se ha realizado en Colombia, se identifica la necesidad de conocer las variantes genéticas que interactúan entre ellas e inciden en DD/ID sobre la población colombiana, teniendo en

cuenta un enfoque de diseño familiar, dado el aumento de hallazgos en variantes *de novo* (Chen et al., 2018; Iossifov et al., 2014), de modo que, esta etapa de detección de variantes enriquezca la capacidad de desarrollo de estrategias de diagnóstico y dianas terapéuticas de DD/ID.

1.3 Pregunta de investigación

¿Cómo se pueden identificar las interacciones de SNPs relacionadas al retraso del desarrollo/discapacidad intelectual en un grupo de pacientes pediátricos colombianos, usando un modelo de epistasis basado en aprendizaje automático?

2. Marco teórico

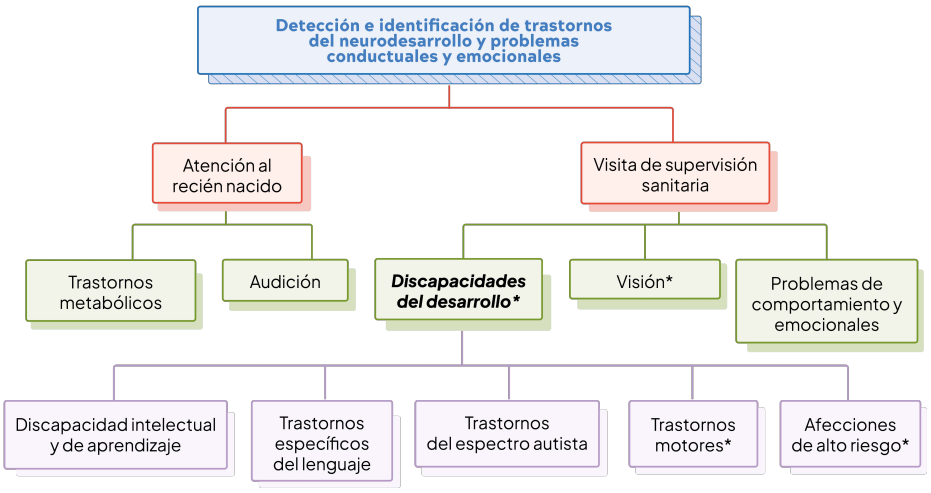
2.1 Trastornos del neurodesarrollo

Según el Manual Diagnóstico y Estadístico de los Trastornos Mentales 5ª edición (DSM-5), los TND se caracterizan por déficits de inicio temprano de gravedad variable en el funcionamiento personal, social, académico o laboral. Estos trastornos incluyen, entre otros, la discapacidad intelectual, el ASD, ADHD y los trastornos de la comunicación (American Psychiatric Association, 2013). La experiencia clínica con pacientes afectados por TND ha mostrado la existencia de cuadros clínicos intermedios y la frecuente comorbilidad entre estos trastornos enfatiza el vínculo entre ellos. Por ejemplo, en un porcentaje que oscila entre el 11% y el 65% de los pacientes con ASD existe una discapacidad intelectual asociada (Lord et al., 2018), mientras que en el 33%-37% de ellos existe un ADHD asociado (Posar & Visconti, 2017). Los TND, aun siendo entidades diagnósticas independientes, comparten manifestaciones comunes a las que presentan personas con daño cerebral o disfunción en la corteza prefrontal, es decir, presentan diferentes alteraciones de las funciones ejecutivas (Bausela-Herreras et al., 2019).

El Retraso en el Desarrollo (DD)/Discapacidad Intelectual (ID) se concibe como un TND común, cuya prevalencia se estima que varía del 1 al 3% en todo el mundo (Leonard & Wen, 2002), y está caracterizado por un cociente intelectual (CI) de 70 o menos que explica limitaciones sustanciales a nivel de funcionamiento intelectual y comportamiento adaptativo. Específicamente, la Organización Mundial de la Salud la define actualmente como "... una capacidad significativamente reducida para comprender información nueva o compleja y para aprender y aplicar nuevas habilidades (inteligencia deficiente)", dando como resultado una capacidad reducida para afrontar situaciones debido al comportamiento social deteriorado.

La ID se distingue por una gran heterogeneidad clínica y genética, donde la mayoría de las personas con la enfermedad son identificadas en edad temprana, con un diagnóstico en la infancia indicado por retrasos en el desarrollo. Así mismo, la afección puede ocurrir aisladamente o a partir de comorbilidades con malformaciones congénitas y factores neurológicos como el ASD, epilepsia y deterioros sensoriales. Asimismo, sus niveles de severidad están dados por gravedades leves, moderadas, graves y profundas y aproximadamente 60% de los casos no tienen etiología conocida (Rauch et al., 2006). La Academia de Pediatría Americana (Lipkin & Macias, 2020) propone exámenes médicos (*screening*) y clasificación de los pacientes en una etapa temprana, de acuerdo con categorías amplias que pueden ser útiles para el diagnóstico y seguimiento (**Figura 2-1**).

Figura 2-1: Pruebas de detección en la primera infancia TNDs, adaptado de (Lipkin & Macias, 2020)



Explícitamente, estos trastornos se definen como una deficiencia cuantitativa (menor a 70) en el CI en niños mayores de 5 años, tras haber sido examinados mediante un cuestionario de evaluación del CI. Se asocia a una capacidad cognitiva limitada con problemas de adaptación en la sociedad debido a deficiencias comunicativas, de empatía, en la autonomía, para el trabajo y el aprendizaje. En términos operativos se categoriza en leve (CI, 50-70), moderado (CI, 35-50) y grave (CI, 20-35). Los casos con un CI inferior a 20 son catalogados como retraso profundo (González Alvaredo et al., 2008).

2.2 Etiología de la enfermedad

Dentro de la etiología del TND severo/grave se ha encontrado que las malformaciones cromosómicas causan un alto porcentaje (4-28%) y se identifican algunos síndromes en el 3-7% de los casos (Biancotti & Benvenisty, 2011; Froukh, 2017; González Alvaredo et al., 2008; Griffin, 1996). Algunas condiciones monogénicas (de segregación mendeliana) también son causa importante y representan hasta un 9% (González Alvaredo et al., 2008). Otros defectos genéticos, como el síndrome de Zwellweger, las adrenoleucodistrofias y la aciduria glutárica tipo II, representan hasta el 17% (González Alvaredo et al., 2008). El resto de los casos de los TND severos son atribuidos a nacimientos prematuros, y otros efectos ambientales. Está claro que para una gran proporción de pacientes con TND el diagnóstico de la causa de su enfermedad permanece desconocido (González Alvaredo et al., 2008).

Dentro de la etiología del TND leve hay causas cromosómicas como la trisomía 21 o síndrome de Down que puede representar hasta el 25% de los casos (Biancotti & Benvenisty, 2011; Froukh, 2017; Griffin, 1996) y la enfermedad por expansión de tripletas nucleotídicas X frágil que representa el 7%, aunque las cifras que se conocen para este último, el síndrome de X-Fragil, no son precisas puesto que el diagnóstico es particularmente difícil y solo se logra una mejoría en la sensibilidad y especificidad de su diagnóstico cuando se usan pruebas combinadas y evaluación exhaustiva con PCR de tiempo real, Southern Blot, y evaluación de la metilación del promotor del gen (Saldarriaga et al., 2014, 2018).

Además del rol destacable de las diferentes causas genéticas, la etiología de DD/ID puede estar ocasionada por factores exógenos, como el abuso materno de alcohol o complicaciones durante el embarazo, infecciones y desnutrición, entre otros. Debido al gran número de proteínas participes en el proceso complejo del desarrollo cerebral y funcionamiento cognitivo, puede llegar a influir cualquier pequeña mutación en alguno de los nucleótidos que hacen parte de la traducción del Ácido Ribonucleico (ARN) a proteínas, ya sea un indel (inserción o deleción) o reordenamientos. De hecho, los estudios de inteligencia a nivel familia y población demuestran una alta heredabilidad, no obstante, no existe evidencia científica confiable o significativa sobre la heredabilidad de DD/ID, donde la heterogeneidad clínica de la enfermedad se refleja con una heterogeneidad genética

extrema. Asimismo, hacen falta desarrollos y estrategias de diagnóstico genético en la mayoría de los casos. Aun así, DD/DI se ha convertido en el motivo más frecuente de consulta a los servicios de genética pediátrica (Vissers et al., 2016).

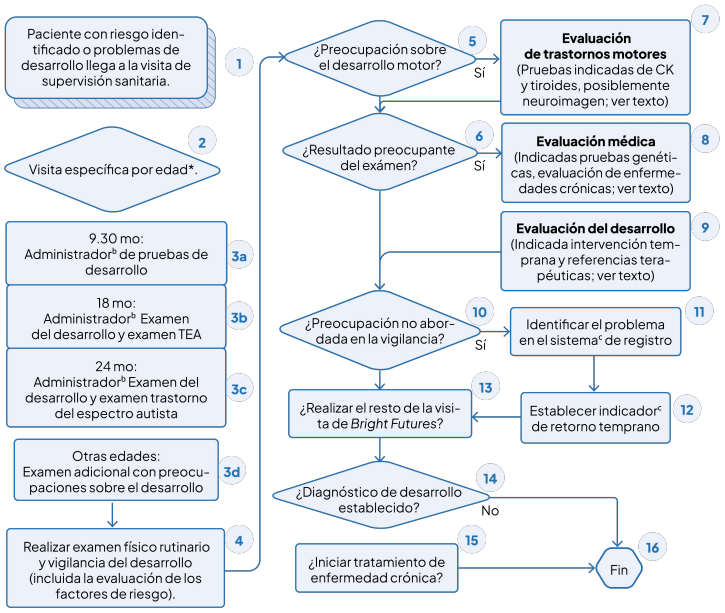
Así pues, los estudios de microarreglos genómicos han revelado una fuerte relación entre el número de genes afectados por variantes CNV y la gravedad en DD/DI. Se han detectado CNV clínicamente relevantes, que varían entre varios tamaños de bases y han llevado a la identificación de varias microdeleciones y micro duplicaciones nuevas asociadas con DD/ID (Slavotinek, 2008; Vissers & Stankiewicz, 2012), por ejemplo, que involucran a los cromosomas 1q41q42, 9q22.3, 15q13.3, 15q24, 16p11.2 y 17q21.31 (Ballif et al., 2007; Koolen et al., 2006; Redon et al., 2006; Shaffer et al., 2007; Sharp et al., 2008; Shaw-Smith et al., 2006). Además, recientemente se ha demostrado que las mutaciones de novo raras de acción dominante son una de las principales causas de DI grave y trastornos del desarrollo asociados (de Ligt et al., 2012; Hamdan et al., 2011; “Large-Scale Discovery of Novel Genetic Causes of Developmental Disorders.,” 2015). La mayoría de estos conocimientos se han adquirido gracias a estudios centrados principalmente en las formas sindrómicas de DI, mientras que la discapacidad intelectual no sindrómica, caracterizada por el deterioro cognitivo como característica clínica única, todavía está poco investigada. Comprender la genética detrás de es uno de los temas más desafiantes en el campo de los trastornos neuropsiquiátricos comunes porque sería relevante tanto para el cuidado de los pacientes como para explorar los mecanismos básicos subyacentes a la cognición y el intelecto humanos.

Con relación a su diagnóstico, la anotación en la historia clínica debe estar acompañada por la interpretación de la herencia, si los antecedentes en la familia son suficientes o lo permiten (González Alvaredo et al., 2008). Contar con información de edades maternas y paternas en el momento de la concepción, hábitos y riesgos preconcepcionales. Es importante definir si fue una concepción normal o asistida, la presencia o ausencia de movimientos fetales, comparación con otras concepciones y los resultados de los análisis ultrasonográficos, también deben anotarse la presencia de infecciones prenatales y la duración de la gestación con la mayor precisión posible (González Alvaredo et al., 2008).

De acuerdo con Posar y Visconti, queda claro que la “clasificación de los trastornos del neurodesarrollo es muy compleja y, en la actualidad, debe considerarse que se encuentra

en desarrollo, sujeta a cambios significativos, porque debe responder a diferentes necesidades, a veces divergentes" (Posar & Visconti, 2017). Existe la necesidad de un diagnóstico preciso que defina el cuadro clínico de cada individuo, en primer lugar, con fines pronósticos y terapéuticos y, en segundo lugar, con fines de investigación para encontrar una visión unificadora que pueda resaltar los aspectos comunes y los puntos de contacto de estos trastornos con fines especulativos, con el objetivo de construir nuevos modelos de interpretación del neurodesarrollo, y de inspirar nuevas hipótesis de investigación (Posar & Visconti, 2017). Se presenta un diagrama de flujo con una visión que contiene cómo abordar desde el punto de vista genético un paciente en el que se ha definido el diagnóstico de TND.

Figura 2-2: Algoritmo de atención de pacientes con TND, adaptado de (Lipkin & Macias, 2020).



2.3 Secuenciación de exoma NGS y bioinformática

Como se ha descrito anteriormente, existen diferentes pruebas genómicas que han permitido ayudar al diagnóstico de los TND, tales como: 1) cariotipo, que permite organizar y visualizar los cromosomas para realizar un conteo e identificar anormalidades visibles al microscopio, determinando el sexo cromosómico de los pacientes y si hay una cantidad de cromosomas normal; 2) hibridación genómica comparada (microarreglos) para evaluar

variantes en el número de copias, que pueden corresponder a deleciones o duplicaciones de material genético; y 3) secuenciación de exoma completo por NGS. Este último es de especial de interés, ya que teniendo en cuenta que por medio del cariotipo y la implementación de aCGH se pueden realizar evaluaciones de anomalías cromosómicas, las cuales no son aplicables para la identificación de variantes puntuales como variantes de un solo nucleótido (SNV) y variantes de inserciones-deleciones (INDEL), mientras que la secuenciación de próxima generación (NGS) puede realizar esta identificación debido a que realiza lecturas de cada uno de los nucleótidos de la secuencia del genoma (Goodwin et al., 2016).

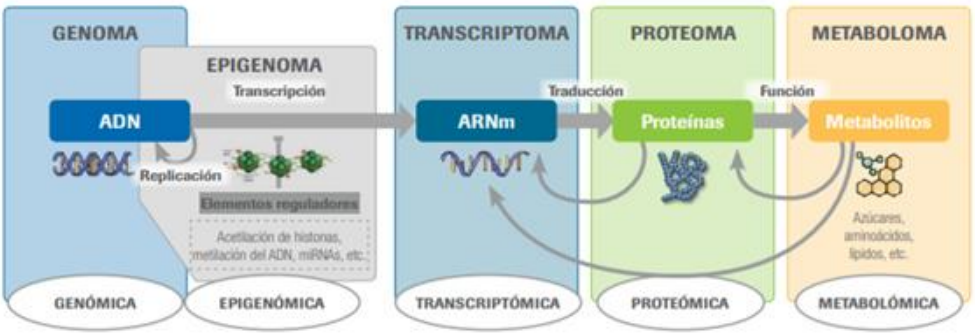
Esta técnica de NGS para exoma se usa actualmente en múltiples estrategias de análisis para el diagnóstico clínico, está implementada para la evaluación de paneles de genes, para la secuenciación de exoma, de genoma y para la secuenciación de una sola célula. Incluso con un análisis bioinformático de alto rendimiento y precisión es posible evaluar CNVs (Qin, 2019). En el ámbito clínico, la evaluación de exoma completo se usa en caso de sospecha de enfermedades de origen monogénico en el que se evalúa un aproximado de 20.000 genes a una cobertura entre un 95-98% con el fin de detectar variantes genéticas que expliquen el fenotipo del paciente (Adams & Eng, 2018; Dillon et al., 2018).

Actualmente se considera una técnica costo-efectiva y los mayores retos se relacionan con el almacenamiento y procesamiento de datos provenientes de la secuenciación, para darles significancia clínica y correlacionar los hallazgos con el fenotipo del paciente evaluado. La clasificación de variantes y CNVs está delimitada por las guías del ACMG. A partir de estas guías y el criterio PS2 que evalúa si la variante identificada es de novo, de forma que no es heredada de los padres, pero se cuenta con confirmación de paternidad, toma relevancia, no solo la evaluación de la paciente afectada, sino también la de los padres, a lo que se le conoce como exoma completo en trío (Wright et al., 2018).

En este sentido, las diferentes técnicas se desarrollan alrededor de los diferentes tipos o derivados de sistemas biológicos que hacen parte de los procesos internos de las células (Palsson, 2002). Entre las diferentes categorías de las ómicas, se encuentran la transcriptómica, genómica, proteómica, exómica, metabolómica y metagenómica, entre otras. Estas tienen como propósito el estudio de diferentes procedimientos que ocurren a diferentes niveles genéticos, o incluso la evaluación de propiedades en un nicho biológico

específico. A todo esto y muchas más características similares se les conoce como genómica funcional (Hieter & Boguski, 1997; Manzoni et al., 2018). Sin embargo, cada aproximación ómica es singular respecto a las demás, especialmente porque van dirigidas a nivel moleculares diferentes y, a pesar de esto, individualmente no son capaces de capturar toda la información relevante de su grado molecular debido a que la estructura molecular es compleja y los niveles se relacionan entre sí (ver **Figura 2-3**) (Orfao et al., 2019).

Figura 2-3: Dogma de la biología molecular y dominio de acción de las principales ciencias ómicas, tomado de (Manzoni et al., 2018).



Para el estudio y análisis de datos ómicos se concibe la biología computacional; un campo interdisciplinario que surge de la necesidad de trabajar en conjunto entre las ciencias de la computación y las ciencias biológicas. Y dentro de este la ramificación de minería de datos por bioinformática, que pretende el estudio de las estructuras ómicas en los organismos. Este campo es el interés fundamental en este proyecto, debido a los tratamientos clínicos orientados a partir de la búsqueda y anotación de genes y variantes en los pacientes, como organismos humanos (Xuan et al., 2013).

2.4 Inteligencia artificial y aprendizaje de máquina

La inteligencia artificial (IA) se ha justificado como el campo que comprende el estudio y desarrollo de agentes y máquinas con la capacidad de percibir y razonar, de manera que puedan desempeñar funciones de manera similar a las ejecutadas por el ser humano. De esta manera, la máquina puede abstraer el conocimiento mediante un espectro de

pensamiento racional, definido por (Winston, 1992) como “El estudio de computaciones que hace posible percibir, razonar y actuar”. Sin embargo, como se ha visto a lo largo de los años esta capacidad de razonamiento o singularidad ha estado limitada, donde el agente es capaz de llevar a cabo una única tarea.

Desde el auge de información y los datos en principios de los 90s y, particularmente, en 2002 y 2008 con los comienzos de los proyectos *HapMap* y *1000 Genomes,* respectivamente, asociado a la capacidad computacional de las máquinas en crecimiento, la IA ha sido capaz de probar la eficacia de sus algoritmos para la resolución de problemas específicos a partir de esta gran disponibilidad de datos estructurados o no estructurados. Dicho volumen y estructura de los datos ha permitido desarrollar nuevos algoritmos y escenarios en la IA que comprendan y razonen sobre ellos, y así realizar una tarea específica que resuelve o ayuda a resolver el problema (Dias & Torkamani, 2019).

Muchas de las tareas específicas han sido abordadas por un conjunto de algoritmos particulares de un área dentro de IA conocida como aprendizaje de máquina, la cual busca desarrollar sistemas que se ajusten a partir de un conjunto de datos. Así desde un punto de vista cognitivo, el conjunto de datos representa la experiencia que se le suministra al sistema mediante la presentación iterativa de dichos datos, de los cuales extrae modelos o patrones. En la última década, han tenido especial auge algunos de estos algoritmos más complejos acuñados bajo el término aprendizaje profundo, capaces de aprender sobre características embebidas en un gran y complejo volumen de datos (Dias & Torkamani, 2019).

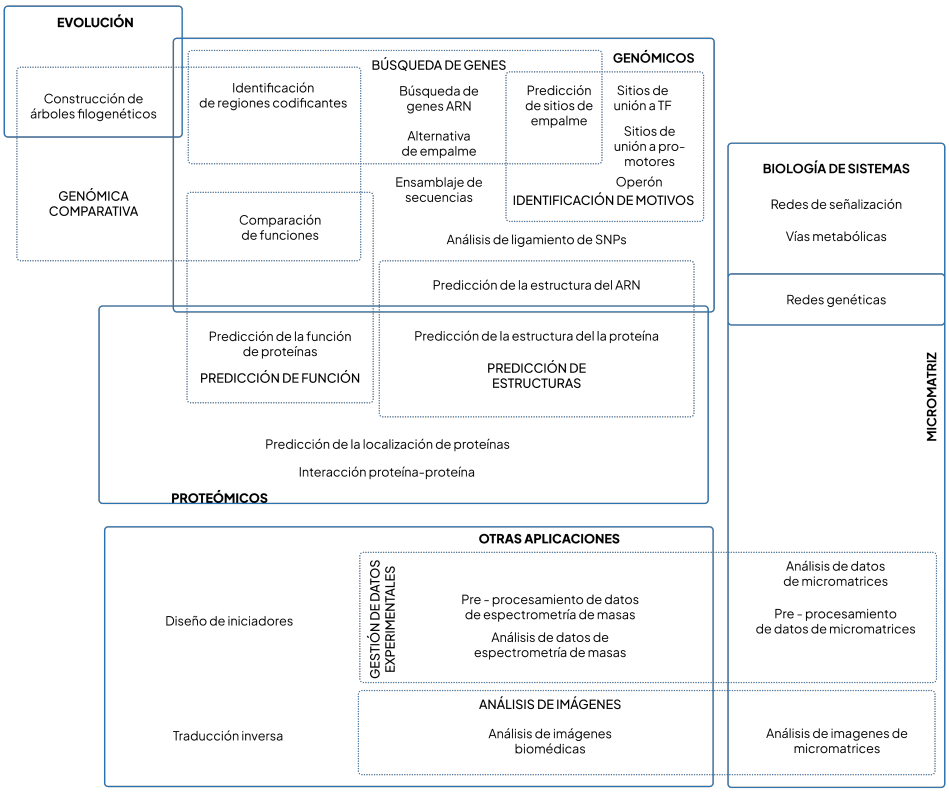
Desde la perspectiva de los datos ómicos, su amplia información provee los mayores insumos para la identificación de factores asociados a enfermedades complejas, en particular para aquellas como el retraso del neurodesarrollo y la discapacidad intelectual con una etiología multifactorial. Esta asociación se puede identificar a partir de diferentes procesos biológicos relacionados a las diferentes ómicas, tales como; expresión de genes en transcriptómica, síntesis de proteínas, variantes genómicas como CNV y SNP, entre otros, donde se puede extraer información relacionada a un factor significativo, a partir de categorías funcionales, como procesos biológicos, componentes celulares, funciones moleculares, participación en vías metabólicas, dominios de proteína, entre muchos otros.

Este tipo de análisis desde diferentes ómicas permite destacar la importancia de un factor en la patogenicidad de la enfermedad (Dias & Torkamani, 2019).

Sin embargo, y como se ha descrito anteriormente, es fundamental tener en cuenta la etiología, ya que ciertos procesos biológicos relacionados son susceptibles a determinados factores que cambian en el tiempo, como estilos de vida y variables del entorno (Perakakis et al., 2018), y es ahí, cuando el monitoreo y la integración con variables clínicas juega un papel sustancial. De esta manera, la necesidad de integración de multi ómicas y datos clínicos, así como, la comprensión de los patrones de asociación de los factores con la patología, se incluye el uso de aproximaciones de aprendizaje de máquina.

De este modo y a lo largo de los años, las diferentes categorías de algoritmos en IA se han enfocado en resolver varios problemas biológicos. Estas diferentes problemáticas se pueden clasificar en diferentes dominios ómicos (Larranaga, 2006) como: genómicas, proteómicas y sistemas biológicos, entre otros (ver **Figura 2-4**). Respecto a las genómicas y con especial interés en los factores que las caracterizan, los genes, y donde se aplican principalmente algoritmos de predicción, se busca a partir de una secuencia encontrar aquellas regiones de genes que codifican para proteínas, segmentando sus intrones y exones, donde estos últimos están enlazados a sensores de contenido para clasificar dichas regiones de ADN en codificantes y no codificantes (Mathé et al., 2002).

Figura 2-4: Clasificación de los temas donde se aplican los métodos de aprendizaje automático, adaptado de (Larranaga, 2006).



2.5 Estudios de asociación: interacciones epistáticas

Así como los GWAS, los estudios de asociación del exoma (EWAS) han exhibido un sin número de genes y variantes genéticas que explican la expresión y la compleja biología de muchos fenotipos o rasgos humanos, y han permitido una reducción de costos, ya que gran parte de los factores del genoma no son secuenciados, enfocándose en este caso, a nivel exoma. De esta manera, estos estudios permiten para ciertas enfermedades humanas interpretar su relación con ciertas asociaciones genéticas, donde habitualmente son identificados SNPs y CNVs. Durante los últimos años, el desarrollo de estos estudios ha permitido a la humanidad comprender la importancia y el impacto que tienen las alteraciones o variaciones genéticas sobre la biología y el desarrollo del cerebro humano, como trastornos o desórdenes psiquiátricos, entre los cuales se incluyen las enfermedades neurodegenerativas, discapacidades intelectuales, influenciadas por un rasgo de retraso en el neurodesarrollo. Concretamente, a partir de estos estudios se han encontrado genes o variaciones que influencian el desarrollo de los TND, como es el caso del gen NRGN en

esquizofrenia (Ohi et al., 2013), variantes como rs2060546 acerca del gen NTN4 que explican el síndrome de Tourette (Paschou et al., 2014), el desarrollado de estrategias de genómica funcional convergentes para identificar variantes familiares (estudios trío) para desórdenes del espectro autista (Carayol et al., 2014) e incluso conjuntos de genes como ARTN, PIDD1 y C2orf82 que influyen en la expresión genética del Desorden Hiperactivo y Déficit de Atención (ADHD) (Pineda-Cirera et al., 2019), para la cual también se han considerado factores de discapacidades de lectura que permitieron identificar los genes ARHGAP23 y PNC (Price et al., 2020).

A pesar del gran beneficio que han demostrado los estudios de asociación en detectar SNP, las variantes encontradas revelan relaciones leves sobre la enfermedad, explicando de una manera muy modesta la capacidad de herencia de la enfermedad (Manolio et al., 2009). Del mismo modo, las enfermedades complejas pueden ser causadas por efectos conjuntos de múltiples genes, entre los que se encuentran interacciones gen-ambiente y gen-gen, con este último explicando expresión genética y epistasis, que pueden llegar a sustentar los motivos y orígenes de la enfermedad. Así, la epistasis se entiende como la interacción gen a gen más influyente en la comprensión de comportamientos biológicos de herencia, aspecto del que carece GWAS (Moore, 2003; Van steen, 2012). Por lo tanto, vista como parte de un GWAS o incluso como un tema totalmente distinto, la epistasis, precisada como la interacción entre diferentes factores genéticos, se ha transformado en un tema imprescindible para el entendimiento de la expresión de enfermedades complejas en los últimos años, como es el caso de los TND, con el reto de detectar realmente la influencia de diferentes loci sobre el fenotipo o la enfermedad (Cordell, 2002).

Para abordar en mayor profundidad el estudio de asociación, se describen algunos términos de la genética poblacional y la estadística genómica:

Polimorfismo de un solo nucleótido (SNP): Hace alusión a la variación en un solo nucleótido (SNV) que ocurre en una posición específica en el genoma. Para que esta variación sea catalogada como un polimorfismo (SNP), la variante debe estar presente (reportada) en al menos el 1% de la población.

Llamado de genotipos: Con el fin de determinar el genotipo de cada individuo, se realiza un proceso para posiciones en las que ya se ha evaluado la presencia de SNPs u otro tipo de variantes. El proceso de llamado se refiere a la estimación de un SNP o genotipo único.

Co-heredabilidad: Se conoce como la medición de la relación genética entre varios trastornos, esto es, la proporción de covarianza entre pares de trastornos explicada por los diferentes SNP.

Cigocidad: Es la igualdad o diferencia entre las dos copias para un segmento de bases (alelos). Este grado de similitud puede ser principalmente homocigoto y heterocigoto; el primero indica la portación de dos alelos iguales, mientras el segundo señala dos alelos diferentes de un SNP específico. Esta diferencia puede estar representada en un conjunto de SNPs como una tasa de heterocigocidad de un individuo, es decir, la proporción de genotipos heterocigotos. Niveles altos de heterocigocidad pueden indicar baja calidad de la muestra del individuo.

Ausencia de genotipos: A nivel de individuo, esta métrica explica el número de SNP que le hacen falta. Los altos niveles de faltantes indican mala calidad del ADN o mala manipulación de la muestra. Así mismo, se puede evaluar la ausencia de un SNP para todos los individuos.

Frecuencia alélica menor (MAF): Frecuencia del alelo que ocurre con menos frecuencia en un sitio particular del genoma (locus) en una población específica. Un MAF bajo explica una variación genética muy rara en la población y por ende los estudios no tienen el poder suficiente para detectar asociaciones en estas.

Desequilibrio de ligamiento (LD): Medida de asociación no aleatoria entre alelos (correlación) en diferentes ubicaciones en el mismo cromosoma. Se dice que un SNP se encuentra en LD cuando la frecuencia de esta correlación entre sus alelos es mayor de lo esperado en una distribución aleatoria.

