CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESIS DE DOCTORADO

BASES GENÉTICAS DE LOS TRASTORNOS DEL ESPECTRO AUTISTA: ESTUDIO DE LA VARIACIÓN COMÚN Y RARA

Aitana Alonso González

ESCOLA DE DOUTOTAMENTO INTERNACIONAL
PROGRAMA DE DOCTORADO EN MEDICINA MOLECULAR

SANTIAGO DE COMPOSTELA
AÑO 2020

CENTRO INTERNACIONAL DE ESTUDOS DE DOUTORAMENTO E AVANZADOS DA USC (CIEDUS)

DECLARACIÓN DEL AUTOR DE LA TESIS

Bases genéticas de los trastornos del espectro autista: estudio de la variación común y rara

D./Dña. Aitana Alonso González

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento, y declaro que:

1)La tesis abarca los resultados de la elaboración de mi trabajo.

2)En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.

3) La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.

4)Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.

En Santiago de Compostela, a

27 de mayo de 2020

Fdo. Aitana Alonso González

CENTRO INTERNACIONAL DE ESTUDOS DE DOUTORAMENTO E AVANZADOS DA USC (CIEDUS)

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

Bases genéticas de los trastornos del espectro autista: estudio de la variación común y rara

D./Dña. Ángel Carracedo Álvarez

D./Dña. Cristina Rodriguez Fontenla

INFORMAN:

Que la presente tesis, corresponde con el trabajo realizado por D/Dña. ***Aitana Alonso González*** *bajo mi dirección, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director de ésta no incurre en las causas de abstención establecidas en Ley 40/2015.*

En Santiago de Compotela a 27de mayo de 2020.

Fdo. Ángel Carracedo Álvarez

Fdo. Cristina Rodríguez Fontenla

CENTRO INTERNACIONAL DE ESTUDOS DE DOUTORAMENTO E AVANZADOS DA USC (CIEDUS)

Yo, Aitana Alonso González con DNI 54063597G declaro que esta tesis no presenta conflictos de interés. Y para que así conste firmo a día 27 de mayo de 2020

Fdo: Aitana Alonso González.

"Me doy cuenta que, si fuera estable, prudente y estático, viviría en la muerte. Por consiguiente, acepto la confusión, la incertidumbre, el miedo y los altibajos emocionales, porque ese es el precio que estoy dispuesto a pagar por una vida fluida, perpleja y excitante"

Carl Rogers

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AGRADECIMIENTOS

Una tesis, es como una carrera de fondo. A lo largo del recorrido, quieres abandonar mil veces, pero las ganas de alcanzar tu objetivo son aún mayores, y por eso, continuas. Sabes que, una vez llegues a la meta, la sensación de superación y satisfacción, harán que el esfuerzo haya merecido la pena. Debo decir, que poca experiencia tengo yo en carreras de fondo, pero al igual que me ha pasado con esta tesis, si conseguí terminar una en mi vida, fue gracias a la gente que corría a mi lado y me acompañó hasta el final, sin quedarse en el camino.

Por eso, en primer lugar, quiero agradecer a Ángel por haberme dado la oportunidad de iniciar este recorrido, apostar por mí y permitirme dar mis primeros pasos en el mundo de la genética bajo su dirección. Otra cosa que debo agradecerle, es poner a Cris es mi camino. Espero que recuerdes tu primer "experimento" de dirección de una tesis con cariño, como yo lo haré. Ambas sabemos que, sin ti, esta tesis no habría sido posible.

Gracias a todo el equipo de Xenómica, tanto el de la Fundación, como el del CIMUS por ayudarme en todo lo que he necesitado. Gracias a Inés, por haber creído en mí y no haberme dejado tirar la toalla. Gracias a Pancho por todo su tiempo y todo lo que he aprendido con él. Gracias a Cata por darle vida al CIMUS. Gracias a Elena, la mejor compañera de mesa que se puede tener. Como te echo de menos cuando te bajas a cultivos. Gracias a Raquel, por estar siempre dispuesta a resolver dudas. Gracias a Manuel Calaza por habernos ayudado tanto a Cris y a mí al principio. Gracias a Xulio y a Carolina. Os dije que solo estaría unos meses, pero es que he estado muy bien allí. Gracias a Vicente y Cris, mis otros dos grandes compañeros de despacho en la Fundación. Gracias al grupo *Coffe*, del que forman parte todos los doctorandos, con los que he compartido cabreos y risas. Gracias Montse, Olalla, Jorge, Joja, Pili, Sole y todos los compañeros que siempre han estado ahí para echar una mano.

Gracias a todos mis amigos. Si nos los hubiese elegido tan buenos, creo que esta tesis habría quedado en un proyecto sin acabar. Gracias a las doctorandas, Rita, Anael, Marinela y Cris. Esos cafés/cañas me han salvado la vida en más de una ocasión y lo sabéis. Gracias a Cebro. Menudo año te he dado, pero menos mal que nuestros laboratorios no tienen puerta. Gracias a Presidenta, Laura y Cris por haberme ayudado tanto al principio. Gracias a todos y cada uno de los miembros de *Agenda Picheleira*. No os puedo nombrar a todos, porque me quedo sin sitio, pero habéis sido una de las principales cosas por las que ha merecido la pena esta etapa. Gracias al grupo de *Tropical*, un reciente descubrimiento sin el que no hubiese podido sobrevivir esta cuarentena ni escribir esta tesis. Gracias a mis *Chicas*, que son, más que unas amigas, mi familia. Gracias a las *Doctoras Sexys*, por aceptarme, aunque sea la oveja negra de todos los médicos.

Por último, gracias a mi familia, que llegados a este punto estarán pensando que en qué momento me convertí en una cursi. Gracias a mis padres por haberme dado la oportunidad de poder llegar a donde yo quiera. Son, cada uno a su manera, un ejemplo a seguir. Gracias a mi hermana, por ser mi mayor apoyo y siempre estar ahí para cuando la necesito.

A todos, GRACIAS

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

GLOSARIO DE TÉRMINOS

AAF: Frecuencia del alelo alternativo (del inglés, *Alternate Allele Frequency*)
ACE: *Autism Center of Excellence*
ACMG: *American College of Medical Genetics*
AD: Autosómico dominante
ADDM: Red de Vigilancia del Autismo y las Discapacidades del Desarrollo (del inglés, *Early Autism and Developmental Disabilities Monitoring*)
ADI-R: *Autism Diagnostic Interview-Revised*
ADN: Ácido desoxirribonucleico
ADOS: *Autism Diagnostic Observation Schedule*
AGP: *Autism Genome Project Consortium*
AGRE: *Autism Genetics Resource Exchange*
Annovar: *ANNOtate VARiation*
AR: Autosómico recesivo
ARN: Ácido ribonucleico
ASD: *Autism Sequencing Consortium*
CGH: Hibridación genómica comparada (del inglés, *Comparative Genome Hybridization*)
CNV: Variante del número de copias (del inglés, *Copy Number Variant*)
DDD: *Deciphering Developmental Disorders*
DE: Desviación estándar
DI: Discapacidad intelectual
DSM-5: Manual Diagnóstico y Estadístico de los Trastornos Mentales, 5º edición (del inglés, *Diagnostic and Statistical Manual of Mental Disorders, fifth edition*)
eQTL: *Expression Quantitative Trait Loci*
EWCE: *Expression Weighted Cell-type Enrichment*
ExAC: *Exome Aggregation Consortium*
FDR: *False Discovery Rate*
FISH: Hibridación fluorescente *in situ* (del inglés, *Fluorescence In Situ Hybridization*)
FMRP: *Fragile X mental retardation 1 protein*
FSHD: *Facioscapulohumeral muscular dystrophy*
FUMA: *Functional Mapping and Annotation of Genome-Wide Association Studies*
GBA: Análisis basados en genes (del inglés, *Gene-Based Analysis*)
GEO*: Gene Expression Omnibus*
GnomAD: *The Genome Aggregation Database*
GO: Ontología génica (del inglés, *Gene Ontology*)

GQ: *Genotype Quality*
GSEA: Análisis de enriquecimiento de genes (Del inglés, *Gene Set Enrichment Analysis)*
GTEx: *The Genotype-Tissue Expression Project*
GWAS: Estudios de asociación de genoma completo (del inglés, Genome Wide Association Study)
ICD10: *International Classification of Diseases, 10th revision*
KEEG: *The Kyoto Encyclopaedia of Genes and Genome Elements*
LD: Desequilibrio de ligamiento (del inglés, *Linkage Disequilibrium*)
LoF: Mutaciones con pérdida de función (del inglés, *Loss-of-Function*)
MAF: Frecuencia del alelo menor (del inglés, *Minor Allele Frequency*)
Mb: Megabase
MES: *MaxEntScan*
MOMBOS: *NIMH Repository, the Montreal/Boston Collection*
MPC: *Missense Badness, PolyPhen-2, Constraint*
NALT: *Number of non-reference genotypes,*
NGS: Secuenciación de nueva generación (del inglés, *Next Generation Sequencing)*
NHET: *Number of heterozygous genotypes for individual*
NHGRI: Instituto Nacional de Investigación del Genoma Humano (del inglés, *National Human Genome Research Institute Home*)
NMIN*: Number of genotypes with a minor allele*
NNS: *NNSPLICE*
NVAR: *Total number of called variants for individual*
OMIM: *Online Mendelian Inheritance in Man*
OR: *Odds ratio*
Pb: Pares de bases
PCA: Análisis de componenteS principales (del inglés, *Principal Component Analysis*)
PCR: Reacción en cadena de la polimerasa (del inglés, *Polymerase Chain Reaction*)
PGC: *Psychiatric Genomics Consortium*
pLI: *Probability of loss-of-function intolerance*
PPI: Interacción proteína-proteína (del inglés, *Protein-Protein Interaction*)
pSI: *Specificity index statistic*
PZM: Mutación postcigótica (del inglés, *Postzygotic Mutation*)
RATE: *Genotyping rate for individual*
RGD: Retraso global del desarrollo
RPKM: Lecturas por kilobase por millón (del inglés, *Read Per Kilobase per Million)*
RR: Recurrencia relativa
RVIS: *Residual Variation Intolerance Score*
scRNA-Seq: Single-cell RNA-Seq
SFARI: *Simons Foundation Autism Research Initiative*

SNP: Polimorfismo de un solo nucleótido (del inglés, *Single Nucleotide Polimorphism)*
SNV: Variante de un solo nucleótido (del inglés, *Single Nucleotide Variant*)
SPARK: *Simons Foundation Autism Research Initiative*
SSC: *Simons Simplex Collection*
SSF: *SpliceSiteFinder*
TAD: Dominios topológicamente asociados (del inglés, *Topologically Associating Domain*)
TDAH: Trastorno por déficit de atención e hiperactividad
TEA: Trastornos del espectro autista
TF: Factor de transcripción (del inglés, *Transcriptional Factor*)
TND: Trastornos del neurodesarrollo
TPM: Tránscrito por millón (del inglés, *Transcripts Per Million*)
VCF: *Variant Calling Format*
WES: Secuenciación de exoma completo (del inglés, *Whole Exome Sequencing*)
WGCNA: *Weighted Gene Co-Expression Red Analysis*
WGS: Secuenciación de genoma completo (del inglés *Whole Genome Sequencing*)

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 RESUMEN

En esta tesis se ha abordado el estudio de las bases genéticas de los trastornos del espectro autista (TEA) desde diferentes perspectivas con el objetivo de analizar la contribución al riesgo, tanto de la variación común, como de la variación rara. En primer lugar, se ha llevado a cabo un GBA, usando como *input* los *summary statistics* del mayor metaanálisis de GWAS llevado a cabo en los TEA, que ha permitido identificar nuevos *loci* asociados. Dichos *loci* fueron posteriormente anotados funcionalmente. En segundo lugar, se ha comprobado que las mutaciones postcigóticas (PZMs) tienen un papel importante en la etiología de los TEA y que los genes con PZMs están implicados en mecanismos biológicos diferentes a los de los genes con mutaciones germinales. Finalmente, se ha estimado el rendimiento diagnóstico de la secuenciación de exoma completo en una cohorte gallega de TEA y se ha comprobado su utilidad en el diagnóstico genético de individuos con sospecha previa de un síndrome genético.

In this thesis it has been addressed the study of the genetic basis of Autism Spectrum Disorders (ASD) from different perspectives with the goal of analyzing the contribution of common and rare variation to the risk. First, it has been performed a GBA, using as input the summary statistics from the biggest GWAS meta-analysis in ASD, which has allowed the identification of new associated loci. These loci were then functionally annotated. Secondly, it has been demonstrated that postzygotic mutations (PZMs) play an important role in the etiology of ASD and genes carrying PZMs participate in biological process different from those of genes with germinal mutations. Finally, it was estimated the diagnostic yield of whole exome sequencing and it was also verified its utility in the diagnosis of patients with a prior suspicion of a genetic syndrome.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

2 INTRODUCCIÓN

2.1 LOS TRASTORNOS DEL ESPECTRO AUTISTA (TEA)

2.1.1 Definición y criterios diagnósticos

Los TEA fueron definidos por primera vez en 1943 por el psiquiatra infantil Leo Kanner al observar a 11 niños que compartían dos características comunes:

1. Falta de interés por el mundo social.
2. Comportamiento definido como de resistencia al cambio o adherencia a la monotonía[1].

La descripción clínica que ofreció Kanner entonces se ha mantenido a lo largo de los años y actualmente algunas de estas características se reconocen como criterios diagnósticos en los TEA.

De acuerdo con la quinta edición del Manual Diagnóstico y Estadístico de los Trastornos Mentales o DSM-5 (de sus siglas en inglés, *Diagnostic and Statistical Manual of Mental Disorders, fifth edition*), los TEA se definen como un grupo heterogéneo de trastornos del neurodesarrollo (TND) que se caracterizan por su aparición en etapas tempranas del desarrollo[2]. Su diagnóstico se basa en el cumplimiento de los criterios que se resumen en la Tabla 1.

Criterio A. **Déficits persistentes en la comunicación e interacción social en múltiples contextos que se manifiesten por los siguientes síntomas, actualmente o por los antecedentes:** **A.1 Déficits en la reciprocidad socio-emocional.** **A.2 Déficits en los comportamientos de comunicación no verbales usados para la interacción social.** **A.3 Déficits en el desarrollo, establecimiento y comprensión de las relaciones sociales.**
Criterio B. **Patrones restrictivos y repetitivos de comportamiento, actividades e intereses, que se manifiestan por al menos dos de los siguientes puntos:** **B.1 Movimientos, uso de objetos y habla de manera estereotipada y repetitiva.** **B.2 Resistencia al cambio, adherencia inflexible a las rutinas, patrones ritualizados de comportamiento verbal y no verbal.** **B.3 Intereses muy restringidos y fijos que son anómalos en intensidad y foco de interés.** **B.4 Hipo reactividad o hiper reactividad a estímulos sensoriales o interés anormal en aspectos sensoriales del ambiente.**
Criterio C. **Estos síntomas deben estar presentes en periodos tempranos del desarrollo aunque hay que tener en cuenta que puedan no manifestarse plenamente hasta que las capacidades limitadas no permitan responder a las exigencias sociales, o bien queden enmascarados por estrategias aprendidas en fases posteriores de la vida.**
Criterio D. **Estos síntomas suponen un deterioro clínico en el ámbito social, ocupacional y en otras áreas del funcionamiento.**
Criterio E. **Estas alteraciones no se explican mejor por el hecho de presentar discapacidad intelectual (DI) o retraso global del desarrollo (RGD). Para hacer un diagnóstico de comorbilidades de los TEA y DI, la comunicación social debe estar por debajo del nivel de desarrollo esperado.**

Tabla 1. Criterios diagnósticos para los TEA de acuerdo al DSM-5.

El DSM-5 establece el nivel de gravedad de los TEA en base al grado de alteración de la comunicación social y en base a los comportamientos restringidos y repetitivos, diferenciando tres niveles de gravedad que se resumen en la Tabla 2.

Niveles de gravedad en los TEA en base al DSM-5	Comunicación social	Comportamientos restringidos y repetitivos
Nivel 1	Dificultad en iniciar la interacción social y una respuesta atípica o sin éxito a la apertura social con los otros.	La inflexibilidad en el comportamiento causa interferencia con el normal funcionamiento en uno o más contextos.
Nivel 2	Déficits marcados en las habilidades de comunicación verbales y no verbales, que persisten incluso con apoyo. Hay una dificultad en la iniciación de la interacción social y una respuesta inadecuada a la apertura con los otros	Los comportamientos restringidos y repetitivos aparecen con suficiente frecuencia como para ser obvios para un observador y para interferir en el funcionamiento normal.
Nivel 3	Déficits graves en las habilidades de comunicación verbales y no verbales que causan una grave alteración en el funcionamiento, graves limitaciones en la iniciación de la interacción social y mínima respuesta a la apertura social con los otros	La inflexibilidad en el comportamiento y a los cambios o el comportamiento restringido y repetitivo causan una grave interferencia en el funcionamiento en todas las áreas

Tabla 2. Niveles de gravedad de los TEA de acuerdo al DSM-5.

2.1.2 Prevalencia de los TEA

En epidemiología, se denomina prevalencia al número o proporción de individuos de un grupo o una población en un momento o en un período de tiempo determinado. La prevalencia, por tanto, describe la proporción de individuos con TEA en la población de estudio.

Entre 1960 y 1970, la prevalencia de los TEA se estimaba en 2-4 casos por cada 10000 niños no afectos. Desde entonces, las estimas de prevalencia que se han realizado han aumentado casi de una manera exponencial[3,4]. El estudio más reciente y completo, llevado a cabo por la Red de Vigilancia del Autismo y las Discapacidades del Desarrollo o ADDM (de sus siglas en inglés, *Early Autism and Developmental*

Disabilities Monitoring), estimó en 2014 que la prevalencia de los TEA era de 1 afecto por cada 59 no afectos en niños de 8 años en los Estados Unidos. Esta estima significa un incremento de alrededor de un 15% en comparación con los datos ofrecidos por el ADDM en 2012, cuando se calculó una prevalencia de 1 afecto por cada 68 no afectos[5]. Este aumento en la estima de prevalencia puede deberse a la mejora de las herramientas diagnósticas y al reconocimiento temprano de los síntomas nucleares de los TEA[4]. Dado que estos criterios son difíciles de homogeneizar entre estudios resulta difícil hacer una estimación de su prevalencia mundial, pero estudios recientes calculan que esta es cercana al 1%[6].

Por otro lado, la proporción de hombres afectos y mujeres afectas es diferente[4,5]. Los últimos datos estiman que el ratio por sexo es 3:1, siendo mayor la afectación en hombres que en mujeres[7].

2.1.3 Comorbilidades en los TEA

La comorbilidad es un término usado para describir la presencia simultánea de dos o más condiciones nosológicas (enfermedades o trastornos) en un mismo individuo.

En los TEA, más de un 70% de los individuos afectos tienen comorbilidad con algún otro trastorno clínico o psiquiátrico[7]. Las condiciones médicas más frecuentes en los TEA son la epilepsia, las alteraciones gastrointestinales y los trastornos del sueño. A su vez, los TEA se acompañan frecuentemente de otros TND, incluyendo la DI, el trastorno por déficit de atención e hiperactividad (TDAH) y trastornos motores, así como trastornos psiquiátricos como la ansiedad, la depresión o el trastorno obsesivo compulsivo (TOC)[8] (Tabla 3). El DSM-5 aconseja registrar la información de comorbilidades a la hora de establecer el diagnóstico, de manera que quede recogida qué alteración médica, trastorno genético, TND o trastorno psiquiátrico tiene asociado el paciente con TEA. A su vez, también contempla la necesidad de especificar el nivel cognitivo (presencia de DI o no), el nivel lingüístico, y la presencia o no de catatonia en el paciente[2].

	Proporción de individuos afectos	*Comentarios*
	Trastornos del neurodesarrollo	
Discapacidad intelectual	45%	Las estimas de prevalencia dependen de los criterios diagnósticos y de la definición de inteligencia usada en cada caso (por ejemplo, si las habilidades verbales se usan o no como criterio).
Trastornos del lenguaje	Variable	En el DSM-IV, el retraso en la adquisición del lenguaje se consideraba un criterio diagnóstico de los TEA. Sin embargo, en el DSM-5 ya no se incluye dicho criterio aunque se acepta que los TEA tengan un desarrollo del lenguaje atípico.
Trastorno por déficit de atención e hiperactividad	28-44%	En el DSM-IV no se aceptaba la posibilidad de un diagnóstico comórbido de los TEA y el TDAH. Esto ya no ocurre en el DSM-5.
Tics	14-38%	Aproximadamente un 6.5% presenta síndrome de Tourette.
Alteraciones motoras	<79%	Se incluye el retraso motor, la hipotonía, la catatonia, los déficits en la coordinación, planificación y preparación del movimiento, y en la marcha.
	Alteraciones médicas	
Epilepsia	8-30%	Su frecuencia está aumentada en individuos con DI o trastornos genéticos.
Problemas gastrointestinales	9-70%	Se incluye la constipación crónica, el dolor abdominal, la diarrea crónica, el reflujo gastro-esofágico. Se incluyen también trastornos como gastritis, esofagitis, enfermedad celiaca, enfermedad de Crohn, enfermedad intestinal inflamatoria o colitis.
Alteraciones del sistema inmune	<38%	Alergias o trastornos autoinmunitarios.
Trastornos genéticos	5%-10%	Algunos ejemplos son el síndrome X frágil, síndrome de Rett, esclerosis tuberosa, síndrome de Down o fenilcetonuria.
Alteraciones del sueño	50-80%	Insomnio es el más común.
	Trastornos psiquiátricos	
Ansiedad	42-56%	Los trastornos más comunes son el trastorno de ansiedad social (fobia social) y el trastorno de ansiedad generalizada.
Depresión	12-70%	Más frecuente en pacientes adultos que en niños.

Trastorno obsesivo compulsivo	12-17%	Es importante distinguir entre comportamientos repetitivos que no se acompañan de pensamientos obsesivos o intrusivos (criterio diagnóstico en los TEA) de los que sí lo hacen (parte del TOC).
Trastorno psicótico	12-17%	Más frecuente en adultos.
Abuso de sustancias	16%	Se debe, principalmente, al abuso de sustancias para aliviar la ansiedad.
Trastorno oposicionista-desafiante	16-28%	Los comportamientos oposicionistas pueden ser una manifestación de la ansiedad, resistencia al cambio, la dificultad para entender el punto de vista del otro y la escasa preocupación del efecto que pueda tener el comportamiento propio en los demás.
Trastornos alimenticios	4-5%	Puede haber un diagnóstico erróneo de TEA, sobre todo en mujeres, puesto que los trastornos alimenticios se acompañan también de comportamientos rígidos, inflexibilidad cognitiva, alta preocupación por los detalles y por uno mismo.

Tabla 3. Condiciones comórbidas más frecuentes en los TEA. (Adaptada de "Autism", Lancet, 2014[8]).

2.1.4 Factores de riesgo

Los TEA son TND complejos que presentan una etiología multifactorial en la cual participan tanto factores genéticos como ambientales.

Se calcula que aproximadamente un 17% de su etiología se explica por factores ambientales[9]. Así, una edad paterna o materna avanzada en el momento de la concepción es un factor de riesgo conocido. Por otro lado, la exposición prenatal a sustancias exógenas (ácido valproico, talidomida, pesticidas químicos o fármacos psiquiátricos) o la presencia de ciertas condiciones biológicas como la obesidad materna o infeccionas maternas durante el embarazo, aumentan el riesgo de TEA. Por último, se han descrito también factores de riesgo postnatales que actúan en los primeros años de vida como infecciones, alergias o exposición de los niños a ciertas drogas[10,11].

El porcentaje restante de su etiología se explica principalmente por factores genéticos que se tratarán en detalle en el apartado siguiente.

2.2 GENÉTICA DE LOS TEA

2.2.1 Epidemiología genética.

Para cuantificar la importancia que tienen los factores genéticos en la etiología de los TEA se usa el concepto de heredabilidad. La heredabilidad se define como la proporción de la varianza fenotípica que se atribuye a la variación genética en una población determinada[12]. Las estimas de la heredabilidad varían en un rango de 0 a 1. Un valor cercano a 0 indica que la varianza observada para un determinado carácter biológico en una población se atribuye muy poco a la variación genética existente entre sus individuos. Un valor cercano a 1, en cambio, indica que la varianza observada para un carácter biológico en una población se atribuye en gran medida a la variación genética existente entre sus individuos[13].

La estima de la heredabilidad se basa en el hecho de que los individuos relacionados por parentesco comparten parte de su genoma. En función del grado de compartición de su genoma (por ejemplo, 50% para parejas de hermanos de mismos padres o 25% para abuelos-nietos), se espera también un grado de compartición fenotípica entre individuos. Sin embargo, se debe tener en cuenta que la expresión de un fenotipo es el resultado de una interacción entre factores genéticos y ambientales.

Los estudios en gemelos monocigóticos (MC) y dicigóticos (DC) permiten separar con gran precisión los efectos ambientales de los genéticos. En ese sentido, los gemelos MC comparten el 100% de su genoma, y, por tanto, cualquier variación fenotípica entre parejas de gemelos se debe exclusivamente a factores ambientales. Por otro lado, los gemelos DC comparten el 50% de su genoma. A partir del porcentaje de concordancia fenotípica en gemelos MC se puede calcular la heredabilidad. La concordancia entre gemelos DZ se usa para corregir la estimación anterior[13].

En los TEA, los estudios en gemelos MC revelan que la concordancia del fenotipo autista es significativamente superior (70-90%) a la concordancia de gemelos DC (0-30%)[14,15]. Las estimas de la heredabilidad, sin embargo, difieren mucho entre un estudio y otro, aunque un metaanálisis reciente, que incluyó los principales estudios de

gemelos llevados a cabo en los TEA, la sitúa en torno a un 64-91% (Figura 1)[16].

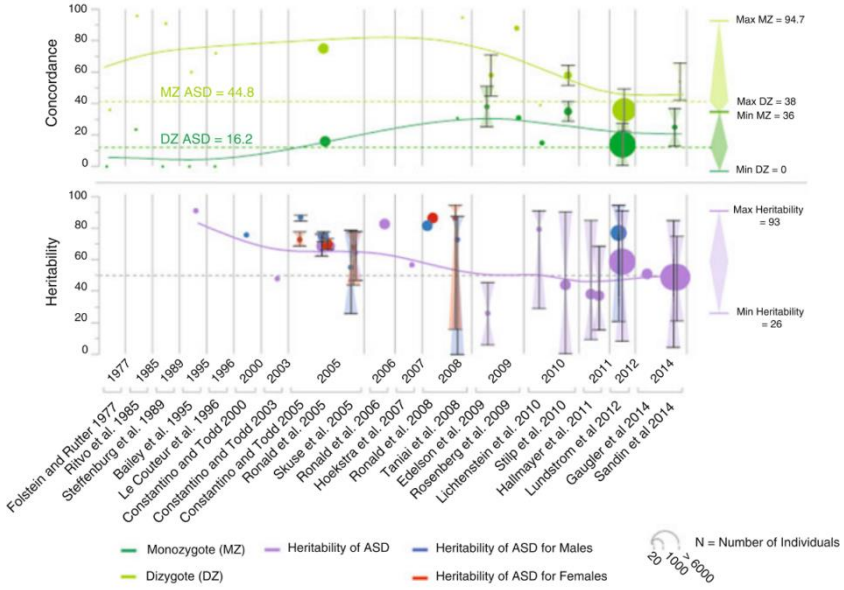

Figura 1. Principales estudios de gemelos MC y DC llevados a cabo en los TEA entre 1977 y 2014. En la parte superior de la figura se representan los valores de concordancia fenotípica entre gemelos MC y DC, mientras que en la parte inferior se representan las estimas de heredabilidad obtenidas en cada estudio. (Extraída de Huget and Bourgeron, 2016[310], permitido por Springer Nature).