Ley de equilibrio de Hardy-Weinberg (HWE): Este concepto de genética poblacional se refiere a la relación entre las frecuencias alélicas y genotípicas. Su hipótesis establece que una población grande y sin mutaciones, el genotipo y las frecuencias alélicas son

constantes a lo largo de las generaciones. Rechazar la hipótesis HWE significa que las frecuencias del genotipo son significativamente diferentes de las esperadas. En estudios de asociación se utiliza para explicar que las desviaciones de HWE son el resultado de errores de llamado de genotipos.

2.5.1 Epistasis

En esencia, el término epistasis fue formalizado por Cordell (Cordell, 2002, 2009), esclareciendo los problemas en su concepción biológica e interpretándolo mejor como un concepto estadístico, basándose en la interpretación de Fisher (Fisher, 1919), donde se refiere a la epistasis como la desviación de la independencia de los efectos de diferentes loci genéticos, de manera que su combinación explica o no una enfermedad. Sin embargo, como él mismo lo sustenta, el concepto de independencia no siempre se establece con precisión, presentándose con múltiples percepciones e interpretaciones, particularmente entre las definiciones asumidas por biólogos, epidemiólogos, y estadísticos, donde habitualmente, su abstracción se emplea simplemente para la explicación de una interacción estadística entre genes.

Sin embargo, la palabra tenía originalmente un significado algo diferente, su noción se acuña al año 1909, donde Bateson (Bateson & Mendel, 2013) lo usó para describir casos en los que el efecto de una variante genética particular estaba enmascarado por una variante en otro locus, de manera que la variación del fenotipo o rasgo a partir de un genotipo dependía de la relación o influencia de otro genotipo en otro locus sobre el primero.

Este comportamiento de epistasis, es muy similar a las interacciones biológicas que interpreta un biólogo, donde se da una situación, en la cual, la naturaleza de un mecanismo de acción para un factor se ve afectada o alterada por la simple presencia de otro. Además de la postura de epistasis recesiva, hacen parte dentro del espectro biológico: epistasis dominante, duplicada dominante, duplicada recesiva y dominante recesiva. Dada la confusión que se ha presentado y los problemas inherentes a las definiciones biológicas de epistasis identificados (Cordell, 2002), se adecuó un concepto matemático para la identificación e interpretación de epistasis, donde el concepto genético cuantitativo de epistasis se puede representar para dos loci mediante un modelo lineal. Los métodos para

la detección de la epistasis varían según el tipo de análisis que se esté realizando; de asociación o de ligamiento, y considerando el tipo de rasgo; cuantitativo o cualitativo. De igual manera, hay que tener en cuenta que estos procedimiento estadísticos asumen que los loci han sido genotipados o, en otros términos, son las variantes etiológicas que explican el fenotipo y no están en ningún tipo de desequilibrio de ligamiento (LD, por sus siglas en inglés) con otras variantes que se relacionan con el rasgo de estudio, en caso de que dicha suposición no se cumpliera, se tendría que desarrollar un previo análisis de conocimiento y etiología, como el caso de ATHENA (Turner et al., 2010).

En su estudio, Zhang y compañía (X. Zhang et al., 2010), enmarcan los grandes retos de esta técnica. El primero hace referencia al desarrollo de pruebas estadísticas fuertes que puedan capturar de manera efectiva la interacción entre variantes o genes (especialmente SNPs), donde, como primeras aproximaciones de pruebas estadísticas en la detección de epistasis, aparte del principio empírico (Cole et al., 2017; Vanderweele, 2010), se han propuesto; la prueba de chi-cuadrado de Pearson (Chi-2), prueba exacta de Fisher (FET) y la regresión logística (LR) (Manavalan & Priya, 2020). El otro desafío importante en la técnica de asociación entre variantes es la intensa carga computacional impuesta por el enorme espacio de búsqueda. Esto hace referencia a que es computacionalmente inviable medir todas las combinaciones de SNP, por ejemplo, el espacio de búsqueda exhaustivo y medio entre apenas dos variantes es $\big(PN(N-1)\big)/2$, donde N es el número de SNPs y P el número de pacientes o casos en el estudio. Es importante tener en cuenta que, para mantener una estructura correlacional del genotipo, se debe contar con una prueba de permutaciones que distribuyan el fenotipo en la población de estudio, es decir, validar pruebas múltiples, y este caso al hacer K permutaciones, el espacio de búsqueda se incrementaría a $(KPN(N-1))/2$. Por último, el tercer reto mencionado, hace referencia a un dominio sobre errores en estos estudios, el cual puede ser validado a partir de un control que minimice la tasa de error familiar (FWER) y la tasa de descubrimiento falso (FDR).

2.5.2 Métodos de detección de epistasis

Desde la invención del término epistasis, se han venido desarrollando métodos y técnicas cada vez más capaces que detecten las interacciones entre genes y SNPs. La **Tabla 2-1** muestra varios de los algoritmos más notables y de mejor desempeño, particularmente para aproximaciones de caso control.

Tabla 2-1: Métodos prometedores para la detección de interacciones epistáticas.

Algoritmo	Caso de uso	Resultados	Autor
EACO: Ant Colony Optimization	Simulación: degeneración macular	EACO se comparó contra BOOST, SNPRuler y epiMODE. Se indican resultados donde EACO es prometedor en la identificación de epistasis. La complejidad temporal de EACO es O(LJ+nm2), complejidad mayor a otros algoritmos de Colonia de Hormigas [20]	(Sun et al., 2018) Sun et al.
Random Forest	Alzheimer	Se demostró que este enfoque de aprendizaje automático podría usarse para descubrir interacciones gen-gen significativos a partir de conjuntos de datos de expresión genética heterogéneos. En comparación con otros métodos como SVM, ANN y PART, el algoritmo presenta unos mejores desempeños con un AUC de 0.87, incluso mejor que AdaBoost [22]	(C. Park et al., 2018) Park et al.
GCORE: correlaciones	Autismo, familias tríos	La principal ventaja del GCORE sobre otras pruebas es que se puede utilizar para realizar un estudio de interacción de todo el genoma sin recursos informáticos a gran escala. Sin embargo, su potencia puede ser significativamente menor que MDR en algunas situaciones. Presenta un P-value de 3.8x10-7	(Sung et al., 2016) Sung et al
Deep Learning	Simulación: GAMETES	Este enfoque es más robusto y estable frente a diferentes formas de ruido en comparación con los enfoques MDRAC y MDR [25]. Además, se reconoce que los efectos del ruido de fenocopia y heterogeneidad tienen mayor impacto en la precisión. Se reporta una exactitud (accuracy) de 0.94	(Ghanem et al., 2019) Ghanem et al.
EF-MDR: MDR Difuso Empírico	Simulación: Enfermedad de Crohn y trastorno bipolar	Se demostró que EF-MDR, tiene mayor potencia respecto a Fuzzy MDR y MDR en varios modelos de simulación. Con datos reales, EF-MDR demostró su capacidad de proporcionar una interpretación más flexible de interacciones biológicamente significativas.	(Leem & Park, 2017) Leem et al.
Permutation Random Forest	Cáncer de vejiga	Esta metodología logró altas tasas de éxito para la detección de SNPs interactivos para un conjunto de datos simulados por GAMETES. Se obtuvo un error promedio 5E de 7;23 %	(J. Li et al., 2016) Li et al.
Regresión logística funcional	Simulación: arteriopatía coronaria	Se encontraron variantes raras que se comprimieron en unos pocos componentes principales funcionales. El valor P para probar la interacción entre TMX4 y C20orf7 fue muy pequeño (1.09x10-19).	(Zhao et al., 2016) Zhao et al.
CRC: Red asociativa	Cáncer colorrectal	La construcción de esta red constó de un número significativo de SNP y se utilizaron varias propiedades de la red para resaltar algunos SNP clave con una posible asociación con la enfermedad. P-value de 0.001.	(Kafaie et al., 2019) Kafaie et al.
GenEpi: Regresión regularizada L1	Alzheimer	Los resultados de los datos de simulación y de Alzheimer demostraron que GenEpi tiene la capacidad de detectar la epistasis asociada con los fenotipos de manera eficaz y eficiente. Se reporta un AUC de 0.85.	(Y. C. Chang et al., 2020) Chang et al.
iLOCi: priorización de interacción	Simulación: WTCCC	El modelo demuestra que es computacionalmente eficiente y adecuado para una búsqueda exhaustiva de interacciones a lo largo de marcadores en un conjunto	(Piriyapongsa et al., 2012)

			de datos GWAS típico. Se obtienen una serie de p-values que varían de 2.22x10-16 a 1.14x10-7 en SNP.	Piriyapongsa et al.
	AGGrEGATOr: min-P	Artritis reumatoide	El método logra que el control de la tasa de falsos positivos sea robusto a la presencia de efectos marginales y diferentes patrones de LD dentro y entre los dos genes investigados.	(Emily, 2016) Emily et al.

De acuerdo con Cole y colaboradores (Cole et al., 2017) "… la epistasis golpea la esencia de la inmensa complejidad e interconexión de los sistemas biológicos y, por lo tanto, ha estimulado el desarrollo de métodos estadísticos y de computación. Enfoques y técnicas, que probablemente serán aplicables a otras áreas en las que las interacciones son importantes …". En concordancia, ha sido un área que se ha complementado mutuamente con otras, como la minería de datos, la inteligencia artificial y el aprendizaje automático.

En sus comienzos varios análisis estadísticos paramétricos en estos estudios de asociación, tanto para individuos no relacionados, como familiares. Especialmente, se dio origen al concepto con métodos de regresión lineal multivariable y generalizada, incluso añadiendo modelos de efectos mixtos para agregar efectos aleatorios como la relación heterogénea. En este sentido los análisis de regresión de epistasis se basan en el modelo lineal motivado por Fisher (Fisher, 1919), donde expresa un fenotipo como una función lineal de la combinación de genotipos.

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_1 G_1 + \in \tag{2.1}$$

Siendo así, los modelo lineales proporcionan ventajas importantes para el descubrimiento de interacciones entre genotipos, incluso permitiendo agregar covariables. Sin embargo, la estimación de los coeficientes (β) para las interacciones se dificulta por la falta de observaciones en la población, especialmente cuando varios genotipos son poco comunes o raros y se evalúa su presencia simultánea (Moore et al., 2004). De esta manera, para resolver la necesidad de grandes tamaños de muestra y el reto de estimación de la interacción entre n genotipos, se desarrollaron y adaptaron otros modelos.

Tal es el caso de la relación con el aprendizaje automático, que diferencia de los métodos estadísticos paramétricos, ofrecen varias ventajas clave para problemas de alta dimensión (Martin et al., 2006). Las estrategias de selección de características pueden reducir el

espacio de búsqueda incorporando técnicas de algoritmos evolutivos, como EACO por optimización (Sun et al., 2018). Métodos de clasificación y regresión no paramétricos, incluidos los métodos basados en árboles (C. Park et al., 2018), o en reglas, pueden generar predicciones y extrayendo la cuantificación de la importancia relativa de combinaciones de genotipos. Asimismo, los enfoques de aprendizaje automático basados en reducción de dimensionalidad han reportado grandes desempeños con respecto a la detección de epistasis, y su gran precursor es MDR, quien ha evolucionado a través del tiempo en diferentes variaciones como EF-MDR (Leem & Park, 2017). Este modelo se analiza más adelante, dada su relevancia en estudios familiares.

Por otro lado, el modelo de optimización multiobjetivo de Yang (Yang et al., 2017) indica interacciones significativas y fuertes entre SNP, a partir de un valor-P < 0.0001, sin embargo, no es muy claro respecto a la tasa de éxito de detección comparativa. Un estudio del desempeño de LASSO con penalización sobre datos sintéticos por Zhou (Zhou et al., 2014) presenta de manera didáctica en múltiples pasos, un procedimiento eficaz para satisfacer dos efectos en la selección de características simultáneamente, generalización y escasez, pero no presenta resultados de desempeño contundentes, con la comparación respecto a el método Screen and Clean (SC).

En cuanto al uso de métodos específicos en los últimos años, se demostró que a partir de un aprendizaje global de Markov estocástico, optimizado a partir de un algoritmo de colonias de hormigas (ACO), se mejora el tiempo de complejidad temporal, y desempeños ligeramente mejores respecto a otros algoritmos de Markov (Sinoquet & Niel, 2019). También es el caso de la propuesta de Ansarifar (Ansarifar et al., 2019), con un algoritmo heurístico con uso de funciones de error (RMSE) y escasez para epistasis de orden alto [50], donde se mejoran rendimiento de complejidad, con errores bajos similares al algoritmo de Ground Truth (3%). Como aproximación especial, Park (M. Park et al., 2020) implementa un modelo Kaplan Meir con MDR para un caso de sobrevida, aplicado a cáncer de ovario, al ser un modelo simple, requiere menos cálculo, y tiene gran ventaja para tratar con datos biológicos de alta dimensión.

Entre los principales desarrollos de epistasis en esta área, se destacan en gran medida los asociados a trastornos de autismo, esquizofrenia o desordenes psicóticos y ADHD. En cuanto al autismo, uno de los trabajos más notables es la aproximación de una red

neuronal en términos de aprendizaje profundo denominada PANDA (Y. Zhang et al., 2020), que tiene por objeto, identificar aquellas interacciones gen a gen, para un conjunto de genes asociados a autismo, la característica significativa de esta contribución, recae en el uso etiológico de una red ontológica que expresa las interacciones moleculares humanas, para asociar los diferentes genes de autismo, entre ellos mismos, obteniendo buenos resultados con una exactitud de 89%. Sung y compañía (Sung et al., 2016) como se ha nombrado anteriormente, demuestra que a partir de una prueba estadística especial (como GCORE) se puede lograr un estudio de interacción de todo el genoma sin recursos superiores como los usados en PANDA, pero su potencia se puede ver disminuida respecto a otras aproximaciones.

Desde los inicios en el desarrollo de este tipo de estudios de asociación en epistasis y como estrategia que con los años ha evolucionado y se ha mantenido como la más usada; es la aproximación no paramétrica denominada Reducción de Dimensionalidad Multifactorial (MDR). El método MDR fue el primer enfoque de aprendizaje automático construido específicamente para identificar e interpretar interacciones no aditivas entre variantes o genes en ausencia de efectos independiente (M D Ritchie et al., 2001). En el caso de estudios familiares, también denominados trio o genealógicos (pedigrí), este es el único método extendido y adaptado que ha demostrado un buen desempeño detectando la relación entre los genotipos trasmitidos en la familia. Actualmente, el estado del arte evidencia tres métodos extendidos para este propósito; MDR Generalizado basado en Pedigrí (PG-MDR) (Lou et al., 2008), FAMily Multifactor Dimensionality Reduction (FAM-MDR) (Cattaert et al., 2010) y Prueba de Desequilibrio de Pedigrí MDR (MDR-PDT) (Martin et al., 2006). Este último es de especial interés para este trabajo debido a su desempeño respecto a los demás, su actualización durante los años y su desarrollo llevado a cabo por parte de los mismos autores al método MDR.

2.5.3 Enfoques para control de calidad de SNPs

Un factor importante en el desarrollo de estudios de asociación, sean sobre epistasis o GWAS, es la aplicación de controles de calidad adecuados, ya que, sin este tipo de pruebas meticulosas, los estudios no demostrarán resultados confiables al analizar datos o genotipos crudos con errores. Estos errores en los genotipos pueden surgir por varias

razones: mala calidad de las muestras de ADN, una mala hibridación, un rendimiento deficiente de las sondas y por supuesto, mezclas o contaminación de las muestras.

En el flujo de trabajo o pipeline habitual para el procesamiento de muestras por NGS para estudios de variación genética humana, uno de los procesos finales es el análisis e identificación de variantes asociadas con un rasgo o población específica. Este proceso es denominado como llamado de variantes, con el objetivo de identificar variantes a partir de datos de secuencia alineados, esto es, determinar dónde difieren las lecturas alineadas respecto al genoma de referencia y reportar estas variantes en un Archivo de Llamado de Variantes (VCF). A partir de los VCF de la población se debe relacionar cada una de estas variantes con lo reportado por la comunidad clínica y científica, a este proceso se le llama anotación.

Una de las herramientas más usadas es ANNOVAR, una herramienta rápida y eficaz para anotar las características funcionales de la variación genética a partir de datos de secuenciación, proporcionando ventajas de reducción de variantes que ayudan a identificar subconjuntos específicos de variantes con mayor probabilidad de ser causales de la enfermedad o fenotipo (Wang et al., 2010). ANNOVAR[1] ofrece tres tipos de anotaciones: anotación basada en genes, anotación basada en regiones y anotación basada en filtros, y se utiliza ampliamente en estudios clínicos; por ejemplo, el desafío CLARITY reporta que el 63 % de los finalistas utilizaron ANNOVAR para realizar hallazgos de mutaciones asociadas a enfermedades (Brownstein et al., 2014).

Una vez anotadas las variantes y teniendo en cuenta los desafíos de los estudios de asociación, se señalan a continuación los criterios de control de calidad con sus recomendaciones para estudios de asociación, según Andries Marees y compañía (Marees et al., 2018).

Filtro de SNPs faltantes: Excluir los SNP que faltan en una gran proporción de los individuos. Así mismo, se debe excluir los individuos con altas tasas de ausencia de

[1] ANNOVAR puede accederse en http://annovar.openbioinformatics.org/. Es una herramienta de línea de comandos escrita en el lenguaje de programación Perl.

genotipos. De esta manera, se requiere eliminar los SNP e individuos con llamadas bajas de genotipo. Se suele recomendar filtrar primero los SNP y los individuos en función de un límite moderado (>20% de ausencia).

Filtro de MAF: Evaluar e incluir solo aquellos SNP por encima del umbral MAF establecido. SNPs con un MAF bajo son raros, y, por lo tanto, es muy complicado detectar sus asociaciones con la enfermedad. El umbral MAF debe depender del tamaño de su muestra, por ejemplo, para muestras grandes se pueden usar umbrales MAF más bajos y 0,05 es recomendado muestras intermedias.

Filtro Equilibrio Hardy–Weinberg: Este filtro excluye aquellos SNPs que presentan un desvío del equilibrio de Hardy-Weinberg. Se recomienda un umbral de p-valor de la prueba HWE menor que $1e^{-10}$ en los casos y menor que $1e^{-6}$ en los controles, de esta manera, se evita descartar los SNPs asociados a enfermedades en riesgo.

Filtro de Heterocigocidad: Excluir aquellos individuos con tasas de heterocigocidad altas o bajas, con umbrales de $\mu \pm 3\sigma$ sobre las muestras. Un exceso de heterocigocidad indica contaminación de las muestras. Habitualmente el llamado de genotipos y la heterocigocidad se visualizan y evalúan juntos.

Filtro de Poda por LD: Seleccionar un subconjunto de variantes que se encuentran en LD. De esta manera, se evalúa la fuerza de LD por regiones, seleccionando los SNPs que no están aproximadamente correlacionados, a partir de un umbral de LD. En el caso de los estudios de epistasis se ha evidenciado que puede ser contraproducente, ya que excluye la detección de sitios en LD que pueden subyacer a la arquitectura genética de rasgos complejos.

2.6 Métodos de asociación

2.6.1 MDR-PDT

El modelo MDR-PDT es una adaptación del clásico MDR como medida para estudios familiares de asociación entre el genotipo y el estado de la enfermedad. La aproximación está basada en el estadístico de Prueba de Desequilibrio de Pedigrí en genotipos (geno-

PDT) (Martin et al., 2003). La esencia del modelo original MDR radica en un algoritmo de inducción constructiva de características, el cual crea una nueva variable o atributo agrupando, por ejemplo, genotipos de múltiples SNP. De esta manera, realiza una búsqueda exhaustiva por clasificación para obtener una combinatoria óptima de los SNPs que predicen el riesgo de la enfermedad. El cambio principal del modelo MDR-PDT consiste en evaluar la transmisión de los genotipos entre la familia, en lugar de frecuencias entre casos y controles, esto a partir del estadístico de Pedigrí.

Para este estadístico, se establece que dos tipos de familias aportarán información de la asociación genotipo-fenotipo: Núcleos Familiares Informativos (INF) (trio; hijo afectado y ambos padres son genotipados) y Hermanos Discordantes Informativos (DSP) (hermano afectado y hermano no afectado). Para un locus con dos alelos, α1 y α2, considere sus diferentes genotipos. Para el genotipo $\{\alpha_i \alpha_j\} = g$ se definen dos variables aleatorias para INF y DSP. En el caso de INF, se estiman el par de alelos que se han transmitido y el par de alelos que no se ha transmitido, denotado como $X_T(g) = (\#\ veces\ que\ g\ es\ trasmitido) - (\#\ veces\ que\ g\ no\ es\ trasmitido)$. Para DSP, se observan los genotipos trasmitidos a cada hijo, donde $X_S(g) = (\#\ veces\ que\ g\ ocurre\ en\ el\ afectado) - (\#\ veces\ que\ g\ ocurre\ en\ el\ no\ afectado)$. Para un Pedigrí que contiene INF nT de familias nucleares informativas y DSP nS de hermanos discordantes informativos, se define la variable aleatoria:

$$\mathrm{D(g)} = \left[\sum_{j=1}^{n_T} X_{Tj}(g) + \sum_{j=1}^{n_S} X_{Sj}(g) \right] \tag{2.2}$$

Si N es el número total de Pedigrís informativos no relacionados en la muestra y Di(g) es la variable aleatoria para el i-ésimo Pedigrí, entonces la estadística de genotipo-PDT para el genotipo g distribuido en las familias no relacionadas es:

$$\mathrm{T}(g) = \frac{\sum_{i=1}^{N} D_i(g)}{\left\lfloor \sqrt{\sum_{i=1}^{N} D_i(g)^2} \right\rfloor} \tag{2.3}$$

En este sentido, el modelo MDR-PDT involucra varios pasos y para facilitar la comprensión se describe para combinatorias de 2 locus: 1) Se seleccionan un par de SNPs de un

conjunto de η SNPs; 2) Los genotipos posibles se representan en un espacio bidimensional donde cada combinación o multifactor se representa en cada celda, y para cada uno de ellos se calcula el estadístico geno-PDT; 3) Para cada multifactor/celda (combinación de genotipos de SNPs) se etiqueta como alto riesgo o bajo riesgo, clasificándolo a partir del umbral τ, de manera que es de alto riesgo si su estadística PDT es mayor; $geno - PDT > \tau$ y de bajo riesgo en caso contrario; 4) Los genotipos de alto y bajo riesgo se agrupan en dos clases genotípicas, al poder representar una clase (alto riesgo) como el valor negativo de la otra, se reduce el problema a calcular geno-PDT para la clase de genotipo de alto riesgo, a este nuevo estadístico basado en genotipos agrupados de alto riesgo, se le denomina MDR-PDT. Este estadístico proporciona una medida de la asociación genotipo-enfermedad para el par de SNPs elegidos; y 5) Este procedimiento se repite para cada combinatoria de SNPs y se selecciona el modelo que maximiza MDR-PDT.

Así como otros métodos de aprendizaje automático, MDR generalmente se implementa usando validación cruzada (CV) para evaluar la precisión predictiva de los modelos, permitiendo evaluar la generalización de los modelos MDR a partir de dividir el grupo de SNPs en varios subconjuntos divididos para entrenamiento y validación. La significancia estadística se evalúa mediante pruebas de permutación (Pattin et al., 2009), donde realizarlas dentro de las familias asegura que se mantengan las correlaciones. De manera que repetir el algoritmo MDR-PDT en todos los conjuntos de k combinatorias de SNPs proporciona un valor máximo del estadístico MDR-PDT aproximándolo a la distribución de hipótesis nula, y observando el valor p de la permutación.

2.6.2 TDT

Adicional al desarrollo de métodos eficientes como MDR para investigar epistasis, el creciente conjunto de datos moleculares funcionales ha permitido la formulación de hipótesis que evalúan la significancia de variantes genéticas frente al fenotipo o enfermedad, siendo esencial desarrollar estrategias estadísticas para probar los efectos de transmisión en alelos específicos, y más aún en estudios trio o familiares. Es de esta manera como la prueba de desequilibrio de transmisión (TDT) (Spielman et al., 1993) toma importancia, debido a que identifica alteraciones y evalúa la frecuencia con la que un alelo o su alternativo se transmite a la descendencia afectada. De esta manera, la prueba considera para un locus de enfermedad *A*, la presencia del alelo A_1 de la enfermedad y un

alelo A_2 normal, y para un locus marcador los alelos codominantes (cada uno expresa un fenotipo) M_1 y M_2. En el caso particular de familias únicamente trio, y por lo tanto el más simple, se cuenta con una muestra de n familias. Para el locus marcador *M* habrá un total de *4n* alelos parentales; donde *2n* se transmiten. De esta manera, la trasmisión de los alelos M_1 y M_2 en los hijos afectados se pueden representar, siendo, *w* y *y* las frecuencias del alelo M_1 trasmitido y no trasmitido, respectivamente, y en el caso de M_2 como la frecuencia restante, *2n-w* para trasmitido y *2n-y* para no trasmitido. Sobre estas frecuencias representadas como una tabla 2x2 se puede utilizar la prueba X^2 habitual de significancia, llevando a cabo el estadístico estándar para una tabla de contingencia con un grado de libertad:

$$\frac{4n(w-y)^2}{(w+y)(4n-w-y)} \tag{2.4}$$

En este sentido, se establece y demuestra (Spielman et al., 1993) la hipótesis sobre el estadístico, donde estipula que: no hay asociación entre el marcador y la enfermedad/fenotipo ($\delta = 0$). A partir de esta prueba y en validación de la hipótesis de no ligamiento ($1-2\theta = 0$, utilizando datos de padres heterocigotos), determinaron que el estadístico de prueba X_2 es la aproximación estándar a una prueba binomial de igualdad de dos proporciones, a partir de la tabla de contingencia de combinatorias de marcadores transmitidos y no transmitidos.

Tabla 2-2: Combinaciones alelos marcadores M_1 y M_2 transmitidos y no transmitidos

	Alelos no transmitidos		
Alelos transmitidos	M_1	M_2	**Total**
M_1	A	b	a+b
M_2	C	d	c+d
Total	a+c	b+d	2n

Los autores denominaron al estadístico X^2 "transmisión/desequilibrio X^2" o TDT, denotado como X^2_{td}, y de esta manera se usa para evaluar la asociación entre los loci *A* y *M*. Cuando la hipótesis es verdadera con media *0* y varianza (*b* + *c*), se le da la definición al estadístico:

$$X^2_{td} = \frac{(b-c)^2}{b+c} \tag{2.5}$$

2.7 Métodos de interpretación de asociaciones

2.7.1 Redes de epistasis estadísticas

Cuando se conocen las asociaciones entre variantes es importante también comprender las bases genéticas de la susceptibilidad y la etiología de las enfermedades. Adicional a los modelos de asociación y la evaluación de la transmisión de alelos, se puede a partir de variantes significativas encontradas evaluar su efecto principal, es decir, como las variantes asociadas pueden influir individualmente en la enfermedad. El método más efectivo y por ende más usado para este propósito es la red de epistasis estadística (SEN), la cual infiere redes de interacción genética que están basadas en los efectos principales, pero van más allá a partir de la estimación de la ganancia de información en la topología de la red. Este enfoque desarrollado por Ting Hu y colaboradores (Hu et al., 2011) clasifica las interacciones entre SNPs de acuerdo con su fuerza relativa y genera redes de epistasis estadísticas que plasman las fuerzas significativas.

Como se ha mencionado, esté método está basado en la teoría de redes y en la ganancia de información, de manera que la estrategia se formula mediante grafos, lo cuales están compuestos por los conjuntos *V* de vértices y *E* de aristas. Los vértices corresponden a SNPs (V_A corresponde a SNP A), mientras las artistas corresponden a la interacción entre los SNP. De esta manera, a cada SNP y cada interacción (arista) se le asigna un peso para cuantificar de qué manera estos explican el estado de la enfermedad. Esta cuantificación se realiza mediante la teoría de ganancia de información (o información mutua), donde específicamente el peso de V_A es $I(V_A; \mathrm{T})$, la información mutua entre el genotipo del SNP A y T, el estado del fenotipo o enfermedad, siendo $I(V_A; \mathrm{T}) = H(\mathrm{T}) - H(\mathrm{T}|V_A)$, donde H corresponde a la entropía, o de manera más explícita, $H(\mathrm{T}|V_A)$ como la medición de la incertidumbre del fenotipo debido a la expresión del genotipo V_A, y todo traducido en la teoría respectiva como:

$$I(V_A; \mathrm{T}) = \sum_{\tau} p(\tau) \log \frac{1}{p(\tau)} - \sum_{a,\tau} p(a, \tau) \log \frac{1}{p(a|\tau)} \tag{2.6}$$

donde $p(\tau)$ es la probabilidad de que un individuo presente el fenotipo T, $p(a, \tau)$ es la probabilidad de que el genotipo *a* se exprese y el individuo presente el fenotipo τ, y $p(a|\tau)$ como la probabilidad de presentar el fenotipo τ, dada la expresión del genotipo *a*. Dado

que en la mayoría de los casos un SNP tiene dos alelos y, en consecuencia, hay tres genotipos posibles para cada SNP, la segunda sumatoria considera las seis combinatorias de genotipos y fenotipo. Según la teoría respectiva, si $I(V_A;\mathrm{T}) = 0$, el fenotipo es independiente del SNP A, y, por lo tanto, el SNP A no predice el estado de la enfermedad. Por el contrario, si $I(V_A;\mathrm{T}) > 0$, hay correlación entre el genotipo y el fenotipo, y entre más grande sea la ganancia de información, se explica una mayor correlación.