Otra aproximación para estimar la heredabilidad consiste en medir la correlación de la enfermedad entre parientes de diferente parentesco procedentes de una muestra aleatoria de la población. Este tipo de estudios tienen la ventaja, con respecto a los estudios de gemelos, de estimar también el riesgo de recurrencia relativa (RR), que se define como el riesgo de padecer la enfermedad si un miembro de la familia se encuentra ya afecto, en comparación con una población control.

En los TEA se ha calculado que el RR es de 153 para gemelos MC, 8.2 para gemelos DC, 10.3 para hermanos completos, 3.3 para hermanos maternos, 2.9 para hermanos paternos y 2 para primos[17,18]. La heredabilidad de los TEA calculada por esta aproximación se estima en un 83%[9].

A partir de todos estos estudios se ha demostrado que en la etiología de los TEA intervienen principalmente factores genéticos, aunque no se puede excluir la influencia, en menor medida, de factores ambientales.

Las estimaciones de la elevada heredabilidad de los TEA en los estudios de gemelos y los estudios en familias, han motivado la realización de un número elevado de estudios genéticos cuyo objetivo principal ha sido dilucidar las bases genéticas de estos TND, usando diferentes aproximaciones.

2.2.2 Factores genéticos

2.2.2.1 Arquitectura genética

La secuenciación del genoma humano reveló que un 99.9% de nuestro material genético es compartido mientras que solo un 0.1% difiere de un individuo a otro y es lo que se conoce como variación genética[19].

La variación genética existente entre individuos puede ser respecto a la secuencia del ADN o respecto a la estructura submicroscópica de los cromosomas. Según su frecuencia poblacional puede ser común (frecuencia del alelo menor o MAF > 1%) o rara (MAF < 1%) y según el tipo de herencia puede ser heredada o *de novo* (Tabla 4)[20,21].

En general, se asume que la variación común contribuye de manera pequeña o moderada al riesgo de la enfermedad, mientras que las variantes raras tienen un impacto mayor en la aparición de la enfermedad y sus frecuencias alélicas se mantienen bajas como consecuencia de la selección natural negativa. El caso extremo lo constituyen las variantes *de novo*, que, al surgir espontáneamente en una nueva generación, no tienen tiempo de propagarse en una población si su impacto en el individuo es muy negativo[22].

Criterio de clasificación	*Tipo de variante*
Frecuencia poblacional	**Variante común**: Cualquier cambio en el genoma de referencia que esté presente en más de un 1% de los individuos de una población. El ejemplo clásico son los SNPs (de sus siglas en inglés *Single Nucleotide Polymorphisms)* que son variantes de una sola base que se encuentran a una frecuencia superior del 1% en la población.
	Variante rara: Cualquier cambio en el genoma de referencia que esté presente en menos de un 1% de los individuos de una población.
Herencia	**Variante heredada**: Cualquier cambio en el genoma de referencia que esté presente en alguno de los progenitores y se transmita a sus hijos.
	Variante *de novo:* Mutación espontánea que surge en el espermatozoide, óvulo o primeras divisiones del cigoto y por tanto, están en el hijo y en ninguno de sus progenitores.
Estructura	**Variantes estructurales:** Las variantes estructurales son cambios con respecto a la estructura submicroscópica de los cromosomas. En este grupo se incluyen variantes en el número de copias o CNVs (del inglés, *Copy Number Variation*) elementos móviles, inversiones y translocaciones. Las CNVs son variantes estructurales que presentan genralmente un tamaño superior a 1000 pb.
	SNVs (de sus siglas en inglés, *Single Nucleotide Variants)*: Los SNVs son cambios de una sola base con respecto al genoma de referencia que no son comunes en la población.
	***Indels*:** Los *Indels son* cambios en el genoma de referencia que incluyen la inserción o deleción de un número pequeño de bases (generalmente su número es inferior a 1000 pb).
Localización	**Codificantes:** Cualquier cambio en el genoma de referencia localizado en el 1% del genoma que codifica funcionalmente para proteínas.
	No codificantes: Cualquier cambio en el genoma de referencia localizado fuera del 1% del genoma que codifica funcionalmente para proteínas y que incluye intrones, elementos reguladores y segmentos intergénicos.

Tabla 4. Tipos de variantes genéticas clasificadas en función de su frecuencia poblacional, herencia, estructura y localización en el genoma. (Adaptada de Geschwind *et al.*, 2015[23])

La arquitectura genética es la base genética subyacente a la variabilidad fenotípica de un trastorno. El estudio de la arquitectura genética de los TEA, comprende pues, la identificación de todas las variantes genéticas existentes en la población que confieren riesgo en este TND[24]. Tanto la variación común, como la variación rara, contribuyen al riesgo en el caso de los TEA. Así, un 50% del componente genético de los TEA, se puede explicar por la variación genética común. Las variantes raras, aunque en conjunto tienen una contribución limitada a su heredabilidad, confieren un alto riesgo a nivel individual, lo cual explica la importancia de su estudio (Figura 2)[25].

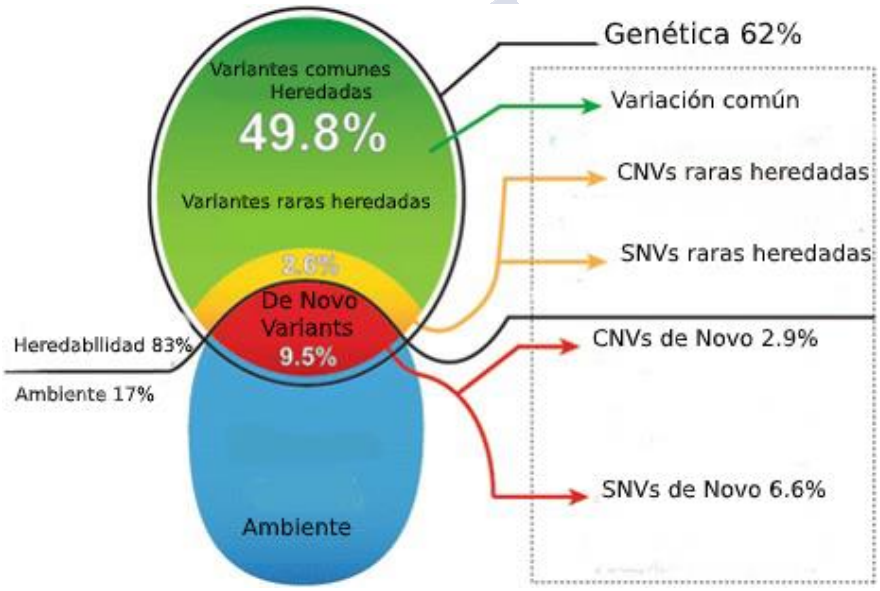

Figura 2. La arquitectura genética de los TEA. (Adaptada de Huget and Bourgeron, 2016[310], permitido por Springer Nature).

Los estudios genéticos que han intentado elucidar la arquitectura genética de los TEA se desarrollarán en los apartados siguientes.

2.2.2.2 Estudios genéticos llevados a cabo en los TEA

2.2.2.2.1 Estudios de ligamiento

Los primeros estudios genéticos llevados a cabo en los TEA con el objetivo de identificar variantes de susceptibilidad, fueron los estudios de ligamiento. En la actualidad, estos estudios ya no se realizan, pero tuvieron un papel relevante en la posterior identificación de los primeros genes asociados a los TEA.

En los estudios de ligamiento se buscan marcadores genéticos (cuya localización física es conocida) que cosegregan con la enfermedad en familias con varios miembros afectos. Es decir, se identifica qué variante alélica de un marcador genético comparten todos los miembros afectos.

Los marcadores seleccionados se encuentran distribuidos aleatoriamente por todo el genoma, lo cual permite no contar con una hipótesis previa de qué regiones o genes podrían estar implicados. Sin embargo, una vez se identifica una región genética de interés, esta se puede delimitar todavía más al cubrirla con más marcadores, lo que permite priorizar después los genes de la región por su función. Inicialmente, los marcadores genéticos empleados en los estudios de ligamiento eran los microsatélites, que son secuencias nucleotídicas de 2 a 6 pares de bases que se repiten de manera consecutiva. Sin embargo, estos han sido sustituidos progresivamente por SNPs en los estudios de ligamiento posteriores[26].

Los estudios de ligamiento se basan en la transmisión conjunta o cosegregación de marcadores genéticos que se encuentren próximos entre sí, propiedad denominada ligamiento. El ligamiento depende a su vez de la frecuencia de recombinación. El estadístico usado para medir la existencia de ligamiento o no es el LOD score. Este estimador, permite detectar la existencia de ligamiento entre un marcador genético y la variante de susceptibilidad al estimar la frecuencia de recombinación más probable entre los genes de la especie humana[26].

Los primeros estudios de ligamiento en los TEA incluyeron tamaños muestrales de menos de 300 familias. Estos estudios, pese a su escasa potencia estadística, identificaron numerosas regiones de interés asociadas a los TEA, pero muy pocas se replicaron (Tabla 5)[27].

Estudio	Número de familias	Región cromosómica identificada
Ashley-Koch *et al*. (1999)	76	7q22.1-q31.2
Auranen *et al*. (2002	38	3q25-27
Barrett *et al*. (1999)	75	7q31-33 13
Bartlett *et al*. (2005)	303	1q23-24 17q11
Buxbaum *et al*. (2001)	95	2q
Cantor *et al*. (2005)	91	3p14-12 17q11-23
Coon *et al*. (2005)	1	3q25-27
IMGSAC (2001)	152	2q24-q33 7q 16p
IMGSAC (1998)	99	7q
Lamb *et al*. (2005)	219	2,7,9
Liu *et al*. (2001)	110	5, 19, X
McCauley *et al*. (2005)	158	3p25 17q11 19p13
Philippe *et al*. (1999)	51	6q21
Risch *et al*. (1999)	90 (fase 1) y 49 (fase 2)	1
Shao *et al*. (2002a)	99	3
Vincent *et al*. (2005)	22	Xq27-28
Yonan *et al*. (2003)	345	5,11,17
Ylisaukko-Oja *et al*. (2004)	17	1q21-22 3p14-24 D13q31-33
Ylisaukko-Oja *et al*. (2006)	314	1p12-q25 3p24-26 4q21-31 6q14-21 7q33-36 8q22-24 17p12-q21

Tabla 5. Principales estudios de ligamiento llevados a cabo en los TEA.(Adaptada de C. Freitag, 2007[27]).

Posteriormente, se crearon grandes consorcios como el *Autism Genetics Resource Exchange* (AGRE) o el *Autism Genome Project Consortium* (AGP), en los cuales grupos de todo el mundo contribuyeron a la creación de grandes cohortes de TEA. Pese a este incremento del tamaño muestral no se lograron, tampoco, resultados satisfactorios, con la excepción de la identificación de la región 20p13[28–30].

Esta falta de reproducibilidad entre estudios, puso de manifiesto, ya no solo la heterogeneidad genética de los TEA, sino también la enorme heterogeneidad fenotípica que los caracteriza. Para solventar este problema, surgieron estrategias cuyo objetivo era reunir cohortes más homogéneas. Así, la estratificación por sexo del probando, el lenguaje u otras características comportamentales de los individuos afectos con TEA permitió incrementar las señales obtenidas con respecto a estudios previos. De esta manera, se encontró asociación con alelos de riesgo específicos de sexo en la región 17q21[31,32]. Otra estrategia exitosa fue la estratificación de las muestras de estudio en base a endofenotipos. Los endofenotipos constituyen fenotipos estables, heredables y fácilmente medibles, cuya base genética es fácilmente identificable[33]. Al clasificar por endofenotipos en lugar de por categorías diagnósticas, se pueden incluir también familiares de individuos con TEA, que, sin pasar el umbral de significación clínica, presentan algunas características cognitivas o de conducta propias de los TEA (fenotipo ampliado de los TEA). El estudio más exitoso al respecto, fue el llevado a cabo por Alarcon *et al.*, en el cual se encontró una asociación entre el retraso en la adquisición del lenguaje y la región 7q35[34].

2.2.2.2.2 Estudios de asociación con genes candidatos

Los estudios de genes candidatos buscan la asociación entre variantes seleccionadas a *priori* (generalmente SNPs) en genes candidatos, y la enfermedad o trastorno.

La selección de genes candidatos puede realizarse en base a un conocimiento previo de su función y su posible implicación en la fisiopatología de la enfermedad, o bien a su localización en regiones que han sido previamente señaladas en estudios de ligamiento.

El planteamiento de dichos estudios consiste en comparar las frecuencias alélicas de los SNPs seleccionados en estos genes candidatos, entre casos (afectos) y controles (no afectos). Para determinar la asociación de un alelo con el trastorno, se debe determinar si su frecuencia es significativamente superior en casos con respecto a controles. Para ello se realiza un test estadístico, Chi-cuadrado ($\chi 2$), que proporciona un p-valor para cada SNP. Por otro lado, la OR (*odds ratio*) y su intervalo de confianza (95% IC), proporcionan información

cuantificable sobre el riesgo que confiere un alelo. De esa manera, una OR < 1 significa que el alelo confiere protección mientras que una OR > 1 significa que el alelo confiere susceptibilidad[35].

En el caso de los TEA, los estudios de genes candidatos se han realizado seleccionando y priorizando genes por su función en regiones previamente identificadas en estudios de ligamiento. Sin embargo, pese al interés que despertaron en el pasado, su éxito fue muy limitado. Así, incluso los genes más frecuentemente asociados como *RELN* no han sido replicados en otros estudios de genes candidatos (Tabla 6)[36].

Estudio	Gen	Región cromosómica	Replicación de resultados
Yirmiya N, *et al*. (2006)	*AVPR1A*	12q14-12q15	Sin replicación
Strauss KA, *et al*.(2006), Alarcon M, *et al*. (2008)	*CNTNAP2*	7q35	Replicación independiente
Kilpinen H, *et al*. (2007)	*DISC1*	1q42	Sin replicación
Gharani N, *et al*. (2004), Benayed R, *et al*. (2005)	*EN2*	7q36	Replicación independiente
Samaco RC, *et al*. (2005), Cook EH Jr *et al*. (1998), Buxbaum JD, *et al*. (2002)	*GABRB3*	15q11-15q12	Replicación independiente
Jamain S, *et al*. (2002), Shuang M, *et al*. (2004)	*GRIK2*	6q21	Replicación independiente
Weiss LA, *et al*. (2006), Coutinho AM, *et al*. (2007)	*ITGB3*	17q21	Replicación independiente
Campbell DB, *et al*. (2006), Campbell DB, *et al*. (2007)	*MET*	7q31	Replicación interna
Wu S, *et al*. (2005), Lerer E, *et al*. (2007)	*OXTR*	3p25	Su asociación con el estatus afecto ha sido replicada
Persico AM, *et al*. (2001), Li H, *et al*. (2007)	*RELN*	7q22	Replicación independiente
Ramoz N, *et al*. (2004), Segurado R, *et al*. (2005)	*SLC25A12*	2q24	Replicación independiente
Kim SJ, *et al*. (2002), Sutcliffe JS, *et al*. (2005)	*SLC6A44*	17q11	Replicación independiente

Tabla 6. Principales estudios de genes candidatos llevados a cabo en los TEA. (Adaptada de B. Abrahams and Daniel H. Geschwind, 2009[37]).

Este hecho ha propiciado que los estudios de genes candidatos hayan sido progresivamente sustituidos por los estudios de asociación de genoma completo o GWAS (de sus siglas en inglés, *Genome Wide Association Study*) cuya estrategia de análisis es similar, pero permiten estudiar la totalidad de la variación común del genoma humano.

2.2.2.2.3 Estudios de asociación de genoma completo (GWAS) y análisis basados en genes (GBA)

Una teoría ampliamente extendida en el estudio de los factores genéticos que contribuyen al riesgo en los TEA, es el de la variante común-enfermedad común. De acuerdo con esta teoría, la susceptibilidad se debe al efecto de múltiples alelos de bajo riesgo que son comunes en la población. Los TEA, son TND complejos, y como tal, sería un error asumir que el total de su componente genético se debe a variantes comunes. Sin embargo, el peso de estas variantes en su etiología es importante, tal y como apuntan los estudios de heredabilidad[38].

2.2.2.2.3.1 Concepto de GWAS

Los GWAS analizan cientos de miles de variantes genéticas comunes (SNPs) distribuidas a lo largo de todo el genoma sin necesidad de hipótesis previa. La selección de estos SNPs, se realiza en base al desequilibrio de ligamiento o LD (de sus siglas en inglés, *Linkage Disequilibrium*). El LD es la propiedad por la cual los alelos de dos o más marcadores tienden a heredarse juntos en una población. La existencia de LD propicia la organización del genoma humano en bloques de LD que son característicos de cada población humana. Dicha estructura fue conocida gracias al desarrollo del proyecto HapMap, donde se genotiparon miles de SNPs en individuos procedentes de diferentes poblaciones. Gracias a esta propiedad, basta con seleccionar un SNP (tagSNP), con un elevado LD con los restantes SNPs presentes en el bloque haplotípico, para poder inferir a partir de él la variación genética de la región en un proceso llamado imputación, sin necesidad de genotipar todos los SNPs de un bloque[20,36]. Se calcula que gracias a esta estrategia es posible detectar más de un 80% de la variación común presente en el genoma humano usando *arrays* de genotipado que contienen entre 500000 y 1000000 de SNPs representativos.

El diseño típico de los GWAS consiste en seleccionar dos grupos: casos y controles. Todos los individuos se genotipan usando *arrays* de genotipado y a continuación se comparan las frecuencias alélicas de los SNPs entre ambos grupos. Si la frecuencia de algún SNP es superior en casos con respecto a los controles, este se considera asociado a la enfermedad. Esto quiere decir que, para cada alelo, se realiza un test

estadístico que determina su asociación o no con la enfermedad. Como el número de SNPs analizados es muy elevado es necesario realizar una corrección por múltiples test para fijar un nivel de significación y reducir la tasa de posibles falsos positivos. El método más común es la corrección de Bonferroni, en la cual se ajusta el valor estándar de $\alpha = 0.05$ según la fórmula ($\alpha_{corregido} = 0.05 / k$), donde k es el número de test analizados. En términos generales, el umbral de significación estadística en un GWAS típico donde se analizan 10^6 SNPs es de $p < 5 \times 10^{-8}$.

Los niveles de asociación de los SNPs de un GWAS se representan habitualmente en un *Manhattan plot*, en el cual se muestran todas las señales analizadas con respecto a su posición genómica (cada cromosoma tiene un color diferente) (Figura 3). Los resultados de múltiples GWAS que se realicen bajo la misma hipótesis y cuyo diseño sea similar, pueden analizarse conjuntamente en los llamados metaanálisis. De esta manera, al incrementarse el tamaño muestral y la potencia estadística, se pueden identificar nuevos SNPs asociados[39].

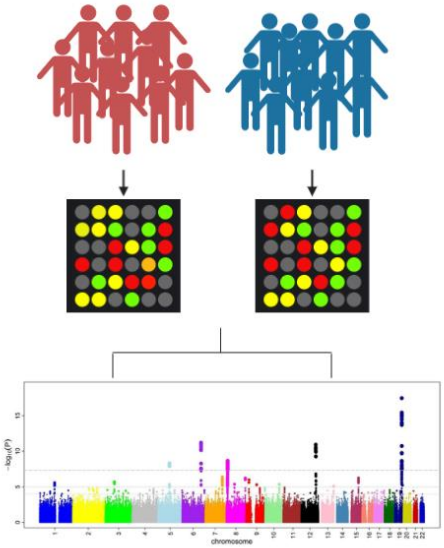

Figura 3. Estudios de asociación de genoma completo o GWAS. En primer lugar, se seleccionan dos grupos (casos y controles) y se genotipan todos los individuos de cada grupo. A continuación, se comparan las frecuencias alélicas de los SNPs entre casos y controles y se representan los SNPs analizados en un *Manhattan plot*.

.

2.2.2.2.3.2 Resultados de los principales GWAS llevados a cabo en los TEA

Los primeros GWAS llevados a cabo en los TEA no fueron lo exitosos que se esperaba debido probablemente a la heterogeneidad fenotípica del trastorno y a un tamaño muestral insuficiente para poder alcanzar la potencia estadística necesaria[28,40,41]. En ellos se encontraron solo algunos SNPs significativos que señalaban a posibles genes candidatos: rs10513025 ($p = 2 \times 10^{-7}$) localizado entre *SEMA5A* y *TAS2R1*[28]; rs4141463 ($p = 4.7 \times 10^{-8}$) localizado en una región intrónica de *MACROD2*[40,42]*;* rs4307059 ($p = 2.1 \times 10^{-10}$) localizado entre *CHD10* y *CHD9* y cerca del pseudogen *MSNP1* que codifica para un ARN no codificante (*MSNP1AS*)[43,44]. Estas señales, sin embargo, no han vuelto a ser replicadas en estudios posteriores de manera que no se ha podido confirmar la contribución de dichos SNPs al riesgo en los TEA[45].

En los últimos años, el PGC (de sus siglas en inglés, *Psychiatric Genomics Consortium*) ha hecho un enorme esfuerzo por incrementar el tamaño muestral de las cohortes incluidas en los GWAS gracias a la colaboración coordinada de multitud de grupos de investigación de todo el mundo.

El primer trabajo llevado a cabo por el PGC en los TEA, consistió en un metaanálisis de 14 cohortes independientes, reuniendo un tamaño muestral de 7387 casos y 8567 controles. Pese a este incremento muestral con respecto a los estudios previos, no se identificó ningún SNP que alcanzara el umbral de significación estadística requerido ($p < 5 \times 10^{-8}$). Sin embargo, algunas de las señales marginalmente significativas señalaban a genes con un papel importante en el neurodesarrollo como *EXT1*, *ASTN2*, *MACROD2*, y *HDAC4*. También se identificó claramente una correlación genética entre la esquizofrenia y los TEA[46]. Este estudio hizo evidente la necesidad de seguir incrementando el tamaño muestral y el desarrollo de *pipelines* informáticos que permitieran realizar un correcto control de calidad de los datos, además de una correcta imputación[47].

El mayor GWAS realizado en los TEA hasta la fecha reunió un tamaño muestral de alrededor de 18000 casos y 28000 controles. Gracias a este incremento se consiguió por primera vez detectar 93 SNPs estadísticamente significativos en 3 *loci* separados, de los cuales 53 fueron replicados en cohortes independientes (Tabla 7).

Index SNPs	Chr	pb	p-valor	A1/A2	Genes cercanos
rs910805	20	21248116	2.04×10^{-9}	A/G	KIZ, XRN2 NKX2-2, NKX2-4
rs10099100	8	10576775	1.07×10^{-8}	C/G	C8orf74 SOX7 PINX1
rs71190156	20	14836243	2.75×10^{-8}	GGTTTTTT /GG	MACROD2

Tabla 7. *Loci* que alcanzan el umbral de significación estadística ($p > 5 \times 10^{-8}$) en el mayor metaanálisis de GWAS llevado a cabo en los TEA. (Adaptada de Grove *et al.*, 2019[48]).

Además, se confirmó un solapamiento genético con esquizofrenia, así como con otros fenotipos, como la depresión mayor o el nivel educativo. La anotación funcional de las variantes asociadas determinó que éstas se localizan principalmente en elementos reguladores involucrados en la regulación de genes expresados durante la corticogénesis[48].

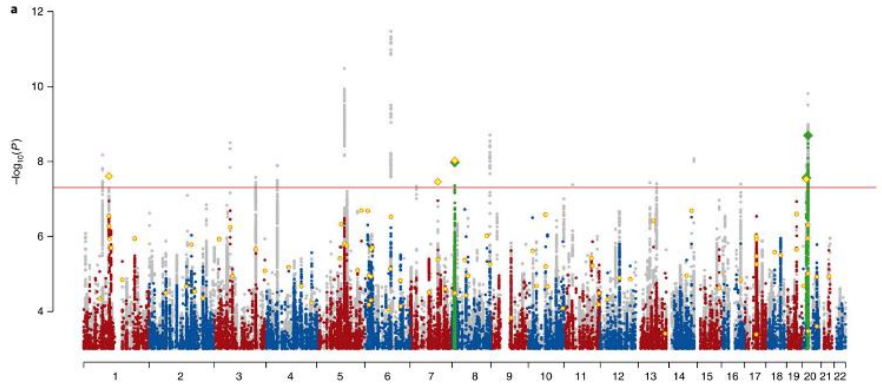

Figura 4. *Manhattan plot* que muestra los resultados del mayor GWAS realizado en los TEA hasta la fecha (Extraída de Grove *et al.*, 2019[48], con permiso de Springer Nature).

2.2.2.2.3.3 Análisis basados en genes (GBA)

El principal objetivo de un GWAS, es la detección de alelos que se asocien con una enfermedad y, por tanto, la unidad básica del análisis es el alelo en sí. Otra manera de analizar los resultados de un GWAS es considerar el gen como la unidad básica de análisis en los llamados análisis basados en genes o GBA (de sus siglas en inglés, *Gene-Based Analysis*). Los GBA, permiten obtener un estadístico para cada uno de los genes del genoma humano (p-valor). Este tipo de análisis tiene ciertas ventajas:

1. Por un lado, se reduce el número de test independientes realizados, ya que en lugar de hacer uno por cada uno de los SNPs analizados en un GWAS se hacen aproximadamente 20000 (uno por cada uno de los genes que hay en nuestro genoma).

2. Por otro lado, al considerar el gen como la unidad básica de estudio, se pueden realizar análisis secundarios, para tratar de identificar los mecanismos moleculares y celulares que están implicados en la enfermedad de estudio[49,50].

Tradicionalmente, los GBA usaban una aproximación basada en permutaciones, que requiere el uso de genotipos, para el cálculo del estadístico de cada gen. Los GBA más recientes, en cambio, solo necesitan los p-valores de cada SNP incluido en el GWAS (*summary statistics*). En general, se distinguen dos pasos principales:

1. Un primer paso de anotación génica, en el cual los SNPs se asignan a genes en función del genoma de referencia que se considere y los límites teóricos que se definen para el gen. Por defecto, los GBA suelen considerar el inicio y fin del gen en 50 kb *up* y *downstream.*
2. Cálculo del estadístico para cada gen teniendo en cuenta todos los SNPs del gen, o solo los más significativos. Para este paso, es necesario considerar la estructura de LD. Para ello, el método más común se basa en la construcción de matrices de correlación SNP-SNP a partir de datos genotípicos de una población externa cuya estructura genética sea lo más parecida a la empleada en el GWAS original.

Algunos ejemplos de GBA son los algoritmos predictivos de VEGAS2[51], MAGMA[52] y PASCAL[53]. Los dos primeros han sido usados con éxito en los dos metaanálisis llevados a cabo en los TEA por el PGC. Todos ellos permiten el uso de *summary statistics* como *input* sin la necesidad de usar la información genotípica y corrigen por LD usando los datos genotípicos del Proyecto 1000 Genomas, aunque el algoritmo empleado es diferente en cada caso. PASCAL aún no se ha usado en datos de TEA, pero su algoritmo ha demostrado ser más rápido y preciso que el de VEGAS o MAGMA[53].