Toda la anterior teoría se extiende a la evaluación de las interacciones entre SNPs expresada en los vértices del grafo, así como integraciones de tercer orden, relacionando la ganancia de información que explican tres SNPs asociados. La extensión de esta teoría y la generación del software ViSEN como una estrategia de visualización fue formalizada por los mismos autores unos años posteriores (Hu et al., 2013).

3. Objetivos

3.1 Objetivo general

Desarrollar un modelo de epistasis basado en aprendizaje automático para una cohorte de pacientes de discapacidad intelectual y retraso en el neurodesarrollo (DD/ID) bajo un diseño de asociación familiar, con el propósito de explicar de manera significativa las asociaciones que inciden entre múltiples interacciones de Polimorfismos de Nucleótidos Únicos (SNP) y la enfermedad.

3.2 Objetivos específicos

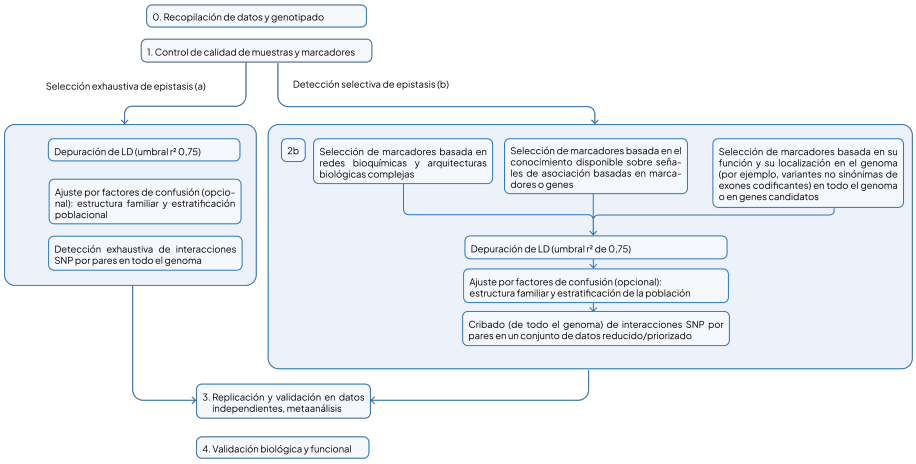
1) Designar los factores genéticos que permitan la obtención del conjunto de casos prevalentes y controles asociados con DD/ID, a partir de una revisión exhaustiva de la literatura y la consecuente extracción de datos retrospectivos que expliquen una heterogeneidad genética.
2) Establecer protocolos estadísticos de análisis y un modelo de epistasis basado en aprendizaje automático capaz de describir las interacciones entre variantes genéticas para DD/ID, a través del conjunto de descriptores genotípicos caracterizados.
3) Interpretar los resultados del modelo de epistasis en función de una red de interacción, que asocie los resultados obtenidos con conocimiento biológico, así como, un metaanálisis por combinación de valores p, para resaltar asociaciones significativas y descartar falsos positivos.

4. Metodología propuesta

Para la elaboración del estudio se establece una aproximación basada en la aplicación de un marco de trabajo propuesto por Ritchie y Van Steen (Marylyn D. Ritchie & Van Steen, 2018) y el protocolo de elaboración de GWAIS por Gusareva y Van Steen (Gusareva & Van Steen, 2014), bajo el dominio de estudios de asociación epistaticos, a partir de recomendaciones, de modo que se logre un progreso significativo y clínicamente relevante para comprender adecuadamente la estructura genética y como sus mecanismos biológicos pueden relacionarse con la enfermedad. Además, como estado inicial se empleará un método de obtención y procesamiento de los datos específico, basado en los marcos de trabajo habituales en estudios con aprendizaje automático, como lo son CRISP-DM y OSEMN (Chapman, 2000; Mason & Wiggins, 2010). Así pues, el flujo de trabajo y diseño del estudio estará condicionado por las recomendaciones, fundamentándose en las siguientes fases:

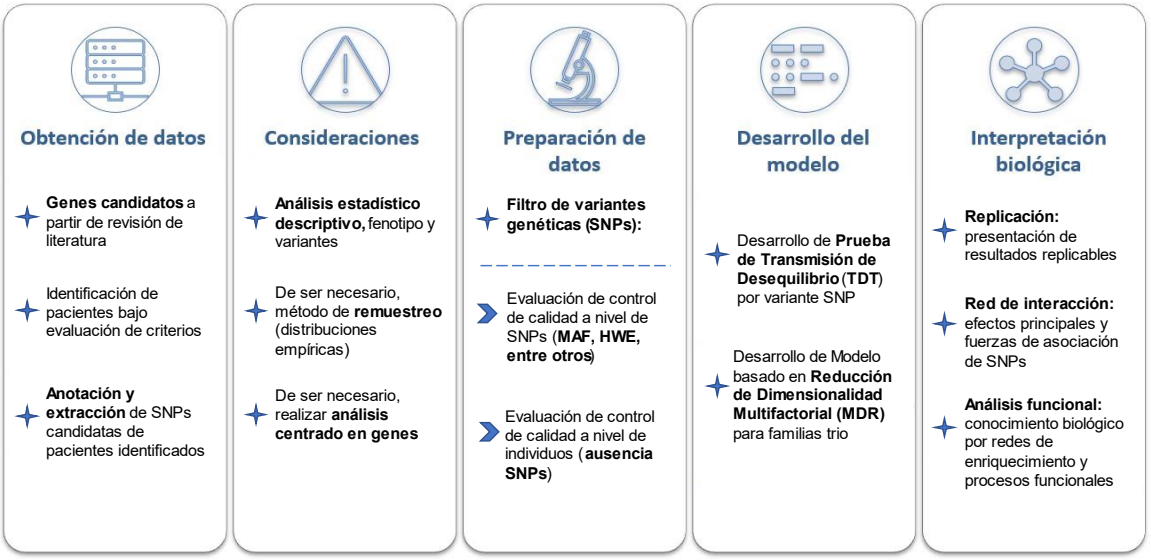
- Obtención de los datos
- Consideraciones: verificación de supuestos y selección del modelo
- Preparación de los datos
- Desarrollo del modelo para detección de epistasis
- Interpretación biológica

Figura 4-1: Marco de trabajo para el análisis de interacciones en estudios genómicos, adaptado de (Gusareva & Van Steen, 2014).



De la misma forma, se resalta qué el análisis de datos basado en la identificación de los genotipos para los SNPs con inclusión de individuos familiares permite validar que la población se encuentra en equilibrio (no existe selección contra ningún genotipo en particular) si las frecuencias de genotipos se mantienen constantes frente a las frecuencias alélicas. En este sentido, se evaluará la composición genética de la población mediante la prueba de equilibrio de Hardy Weinberg, junto con otros controles de calidad de genotipos. Consecuentemente, se indica el diseño del estudio a partir de las anteriores fases, estableciendo una metodología general con el contexto y el esquema de acción secuencial para la elaboración adecuada del modelo y su respectiva interpretación (ver **Figura 4-2.**). Cada uno de estos pasos se describe a continuación.

Figura 4-2: Metodología de estudio poblacional de asociaciones para SNP.



4.1 Fases de elaboración e interpretación del modelo

4.1.1 Obtención de datos

El desarrollo de este trabajo está determinado a partir de un esfuerzo por focalizar y entender aquella fracción de la población colombiana que padece DD/ID, que ha dispuesto de los servicios integrales de salud de Colsanitas y EPS Sanitas, para una evaluación por patologías de etiología genética, a lo largo de los años. Dicha información génica de los pacientes, analizada por el Laboratorio Especializado de Citogenética a partir de exomas clínicos, se presenta almacenada a manera de variantes, en un software de análisis bioinformático y, por lo tanto, se propone un método de extracción retrospectivo de aquellas variantes SNP de interés.

En consecuencia, en primer lugar, la investigación se enfocará en identificar, por un lado, los pacientes con diagnóstico de DD/ID (a partir de los códigos CIE10 en historia clínica electrónica: F70-F79 Discapacidad intelectual y F80-F89 Trastornos generalizados y específicos del desarrollo), y, por otro lado, en identificar el conjunto o panel de genes candidatos asociados a variantes SNP relacionadas al rasgo, que se realiza como un filtrado inicial, a partir de una búsqueda exhaustiva de la literatura. Esta búsqueda pretende reconocer cuales han sido las variantes de tipo SNP que más se han reportado para las

enfermedades relativas a DD/ID en diferentes instancias de publicación como congresos, revistas indexadas o bases de datos y así generar el listado de sus respectivos genes.

De acuerdo con el objetivo determinado, las fuentes de información son definidas de forma tal que puedan proveer la mayor cantidad de documentación posible acerca de los trabajos realizados que reportan algún SNP relacionado a DD/ID. Debido a esto, se plantea una búsqueda sobre los índices de Scopus, Web of Science y Scielo, y los motores de búsqueda como CiteSeerX, Google Scholar, SemanticScholar, Pubmed y OVID.

Dado que el interés del proyecto es reconocer los resultados de trabajos relacionados que realizan estudios poblacionales para la identificación de SNP sobrerrepresentados en la muestra con enfermedades afines a DD/ID, la búsqueda estará relacionada únicamente con patrones en el título, las palabras clave y el resumen (ver **Tabla 4-1**).

Tabla 4-1: Criterios de búsqueda para la identificación de SNPs relacionados.

Categoría de búsqueda	**Descripción**	**Términos**
Enfermedad	Fenotipo enmarcado dentro de las categorías DSM-5 y contemplando el esquema HPO (estándar en de anomalías fenotípicas en enfermedades humanas; Ontología del Fenotipo Humano).	"*intellectual disability*", "*Intellectual development*", "*neurodevelopmental delay*", "*neurodevelopmental abnormality*", "*global developmental delay*", "*learning disability*", "*autism*", "*attention deficit*"
Estudio/Modelo	Modelo aplicado o tipo de estudio llevado a cabo, esto para el respectivo contexto de la búsqueda.	"*epistasis*", "*gene-gene interaction*", "*gene association*", "*GWAS*", "*gene expression*", "*therapeutic targets*"

Los criterios de selección de los documentos al estudio están orientados a identificar la mayor cantidad posible de actividad académica e investigativa relacionada con hallazgos de variantes y genes significativos. Por lo tanto, se consideran los siguientes criterios:

1) El documento debe corresponder a uno de los siguientes tipos: artículo de revista científica, capítulo de libro, artículo sometido a congreso, conferencia o workshop, tesis de posgrado o pregrado con resultados validados por pares.
2) El trabajo debe contar con conclusiones significativas (umbrales de significación en el estudio de asociación con valor p nominal < 5×108). De esta manera, se le dará un mayor valor a trabajos que incluyan experimentos o análisis genéticos que aborden el mecanismo por el cual una variante basada en GWAS da lugar a diferencias fenotípicas.
3) La investigación debe ser rigurosa en sus métodos, esto quiere decir, que controla adecuadamente las comparaciones múltiples, la estratificación de la población, la relación y la calidad técnica. De esta manera, se debe controlar y sustentar tanto el error tipo I, como el tipo II.

En el análisis de exoma completo, para obtener las diferentes variantes SNP y por ende sus respectivos genes, se extraerá la información esencial para el estudio a partir de aplicar algoritmos de análisis bioinformático terciario a partir del exoma alienado, anotando y filtrando las variantes identificadas bajo los criterios anteriormente descritos; específicamente, la información obtenida es gen, coordenada, región, cobertura, frecuencia y variante (tanto secuencia de referencia; *RefSeq ID*, como secuencia de referencia de ADN codificante). Para esto, es necesaria la combinación de los archivos de llamado de variantes (VCF) de los individuos en uno solo, archivos que contienen todas las variantes del exoma completo del probando (hijo) y los padres.

4.1.2 Consideraciones iniciales: verificación de supuestos

En virtud de orientar el estudio de manera consistente a la identificación adecuada de las interacciones, ya con los datos adquiridos, se deben proponer una serie de supuestos apropiados, que deben obedecer al comportamiento de los datos con los que se contará. Por lo tanto, en la metodología se plantea realizar un análisis estadístico descriptivo de los datos que se obtengan, específicamente, en la distribución de la variable dependiente: el fenotipo. Estos análisis van a permitir dictaminar ciertas conjeturas sobre la complejidad, analizando tanto del número de sujetos, como del número de marcadores genéticos, SNPs a estudiar. Por otra parte, la evaluación de distribución a partir de un estadístico de prueba va a permitir evidenciar la distribución del fenotipo en la población, su varianza y la

independencia de las pruebas. En base a lo anterior, se llevará a efecto, de ser necesario un método de remuestreo para obtener distribuciones empíricas sobre las variantes, bajo la adopción de un análisis centrado en genes, de manera que se garantice un control robusto sobre la tasa de error familiar.

4.1.3 Preparación de los datos

Una vez se han evaluado las consideraciones, y en alineamiento con la identificación del problema, se considerará como estrategia de análisis de la asociación, una búsqueda selectiva en un conjunto de marcadores priorizados, de manera que, se aplicarán una serie de filtros a los datos, para seleccionar marcadores en base a una serie de conocimientos reportados científicamente y disponibles a partir de múltiples herramientas, desde un enfoque de dos etapas, considerando factores como anotación respecto a la patología, interacciones a nivel de proteína y de vías metabólicas. Además, de ser necesario, se adoptarán otros tipos de filtrados, diseñados para reducir el espacio de búsqueda.

De esta manera, específicamente, para filtrar los datos bajo los dos enfoques, se estudiarán los resultados de herramientas de anotación, que explican el enfoque en dos etapas, y en cuanto al filtro estadístico, se considerará el uso y evaluación otras estrategias de control de calidad mencionadas en del apartado 2.5.3, exceptuando la poda por LD, ya que a diferencia de otros estudios de asociación de efecto aditivo, la perdida de información es más dramática, donde los SNP filtrados son insuficientes para representar asociaciones complejas (Slim Lotfi AND Chatelain, 2020). Estos métodos de filtrado de datos van a permitir una identificación rápida y exhaustiva de SNP interactuantes potencialmente prometedores.

4.1.4 Desarrollo del modelo para detección de epistasis

Finalmente, para identificar las interacciones entre SNPs de orden superior y debido a que se pretende aprovechar las capacidades computacionales actuales, se implementará un modelo basado en aprendizaje de maquina adaptado para operaciones paralelizadas, a partir de un paradigma no paramétrico y de minería de datos. Por lo cual, se necesitará una previa selección del algoritmo más adecuado frente a la implementación de la prueba de asociación más apropiada, que permita reconocer las interacciones entre las características SNP. La idea de emplear este tipo de modelos recae en la visión de

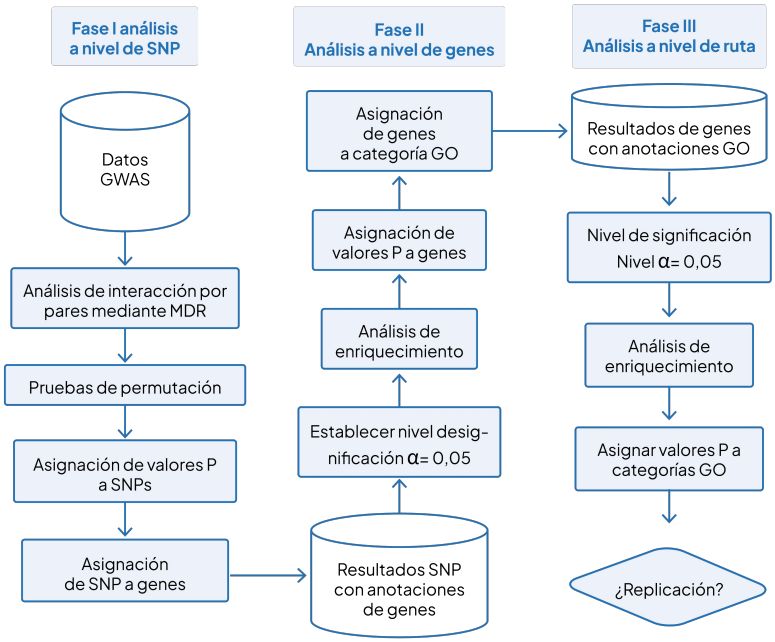
aprovechar los recursos computacionales, sin dejar de lado la eficacia y evitando aquellas señales de efectos que dan lugar a falsos positivos.

4.1.5 Interpretación biológica

En virtud de presentar los resultados de asociación que identifique el modelo, se partirá por expresarlos de una manera clara y metódica, de manera que puedan ser replicados. De este modo, se espera que se identifiquen las mismas asociaciones y SNPs en otros estudios, o al menos la replicación de este estudio se realice para ciertas regiones de interés para los SNPs significativos. Por lo tanto, todo el proceso de replicación se explicará adecuadamente, de manera que pueda efectuarse con el mismo diseño de estudio y dirección.

Seguidamente, para convalidar las variantes implicadas en las asociaciones significativas, se realizará un ejercicio de evaluación con redes de interacción por ganancia de información (ver sección 2.7.1), para las combinatorias potenciales encontradas en el algoritmo MDR. En este sentido, y basados en la metodología de análisis ontológica de Kim y colaboradores (Kim et al., 2012), se cuantificarán los efectos principales de cada SNP de los modelos de epistasis basados en MDR para todas las combinaciones con mejores resultados (medidos en valores p). De esta manera, se seleccionarán aquellas combinatorias asociadas a los modelos MDR con mejor desempeño, agregándoles los resultados del método SEN (como mapas de las interacciones fuertes), y mapeándolos a sus respectivos genes, realizando análisis de funcionalidad y enriquecimiento.

Figura 4-3: Diagrama de flujo para el análisis de la ontología genética en estudios de asociación con SNPs, adaptado de (Kim et al., 2012).



En este sentido, se estaría generando un procedimiento de análisis funcional de los resultados que se obtengan, a partir de conocimiento estructurado en bases de datos (anotaciones funcionales, vías inmunológicas, entre otros), y así el estudio pueda contribuir en cierta medida en el diagnóstico y tratamiento de DD/ID. Para facilitar este proceso, se hará uso de una serie de herramientas que asocien los resultados obtenidos con conocimiento biológico, permitiendo la interpretación biológica de los hallazgos estadísticos de epistasis. Específicamente, se evaluará el uso de ciertas herramientas; como puede ser Cytoscape, incluyendo análisis expresados por el plugin GeneMANIA (Montojo et al., 2010), IMP (Wong et al., 2012) o STRING (Szklarczyk et al., 2023), de manera que se pueda tener una amplia visión de conocimiento genético en los resultados, como proteínas e interacciones genéticas, vías metabólicas, co-expresión, co-localización, similitud de dominio, anotación funcional, entre otros.

4.2 Diseño del estudio

Particularmente, se lleva a cabo un estudio de observacional alusivo a casos y controles del tipo analítico a partir de datos retrospectivos, considerando las triadas familiares, como hijos-casos y padres-controles.

Definición de caso: pacientes con un diagnóstico de trastorno del neurodesarrollo (TND) enfocado a DD/ID, i.e., retardo del desarrollo (DD), la discapacidad intelectual (ID), el trastorno del espectro autista (ASD), el trastorno por déficit de atención/hiperactividad (ADHD) y los trastornos de la comunicación.

Definición de controles: serán aquellos padres del probando caso de la consulta genética remitido a estudio de exoma trio. El padre/madre no debe presentar ningún tipo de trastorno del neurodesarrollo.

Entonces, comprende el estudio de pacientes colombianos a quienes se les haya realizado un estudio de exoma en el Laboratorio Especializado de citogenética y biología molecular de Clínica Colsanitas, debido a diagnostico por consulta a genética clínica, en el periodo de tiempo entre enero de 2018 y diciembre de 2021, siempre y cuando cumplan los criterios de selección. En el estudio se realizará una comparación de la presencia e interacción entre variantes asociadas a DD/ID de los exomas completos en trío para estimar la epistasis a partir del modelo computacional, relacionando los respectivos rendimientos.

4.2.1 Población de estudio

- **Población diana o blanco:** La muestra poblacional estará constituida por pacientes pediátricos (menores de 18 años), a quienes se le estudian alteraciones genéticas mediante prueba de exoma.
- **Población accesible:** Pacientes pediátricos asegurados dentro del sistema de salud colombiano que se encuentran adscritos a la EPS Sanitas, MediSanitas o Colsanitas.
- **Población elegible:** Criterios de selección
    - Criterios de inclusión:
        - Pacientes pediátricos y sus padres que consulten el servicio de genética clínica, remitidos a estudio de exoma en trio/dúo durante el periodo de tiempo de enero de 2018 hasta diciembre de 2021.
    - Criterios de exclusión:

- Paciente con identificación de la etiología del trastorno por evento traumático (noxa perinatal, parto distócico, trauma severo de la infancia o adolescencia con evidencia clara de lesión estructural, infección complicada y sintomática neurológica como encefalitis localizada o generalizada, antecedentes claros de anoxia o hipoxia perinatal o de la infancia y adolescencia por evento traumático).
- Pacientes con epilepsia parcial o generalizada sindrómica con evidencia clara de etiopatogenia genética o hereditaria evidenciada por transmisión familiar o presencia de mutaciones en genes ya descritos asociados a los síndromes epilépticos.
- Pacientes con metabolopatía clara y definida de la etapa perinatal con cuadro severo de sufrimiento neonatal o perinatal conducente a deterioro neurológico.
- Síndromes conocidos y definidos por aneuploidías identificadas claramente por el análisis citogenético como el Síndrome de Down.

4.2.2 Variables

En la **Tabla 4-2** se presentan las variables que se usarán para evaluar de manera descriptiva el comportamiento de los hallazgos; vale anotar que no son predictoras. Las variables para evaluar su interacción corresponden a la presencia o no de un SNP específico.

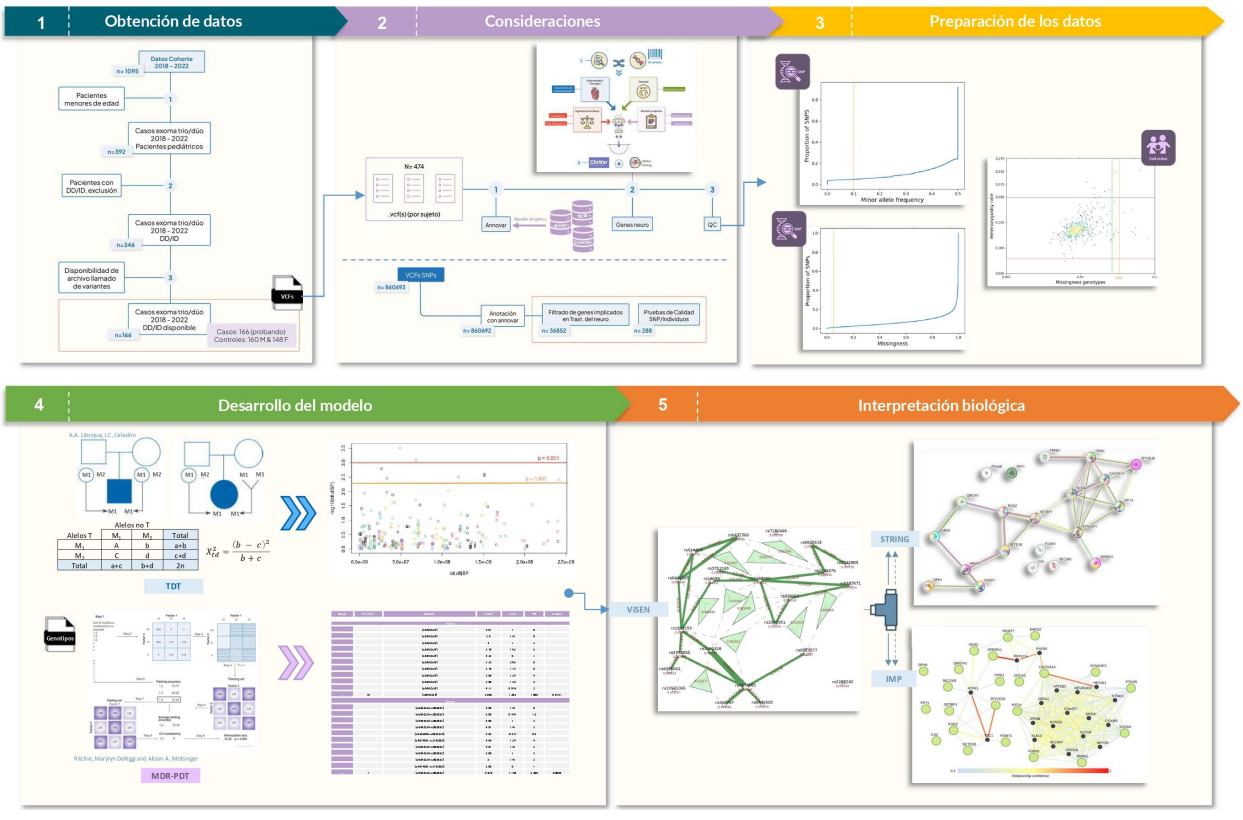
Tabla 4-2: Variables consideradas para el estudio sobre casos exoma trio

Variable	Medición	Escala de medición	Función	Unidades
Edad	Cuantitativa	Razón	Independiente	Años
Indicación clínica	Cualitativa	Nominal	Independiente	DD, ID, ASD, ADHD, Trastornos de la comunicación.
Género	Cualitativa	binomial dicotómica	Independiente	Masculino, Femenino
Variante SNV	Cualitativa	binomial dicotómica	Independiente	A,C,T,G

5. Análisis y resultados

En base a la metodología propuesta, y en función de los procedimientos aplicados, se presenta un pipeline compuesto por los pasos de recolección, integración, anotación, control de calidad, identificación de asociaciones e interpretación. Este pipeline se puede evidenciar en la **Figura 5-1**.

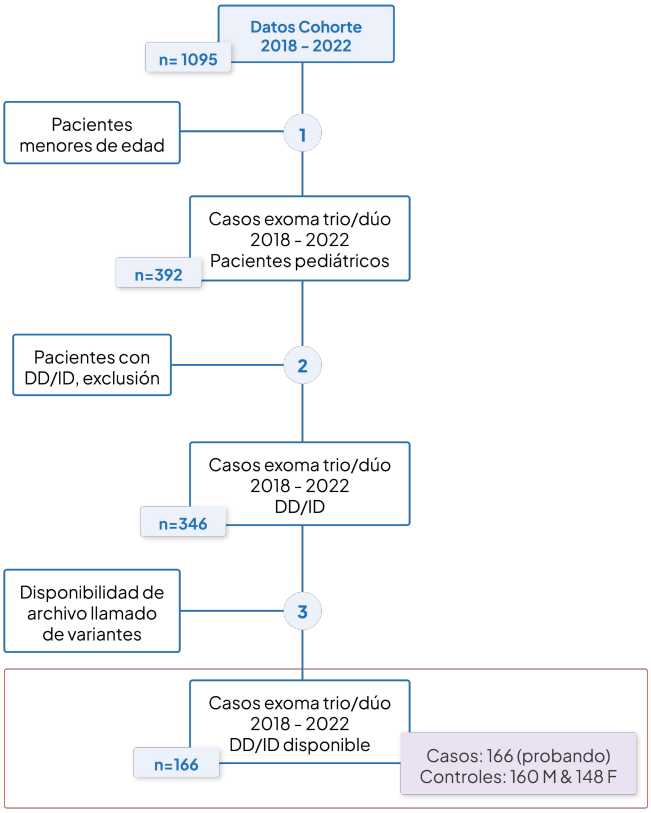
Figura 5-1: Pipeline para el análisis de asociaciones SNP en estudio familiar



5.1 Análisis del conjunto de datos

En función de realizar la construcción de los datos para el estudio, a partir de lo definido en la sección 4.2 sobre la estructura del estudio, se identificaron de manera retrospectiva 1095 casos de estudio exoma trio procesados durante los años 2018 y 2021, esto es, todas las pruebas reportadas por la entidad. Sobre todos estos casos de estudio y teniendo en cuenta los criterios de inclusión, se aplicaron una serie de filtros para obtener el conjunto final (ver **Figura 5-2**): 1) consecuente al comportamiento de la enfermedad se seleccionaron aquellos probandos (paciente al que se le realiza el estudio, y particularmente el hijo en el exoma trio) pediátricos; menores de edad (< 18 años); 2) como fenotipo objeto del estudio, se seleccionaron los probandos con una indicación clínica relacionada con diagnostico o sospecha reportada de DD/ID, esto considerando los criterios de DSM-5; y 3) con una población resultante de 346 casos de estudio, se seleccionaron finalmente los casos para los que se cuenta con disponibilidad del archivo resultante de la prueba de exoma VCF para el probando, obteniendo una población final de 166 casos de exoma trio.

Figura 5-2: Flujo de trabajo para la selección de la población.



De esta manera, se cuenta con una cantidad total de 474 individuos y sus respectivos archivos VCF, de los cuales 166 son probandos, 160 madres y 148 padres. Es importante aclarar que las pruebas de exoma trio, aunque su nombre menciona tres individuos, se puede realizar para el probando y al menos uno de sus padres. Para fines de análisis particulares durante las etapas del estudio, como la prueba de transmisión TDT, se consideran estas familias trio incompletas. Sin embargo, en la aplicación del modelo de aprendizaje de máquina se seleccionan aquellos casos con triada completa.

5.2 Identificación de genes candidatos

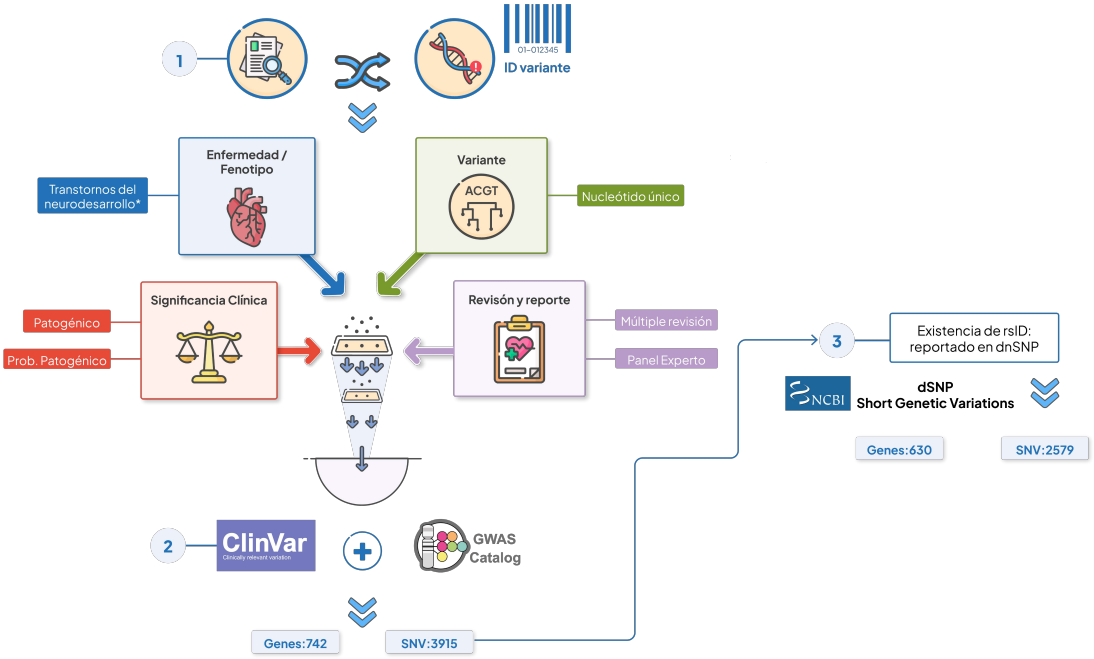
Como principal insumo comentado en la metodología propuesta, se realizó una búsqueda exhaustiva en varios recursos bibliográficos de variantes reportadas alrededor del espectro los trastornos TND. Para este proceso se aplicaron los criterios mencionados en la **Tabla 4-1**. Un ejemplo de la estructura de una consulta llevada a la sintaxis de *Pubmed* mediante términos MesH (contexto clínico) se muestra a continuación:

> *("Intellectual Disability"[Mesh] OR "Language Development Disorders"[Mesh] OR "Neurodevelopmental Disorders"[Mesh] OR "Learning Disabilities"[Mesh] OR "Autism Spectrum Disorder"[Mesh] OR "Attention Deficit Disorder with Hyperactivity"[Mesh]) AND ("Epistasis, Genetic"[Mesh] OR "Gene Expression"[Mesh] OR "Gene Expression Profiling"[Mesh] OR "Gene Regulatory Networks"[Mesh] OR "Gene-Environment Interaction"[Mesh] OR "Genetic Association Studies"[Mesh] OR "Genome-Wide Association Study"[Mesh]).*

En cuanto a la selección de variantes reportadas, se aplican filtros de calidad (ver **Figura 5-3**) para seleccionar los genes potenciales relacionados a SNPs confirmados clínica y estadísticamente, estos filtros consisten en: 1) sobre el conjunto de publicaciones encontradas se asocia el código único de la variante reportada sobre las bases de datos *ClinVar* y *GWAS-Catalog*; 2) una vez asociada la variante, aplicar criterios que permiten considerar un SNV como una SNP asociada a la prevalencia de la enfermedad (“Clinical Utility of Genetic and Genomic Services: A Position Statement of the American College of Medical Genetics and Genomics.,” 2015), es decir, que la variante reportada sea de tipo nucleótido único, la enfermedad o fenotipo asociado sea parte del espectro TND, la significancia clínica sea del tipo patogénico o probablemente patogénico y, finalmente, que

la variante sea reportada por múltiple revisión o por un panel experto; y 3) sobre los SNVs obtenidos, validar que se encuentren reportados como polimorfismo (SNP), presentando un identificador asociado (dbSNP).

Figura 5-3: Flujo de trabajo para la selección de genes candidatos.



De esta manera se obtiene finalmente un conjunto de 2579 SNPs asociados a 630 genes candidatos para el estudio. Sobre este conjunto de genes se resaltan algunos de ellos con las respectivas variantes encontradas (ver **Tabla 5-1**). El gen KAT6A ha sido relacionado en varios estudios para las condiciones de discapacidad intelectual autosómica dominante y anomalías craneofaciales, y específicamente la variante relacionada, rs786200960, demuestra una mutación *nonsense* con cambio de C>T que altera la participación en la acetilación y desacetilación de histonas, actuando en la desregulación de la acetilación de H3K9 y H3K18, y dando resultado a varias anomalías congénitas. Por otro lado, el gen PPP2R1A, que comprende subunidades que determinan la especificidad del sustrato y la función fisiológica, bajo la mutación rs1057519946, relaciona una desregulación de la actividad de la proteína fosfatasa como causa de discapacidad intelectual debido a la sobreexpresión de algunas subunidades mutantes.

Tabla 5-1: Genes y variantes asociadas a TND reportadas en la literatura

Gene	Nombre	Condiciones	Significancia clínica	VariationID	dbSNP ID
ATRX	NM_000489.6(ATRX):c.5540A>G (p.Tyr1847Cys)	Síndrome de discapacidad intelectual ligado al cromosoma X-alfa talasemia | Hipotonía neonatal	Probablemente patogénico	1172655	rs1057521987
	NM_000489.6(ATRX):c.109C>T (p.Arg37Ter)	Enfermedades genéticas congénitas | Síndrome de discapacidad intelectual ligado al cromosoma X | Discapacidad intelectual	Patogénico | Probablemente patogénico	11742	rs122445108
CLASP1|RNU4ATAC	NM_001395891.1(CLASP1):c.196-609C>T	Síndrome de Lowry-Wood | Síndrome de Roifman | Enanismo primordial osteodisplásico, tipo 1 | Discapacidad intelectual | Estatura baja	Patogénico	30179	rs575472572
EHMT1	NM_024757.5(EHMT1):c.1647+2T>C	Síndrome de Kleefstra 1 | Esquizofrenia | Retraso global del desarrollo | Discapacidad intelectual | Sinofridia | Rasgos faciales toscos	Patogénico | Probablemente patogénico	374120	rs1057518913
GATAD2B	NM_020699.4(GATAD2B):c.1241G>A (p.Arg414Gln)	Síndrome de discapacidad intelectual grave | lenguaje deficiente | estrabismo | cara con muecas | dedos largos	Patogénico | Probablemente patogénico	381463	rs1057521041
KAT6A	NM_006766.5(KAT6A):c.3385C>T (p.Arg1129Ter)	Síndrome de discapacidad intelectual autosómica dominante | anomalías craneofaciales | defectos cardíacos | Enfermedades genéticas congénitas	Patogénico	180229	rs786200960
MTOR	NM_004958.4(MTOR):c.7500T>G (p.Ile2500Met)	Síndrome de macrocefalia | discapacidad intelectual | trastorno del neurodesarrollo | tórax pequeño	Probablemente patogénico	376455	rs1057519915
MYT1L	NM_001303052.2(MYT1L):c.1678C>T (p.His560Tyr)	Discapacidad intelectual	Probablemente patogénico	617493	rs1558371790
	NM_001303052.2(MYT1L):c.1706G>A (p.Arg569Gln)	Discapacidad intelectual autosómica dominante |	Patogénico | Probablemente patogénico	235469	rs878853045

		Discapacidad intelectual			
POLA1	NM_001330360.2(POLA1):c.463-2A>T	Discapacidad intelectual ligada al cromosoma X, tipo van Esch	Patogénico	1172688	rs2148341850
PPP2R1A	NM_014225.6(PPP2R1A):c.547C>T (p.Arg183Trp)	Síndrome de microcefalia | Agenesia del cuerpo calloso | discapacidad intelectual | dimorfismo facial	Patogénico | Probablemente patogénico	376505	rs1057519946
PTEN	NM_000314.8(PTEN):c.203A>G (p.Tyr68Cys)	Autismo | Retardo global del desarrollo | Espasmos infantiles | Convulsiones | Retardo del desarrollo | Deterioro visual cerebral | Deterioro cognitivo	Patogénico | Probablemente patogénico	233777	rs876660634
TAF1	NM_004606.5(TAF1):c.3508C>T (p.Arg1170Cys)	Discapacidad intelectual, ligada al cromosoma X, sindrómica 33	Patogénico	599303	rs1569301036
	NM_004606.5(TAF1):c.3950T>C (p.Ile1317Thr)	Discapacidad intelectual, ligada al cromosoma X, sindrómica 33	Patogénico	219114	rs864321627

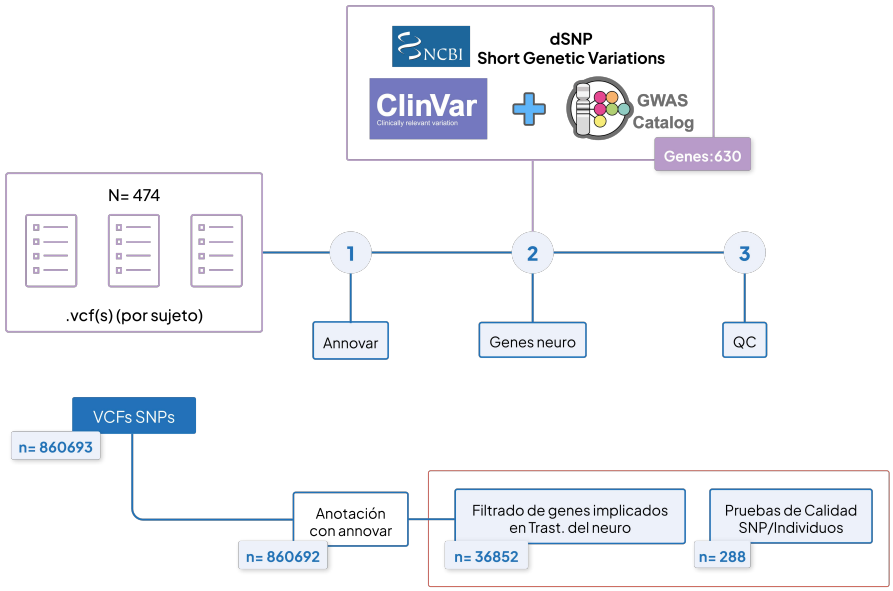
5.3 Procesamiento y control de calidad

Una vez se obtuvieron el conjunto de casos de exoma, con 474 individuos y los genes candidatos a partir de la revisión de literatura, se consolidaron los datos agrupándolos en un único archivo multi muestra VCF y se seleccionaron únicamente las variantes SNV con el uso de la herramienta BCFtools (Danecek et al., 2021), donde se realizaron las modificaciones pertinentes para relacionar los identificadores de los individuos bajo el formato FID_IID_(P/M/F), el cual permite relacionar las familias y cada uno de sus miembros, de tal forma que FID es el identificador de familia (de 1 hasta 166), IID es el identificador del caso de exoma procesado y (P/M/F) representa que miembro de la familia es, es decir, Probando (P), Madre (M) o Padre (F).

Sobre el archivo multi muestra VCF se realiza una análisis terciario de anotación de variantes a partir de ANNOVAR, bajo su funcionalidad de anotación por genes y, seguidamente, se realiza el filtrado de los genes evidenciados en la literatura, a partir de la herramienta VCFtools (Danecek et al., 2011). Al generar la anotación para el VCF, se evidenció la presencia de 860692 variantes SNPs, seguidamente en el filtro de 630 genes

asociados con la enfermedad y encontrados en la literatura, se localizaron en ellos 36852 SNPs y, finalmente, con la aplicación de los criterios de control de calidad de genotipos se concluye con un conjunto de 288 SNPs. Los procedimientos descritos anteriormente se pueden observar en la **Figura 5-4**.

Figura 5-4: Flujo de trabajo para la selección de variantes en las muestras VCF.



Se aplicaron los controles de calidad descritos en la sección 2.5.3 sobre los 36852 SNPs para 474 individuos, con base en los algoritmos definidos por PLINK v1.9 (C. C. Chang et al., 2015). Esta herramienta de código abierto se emplea en procesos de procesamiento de datos establecidos para mapeo de rasgos y estudios genéticos poblacionales, siendo una de las más usadas. Para hacer uso de los algoritmos de la herramienta, se transformó el archivo VCF al formato requerido por PLINK; un archivo tabular de genotipo binario bialélico (.bed), con las representaciones de los llamados genotípicos, acompañado por, un archivo de información de variante extendida (.bim) con información de la localización y alelos de cada variante, y para el contexto del estudio, un archivo de información de las muestras (.fam) con datos de los identificadores FID_IID_(P/M/F), sexo y presencia del fenotipo.

En primer lugar, se evaluó la frecuencia de los alelos MAF para eliminar las variantes muy raras en la población, evaluando y generando con la herramienta un reporte de frecuencias

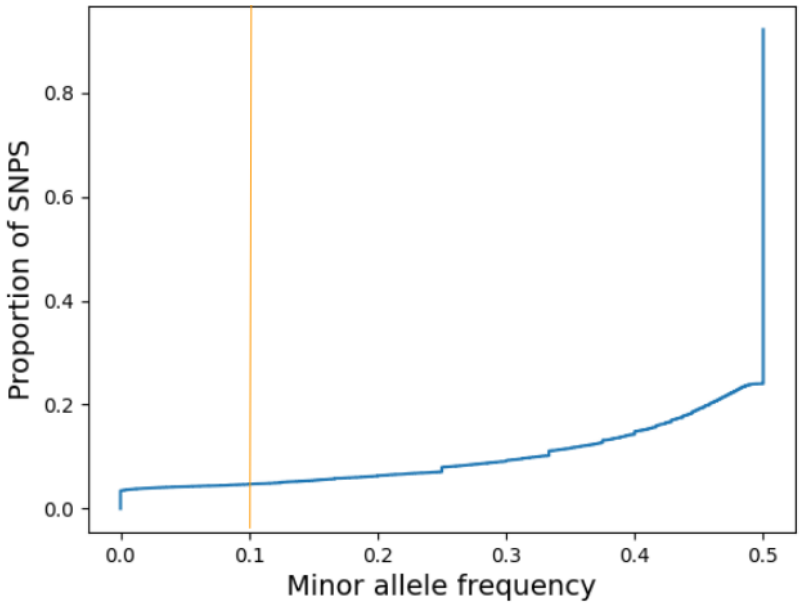
alélicas. En la **Tabla 5-2** se pueden evidenciar las 15 variantes con menor MAF (≠ 0). A partir de esta información se eliminaron de la muestra 1379 variantes bajo el criterio recomendado de excluir variantes con MAF < 0.02, debido a la baja fuerza de detección de asociaciones en SNPs muy raros. De manera más gráfica y con la ayuda del pipeline de calidad H3Agwas (Brandenburg et al., 2022), la distribución del MAF en los SNPs se puede observar en la **Figura 5-5**. El filtro de frecuencia alélica redujo los 36852 SNPs a 35473.

Tabla 5-2: Frecuencia alélica MAF de las 15 variantes con menor tasa.

SNP	**A1**	**A2**	**MAF**	**NCHROBS[1]**
rs6593795	C	T	0.001613	620
rs7688609	G	A	0.001613	620
rs4764010	A	G	0.001613	620
rs9898024	T	C	0.001613	620
rs7217707	T	C	0.001613	620
rs611326	C	G	0.001613	620
rs4838864	C	G	0.001613	620
rs9882534	G	C	0.001618	618
rs6503030	G	A	0.001618	618
rs9877581	A	G	0.001623	616
rs852422	A	G	0.001634	612
rs4819371	C	T	0.001684	594
rs548858	A	G	0.001761	568
rs4806260	G	A	0.001887	530
rs4865476	G	C	0.001992	502

[1] número de alelos observados

Figura 5-5: Espectro de MAF acumulado respecto a los SNPs.



En relación con la evaluación de la falta de llamados genotípicos, se estimó la distribución de cada SNPs en la población, cuantificando la frecuencia de ausencia (1 menos la tasa de llamados). En la **Tabla 5-3** se puede evidenciar la tasa de ausencia de llamados (F_miss) de las 15 variantes con mayores tasas, así mismo, en la **Figura 5-6** se puede visualizar todo el espectro de la proporción de SNPs respecto a las tasas de ausencia. Como indica esta última ilustración, se eligió un umbral de 0.05 para eliminar los SNPs, en otras palabras, se eliminaron todas las variantes con F_miss > 5%. Habitualmente, se suelen usar umbrales entre 1% a 5%. Este es uno de los filtros más impactantes para los datos del estudio, reduciendo el conjunto de datos de 35473 SNPs a 305 (99,14% variantes eliminadas).

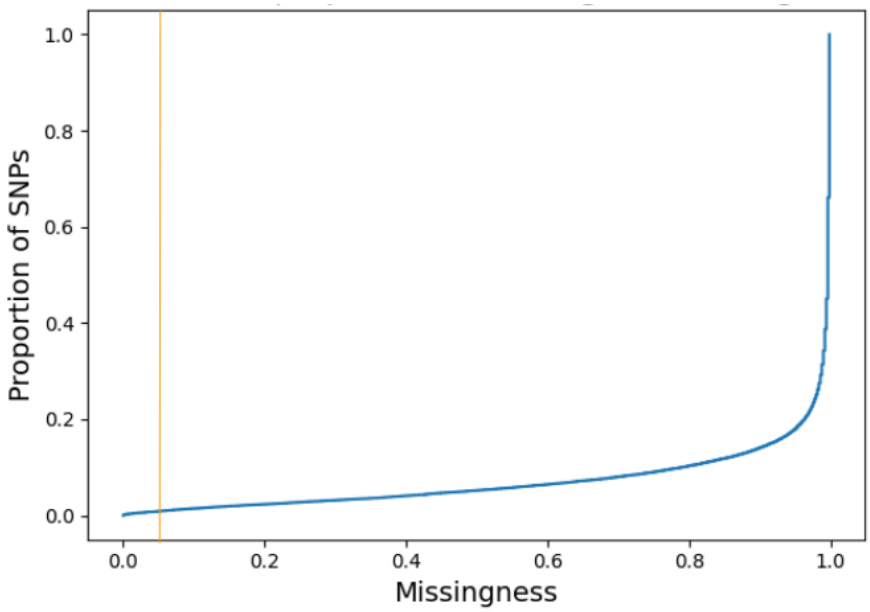
Tabla 5-3: Tasa de ausencia de llamados de las 15 variantes con mayor proporción.

CHR	SNP	N_MISS[1]	N_GENO[2]	F_MISS[3]
1	rs115850563	473	474	0.9979
1	rs116733754	473	474	0.9979
1	rs143955960	473	474	0.9979
1	rs10796395	473	474	0.9979
1	rs1452987372	473	474	0.9979
1	rs139000965	473	474	0.9979

1	rs758197716	473	474	0.9979
1	rs2144688	473	474	0.9979
1	1433455380	473	474	0.9979
1	rs141738594	473	474	0.9979
1	rs146880714	473	474	0.9979
1	rs1007805789	473	474	0.9979
1	rs887936948	473	474	0.9979
1	rs12070592	473	474	0.9979
1	rs116487972	473	474	0.9979

[1] número de llamadas genotípicas faltantes
[2] número de llamadas potencialmente válidas (N muestra)
[3] tasa de ausencia de llamados

Figura 5-6: Distribución de proporción de SNPs en función de su tasa de ausencia de llamados genotípicos.



El último control de calidad a nivel de variantes aplicado fue la prueba de Hardy-Weinberg (HWE). Para esta prueba y como lo sustenta el principio de la sección 2.5, se pretende evaluar que tan constantes son las frecuencias alélicas y genotípicas, a partir de calcular las frecuencias genotípicas esperadas a partir de las frecuencias alélicas observadas y por medio de la prueba X^2 compararlas con los conteos genotípicos observados, los resultados de la prueba se evidencian en la gráfica de proporción de SNPs respecto a su significancia en la prueba HWE (ver **Figura 5-7**). Bajo el análisis sobre el valor-P de la prueba y con las

recomendaciones respectivas (ver 2.5.3), se eligió para la prueba caso/control un umbral de 1e-6, donde las variantes con un valor menor fueron eliminadas, siendo estas 17 con sus respectivos valores evidenciados en la **Tabla 5-4**. De esta manera, se finalizó con un conjunto de 288 SNPs.

Figura 5-7: Proporción de SNPs en función de su significancia HWE

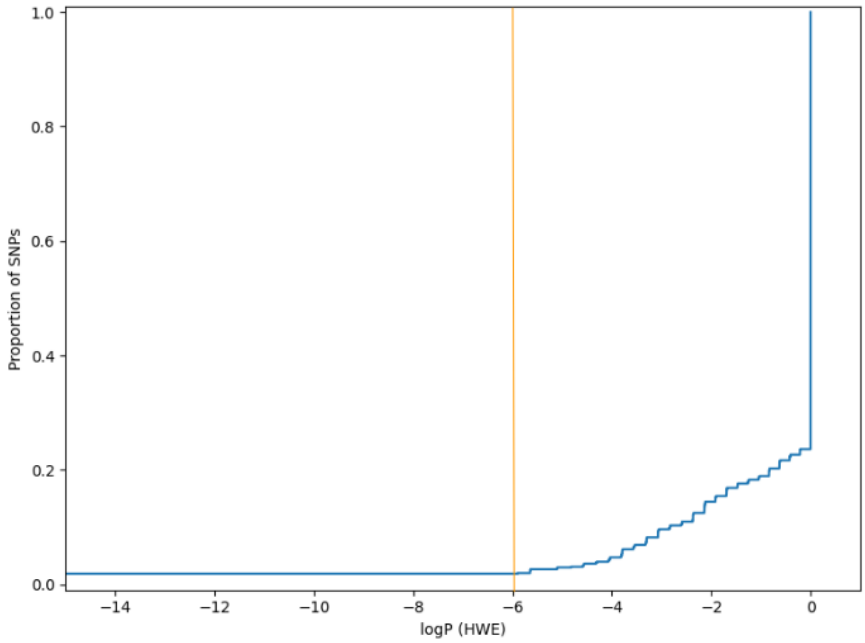


Tabla 5-4: Prueba Hardy-Weinberg sobre SNPs eliminados.

CHR	SNP	A1	A2	GENO	O(HET)	E(HET)	P
7	rs62481502	G	A	0/310/0	1	0,5	1,564E-92
7	rs62478355	G	C	0/310/0	1	0,5	1,564E-92
7	rs10454320	T	A	0/310/0	1	0,5	1,564E-92
7	rs112326730	C	T	0/309/0	1	0,5	2,989E-92
7	rs4024453	A	G	0/308/0	1	0,5	6,239E-92
7	rs62478356	A	T	0/308/0	1	0,5	6,239E-92
7	rs201948579	A	G	0/308/0	1	0,5	6,239E-92
7	rs2537264	T	G	0/306/0	1	0,5	2,488E-91
7	rs3896406	C	G	0/302/0	1	0,5	3,955E-90
7	rs62481501	G	A	0/299/0	1	0,5	3,010E-89
7	rs2479172	C	T	0/305/1	0,9967	0,5	3,985E-89
7	rs77735469	G	A	0/297/0	1	0,5	1,200E-88

7	rs2537263	T	C	0/297/0	1	0,5	1,200E-88
7	rs28439884	T	C	0/308/2	0,9935	0,5	2,143E-88
7	rs879080169	T	C	0/308/2	0,9935	0,5	2,143E-88
7	rs28522267	C	A	0/303/7	0,9774	0,4997	6,998E-81
6	rs672648	C	A	0/228/78	0,7451	0,4675	7,297E-32

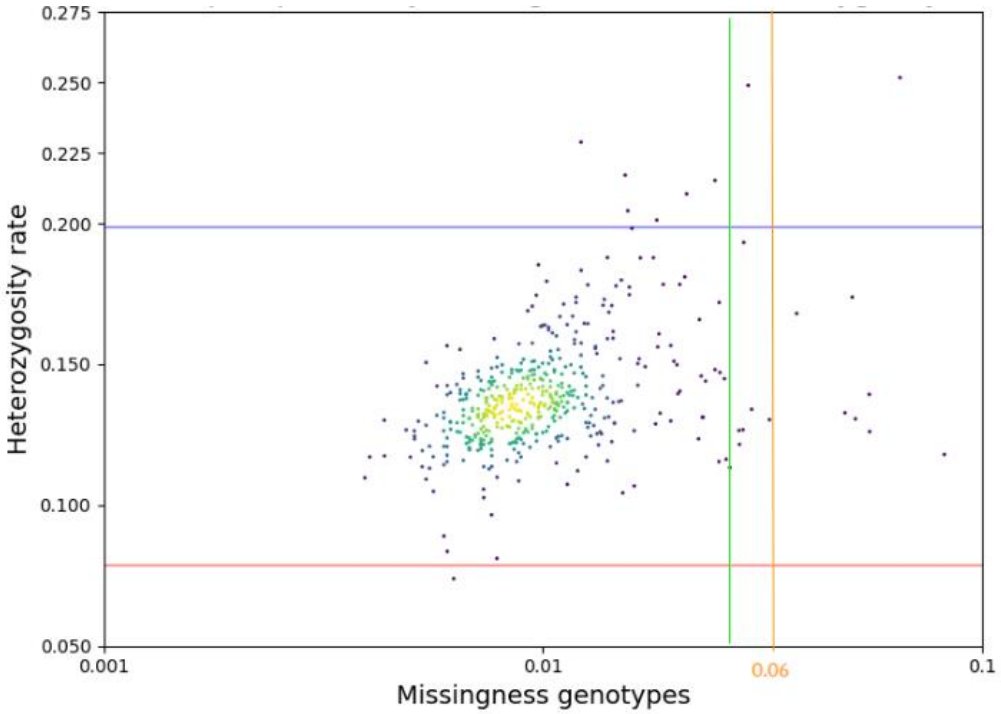
Finalmente, y como último control de calidad, se evaluaron las frecuencias de SNPs para cada individuo o muestra, teniendo en cuenta las tasas de heterocigocidad y la tasa de llamados genotípicos (ver **Tabla 5-5**). De esta manera, se excluyeron los individuos que no se encontraban dentro del límite de tasa de heterocigocidad ± 3 desviaciones estándar de la media o una tasa de ausencia mayor a 0.06 (ver **Figura 5-8**). Estos dos criterios excluyen 16 muestras, sin embargo, también se elimina toda la familia del estudio si 1) se elimina el probando o hijo (-P) o 2) el caso exoma trio se queda sin padres (p.ej. la familia F24, donde se elimina al padre). Lo anterior supone reducir las muestras de 474 a 450 (excluidos 24 individuos, 6 familias).

Tabla 5-5: Tasa de ausencia de llamados y heterocigocidad por muestra, excluidos.

FID	IID	Llamados genotípicos			Heterocigocidad			
		N_MISS	N_GENO	F_MISS	O(HOM)	E(HOM)	N(NM)	F
F40	51610280-P	29	288	0.1007	214	214.6	257	-0.01392
F119	62604549-M	27	288	0.09375	221	214.2	257	0.1585
F175	102009856-F	25	288	0.08681	196	216.8	259	-0.4922
F80	52101591-P	25	288	0.08681	219	217.2	260	0.04288
F103	81301681-M	25	288	0.08681	164	215.5	259	-1.184
F212	100210526-F	24	288	0.08333	192	218.9	260	-0.6551
F139	92509033-F	24	288	0.08333	188	218.3	261	-0.7097
F145	90705584-M	21	288	0.07292	220	220.5	263	-0.0113
F24	20512919-F	20	288	0.06944	219	221.3	264	-0.05282
F69	51100904-M	20	288	0.06944	219	220.9	264	-0.04407
F141	90118090-P	20	288	0.06944	215	221.3	266	-0.1401
F134	90206316-F	20	288	0.06944	201	220.9	264	-0.4625
F127	91110455-M	20	288	0.06944	235	219.8	264	0.3432
F92	62303300-F	19	288	0.06597	228	221.7	265	0.145

F40	51610280-M	18	288	0.0625	200	222.9	266	-0.5329
F140	92803566-M	18	288	0.0625	220	222.8	266	-0.06544

Figura 5-8: Muestras en función de ausencia genotípica vs heterocigocidad.