En el primer trabajo del PGC, en el cual se empleó el algoritmo VEGAS2, no se detectó ningún gen asociado tras la corrección por Bonferroni ($p < 2.89 \times 10^{-6}$). Sin embargo, merece la pena destacar que la región 6p21.1, mostró 3 genes con p-valor = 7×10^{-6} (*ENPP4, ENPP5*, y *CLIC5)*[46].

En el segundo trabajo del PGC, se empleó el algoritmo MAGMA, y se detectaron 14 genes asociados (Tabla 8)[48]:

Gen	Cromosoma	MAGMA p-valor	tagSNP
XRN2	20	$9{,}69 \times 10^{-10}$	rs 910805
KCNN2	5	1.02×10^{-9}	rs 13188074
PLK1S1	20	5.17×10^{-9}	rs 910805
MACROD2	20	1.40×10^{-7}	rs 71190156
WNT3	17	4.03×10^{-7}	rs 142920272
MAPT	17	5.01×10^{-7}	rs 142920272
MFHAS1	8	5.58×10^{-7}	rs 11249905
XKR6	8	8.01×10^{-7}	rs 10099100
MSRA	8	9.15×10^{-6}	rs 10099100
CRHR1	17	1.07×10^{-6}	rs 142920272
SOX7	8	1.24×10^{-6}	rs 10099100
NTM	11	1.32×10^{-6}	rs 549507
MMP12	11	2.28×10^{-6}	rs 102751102
BLK	8	2.45×10^{-6}	rs 2736342

Tabla 8. Resultados del GBA realizado con MAGMA. (Adaptada de Grove *et al.*, 2019[48]).

Los GBA no deben entenderse nunca como un sustituto de los estudios GWAS clásicos, sino como herramientas complementarias que permiten acceder a otro nivel de información biológica, el nivel gen.

2.2.2.2.4 Estudios de CNVs

2.2.2.2.4.1 Definición de CNV y herramientas de detección

Las CNVs son variaciones estructurales del genoma, las cuales pueden ser ganancias (inserciones, duplicaciones o amplificaciones) o pérdidas (deleciones), en comparación con la secuencia genómica de referencia (Figura 5)[54].

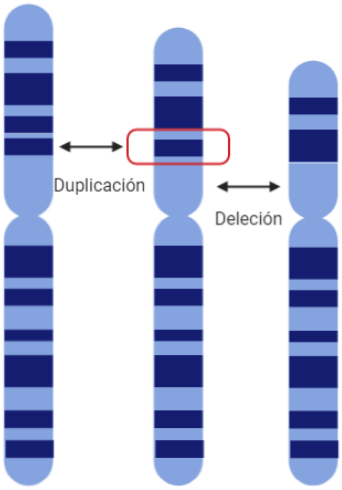

Figura 5. Variantes en el número de copias o CNVs. Se muestra una región duplicada a la izquierda y una región delecionada a la derecha, como ejemplos de CNV.

La implantación durante la última década de nuevas tecnologías de cariotipado molecular (*microarrays* cromosómicos), como la hibridación genómica comparada o CGH (de sus siglas en inglés, *Comparative Genome Hybridization*) y los paneles o *arrays* de SNPs, ha mejorado de manera significativa el poder de detección de este tipo de variantes que hasta entonces habían pasado desapercibidas a las técnicas citogenéticas convencionales, como el cariotipo o la hibridación fluorescente *in situ* o FISH (de sus siglas en inglés, *Fluorescence In Situ Hybridization*), que presentan muy baja resolución (orden de megabases (Mb)).

De especial relevancia fueron los *arrays* de SNPs, que, aunque en su diseño original estaban pensados para el genotipado en estudios de GWAS, también permiten detectar con gran precisión CNVs de pequeño tamaño en función de la densidad de SNPs incluidos en el *array*. En este tipo de *arrays*, los fragmentos de ADN de la muestra diana hibridan con sondas de oligonucleótidos correspondientes a variantes alélicas de SNPs seleccionados. La intensidad de la fluorescencia emitida cuando dichas sondas hibridan con fragmentos del ADN diana se compara con valores de referencia procedentes de un conjunto de individuos control de la misma población y genotipados con la misma tecnología (comparación indirecta), y esto permite identificar deleciones (intensidad inferior en comparación con individuos control) y duplicaciones (intensidad superior en comparación con individuos control).

En el caso de los *arrays* CGH los fragmentos de la muestra diana y la muestra control se marcan con diferentes fluorocromos e hibridan con oligonucleótidos inmovilizados en los *arrays*. La intensidad de la fluorescencia emitida por los dos tipos de fluorocromos determina los cambios en el número de copias: cuando solo se detecta fluorescencia del fluorocromo control significa que hay una deleción en esa posición de la muestra diana. Cuando la fluorescencia del fluorocromo de la muestra se detecta con mayor intensidad que la del control, significa que hay una duplicación en esa posición en la muestra diana[55].

No obstante, en los últimos años se han desarrollado nuevas herramientas que permiten la detección de toda clase de variantes estructurales (inserciones, deleciones, duplicaciones, inversiones y translocaciones) a partir de datos de secuenciación de genoma o exoma completo. Además de cubrir todo el espectro la variación estructural, esta aproximación permite detectar variantes de muy pequeño tamaño (< 50 pb)[56–59].

2.2.2.2.4.2 Principales estudios de CNVs llevados a cabo en los TEA

Las CNVs que se han visto implicadas en la etiología de los TEA han sido CNVs raras (MAF < 1%) ya sean heredadas o *de novo*.

Previamente al desarrollo de los *microarrays* cromosómicos, las técnicas citogenéticas clásicas pusieron de manifiesto la implicación de grandes aberraciones cromosómicas en casos graves de TEA[60]. Sin embargo, no fue hasta 2007 y 2008 cuando dos grupos publicaron importantes trabajos en los cuales se establecía por primera vez una relación causal entre las CNVs *de novo* y los TEA[61,62].

Entre 2010 y 2011 tres publicaciones corroboraron estos resultados previos, tras realizar un análisis de CNVs aplicando tecnologías de cariotipado de mayor resolución en dos grandes cohortes de TEA: la del AGP y la Simons Simplex Collection (SSC). Los tres estudios realizaron diseños caso-control incluyendo familias con uno o varios individuos afectos (familias con TEA esporádico o familiar respectivamente), y analizaron el papel de las CNVs raras *de novo* (≥ 20 kb) en los TEA[63–65].

Estos primeros trabajos sirvieron para proponer una hipótesis alternativa a la de enfermedad común-variante común, otorgándole mayor peso a la contribución de las variantes raras. De ellos se extrajeron una serie de conclusiones:

1) La frecuencia de CNVs *de novo* en afectos (5-10%) es significativamente superior a la de no afectos (1-2%), lo cual sugiere una implicación directa de estas variantes en el riesgo de los TEA.

2) La frecuencia de CNVs *de novo* es superior en individuos procedentes de familias con TEA esporádico que en individuos procedentes de familias con TEA familiar. Este hecho señala la implicación de diferentes mecanismos etiológicos en ambos tipos de TEA, siendo la contribución de CNVs raras heredadas superior en los TEA familiar que los esporádicos.

3) El riesgo que confiere una CNV en los individuos con TEA incrementa a medida que también lo hace su tamaño e implica a un mayor número de genes. Esto ocurre tanto para CNVs *de novo* como heredadas[63–65].

En los años posteriores, la inclusión de nuevas familias tanto en la SSC como en la cohorte del AGP, permitió reportar nuevos datos con respecto al papel de las CNVs en los TEA. En el trabajo de Sanders *et al*., se incluyeron los resultados correspondientes al análisis de la cohorte de la SSC al completo (10220 individuos con TEA procedentes

de 2591 familias) Esta cohorte, incluye principalmente casos de TEA esporádico, y por tanto las conclusiones que se extraen de él se relacionan específicamente con este tipo. En este trabajo, se observó, una mayor carga de CNVs *de novo* en mujeres con respecto a los hombres ($p = 0.04$). Además, las CNVs *de novo* en mujeres, eran, por lo general, de mayor tamaño y contenían más genes que las CNVs *de novo* detectadas en hombres ($p = 0.01$)[66]. Esta observación apoya la hipótesis propuesta por Jacquemont *et al.,* tan solo un año antes, conocida como el "modelo protector femenino", en la cual se sugiere que las mujeres requieren una carga mutacional superior que los hombres para lograr el estatus afecto[67]. Por otro lado, se propusieron dos mecanismos diferentes para entender el impacto que tiene una CNV en la expresión del fenotipo. En el caso de CNVs de gran tamaño (> 7 genes implicados) se sugiere que sean múltiples genes, cada uno con un efecto moderado, los que contribuyan al fenotipo. En el caso de CNVs *de novo* más pequeñas (≤ 7 genes implicados), es más posible que un único gen sea el responsable de la expresión del fenotipo[66].

Con respecto a los TEA familiares destaca un estudio desarrollado en 1532 familias con múltiples miembros afectos pertenecientes al AGRE. En él se identificó que en más de dos tercios de las familias en las cuales se detectaba una CNV de alto riesgo, la variante no era compartida por todos los miembros afectos. Esto indica, que al contrario de lo que se esperaba, las CNVs raras heredadas no son causa suficiente para desencadenar los TND y otros eventos genéticos o ambientales han de estar presentes para la expresión del fenotipo[68].

Tanto en el estudio de Sanders *et al.,* como en otros trabajos previos, se han identificado numerosas CNVs que aparecen de manera recurrente en individuos no emparentados entre sí[66,69,70]. Estas CNVs se caracterizan por presentar puntos de ruptura idénticos, lo que significa que tienen el mismo tamaño y el mismo contenido genómico[71,72]. Muchas de estas CNVs, denominadas recurrentes, se han detectado tanto en individuos sanos como en individuos afectos, así como en múltiples trastornos psiquiátricos, lo cual indica que su penetrancia para la expresión del fenotipo es incompleta y su expresividad es variable[73,74]. El análisis conjunto de los datos del AGP y la SSC,

permitió detectar asociación de diferentes CNVs con los TEA (Tabla 9)[66].

Banda	Región	Genes *RefSeq*	Genes	p-valor
1q21.1	chr1:146,467,203-147,801,691	13		6×10^{-9}
2p16.3	chr2:50,145,643-51,259,674	1	*NRXN1*	1×10^{-7}
3q29	chr3:195,747,398-196,191,434	7		0.07
7q11.23	chr7:72,773,570-74,144,177	22		0.005
7q11.23	chr7:72,773,570-73,158,061	10		0.0002
7q11.23	chr7:73,978,801-74,144,177	2	*GTF2I*, *GTF2IRD1*	0.0002
15q12	chr15:26,971,834-27,548,820	3	*GABRA5*, *GABRB3* *GABRG3*	1×10^{-10}
15q11.2-13.1	chr15:23,683,783-28,446,765	13		1×10^{-10}
15q13.2-13.3	chr15:30,943,512-32,515,849	7		0.005
16p11.2	chr16:29,655,864-30,195,048	27		1×10^{-10}
22q11.21	chr22:18,889,490-21,463,730	45		1×10^{-7}
22q13.33	chr22:51,123,505-51,174,548	1	*SHANK3*	0.07

Tabla 9. CNVs recurrentes asociadas a los TEA en la SSC y la cohorte del AGP. En la tabla se muestran las CNVs recurrentes identificadas en el análisis conjunto de los datos de la SSC y el AGP. Se especifica la banda y región cromosómica donde se localiza cada CNV, número de genes implicados en cada caso y se detallan los genes interrumpidos solo en aquellas CNVs que implican ≤ 3 genes. El p-valor mostrado es un p-valor corregido para múltiples comparaciones (Adaptada de Sanders *et al.*,2015[66]).

Los trabajos mencionados previamente demuestran que los *arrays* cromosómicos son herramientas eficaces en la detección de CNVs. Sin embargo, la limitación de esta aproximación reside en los intervalos genómicos que hay entre marcador y marcador. Por ese motivo, resulta muy difícil determinar con exactitud los límites de una CNV, o bien detectar CNVs de muy pequeño tamaño.

En lo últimos años se han desarrollado numerosos algoritmos que permiten la detección de todo tipo de variantes estructurales a partir de datos de secuenciación de exoma o genoma completo. Destacan, por ejemplo, los trabajos de Poultney *et al.*, y Brandler, *et al.*, donde se analizaron CNVs detectadas a partir de datos generados por secuenciación de exoma completo y genoma completo respectivamente. Con la detección de deleciones y duplicaciones muy

pequeñas (< 500pb), surgen, además, nuevas hipótesis en lo referente al impacto de las CNVs en el riesgo de los TEA. Brandler *et al.*, por ejemplo, propusieron, a partir de sus resultados, que la diferencia entre casos y controles no esté marcada por un número superior de CNVs *de novo* en afectos frente a no afectos, como los primeros estudios señalaban, sino por una proporción mayor de genes interrumpidos[56,57].

2.2.2.2.5 Secuenciación de nueva generación y estudios de secuenciación.

2.2.2.2.5.1 Evolución de las tecnologías de secuenciación de nueva generación.

El precedente para el desarrollo de las tecnologías de secuenciación tuvo lugar en 1977 cuando Frederick Sanger desarrolló la técnica de la "terminación de cadena" para secuenciar la molécula del ADN. Esta tecnología se denomina ahora secuenciación de primera generación o secuenciación Sanger[75].

La secuenciación Sanger fue la técnica de secuenciación empleada en el Proyecto Genoma Humano, que se desarrolló durante 13 años y culminó con la secuenciación del primer genoma humano en el año 2004[76,77]. Pese al gran avance en el campo de la genética que suponía este éxito, este proyecto puso de manifiesto la necesidad de desarrollar tecnologías más baratas, rápidas y de mayor rendimiento. Por ese motivo, el Instituto Nacional de Investigación del Genoma Humano o NHGRI (de sus siglas en inglés, *National Human Genome Research Institute Home*) inició un programa cuyo objetivo era la reducción del coste de la secuenciación del genoma humano (en ese momento de 1000$ por genoma) en un plazo de 10 años. Esto permitió el desarrollo y comercialización de las técnicas de secuenciación de nueva generación o NGS (de sus siglas en inglés, *Next Generation Sequencing*) como alternativa a la secuenciación de Sanger y que el coste del análisis de paneles de genes, exomas completos y genomas completos bajara exponencialmente. Las principales ventajas de las técnicas de NGS con respecto a la secuenciación de primera generación eran las siguientes:

1. El proceso de clonado en bacterias quedó sustituido por la preparación de librerías.

2. Se producían miles y hasta millones de reacciones de secuenciación en paralelo.
3. Dejó de ser necesario el paso de electroforesis.

Desde el lanzamiento de la primera plataforma de NGS en 2005, emergieron multitud de tecnologías que, pese a diferir en aspectos técnicos y funcionales, comparten un mismo flujo de trabajo[78,79]. Así, un paso esencial en las tecnologías de NGS es la preparación de la librería. Existe una gran variedad de protocolos específicos para ello dependiendo de la plataforma de secuenciación, pero todos tienen en común la fragmentación del ácido nucleico seguido de la fusión de adaptadores. A continuación, se seleccionan los fragmentos del tamaño deseado en un paso de enriquecimiento, en el cual también se eliminan los adaptadores que han quedado libres. Habitualmente, los fragmentos se amplifican por reacción en cadena de la polimerasa o PCR (de sus siglas en inglés, *Polymerase Chain Reaction*) para seleccionar moléculas que contengan dos adaptadores a cada extremo y generar suficiente material para la secuenciación. En sus extremos, estos adaptadores suelen contener elementos que permiten la inmovilización de los fragmentos en una superficie sólida donde tiene lugar su amplificación clonal. Las secuencias adyacentes a los adaptadores sirven como sitio de unión de los *primers* que iniciarán la secuenciación de los productos amplificados. El resultado es la generación de millones de pequeñas secuencias denominadas lecturas o *reads* que posteriormente se procesarán por métodos bioinformáticos.

En primer lugar, los *reads* se alinean contra un genoma de referencia. A continuación, se compara el genoma diana con el genoma de referencia y se detectan las diferencias entre ambos para cada posición, generando un listado de variantes. Este proceso se denomina *variant calling*. Un aspecto importante de la NGS es la profundidad de lectura. Este concepto se refiere al número de veces que una base es leída por el secuenciador. Por tanto, cuanto mayor sea su valor, con mayor fiabilidad podemos asignar a cada nucleótido una posición determinada. Consecuentemente, el listado final de variantes, será más exacto y contendrá un número menor de artefactos que se pueden confundir con variantes genómicas. Finalmente, las variantes se anotan funcionalmente con toda la información disponible (frecuencia

poblacional, predictores *in silico* de patogenicidad, etc…) para facilitar su interpretación (Figura 6)[81].

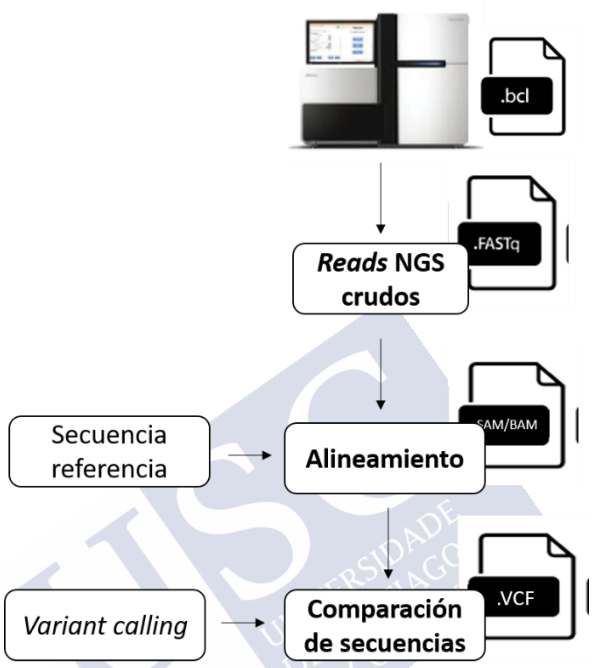

Figura 6. Esquema del procesamiento bioinformático en la NGS. Los datos crudos generados directamente por el secuenciador se encuentran en un formato .bcl. El formato .FASTq, generado a continuación, contiene la secuencia de cada *read* generado por el secuenciador, así como un *score* de calidad para cada una de las bases que conforman dicha secuencia. El alineamiento de los *reads* contra un genoma de referencia da lugar a un archivo en formato.SAM o su formato comprimido .BAM. Tras el proceso de *variant calling*, en el cual se compara la secuencia diana con una de referencia, se obtiene un archivo de variantes cuyo formato más extendido es el .VCF. Este archivo se puede emplear en visores genómicos como el *Integrative Genomics Viewer* (IGV), o se puede anotar funcionalmente.

La aparición de las tecnologías de NGS ha supuesto una revolución de la genómica. A medida que las nuevas tecnologías emergentes incrementan su rendimiento y disminuyen su coste, la secuenciación del genoma se está convirtiendo en algo habitual incluso en laboratorios pequeños. Esto ha posibilitado su implantación en diferentes ámbitos,

como el diagnóstico clínico o la investigación forense, que hasta entonces se habían apoyado en la secuenciación de Sanger, con las limitaciones que ello conlleva[81].

En los últimos años han surgido además nuevas aplicaciones de esta tecnología. Gracias a la NGS, ha sido posible el estudio de los mecanismos de regulación del genoma humano. Así, se pueden estudiar interacciones ADN-proteína mediante técnicas de inmunoprecipitación de la cromatina seguidas de la secuenciación (ChIP-seq)[82].

En la misma línea, el campo de la transcriptómica también se ha beneficiado de la NGS gracias a la tecnología de RNA-seq que permite secuenciar el ARN con mayor sensibilidad que los *microarrays* de expresión. El RNA-seq ha evolucionado permitiendo secuenciar el ARN de una sola célula (scRNA-Seq). Esta aproximación permite, en el caso de los TEA, caracterizar diferentes poblaciones celulares neuronales en diferentes periodos del neurodesarrollo[83].

Otro de los grandes beneficios de la NGS es la oportunidad que ofrece para desarrollar proyectos a gran escala como el Proyecto 1000 Genomas o la actual iniciativa europea *1 + Million Genomes* (https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative). Sin lugar a dudas, la secuenciación de genomas a nivel poblacional está siendo una herramienta esencial para el estudio de la variabilidad genética humana y para entender los mecanismos patológicos de muchas enfermedades[19,84].

A pesar de las ventajas de la NGS, anteriormente citadas, una de sus mayores limitaciones es la generación de *reads* muy cortos. Esto supone un problema a la hora de analizar regiones del genoma que contienen muchas repeticiones y que permanecen aún sin caracterizar. Además, aunque las SNVs e *indels* se pueden detectar con precisión y fiabilidad, la detección de variantes estructurales es más difícil de realizar con este tipo de tecnologías. Por ese motivo, en los últimos años están emergiendo las llamadas tecnologías de tercera generación, que pretenden mejorar esta limitación permitiendo la secuenciación de una molécula única y generando *reads* muy largos. Además, estas tecnologías no precisan de una PCR previa, con la consiguiente disminución de errores que este paso introduce. Sin lugar a dudas, estas tecnologías prometen revolucionar de nuevo el campo de la genómica

y la transcriptómica al permitir el análisis de genomas a una resolución sin precedentes[75].

2.2.2.2.5.2 Estudios de secuenciación llevados a cabo en los TEA

Los estudios de secuenciación en los TEA se han llevado a cabo gracias a la creación de grandes consorcios que, a lo largo de los años, han secuenciado el genoma de miles de familias.

Una de las primeras cohortes de TEA, fue la SSC, creada por la "*Simons Foundation Autism Research Initiative* (SFARI)". Como se ha señalado con anterioridad, su diseño permite fundamentalmente la detección de variantes *de novo*. En esta cohorte se han reunido aproximadamente 2600 familias que se han genotipado y posteriormente secuenciado en diferentes fases del proyecto[85]. Recientemente, SFARI ha fundado un segundo consorcio, "*Simons Foundation Powering Autism Research for Knowledge* (SPARK)" en el cual se pretende reunir 50000 familias con TEA para genotiparlas y secuenciarlas[86].

Uno de los consorcios que más ha contribuido a la generación de conocimiento sobre la arquitectura genética de los TEA es el "*Autism Sequencing Consortium*" (ASC). Este consorcio fue creado en 2010 y en la actualidad, ya cuenta con aproximadamente 35000 individuos cuyo exoma completo se ha secuenciado (12000 afectos)[87].

Por último, destaca el MSSNG, que surge a raíz de una colaboración entre Google y Autism Speaks y cuyo objetivo es crear la mayor base de datos genómicos disponible de este TND. La primera fase del proyecto acaba de finalizar con la secuenciación del genoma completo de 10000 individuos con TEA[88].

En dichas cohortes se han usado fundamentalmente dos aproximaciones: la secuenciación de exoma completo o WES (de sus siglas en inglés *Whole Exome Sequencing)* y la secuenciación de genoma completo o WGS (de sus siglas en inglés, *Whole Genome Sequencing*).

2.2.2.2.5.2.1 Estudios de secuenciación de exoma completo

La secuenciación de exoma completo consiste en secuenciar de forma selectiva las regiones codificantes para proteína (exones) que representan un 1-2% de todo el genoma. Teniendo en cuenta que un 85% de las mutaciones ligadas a una enfermedad se localizan en regiones codificantes y funcionales del genoma, esta estrategia parece un método efectivo y costo-eficiente para detectar mutaciones deletéreas asociadas a los TEA[89].

El objetivo fundamental de los estudios de exoma completo es la detección de SNVs e *indels*. Estos cambios pueden ser sinónimos, si no producen una alteración de la secuencia de aminoácidos que conforman la proteína o no sinónimos cuando ocurre lo contrario. Las mutaciones no sinónimas se clasifican a su vez en:

- Mutación con cambio de sentido o *missense*: resultan en un cambio de aminoácido.
- Mutación sin sentido o *nonsense*: mutación que introduce un codón de parada o *stop* prematuro, lo cual produce un producto proteico truncado.
- Mutación por desplazamiento del marco de lectura o *frameshift*: mutación que ocurre cuando la deleción o adición de nucleótidos cambia el marco de lectura (Figura 7).

Las mutaciones de *splicing*, aunque teóricamente se localizan en regiones intrónicas también se detectan de manera eficiente en los estudios de exomas, debido al efecto que tienen en el procesamiento del ARN maduro.

A su vez, las mutaciones *nonsens*e, *frameshift* y de *splicing*, se denominan conjuntamente mutaciones con pérdida de función o LoF, (de sus siglas en inglés, *Loss-of-Function*) y son consideradas como potencialmente patogénicas.

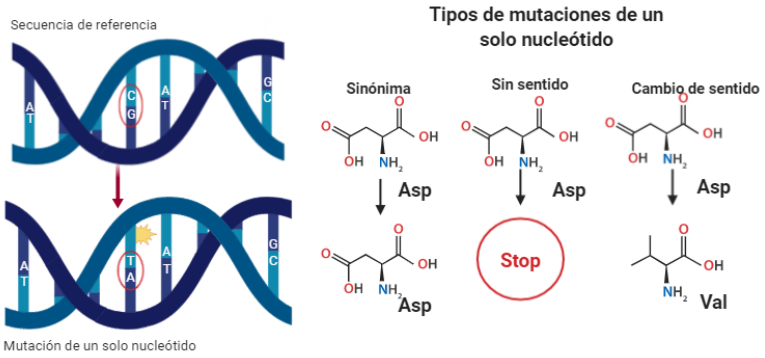

Figura 7. Ejemplos de mutaciones en regiones codificantes.

Los primeros estudios de secuenciación de exoma completo se realizaron en la SSC con el objetivo de detectar mutaciones *de novo* en casos de TEA esporádico[90–93]. En ellos se observó que entre un 3.6 y un 8.8% de los individuos presentaba una mutación causal *de novo*[91,93–95]. El origen de estas variantes era frecuentemente paterno y su aparición se correlacionaba positivamente con la edad paterna[92,95,96]. A partir de estos primeros estudios, se calculó que las mutaciones LoF contribuyen al riesgo en los TEA en más de un 10% de los casos[92–94], mientras que las variantes *missense*, lo hacen en menor grado (<10%). Las variantes sinónimas, en cambio, no han mostrado tendencia alguna de conferir riesgo (Figura 8)[97].

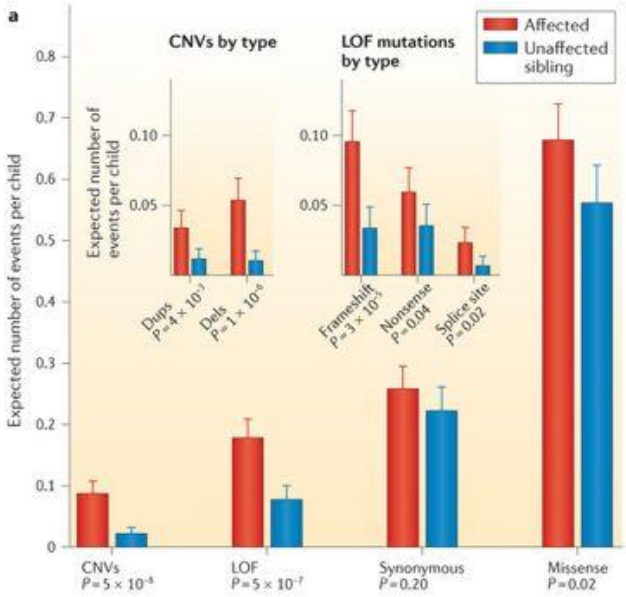

Figura 8. Contribución de las mutaciones *de novo* al riesgo en los TEA en la SSC. Esta figura, extraída de Ronemus *et al.*, 2014, resume la incidencia de CNVs *de novo* (tanto deleciones como duplicaciones) y SNVs *de novo* (mutaciones LoF, mutaciones *missense* y mutaciones sinónimas) en todos los casos de TEA y controles (hermanos no afectos) pertenecientes a la SSC[97]. (Extraída de Ronemus *et al.*, 2014[97], con permiso de Springer Nature).