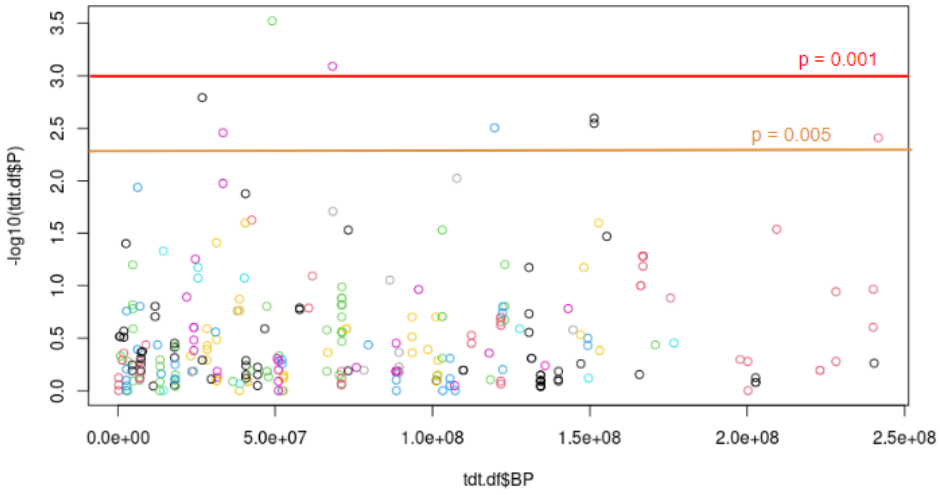


El procesamiento de muestras y control de calidad concluyó con 288 SNPs y 450 muestras, donde exclusivamente 142 de 159 familias son triadas completas, correspondiendo a 426 muestras.

5.4 Prueba de TDT para asociación SNP-fenotipo

Con el fin de evaluar la importancia de cada SNP frente a la enfermedad, se realizó una prueba X^2 de desequilibrio de transmisión (ver **Figura 5-9**). Nuevamente, para este procedimiento se hace uso de la herramienta PLINK, con la implementación de (2.) y sobre los 288 SNPs con 450 muestras. Es importante tener en cuenta que esta prueba de transmisión/desequilibrio evalúa únicamente la transmisión en los probandos para tener una idea del efecto inicial de las variantes. Si se tuviera información del fenotipo en los padres, entonces esta prueba podría agregar un poder considerable al análisis de asociación basado en la familia.

Figura 5-9: Significancia de SNPs para TDT



Los resultados de la prueba (ver **Tabla 5-6**) evidencian ocho variantes relevantes (< 5e-3) y se resaltan las variantes con un valor-P significativo (< 1e-3). Estas dos variantes corresponden a rs3742883 y rs4974081. La primera de ellas, una mutación *missense* reportada como benigna para un estudio de Parkinson esporádico, localizada en el gen ZFYVE26, el cual codifica una proteína que contiene un dominio de unión a dedos de zinc FYVE, encargado de dirigir estas proteínas a los lípidos de la membrana y por lo tanto sus mutaciones suelen ser asociadas a espasmos. En el caso de rs4974081, se le conoce como una variante en el gen QRICH1; implicado en procesos de unión del ADN y en la respuesta de proteína desplegada para las vías de señalización apoptótica intrínseca en respuesta al estrés del retículo endoplásmico. Además, es importante mencionar que las variantes relevantes rs3748550 y rs6587577 han sido asociadas como benignas para el síndrome de discapacidad intelectual, microcefalia, estrabismo y alteraciones del comportamiento.

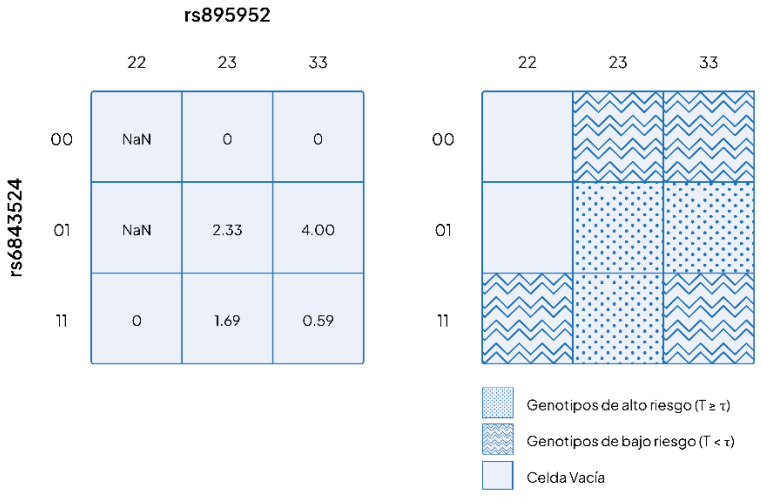
Tabla 5-6: Prueba de transmisión/desequilibrio X^2, variantes relevantes

CHR	SNP	BP	A1	A2	T	U	OR	L95	U95	CHISQ	P
3	rs4974081	49070499	C	T	16	44	0.3636	0.2052	0.6444	13.07	0.0003006
14	rs3742883	68234539	T	C	19	46	0.413	0.242	0.7049	11.22	0.0008112
17	rs614434	26851501	G	A	21	47	0.4468	0.2671	0.7474	9.941	0.001616
1	rs3748550	151384733	G	A	29	57	0.5088	0.3254	0.7956	9.116	0.002533
1	rs6587577	151402045	A	G	30	58	0.5172	0.3329	0.8037	8.909	0.002838
4	rs6843524	119736598	C	T	45	21	2.143	1.277	3.597	8.727	0.003135
6	rs3119019	33414637	G	T	7	23	0.3043	0.1306	0.7093	8.533	0.003487
2	rs6437368	241659368	C	A	25	50	0.5	0.3094	0.8081	8.333	0.003892

5.5 Entrenamiento y prueba del modelo de inducción constructiva

Posteriormente, se realiza el desarrollo del modelo de aprendizaje automático MDR-PDT bajo el principio de inducción constructiva. Para este algoritmo, descrito en la sección 2.6.1 se llevó a cabo a partir de la implementación pMDR (MDR paralelo) del laboratorio Ritchie[2]. En la ejecución del algoritmo se usaron los datos de triadas completas (426 muestras y 288 SNPs), donde se generaron los procesos de validación cruzada (CV) seleccionando subconjuntos de combinaciones de variantes. En la **Figura 5-10** se puede visualizar un caso de la construcción inductiva k=2 por tablas de contingencia para el cálculo del estadístico geno-PDT (2.) para cada multifactor o combinación de genotipos. Es importante resaltar que al no presentar hermanos discordantes informativos (DSP) en el conjunto de datos, la variable aleatoria resumida se simplifica a $\mathrm{D(g)} = \sum_{j=1}^{n_T} X_{Tj}(g)$. Las combinatorias son catalogadas alto riesgo cuando superan el umbral; $\mathrm{T}(g) > (\tau = 1)$.

Figura 5-10: Caso de tablas de contingencia k=2 para estadístico genotipo-PDT



[2] Ritchie Lab, University of Pennsylvania. Software pMDR (https://ritchielab.org/software/mdr-downloads-1)

En el caso del modelo rs6843524-rs895952, se obtuvo para el Fold 1 de la CV un 15.38% de ausencia en los genotipos, donde para las celdas vacías no se encuentra ninguno y, por ende, no se puede medir la variable resumida D(g). Los resultados de la prueba reflejaron un estadístico-T (estadístico MDR-PDT) de 4.99; este estadístico es el resultado de calcular geno-PDT a los genotipos agrupados para alto riesgo, representado la medida de asociación entre la combinación de genotipos y la enfermedad. De manera similar al efecto principal, entre mayor sea el estadístico mayor asociación expresa.

De esta manera, el modelo de MDR-PDT fue realizado para 1, 2 y 3 locus combinados (k), generando varios modelos con sus respectivos valores de estadístico-T y valor-P medido a partir de 1000 pruebas de permutación. Estas pruebas de permutación fueron paralelizadas para la optimización de ejecución del algoritmo. Los modelos para cada k fueron clasificados de mejor a peor, a partir la consistencia en la validación cruzada, es decir, el número de veces que se identifica el mismo modelo MDR para los subconjuntos o Folds. Los mejores dos modelos de cada k se presentan en la **Tabla 5-7**.

Tabla 5-7: Mejores modelos seleccionados para MDR-PDT

Rango	XV Cons[1]	Modelo	T-Train[2]	T-Test	OR	P-value
1-locus						
		[rs6843524]	4.01	1	0	
		[rs6843524]	3.9	1.41	0	
		[rs6843524]	4	1	3	
		[rs6843524]	3.78	1.63	5	
		[rs6843524]	4.32	0	1	
		[rs6843524]	3.33	2.65	0	
		[rs6843524]	3.79	1.73	0	
		[rs6843524]	3.89	1.34	4	
		[rs6843524]	3.89	1.34	4	
		[rs6843524]	4.11	0.816	2	
1	**10**	**[rs6843524]**	**3.905**	**1.292**	**1.900**	**0.0151**
		[rs3742883]	3.7	-0.447	0.667	
		[rs6587577]	3.1	0.632	1.5	
		[rs3742883]	3.28	0.816	2	
		[rs4974081]	3.32	0	1	
		[rs3742883]	3.62	-0.577	0.5	

		[rs614434]	3.32	-0.447	0.667	
		[rs6587577]	3.25	0	1	
		[rs3742883]	3.32	0.577	2	
		[rs4974081]	3.13	1	0	
		[rs3742883]	3.7	-0.447	0.667	
2	**5**	**[rs3742883]**	**3.524**	**-0.016**	**1.000**	**0.0622**

2-locus						
		[rs6843524-rs895952]	4.99	1.41	0	
		[rs6843524-rs895952]	5.08	0.447	1.5	
		[rs6843524-rs895952]	5.08	1	3	
		[rs6843524-rs895952]	4.81	1.41	3	
		[rs6442905-rs6843524]	4.92	-0.577	0.5	
		[rs4974081-rs3742883]	4.64	1.34	4	
		[rs6843524-rs895952]	4.81	1.41	3	
		[rs6843524-rs895952]	5.08	1	3	
		[rs6843524-rs895952]	5	1.41	3	
		[rs4974081-rs3742883]	5.08	0	1	
1	**7**	**[rs6843524-rs895952]**	**4.979**	**1.158**	**2.200**	**0.0084**

		[rs4974081-rs3742883]	4.99	0.378	1.33	
		[rs6843524-rs1168666]	4.9	0.447	1.5	
		[rs6843524-rs1168666]	4.99	0.577	2	
		[rs6843524-rs1168666]	4.62	1.41	3	
		[rs6843524-rs895952]	4.9	1.63	5	
		[rs6605618-rs3742883]	4.53	1.13	2.5	
		[rs3748550-rs3820900]	4.72	0	1	
		[rs6843524-rs1168666]	4.9	1	3	
		[rs6843524-rs1168666]	5	1	2	
		[rs6843524-rs3742883]	4.9	0	1	
2	**5**	**[rs6843524-rs1168666]**	**4.882**	**0.888**	**2.233**	**0.0079**

3-locus						
		[rs4974081-rs3742883-rs2073951]	5.72	-0.816	0.5	
		[rs6843524-rs895952-rs16039]	5.77	0.447	1.5	
		[rs6843524-rs1168666-rs16039]	5.86	0.577	2	
		[rs3122428-rs4974081-rs2288242]	5.83	0	1	
		[rs6437368-rs6843524-rs303767]	5.69	-0.447	0.667	
		[rs4974081-rs3742883-rs3751526]	5.68	-0.378	0.75	

		[rs6587577-rs2236375-rs8096662]	5.77	0.378	1.33	
		[rs6843524-rs895952-rs16039]	5.77	1	3	
		[rs6843524-rs895952-rs16039]	5.73	1.41	3	
		[rs6437368-rs4974081-rs3742883]	5.77	0	1	
1	**3**	**[rs6843524-rs895952-rs16039]**	**5.759**	**0.954**	**1.475**	**0.0406**

		[rs4974081-rs3742883-rs6497671]	5.72	-0.816	0.5	
		[rs6843524-rs1168666-rs16039]	5.77	0.447	1.5	
		[rs6843524-rs895952-rs16039]	5.77	1	3	
		[rs6442895-rs4974081-rs1868159]	5.52	0.447	1.5	
		[rs6442905-rs6843524-rs895952]	5.55	1.34	4	
		[rs4974081-rs3742883-rs7180446]	5.68	-0.378	0.75	
		[rs3748550-rs3820900-rs6843524]	5.69	0.707	1.67	
		[rs6843524-rs1168666-rs16039]	5.77	1	3	
		[rs6843524-rs1168666-rs16039]	5.73	1.41	3	
		[rs4974081-rs10901065-rs3742883]	5.74	0	1	
2	**3**	**[rs6843524-rs1168666-rs16039]**	**5.759**	**0.954**	**1.992**	**0.0165**

[1] consistencia de validación cruzada

[2] estadístico MDR-PDT para entrenamiento

5.6 Interpretación de variantes

De esta manera, se designó un conjunto de SNPs a partir de aquellas reportadas en combinatorias de los mejores modelos MDR-PDT, teniendo en cuenta opcionalmente su importancia bajo la prueba TDT como relación individual frente a la enfermedad. Este conjunto final estaba compuesto por 26 SNPs localizados en 21 genes. Para cada muestra (sobre la población sin completitud de triada; 450 individuos) se reportó su cigocidad y estado de la enfermedad, de manera que el algoritmo SEN determinara la ganancia de información correspondiente para cada variante, conocido como efecto principal (ver **Tabla 5-8**) y cada posible combinación, como fuerza de interacción.

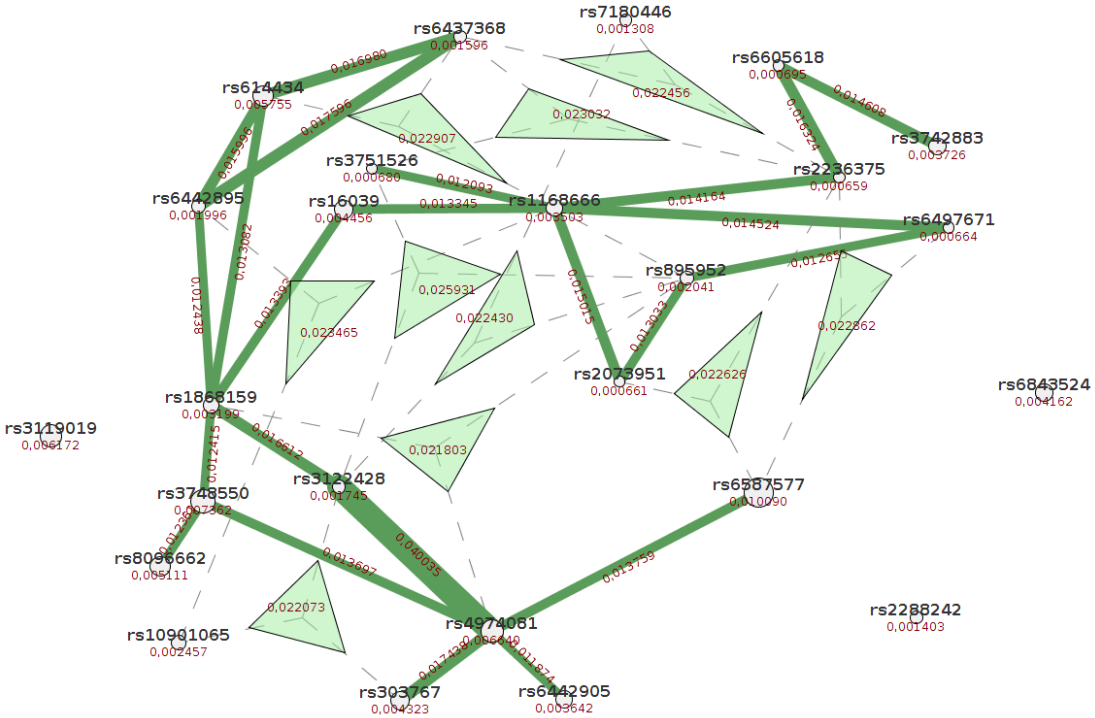
Tabla 5-8: Evaluación del efecto principal de las variantes significativas

Gen	SNP	Significancia (valor-P)	$I(V_A; \mathbf{T})$
SEC24D	rs6843524	0,0079 (MDR) | 0,003135 (TDT)	0,004162
SETD1B	rs895952	0,0084 (MDR)	0,002041
	rs1168666	0,0079 (MDR)	0,003503
ITPR1	rs6442905	0,0425 (MDR)	0,003642
	rs6442895	0,0165 (MDR)	0,001996

QRICH1	rs4974081	0,0179 (MDR) | 0,0003006 (TDT)	0,006640
ZFYVE26	rs3742883	0,0622 (MDR) | 0,0008112 (TDT)	0,003726
CA2	rs6605618	0,0297 (MDR)	0,000695
POGZ	rs3748550	0,0283 (MDR) | 0,002533 (TDT)	0,007362
	rs6587577	0,0143 (MDR) | 0,002838 (TDT)	0,010090
PTH2R	rs3820900	0,0109 (MDR)	0,001950
EARS2	rs2073951	0.0406 (MDR)	0,000661
	rs6497671	0,0165 (MDR)	0,000674
CACNA1A	rs16039	0,0165 (MDR)	0,004456
PPT1	rs3122428	0.0143 (MDR)	0,001745
TRPM1	rs2288242	0,0902 (MDR)	0,001403
KIF1A	rs6437368	0,0143 (MDR) | 0,003892 (TDT)	0,001596
SCN8A	rs303767	0,0406 (MDR)	0,004323
SPRED1	rs3751526	0,0406 (MDR)	0,000680
	rs7180446	0,0165 (MDR)	0,001308
DPH1	rs2236375	0,0250 (MDR)	0,000659
SETBP1	rs8096662	0,0250 (MDR)	0,005111
PRMT7	rs1868159	0,0165 (MDR)	0,003199
POMT1	rs10901065	0,0165 (MDR)	0,002457
SYNGAP1	rs3119019	0,0396 (MDR) | 0,003487 (TDT)	0,006172
FOXN1	rs614434	0,0179 (MDR) | 0,001616 (TDT)	0,005755

La información mutua entre variantes se puede entender ya sea como una sinergia o como una correlación respecto a la enfermedad. Como se puede ver en la **Figura 5-11**, las variantes que reportaron la mayor cantidad de información mutua entre ellas fueron rs4974081 y rs3122428 con un valor $\boldsymbol{I = 0,040035}$. En cuanto al efecto principal, las variantes que reportaron mayor asociación con la enfermedad por sí solas fueron rs6587577, rs3748550 y rs4974081, localizadas en los genes POGZ y QRICH1, y las cuales fueron reportadas como significativas en el análisis por TDT anteriormente. En el caso de información mutua entre tres variantes, no se encontró ninguna relevante.

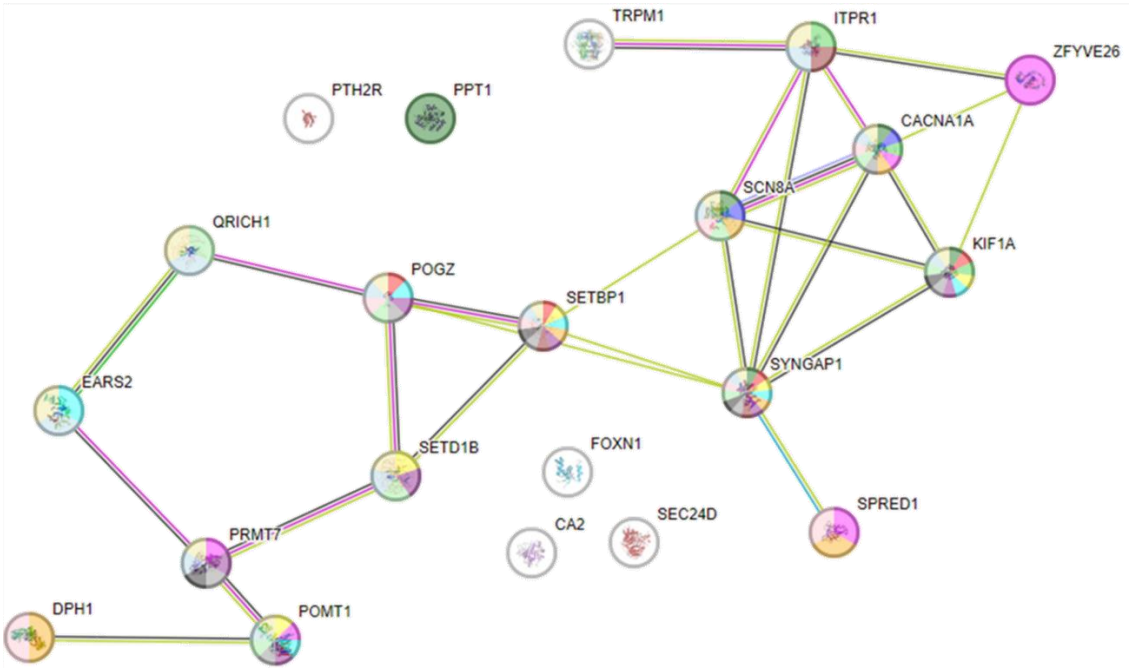
Figura 5-11: Ganancia de información de las variantes significativas



A partir de los resultados conjuntos entre los modelos MDR-PDT, la prueba TDT y la información mutua por SEN, se relacionaron las variantes con conocimientos biológicos, a partir de análisis de redes biológicas. Para llevar a cabo este proceso y a partir de la recomendación de selección de variantes (ver **Figura 5-11**), se asignaron las variantes significativas a sus genes y se relacionó su información y comportamientos por medio de STRING e IMP (Predicción integradora de múltiples especies).

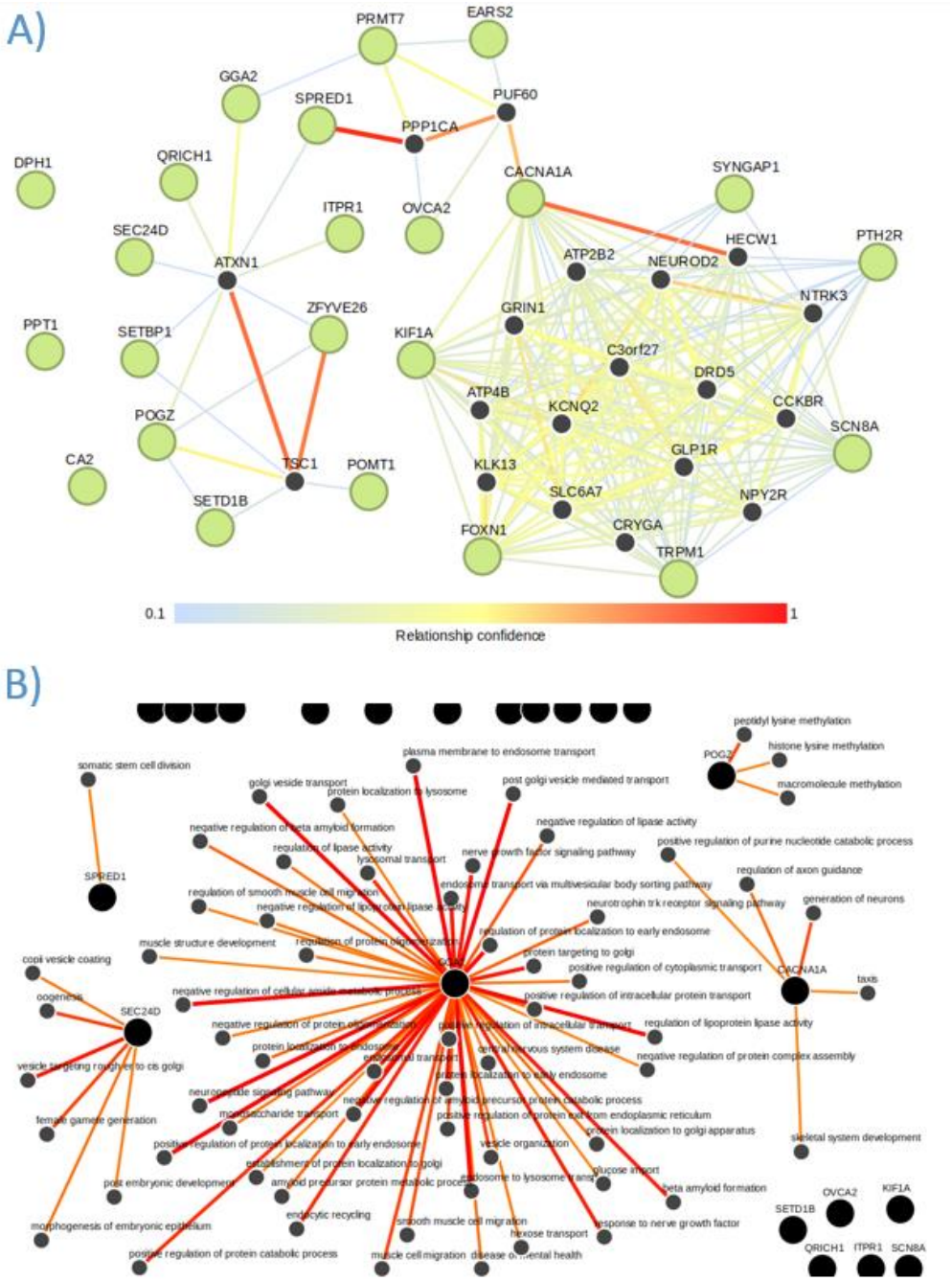
La relevancia biológica de los genes identificados se estimó mediante la evaluación de la interacción proteína-proteína (PPI) y las redes moleculares, mediante el análisis realizado con STRING, donde se encontró por medio de asociaciones funcionales y físicas una red (ver Figura 5-12) con 26 aristas (relaciones) de un esperado de 14 asociaciones, expresando un notable enriquecimiento de la interacción (valor-P enriquecimiento 0.00367). Así mismo, esta red se caracterizó por, un grado promedio de los nodos de 2.48, es decir, cada variante tiene en promedio 2 relaciones, y no se encontró una clara distinción de grupos, dado un coeficiente de agrupamiento local promedio de 0.359.

Figura 5-12: Análisis de enriquecimiento funcional STRING



Como análisis adicional se hizo uso de IMP para interpretar resultados en el contexto de un gran compendio de predicciones y redes funcionales entre organismos. Esta evaluación adicional fue motivada debido a que IMP identifica homólogos con roles funcionales conservados, permitiendo conocer anotaciones experimentales por medio de predicciones de funciones precisas. La selección de esta herramienta está principalmente fundamentada en que los conjuntos de genes pequeños se benefician particularmente de este análisis, debido a que puede ampliar el análisis con genes funcionalmente similares para mejorar la interpretación biológica (Wong et al., 2012). La ampliación de esta red permitió evidenciar que hay un alto coeficiente de asociación entre; el gen SPRED1 con PPP1CA, CACNA1A con HECW1 y ZFYVE26 con TSC1 (ver **Figura 5-13**).

Figura 5-13: Análisis de redes funcionales con IMP; enriquecimiento y procesos.



6. Discusión

En justificación del análisis de enriquecimiento derivado con STRING, en la **Tabla 6-1** se pueden observar los resultados de las anotaciones funcionales encontradas.

Tabla 6-1: Enriquecimientos funcionales en la red por medio de STRING

Categoría	ID termino	Descripción	Fuerza	FDR	Genes asociados	Color[1]
Enfermedad	DOID:0060307	Autosomal dominant non-syndromic intellectual disability	1.82	0.00055	POGZ,KIF1A,SYNGAP1,SETBP1	
Fenotipo	HP:0000750	Delayed speech and language development	1.15	4.03e-09	POGZ,ITPR1,PRMT7,SCN8A,POMT1,QRICH1,KIF1A,EARS2,SETD1B,CACNA1A,SYNGAP1,SETBP1	
Fenotipo	HP:0002463	Language impairment	1.14	4.03e-09	POGZ,ITPR1,PRMT7,SCN8A,POMT1,QRICH1,KIF1A,EARS2,SETD1B,CACNA1A,SYNGAP1,SETBP1	
Fenotipo	HP:0000729	Autistic behavior	1.21	3.20e-06	POGZ,SCN8A,POMT1,QRICH1,KIF1A,SETD1B,CACNA1A,SYNGAP1	
Fenotipo	HP:0001256	Intellectual disability, mild	1.27	1.11e-05	POGZ,PRMT7,POMT1,KIF1A,SETD1B,SYNGAP1,SETBP1	
Fenotipo	HP:0010864	Intellectual disability, severe	1.23	1.96e-05	POGZ,PRMT7,POMT1,KIF1A,CACNA1A,SYNGAP1,SETBP1	
Fenotipo	HP:0000752	Hyperactivity	1.19	2.97e-05	DPH1,POGZ,SPRED1,SCN8A,CACNA1A,SYNGAP1,SETBP1	
Fenotipo	HP:0001344	Absent speech	1.3	5.18e-05	POGZ,POMT1,KIF1A,EARS2,SYNGAP1,SETBP1	
Fenotipo	HP:0007018	Attention deficit hyperactivity disorder	1.3	5.34e-05	DPH1,SPRED1,SCN8A,CACNA1A,SYNGAP1,SETBP1	
Fenotipo	HP:0002187	Intellectual disability, profound	1.49	6.55e-05	POMT1,KIF1A,SETD1B,SYNGAP1,SETBP1	
Fenotipo	HP:0001328	Specific learning disability	1.34	0.00025	ZFYVE26,SPRED1,PRMT7,POMT1,CACNA1A	
Fenotipo	HP:0006855	Cerebellar vermis atrophy	1.89	0.00071	ITPR1,KIF1A,CACNA1A	

Fenotipo	HP:0002342	Intellectual disability, moderate	1.22	0.0052	PRMT7,KIF1A,SYNGAP1,SETBP1	
Fenotipo	HP:0033348	Epileptic aura	2.07	0.0081	SCN8A,CACNA1A	
Fenotipo	HP:0012433	Abnormal social behavior	1.24	0.0272	ITPR1,SYNGAP1,SETBP1	
Publicación	PMID:34469436	(2021) Next-generation sequencing in childhood-onset epilepsies: Diagnostic yield and impact on neuronal ceroid lipofuscinosis type 2 (CLN2) disease diagnosis.	2.09	0.0034	SCN8A,KIF1A,CACNA1A,PPT1,SYNGAP1	

[1] paleta de colores en la red de asociación de la

Figura 5-12

Como se mencionó en un principio, los trastornos del neurodesarrollo presentan una etiología bastante compleja, comprendida como una patología multifactorial, es decir, que está ocasionada por un efecto combinado (Vissers et al., 2016) o de interacción entre genes o proteínas (participes en el desarrollo cerebral o cognitivo), y factores exógenos y ambientales (complicaciones en el embarazo, desnutrición, polidactilia, etc). Es importante tener en cuenta que una evidencia robusta en los factores puede explicar la presencia de la enfermedad, como es el caso de una mutación reportada en la literatura como probablemente patogénica o patogénica (PP/P). Sin embargo, cuando no se encuentran mutaciones en el paciente, para esclarecer el diagnostico se puede describir a partir del efecto combinado entre los factores exógenos y las asociaciones de múltiples variantes con la enfermedad (“Large-Scale Discovery of Novel Genetic Causes of Developmental Disorders.,” 2015).

Teniendo en cuenta que este estudio está ligado a pruebas clínicas analizadas previamente, se evidencia que entre el 20%-30% de los casos suelen ser positivos, es decir, contaron con mutaciones PP/P reportadas en la literatura. Particularmente en la población elegible de este estudio, de los 346 casos, 87 fueron positivos (25,14%). De esta manera, el propósito de estos estudios de asociación es encontrar esas variantes SNV (no mutaciones) que puedan asociarse con la patología y así, ayuden junto con otros factores por efecto combinado, a explicar la presencia de la enfermedad cuando no se encuentre una mutación reportada. En base a lo anterior, se discuten las asociaciones halladas de manera individual (TDT en **Tabla 5-6**) y entre ellas (MDR-PDT en **Tabla 5-7**).

Desde el punto de partida del análisis de asociación directa SNP-Fenotipo, partiendo con los resultados del modelo TDT, se verificó la importancia de los SNP rs6587577, rs3748550 (en gen POGZ) y rs4974081 (en gen QRICH1) (valor-P 0.002838, 0.002533 y 0.0003006 respectivamente) a partir de su efecto principal reportado en SEN (0.010090, 0,007362 y 0,006640). Así mismo en el análisis de enriquecimiento se puede evidenciar que estos dos genes son conocidos por estar asociados en HPO con trastorno del lenguaje, comportamiento autista y retardo en el desarrollo del habla y el lenguaje (Baruch et al., 2021; Stessman et al., 2016), con FDR en el enriquecimiento de 6.55e-05, 3.20e-06 y 4.03e-09, respectivamente. También, en los modelos MDR-PDT (de 1-locus) estas variantes fueron reportadas con un valor-P de 0,0143 para rs6587577, 0,0283 para rs3748550 y 0,0179 para rs4974081.

A partir de las asociaciones SNP+SNP-Fenotipo generadas por los modelos MDR-PDT (2-locus), aunque no fue reportada ninguna asociación significativa, en los dos mejores modelos rs6843524-rs895952 y rs6843524-rs1168666, se destaca la variante rs6843524 localizada en el gen SEC24D con resultados de asociación individual significativos en TDT (valor-P 0.003135) y un efecto principal levemente destacable en SEN con 0,004162. La única relación obtenida para este gen es su asociación con ID y síndromes de epilepsia hereditarios. Sin embargo, la evidencia es baja a nivel de paneles. A pesar de esto, las variantes con las que se encuentra asociada; rs895952 y rs1168666, se localizan en el gen SETD1B, gen asociado con los mismos fenotipos de QRICH1 más Discapacidad Intelectual (ID) leve y profunda (Roston et al., 2021) (FDR 1.11e-05 y 6.55e-05).

De igual manera, en las asociaciones de tres variantes (3-locus) frente al fenotipo, resultantes de los modelos MDR-PDT, se resaltan rs6843524-rs1168666-rs16039 y rs3122428-rs4974081-rs2288242; con un valor-P en esta prueba de 0,0165 y 0,0902, respectivamente. Se destacan estos modelos de asociación debido a que presentaron un patrón similar en el modelo de entropía SEN, relacionándose rs1168666-rs16039 con una información mutua de 0,013345 y rs3122428-rs4974081 con 0,040035, esta última como la sinergia entre variantes más alta identificada. Nuevamente se destacan las variantes rs1168666 (en SETD1B) y rs4974081 (en QRICH1). Las nuevas variantes evidenciadas en las asociaciones son rs16039 y rs3122428, localizadas respectivamente en CACNA1A y PPT1. En los análisis de los enriquecimientos funcionales en la red, se encontró que PPT1 fue asociada por una variante CNV (c.451C>T) con Retardo en el Desarrollo (DD) y

enfermedad progresiva (Gall et al., 2021) y se vincula la neurodegeneración (UniProt). En cuanto a CACNA1A, se determinó como uno de los genes con mayores asociaciones a fenotipos derivados de TND (Kessi et al., 2023); además de estar incluido en el mismo reporte y con los mismos fenotipos asociados de PPT1, se relaciona con auras epilépticas, atrofias cerebrales, discapacidad de aprendizaje, ID severa, hiperactividad, y ADHD. De igual forma, IMP muestra que es un gen involucrado en la generación de neuronas y encargado de regular la guía y crecimiento de axones, así como en asociación con GRIN1 en la maduración del sistema nervioso central y el rombencéfalo.

En relación con lo antes expuesto y los resultados obtenidos, se destacan como posibles variantes significativas de asociación ($p<0.001$) con la patología las variantes rs4974081, y rs3742883 de manera individual por TDT, y en cuanto a epistasis se hallaron las interacciones rs6843524-rs895952, y rs6843524-rs1168666 como potenciales ($p<0.05$; XV>=5) con la observación de requerir una revisión más exhaustiva para validar su significancia.

La variante rs4974081 está ubicada en el gen QRICH1 (Glutamine-rich protein 1; MIM# 617387), aunque la función de la expresión de esté ha sido desconocida, un estudio reciente le atribuye el control por proteostasis de la respuesta desplegada del estrés del retículo endoplásmico (RE), y en consecuencia, un estudio sugirió que las variantes de QRICH1 pueden conducir a características del desarrollo neurológico debido la desregulación de las respuestas al estrés del RE, donde el deterioro de la vía secretora puede afectar la formación o función de las sinapsis (Kumble et al., 2022). Por otro lado, el gen ZFYVE26 (Zinc finger FYVE domain-containing protein 26; MIM# 612012) que contiene a la variante rs3742883, en su expresión actúa como regulador de la abscisión en el proceso de citocinesis y hace parte del adaptador-relacionado con el complejo proteico 5 (AP5), significativo en la reformación de lisosomas autofágicos, y en el cual han sido reportadas más de 30 mutaciones, causales de la paraplejia espástica tipo 15, condición autosómica recesiva caracterizada por problemas progresivos de movimiento, discapacidad intelectual y de visión (Vantaggiato et al., 2014).

En cuanto a las interacciones destacadas, rs6843524-rs895952 y rs6843524-rs1168666 tienen en común que ocurren en los mismos dos genes SEC24D (SEC24 homolog D; MIM:607186) - SETD1B (SET domain containing 1B; MIM:611055). SEC24D hace parte

del complejo de proteínas de cubierta vesicular (COPII) que promueve la formación de vesículas de transporte desde RE, deformando la membrana en vesículas y la selección para su trasporte al aparato de Golgi. De manera similar a QRICH1, un estudio demostró que las neuronas postmitóticas son extremadamente sensibles a la pérdida de SEC24C, está perdida condicional en progenitores neuronales durante la embriogénesis causó microcefalias y muerte celular apoptótica de neuronas posmitóticas en la corteza cerebral, actividades que pueden sustentar los reportes de mutaciones en este gen con DD, como el caso de rs730882211 (Alazami et al., 2015). En SETD1B (una histona metiltransferasa) se conoce que está involucrado en la producción histona trimetilada H3 en Lys4 a partir de catalizar la transferencia del grupo metilo de la S-adenosil-L-metionina, así mismo, se ha demostrado que controla epigenéticamente la expresión génica y el estado de la cromatina, debido al enriquecimiento en los sitios promotores (H3K4me3) y potenciadores (H3K4me1 y H3K4me2) (Abay-Nørgaard et al., 2020). Sin embargo, a fin de comprender la interacción entre las variantes, no se encuentra relación funcional (como se puede ver en **Figura 5-13**) entre la expresión de estos dos genes.

Por último, el objetivo de este estudio es el análisis de asociación entre variantes, partiendo de un conjunto de genes que han reportado causalidad sobre los trastornos del neurodesarrollo (patogénico) o al menos una sospecha (VUS). Bajo esta finalidad la mayoría de los análisis se han centrado en SNP individuales o aditivos, pasando por alto el potencial de la epistasis o no-aditivo, debido a limitaciones metodológicas (Nodzenski et al., 2022). Como limitación adicional, como se planteó en el marco teórico, pocos de los métodos son aplicables a estudios familiares que evalúen la transmisión de genotipos. Los métodos conocidos para este paradigma son MDR-PDT (Martin et al., 2006), FAM-MDR (Cattaert et al., 2010), PG-MDR (Lou et al., 2008), como adaptaciones del método más notable en el área y GADGETS (Nodzenski et al., 2022), una aproximación reciente basada en algoritmos genéticos.

Dicho lo anterior y dadas las limitantes, se analizó el desempeño de estos modelos en otros trastornos o fenotipo. Newman y colaboradores (Newman et al., 2023), a partir de la observación de una interacción bioquímica y efectos funcionales en transcripción entre UBASH3A y PTPN22, confirmaron a partir de MDR-PDT ($p=0.001$; análisis de 2-SNP) que la interacción genética entre estos dos loci contribuye a la diabetes tipo 1. Para el caso de MDR-PDT aplicado a estudios de enfermedades mentales, en un trabajo sobre

esquizofrenia, los análisis revelaron para la cohorte del Estudio Irlandés de Familias con Esquizofrenia de Alta Densidad (ISHDSF), un modelo significativo de 3 locus (p=0,003) incluyendo IL3 SNP rs2069803, DTNBP1 SNP rs2619539 y RGS4 SNP rs2661319, así como una interacción significativa de 2 locus entre IL3 SNP rs31400 y DTNBP1 SNP rs760761 (p=0,019) para los datos del Estudio Irlandés de Casos y Controles sobre Esquizofrenia (ICCSS) (Edwards et al., 2008). Así mismo en alzheimer, este modelo detectó el efecto principal de APOE en tres conjuntos de datos diferentes, hallándolo en el mejor modelo con consistencia CV perfecta (5/5) y siendo estadísticamente significativo (P<0,05) (Thornton-Wells et al., 2008).

En la aplicación de otros modelos familiares, Mascheretti y colaboradores basados en el algoritmo de Li Qing (Q. Li et al., 2010), generaron una implementación de regresión lineal para estudios trio (Mascheretti et al., 2015), donde reportaron asociaciones significativas (p<0.001) para dislexia entre DYX1C1 (1259C/G) y GRIN2B (rs2268119A/T). Sobre el mismo algoritmo, para labio leporino y paladar hendido se encontró una asociación entre rs17252114 (CTNNB1) y rs1274944 (ACTN1) con una significancia de p=0,0002 (Liu et al., 2019). Otros métodos novedosos, como GADGETS (Nodzenski et al., 2022), GCORE (Sung et al., 2016) y EPISFA (Xiang et al., 2020) no reportan aplicaciones. Por otro lado, las variantes encontradas en este estudio no han sido informadas en otros estudios de epistasis, caso-control o familiar.

7. Conclusiones

En el presente estudio se llevó a cabo una metodología basada en varios marcos de trabajo que permitió realizar el procesamiento y análisis de variantes de un solo nucleótido (SNV), interpretando sus interacciones en asociación a la enfermedad de discapacidad intelectual y retraso en el neurodesarrollo (DD/ID). Las variantes analizadas se derivaron de una cohorte de pacientes colombianos con un estudio de exoma de triada familiar, realizado en un periodo de tiempo entre enero de 2018 hasta diciembre de 2021 y para aquellos con indicación clínica de DD/ID. El estudio desarrollado con enfoque sobre el modelo MDR-PDT, con verificación de la prueba TDT y el algoritmo SEN por ganancia de información, permitió identificar la asociación individual de variantes, así como interacciones no significativas entre varios SNPs, respecto a la enfermedad, validadas a partir de análisis de enriquecimiento a nivel de genes.

Previamente al análisis de asociación con los tres algoritmos, se realizaron varios procesos dispendiosos sobre los datos VCF retrospectivos; procedimientos de recolección, integración, anotación y control de calidad. El hecho de partir por exomas completos, sin un panel puntual de variantes o genes, implicó la evaluación minuciosa de los genotipos de las variantes candidatas; variantes procedentes de un proceso de revisión de la literatura, finalizando con un conjunto de genes candidatos, que posteriormente fueron identificados en el VCF integrado, anotando todas las variantes SNV localizadas en estos.

Se realizaron varios procesos de control de calidad sobre el conjunto de variantes potenciales. Entre los cuales se incluyeron: frecuencia alélica MAF para descartar variantes muy raras y por ende difíciles de detectar asociación; desviaciones de HWE para evitar llamados de genotipos erróneos como artefactos; heterocigocidad atípica para descartar errores de contaminación en las muestras; y ausencia de genotipos en las

muestras, para excluir variantes sin suficiente presencia o llamado genotípico en la población. Especialmente este último paso, de ausencia en los llamados genotípicos, implicó una reducción sustancial del 99,14% en el número de variantes, esto debido a la manera como fueron sustraídas las variantes candidatas de las muestras.

Finalmente, a pesar de la significancia de los modelos MDR-PDT no fue muy importante, se logró identificar por medio de las otras pruebas y validar desde la interpretación biológica varias variantes; entre ellas se destacan rs1168666, en el gen SETD1B y rs4974081 en QRICH1, las cuales fueron evidencia en varios modelos, evidenciando una fuerte asociación en el riesgo de padecer DD/ID. De la misma manera, rs16039 en el gen CACNA1A sobresale tanto en las significancias, como en los análisis de enriquecimiento debido a la relevancia que ha tenido en los últimos años este gen frente a los trastornos del neurodesarrollo. En consecuencia, se concluye que es esencial explorar más a fondo este tipo de análisis de epistasis en enfermedades con etiologías complejas, así como, conectar su evaluación con datos adicionales, como paraclínicos, datos de laboratorio, y datos del entorno en cuanto a estilos de vida, para llegar a conclusiones más determinantes.

Además de lo anterior, se evidencia que culturalmente y en la particularidad de evaluar la asociación por interacción entre variantes, epistasis, este es el primer estudio sobre una cohorte de pacientes pediátricos. Sin embargo, se sugiere en estudios futuros realizar análisis con una población más grande, tanto a nivel de número de familias como en la inclusión de hermanos discordantes; determinar las variantes candidatas desde la secuenciación o previamente al alineamiento en caso de considerar todo el exoma; y por supuesto, encaminar esfuerzos en el desarrollo de nuevos algoritmos de epistasis centrados en la inteligencia artificial, para los casos de estudio familiar.

A. Anexo: MDR-PDT mejores cinco modelos de asociaciones entre SNPs para uno, dos y tres locus

Rango	XV Cons[1]	Modelo	T-Train[2]	T-Test	OR	P-value
1-locus						
		[rs6843524]	4.01	1	0	
		[rs6843524]	3.9	1.41	0	
		[rs6843524]	4	1	3	
		[rs6843524]	3.78	1.63	5	
		[rs6843524]	4.32	0	1	
		[rs6843524]	3.33	2.65	0	
		[rs6843524]	3.79	1.73	0	
		[rs6843524]	3.89	1.34	4	
		[rs6843524]	3.89	1.34	4	
		[rs6843524]	4.11	0.816	2	
1	**10**	**[rs6843524]**	**3.905**	**1.292**	**1.900**	**0.0151**
		[rs3742883]	3.7	-0.447	0.667	
		[rs6587577]	3.1	0.632	1.5	
		[rs3742883]	3.28	0.816	2	
		[rs4974081]	3.32	0	1	
		[rs3742883]	3.62	-0.577	0.5	
		[rs614434]	3.32	-0.447	0.667	
		[rs6587577]	3.25	0	1	
		[rs3742883]	3.32	0.577	2	
		[rs4974081]	3.13	1	0	
		[rs3742883]	3.7	-0.447	0.667	
2	**5**	**[rs3742883]**	**3.524**	**-0.016**	**1.000**	**0.0622**
		[rs4974081]	2.97	1.34	4	
		[rs3748550]	3.1	0.632	1.5	

		[rs3748550]	3.25	0	1	
		[rs6587577]	2.84	1.34	4	
		[rs614434]	3.13	0	1	
		[rs4974081]	3.32	0	1	
		[rs3748550]	3.25	0	1	
		[rs4974081]	3.24	0.707	1.67	
		[rs614434]	3.09	0.378	1.33	
		[rs4974081]	3.36	0.378	1.33	
3	**4**	**[rs4974081]**	**3.223**	**0.607**	**1.783**	**0.0179**
		[rs6587577]	2.94	1	3	
		[rs6437368]	3.05	-1.13	0.4	
		[rs614434]	3.2	0	1	
		[rs3748550]	2.84	1.34	4	
		[rs3122428]	2.84	-1.63	0.2	
		[rs3742883]	3.09	1.34	4	
		[rs4974081]	3.09	1	3	
		[rs6587577]	3.15	0.378	1.33	
		[rs7997]	2.95	-0.447	0.667	
		[rs6587577]	3.05	0.707	1.67	
4	**3**	**[rs6587577]**	**3.048**	**0.695**	**1.927**	**0.0143**
		[rs3748550]	2.94	1	3	
		[rs4974081]	2.97	1.34	4	
		[rs6587577]	3.1	0.447	1.5	
		[rs7997]	2.71	0	1	
		[rs6437368]	2.73	-0.447	0.667	
		[rs7997]	2.71	0.378	1.33	
		[rs3742883]	2.9	2	0	
		[rs3748550]	3.15	0.378	1.33	
		[rs3742883]	2.9	2	0	
		[rs3748550]	2.89	1.13	2.5	
5	**3**	**[rs3748550]**	**2.995**	**0.837**	**1.533**	**0.0283**
2-locus						
		[rs6843524-rs895952]	4.99	1.41	0	
		[rs6843524-rs895952]	5.08	0.447	1.5	
		[rs6843524-rs895952]	5.08	1	3	
		[rs6843524-rs895952]	4.81	1.41	3	
		[rs6442905-rs6843524]	4.92	-0.577	0.5	
		[rs4974081-rs3742883]	4.64	1.34	4	

		[rs6843524-rs895952]	4.81	1.41	3	
		[rs6843524-rs895952]	5.08	1	3	
		[rs6843524-rs895952]	5	1.41	3	
		[rs4974081-rs3742883]	5.08	0	1	
1	**7**	**[rs6843524-rs895952]**	**4.979**	**1.158**	**2.200**	**0.0084**
		[rs4974081-rs3742883]	4.99	0.378	1.33	
		[rs6843524-rs1168666]	4.9	0.447	1.5	
		[rs6843524-rs1168666]	4.99	0.577	2	
		[rs6843524-rs1168666]	4.62	1.41	3	
		[rs6843524-rs895952]	4.9	1.63	5	
		[rs6605618-rs3742883]	4.53	1.13	2.5	
		[rs3748550-rs3820900]	4.72	0	1	
		[rs6843524-rs1168666]	4.9	1	3	
		[rs6843524-rs1168666]	5	1	2	
		[rs6843524-rs3742883]	4.9	0	1	
2	**5**	**[rs6843524-rs1168666]**	**4.882**	**0.888**	**2.233**	**0.0079**
		[rs2692185-rs3742883]	4.87	-0.447	0.667	
		[rs6437368-rs2274424]	4.52	-0.577	0.5	
		[rs6605618-rs3742883]	4.53	1.13	2.5	
		[rs4974081-rs3119019]	4.52	-0.577	0.5	
		[rs1838846-rs6843524]	4.8	0	1	
		[rs6843524-rs895952]	4.43	2.83	0	
		[rs6587577-rs2236375]	4.71	-0.378	0.75	
		[rs4974081-rs3742883]	4.81	1	2	
		[rs6843524-rs7997]	4.81	0	1	
		[rs6843524-rs895952]	4.81	1.89	6	
3	**2**	**[rs6843524-rs895952]**	**4.618**	**2.359**	**1.492**	**0.0297**
		[rs6843524-rs1168666]	4.72	1.73	0	
		[rs6437368-rs7997]	4.52	-0.577	0.5	
		[rs6843524-rs1868159]	4.46	0.577	2	
		[rs6843524-rs4779817]	4.44	0.447	1.5	
		[rs2121371-rs6843524]	4.8	0	1	
		[rs154001-rs614434]	4.43	-0.816	0.5	
		[rs6843524-rs1168666]	4.62	1.41	3	
		[rs6843524-rs3742883]	4.46	1.34	4	
		[rs6843524-rs6781]	4.63	0	1	
		[rs2692185-rs3742883]	4.74	-0.577	0.5	

4	**2**	**[rs6843524-rs1168666]**	**4.670**	**1.573**	**1.400**	**0.0341**
		[rs6605618-rs3742883]	4.71	0.447	1.5	
		[rs1838846-rs6843524]	4.44	0	1	
		[rs1838846-rs6843524]	4.46	1	3	
		[rs6843524-rs4779816]	4.44	0.447	1.5	
		[rs6843524-rs154001]	4.77	-0.577	0.5	
		[rs4974081-rs1868159]	4.32	-1	0	
		[rs3748550-rs2236375]	4.53	-0.378	0.75	
		[rs2692185-rs3742883]	4.43	0.577	2	
		[rs4974081-rs3742883]	4.56	1.73	0	
		[rs6442905-rs6843524]	4.63	0.816	2	
5	**2**	**[rs1838846-rs6843524]**	**4.449**	**0.500**	**1.225**	**0.0455**
		3-locus				
		[rs4974081-rs3742883-rs2073951]	5.72	-0.816	0.5	
		[rs6843524-rs895952-rs16039]	5.77	0.447	1.5	
		[rs6843524-rs1168666-rs16039]	5.86	0.577	2	
		[rs3122428-rs4974081-rs2288242]	5.83	0	1	
		[rs6437368-rs6843524-rs303767]	5.69	-0.447	0.667	
		[rs4974081-rs3742883-rs3751526]	5.68	-0.378	0.75	
		[rs6587577-rs2236375-rs8096662]	5.77	0.378	1.33	
		[rs6843524-rs895952-rs16039]	5.77	1	3	
		[rs6843524-rs895952-rs16039]	5.73	1.41	3	
		[rs6437368-rs4974081-rs3742883]	5.77	0	1	
1	**3**	**[rs6843524-rs895952-rs16039]**	**5.759**	**0.954**	**1.475**	**0.0406**
		[rs4974081-rs3742883-rs6497671]	5.72	-0.816	0.5	
		[rs6843524-rs1168666-rs16039]	5.77	0.447	1.5	
		[rs6843524-rs895952-rs16039]	5.77	1	3	
		[rs6442895-rs4974081-rs1868159]	5.52	0.447	1.5	
		[rs6442905-rs6843524-rs895952]	5.55	1.34	4	
		[rs4974081-rs3742883-rs7180446]	5.68	-0.378	0.75	
		[rs3748550-rs3820900-rs6843524]	5.69	0.707	1.67	
		[rs6843524-rs1168666-rs16039]	5.77	1	3	
		[rs6843524-rs1168666-rs16039]	5.73	1.41	3	
		[rs4974081-rs10901065-rs3742883]	5.74	0	1	
2	**3**	**[rs6843524-rs1168666-rs16039]**	**5.759**	**0.954**	**1.992**	**0.0165**
		[rs4974081-rs3742883-rs1468138]	5.72	-0.816	0.5	

		[rs3748550-rs6843524-rs895952]	5.69	0	1	
		[rs6843524-rs1168666-rs11070320]	5.55	-1.41	0	
		[rs4974081-rs4600135-rs9318917]	5.51	0.447	1.5	
		[rs6437368-rs30612-rs3742883]	5.52	0.577	2	
		[rs4974081-rs3742883-rs7182445]	5.68	-0.378	0.75	
		[rs6587577-rs3742883-rs2236375]	5.61	-1	0.333	
		[rs4974081-rs10747050-rs3742883]	5.63	0.707	1.67	
		[rs3122428-rs4974081-rs2288242]	5.67	NA	0	
		[rs4974081-rs3742883-rs3020959]	5.7	-0.378	0.75	
3	**1**	**[rs3122428-rs4974081-rs2288242]**	**5.667**	**NA**	**0.850**	**0.0902**
		[rs4974081-rs3119019-rs3742883]	5.72	0.378	1.33	
		[rs6587577-rs6843524-rs895952]	5.6	0	1	
		[rs6442905-rs6843524-rs895952]	5.55	1.34	4	
		[rs6843524-rs1168666-rs11621299]	5.48	1	2	
		[rs3122428-rs4974081-rs2288242]	5.49	1.41	0	
		[rs3742883-rs13078-rs7180446]	5.61	-1.13	0.4	
		[rs6587577-rs572126-rs2236375]	5.57	-0.816	0.5	
		[rs4974081-rs10870196-rs3742883]	5.63	0.707	1.67	
		[rs6442905-rs6843524-rs895952]	5.66	1.13	2.5	
		[rs4974081-rs3119019-rs3742883]	5.7	0	1	
4	**2**	**[rs6442905-rs6843524-rs895952]**	**5.602**	**1.238**	**1.440**	**0.0425**
		[rs6843524-rs895952-rs16039]	5.66	1.41	0	
		[rs6587577-rs2236375-rs8096662]	5.57	1.13	2.5	
		[rs6843524-rs1168666-rs3213696]	5.53	0	1	
		[rs4974081-rs1017361-rs9318917]	5.43	0.447	1.5	
		[rs6843524-rs1168666-rs11621299]	5.46	1.34	4	
		[rs3742883-rs13078-rs7182445]	5.61	-1.13	0.4	
		[rs3748550-rs2236375-rs8096662]	5.57	0.378	1.33	
		[rs3748550-rs6843524-rs895952]	5.6	1	3	
		[rs6843524-rs1168666-rs11621299]	5.57	1.13	2.5	
		[rs6587577-rs3742883-rs2236375]	5.66	0.378	1.33	
5	**2**	**[rs6843524-rs1168666-rs11621299]**	**5.516**	**1.238**	**1.757**	**0.0250**

[1] consistencia de validación cruzada

[2] estadístico MDR-PDT para entrenamiento

B. Anexo: Prueba de transmisión/desequilibrio X^2

CHR	SNP	BP	A1	A2	T	U	OR	L95	U95	CHISQ	P
3	rs4974081	49070499	C	T	16	44	0.3636	0.2052	0.6444	13.07	0.0003006
14	rs3742883	68234539	T	C	19	46	0.413	0.242	0.7049	11.22	0.0008112
17	rs614434	26851501	G	A	21	47	0.4468	0.2671	0.7474	9.941	0.001616
1	rs3748550	151384733	G	A	29	57	0.5088	0.3254	0.7956	9.116	0.002533
1	rs6587577	151402045	A	G	30	58	0.5172	0.3329	0.8037	8.909	0.002838
4	rs6843524	119736598	C	T	45	21	2.143	1.277	3.597	8.727	0.003135
6	rs3119019	33414637	G	T	7	23	0.3043	0.1306	0.7093	8.533	0.003487
2	rs6437368	241659368	C	A	25	50	0.5	0.3094	0.8081	8.333	0.003892
8	rs1784468	107754583	T	G	11	27	0.4074	0.2021	0.8213	6.737	0.009444
6	rs3119027	33414626	T	C	8	22	0.3636	0.1619	0.8168	6.533	0.01059
4	rs1801212	6302519	G	A	30	53	0.566	0.3617	0.8858	6.373	0.01158
1	rs3122428	40545964	T	G	35	59	0.5932	0.3905	0.9012	6.128	0.01331
16	rs1868159	68387496	A	C	30	51	0.5882	0.3747	0.9234	5.444	0.01963
18	rs8096662	42533130	A	G	10	23	0.4348	0.2069	0.9135	5.121	0.02364
23	rs3112299	40456961	G	C	0	5	0	0	NA	5	0.02535
23	rs3020959	152825414	G	T	0	5	0	0	NA	5	0.02535
2	rs3820900	209358027	C	T	25	43	0.5814	0.3551	0.9518	4.765	0.02905
9	rs4744605	73213610	A	G	22	39	0.5641	0.3345	0.9513	4.738	0.02951
11	rs589623	103082590	G	A	22	39	0.5641	0.3345	0.9513	4.738	0.02951
1	rs4971053	155408635	T	C	1	7	0.1429	0.01758	1.161	4.5	0.03389
15	rs2288242	31330280	A	G	22	38	0.5789	0.3425	0.9788	4.267	0.03887
17	rs3213696	2573652	C	T	11	23	0.4783	0.2331	0.9811	4.235	0.03959
5	rs30612	14420027	A	C	28	45	0.6222	0.3882	0.9973	3.959	0.04662
2	rs7601520	166893081	G	A	34	52	0.6538	0.4244	1.007	3.767	0.05226
2	rs2020318	166894230	T	C	34	52	0.6538	0.4244	1.007	3.767	0.05226
2	rs2126152	166896143	T	G	34	52	0.6538	0.4244	1.007	3.767	0.05226
2	rs6432860	166897864	A	G	34	52	0.6538	0.4244	1.007	3.767	0.05226
2	rs1461193	166904346	G	A	34	52	0.6538	0.4244	1.007	3.767	0.05226
6	rs1129644	24653376	G	A	31	48	0.6458	0.4111	1.014	3.658	0.05579