Las SNVs heredadas, aunque de manera más limitada, también contribuyen al riesgo genético en los TEA. Según un estudio de exoma completos donde se analizaron 933 casos y 869 controles, las mutaciones raras autosómicas recesivas (mutaciones LoF homocigotas o heterocigotas compuestas) podrían explicar un 3% de los casos, mientras que las mutaciones raras ligadas al cromosoma X explicarían un 2% de los casos de TEA en varones[98]. Un estudio reciente, realizado en una cohorte mayor perteneciente al ASC, calcula que este porcentaje es aún superior, y que las mutaciones autosómicas recesivas podrían explicar hasta un 5% del total de los casos[99].

Además, el análisis de variantes raras bialélicas en los TEA ha permitido identificar genes causales de TEA familiar como: *AMT*, *PEX7*, *SYNE1*, *VPS13B*, *PAH* y *POMGNT1 CA2*, *DDHD1*, *NSUN2*, *RARB*, *ROGD1*, *SLC1A1*, *USH2A* y *FEV*[99,100] .

Las variantes raras heredadas *LoF* heterocigotas confieren riesgo tanto en TEA familiar como esporádico. Sin embargo, los estudios genéticos sugieren que por sí solas no son causa suficiente para que se expresen las características fenotípicas propias de los TEA y han de estar presentes otras variantes de riesgo, que interaccionen con ellas, para que esto ocurra. La transmisión de dichas variantes se produce frecuentemente de madres sanas a varones afectos, lo cual apoyaría la hipótesis del "modelo protector femenino" [67,101].

Una de las ventajas que ofrecen los estudios de exoma completo frente al cariotipado molecular, es la posibilidad de detectar genes de riesgo. Las CNVs a menudo afectan a numerosos genes y resulta complejo identificar el gen, o los genes causales, implicados en la expresión del fenotipo. Por el contrario, las SNVs permiten indentificar, con mayor precisión, genes de susceptibilidad en los TEA.

A partir de los primeros estudios de exoma completo se identificaron 10 genes de riesgo portadores de mutaciones *de novo* recurrentes: *CHD8, GRIN2B, SCN1A, DYRK1A, KATNAL2, RIMS1, SCN2A, ANP ARID1B* y *TBR1*[90–95]. Además, el desarrollo de nuevas aproximaciones para mejorar la detección de *indels de novo* a partir de los datos de exoma completo de la SSC permitió incorporar dos nuevos genes al listado previo: *KMT2E* y *RIMS1*[96]. Sin embargo, el número de genes de riesgo identificados era muy reducido, debido a que solo se estudiaron genes con mutaciones LoF de alto impacto, obviando genes con mutaciones de efecto más moderado.

Por ese motivo, los estudios de exoma completo más recientes han incorporado en sus análisis modelos estadísticos capaces de detectar nuevos genes de riesgo, con precisión, a partir de los datos de secuenciación. Uno de los mejores ejemplos, lo constituye el modelo estadístico TADA (*Transmission And De Novo Association*). La novedad del TADA reside en su capacidad de integración, en el mismo análisis, no solo variantes *de novo*, sino también variantes heredadas y variantes detectadas en estudios caso-control, al asumir que todas ellas confieren, en mayor o menor medida, riesgo genético[102]. TADA detecta genes asociados a la enfermedad de estudio teniendo en cuenta su carga mutacional y el riesgo relativo que confiere cada tipo de variante. De

tal manera que, el peso que se le confiere a una variante *de novo* LoF es superior al de una variante *de novo missense*, y el peso de esta última es superior al de una variante LoF heredada[102]. Aunque TADA puede ser usado en cualquier enfermedad, su uso se ha restringido casi en su totalidad a cohortes de TEA, principalmente en la SSC y la cohorte del ASC[66,103,104].

Así, TADA fue empleado en el primer estudio llevado a cabo por el ASC identificando 33 genes asociados a TEA con un FDR (de sus siglas en inglés *False Discovery Rate*) < 0.1 (Tabla 10) y 107 genes con un FDR < 0.3[102].

Mut. LoF	FDR < 0.01	0.01< FDR ≤ 0.05	0.05 < FDR ≤ 0.1
≥2	*ADNP, ANK2, ARID1B, CHD8, CUL3, DYRK1A, GRIN2B, KATNAL2, POGZ, SCN2A, SUV420H1, SYNGAP1, TBR1*	*ASXL3, BCL11A, CACNA2D3, MLL3*	*ASH1L*
1		*CTTNBP2, GABRB3, PTEN, RELN*	*APH1A, CD42BPB, ETFB, NAA15, MYO9B, MYT1L, NR3C2, SETD5, TRIO*
0		*MIB1*	*VIL1*

Tabla 10. Genes de riesgo en los TEA identificados por TADA ordenador por rangos de FDR y número de mutaciones LoF. (Adaptada de Rubeis *et al*., 2014[105])

En el último estudio llevado a cabo por el ASC, que constituye el mayor estudio de exoma completo realizado hasta la fecha (12000 afectos), se empleó un modelo de TADA mejorado en base a dos premisas:

1. No todas las variantes *de novo* LoF tienen un efecto patogénico. Para poder clasificar las variantes LoF en función de su efecto, se usa un *score* llamado pLI (de sus siglas en inglés, *Probability Of Loss-Of-Function Intolerance*) que indica la probabilidad de que un gen sea intolerante a una mutación LoF. Cuanto más intolerante sea el gen, más alto será su pLI (pLI > 0.9)[106].

2. La mayoría de las variantes en regiones codificantes son variantes *missense*. Sin embargo, solo un pequeño porcentaje de estas variantes tiene un efecto funcional, siendo algunas de ellas tan patogénicas como las variantes *de novo* LoF. Para poder indentificar variantes *missense* patogénicas se usa el *score* MPC (de sus siglas en inglés, *Missense Badness, PolyPhen-2, Constraint*)[107].

Así, TADA incluye ahora ambos scores (pLI para variantes LoF y MPC para variantes *missense*) mejorando la sensibilidad y precisión del modelo original. Con esta versión mejorada de TADA, se detectaron 102 genes de riesgo con un FDR < 0.1 (Tabla 11)[104].

Categoría	Genes
Genes enriquecidos con variantes raras en cohortes de TEA o de TND	*GIGYF1, KDM6B, PAX5, TCF7L2, PHF2, CACNA2D3, NR3C2, PTK7, DIP2A*
Genes previamente asociados a TND (Autosómicos dominantes o autosómicos recesivos)	*CHD8, SCN2A, SYNGAP1, ADNP, FOXP1, POGZ, ARID1B, SUV420H1, DYRK1A, SLC6A1, GRIN2B, PTEN, SHANK3, MED13L, CHD2, ANKRD11, ANK2, ASH1L, TLK2, CTNNB1, DEAF1, DSCAM, SETD5, KCNQ3, KDM5B, WAC, SHANK2, NRXN1, TBL1XR1, DNMT3A, MYT1L, BCL11A, RAI1, DYNC1H1, KMT2C, GABRB3, SIN3A, MBD5, STXBP1, TBR1, PPP2R5D, PHF21A, SKI, ASXL3, SPAST, SMARCC2, TRIP12, CREBBP, TCF4, CACNA1E, GNAI1, TCF20, FOXP2, NSD1, GFAP, IRF2BPL, SCN1A, TRAF7, KMT2E, NACC1, GABRB2, KCNMA1*
Genes nuevos	*SRPR, RORB, DPYSL2, AP2S1, MKX, MAP1A, CELF4, PHF12, TM9SF4, PRR12, LDB1, EIF3G, KIAA0232, VEZF1, ZMYND8, SATB1, RFX3, PPP5C, TRIM23, ELAVL3, GRIA2, LRRC4C, NUP155, PPP1R9B, HDLBP, TAOK1, UBR1, TEK, CORO1A, HECTD4, NCOA1*

Tabla 11. Genes de riesgo en los TEA con FDR < 0.1 detectados en el mayor estudio de secuenciación de exoma completo llevado a cabo en los TEA (Adaptada de Satterstrom *et al.*, 2020[104]).

2.2.2.2.5.2.1.1 Mutaciones postcigóticas y mosaicismos detectados en estudios de secuenciación de exoma completo.

Las mutaciones *de novo* puede ser de dos tipos: germinales, si la mutación aparece de manera espontánea en células germinales (óvulo o espermatozoide), o postcigóticas (PZMs, de sus siglas en inglés, *postzygotic mutations*) si surgen espontáneamente durante las primeras divisiones mitóticas de un cigoto.

En el caso de las mutaciones germinales el embrión resultante portará la mutación en cada una de sus células. Por el contrario, en el caso de las mutaciones PZMs, el individuo resultante es un mosaico en el cual un número variable de sus células porta la mutación (Figura 9)[108].

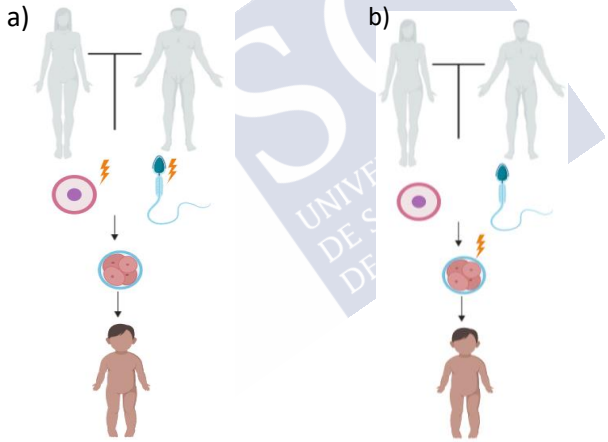

Figura 9. Mutaciones *de novo*. a) Mutaciones germinales. b) Mutaciones postcigóticas.

Las PZMs se han visto implicadas en diferentes TND, como la epilepsia, malformaciones corticales o rasopatías[109–111]. En estos trastornos se ha observado una relación entre la gravedad de los síntomas que padece el individuo afecto y el momento del desarrollo en el cual aparece la mutación, así como los tipos celulares afectos. Por ejemplo, las mutaciones en el gen *MECP2* que causan el síndrome de Rett, suelen ser letales en hombres y dominantes en mujeres, pero se

han descrito casos donde las mutaciones en mosaico en varones son compatibles con la vida[112].

A pesar de que las PZMs han demostrado jugar un papel importante en la etiología de algunos TND, su detección constituye todo un desafío, debido a que son generalmente específicas de tejido, y en concreto, las muestras de tejido cerebral son muy difíciles de obtener. Las técnicas de NGS, sin embargo, han demostrado ser muy eficaces en la detección de mutaciones mosaico a partir de sangre periférica al poseer una mayor profundidad de lectura. La frecuencia del alelo alternativo o AAF (de sus siglas en inglés *Alternate Allele Frequency*) se puede calcular si se obtiene un número suficiente de *reads* que contengan el alelo de referencia y el alelo mutado. En condiciones normales, se esperaría un valor de AAF cercano al 50%, pero en el caso de una mutación PZM, este valor se reduce (AAF < 40%), al detectarse un número menor de *reads* con el alelo mutado. Así, la secuenciación de exoma completo a una profundidad de lectura superior a 200X (cada base se lee 200 veces), es lo suficientemente sensible para detectar PZMs cuyo valor de AAF es solo de un 15%, lo cual significa que la mutación está presente en solo un 25-30% de las células[113,114].

Los estudios de exoma completo en los TEA han obviado frecuentemente el análisis de PZMs debido a la falta de *pipelines* adecuados para su detección. Un reanálisis de la SSC al completo, diseñado específicamente para detectar PZMs, determinó que un 22% de las mutaciones *de novo* son PZMs y que su detección no se estaba realizando[115].

En la misma línea, un trabajo del ASC en el cual se reanalizaron los datos de exoma completo del consorcio aplicando un *pipeline* adecuado para la detección de PZMs, determinó que un 80% de estas variantes habían sido obviadas en estudios previos (Tabla 12)[116] .

Estudio	Krupp *et al.*, 2017	Lim *et al.*, 2017
Número de familias analizadas	2264	5947
% de PZMs detectadas aplicando nuevos *pipelines* informáticos	22 %	9.7 %
% de PZMs sin publicar en estudios previos	70.64 %	83.3 %

Tabla 12. Resultados obtenidos en los principales estudios en cohortes de TEA donde se analizaron PZMs. (Adaptada de de Alonso-Gonzalez *et al.*, 2018[103]).

A partir de estos innovadores estudios, se ha calculado que las PZMs contribuyen aproximadamente en un 4% al riesgo genético de los TEA[115,116]. Así, las PZMs no solo han sido identificadas en genes de riesgo en los TEA, sino que, además, su detección ha permitido identificar nuevos genes candidatos como *KLF16* y *MSANTD* (Tabla 13).

Estudio	Genes identificados
Lim *et al.*, 2017	*KLF16, MSANTD2, POLA2, SMARCA4, AZGP1, CNGB3, HNRNPU, SCN2A, EPPK1, CARD11*
Krupp *et al.*, 2017	*CHD2, CTNBB1, SCN2A, SYNGAP1, ACTL6B, BAZ2B, COL5A3, SSRP1, UNC79*
Freed and Pevsner, 2016	*KMT2C, NCKAP1, MYH10*

Tabla 13. Genes portadores de mutaciones PZMs en los TEA. (Adaptada de Alonso-Gonzalez *et al.*, 2018[103]).

Un análisis en detalle de PZMs no sinónimas, ha revelado que estas se localizan con mayor frecuencia en genes expresados en cerebro y en exones poco tolerantes a variantes LoF. Además, el análisis de expresión espacio temporal en diferentes periodos del neurodesarrollo y en diferentes áreas cerebrales de los genes portadores de PZMs, señaló la amígdala como una nueva área del cerebro de interés en el estudio de los TEA. Esta región no había sido identificada previamente en estudios de secuenciación de exoma completo[116].

En conclusión, estos estudios preliminares señalan el importante papel que tienen las PZMs en la patogénesis de los TEA. Quedan aún, sin embargo, muchas cuestiones por resolver en relación a estas

variantes, como, por ejemplo, los procesos biológicos en los cuales están implicadas, o cuál es su impacto en la expresión del fenotipo.

2.2.2.2.5.2.2 Estudios de secuenciación de genoma completo

Los estudios de secuenciación de genoma completo son aquellos que secuencian la totalidad del genoma, incluyendo las regiones no codificantes que no están cubiertas en los estudios de exoma completo. Gracias a esta aproximación se puede detectar todo el espectro de la variación genética conocida, incluyendo SNPs, SNVs, *indels* y variantes estructurales incluyendo; translocaciones, inversiones y CNVs[57,117].

Tradicionalmente, se pensaba que la variación genética en regiones no codificantes no contribuía al riesgo genético de las enfermedades debido a que no codifica para proteínas funcionales. Sin embargo, el proyecto ENCODE (*Encyclopedia of DNA Elements*), financiado por el NHGRI, y con el propósito de catalogar y descifrar todos los elementos funcionales del genoma, demostró que el 80.4% del genoma participa en alguna actividad reguladora que implica al RNA y/o a la cromatina. Así pues, gran parte de la variación genética que se encuentra en regiones no codificantes participa en la regulación genética de diferentes procesos relevantes en el neurodesarrollo, como la neurogénesis o la diferenciación celular (Figura 10)[118]. Además, también se ha comprobado que las regiones reguladoras están sometidas al efecto de la selección natural negativa lo cual sugiere su potencial efecto deletéreo[119].

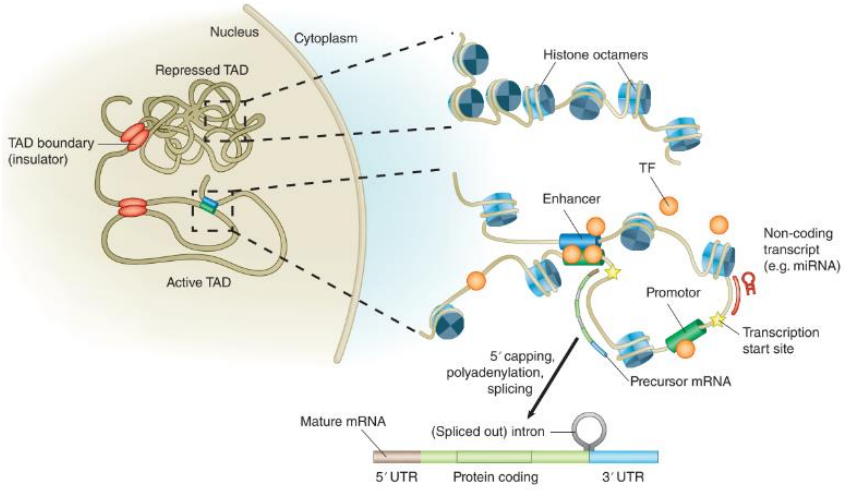

Figura 10. Representación de elementos no codificantes con función reguladora. En la figura se pueden ver ampliados algunos detalles importantes del ADN compacto (eucromatina) en el núcleo de una célula eucariota. Los TAD (de sus siglas en inglés, *Topologically Associating Domain*) son regiones del genoma, donde las secuencias de ADN interactúan entre sí con más frecuencia. Dichas regiones se separan unas de otras gracias gracias a los *insulators* que actúan de barrera. Los TAD pueden estar inactivos. En ese caso, la cromatina de la región se encontrará en un estado muy compacto gracias a las proteínas de histona. En el caso de estar activos, se favorece la interacción en la región de elementos reguladores, y los genes son activamente transcritos. Los promotores son regiones que se localizan cerca del gen (± 20 pb) y juegan un papel crítico a la hora de iniciar la transcripción del gen. Los *enhancers* controlan también la expresión del gen, aunque se encuentran muy alejados del mismo (> 100 pb). Los factores de transcripción o TF (de sus siglas en inglés, *transcription factor)* se unen a ambos reconociendo secuencias específicas y permiten reclutar la maquinaria necesaria para facilitar la transcripción de un gen. Cuando se inicia la transcripción de un gen en regiones codificantes, se genera inicialmente una molécula de ARN precursor que debe ser procesado para dar lugar a ARN maduro que se traducirá a proteína. Sin embargo, hay tipos de ARN que no con codificantes y que tienen también una función reguladora. (Extraída de A. Takata, 2019[120], con permiso de John Wiley and sons).

La secuenciación de genoma completo es especialmente útil en los casos de TEA en los cuales no se ha encontrado una CNV o una SNV *de novo* LoF causal. Estos casos, suponen, por tanto, un reto a la hora de esclarecer la causa genética subyacente al TND que padecen.

Por ese motivo, en muchos estudios de secuenciación de genoma completo las cohortes estudiadas están enriquecidas con este tipo de casos, lo cual incrementa la posibilidad de encontrar variantes causales en regiones no codificantes[121–125]. Aun así, se estima que solo un 1-2.8% de los casos con un exoma completo o *array* negativo portan una mutación patogénica en regiones reguladoras, por lo que muy pocos se pueden beneficiar de un diagnóstico mediante el estudio de su genoma completo[121].

Por otro lado, los estudios de genoma completo tienen la capacidad de analizar la totalidad de la arquitectura genética en un solo experimento[124,126]. Esta posibilidad, es importante de acuerdo a un modelo poligénico recientemente propuesto en los TEA, en el cual, el riesgo genético para este TND surge por el efecto acumulativo de 2 o más variantes *de novo* localizadas tanto en regiones codificantes como no codificantes. Este modelo rechazaría, por tanto, el modelo monogénico, más estudiado por estudios de exoma completo, en el cual toda la causalidad del trastorno se atribuye a una variante "patogénica"[124]. Además, se ha visto que la secuenciación de genoma completo aporta una mayor cobertura media de las regiones codificantes[88] y esta característica permite identificar mutaciones puntuales que han pasado desapercibidas en un 2.5% de los estudios de exoma completo[124].

Sin embargo, existen diversos problemas a la hora de analizar estas variantes, como son la cantidad de datos generados en estos estudios (se analizan aproximadamente 3 billones de nucleótidos por genoma) y la dificultad existente para anotarlas funcionalmente al no poder aplicar el código de tripletes, tal y como se hace en las regiones codificantes. Para la anotación funcional de estas variantes, se recurre a múltiples criterios, como pueden ser el grado de conservación genética, el tipo de variante (SNV, *indel*, o variante estructural) o su localización según la definición de gen de GENCODE (promotores, UTRs, regiones de *splicing*, etc), pudiendo surgir un gran número de categorías posibles si se combinan todos ellos.

Uno de los primeros trabajos en los que se estudió el papel de los elementos reguladores en los TND fue el llevado a cabo por Short, *et al*. En él, se hizo una secuenciación dirigida de regiones no codificantes

altamente conservadas y en regiones promotoras y potenciadoras (del inglés, *enhancer*) para detectar SNVs *de novo* en 8000 individuos pertenecientes al proyecto "Deciphering Developmental Disorders (DDD)". Así, se encontró un enriquecimiento de SNVs *de novo* en casos con respecto a controles, en regiones reguladoras que están activas durante el desarrollo fetal[121]. Aunque en este trabajo no se hizo una secuenciación de genoma completo, ni la cohorte de estudio incluyó exclusivamente individuos con TEA, sus hallazgos justificaron la necesidad de estudiar en detalle estas regiones en estudios posteriores de genoma completo.

Así, en el trabajo de Turner *et al*., se analizó el genoma completo de 516 familias con TEA de la SSC focalizándose solo en aquellas regiones no codificantes en las que una mutación puede tener un mayor impacto funcional (promotores, UTRs y elementos reguladores activos en el cerebro fetal). En este estudio, se atribuyó un papel importante a las variantes reguladoras localizadas en promotores activos del cerebro fetal y *enhancers* de células embriónicas en el riesgo de los TEA[124].

De manera paralela, el genoma de las mismas familias fue analizado siguiendo una estrategia diferente en el trabajo de Werling *et al*. En este caso, el análisis no se restringió a regiones concretas y se consideraron 5 categorías de anotación definidas por el grado de conservación entre especies, el tipo de variante, las definiciones de gen según GENCODE, listas de genes asociados a TEA previamente y anotaciones funcionales. En este estudio resultó llamativo que no se hallara ninguna asociación entre variantes en regiones no codificantes y el riesgo en los TEA. Así pues, a partir de estos datos se concluyó que su contribución es muy modesta en comparación con variantes *de novo* en regiones codificantes. Además, se estimó que esta estrategia requería cohortes de mayor tamaño para poder obtener resultados significativos y se hizo evidente la necesidad de desarrollar nuevas estrategias de análisis que consideren el elevado número efectivo de test que se realiza en la anotación funcional de variantes no codificantes[123].

Una solución al problema de la anotación funcional se propuso en el trabajo de Zhou *et al*. En él, se desarrolló una estrategia basada en *machine learning* que permitió predecir el efecto específico en la regulación de las mutaciones *de novo* detectadas. Aplicando esta

aproximación en una cohorte de 1790 familias de la SSC, se confirmó el papel de las variantes no codificantes en los TEA. Así, se encontró que, en casos, había un enriquecimiento de variantes con alto impacto funcional en mecanismos de regulación transcripcional y postranscripcional. Estas variantes, afectaban con mayor frecuencia a genes expresados en cerebro que convergían también en rutas relacionadas con el desarrollo neuronal (transmisión sináptica y regulación de la cromatina) señaladas previamente en el estudio de variantes codificantes LoF[105,122].

Además del estudio de las variantes *de novo* en regiones reguladoras, los estudios de secuenciación de genoma completo en los TEA han destacado por su capacidad de analizar con detalle todo tipo de variantes estructurales. Se ha estimado que aproximadamente un 95.6 % de las CNVs de menor tamaño detectadas por estudios de genoma completo no pueden ser detectadas mediante *microarrays* de alta resolución[88]. Además, las variantes estructurales en regiones no codificantes afectan con mayor frecuencia a la regulación génica que las SNVs, debido su capacidad de interrumpir y reorganizar el genoma. Así pues, se han encontrado variantes estructurales recurrentes localizadas en regiones promotoras de los genes *DLG2* y *NR3C2* en individuos con *TEA* con una posible implicación directa en la expresión de su fenotipo[126]. Además, también se ha detectado que la transmisión de padres a hijos de variantes estructurales en regiones reguladoras en *cis* es más frecuente en casos que controles, estimando así, una contribución de las mismas al riesgo genético de un 0.77%[125].

Los estudios de secuenciación de genoma completo llevados a cabo en los TEA, son aún escasos en comparación con el elevado número de estudios de secuenciación de exoma desarrollados en la última década. Sin embargo, la disminución del coste de la secuenciación y el desarrollo de herramientas bioinformáticas que permitan trabajar con la inmensa cantidad de datos que estos estudios generan, hará posible, en un futuro próximo, realizar proyectos a gran escala, y secuenciar el genoma completo de miles de individuos con TND. Sin lugar a dudas, estos trabajos serán clave en el estudio de la arquitectura genética de los TEA, gracias a su capacidad de estudiar toda la variabilidad genética en un solo experimento.

2.3 NEUROBIOLOGÍA DE LOS TEA

Los TEA son TND extremadamente heterogéneos en términos fenotípicos, caracterizados también por una enorme variabilidad genética, tal y como se ha descrito en los apartados anteriores. Se calcula que alrededor de 1000 genes pueden estar involucrados en su etiología, lo cual significa que, individualmente, cada uno de ellos explica menos del 1% de todos los casos[105]. Teniendo en cuenta este hecho, resulta imposible llevar a cabo todos los estudios funcionales que serían necesarios para conocer el efecto que causa una mutación en cada uno de estos genes de riesgo.

Para entender cómo los genes contribuyen al fenotipo resultante de los TEA, la aproximación más coherente consiste en buscar las rutas biológicas en las cuales, converge toda la variabilidad genética que confiere riesgo en este TND. Para ello, son necesarias herramientas que operen usando la biología de sistemas y que permitan estudiar los mecanismos que subyacen a estos trastornos en diferentes niveles: molecular, celular y anatómico (Figura 11).

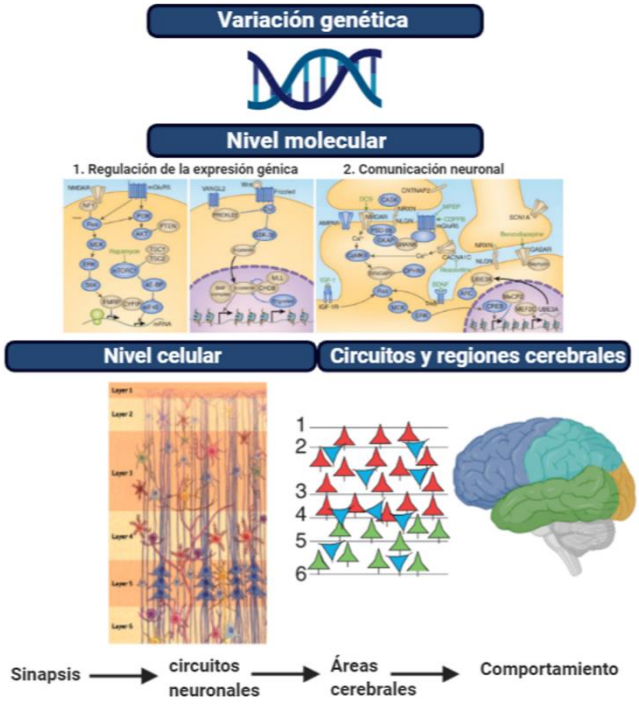

Figura 11. Representación de los diferentes niveles de estudio de la neurobiología de los TEA. Los estudios genéticos identifican variantes causales en regiones codificantes de genes de riesgo en los TEA. El impacto que tienen estas variantes, se mide en la funcionalidad de su producto proteico (nivel molecular). A nivel celular, los cambios a nivel molecular pueden resultar en cambios a la hora de establecer los correctos circuitos neuronales, lo que puede conllevar cambios en el siguiente nivel (nivel anatómico).

El éxito de los análisis que integran información biológica a nivel genético, molecular, celular o anatómico se debe principalmente a la creación de grandes consorcios que han contribuido a generar conocimiento en diferentes áreas como la transcriptómica, la epigenómica y la proteómica. Entre ellos destaca el proyecto *BrainSpan Atlas*, que contiene información referente al patrón de expresión de los genes en 16 regiones corticales y subcorticales a lo largo de las diferentes etapas del neurodesarrollo (prenatal y postnatal)[127,128]. También es importante el proyecto GTEx (de sus siglas en inglés, *The*

Genotype-Tissue Expression Project) que ha caracterizado el patrón de expresión génico en 54 tejidos diferentes, además de caracterizar *loci* genómicos que intervienen en la regulación de la expresión denominados eQTLs (de sus siglas en inglés, *Expression Quantitative Trait Loci*)[129].

Gracias a estos repositorios de información, se han hecho enormes avances en neurociencia y se ha podido crear, por primera vez, un nexo entre variación genética, comportamiento y cognición.

2.3.1 Herramientas utilizadas en el estudio de la neurobiología de los TEA: redes génicas, análisis de expresión diferencial y análisis de enriquecimiento

Una de las aproximaciones más utilizadas para entender los mecanismos biológicos que subyacen a los TEA son los análisis de redes génicas. Este tipo de análisis permite estudiar e integrar múltiples niveles de organización (transcriptómica, proteómica, genómica o rutas biológicas) para identificar los procesos biológicos involucrados en la fisiopatología de una enfermedad.

En una red génica se distinguen dos elementos fundamentales: los nodos y las aristas. Los nodos, son entidades moleculares, generalmente genes o sus productos proteicos, que se conectan entre sí por aristas en base a las relaciones que mantienen entre sí. Para definir estas relaciones se requiere información externa obtenida experimentalmente o computacionalmente que evalúa las interacciones entre nodos (por ejemplo, en base a su patrón de expresión o en base a interacciones físicas).

Una red génica se puede subdividir en módulos o clústeres que son subgrupos de genes altamente conectados entre sí. A su vez, dentro de cada módulo podemos identificar *hubs* que son moléculas que están muy conectadas entre sí, lo que indica que tienen un papel relevante en un determinado proceso biológico. Para anotar funcionalmente e identificar procesos biológicos sobrerrepresentados en cada módulo se emplea la ontología génica usando términos GO *(Gene Ontology)* y rutas KEEG *(The Kyoto Encyclopaedia of Genes and Genome Elements)* (Figura 12)[130].

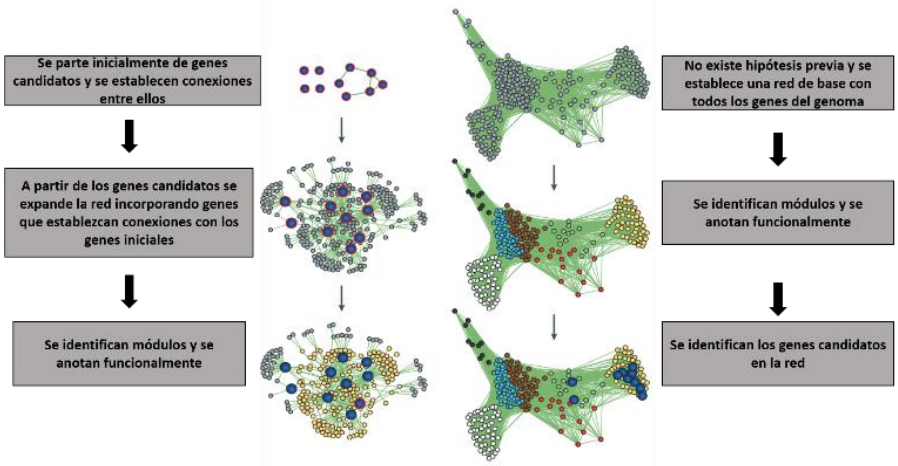

Figura 12. Esquema de la construcción de una red génica usando dos aproximaciones diferentes. A la izquierda, la construcción de la red se realiza con una hipótesis previa a partir de unos genes de interés. A la derecha, la red se construye sin hipótesis previa considerando todos los genes con información funcional disponible. En ambos casos los módulos se anotan usando información externa. (Adaptada de Parikshak *et al.*, 2015[130], con permiso de Springer Nature).

En el estudio de los TND se han usado frecuentemente dos fuentes de información para crear redes génicas:

1. Datos tejido-específico generados experimentalmente mediante aproximaciones de transcriptómica.
2. Bases de datos biológicas con información sobre ontologías génicas o de interacciones proteína-proteína o PPI (de sus siglas en inglés, *Protein-Protein Interaction*).

En el primer caso, las conexiones entre los nodos que conforman la red se establecen de acuerdo a datos de transcriptoma procedentes de tejido o de célula única. Estos datos pueden proceder directamente de experimentos RNA-seq o scRNA-seq, *microarrays* de expresión o bien usar bases de datos externas como *BrainSpan*, que contiene datos de expresión obtenidos experimentalmente[131].

Para generar redes de coexpresión se pueden usar múltiples aproximaciones estadísticas, pero todas ellas tienen en común dos pasos críticos: el cálculo de una medida de coexpresión y el establecimiento de un umbral de significación estadística[132]. Un método muy usado es

WGCNA (*Weighted Gene Co-expression Red Analysis*). WGCNA construye una red de coexpresión en la que se identifican módulos de genes con un patrón de expresión muy similar. En cada módulo se identifican los denominados *eigengenes* mediante herramientas de reducción de variables como el análisis de componentes principales o PCA (de sus siglas en inglés, *principal component analysis*). Los *eigengenes* son genes cuyo patrón de expresión es representativo del módulo y por tanto permiten entender el significado biológico del mismo[131].

Por otro lado, las redes construidas a partir de bases de datos biológicas tienen la desventaja de que a menudo dichos datos no son tejido-específicos, y, por tanto, las interacciones entre genes que se establecen a partir de ellos pueden no reflejar exactamente las condiciones existentes en el tejido cerebral.

El ejemplo más común lo constituyen las redes de interacción proteína-proteína o PPI (de sus siglas en inglés, *Protein-Protein Interaction*). Los datos de PPI proporcionan información de cómo las proteínas interactúan entre sí en diferentes procesos celulares[133]. Dichos datos se obtienen usando diferentes aproximaciones (bioquímica, química cuántica, dinámica molecular…) y esta información se deposita en diferentes bases de datos como BioGRID, STRING, MINT, KEEG, DIP, HPRD, o IntACt[134]. La ventaja fundamental de este tipo de redes es la posibilidad de mapear genes de interés en ellas e identificar nuevas asociaciones con otros genes que no habían sido señalados previamente y así ampliar la lista de genes candidatos asociados a una enfermedad.

En los últimos años, se han desarrollado herramientas que permiten la construcción de redes génicas a partir de varias fuentes de información existentes[130]. NETBAG y MAGI, son ejemplos de ello y ambas se han usado con éxito en TND como los TEA y enfermedades psiquiátricas como la esquizofrenia[69,135–137].

En el caso de NETBAG, la red se construye en base a la probabilidad de que dos genes participen en un mismo fenotipo, como pueden ser los TEA y DI. Para su construcción se usa la ontología génica (GO y KEEG) o información de PPI[137]. MAGI, por otro lado,

integra de manera simultánea información de PPI y datos de expresión con datos genéticos[136].

Los análisis de genes diferencialmente expresados o DEG (de sus siglas en inglés *Differentially Expressed Genes*) comparan la expresión de miles de genes entre dos grupos (casos y controles) para identificar genes asociados a una enfermedad analizando cada gen individualmente sin tener en cuenta las relaciones existentes entre ellos. Esto da lugar a listas muy largas de genes diferencialmente expresados cuya interpretación es muy compleja. Además, hay que sumarle la dificultad de obtener tejido cerebral humano *post-mortem* para crear tamaños de muestra lo suficientemente grandes como para diferenciar la variabilidad interindividual de las diferencias de expresión relativas a la enfermedad[130].

Los análisis de enriquecimiento de genes o GSEA (de sus siglas en inglés, *Gene Set Enrichment Analysis*) comprueban la sobrerepresentación estadística de un grupo de genes en un *set* de genes definido *a priori*. En el análisis se distinguen 3 pasos fundamentales: el cálculo de un *score* de enriquecimiento, la estimación de la significación estadística del *score* y la corrección por múltiples test si se analizan muchos *sets* de genes simultáneamente. Los GSEA utilizan listados de genes que han sido categorizados previamente, generalmente en base a su implicación en rutas biológicas o bien a su identificación como genes de riesgo en una patología. Así, en los estudios de TEA, los genes de riesgo identificados se comparan con listados de genes clasificados según alguno de esos criterios y se comprueba si existe una sobrerrepresentación de estos genes en alguna ruta biológica, para así determinar su posible asociación con el fenotipo[130].

2.3.2 Resultados de los principales estudios

Los estudios de TEA donde se han realizado análisis de redes génicas han permitido identificar las principales rutas biológicas que participan en la patogénesis de los TEA

El trabajo de O´Roak *et al.*, fue pionero a la hora demostrar que los genes de riesgo en TEA participan en procesos biológicos comunes. En él se demostró, que un 39% de los genes de riesgo identificados en

individuos con TEA, interactuaban entre sí en una red de tipo PPI. Dicha red, estaba enriquecida para funciones biológicas relacionadas con el remodelado de la cromatina y la vía de señalización *Wnt/β-catenina*[95]. En este primer trabajo, solo se incluyeron genes con mutaciones *de novo* LoF, pero en estudios posteriores se tuvieron en cuenta otro tipo de variantes que también confieren riesgo en los TND. Así, en el trabajo de Ruzzo *et al.*, se comprobó que tanto la variación rara *de novo* como la heredada converge en una red de proteínas interconectadas entre sí. Esta red se enriquece en componentes de la familia SWI/SNF, un complejo remodelador de la cromatina durante la neurogénesis cortical[126].

El análisis de CNVs identificadas en la SSC y la cohorte del AGP, permitió comprobar que también los genes implicados por estas variantes convergen en procesos biológicos relacionados con el neurodesarrollo. Así, en el trabajo de Gilman *et al.*, se observó un enriquecimiento para funciones relacionadas con el desarrollo de las sinapsis, la motilidad neuronal y la guía de los axones[137] mientras que en el trabajo de Pinto *et al.*, se detectó un enriquecimiento para funciones ya descritas anteriormente como la transcripción, el remodelado de la cromatina y el desarrollo de las sinapsis, además de otras nuevas como la vía de señalización MAPK[69].

El estudio de Rubeis *et al.*, llevado a cabo por el ASC, reportó resultados similares identificando 4 módulos enriquecidos para funciones biológicas relacionadas con la formación de las sinapsis, la transcripción y el remodelado de la cromatina (Figura 13)[105].

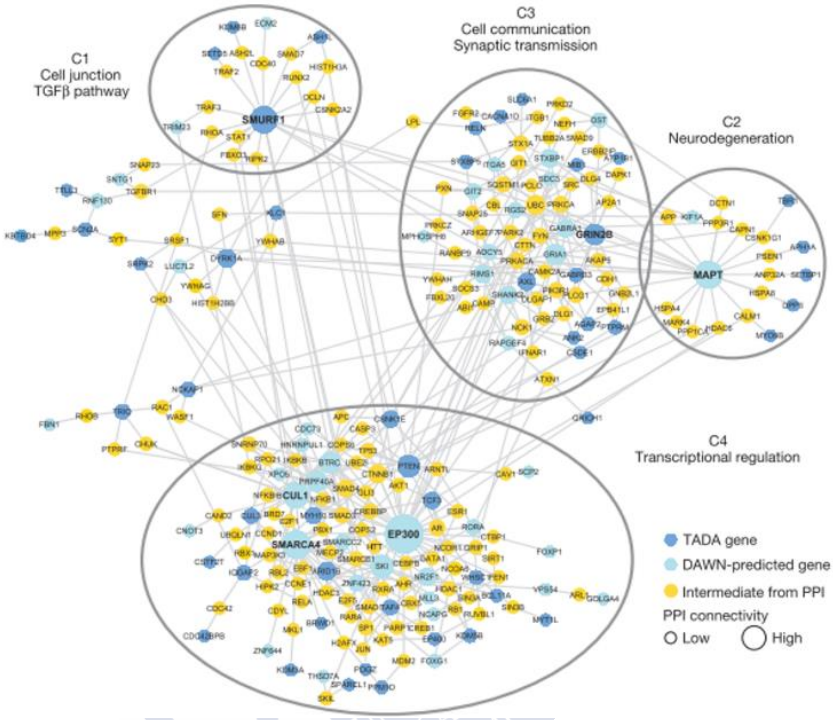

Figura 13. Análisis de redes génicas (PPI) realizado a partir de los genes de riesgo para los TEA identificados por los algoritmos TADA y DAWN. (Extraída de Rubeis *et al.*, 2014[105], con permiso de Springer Nature).

Los resultados de estos estudios señalan la existencia de dos grupos funcionales de genes: genes involucrados en la regulación de la expresión génica, incluyendo la regulación de la transcripción y el remodelado de la cromatina, y genes involucrados en la comunicación neuronal. Esta división fue corroborada en el trabajo llevado a cabo por Satterstrom *et al.*, en el cual se realizó un análisis de enriquecimiento basándose en ontologías para un listado de 102 genes obtenido con TADA (FDR < 0.1). En efecto, la mayoría de los genes fueron clasificados en alguna de las dos categorías principales (regulación génica y comunicación neuronal). Sin embargo también se identificaron otros módulos relacionados con la organización del citoesqueleto y con cascadas de señalización y/o ubiquitinación (Tabla 14)[104].

Categorías	*Genes*
Regulación de la expresión	*CHD8, ADNP, FOXP1, POGZ, ARID1B, SUV420H1, MED13L, CHD2, ANKRD11, ASH1L, TLK2, DNMT3A, DEAF1, CTNNB1, KDM6B, SETD5, KDM5B, WAC, TBL1XR1, MYT1L, BCL11A, RORB, RAI1, KMT2C, PAX5, MKX, SIN3A, MBD5, CELF4, PHF12, TBR1, PPP2R5D, PHF21A, SKI, ASXL3, SMARCC2, TRIP12 CREBBP, TCF4, TCF20, FOXP2, NSD1, TCF7L2, LDB1, EIF3G, PHF2, VEZF1, IRF2BPL, ZMYND8, SATB1, RFX3, TRAF7, ELAVL3, KMT2E, NR3C2, NACC1, HDLBP, NCOA1*
Comunicación neuronal	*SCN2A, SYNGAP1, SLC6A1, GRIN2B, PTEN, SHANK3, ANK2, DSCAM, KCNQ3, SHANK2, NRXN1, AP2S1, GABRB3, STXBP1, PRR12, CACNA1E, SCN1A, GRIA2, LRRC4C, CACNA2D3, PPP1R9B, GABRB2, KCNMA1, DIP2A*
Organización del citoesqueleto	*DYRK1A, DYNC1H1, DPYSL2, MAP1A, SPAST, GFAP, PTK7, TAOK1, CORO1A*
Otros	*GIGYF1, SRPR, TM9SF4, GNAI1, KIAA0232, PPP5C, TRIM23, NUP155, UBR1, TEK, HECTD4*

Tabla 14. Clasificación funcional de los genes identificados por TADA (FDR < 01). (Adaptada de Satterstrom *et al*., 2020[104]).

Una vez identificadas las principales rutas biológicas en las cuales intervienen los genes de riesgo de los TEA, la siguiente cuestión que se debe abordar es averiguar cuándo, dónde y en qué tipo celular tienen lugar estos procesos biológicos. Para ello han sido cruciales los estudios donde se han llevado a cabo análisis de redes de coexpresión ya sea con datos derivados directamente de muestras de cerebro *post-mortem* o bien usando datos de *BrainSpan* (Tabla 15).

De acuerdo a la primera estrategia, en el estudio de Voineagu *et al*., se buscaron genes diferencialmente expresados entre casos y controles. Así, se hallaron dos módulos de genes diferencialmente expresados en el córtex: en el primer módulo se identificaron genes con menor expresión en los TEA relacionados con la sinapsis, el transporte de vesículas y las proyecciones neuronales; en el segundo módulo se identificaron genes sobreexpresados en los TEA que se relacionan con procesos inflamatorios[138]. La misma estrategia fue llevada a cabo en una cohorte mayor de casos de TEA obteniéndose resultados muy similares[139].

Así pues, en el estudio con mayor número de muestras de cerebro analizadas se identificaron 24 módulos de coexpresión. De ellos, solo 6 estaban asociados a los TEA. Tres de ellos estaban enriquecidos en genes que participan en funciones sinápticas, mientras que los tres restantes incluyeron genes que participan en funciones relacionadas con procesos inflamatorios[140].

A partir de estos datos se ha propuesto que una actividad neuronal deficiente pueda estar implicada en la etiología de los TEA mientras que el aumento en la respuesta inflamatoria sea probablemente un proceso biológico secundario[138,139].

Paralelamente, la construcción de redes de coexpresión a partir de datos disponibles públicamente, ha revelado cómo los genes de riesgo para los TEA se expresan diferencialmente en etapas muy tempranas del neurodesarrollo en el córtex prefrontal, temporal y cerebeloso (Figura 14)[141–143]. No obstante, otros estudios también señalan otras áreas cerebrales como el cerebelo, el estriado, la amígdala o el tálamo. Así pues, el hecho de que la actividad de los genes de riesgo no se restrinja al córtex explicaría el amplio espectro de síntomas que caracteriza a los TEA[142,143].

A nivel celular, la expresión de genes de riesgo en los TEA se ha observado preferentemente en neuronas glutamatérgicas del córtex. Sin embargo, dos estudios difirieron localizándolas en su estructura laminar. En el estudio de Willsey *et al.*, se identificaron en las capas 5 y 6[141], mientras que en el de Parikshak *et al.*, se observaron en capas superiores[144].

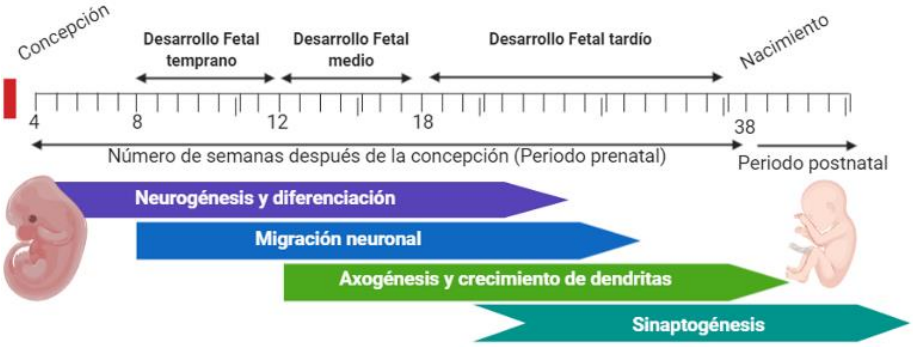

Figura 14. Representación esquemática de los principales eventos que tienen lugar durante el neurodesarrollo. Se detalla el inicio y fin aproximado de los diferentes procesos que participan en el correcto desarrollo neurológico: neurogénesis y diferenciación neuronal, migración neuronal, axogénesis y sinaptogénesis

El último estudio de exoma completo llevado a cabo por el ASC, integra resultados de diversos estudios y alcanza conclusiones muy similares, destacando la expresión de genes de riesgo en los TEA en el córtex durante etapas tempranas del neurodesarrollo. Sin embargo, al hacer una distinción entre los dos principales grupos de genes (genes reguladores de la expresión génica y genes relacionados con la comunicación neuronal) se distinguieron dos subperiodos de susceptibilidad en los TEA: uno durante el desarrollo fetal temprano, en el que tiene lugar la máxima expresión de los genes reguladores, y otro durante el desarrollo fetal tardío en el que se expresan los genes relacionados con la comunicación neuronal (Figura 14). Una tendencia similar se había observado previamente en otro trabajo, en el cual se distinguieron dos módulos de genes: uno expresado entre las semanas 8 y 14 tras la concepción que incluía genes relacionados con la vía de señalización *Wnt/β-catenina* y otro expresado en periodos postnatales que incluyó genes relacionados con la función sináptica[136].

A nivel celular, en el trabajo del ASC, la expresión de los genes de riesgo se localizó en tipos neuronales que incluyeron neuronas excitatorias e inhibitorias, tanto maduras como inmaduras. Este hecho, apoya la teoría de que una función deficiente de los genes relacionados

con la comunicación neuronal pueda resultar en una alteración del balance excitatorio /inhibitorio, tal y como ha sido defendido por otros autores[145]. Sin embargo, teniendo en cuenta que los genes reguladores no tienen actividad directa sobre la expresión de genes involucrados en la comunicación neuronal y que su patrón de expresión es diferente a estos, se presupone que los genes reguladores de la expresión génica debe estar involucrados en otros procesos diferentes que tienen lugar en etapas más tempranas del neurodesarrollo como son la neurogénesis o la migración neuronal (Figura 14)[104]. Los hallazgos de los estudios de genoma completo y GWAS corroboran esta teoría, pues en ellos se han detectado multitud de variantes que afectan a la expresión de genes que desempeñan su función esencialmente durante el desarrollo fetal[48,121,124].

Estudio	*Tipo de Red*	*Alias*	*Principales procesos biológicos*
O´Roak et al., 2012	PPI	NA	Remodelado de la cromatina Via de señalización *Wnt/β-catenina*
Krumm et al., 2014	PPI	NA	Función sináptica, Remodelado de la cromatina Vía de señalización *Wnt/β-catenina t*
Ruzzo et al., 2018	PPI	NA	complejo SWI/SNF
Gilman et al., 2011	PPI	NETBAG	Desarrollo de las sinapsis Motilidad neuronal Guía axonal
Pinto et al., 2014	PPI	NETBAG	Transcripción y el remodelado de la cromatina, Vía de señalización MAPK Desarrollo de las sinapsis.
Rubeis et al., 2013	PPI	NA	Transcripción y remodelado de la cromatina Función sináptica
Voineagu et al., 2011	Co-expresión	NA	Función sináptica Respuesta inflamatoria
Parikshak et al., 2016	Co-expresión	NA	Función sináptica Respuesta inflamatoria
Hormozdiari et al., 2015	PPI	MAGI	Función sináptica Vía de señalización *Wnt/β-catenina*
Willsey et al., 2013	Co-expresión	NA	Neuronas glutamatérgicas en láminas 5 y 6
Parikshak et al., 2013	Co-expresión	NA	Neuronas glutamatérgicas en láminas 2-4
Satterstrom et al., 2019	Co-expresión	NA	Transcripción y remodelado de la cromatina Función sináptica

Tabla 15. Principales estudios llevados a cabo en los TEA en los que se ha realizado análisis de redes génicas. Se indican los principales procesos biológicos que fueron destacados en cada uno de ellos (Adaptada de Joon Yong An and Charles Claudianos, 2016[146]).

Los análisis de redes génicas y los análisis de expresión llevados a cabo en los estudios de TEA, ofrecen pues, en su conjunto, una visión

clara de los mecanismos biológicos implicados en su patogénesis. Así pues, todos ellos apuntan a que la alteración de procesos biológicos relacionados con la neurogénesis, la migración celular, el establecimiento de las sinapsis y su correcto funcionamiento durante las etapas tempranas del neurodesarrollo están ligados a las manifestaciones cognitivas que aparecen en los TEA.

2.4 AVANCES EN EL DIAGNÓSTICO MOLECULAR DE LOS TEA

Los estudios de investigación llevados a cabo en los TEA han generado un enorme conocimiento de sus bases genéticas. Se sabe que los TEA son trastornos complejos en cuya etiología intervienen tanto variantes comunes como variantes raras altamente penetrantes. Además, también se conocen los principales mecanismos biológicos implicados en su patogénesis.

Sin embargo, trasladar todo este conocimiento a la práctica clínica sigue suponiendo un reto. En un contexto clínico, el diagnóstico molecular de los TEA se limita a la búsqueda de variantes raras altamente penetrantes en genes con un papel firmemente demostrado en los TND.

2.4.1 Algoritmos diagnósticos en los TEA

2.4.1.1 TEA sindrómico

Los TEA sindrómicos son aquellos que se acompañan de un conjunto de características clínicas o psiquiátricas asociadas a un trastorno genético. Aproximadamente un 10% de los pacientes con TEA, pueden presentar un síndrome genético con causa conocida[147].

Los TEA sindrómicos pueden deberse a alteraciones cromosómicas microscópicas o mutaciones en genes únicos que causan trastornos autosómicos dominantes (AD), recesivos (AR), o ligados al cromosoma X (Tabla 16).

Síndrome	*Gen/ Región*	*Incidencia del síndrome en los TEA*	*Incidencia de los TEA en el síndrome*
X frágil	*FMR1*	2.1%	18-33%
Esclerosis tuberosa	*TSC1, TSC2*	1-4%	25-60%
Neurofibromatosis tipo 1	*NF1*	<1.4%	4%
Fenilcetonuria sin tratar	*PAH*	NA	5.70%
Deficiencia de adenilosuccinato liasa	*ADSL*	<1%	80-100%
Smith-Lemli-Opitz	*DHCR7*	<1%	46-52%
Cohen	*COH-1,* desconocido	<1%	48%
Cornelia Lange	*NIPBL, SMC1A, SMC3*	<1%	46-67%
Sotos	*NSD1*	<1%	NA
Cole-Hughes	*NA*	<1%	NA
Lujan-Fryns	*UPF3B, MED12*	<1%	62.5%
San Filipo			
A	*SGSH*	<1%	NA
B	*NAGLU*	<1%	35%
C	*HGSNAT*	<1%	7%
D	*GNS*	<1%	35.7%
ARX	*ARX*	<1%	50-81%
Deleción 2q37	2q37.3	<1%	35%
Williams-Beuren	Del7q11.23	<1%	7%
Duplicación Williams-Beuren	Dup7q11.23	<1%	35.7%
Angelman	Del/mut en alelo materno UBE3A	<1%	50-81%
Prader-Willi	Del. alelo paterno 15q11q13	NA	19-36.5%
Smith-Magenis	Del. 17p11.2	<1%	93%
Potocki-Lupsky	Dup 17q11.2	<1%	90%
Down	Trisomía cromosoma 21	1.7-3.7%	5.6-8%
Velocardiofacial	Del 22q11.2	<1%	20-31%

/Di George			
Duplicación 22q11	Dup 22q11.2	<1%	NA
Phelan-McDermid	Del 22q13.3	<1%	50-70%

Tabla 16. Principales síndromes genéticos asociados a los TEA. (Adaptada de de A. Persico and V. Napolini, 2013[148]).