3	rs9881951	123014877	C	T	29	45	0.6444	0.4041	1.028	3.459	0.06289
3	rs4271897	4687548	A	G	18	31	0.5806	0.3248	1.038	3.449	0.06329
2	rs6753355	166900606	A	C	34	51	0.6667	0.432	1.029	3.4	0.0652
7	rs1637841	148080647	T	G	26	41	0.6341	0.388	1.037	3.358	0.06687
9	rs6781	130698043	A	G	41	26	1.577	0.9647	2.578	3.358	0.06687
13	rs9318917	25466774	T	C	26	41	0.6341	0.388	1.037	3.358	0.06687
10	rs10733757	61819049	T	G	34	50	0.68	0.4398	1.051	3.048	0.08086
13	rs7330016	40233691	C	T	35	51	0.6863	0.4463	1.055	2.977	0.08447
13	rs1530876	25458650	G	A	26	40	0.65	0.3967	1.065	2.97	0.08484
8	rs6605618	86389586	A	C	7	15	0.4667	0.1903	1.145	2.909	0.08808
2	rs2121371	166170127	T	C	34	49	0.6939	0.448	1.075	2.711	0.09967
2	rs1838846	166172317	A	G	34	49	0.6939	0.448	1.075	2.711	0.09967
11	rs1629304	71174589	T	C	5	1	5	0.5842	42.8	2.667	0.1025
2	rs11124182	240085403	C	A	31	45	0.6889	0.436	1.088	2.579	0.1083
14	rs13078	95556747	A	T	22	34	0.6471	0.3785	1.106	2.571	0.1088
2	rs6706656	228194570	G	A	15	25	0.6	0.3163	1.138	2.5	0.1138
14	rs8022395	21871653	C	T	30	43	0.6977	0.4377	1.112	2.315	0.1281
2	rs2646159	175613600	T	C	18	10	1.8	0.8309	3.899	2.286	0.1306
11	rs1792268	71146952	G	A	36	50	0.72	0.4691	1.105	2.279	0.1311
11	rs1792265	71149856	C	T	36	50	0.72	0.4691	1.105	2.279	0.1311
15	rs7180446	38631930	C	A	26	38	0.6842	0.4155	1.127	2.25	0.1336
3	rs6442895	4693937	G	C	29	41	0.7073	0.4396	1.138	2.057	0.1515
11	rs949177	71152461	A	G	29	41	0.7073	0.4396	1.138	2.057	0.1515
11	rs1790334	71155153	A	G	29	41	0.7073	0.4396	1.138	2.057	0.1515
11	rs760241	71146691	A	G	35	48	0.7292	0.4717	1.127	2.036	0.1536
1	rs2077360	11848879	A	G	2	6	0.3333	0.06728	1.652	2	0.1573
3	rs4482616	123018963	A	G	30	42	0.7143	0.4471	1.141	2	0.1573
11	rs896818	47364762	A	G	2	6	0.3333	0.06728	1.652	2	0.1573
12	rs2019743	7054859	G	A	6	2	3	0.6055	14.86	2	0.1573
12	rs1168666	122243905	C	T	36	25	1.44	0.8645	2.399	1.984	0.159
17	rs8071749	57725043	T	G	31	43	0.7209	0.4543	1.144	1.946	0.163
2	rs7569946	60687959	A	G	37	50	0.74	0.4838	1.132	1.943	0.1634
3	rs6442905	4817057	T	C	9	4	2.25	0.6929	7.306	1.923	0.1655
6	rs109836	143091263	A	G	4	9	0.4444	0.1369	1.443	1.923	0.1655
17	rs8065248	57750982	T	C	32	44	0.7273	0.4613	1.147	1.895	0.1687
15	rs7182445	38614525	G	A	27	38	0.7105	0.4338	1.164	1.862	0.1724
15	rs3751526	38643574	T	C	27	38	0.7105	0.4338	1.164	1.862	0.1724
8	rs2843740	37985897	A	G	22	32	0.6875	0.3995	1.183	1.852	0.1736
12	rs216045	2760708	G	A	33	45	0.7333	0.468	1.149	1.846	0.1742
12	rs895952	122248582	C	G	33	23	1.435	0.8425	2.443	1.786	0.1814
9	rs7997	130698029	G	C	40	29	1.379	0.8552	2.225	1.754	0.1854

1	rs4846051	11854457	G	A	5	10	0.5	0.1709	1.463	1.667	0.1967
11	rs688094	103029373	G	A	5	10	0.5	0.1709	1.463	1.667	0.1967
11	rs585692	103047007	C	A	5	10	0.5	0.1709	1.463	1.667	0.1967
11	rs949176	71149069	A	G	31	42	0.7381	0.464	1.174	1.658	0.1979
15	rs1493652	101112342	T	C	31	42	0.7381	0.464	1.174	1.658	0.1979
15	rs11074121	93521604	A	G	20	29	0.6897	0.3901	1.219	1.653	0.1985
2	rs2592591	121729490	C	T	8	14	0.5714	0.2397	1.362	1.636	0.2008
2	rs2592590	121729686	T	C	8	14	0.5714	0.2397	1.362	1.636	0.2008
3	rs9855969	123047666	A	G	27	37	0.7297	0.4443	1.198	1.562	0.2113
2	rs3106962	121730138	C	G	9	15	0.6	0.2626	1.371	1.5	0.2207
2	rs2592595	121726447	G	A	10	16	0.625	0.2836	1.377	1.385	0.2393
2	rs7585225	240029736	T	C	28	20	1.4	0.7887	2.485	1.333	0.2482
22	rs5760030	24145395	A	G	15	22	0.6818	0.3537	1.314	1.324	0.2498
22	rs5751738	24145675	G	C	15	22	0.6818	0.3537	1.314	1.324	0.2498
1	rs6659553	46655158	T	C	2	5	0.4	0.07761	2.062	1.286	0.2568
3	rs9844268	4859725	G	A	2	5	0.4	0.07761	2.062	1.286	0.2568
15	rs4778245	28357230	T	C	2	5	0.4	0.07761	2.062	1.286	0.2568
15	rs4777502	72637795	T	C	2	5	0.4	0.07761	2.062	1.286	0.2568
15	rs1800431	72638892	T	C	2	5	0.4	0.07761	2.062	1.286	0.2568
5	rs154001	127685135	C	T	34	44	0.7727	0.4939	1.209	1.282	0.2575
8	rs4874160	144671244	C	A	35	45	0.7778	0.5001	1.21	1.25	0.2636
11	rs532439	66469032	T	C	28	37	0.7568	0.4632	1.236	1.246	0.2643
17	rs2236375	1943888	T	C	29	38	0.7632	0.4707	1.237	1.209	0.2715
11	rs2282619	71183645	T	C	37	47	0.7872	0.5117	1.211	1.19	0.2752
11	rs2276354	71185479	T	C	37	47	0.7872	0.5117	1.211	1.19	0.2752
11	rs2276353	71189436	T	C	37	47	0.7872	0.5117	1.211	1.19	0.2752
20	rs2295765	31019024	C	T	23	31	0.7419	0.4326	1.272	1.185	0.2763
9	rs2274424	130699917	G	A	39	30	1.3	0.8077	2.092	1.174	0.2786
11	rs3794065	71188502	G	A	38	48	0.7917	0.5173	1.212	1.163	0.2809
7	rs2692185	146997203	T	A	32	41	0.7805	0.4916	1.239	1.11	0.2922
10	rs7075340	112356331	G	A	19	26	0.7308	0.4045	1.32	1.089	0.2967
17	rs2740349	648498	C	T	34	43	0.7907	0.5043	1.24	1.052	0.3051
17	rs2740348	649935	G	C	34	43	0.7907	0.5043	1.24	1.052	0.3051
15	rs4777755	93510603	A	G	20	27	0.7407	0.4155	1.321	1.043	0.3072
17	rs7216804	1945041	A	G	27	35	0.7714	0.4669	1.274	1.032	0.3096
4	rs5522	149357475	C	T	36	45	0.8	0.5161	1.24	1	0.3173
15	rs2241493	31362352	C	T	37	29	1.276	0.7847	2.074	0.9697	0.3248
22	rs5760032	24145727	C	T	38	47	0.8085	0.5272	1.24	0.9529	0.329
11	rs2186778	71185518	T	C	40	49	0.8163	0.5376	1.24	0.9101	0.3401
5	rs28580074	176721198	T	C	33	41	0.8049	0.5089	1.273	0.8649	0.3524
1	rs905389	18021406	T	C	25	32	0.7812	0.463	1.318	0.8596	0.3538

14	rs2119703	88453424	G	A	32	25	1.28	0.7586	2.16	0.8596	0.3538
10	rs2039874	112362019	G	T	18	24	0.75	0.4071	1.382	0.8571	0.3545
2	rs4669338	8916847	A	T	4	7	0.5714	0.1673	1.952	0.8182	0.3657
3	rs6444960	170825905	C	G	4	7	0.5714	0.1673	1.952	0.8182	0.3657
12	rs967648	12618715	T	C	4	7	0.5714	0.1673	1.952	0.8182	0.3657
12	rs2037743	79611374	C	T	45	54	0.8333	0.5611	1.238	0.8182	0.3657
4	rs5525	149356516	A	G	36	44	0.8182	0.5267	1.271	0.8	0.3711
15	rs7495441	28359744	G	A	8	12	0.6667	0.2725	1.631	0.8	0.3711
19	rs273269	18279638	T	C	37	45	0.8222	0.5322	1.27	0.7805	0.377
1	rs2270978	18023365	C	T	21	27	0.7778	0.4397	1.376	0.75	0.3865
15	rs7494786	28414665	T	C	5	8	0.625	0.2045	1.91	0.6923	0.4054
4	rs10937714	6279047	C	T	32	39	0.8205	0.5141	1.31	0.6901	0.4061
7	rs160384	98558880	G	C	32	39	0.8205	0.5141	1.31	0.6901	0.4061
22	rs738797	24143502	C	G	24	30	0.8	0.4677	1.368	0.6667	0.4142
23	rs2071127	153129556	G	A	2	4	0.5	0.09158	2.73	0.6667	0.4142
17	rs4239111	7811998	T	C	31	25	1.24	0.7322	2.1	0.6429	0.4227
16	rs3102350	89379936	T	C	32	26	1.231	0.7336	2.065	0.6207	0.4308
17	rs2277638	7402556	G	A	36	43	0.8372	0.5377	1.304	0.6203	0.431
15	rs6494573	66735551	C	T	15	11	1.364	0.6263	2.969	0.6154	0.4328
15	rs12915582	93552330	C	T	18	23	0.7826	0.4224	1.45	0.6098	0.4349
2	rs1529667	1926437	G	T	6	9	0.6667	0.2373	1.873	0.6	0.4386
6	rs9767451	117996818	T	C	9	6	1.5	0.5339	4.214	0.6	0.4386
3	rs4301044	51246360	A	G	21	26	0.8077	0.4545	1.435	0.5319	0.4658
7	rs2058275	23164526	C	G	7	10	0.7	0.2665	1.839	0.5294	0.4669
11	rs28498421	802125	T	C	7	10	0.7	0.2665	1.839	0.5294	0.4669
1	rs2270977	18023509	T	C	23	28	0.8214	0.4732	1.426	0.4902	0.4838
9	rs3750330	131454120	C	T	11	8	1.375	0.5531	3.418	0.4737	0.4913
9	rs4836618	131455903	C	T	11	8	1.375	0.5531	3.418	0.4737	0.4913
11	rs1784259	103104930	C	A	8	11	0.7273	0.2925	1.808	0.4737	0.4913
12	rs1663564	105546172	G	A	11	8	1.375	0.5531	3.418	0.4737	0.4913
17	rs2228130	7404991	T	C	8	11	0.7273	0.2925	1.808	0.4737	0.4913
22	rs6537642	50658053	A	G	11	8	1.375	0.5531	3.418	0.4737	0.4913
17	rs2228129	7402600	T	C	36	42	0.8571	0.5492	1.338	0.4615	0.4969
2	rs12693800	197777589	C	T	20	16	1.25	0.6478	2.412	0.4444	0.505
12	rs4761829	52080965	C	T	31	26	1.192	0.708	2.008	0.4386	0.5078
19	rs16008	13443770	T	C	31	26	1.192	0.708	2.008	0.4386	0.5078
2	rs2070881	1457364	A	G	12	9	1.333	0.5618	3.164	0.4286	0.5127
2	rs4927610	1459806	A	G	12	9	1.333	0.5618	3.164	0.4286	0.5127
15	rs7495875	28362459	G	A	9	12	0.75	0.316	1.78	0.4286	0.5127
17	rs483434	26850929	A	C	9	12	0.75	0.316	1.78	0.4286	0.5127
7	rs377587	101713567	T	C	27	32	0.8438	0.5056	1.408	0.4237	0.5151

17	rs2071046	40689613	G	C	40	46	0.8696	0.5692	1.328	0.4186	0.5176
22	rs1557620	51137249	T	C	28	33	0.8485	0.5128	1.404	0.4098	0.5221
2	rs10153730	200173684	T	G	4	6	0.6667	0.1881	2.362	0.4	0.5271
2	rs4442987	228204982	C	T	18	22	0.8182	0.4389	1.525	0.4	0.5271
3	rs1705789	3197871	T	C	4	6	0.6667	0.1881	2.362	0.4	0.5271
18	rs647595	7037866	T	C	6	4	1.5	0.4233	5.315	0.4	0.5271
1	rs12732924	240371554	A	G	24	20	1.2	0.6629	2.172	0.3636	0.5465
1	rs2153463	147124310	T	G	33	38	0.8684	0.5447	1.384	0.3521	0.5529
12	rs303767	52156255	A	C	21	25	0.84	0.4702	1.501	0.3478	0.5553
17	rs12936464	7403942	A	C	34	39	0.8718	0.5504	1.381	0.3425	0.5584
20	rs2747404	18167977	T	C	7	5	1.4	0.4443	4.411	0.3333	0.5637
9	rs7022772	4566210	A	C	41	36	1.139	0.7279	1.782	0.3247	0.5688
6	rs2757645	135763866	G	A	13	16	0.8125	0.3908	1.689	0.3103	0.5775
17	rs86312	40696233	C	G	13	16	0.8125	0.3908	1.689	0.3103	0.5775
19	rs2248069	13445208	C	T	45	40	1.125	0.7348	1.722	0.2941	0.5876
16	rs2277908	89371839	G	A	47	42	1.119	0.7381	1.697	0.2809	0.5961
1	rs3120803	44386615	A	G	27	31	0.871	0.5199	1.459	0.2759	0.5994
6	rs594012	75841722	A	T	15	18	0.8333	0.42	1.653	0.2727	0.6015
12	rs2239127	2757756	T	C	31	35	0.8857	0.5462	1.436	0.2424	0.6225
12	rs2247291	122284715	T	C	36	32	1.125	0.6988	1.811	0.2353	0.6276
22	rs762672	51064818	T	C	33	37	0.8919	0.5578	1.426	0.2286	0.6326
1	rs454107	109794252	T	C	10	8	1.25	0.4933	3.167	0.2222	0.6374
1	rs413380	109795026	T	C	10	8	1.25	0.4933	3.167	0.2222	0.6374
1	rs437444	109795608	T	C	10	8	1.25	0.4933	3.167	0.2222	0.6374
16	rs2303191	78198192	T	C	34	38	0.8947	0.5633	1.421	0.2222	0.6374
2	rs12623857	223161889	A	G	22	19	1.158	0.6267	2.139	0.2195	0.6394
2	rs6754024	223162024	T	G	22	19	1.158	0.6267	2.139	0.2195	0.6394
18	rs607230	6980523	T	C	39	35	1.114	0.706	1.759	0.2162	0.6419
9	rs301979	4576851	G	C	36	40	0.9	0.5737	1.412	0.2105	0.6464
9	rs7033976	73151715	C	T	23	20	1.15	0.6316	2.094	0.2093	0.6473
14	rs2297124	89063167	A	G	23	20	1.15	0.6316	2.094	0.2093	0.6473
19	rs2217342	42489516	A	C	23	20	1.15	0.6316	2.094	0.2093	0.6473
12	rs7980561	23998888	T	C	41	37	1.108	0.7105	1.728	0.2051	0.6506
17	rs8075218	7405074	C	T	37	41	0.9024	0.5786	1.407	0.2051	0.6506
3	rs900688	47449058	T	C	3	2	1.5	0.2506	8.977	0.2	0.6547
3	rs900689	47452087	G	A	3	2	1.5	0.2506	8.977	0.2	0.6547
9	rs10747050	140055876	G	A	9	11	0.8182	0.339	1.974	0.2	0.6547
10	rs906220	71060610	A	G	21	24	0.875	0.4872	1.572	0.2	0.6547
11	rs615536	66481264	C	T	9	11	0.8182	0.339	1.974	0.2	0.6547
14	rs11621299	31549745	T	G	21	24	0.875	0.4872	1.572	0.2	0.6547
14	rs421262	88401213	T	C	3	2	1.5	0.2506	8.977	0.2	0.6547

14	rs417276	88406404	T	C	3	2	1.5	0.2506	8.977	0.2	0.6547
14	rs448805	88407734	G	C	3	2	1.5	0.2506	8.977	0.2	0.6547
14	rs421466	88407875	T	A	3	2	1.5	0.2506	8.977	0.2	0.6547
14	rs444902	88411803	T	C	3	2	1.5	0.2506	8.977	0.2	0.6547
14	rs367327	88411947	T	C	3	2	1.5	0.2506	8.977	0.2	0.6547
14	rs398343	88414283	G	A	3	2	1.5	0.2506	8.977	0.2	0.6547
16	rs1468138	23535680	A	C	39	43	0.907	0.588	1.399	0.1951	0.6587
16	rs6497671	23536684	T	C	39	43	0.907	0.588	1.399	0.1951	0.6587
16	rs2073951	23546561	G	C	39	43	0.907	0.588	1.399	0.1951	0.6587
16	rs3114912	89371827	G	A	47	43	1.093	0.7228	1.653	0.1778	0.6733
12	rs1805199	13769664	T	C	30	27	1.111	0.6606	1.869	0.1579	0.6911
1	rs3120802	44386305	C	T	29	32	0.9062	0.5483	1.498	0.1475	0.7009
1	rs2790053	165737704	C	G	32	29	1.103	0.6676	1.824	0.1475	0.7009
7	rs377612	101713590	T	C	30	33	0.9091	0.5545	1.491	0.1429	0.7055
11	rs7110158	70349061	A	G	3	4	0.75	0.1679	3.351	0.1429	0.7055
15	rs1724623	52681573	A	G	3	4	0.75	0.1679	3.351	0.1429	0.7055
17	rs659497	40689455	T	C	3	4	0.75	0.1679	3.351	0.1429	0.7055
19	rs2006885	18273410	A	G	33	30	1.1	0.6709	1.804	0.1429	0.7055
9	rs10901065	134386744	T	C	31	34	0.9118	0.5604	1.483	0.1385	0.7098
9	rs10901068	134387315	C	G	31	34	0.9118	0.5604	1.483	0.1385	0.7098
3	rs2625282	101484335	G	A	15	17	0.8824	0.4407	1.767	0.125	0.7237
15	rs4774621	52675261	T	C	5	4	1.25	0.3357	4.655	0.1111	0.7389
19	rs889167	47885180	A	G	5	4	1.25	0.3357	4.655	0.1111	0.7389
14	rs7155228	31538984	C	G	42	45	0.9333	0.6129	1.421	0.1034	0.7477
10	rs4600135	282897	C	T	19	21	0.9048	0.4864	1.683	0.1	0.7518
12	rs915997	7053362	A	G	19	21	0.9048	0.4864	1.683	0.1	0.7518
1	rs1141109	202705455	G	C	47	44	1.068	0.7081	1.611	0.0989	0.7532
18	rs693360	6980457	T	C	22	20	1.1	0.6004	2.015	0.09524	0.7576
5	rs2241694	149602608	A	G	5	6	0.8333	0.2543	2.731	0.09091	0.763
10	rs906221	71060634	G	A	21	23	0.913	0.5053	1.65	0.09091	0.763
15	rs2414145	52667552	G	A	6	5	1.2	0.3662	3.932	0.09091	0.763
17	rs630539	40693344	C	T	5	6	0.8333	0.2543	2.731	0.09091	0.763
12	rs1196802	105543326	T	G	24	22	1.091	0.6117	1.946	0.08696	0.7681
18	rs617573	6985655	C	T	22	24	0.9167	0.514	1.635	0.08696	0.7681
17	rs7405740	29670190	C	G	26	24	1.083	0.622	1.887	0.08	0.7773
9	rs968733	139981627	A	G	6	7	0.8571	0.2881	2.55	0.07692	0.7815
20	rs1205190	18118740	T	G	6	7	0.8571	0.2881	2.55	0.07692	0.7815
20	rs1205193	18143117	T	G	6	7	0.8571	0.2881	2.55	0.07692	0.7815
11	rs572126	118359161	G	A	28	26	1.077	0.6315	1.837	0.07407	0.7855
12	rs2468245	88473049	C	T	29	27	1.074	0.6359	1.814	0.07143	0.7893
3	rs342034	6903601	A	G	7	8	0.875	0.3173	2.413	0.06667	0.7963

12	rs215983	2706720	G	C	8	7	1.143	0.4144	3.152	0.06667	0.7963
19	rs1011320	18273047	T	C	7	8	0.875	0.3173	2.413	0.06667	0.7963
9	rs2808557	101243364	T	G	33	31	1.065	0.652	1.738	0.0625	0.8026
15	rs4779816	31369123	A	G	33	31	1.065	0.652	1.738	0.0625	0.8026
15	rs4779817	31369223	C	T	33	31	1.065	0.652	1.738	0.0625	0.8026
3	rs711631	4856180	T	C	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs11243404	134384277	T	C	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs2018621	134385599	A	G	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs10901066	134386903	G	A	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs3739494	134387488	T	C	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs1547768	134394163	A	G	32	34	0.9412	0.5808	1.525	0.06061	0.8055
9	rs3739495	134398534	T	C	32	34	0.9412	0.5808	1.525	0.06061	0.8055
2	rs11681136	121709091	G	A	9	8	1.125	0.4341	2.916	0.05882	0.8084
9	rs10870196	140040156	G	A	8	9	0.8889	0.343	2.304	0.05882	0.8084
15	rs11070320	41339481	C	T	9	8	1.125	0.4341	2.916	0.05882	0.8084
15	rs12719734	101113862	C	T	8	9	0.8889	0.343	2.304	0.05882	0.8084
22	rs762674	51063987	G	C	35	37	0.9459	0.5959	1.502	0.05556	0.8137
19	rs12608517	36558113	T	C	36	38	0.9474	0.6005	1.494	0.05405	0.8162
15	rs2899010	41346092	T	C	10	9	1.111	0.4515	2.734	0.05263	0.8185
19	rs16018	13411482	G	A	41	39	1.051	0.6781	1.63	0.05	0.8231
1	rs1141108	202715284	G	A	44	46	0.9565	0.6327	1.446	0.04444	0.833
1	rs3196669	202733238	C	T	46	44	1.045	0.6915	1.581	0.04444	0.833
3	rs7638391	71015021	G	T	15	16	0.9375	0.4635	1.896	0.03226	0.8575
21	rs928763	38850640	G	A	15	16	0.9375	0.4635	1.896	0.03226	0.8575
2	rs280196	121740505	C	T	17	18	0.9444	0.4867	1.833	0.02857	0.8658
10	rs906223	71060707	T	A	20	21	0.9524	0.5163	1.757	0.02439	0.8759
10	rs906222	71060696	G	A	21	22	0.9545	0.5249	1.736	0.02326	0.8788
10	rs2303990	292927	C	A	23	24	0.9583	0.5409	1.698	0.02128	0.884
1	rs2270976	18023690	A	G	24	25	0.96	0.5483	1.681	0.02041	0.8864
12	rs1196808	105534025	G	A	27	26	1.038	0.606	1.779	0.01887	0.8907
1	rs783303	44365529	G	A	28	29	0.9655	0.5744	1.623	0.01754	0.8946
6	rs6568446	107050867	A	T	29	30	0.9667	0.5802	1.61	0.01695	0.8964
9	rs2010635	134394646	A	G	32	33	0.9697	0.5963	1.577	0.01538	0.9013
12	rs2239128	2757769	T	C	32	33	0.9697	0.5963	1.577	0.01538	0.9013
1	rs1064261	11288758	G	A	33	34	0.9706	0.6012	1.567	0.01493	0.9028
9	rs6597501	134382673	C	T	33	34	0.9706	0.6012	1.567	0.01493	0.9028
9	rs2277153	134395687	C	A	33	34	0.9706	0.6012	1.567	0.01493	0.9028
19	rs4808755	18276856	T	C	36	37	0.973	0.6149	1.539	0.0137	0.9068
9	rs2277152	134395628	C	G	37	38	0.9737	0.6192	1.531	0.01333	0.9081
9	rs4740165	134397624	C	T	37	38	0.9737	0.6192	1.531	0.01333	0.9081
2	rs7596597	200298313	A	G	4	4	1	0.2501	3.998	0	1

3	rs1669321	3215954	A	G	7	7	1	0.3508	2.851	0	1
4	rs3775091	107168431	G	C	3	3	1	0.2018	4.955	0	1
5	rs26186	14472626	T	C	5	5	1	0.2895	3.454	0	1
10	rs1017361	294953	A	G	21	21	1	0.5462	1.831	0	1
12	rs10848675	2659848	T	C	14	14	1	0.4767	2.098	0	1
12	rs303808	52163789	G	A	29	29	1	0.5977	1.673	0	1
12	rs2468255	88505078	T	C	29	29	1	0.5977	1.673	0	1
12	rs772897	103237468	G	C	31	31	1	0.6078	1.645	0	1
15	rs7181472	38632148	A	G	5	5	1	0.2895	3.454	0	1
15	rs1724577	52689631	T	G	4	4	1	0.2501	3.998	0	1
19	rs16039	13355900	T	G	5	5	1	0.2895	3.454	0	1
22	rs762673	51064141	G	A	36	36	1	0.63	1.587	0	1
23	rs3020957	152821887	C	T	0	0	NA	NA	NA	NA	NA

Bibliografía

Abay-Nørgaard, S., Attianese, B., Boreggio, L., & Salcini, A. E. (2020). Regulators of H3K4 methylation mutated in neurodevelopmental disorders control axon guidance in Caenorhabditis elegans. *Development (Cambridge, England)*, *147*(15). https://doi.org/10.1242/dev.190637

Adams, D. R., & Eng, C. M. (2018). Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. *New England Journal of Medicine*, *379*(14), 1353–1362. https://doi.org/10.1056/nejmra1711801

Alazami, A. M., Patel, N., Shamseldin, H. E., Anazi, S., Al-Dosari, M. S., Alzahrani, F., Hijazi, H., Alshammari, M., Aldahmesh, M. A., Salih, M. A., Faqeih, E., Alhashem, A., Bashiri, F. A., Al-Owain, M., Kentab, A. Y., Sogaty, S., Al Tala, S., Temsah, M.-H., Tulbah, M., … Alkuraya, F. S. (2015). Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Reports*, *10*(2), 148–161. https://doi.org/10.1016/j.celrep.2014.12.015

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Ansarifar, J., Wang, L., & Hancock, J. (2019). New algorithms for detecting multi-effect and multi-way epistatic interactions. *Bioinformatics*, *35*(24), 5078–5085. https://doi.org/10.1093/bioinformatics/btz463

Ballif, B. C., Hornor, S. A., Jenkins, E., Madan-Khetarpal, S., Surti, U., Jackson, K. E., Asamoah, A., Brock, P. L., Gowans, G. C., Conway, R. L., Graham, J. M. J., Medne, L., Zackai, E. H., Shaikh, T. H., Geoghegan, J., Selzer, R. R., Eis, P. S., Bejjani, B. A., & Shaffer, L. G. (2007). Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nature Genetics*, *39*(9), 1071–1073. https://doi.org/10.1038/ng2107

Baruch, Y., Horn-Saban, S., Plotsky, Y., Bercovich, D., & Gershoni-Baruch, R. (2021). A case of Ververi-Brady syndrome due to QRICH1 loss of function and the literature

review. In *American journal of medical genetics. Part A* (Vol. 185, Issue 6, pp. 1913–1917). https://doi.org/10.1002/ajmg.a.62184

Bateson, W., & Mendel, G. (2013). *Mendel's principles of heredity*. Courier Corporation.

Bausela-Herreras, E., Tirapu-Ustárroz, J., & Cordero-Andrés, P. (2019). Deficits and neurodevelopmental disorders in childhood and adolescence. *Revista de Neurologia*, *69*(11), 461–469. https://doi.org/10.33588/RN.6911.2019133

Biancotti, J.-C., & Benvenisty, N. (2011). Aneuploid human embryonic stem cells: origins and potential for modeling chromosomal disorders. *Regenerative Medicine*, *6*(4), 493–503. https://doi.org/10.2217/rme.11.27

Bitta, M., Kariuki, S. M., Abubakar, A., & Newton, C. R. J. C. (2018). Burden of neurodevelopmental disorders in low and middle-income countries: A systematic review and meta-analysis [version 3; referees: 1 approved, 2 approved with reservations]. *Wellcome Open Research*, *2*. https://doi.org/10.12688/wellcomeopenres.13540.3

Brandenburg, J.-T., Clark, L., Botha, G., Panji, S., Baichoo, S., Fields, C., & Hazelhurst, S. (2022). H3AGWAS : A portable workflow for Genome Wide Association Studies. *BioRxiv*, 2022.05.02.490206. https://doi.org/10.1101/2022.05.02.490206

Brownstein, C. A., Beggs, A. H., Homer, N., Merriman, B., Yu, T. W., Flannery, K. C., DeChene, E. T., Towne, M. C., Savage, S. K., Price, E. N., Holm, I. A., Luquette, L. J., Lyon, E., Majzoub, J., Neupert, P., McCallie, D. J., Szolovits, P., Willard, H. F., Mendelsohn, N. J., … Margulies, D. M. (2014). An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biology*, *15*(3), R53. https://doi.org/10.1186/gb-2014-15-3-r53

Carayol, J., Schellenberg, G. D., Dombroski, B., Amiet, C., Génin, B., Fontaine, K., Rousseau, F., Vazart, C., Cohen, D., Frazier, T. W., Hardan, A. Y., Dawson, G., & Frio, T. R. (2014). A scoring strategy combining statistics and functional genomics supports a possible role for common polygenic variation in autism. *Frontiers in Genetics*, *5*(FEB). https://doi.org/10.3389/fgene.2014.00033

Cardoso, A. R., Lopes-Marques, M., Silva, R. M., Serrano, C., Amorim, A., Prata, M. J., & Azevedo, L. (2019). Essential genetic findings in neurodevelopmental disorders. *Human Genomics*, *13*(1), 31. https://doi.org/10.1186/s40246-019-0216-4

Cattaert, T., Urrea, V., Naj, A. C., De Lobel, L., De Wit, V., Fu, M., Mahachie John, J. M., Shen, H., Calle, M. L., Ritchie, M. D., Edwards, T. L., & Van Steen, K. (2010). FAM-

MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. *PloS One*, *5*(4), e10304. https://doi.org/10.1371/journal.pone.0010304

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chang, Y. C., Wu, J. T., Hong, M. Y., Tung, Y. A., Hsieh, P. H., Yee, S. W., Giacomini, K. M., Oyang, Y. J., & Chen, C. Y. (2020). GenEpi: Gene-based epistasis discovery using machine learning. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-3368-2

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://api.semanticscholar.org/CorpusID:59777418

Chen, X., Li, H., Chen, C., Zhou, L., Xu, X., Xiang, Y., & Tang, S. (2018). Genome-Wide Array Analysis Reveals Novel Genomic Regions and Candidate Gene for Intellectual Disability. *Molecular Diagnosis & Therapy*, *22*(6), 749–757. https://doi.org/10.1007/s40291-018-0358-4

Clinical utility of genetic and genomic services: a position statement of the American College of Medical Genetics and Genomics. (2015). *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(6), 505–507. https://doi.org/10.1038/gim.2015.41

Cole, B. S., Hall, M. A., Urbanowicz, R. J., Gilbert-Diamond, D., & Moore, J. H. (2017). Analysis of Gene-Gene Interactions. *Current Protocols in Human Genetics*, *95*(1), 1.14.1-1.14.10. https://doi.org/10.1002/cphg.45

Cordell, H. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*, 2463–2468. https://doi.org/10.1093/hmg/11.20.2463

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews. Genetics*, *10*, 392–404. https://doi.org/10.1038/nrg2579

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Ligt, J., Willemsen, M. H., van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B. B. A., Brunner, H. G., Veltman, J. A., & Vissers, L. E. L. M. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *The New England Journal of Medicine*, *367*(20), 1921–1929. https://doi.org/10.1056/NEJMoa1206524

Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, *11*(1), 70. https://doi.org/10.1186/s13073-019-0689-8

Dillon, O. J., Lunke, S., Stark, Z., Yeung, A., Thorne, N., Gaff, C., White, S. M., & Tan, T. Y. (2018). Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. *European Journal of Human Genetics*, *26*(5), 644–651. https://doi.org/10.1038/s41431-018-0099-1

Edwards, T. L., Wang, X., Chen, Q., Wormly, B., Riley, B., O'Neill, F. A., Walsh, D., Ritchie, M. D., Kendler, K. S., & Chen, X. (2008). Interaction between interleukin 3 and dystrobrevin-binding protein 1 in schizophrenia. *Schizophrenia Research*, *106*(2), 208–217. https://doi.org/https://doi.org/10.1016/j.schres.2008.07.022

Emily, M. (2016). AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies. *Statistical Applications in Genetics and Molecular Biology*, *15*(2), 151–171. https://doi.org/10.1515/sagmb-2015-0074

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Froukh, T. J. (2017). Next Generation Sequencing and Genome-Wide Genotyping Identify the Genetic Causes of Intellectual Disability in Ten Consanguineous Families from Jordan. *The Tohoku Journal of Experimental Medicine*, *243*(4), 297–309. https://doi.org/10.1620/tjem.243.297

Gall, K., Izzo, E., Seppälä, E. H., Alakurtti, K., Koskinen, L., Saarinen, I., Singh, A., Myllykangas, S., Koskenvuo, J., & Alastalo, T.-P. (2021). Next-generation sequencing in childhood-onset epilepsies: Diagnostic yield and impact on neuronal

ceroid lipofuscinosis type 2 (CLN2) disease diagnosis. *PloS One*, *16*(9), e0255933. https://doi.org/10.1371/journal.pone.0255933

Ghanem, S. I., Ghanem, N. M., & Ismail, M. A. (2019). *Noisy Epistasis Using Deep Learning*. 165–168. https://doi.org/10.1109/JEC-ECC.2018.8679568

González Alvaredo, S., Sanz Rojo, R., García Santiago, J., Gaztañaga Expósito, R., Bengoa, A., & Pérez-Yarza, E. G. (2008). [Genetic diagnostic criteria in cases of mental retardation and development of idiopathic origin]. *Anales de pediatria (Barcelona, Spain : 2003)*, *69*(5), 446–453. https://doi.org/10.1157/13128001

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Griffin, D. K. (1996). The incidence, origin, and etiology of aneuploidy. *International Review of Cytology*, *167*, 263–296. https://doi.org/10.1016/s0074-7696(08)61349-2

Gusareva, E. S., & Van Steen, K. (2014). Practical aspects of genome-wide association interaction analysis. *Human Genetics*, *133*(11), 1343–1358. https://doi.org/10.1007/s00439-014-1480-y

Hamdan, F. F., Gauthier, J., Araki, Y., Lin, D.-T., Yoshizawa, Y., Higashi, K., Park, A.-R., Spiegelman, D., Dobrzeniecka, S., Piton, A., Tomitori, H., Daoud, H., Massicotte, C., Henrion, E., Diallo, O., Shekarabi, M., Marineau, C., Shevell, M., Maranda, B., … Michaud, J. L. (2011). Excess of de novo deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. *American Journal of Human Genetics*, *88*(3), 306–316. https://doi.org/10.1016/j.ajhg.2011.02.001

Hieter, P., & Boguski, M. (1997). Functional genomics: it's all how you read it. *Science (New York, N.Y.)*, *278*(5338), 601–602. https://doi.org/10.1126/science.278.5338.601

Hu, T., Chen, Y., Kiralis, J. W., & Moore, J. H. (2013). ViSEN: methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology*, *37*(3), 283–285. https://doi.org/10.1002/gepi.21718

Hu, T., Sinnott-Armstrong, N. A., Kiralis, J. W., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2011). Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, *12*, 364. https://doi.org/10.1186/1471-2105-12-364

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman,

H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., … Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. https://doi.org/10.1038/nature13908

Kafaie, S., Chen, Y., & Hu, T. (2019). A network approach to prioritizing susceptibility genes for genome-wide association studies. *Genetic Epidemiology*, *43*(5), 477–491. https://doi.org/10.1002/gepi.22198

Kessi, M., Chen, B., Pang, N., Yang, L., Peng, J., He, F., & Yin, F. (2023). The genotype-phenotype correlations of the CACNA1A-related neurodevelopmental disorders: a small case series and literature reviews. *Frontiers in Molecular Neuroscience*, *16*, 1222321. https://doi.org/10.3389/fnmol.2023.1222321

Khoury, M. J., Gwinn, M. L., Glasgow, R. E., & Kramer, B. S. (2012). A population approach to precision medicine. *American Journal of Preventive Medicine*, *42*(6), 639–645. https://doi.org/10.1016/j.amepre.2012.02.012

Kim, N. C., Andrews, P. C., Asselbergs, F. W., Frost, H. R., Williams, S. M., Harris, B. T., Read, C., Askland, K. D., & Moore, J. H. (2012). Gene ontology analysis of pairwise genetic associations in two genome-wide studies of sporadic ALS. *BioData Mining*, *5*(1), 9. https://doi.org/10.1186/1756-0381-5-9

Koolen, D. A., Vissers, L. E. L. M., Pfundt, R., de Leeuw, N., Knight, S. J. L., Regan, R., Kooy, R. F., Reyniers, E., Romano, C., Fichera, M., Schinzel, A., Baumer, A., Anderlid, B.-M., Schoumans, J., Knoers, N. V, van Kessel, A. G., Sistermans, E. A., Veltman, J. A., Brunner, H. G., & de Vries, B. B. A. (2006). A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genetics*, *38*(9), 999–1001. https://doi.org/10.1038/ng1853

Kumble, S., Levy, A. M., Punetha, J., Gao, H., Ah Mew, N., Anyane-Yeboa, K., Benke, P. J., Berger, S. M., Bjerglund, L., Campos-Xavier, B., Ciliberto, M., Cohen, J. S., Comi, A. M., Curry, C., Damaj, L., Denommé-Pichon, A.-S., Emrick, L., Faivre, L., Fasano, M. B., … Tümer, Z. (2022). The clinical and molecular spectrum of QRICH1 associated neurodevelopmental disorder. *Human Mutation*, *43*(2), 266–282. https://doi.org/10.1002/humu.24308

Large-scale discovery of novel genetic causes of developmental disorders. (2015). *Nature*, *519*(7542), 223–228. https://doi.org/10.1038/nature14135

Larranaga, P. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*,

86–112. https://doi.org/10.1093/bib/bbk007

Leem, S., & Park, T. (2017). An empirical fuzzy multifactor dimensionality reduction method for detecting gene-gene interactions. *BMC Genomics*, *18*. https://doi.org/10.1186/s12864-017-3496-x

Leonard, H., & Wen, X. (2002). The epidemiology of mental retardation: Challenges and opportunities in the new millennium. *Mental Retardation and Developmental Disabilities Research Reviews*, *8*(3), 117–134. https://doi.org/https://doi.org/10.1002/mrdd.10031

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2016). Detecting gene-gene interactions using a permutation-based random forest method. *BioData Mining*, *9*(1). https://doi.org/10.1186/s13040-016-0093-5

Li, Q., Fallin, M. D., Louis, T. A., Lasseter, V. K., McGrath, J. A., Avramopoulos, D., Wolyniec, P. S., Valle, D., Liang, K.-Y., Pulver, A. E., & Ruczinski, I. (2010). Detection of SNP-SNP interactions in trios of parents with schizophrenic children. *Genetic Epidemiology*, *34*(5), 396–406. https://doi.org/10.1002/gepi.20488

Lipkin, P. H., & Macias, M. M. (2020). Promoting Optimal Development: Identifying Infants and Young Children With Developmental Disorders Through Developmental Surveillance and Screening. *Pediatrics*, *145*(1). https://doi.org/10.1542/peds.2019-3449

Liu, D., Wang, M., Yuan, Y., Schwender, H., Wang, H., Wang, P., Zhou, Z., Li, J., Wu, T., Zhu, H., & Beaty, T. H. (2019). Gene-gene interaction among cell adhesion genes and risk of nonsyndromic cleft lip with or without cleft palate in Chinese case-parent trios. *Molecular Genetics & Genomic Medicine*, *7*(10), e00872. https://doi.org/10.1002/mgg3.872

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *The Lancet*, *392*(10146), 508–520. https://doi.org/10.1016/S0140-6736(18)31129-2

Lou, X.-Y., Chen, G.-B., Yan, L., Ma, J. Z., Mangold, J. E., Zhu, J., Elston, R. C., & Li, M. D. (2008). A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *American Journal of Human Genetics*, *83*(4), 457–467. https://doi.org/10.1016/j.ajhg.2008.09.001

Manavalan, R., & Priya, S. (2020). Epistasis effects of complex diseases from simulated models through computational approaches: A review. *International Journal of*

Scientific and Technology Research, *9*(3), 5444–5462. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082683196&partnerID=40&md5=5bf6555bd63be58e53cb6939d425e8fb

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, *19*(2), 286–302. https://doi.org/10.1093/BIB/BBW114

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608. https://doi.org/10.1002/mpr.1608

Martin, E. R., Bass, M. P., Gilbert, J. R., Pericak-Vance, M. A., & Hauser, E. R. (2003). Genotype-based association test for general pedigrees: the genotype-PDT. *Genetic Epidemiology*, *25*(3), 203–213. https://doi.org/10.1002/gepi.10258

Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S., & Moore, J. H. (2006). A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genetic Epidemiology*, *30*(2), 111–123. https://doi.org/10.1002/gepi.20128

Mascheretti, S., Bureau, A., Trezzi, V., Giorda, R., & Marino, C. (2015). An assessment of gene-by-gene interactions as a tool to unfold missing heritability in dyslexia. *Human Genetics*, *134*(7), 749–760. https://doi.org/10.1007/s00439-015-1555-4

Mason, H., & Wiggins, C. (2010). *A taxonomy of data science*. Http://Www.Dataists.Com/2010/09/a-Taxonomy-of-Data-Science/.

Mathé, C., Sagot, M., Schiex, T., & Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, *30*(19), 4103–4117. https://doi.org/10.1093/nar/gkf543

Moeschler, J. B., & Shevell, M. (2014). Comprehensive evaluation of the child with intellectual disability or global developmental delays. *Pediatrics*, *134*(3), e903-18. https://doi.org/10.1542/peds.2014-1839

Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics (Oxford, England)*, *26*(22), 2927–2928. https://doi.org/10.1093/bioinformatics/btq562

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, *56*(1–3), 73–82. https://doi.org/10.1159/000073735

Moore, J. H., Hahn, L. W., Ritchie, M. D., Thornton, T. A., & White, B. C. (2004). Routine discovery of complex genetic models using genetic algorithms. *Applied Soft Computing Journal*, *4*(1), 79–86. https://doi.org/10.1016/j.asoc.2003.08.003

Morgan, A., Gandin, I., Belcaro, C., Palumbo, P., Palumbo, O., Biamino, E., Dal Col, V., Laurini, E., Pricl, S., Bosco, P., Carella, M., Ferrero, G. B., Romano, C., d'Adamo, A. P., Faletra, F., & Vozzi, D. (2015). Target sequencing approach intended to discover new mutations in non-syndromic intellectual disability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *781*, 32–36. https://doi.org/https://doi.org/10.1016/j.mrfmmm.2015.09.002

Morris-Rosendahl, D. J., & Crocq, M. A. (2020). Neurodevelopmental disorders-the history and future of a diagnostic concept. *Dialogues in Clinical Neuroscience*, *22*(1), 65–72. https://doi.org/10.31887/DCNS.2020.22.1/macrocq

Newman, J. R. B., Concannon, P., & Ge, Y. (2023). UBASH3A Interacts with PTPN22 to Regulate IL2 Expression and Risk for Type 1 Diabetes. *International Journal of Molecular Sciences*, *24*(10). https://doi.org/10.3390/ijms24108671

Nodzenski, M., Shi, M., Krahn, J. M., Wise, A. S., Li, Y., Li, L., Umbach, D. M., & Weinberg, C. R. (2022). GADGETS: a genetic algorithm for detecting epistasis using nuclear families. *Bioinformatics (Oxford, England)*, *38*(4), 1052–1058. https://doi.org/10.1093/bioinformatics/btab766

Ohi, K., Hashimoto, R., Yasuda, Y., Fukumoto, M., Yamamori, H., Umeda-Yano, S., Fujimoto, M., Iwase, M., Kazui, H., & Takeda, M. (2013). Influence of the NRGN gene on intellectual ability in schizophrenia. *Journal of Human Genetics*, *58*(10), 700–705. https://doi.org/10.1038/jhg.2013.82

Orfao, A., Benítez, J., Corrales, F., Martín-Subero, I., Ordovás, J. M., Carracedo, Á., & Lapunzina, P. (2019). *Ciencias ónicas*. 32. www.institutoroche.es

Palsson, B. (2002). In silico biology through "omics." *Nature Biotechnology*, *20*(7), 649–

650. https://doi.org/10.1038/nbt0702-649

Park, C., Kim, J., Kim, J., & Park, S. (2018). Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0201056

Park, M., Lee, J. W., Park, T., & Lee, S. (2020). Gene-Gene Interaction Analysis for the Survival Phenotype Based on the Kaplan-Meier Median Estimate. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/5282345

Paschou, P., Yu, D., Gerber, G., Evans, P., Tsetsos, F., Davis, L. K., Karagiannidis, I., Chaponis, J., Gamazon, E., Mueller-Vahl, K., Stuhrmann, M., Schloegelhofer, M., Stamenkovic, M., Hebebrand, J., Noethen, M., Nagy, P., Barta, C., Tarnok, Z., Rizzo, R., … Scharf, J. M. (2014). Genetic association signal near NTN4 in Tourette syndrome. *Annals of Neurology*, *76*(2), 310–315. https://doi.org/10.1002/ana.24215

Pattin, K. A., White, B. C., Barney, N., Gui, J., Nelson, H. H., Kelsey, K. T., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2009). A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genetic Epidemiology*, *33*(1), 87–94. https://doi.org/10.1002/gepi.20360

Perakakis, N., Yazdani, A., Karniadakis, G. E., & Mantzoros, C. (2018). Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. In *Metabolism: clinical and experimental* (Vol. 87, pp. A1--A9). https://doi.org/10.1016/j.metabol.2018.08.002

Pineda-Cirera, L., Shivalikanjli, A., Cabana-Domínguez, J., Demontis, D., Rajagopal, V. M., Børglum, A. D., Faraone, S. V, Cormand, B., & Fernàndez-Castillo, N. (2019). Exploring genetic variation that influences brain methylation in attention-deficit/hyperactivity disorder. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0574-7

Piriyapongsa, J., Ngamphiw, C., Intarapanich, A., Kulawonganunchai, S., Assawamakin, A., Bootchai, C., Shaw, P. J., & Tongsima, S. (2012). iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies. *BMC Genomics*, *13 Suppl 7*. https://doi.org/10.1186/1471-2164-13-s7-s2

Posar, A., & Visconti, P. (2017). Neurodevelopmental Disorders between Past and Future. *Journal of Pediatric Neurosciences*, *12*(3), 301–302. https://doi.org/10.4103/jpn.JPN_95_17

Price, K. M., Wigg, K. G., Feng, Y., Blokland, K., Wilkinson, M., He, G., Kerr, E. N., Carter, T.-C., Guger, S. L., Lovett, M. W., Strug, L. J., & Barr, C. L. (2020). Genome-

wide association study of word reading: Overlap with risk genes for neurodevelopmental disorders. *Genes, Brain and Behavior*, *19*(6). https://doi.org/10.1111/gbb.12648

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology and Medicine*, *16*(1), 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Hüffmeier, U., Thiel, C., Rüschendorf, F., Nürnberg, P., Reis, A., & Trautmann, U. (2006). Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *American Journal of Medical Genetics. Part A*, *140*(19), 2063–2074. https://doi.org/10.1002/ajmg.a.31416

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. https://doi.org/10.1038/nature05329

Ritchie, M D, Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, *69*(1), 138–147. https://doi.org/10.1086/321276

Ritchie, Marylyn D., & Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*, *6*(8), 157–157. https://doi.org/10.21037/atm.2018.04.05

Roston, A., Evans, D., Gill, H., McKinnon, M., Isidor, B., Cogné, B., Mwenifumbo, J., van Karnebeek, C., An, J., Jones, S. J. M., Farrer, M., Demos, M., Connolly, M., & Gibson, W. T. (2021). SETD1B-associated neurodevelopmental disorder. *Journal of Medical Genetics*, *58*(3), 196–204. https://doi.org/10.1136/jmedgenet-2019-106756

Saldarriaga, W., Forero-Forero, J. V., González-Teshima, L. Y., Fandiño-Losada, A., Isaza, C., Tovar-Cuevas, J. R., Silva, M., Choudhary, N. S., Tang, H.-T., Aguilar-Gaxiola, S., Hagerman, R. J., & Tassone, F. (2018). Genetic cluster of fragile X syndrome in a Colombian district. *Journal of Human Genetics*, *63*(4), 509–516. https://doi.org/10.1038/s10038-017-0407-6

Saldarriaga, W., Tassone, F., González-Teshima, L. Y., Forero-Forero, J. V., Ayala-

Zapata, S., & Hagerman, R. (2014). Fragile X syndrome. *Colombia Medica (Cali, Colombia)*, *45*(4), 190–198.

Savatt, J. M., & Myers, S. M. (2021). Genetic Testing in Neurodevelopmental Disorders. *Frontiers in Pediatrics*, *9*, 526779. https://doi.org/10.3389/fped.2021.526779

Shaffer, L. G., Bejjani, B. A., Torchia, B., Kirkpatrick, S., Coppinger, J., & Ballif, B. C. (2007). The identification of microdeletion syndromes and other chromosome abnormalities: cytogenetic methods of the past, new technologies for the future. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *145C*(4), 335–345. https://doi.org/10.1002/ajmg.c.30152

Sharp, A. J., Mefford, H. C., Li, K., Baker, C., Skinner, C., Stevenson, R. E., Schroer, R. J., Novara, F., De Gregori, M., Ciccone, R., Broomer, A., Casuga, I., Wang, Y., Xiao, C., Barbacioru, C., Gimelli, G., Bernardina, B. D., Torniero, C., Giorda, R., … Eichler, E. E. (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, *40*(3), 322–328. https://doi.org/10.1038/ng.93

Shaw-Smith, C., Pittman, A. M., Willatt, L., Martin, H., Rickman, L., Gribble, S., Curley, R., Cumming, S., Dunn, C., Kalaitzopoulos, D., Porter, K., Prigmore, E., Krepischi-Santos, A. C. V, Varela, M. C., Koiffmann, C. P., Lees, A. J., Rosenberg, C., Firth, H. V, de Silva, R., & Carter, N. P. (2006). Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nature Genetics*, *38*(9), 1032–1037. https://doi.org/10.1038/ng1858

Sinoquet, C., & Niel, C. (2019). *Ant colony optimization for markov blanket-based feature selection. Application for precision medicine: Vol. 11331 LNCS* (pp. 217–230). https://doi.org/10.1007/978-3-030-13709-0_18

Slavotinek, A. M. (2008). Novel microdeletion syndromes detected by chromosome microarrays. *Human Genetics*, *124*(1), 1–17. https://doi.org/10.1007/s00439-008-0513-9

Slim Lotfi AND Chatelain, C. A. N. D. A. C.-A. A. N. D. V. J.-P. (2020). Novel methods for epistasis detection in genome-wide association studies. *PLOS ONE*, *15*(11), 1–18. https://doi.org/10.1371/journal.pone.0242927

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, *52*(3), 506–516.

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W.

K., Firth, H. V, Frazier, T., Hansen, R. L., Prock, L., Brunner, H., Hoang, N., Scherer, S. W., Sahin, M., & Miller, D. T. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *21*(11), 2413–2421. https://doi.org/10.1038/s41436-019-0554-6

Stessman, H. A. F., Willemsen, M. H., Fenckova, M., Penn, O., Hoischen, A., Xiong, B., Wang, T., Hoekzema, K., Vives, L., Vogel, I., Brunner, H. G., van der Burgt, I., Ockeloen, C. W., Schuurs-Hoeijmakers, J. H., Klein Wassink-Ruiter, J. S., Stumpel, C., Stevens, S. J. C., Vles, H. S., Marcelis, C. M., … Kleefstra, T. (2016). Disruption of POGZ Is Associated with Intellectual Disability and Autism Spectrum Disorders. *The American Journal of Human Genetics*, *98*(3), 541–552. https://doi.org/https://doi.org/10.1016/j.ajhg.2016.02.004

Sun, Y., Wang, X., Shang, J., Liu, J., Zheng, C., & Lei, X. (2018). Introducing Heuristic Information into Ant Colony Optimization Algorithm for Identifying Epistasis. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2018.2879673

Sung, P.-Y., Wang, Y.-T., Yu, Y.-W., & Chung, R.-H. (2016). An efficient gene-gene interaction test for genome-wide association studies in trio families. *Bioinformatics*, *32*(12), 1848–1855. https://doi.org/10.1093/bioinformatics/btw077

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000

Tărlungeanu, D. C., & Novarino, G. (2018). Genomics in neurodevelopmental disorders: an avenue to personalized medicine. *Experimental and Molecular Medicine*, *50*(8), 100. https://doi.org/10.1038/s12276-018-0129-7

Thapar, A., Cooper, M., & Rutter, M. (2017). Neurodevelopmental disorders. *The Lancet Psychiatry*, *4*(4), 339–346. https://doi.org/10.1016/S2215-0366(16)30376-5

Thornton-Wells, T. A., Moore, J. H., Martin, E. R., Pericak-Vance, M. A., & Haines, J. L. (2008). Confronting complexity in late-onset Alzheimer disease: application of two-stage analysis approach addressing heterogeneity and epistasis. *Genetic*

Epidemiology, *32*(3), 187–203. https://doi.org/https://doi.org/10.1002/gepi.20294

Turner, S. D., Dudek, S. M., & Ritchie, M. D. (2010). ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. *BioData Mining*, *3*(1). https://doi.org/10.1186/1756-0381-3-5

Van steen, K. (2012). Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, *13*(1), 1–19. https://doi.org/10.1093/bib/bbr012

Vanderweele, T. J. (2010). Epistatic interactions. *Statistical Applications in Genetics and Molecular Biology*, *9*(1). https://doi.org/10.2202/1544-6115.1517

Vantaggiato, C., Clementi, E., & Bassi, M. T. (2014). ZFYVE26/SPASTIZIN: a close link between complicated hereditary spastic paraparesis and autophagy. *Autophagy*, *10*(2), 374–375. https://doi.org/10.4161/auto.27173

Vissers, L. E. L. M., Gilissen, C., & Veltman, J. A. (2016). Genetic studies in intellectual disability and related disorders. *Nature Reviews. Genetics*, *17*(1), 9–18. https://doi.org/10.1038/nrg3999

Vissers, L. E. L. M., & Stankiewicz, P. (2012). Microdeletion and microduplication syndromes. *Methods in Molecular Biology (Clifton, N.J.)*, *838*, 29–75. https://doi.org/10.1007/978-1-61779-507-7_2

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Webber, C. (2017). Epistasis in Neuropsychiatric Disorders. *Trends in Genetics : TIG*, *33*(4), 256–265. https://doi.org/10.1016/j.tig.2017.01.009

Winston, P. H. (1992). *Artificial Intelligence (3rd Ed.)*. Addison-Wesley Longman Publishing Co., Inc.

Wong, A. K., Park, C. Y., Greene, C. S., Bongo, L. A., Guan, Y., & Troyanskaya, O. G. (2012). IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Research*, *40*(Web Server issue), W484-90. https://doi.org/10.1093/nar/gks458

World Health Organization. (2021). *Autism spectrum disorders*. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders

Wright, C. F., FitzPatrick, D. R., & Firth, H. V. (2018). Paediatric genomics: Diagnosing rare disease in children. *Nature Reviews Genetics*, *19*(5), 253–268. https://doi.org/10.1038/nrg.2017.116

Xiang, X., Wang, S., Liu, T., Wang, M., Li, J., Jiang, J., Wu, T., & Hu, Y. (2020). Exploring gene–gene interaction in family-based data with an unsupervised machine learning method: EPISFA. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.22342

Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. (2013). Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters*, *340*(2), 284–295. https://doi.org/10.1016/j.canlet.2012.11.025

Yang, C.-H., Chuang, L.-Y., & Lin, Y.-D. (2017). Multiobjective differential evolution-based multifactor dimensionality reduction for detecting gene-gene interactions. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12773-x

Zhang, X., Huang, S., Zou, F., & Wang, W. (2010). TEAM: Efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, *26*(12), i217–i227. https://doi.org/10.1093/bioinformatics/btq186

Zhang, Y., Chen, Y., & Hu, T. (2020). PANDA: Prioritization of autism-genes using network-based deep-learning approach. *Genetic Epidemiology*, *44*(4), 382–394. https://doi.org/10.1002/gepi.22282

Zhao, J., Zhu, Y., & Xiong, M. (2016). Genome-wide gene-gene interaction analysis for next-generation sequencing. *European Journal of Human Genetics*, *24*(3), 421–428. https://doi.org/10.1038/ejhg.2015.147

Zhou, Z. H., Liu, G. X., Su, L. T., Han, L., & Yan, L. (2014). *Detecting epistasis by LASSO-penalized-model search algorithm in human Genome-Wide Association Studies* (Vols. 989–994, pp. 2426–2430). https://doi.org/10.4028/www.scientific.net/AMR.989-994.2426