Los trastornos monogénicos más frecuentes y con alta penetrancia en los TEA son el síndrome del X frágil (*FMR1*), esclerosis tuberosa (*TSC1* y *TSC2*), neurofibromatosis (*NF1*), síndrome de Rett (*MECP2*) y síndromes asociados a mutaciones en el gen *PTEN*[148–150].

Síndromes frecuentes asociados a alteraciones cromosómicas micróscopicas o submicroscópicas son la duplicación 15q11.q13 en el alelo materno que contiene la región relacionada con los síndromes Prader Willi/Angelman (detectada entre un 1-3% de los casos de TEA) las deleciones 2q31, 22q11.2 (síndrome Velo-cardio facial) y 22q13.3 (Phelan Mcdermit)[147]. También se incluyen aneuploidías como el síndrome de Down (trisomía del cromosoma 21), síndrome de Turner (45,X0), síndrome 47 (XXX), síndrome Klinefelter (47, XXY) y síndrome XYY (47,XYY)[149,151,152].

El diagnóstico molecular de los TEA sindrómicos depende directamente de la pericia y experiencia del clínico (Figura 15). Una correcta y detallada exploración del paciente y de su historial médico es esencial durante este proceso. Así, es necesario la revisión del historial clínico en busca de pruebas previas de neuroimagen, comorbilidades neuropsiquiátricas (DI, trastornos de ansiedad, trastornos del sueño…) y comorbilidades médicas (epilepsia, anomalías congénitas…). Durante el examen físico debe recogerse información sobre el peso, la altura, el perímetro craneal o la existencia de rasgos dismórficos poniendo especial atención en la cara, manos y pies. En el caso de la piel, es importante buscar anomalías dermatológicas que se puedan relacionar con neurofibromatosis (manchas café con leche, neurofibromas y pliegue inguinal) o con esclerosis tuberosa (máculas hipopigmentadas, angiofibromas faciales y fibromas)[153].

Tras un examen clínico adecuado, muchas de las formas sindrómicas de los TEA pueden identificarse simplemente por sus características clínicas o por algún marcador biológico como es el caso de las metabolopatías. Estos casos, se confirman con un test diagnóstico

dirigido a detectar la mutación genética de sospecha. Las formas sindrómicas más fácilmente identificables clínicamente, incluso sin experiencia previa, son el síndrome de Down, la neurofibromatosis, la esclerosis tuberosa, el síndrome del X frágil y síndromes asociados a mutaciones en *PTEN*[153].

En caso de no confirmarse el diagnóstico de sospecha o no identificar el síndrome a través de la exploración física y las pruebas médicas, lo más indicado será aplicar los algoritmos diagnósticos reservados para los TEA no sindrómicos.

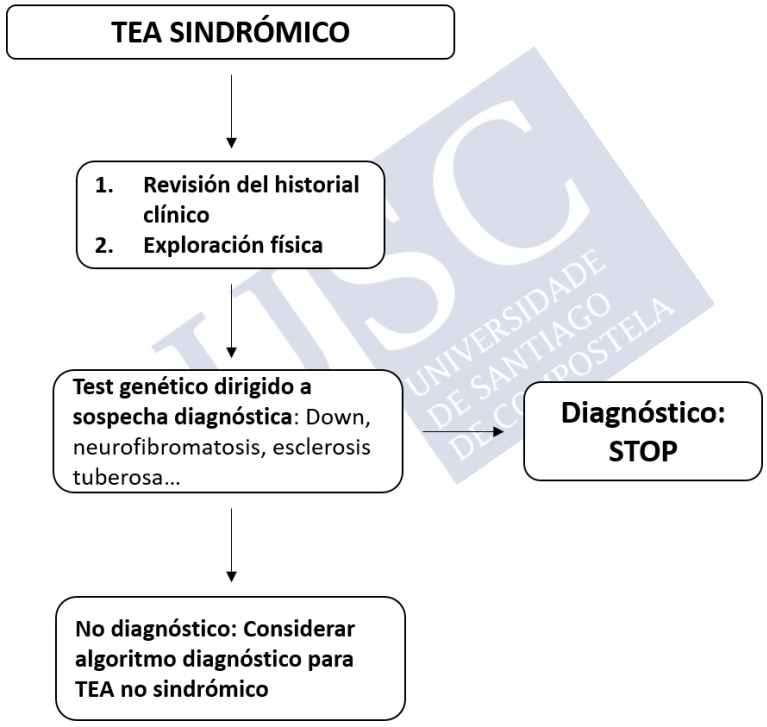

Figura 15. Algoritmo diagnóstico en los TEA sindrómicos

2.4.1.2 TEA no sindrómico

Los TEA no sindrómicos o idiopáticos, son aquellos en los que no es posible detectar un marcador biológico o genético que explique el patrón de características clínicas. No obstante, esto no excluye la

presencia de comorbilidades clínicas y neuropsiquiátricas propias de los TEA.

Entre los algoritmos diagnóstico de los TEA no sindrómicos se distinguen los test de cribado de primera línea (*microarray* cromosómico, cribado del X frágil y cribado para mutaciones en *MECP2* y *PTEN)* y los test de cribado de segunda línea (secuenciación de paneles de genes o secuenciación del exoma completo).

La asignación de un test a un grupo u a otro, se realiza en base a evidencias tanto clínicas como económicas. Estas evidencias pueden ser el coste de la técnica, el rendimiento diagnóstico, o cuestiones técnicas y prácticas sujetas a cada test. Es de esperar, por tanto, que los algoritmos diagnósticos cambien a medida que las herramientas diagnósticas evolucionan.

En el contexto de los TEA, el alto rendimiento diagnóstico de las técnicas de NGS sumado a la disminución de su coste en los últimos años, propicia que empiecen a ser consideradas herramientas diagnósticas de primera línea. No obstante, no existe aún un consenso real que justifique este cambio en su utilización. Por lo tanto, en este apartado se tratarán de abordar en detalle las guías diagnósticas vigentes.

2.4.1.3 Test de cribado de primera línea

En el año 2010, Miller *et al.*, publicaron un trabajo donde se establecía el *microarray* cromosómico como la primera herramienta diagnóstica en pacientes diagnosticados con DI/RGD, TEA, o anomalías congénitas. Así, una revisión exhaustiva de 33 trabajos previos llevó a Miller y colaboradores a comprobar como el rendimiento diagnóstico del *microarray* cromosómico era muy superior al del cariotipo convencional (15-20% vs 3%)[154].

Los *microarrays* cromosómicos no son capaces de detectar alteraciones estructurales balanceadas (translocaciones e inversiones), que sí detectan las técnicas citogenéticas. Sin embargo, se calcula que menos de 1% de los casos con un *microarray* normal presenta algún tipo de alteración estructural balanceada, y, cuando la presenta, es difícil determinar su relación con el fenotipo[155].

No obstante, el cariotipo convencional se debe considerar antes que el *microarray* cromosómico en casos donde haya una sospecha clínica de una aneuploidía cromosómica, o en casos donde exista una historia familiar de reordenamientos cromosómicos o de múltiples pérdidas reproductivas o de infertilidad[154,156].

Además, el cariotipo también es útil para determinar si los padres de un individuo con una translocación no balanceada son portadores de una translocación balanceada, por el riesgo de recurrencia que ello conlleva[155].

El cribado para X frágil se considera también un test de rutina debido a que es la causa más frecuente de DI heredada, y en muchas poblaciones una de las causas genéticas más descritas de TEA, explicando entre el 0.5-2% de los casos.

Este síndrome, se produce por una expansión de más de 200 copias de tripletes de nucleótidos CGG en el gen *FMR1*, localizado en el cromosoma X. Esta expansión conduce a una metilación de las bases de citosina lo que conlleva a un silenciamiento de la transcripción, y, por tanto, a la deficiencia o ausencia de su producto proteico, FMRP (de sus siglas en inglés, *Fragile X Mental Retardation 1 Protein*). FMRP, es una proteína de unión al ARN, que regula la expresión de cientos de genes involucrados en la plasticidad sináptica[157]. Por otro lado, una expansión del triplete en un rango entre 55 y 200 copias se considera una premutación[158].

En varones, un 90% de los individuos con la mutación completa presenta síntomas autistas y un 60% reúne los criterios para ser diagnosticado como TEA. En el caso de las mujeres, sin embargo, el impacto de la mutación es atenuado por la presencia de un segundo cromosoma por el fenómeno de inactivación del cromosoma X, lo cual explica que solo entre un 30-50% presente sintomatología[159]. Cuando hay una premutación, los síntomas característicos del X frágil no suelen estar presentes, aunque se han descrito dificultades en el aprendizaje, problemas emocionales o incluso DI[160]. En el caso de las mujeres, una expansión superior a 80 copias se relaciona con un incremento del riesgo para insuficiencia ovárica prematura[161].

De acuerdo a estas consideraciones clínicas el cribado para el X frágil está especialmente indicado en varones con TEA sin causa

conocida, pero no se considera un test de rutina en mujeres. Sin embargo, sí se recomienda su uso en mujeres que reúnan una serie de requisitos como: un fenotipo compatible con X frágil, una historia familiar de TND ligados al cromosoma X y/o insuficiencia ovárica prematura[156].

Por último, se recomienda también el cribado para mutaciones en *MECP2* y *PTEN*.

Las mutaciones *de novo* en *MECP2* se relacionan con el síndrome de Rett, que ocurre casi de manera exclusiva en mujeres y muy raramente en varones. Las mujeres afectas se desarrollan con normalidad durante los primeros 6-18 meses y luego experimentan un síndrome de regresión, en el cual pierden capacidades adquiridas y aparecen criterios diagnósticos para TEA[162]. Se debe tener en cuenta, sin embargo, que entre un 0.8-1.3% de las mujeres diagnosticadas con TEA pero sin clínica compatible para Rett, portan una mutación en *MECP2*[163].

Por estas características, su cribado de rutina está solo indicado en mujeres con TEA y/o RGD, y en particular en mujeres con características clínicas propias del síndrome de Rett[156].

Las mutaciones en *PTEN* dan lugar a un conjunto de trastornos con síntomas solapantes entre los que se incluye el síndrome de Cowden, el síndrome Bannayan-Riley-Ruvalcaba, el síndrome de Proteus relacionado con *PTEN* y el síndrome de macrocefalia/TEA. Aunque la macrocefalia está descrita en un 20% de los pacientes con TEA, el grado de macrocefalia es muy superior en los pacientes con mutaciones en el gen *PTEN*[164]. Por ese motivo, se recomienda el *screening* en pacientes con TEA y con un perímetro craneal de más de 2.5 desviaciones estándar (DE) por encima de la media[156] (Figura 16).

Figura 16. Algoritmo diagnóstico en los TEA no sindrómicos. Herramientas de cribado de primera línea.

2.4.1.4 Test de cribado de segunda línea

La disminución del coste de las técnicas de NGS ha fomentado su implantación en los laboratorios como una herramienta rutinaria de diagnóstico clínico.

Existen 3 aplicaciones diferentes de las técnicas de NGS: secuenciación de un panel de genes, secuenciación de exoma completo y secuenciación de genoma completo.

La secuenciación de paneles de genes consiste en la secuenciación de un grupo de genes seleccionados por su implicación en una patología de interés. Esta aproximación permite incrementar la profundidad de lectura en las regiones de interés a un bajo coste, así como disminuir la probabilidad de hallazgos incidentales, que no estén relacionados con la enfermedad estudiada[165]. En el caso de los TEA, no existe un consenso a la hora de crear listados de genes relacionados con el trastorno.

En el trabajo de N. Hoang *et al.*, se realizó un estudio comparativo de los paneles ofertados por 21 laboratorios diferentes y encontraron grandes diferencias entre ellos. Los genes incluidos en cada panel variaban en número en un rango entre 11 y 2562 y solo 178 solaparon en al menos 5 paneles. Además, se detectó una falta de información generalizada sobre los criterios empleados para la selección de genes o incluso sobre cuestiones técnicas relacionadas con los análisis llevados a cabo o la interpretación de las variantes (Tabla 17)[166].

Gen	*Paneles que incluyen el gen*	*OMIM*
MECP2	21/21	Encefalopatía (#300673), Discapacidad intelectual (#300260, #300055), síndrome de Rett (#312750), susceptibilidad a TEA (#300496)
NLGN4X	20/21	Discapacidad intelectual (#300495), susceptibilidad a síndrome de Asperger (#300497), susceptibilidad a TEA (#300495)
CACNA1C	19/21	Síndrome de Brugada 3 (#611875), síndrome de Timothy (#601005)
NRXN1	19/21	Síndrome de Pitt-Hopkins-like 2 (#614325), susceptibilidad a esquizofrenia (#614332)
PCDH19	19/21	Encefalopatía epiléptica (#300088)
PTCHD1	19/21	Suceptibilidad a TEA (#300830)
UBE3A	19/21	Síndrome de Angelman (#105830)
NLGN3	18/21	susceptibilidad a síndrome de Asperger (#300494), Suceptibilidad a TEA (#300425)
PTEN	18/21	Síndrome Bannayan-Riley-Ruvalcaba (#153480), síndrome de Cowden/ Lhermitte-Duclos (#158350); syndrome macrocefalia/TEA (#605309), asociación con macrocefalia y ventriculomegalia (#276950), susceptibilidad a glioma (#613028), meningioma (#607174); cáncer de próstata (#176807)
SHANK3	18/21	Síndrome Phelan-McDermid (#606232), esquizofrenia (#613950)
CDKL5	17/21	Encefalopatía epiléptica (#300672)
CNTNAP2	17/21	Síndrome Pitt-Hopkins like 1 (#610042), susceptibilidad a TEA (#612100)
DHCR7	17/21	Síndrome Smith-Lemli-Opitz (#270400)
FOXP1	17/21	Discapacidad intellectual con ausencia de lenguaje con o sin rasgos autistas (#613670)
NSD1	17/21	Leucemia (#601626), síndrome de Sotos 1 (#117550)
ARX	16/21	Encefalopatía epiléptica (#308350), Hidranencefalia con genitales anómalos (#300215), lisencefalia (#300215), Discapacidad intelectual (#300419), síndrome de Partington (#309510), síndrome de Proud (#300004)

Tabla 17. Genes más frecuentemente incluidos en paneles. (Adaptada de Hoang *et al.*, 2018[166]).

Pese a las ventajas que puedan ofrecer los paneles de genes en otros trastornos de herencia mendeliana, en los TEA, los beneficios de esta aproximación son limitados debido a la complejidad del trastorno. No solo existe una falta de consenso sobre qué genes se deben seleccionar,

sino que muchos de los genes que contribuyen al riesgo en los TEA permanecen aún sin identificar. A todo ello se le une la elevada heterogeneidad fenotípica de los TEA. Fuera de las formas sindrómicas comentadas en el apartado anterior, resulta muy difícil plantear una hipótesis previa ante un paciente con TEA que a menudo manifiesta comorbilidades asociadas. Además, puede ocurrir que los síntomas asociados a un posible síndrome no se encuentren presentes en el momento de la exploración clínica dificultando aún más el diagnóstico. Así pues, lo más indicado es hacer un exoma completo donde se analizan todos los genes del genoma caracterizados funcionalmente, sin la necesidad de sospecha previa de un diagnóstico[167].

En la actualidad no existen guías específicas que recomienden cuando usar la secuenciación de exoma completo en los TEA, pero el Colegio Americano de Genética Médica y Genómica o ACMG (de sus siglas en inglés, *American College of Medical Genetics and Genomics*) recomienda esta aproximación en casos en los que la historia familiar o el fenotipo del paciente sea sugestivo de un trastorno genómico, para fenotipos donde haya una alta heterogeneidad genética o en los casos donde hayan fallado otros test genéticos[168]. Así, pues, los TEA pueden beneficiarse de la secuenciación de exoma completo, debido a que a menudo reúnen los criterios anteriormente mencionados.

La secuenciación de exoma completo es usada generalmente como herramienta diagnóstica en casos de TEA cuyo *microarray* cromosómico ha sido negativo, alcanzando un rendimiento diagnóstico de aproximadamente un 8.4% aunque esta estimación varía enormemente de un estudio a otro, según los criterios de inclusión de la cohorte de estudio (Figura 17)[169].

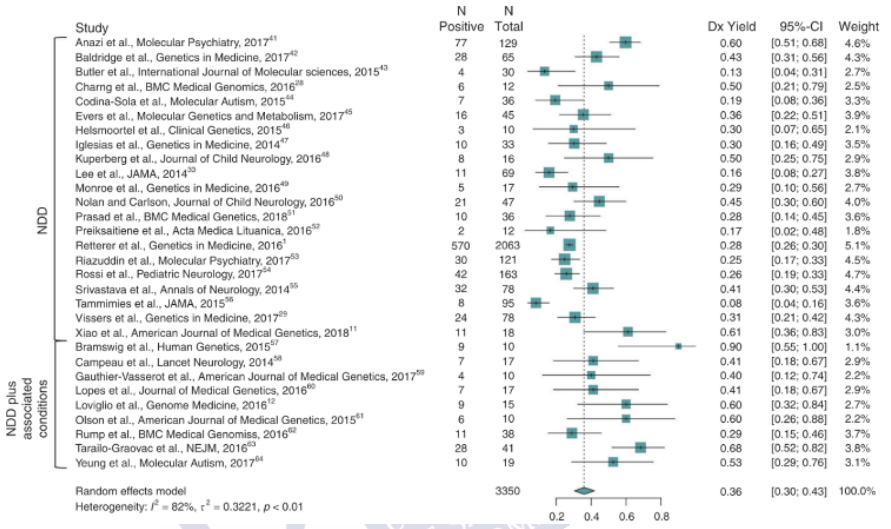

Figura 17. Rendimiento diagnóstico del exoma completo en cohortes con TND y cohortes con TND y otras patologías asociadas, en los estudios seleccionados por Srivastava *et al*. (Extraída de Srivastava *et al.*, 2019[170], permitido por Springer Nature).

Recientemente, un grupo de expertos ha llevado a cabo una revisión sistemática de la literatura con el objetivo de analizar el rendimiento diagnóstico de la secuenciación de exoma completo en los TND frente al de los *microarrays* cromosómicos. En base a sus resultados, en los que se demuestra un rendimiento diagnóstico del exoma completo significativamente superior al de los *microarray* (31-56% vs 15-20%), se recomienda el uso de esta aproximación como primera herramienta diagnóstica en casos de TND cuya causa se desconoce[170]. Faltan, sin embargo, estudios de coste-eficacia que respalden el uso de la secuenciación de exoma completo como herramienta de primera línea.

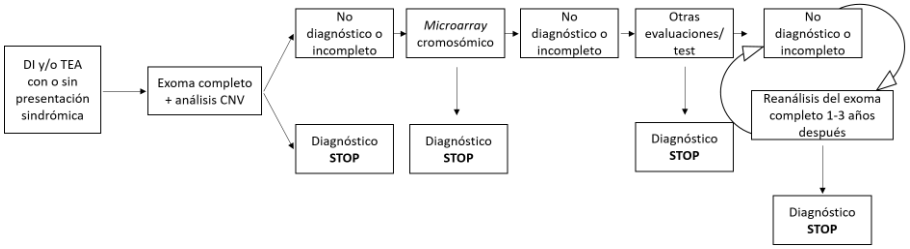

Figura 18. Algoritmo diagnóstico que incorpora la secuenciación de exoma completo en la evaluación clínica de pacientes con TND sin causa conocida propuesto por Srivastava *et al*. (Adaptada de Srivastava *et al.*, 2019[170], permitido por Springer Nature).

En contraposición con la secuenciación de exoma completo, la secuenciación de genoma completo no es una técnica que se emplee de manera rutinaria en el diagnóstico de los TEA. Pese a que ofrece la ventaja de detectar de toda la variabilidad genética en un solo test, su uso aún es incipiente en la práctica clínica. Así pues, el balance coste-beneficio aún no permite colocar esta aproximación en el algoritmo diagnóstico debido a su elevado coste, el tiempo que requiere el análisis de los datos generados y el reto que supone la interpretación de variantes no codificantes, especialmente en el contexto clínico[171].

2.4.2 Priorización e interpretación de variantes en datos de secuenciación de exoma completo.

Aunque el empleo de técnicas de NGS en el diagnóstico de individuos con TEA es una aproximación cada vez más común, la interpretación de las variantes detectadas sigue suponiendo un reto en este trastorno. El motivo es la ausencia de guías específicas en un trastorno tan complejo, donde fenómenos como la penetrancia incompleta o la expresividad variable son más la norma que la excepción. Por ello, en el contexto clínico, se usan las recomendaciones del ACMG para la interpretación de variantes, pese que estas han sido pensadas para trastornos de herencia mendeliana. Estas guías, elaboradas por un grupo de expertos, pretenden fomentar el uso de una terminología estándar a la hora de clasificar variantes en la comunidad clínica y científica[171].

En el análisis de un exoma completo se distinguen dos procesos fundamentales: la priorización y la clasificación de las variantes.

En el proceso de priorización se determina qué variantes, de entre las miles detectadas, se relacionan con mayor probabilidad con la causa de la enfermedad de interés. Para esta tarea, se parte de un archivo de variantes genéticas anotado, que proporciona información sobre la posición de las variantes en el genoma y su efecto funcional. Así, se pueden obviar rápidamente variantes sinónimas, cuyo impacto en la enfermedad se asume menor, y variantes intrónicas o localizadas en regiones 5´UTR o 3´UTR[172].

A continuación, se usan múltiples bases de datos, así como la literatura científica para continuar con el proceso de priorización.

Existen bases de datos poblacionales, que contienen un catálogo de la variación genética existente en la población humana y que son una fuente de información crucial a la hora de interpretar una variante. Estas bases de datos proporcionan información sobre su frecuencia alélica permitiendo distinguir entre variantes raras o comunes (Tabla 18)[173].

Base de datos	*Descripción*
Exome Aggregation Consortium (ExAC)	Contiene información de las variantes detectadas en el exoma completo de aproximadamente 61000 individuos no relacionados entre sí pertenecientes a 6 poblaciones diferentes.
Exome Variant Server	Contiene información de variantes detectadas en el exoma completo de diversas cohortes de individuos de descendencia europea y afroamericana.
1000 Genomes Project	Contiene información de variantes detectadas en la secuenciación del genoma completo de 2504 individuos pertenecientes a 24 poblaciones diferentes.
The Genome Aggregation Database (gnomAD)	Este proyecto sucede al ExAC, y contiene información de variantes detectadas en 125748 secuencias de exoma completo y 15708 secuencias de genoma completo
dbSNP	Contiene información de variantes genéticas de aproximadamente <50 pb
dbVar	Contiene información de variantes estructurales de aproximadamente >50 pb

Tabla 18. Bases de datos poblacionales. (Adaptada de Richards *et al.*, 2015[173]**).**

También existen bases de datos específicas de enfermedad que contienen variantes detectadas en pacientes y suelen incluir una valoración sobre la patogenicidad de la variante. Otras bases de datos son aquellas específicas de genes, que establecen una conexión entre los genes y la enfermedad. Estas bases de datos hay que usarlas con mucha cautela porque a menudo contienen variantes incorrectamente clasificadas (Tabla 19)[172].

Base de datos	*Descripción*
ClinVar	Contiene información de variantes junto a su interpretación clínica y su fenotipo asociado.
OMIM	Contiene información sobre aproximadamente 15000 genes asociados a trastornos genéticos y proporciona información fenotípica de aproximadamente 5000 trastornos.
Human Gene Mutacion Database	Contiene anotaciones de variantes publicadas en la literatura

Tabla 19. Bases de datos específicas de enfermedad y de gen. (Adaptada de Richards *et al.*, 2015[173]).

Por último, existen una gran cantidad de herramientas que predicen computacionalmente el posible impacto funcional de una variante, ya que no es siempre posible en la rutina clínica recurrir a estudios funcionales. Hay dos categorías principales: herramientas que predicen si una variante *missense* tiene un impacto funcional en la proteína resultante y herramientas que predicen el efecto de variantes de *splicing*.

Los criterios para medir el impacto de variantes *missense* se basan generalmente en el grado de conservación entre especies del aminoácido o del nucleótido, la localización del aminoácido en la proteína y las consecuencias bioquímicas de esta sustitución aminoacídica.

Las herramientas para medir el impacto de las variantes de *splicing* predicen la creación o pérdida de sitios de empalme a nivel exónico e intrónico.

Generalmente, los predictores combinan diferentes criterios para aumentar la especificidad y sensibilidad de sus algoritmos. No obstante, es recomendable usar más de uno, pues la información que proporciona cada uno de ellos es complementaria (Tabla 20)[173] .

Categoría	*Nombre*	*Evidencia*
Predictores variantes missense	ConSurf	Conservación evolutiva
	FATHMM	Conservación evolutiva
	MutationAssessor	Conservación evolutiva
	PANTHER	Conservación evolutiva
	PhD-SNP	Conservación evolutiva
	SIFT	Conservación evolutiva
	SNPs&GO	Estructura/función de la proteina
	Align GVGD	Estructura/función de la proteína y conservación evolutiva
	MAPP	Estructura/función de la proteína y conservación evolutiva
	MutationTaster	Estructura/función de la proteína y conservación evolutiva
	MutPred	Estructura/función de la proteína y conservación evolutiva
	Polyphen-2	Estructura/función de la proteína y conservación evolutiva
	PROVEAN	Usa el grado de conservación en la secuencia proteica para predecir el impacto funcional de la variante
	nsSNPAnalyzer	Múltiples alineamiento de secuencia y análisis de estructura proteica
	Condel	Combina SIFT, PolyPhen-2 y MutationAssesor
	CADD	Integra múltiples anotaciones en una sola medida al comparar mutaciones que no son eliminadas mediante selección natural
Predictores de splicing	GeneSplicer	Modelos de Markov
	Human Splicing Finder	Dependiente de la posición
	MaxEntScan	Principio de máxima entropia
	NetGene2	Red neural
	NNSplice	Red neural
	FSPLICE	Predictor para sitios de splicing específico de especies basado en un modelo de matrices
Predictores de nucleótidos conservados	GERP	Mide el grado se conservación de la secuencia en el genoma humano a partir del alineamiento de 43 genomas de otros vertebrados
	PhastCons	Usa múltiples alineamientos de secuencias procedentes de diversas especies para identificar elementos conservados
	PhyloP	Alineamiento y árboles filogenéticos

Tabla 20. Predictores *in silico* de patogenicidad. (Adaptada de S. Richards *et al.*, 2015[173]).

Una vez priorizadas las variantes, el ACMG recomienda su clasificación en 5 categorías: patogénicas, probablemente patogénicas, benignas, probablemente benignas y de significado incierto.

Existen criterios de clasificación de variantes patogénicas (Tabla 21) y de variantes benignas (Tabla 22)[173]:

Evidencia	*Categoría*
Muy alta	PVS1. Variantes que llevan a una disrupción total del producto génico: variantes *nonsense*, variantes *missense*, variantes en sitios de *splicing* canónicos ± 1 y 2, variantes en el codón de inicio de la transcripción, variantes que conducen a la deleción de uno o varios exones. Advertencias: • Hay variantes LoF que no son patogénicas en algunos genes. • Las variantes LoF en el extremo final 3´ se deben interpretar con cautela. • Interpretación con cautela de variantes de *splicing*. Se requiere comprobar su efecto con un análisis funcional del ARNm o el producto proteico. • Considerar la presencia de múltiples transcritos cuyo patrón de expresión puede variar de un tejido a otro.
Alta	PS1. La variante provoca el mismo cambio aminoacídico que otra variante que se ha establecido ya como patogénica. PS2. Variante *de novo* en un paciente con la enfermedad y sin historia familiar. PS3. Estudios funcionales in vitro y/o in vivo han confirmado el efecto patogénico de la variante en el gen o su producto proteico. PS4. La prevalencia de la variante en individuos afectos es significativamente superior que en individuos no afectos.
Moderada	PM1. La mutación se localiza en un dominio funcional crítico sin que se hayan descrito variantes benignas en esa posición. PM2. La variante no se ha identificado nunca en controles, o en caso de estar ligada a una enfermedad autosómica recesiva se ha detectado a frecuencias extremadamente bajas. PM3. En trastornos autosómicos recesivos, la variante se encuentra en *trans*, con respecto a una variante patogénica. PM4. La variante provoca un cambio en la longitud de la proteína como consecuencia de una deleción o inserción en una zona no repetitiva. PM5. Variante que provoca un cambio aminoacídico en una posición donde otro cambio aminoacídico ha sido descrito como patogénico. PM6. La variante se presume *de novo* pero no hay confirmación en los progenitores.

Sugestiva	PP1. La variante cosegrega con la enfermedad en varios miembros afectos de la misma familia y se localiza en un gen que se ha visto implicado en la enfermedad. PP2. Variante *missense* en un gen donde se han detectado pocas variantes benignas de este tipo, y, por el contrario, se han reportado otras variantes *missense* patogénicas. PP3. Muchos predictores de patogenicidad *in silico* predicen que la variante es patogénica. PP4. El fenotipo del paciente o la historia familiar es muy específico para una enfermedad con una etiología genética simple. PP5. Una fuente de información acreditada ha reportado recientemente la variante como patogénica pero la evidencia para hacerlo aún no está disponible.

Tabla 21. Criterios de evidencia ACMG para la clasificación de variantes patogénicas. PVS1: *pathogenic very strong*; PVS1-4: *pathogenic strong* 1-4; PM1-M6: *pathogenic moderate* 1-6; PP1-5: *pathogenic supporting* 1-5. (Adaptada de Richards *et al.*, 2015[173]).

Evidencia	*Categoría*
Suficiente	BA1. MAF > 5% en bases de datos poblacionales
Alta	BS1. MAF superior a lo esperado para la enfermedad BS2. La variante se ha observado en individuos sanos para trastornos autosómicos recesivos (homocigotos), dominantes (heterocigotos) o ligados al cromosoma X (hemicigosis) que tienen alta penetrancia. BS3. Estudios *in vitro/in vivo* han confirmado el efecto benigno de la variante en el gen o su producto proteico. BS4. La variante no cosegrega en múltiples miembros afectos de una misma familia.
Sugestiva	BP1. Variante con cambio de sentido en un gen en el cual se han reportado principalmente variantes LoF como causa de enfermedad. BP2. En trastornos autosómicos dominantes altamente penetrantes, la variante se encuentra en *trans*, con respecto a una variante patogénica, o se observa en *cis* con respecto a una variante patogénica, sea cual sea la herencia del trastorno. BP3. Deleción o inserción que no provoca un cambio en el marco de lectura en una zona repetitiva sin función conocida. BP4. Muchos predictores de patogenicidad *in silico* predicen que la variante es benigna. BP5. Variante detectada en un individuo para el cual ya existe un diagnóstico molecular. BP6. Una fuente de información acreditada ha reportado recientemente la variante como benigna pero la evidencia para hacerlo aún no está disponible. BP7. Variante sinónima para la cual los algoritmos predictores de *splicing* predicen que no hay un impacto funcional.

Tabla 22. Criterios de evidencia ACMG para la clasificación de variantes benignas. BA1: *benign stand-alone*; BS1-4: *benign strong* 1-4; BP1-7: *benign supporting* 1-7. (Adaptada de Richards *et al.*, 2015[173]).

La combinación de los criterios anteriormente mencionados para la clasificación de variantes se resume en la Tabla 23[173]:

Clasificación	*Combinación de criterios*
Patogénica	1. PSV1 y >1 PS1-PS4 o >2 PM1-PM6 o 1 PM1-PM6 y 1 PP1-PP5 o >2 PP1-PP5 2. >2 PS1-PS4 3. 1 PS1-PS4 y >3 PM1-PM6 o 2 PM1-PM6 y >2 PP1-PP5 o 1 PM1-PM6 y >4 PP1-PP5
Probablemente patogénica	1. 1 PVS1 y 1 PM1-PM6 2. 1 PS1-PS4 y 1-2 PM1-PM6 3. 1 PS1-PS4 y >2 PP1-PP5 4. >3 PM1-PM6 5. 2 PM1-PM6 y 2 PP1-PP5 6. 1 PM1-PM6 y >4 PP1-PP5
Benigna	1. BA1 2. >2 BS1-BS4
Probablemente benigna	1. 1 BS1-BS4 y 1 BP1-BP7 2. >2 BP1-BP7
Significado incierto	Otros criterios no mencionados o los criterios de patogenicidad o benignidad contradictorios

Tabla 23. Clasificación de variantes según criterios ACMG. (Adaptada de Richards *et al.*, 2015[173]).

En conclusión, aunque las guías del ACMG se emplean de manera rutinaria en el diagnóstico genético de los pacientes con TEA, se debe tener presente que dichas guías estas pensadas para el diagnóstico de trastornos genéticos de herencia mendeliana, y, por tanto, no se adecúan correctamente a las características de los TEA. Cabe destacar, de que a pesar de que en las últimas décadas se ha estudiado en detalle la arquitectura genética de los TEA, el conocimiento generado ha repercutido muy poco en la práctica clínica. Para mejorar el diagnóstico de estos pacientes, se requiere, por tanto, la elaboración de algoritmos diagnósticos y guías clínicas para la interpretación de variantes que sean específicas y estén adaptadas a nuestro conocimiento actual de los TEA.

3 JUSTIFICACIÓN Y OBJETIVOS

Los estudios genéticos llevados a cabo en las últimas décadas, han demostrado que en la etiología de los TEA intervienen principalmente factores genéticos. Se estima que la variación común puede explicar aproximadamente un 50% del componente genético de los TEA mientras que la variación rara explicaría en torno a un 12%.

Para poder entender la complejidad genética que caracteriza a los TEA, la aproximación más eficaz consiste en buscar aquellas rutas biológicas en las cuales converge toda la variabilidad genética. Para ello se emplean herramientas bioinformáticas capaces de integrar información biológica a nivel molecular, celular o anatómico. Dado que tanto la variación común como la variación rara son factores de riesgo genético en los TEA, el empleo de análisis bioinformáticos que apliquen la biología de sistemas, debería llevarse a cabo con datos procedentes tanto de estudios GWAS como de estudios de secuenciación. De esa forma, se podría ofrecer una visión global y completa de la arquitectura genética de los TEA y de sus bases biológicas.

Aunque la contribución de la variación rara al total de la heredabilidad, es limitada, el riesgo que confiere a nivel individual es muy alto. Por ese motivo, el diagnóstico genético en individuos con TEA se basa en la búsqueda de variantes raras con un alto impacto funcional. Actualmente los algoritmos diagnósticos recomiendan los *array* cromosómicos como primera herramienta diagnóstica seguida de la secuenciación del exoma completo en caso de no encontrar CNVs clínicamente significativas. Sin embargo, el rendimiento diagnóstico de la secuenciación del exoma completo es significativamente superior y si su coste continúa disminuyendo es probable que en un futuro próxima sea la herramienta diagnóstica por excelencia en casos de TEA y otros TND.

Por todo lo anterior el objetivo general de esta tesis es estudiar las bases genéticas de los TEA desde diferentes perspectivas para obtener una visión global de su arquitectura genética. Para ello se estudiará de manera independiente la contribución de la variación común y rara al riesgo genético de los TEA. Por otro lado, se abordará el estudio de los TEA desde un punto de vista clínico y se analizarán los beneficios de la incorporación de las tecnologías de secuenciación masiva al diagnóstico genético de los TEA.

Los objetivos específicos de esta tesis son:

1. Estudio de la variación común y su contribución al riesgo en los TEA:
    - Detección de genes asociados en los TEA empleando una estrategia GBA que usa como *input* los *summary statistic*s del mayor metaanálisis de GWAS de TEA realizado hasta el momento.
    - Caracterización funcional *in silico* de los genes asociados y sus interactores.
2. Estudio de la variación rara y su contribución al riesgo de los TEA en un estudio de secuenciación de exoma completo realizado en una cohorte de 360 tríos:
    - Detección de mutaciones *de novo* (germinales y postcigóticas) e identificación de genes de riesgo en los TEA.
    - Análisis del impacto funcional de las mutaciones germinales y postcigóticas en las diferentes jerarquías biológicas (gen, términos GO, tipos celulares, áreas cerebrales y periodos del desarrollo).
3. Aplicación de tecnologías de secuenciación de exoma completo en el diagnóstico de individuos con TEA.
    - Cálculo del rendimiento diagnóstico de la secuenciación de exoma completo usando una aproximación de trío completo.

Estos tres objetivos específicos se abordan en cada uno de los capítulos de esta tesis.

4 CAPÍTULO 1

4.1 OBJETIVO

El mayor metaanálisis de GWAS de TEA fue llevado a cabo por el PGC y sus resultados se encuentran publicados en el trabajo de Grove *et al.*, 2019. En dicho análisis se empleó también una estrategia de GBA, utilizando el algoritmo MAGMA.

El presente trabajo ha sido publicado en *Frontiers in Genetics*, con el título "*Novel Gene-Based Anlysis of ASD GWAS: Insight Into the Biological Role of Associated Genes*" (https://doi.org/10.3389/fgene.2019.00733). Su objetivo principal ha sido realizar un GBA diferente, empleando el algoritmo PASCAL, utilizando los resultados obtenidos en el metaanálisis de GWAS de TEA de Grove *et al.*, con el fin de encontrar nuevas asociaciones a nivel gen, así como realizar una caracterización funcional *in silico* de los nuevos genes asociados.

4.2 MÉTODOS

4.2.1 Metaanálisis de GWAS de TEA

Los *summary statistics* del último metaanálisis de GWAS de TEA llevado a cabo por el PGC se obtuvieron del repositorio público de este consorcio (https://www.med.unc.edu/pgc/download-results/). El *set* de datos que fue incluido en este estudio fue el siguiente: iPSYCH_PGC_ASD_Nov2017.gz. Este archivo, además de contener los *summary statistics*, incluye información adicional para cada SNP como su posición genómica, alelo de referencia, alelo alternativo, información sobre el *score* de calidad de imputación, OR y error estándar (SE).

Las características de las cohortes incluidas en el metaanálisis se detallan en la Tabla 24. El número total de individuos que reunieron fue de 18381 casos y 27969 controles.

Dado que los datos que se han empleado en este trabajo son públicos y se encontraban anonimizados, no fue preciso contar con la aprobación previa de un comité de ética para llevar a cabo este estudio.

Estudio		Diseño	Tamaño muestra Caso/ control	Descendencia	Instrumento diagnóstico	Total
iPSYCH		Caso/ control	13076/ 22664	Europea	criterios ICD10	
	UCLA ACE	Trío	391/ 391	Europea	ADI-R y/o ADOS	18381 / 27969
PGC	AGP	Trío	2272/ 2272	Europea	ADI-R y ADOS	
	AGRE	Trío	974/ 974	Europea	ADI-R y ADOS	
	MOMBOS	Trío	1396/ 1396	Europea	ADI-R, *Autism Screening questionnaire*, ADOS	
	SSC	Trío	2231/ 2231	Europea	ADI-R y ADOS	

Tabla 24. Características de las cohortes de TEA incluidas en el metaanálisis de TEA llevado a cabo por el PGC. Abreviaturas: ICD10: *International Classification of Diseases*, 10th revision; ADI-R: *Autism Diagnostic Interview-Revised*; ADOS: *Autism Diagnostic Observation Schedule*; ACE: *Autism Center of Excellence*; AGP: *Autism Genome Project*; AGRE: *Autism Genetic Resource Exchange*; MOMBOS: *NIMH Repository, the Montreal/Boston Collection*; SSC: *Simons Simplex Collection*.

4.2.2 Análisis basado en genes (GBA)

Para realizar el GBA se se utilizaron los *summary statistics* del metaanálisis de TEA como *input* para el algoritmo PASCAL. Los siguientes parámetros se emplearon en el análisis:

1. En primer lugar, se asignaron cada uno de los SNPs del metaanálisis a los genes correspondientes. Para ello se emplearon las anotaciones de los genes RefSeq de UCSC (*University of California Santa Cruz*) (versión hg19). El límite teórico para establecer el comienzo y final de cada gen fue de ± 50 kb.
2. Se incluyó un número máximo de 3000 SNPs por cada gen.
3. Se consideró la información de LD procedente del panel europeo de 1000 Genomas.

4. El umbral de significación estadística se estableció en 2.26 x 10^{-6} según la corrección de Bonferroni (0.05/22135 genes inluidos en el análisis).

4.2.3 *Plots* de asociación regionales

La representación visual de los resultados del metaanálisis en las regiones de interés (regiones donde se localizan los genes que PASCAL identificó como asociados) se hizo con la herramienta LocusZoom (http://locuszoom.org/). LocusZoom permite representar gráficamente los *loci* asociados en un GWAS, además de proporcionar información de LD en la región y los puntos de recombinación.

Para ello se usó el archivo original que contiene los *summary statistics* del metaanálisis (iPSYCH_PGC_ASD_Nov2017.gz), y se especificó la región genómica a representar de acuerdo a su posición cromosómica (indicando comienzo y final). La información de LD se obtuvo a partir del panel europeo de 1000 Genomas (hg19/1000 Genomes Nov 2014 EUR), y se usó para la construcción de matrices de correlación SNP-SNP y la obtención de los valores de r^2. El resto de parámetros opcionales de esta herramienta se usaron por defecto.

4.2.4 Análisis de redes

El listado de genes asociados ($p < 2.26 \times 10^{-6}$) se amplió incluyendo los principales interactores funcionales de dichos genes que se identificaron con la herramienta FunCoup v.4.0 (http://funcoup.sbc.su.se/search/).

FunCoup construye redes funcionales a partir de un *set* inicial de genes y diferencia 5 tipos posibles de conexiones funcionales entre ellos: interacción proteína-proteína, co-participación compleja, co-participación en el mismo proceso metabólico, co-participación en la misma ruta de señalización, y *operon* común. Para ello, su base de datos se apoya en 10 tipos de evidencias que prueban la existencia o no de conexiones funcionales entre parejas de genes: interacción física de productos proteicos, co-expresión de mRNA, co-expresión de productos proteicos, perfil similar de interacción genética, factores de transcripción comunes, co-localización en componentes celulares, interacción de dominios, similitud del perfil filogenético, regulación

por miRNA comunes y datos de espectrometría de masas cuantitativa (cantidad total de producto proteico en los diferentes tejidos). La información proporcionada por estos 10 niveles de evidencia se integra para obtener un *score* de confianza con rango entre 0 y 1 que refleja la probabilidad de conexión funcional entre dos genes.

Para la construcción de la red génica se consideraron inicialmente solo 6 de los 8 genes asociados, debido a que dos de ellos fueron eliminados por falta de información en FunCoup (Tabla 25).

Análisis Pascal	Genes principales FunCoup	Interactores	Tipo de red
XRN2	*XRN2*	*DDX21,ILF2,CDKN2AIP,EXOSC8,NOC3L,ACIN1,ILF3,GNL3,KNRNPF,ADAR,LYAR,SNW1,HNRNPK,HNRNPH1,PARN,SUMO2,DHX15,NOP2,UPF1,ALYREF,RBM39,SYNCRIP,PTBP1,C14orf166,NONO,HNRNPUL1*	PPI Co-participación compleja
NKX2-4	*NKX2-2*	*OLIG2*	PPI
KIZ	*KCNN*	*CALM1,CALM2,CALM3*	PPI
KCNN2			
NKX2-2			
CRHR1-IT1			
C8orf74			
LOC644172			

Tabla 25. Genes asociados en el GBA realizado con PASCAL y sus principales interactores identificados por FunCoup. Algunos de los genes asociados no fueron detectados por FunCoup (señalados en negrita). Los genes para los cuales se encontraron 1 o más interactores se denominan genes principales y se muestran en la segunda columna. La última columna indica el tipo de conexión funcional entre los genes que se usó para construir la red. Los genes asociados y los interactores que no fueron detectados por GENE2FUNC (explicado en el siguiente apartado) están subrayados.

La construcción de la red se realizó de acuerdo a 3 parámetros de expansión: el umbral de confianza, el número de pasos para expandir la red y el número de interactores. En primer lugar, solo se incluyeron en la red parejas de genes con un *score* de confianza ≥ a 0.8. En segundo

lugar, la expansión de la red se limitó a un solo paso. Es decir, buscando solo genes con alta conexión funcional con el *set* de genes inicial. Por último, se limitó el número de interactores detectado por FunCoup a 30.

Respecto al tipo de algoritmo usado para expandir la red, se permitió la búsqueda de interactores para cualquiera de los genes incluido en el *set* inicial de manera independiente, sin priorizar la búsqueda en interactores que presentaran una elevada conexión funcional con más de un gen.

La representación gráfica de la red génica resultante se realizó mostrando los genes principales y sus interactores como nodos y las conexiones funcionales entre ellos como aristas. La gráfica muestra solo los principales interactores y el tamaño del nodo refleja la importancia del gen en la red.

Por último, se realizó un análisis de enriquecimiento de funciones biológicas (rutas KEEG, términos GO para funciones biológicas y términos GO para funciones moleculares) para todos los genes incluidos en la red obteniendo los p-valores correspondientes. La información de este análisis se visualizó también en el gráfico construido por FunCoup de tal manera que los genes partícipes en el proceso biológico seleccionado se representaron en negro.

4.2.5 Anotación funcional

GENE2FUNC, una función perteneciente a la herramienta FUMA (*Functional Mapping and Annotation of Genome-Wide Association Studies*) (http://fuma.ctglab.nl/), se utilizó para anotar funcionalmente los genes asociados y sus interactores. Para ello, se usó como *input* un listado total de 36 genes (Tabla 25).

GENE2FUNC realiza múltiples análisis, pero solo dos de ellos se han considerado en este estudio: visualización del perfil de expresión génica de los genes asociados y sus interactores mediante *heatmaps* y un análisis de expresión diferencial de *set* de genes. En ambos casos se emplearon los datos de expresión de *GTEx v7* (incluye datos de expresión de 53 tejidos diferentes) y de *BrainSpan* (incluye datos de expresión en diferentes periodos del desarrollo). En el caso de *GTEx v7* la medida de expresión fue el tránscrito por millón o TPM (de sus siglas en inglés, *Transcripts Per Million*) y en el caso de *BrainSpan* fueron

lecturas por kilobase por millón o RPKM (de sus siglas en inglés, *Read Per Kilobase per Million*).

El *heatmap* es la visualización natural de las matrices de expresión génica. Los datos en el *heatmap* se visualizan en forma de cuadrícula. En ella, cada fila representa un gen y cada columna su condición de expresión, de manera que los colores de cada celda emulan los colores que identifican el nivel de transcripción en los *microarrays.*

Para la construcción de los *heatmaps*, se utilizó el valor promedio de expresión normalizada por categoría (media cero entre muestras). Esta visualización permite comparar la expresión de un gen en diferentes categorías (en el caso de *GTEx*, tejidos, y en el caso de *BrainSpan*, periodos del desarrollo). Así, las celdas rojas muestran una mayor expresión relativa de un gen en esa categoría en comparación con el resto, y el color azul refleja lo contrario.

Para el análisis de expresión diferencial (DEG) los *sets* de genes diferencialmente expresados se calculan previamente realizado un test *t-student* con dos colas para cada gen y tejido frente al resto de categorías (en el caso de *GTEx*, tejidos, y en el caso de *BrainSpan*, periodos del desarrollo). Aquellos genes en los que se obtiene un valor de $p < 0.05$ tras corregir por Bonferroni y un cambio de expresión absoluto de ≥ 0.58 (*log fold change*) se definen como genes diferencialmente expresados. Además, también se considera la sobrexpresión o infraexpresión de los *sets* de genes para cada categoría en función del signo del t-student.

Los 36 genes de nuestro estudio fueron contrastados para cada uno de los *sets* diferencialmente expresados usando un test de probabilidad hipergeométrico. Así, se muestran en rojo aquellos tejidos (*GTEx*) o periodos del neurodesarrollo (*BrainSpan*) en los que existe un enriquecimiento significativo de determinados *sets* de genes (p-valor corregido por Bonferroni ≤ 0.05).

4.2.6 Metaanálisis de expresión génica diferencial en estudios de TEA

dbMDEGA se ha utilizado para evaluar la expresión génica diferencial en tejido cerebral entre individuos control e individuos con TEA para los 36 genes incluidos en la Tabla 25[174].

dbMDEGA utiliza los resultados de un metaanálisis del perfil de expresión de 17741 genes en tejido cerebral humano para realizar un análisis de expresión diferencial para cada de estos genes en individuos con TEA. Los datos de expresión génica proceden de 3 estudios de TEA, cuyos resultados se han en depositado en la base de datos GEO (*Gene Expression Omnibus*): GSE28475[175], GSE28521[138], y GSE38322[176].

Para cada uno de los genes asociados y sus interactores se obtuvieron los siguientes valores: p-valor, FDR y medida de heterogeneidad entre los diferentes datos de expresión (I^2).

4.3 RESULTADOS

4.3.1 Análisis basados en genes (GBA)

El GBA llevado a cabo con PASCAL reveló la asociación de 8 *loci* con TEA (p-valor < 2.26×10^{-6}) (Tabla 26).

NKX2-2 y *NKX2-4* (ambos localizados en el cromosoma 20) mostraron asociación en comparación con los resultados obtenidos al aplicar el algoritmo MAGMA en los mismos datos. Cercano a ambos genes, se encuentra el SNP índice rs910805 (p-valor = 2.04×10^{-9}) (Tabla 27). El *plot* de asociación de la región, en torno a dicho SNP, mostró 3 niveles diferentes de r^2. Así pues, PASCAL logró detectar asociación para ambos genes (*NKX2-2* y *NKX2-4*), aunque se localizan lejos de rs910805. Sin embargo, SNPs de ambos *loci* se encuentran en moderado LD con dicho SNP índice (Figura 19).

PASCAL también detectó asociación de *CRHR1-IT1*, *LOC644172* (ambos localizados en el cromosoma 17) y *C8orf74* (en el cromosoma 8). Cabe destacar, que *CRHR1-IT1* solapa con una región intrónica de *CRHR1* (que había sido previamente identificada por MAGMA) (Figura 20).

Gen	Cromosoma	Posición de inicio	Posición final	Número de SNPs	p-valor
XRN2	20	21283941	21370463	271	$3,53 \times 10^{-9}$
NKX2-4	20	21376004	21378047	143	$9,51 \times 10^{-9}$
PLK1S1	20	21106623	21227258	287	$4,69 \times 10^{-8}$
KCNN2	5	113698015	113832197	540	$3,89 \times 10^{-7}$
NKX2-2	20	21491659	21494664	166	$7,78 \times 10^{-7}$
CRHR1-IT1	17	43716340	43723595	26	$1,69 \times 10^{-6}$
C8orf74	8	10530146	10558103	314	$1,78 \times 10^{-6}$
LOC644172	17	43677490	43679748	24	$2,15 \times 10^{-6}$
LRRC37A	17	44372496	44415160	48	$2,64 \times 10^{-6}$
ARL17A	17	44363861	44657088	64	$2,82 \times 10^{-6}$
KANSL1-AS1	17	44270938	44274089	46	$2,84 \times 10^{-6}$
KANSL1	17	44107281	44302740	153	$3,07 \times 10^{-6}$
MAPT-IT1	17	43973148	43976164	50	$3,68 \times 10^{-6}$
SOX7	8	10581277	10697299	752	$4,84 \times 10^{-6}$
MAPT	17	43971747	44105699	100	$4,86 \times 10^{-6}$
CRHR1	17	43697709	43913194	275	$5,45 \times 10^{-6}$
MAPT-AS1	17	43920721	43972879	224	$6,14 \times 10^{-6}$
STH	17	44076615	44077060	37	$6,27 \times 10^{-6}$
SPPL2C	17	43922255	43924438	193	$6,57 \times 10^{-6}$
PINX1	8	10622883	10697299	684	$7,78 \times 10^{-6}$

Tabla 26. 20 primeros genes que resultan asociados (ordenador por su p-valor) mediande PASCAL, al usar como *input* los *summary statistics* correspondientes al metaanálisis de TEA. Los genes señalados en negrita son aquellos que alcanzaron el umbral de significación estadística tras la corrección por Bonferroni (p-valor < 2.26×10^{-6}). Las columnas muestran gen, cromosoma, así como posición inicial y posición final en pb, el número de SNPs que PASCAL incluyó en el análisis para cada gen, y el p-valor obtenido por PASCAL.

Gen	Cromosoma	Posición inicio	Posición final	p-valor PASCAL /MAGMA	SNP índice
Genes asociados tras aplicar el algoritmo PASCAL					
XRN2	20	21283941	21370463	$3{,}53 \times 10^{-9}$	rs 910805
NKX2-4	20	21376004	21378047	$9{,}51 \times 10^{-9}$	rs 910805
PLK1S1	20	21106623	21227258	$4{,}69 \times 10^{-8}$	rs 910805
KCNN2	5	113698015	113832197	$3{,}89 \times 10^{-7}$	rs 13188074
NKX2-2	20	21491659	21494664	$7{,}78 \times 10^{-7}$	rs 910805
CRHR1-IT1	17	43716340	43723595	$1{,}69 \times 10^{-6}$	rs 142920272
C8orf74	8	10530146	10558103	$1{,}78 \times 10^{-6}$	rs 10099100
LOC644172	17	43677490	43679748	$2{,}15 \times 10^{-6}$	rs 142920272
Genes asociados tras aplicar el algoritmo MAGMA (Grove *et al*. 2019)					
XRN2	20	21283942	21370463	$9{,}69 \times 10^{-10}$	rs 910805
KCNN2	5	113698016	113832197	1.02×10^{-9}	rs 13188074
PLK1S1	20	21106624	21227260	5.17×10^{-9}	rs 910805
MACROD2	20	13976146	16033842	1.40×10^{-7}	rs 71190156
WNT3	17	44841686	44896082	4.03×10^{-7}	rs 142920272
MAPT	17	43971748	44105700	5.01×10^{-7}	rs 142920272
MFHAS1	8	8641999	8751131	5.58×10^{-7}	rs 11249905
XKR6	8	10753654	11058875	8.01×10^{-7}	rs 10099100
MSRA	8	9911830	10286401	9.15×10^{-6}	rs 10099100
CRHR1	17	43697710	43913194	1.07×10^{-6}	rs 142920272
SOX7	8	10581278	10588022	1.24×10^{-6}	rs 10099100
NTM	11	131240371	132206716	1.32×10^{-6}	rs 549507
MMP12	11	102733464	102745764	2.28×10^{-6}	rs 102751102
BLK	8	11351521	11422108	2.45×10^{-6}	rs 2736342

Tabla 27. Genes asociados identificados por PASCAL y MAGMA. Se muestran los genes asociados en este estudio y en el trabajo de Grove *et al*., 2019[48]. Las columnas muestran cromosoma, posición de inicio y final para cada gen, p-valor obtenido por PASCAL o MAGMA y el SNP índice correspondiente a cada gen.

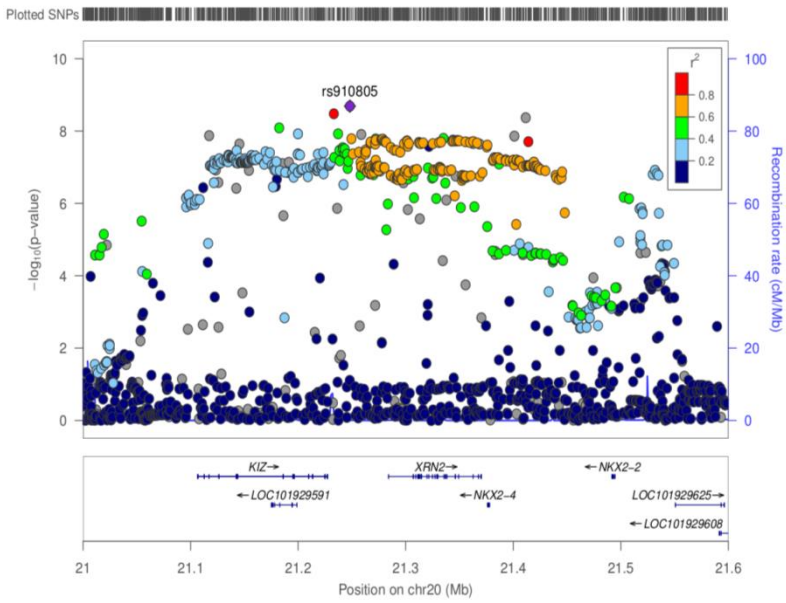

Figura 19. *Plot* de asociación de la región donde se localizan *NKX2-2* y *NXX2-4* realizado con LocusZoom

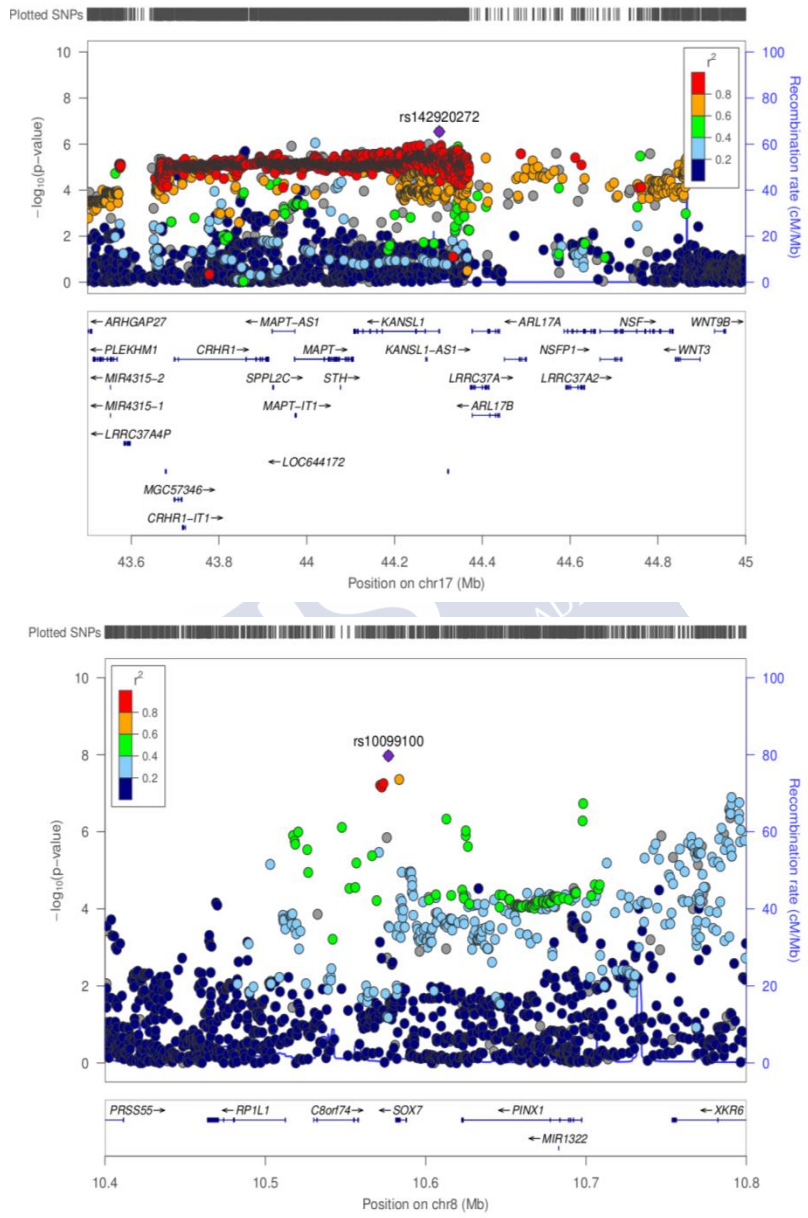

Figura 20. ***Plots*** **de asociación de las regiones donde se localizan *CRHR1-IT1* y *LOC644172* (cromosoma 17) y *C8orf74* (cromosoma 8).**

Aparte de estos hallazgos, el resto de *loci* asociados mediante PASCAL ya habían sido reportados previamente cuando el GBA se realizó con MAGMA (Tabla 27).

4.3.2 Análisis de redes

FunCoup identificó interactores para los *loci* asociados a TEA en el GBA realizado con PASCAL. La red génica se construyó utilizando un total de 36 genes (6 genes asociados y 30 interactores) y se detectaron 120 conexiones funcionales totales entre ellos (Tabla 25).

El análisis de enriquecimiento para rutas KEEG y términos GO llevado a cabo con los genes que conformaron la red, mostró un enriquecimiento para diferentes procesos biológicos como espliceosoma (q-valor = 3×10^{-4}), transporte del ARN (q-valor = 1.09×10^{-3}) y unión a ácidos nucleicos (q-valor = 1.09×10^{-14}) (Tabla 28 y Figura 21).

Términos enriquecidos		Genes	q-valor
Ruta metabólica KEEG	Spliceosome	*DH15, HNRNPK, ACIN1, ALYREF, SNW1*	3×10^{-4}
	RNA transport	*SUMO2, ACIN1, ALYREF, UPF1*	2.8×10^{-3}
Función molecular, térmico GO	Nucleic acid binding	*DDX21, ILF2, CDKN2AIP, EXOSC8, NOC3L, ACIN1, ILF3, GNL3, KNRNPF, ADAR, LYAR, SNW1, HNRNPK, HNRNPH1, PARN, SUMO2, DHX15, NOP2, UPF1, ALYREF, RBM39, SYNCRIP, PTBP1, C14orf166, NONO ,HNRNPUL1, XRN2, NKX2-4, NKX2-2*	1.09×10^{-14}
	Heterocyclid compound binding	*DDX21,ILF2,CDKN2AIP,EXOSC8 ,NOC3L,ACIN1,ILF3,GNL3,KNR NPF,ADAR,LYAR,SNW1,HNRNP K,HNRNPH1,PARN,SUMO2,DHX 15,NOP2,UPF1,ALYREF,RBM39 ,SYNCRIP,PTBP1,C14orf166,N ONO,HNRNPUL1,XRN2,NKX2-4,NKX2-2*	3.88×10^{-10}
	Organic cyclid compound binding	*DDX21, ILF2, CDKN2AIP, EXOSC8, NOC3L, ACIN1, ILF3, GNL3, KNRNPF, ADAR, LYAR,*	3.88×10^{-10}

		SNW1, HNRNPK, HNRNPH1, PARN, SUMO2, DHX15, NOP2, UPF1, ALYREF, RBM39, SYNCRIP,PTBP1,C14orf166,NO NO, HNRNPUL1, XRN2, NKX2-4, NKX2-2	
	Protein binding	*DDX21, ILF2, CDKN2AIP, EXOSC8, ACIN1, ILF3, GNL3, KNRNPF, ADAR, LYAR, SNW1, HNRNPK, HNRNPH1, PARN, SUMO2, DHX15, NOP2, UPF1, ALYREF, RBM39, SYNCRIP, PTBP1, C14orf166, NONO, HNRNPUL1, XRN2, KCNN2, NKX2-2, KIZ, C8orf74*	9.94×10^{-4}
	Binding	*DDX21,ILF2,CDKN2AIP, EXOSC8, ACIN1,ILF3,GNL3, KNRNPF,ADAR,LYAR, SNW1,HNRNPK, HNRNPH1,PARN,SUMO2, DHX15,NOP2, UPF1, ALYREF, RBM39,SYNCRIP,PTBP1,C14orf 166,NONO, HNRNPUL1, XRN2,KCNN2, NKX2-2,KIZ, C8orf74, NKX2-4, NOC3L*	2.81×10^{-2}
	Chromatin binding	*NKX2-2, NONO, NOC3L, UPF1,*	1.03×10^{-1}
	Transcription factor binding	*NKX2-2, HNRNPF, SUMO2, SNW1*	1.03×10^{-1}
	Enzyme binding	*KIZ, SUMO2,C14orf166, PARN, ACIN1, HNRN,UL1, SNW1*	2.54×10^{-1}
	Hydrolase activity, acting on acid anhydrides	*DHX15, DDX21,ACIN1,UPF1*	3.34×10^{-1}
	Identical protein binding	*KCNN2, NONO, EXOSC8, C14orf166,OLIG2*	4.21×10^{-1}
	hydrolase activity	*XRN2, DHX15,ADAR, DDX21,PARN,ACIN1,UPF1*	6.03×10^{-1}

Tabla 28. Enriquecimiento de términos GO y rutas KEEG para los genes asociados identificados por PASCAL y sus interactores, de acuerdo a FunCoup. Se muestran en inglés los términos con p-valores más significativos.

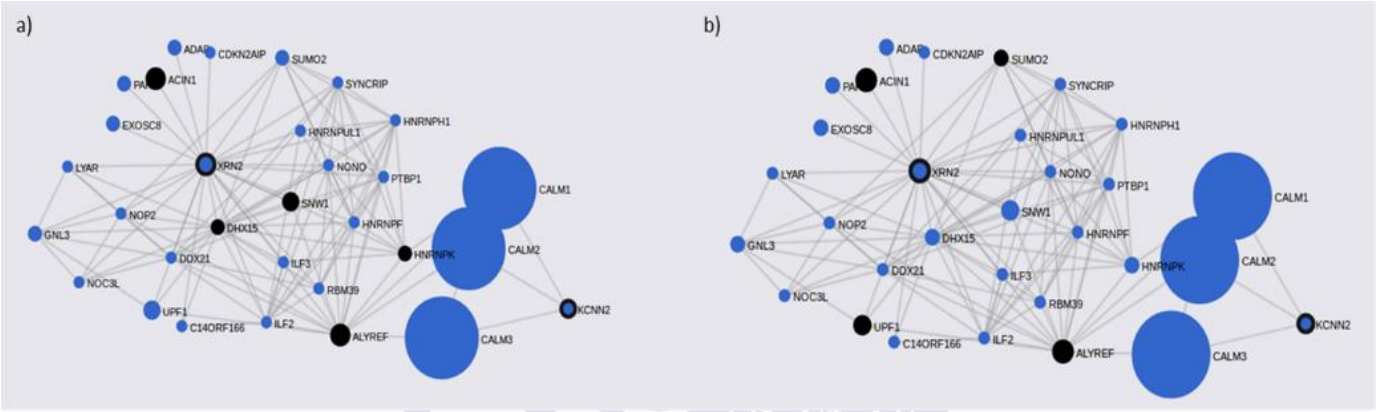

Figura 21. Visualización de la red génica construida con los genes asociados identificados por PASCAL y sus interactores de acuerdo a FunCoup. El tamaño de cada nodo revela la importancia de cada gen en la red, mientras que los nodos que participan en cada ruta KEEG están señalados en negro: a) Spliceosoma; b) mRNA surveillance

4.3.3 Anotación funcional

4.3.3.1 *Heatmaps* de expresión génica y análisis de expresión diferencial (DEG)

El *heatmap* de expresión de los genes asociados identificados por PASCAL y sus interactores, que se construyó con los datos de *GTEx v7*, reveló resultados interesantes:

NKX2-2, *OLIG-2* y *KCNN2* mostraron niveles de expresión elevados en tejido cerebral en comparación con otros tejidos. Otro *set* de genes, que incluyó a *XRN2* y sus interactores, mostró una tendencia opuesta, siendo el nivel de expresión relativo de estos genes más bajo en tejido cerebral que en los restantes tejidos.

El *heatmap* basado en los datos de RNA-seq de *BrainSpan* mostró, que la mayoría de genes infraexpresados en tejido cerebral, mostraron niveles elevados de expresión relativa en etapas tempranas del desarrollo (periodo prenatal temprano, temprano-medio y tardío-medio). *NKX2-2* y *OLIG-2*, en cambio, mostraron niveles de expresión más bajos en periodos prenatales con respecto a los niveles de expresión detectados en el cerebro adulto (Figura 22).

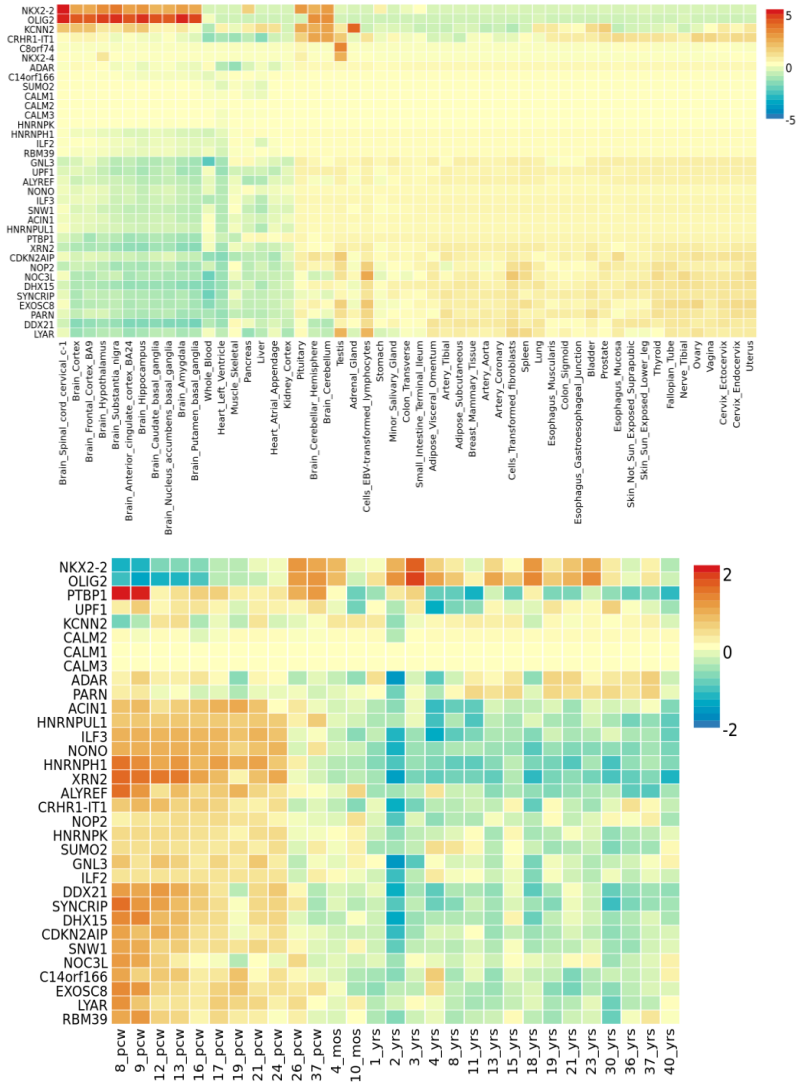

Figura 22. *Heatmaps* de expresión para los genes asociados en el GBA de PASCAL y sus interactores. Se utilizaron los datos de GTEx v7 (53 tejidos) (Figura superior) y los datos de *BrainSpan* (29 periodos del desarrollo) (Figura inferior). Los genes y tejidos están ordenados por clústeres en el caso del *heatmap* de GTEx. En el caso del *heatmap* de *BrainSpan* los genes están ordenados por clústeres de expresión y los periodos del desarrollo están ordenados de manera cronológica.

Estos resultados muestran una clara división, en términos de expresión, de aquellos *sets* de genes expresados en etapas tempranas del desarrollo (semanas posteriores a la concepción) en comparación con aquellos genes cuya expresión es superior durante la edad adulta.

El análisis de expresión diferencial de *sets* de genes usando los datos de *GTEx v7* mostró *sets* de genes diferencialmente expresados en tejido cerebral. El mismo análisis llevado a cabo con datos de *BrainSpan* mostró la existencia de dos *sets* de genes: uno diferencialmente expresado en periodo prenatal (sobreexpresión) y otro diferencialmente expresado en edad adulta (infraexpresión). En particular, la sobreexpresión tiene lugar en las semanas 8, 9 y 13 posteriores a la concepción y la infraexpresión tiene lugar a lo largo de varios periodos postnatales (Figura 23).

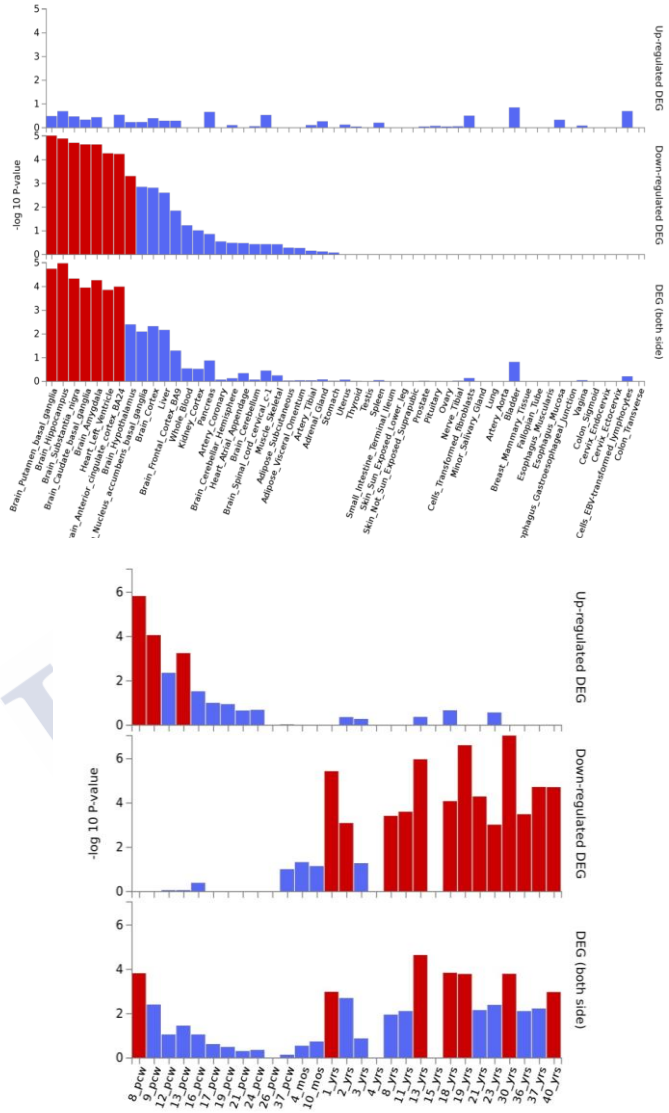

Figura 23. *Plots* representando el análisis de expresión diferencial de *sets* de genes construidos a partir de los datos de *GTEx v7* (53 tejidos) (Figura superior) y los datos de *BrainSpan* (29 periodos del desarrollo) (Figura inferior). Los *sets* de genes diferencialmente expresados y significativamente enriquecidos para el listado de genes de este análisis (p-valor corregido por Bonferroni < 0.05) se muestran en rojo.

4.3.4 Metaanálisis de expresión génica diferencial en estudios de TEA.

Algunos de los genes asociados identificados por PASCAL mostraron una expresión génica diferencial en tejido cerebral de individuos con TEA en comparación con el tejido cerebral de individuos control (*XRN2*, p-valor = 0.02; *KCNN2*, p-valor = 0.01; *C8orf74*, p-valor = 0.03). Además, algunos de los interactores de *XRN2* también mostraron asociación en el metaanálisis de expresión diferencial llevado a cabo con la herramienta dbDMEGA (*XRN2, CDKN2AIP, ILF3, GNL3 ADAR, LYAR, SNW1, RBM39, SYNCRIP,* y *PTBP1*) (Tabla 29).

Gen	p-valor PASCAL	p-valor	FDR	Heterogeneidad
XRN2	$3{,}52863316 \times 10^{-9}$	0.02	0.12	0%
NKX2-4	$9{,}5086814 \times 10^{-}$	NA	NA	NA
PLK1S1	$4{,}68346333 \times 10^{-8}$	NA	NA	NA
KCNN2	$3{,}88782618 \times 10^{-7}$	0.01	0.11	0%
NKX2-2	$7{,}7773789 \times 10^{-7}$	0.23	0.37	0%
CRHR1-IT1	$1{,}6858736 \times 10^{-6}$	NA	NA	NA
C8orf74	$1{,}77348539 \times 10^{-6}$	0.03	0.16	43%
LOC644172	$2{,}15004243 \times 10^{-6}$	NA	NA	NA
DDX21	$8{,}86663103 \times 10^{-1}$	0.3	0.41	69%
ILF2	$5{,}52756293 \times 10^{-1}$	0.38	0.45	28%
CDKN2AIP	$8{,}21802658 \times 10^{-1}$	0.04	0.17	0%
EXOSC8	$9{,}41352461 \times 10^{-2}$	0.1	0.26	64%
NOC3L	$3{,}66793335 \times 10^{-1}$	0.35	0.44	0%
ACIN1	$9{,}65129025 \times 10^{-1}$	0.15	0.3	48%
ILF3	$1{,}0116456 \times 10^{-1}$	0.01	0.1	66%
GNL3	$2{,}39022069 \times 10^{-3}$	0.01	0.07	67%
KNRNPF	NA	NA	NA	NA
ADAR	$1{,}70702835 \times 10^{-2}$	0.01	0.1	72%
LYAR	$2{,}30069599 \times 10^{-1}$	0.03	0.15	0%

SNW1	1,0892706 x 10^{-1}	0.04	0.18	74%
HNRNPK	3,13685174 x 10^{-1}	NA	NA	NA
HNRNPH1	6,07006486 x 10^{-1}	NA	NA	NA
PARN	9,01671731 x 10^{-1}	0	0.03	79%
SUMO2	8,85347296 x 10^{-1}	0.27	0.39	72%
DHX15	2,93250866 x 10^{-1}	0.12	0.38	0%
NOP2	7,85747107 x 10^{-1}	NA	NA	NA
UPF1	2,08668839 x 10^{-1}	0.12	0.27	14%
ALYREF	2,00036414 x 10^{-1}	NA	NA	NA
RBM39	2,90163509 x 10^{-1}	0.06	0.2	18%
SYNCRIP	2,62087002 x 10^{-1}	0.01	0.08	0%
PTBP1	9,27312631 x 10^{-2}	0.02	0.13	0%
C14orf166	2,91369259 x 10^{-1}	0.36	0.44	58%
NONO	NA	0.18	0.33	32%
HNRNPUL1	4,2805356 x 10^{-1}	NA	NA	NA
CALM1	3,30876236 x 10^{-1}	0.4	0.46	27%
CALM2	9,81305815 x 10^{-1}	0.36	0.44	0%
CALM3	7,91761032 x 10^{-1}	0.5	0.5	0%
OLIG2	4,52957535 x 10^{-1}	0.4	0.46	0%

Tabla 29. Genes asociados identificados por PASCAL e interactores y resultados de dbMDEGA. Se muestran los p-valores calculados por PASCAL y los p-valores calculados por dbMDEGA para cada gen. Se muestran también los valores de heterogeneidad entre estudios de expresión (NA = el gen no se encuentra en la base de datos de dbMDEGA).

4.4 DISCUSIÓN

El último metaanálisis de GWAS de TEA identificó asociación de diferentes SNPs en los cromosomas 20 (rs910805), 5 (rs13188074), 17 (rs142920272), y 8 (rs10099100) entre otros. En este mismo estudio también se llevó a cabo un GBA empleando el algoritmo MAGMA, que identificó 15 *loci* que alcanzaron el umbral de significación estadística. Como era de esperar, muchos de estos genes se localizaban muy cerca de los SNPs reportados como asociados en el GWAS, como *XRN2* (rs910805) *KCNN2* (rs13188074) y *KIZ* (o *PLK1S1*) (rs910805)[48]. Además, el GBA llevado a cabo por PASCAL ha detectado nuevos

genes en comparación con aquellos reportados por MAGMA: *NKX2-2*, *NKX2-4*, *CRHR1-IT1*, *C8orf74* y *LOC644172*. Sin embargo, estos hallazgos no deben ser considerados como novedosos del todo. Así pues, los SNPs índice localizados en los cromosomas 20 y 8 también mostraron previamente asociación con TEA al realizarse otros análisis. Así, cuando se realizó el análisis MTAG (de sus siglas en inglés *multi-trait analysis of GWAS*) considerando otros fenotipos genéticamente correlacionados con TEA (como esquizofrenia o nivel educativo), se detectó asociación de dichos *loci*, lo cual sirvió para resaltar la importancia de la región genética y de los genes localizados en ella (*KIZ*, *XRN2*, *NKX2-2*, y *NKX2-4*)[48,177]. Este análisis explica por qué *NKX2-2* y *NKX2-4* se consideraron *loci* probablemente asociados. Sin embargo, MAGMA no los identificó como *loci* estadísticamente significativos[48]. PASCAL, por el contrario, sí fue capaz de detectar estas asociaciones gracias a la aproximación estadística que emplea. Por otro lado, las posiciones cromosómicas de *CRHR1-IT1* y de *CRHR1* solapan parcialmente, aunque codifican para distintos tránscritos. *CRHR1* mostró asociación debido a que está incluido en el listado inicial de genes de referencia que utiliza PASCAL para mapear los SNPs. Por el contrario, el *input* de MAGMA no incluía este gen. *C8orf74* es un hallazgo controvertido, pues, aunque hay SNPs significativos que solapan con el gen, éstos también se encuentran muy cercanos al gen *XKR6*, que sí mostró asociación en el trabajo previo de Grove *et al.*,[48]. Aun así, PASCAL ha demostrado ser una herramienta útil para detectar nuevos genes asociados que se localizan cerca de aquellos reportados previamente por MAGMA. Ambas herramientas, MAGMA y PASCAL, funcionan de manera similar: 1) emplean como *input* del análisis los *summary statistics* del metaanálisis de GWAS en lugar de genotipos, 2) el p-valor para cada gen se calcula teniendo en cuenta todos los p-valores de cada uno de los SNPs localizados a lo largo del gen, 3) la corrección por LD se realiza a partir de información externa procedente del panel europeo 1000 Genomas[52,53]. Sin embargo, la construcción de la matriz de LD (correlación entre SNPs) es ligeramente diferente en ambos métodos. Esto explicaría las pequeñas diferencias encontradas con respecto a los genes que cada GBA señala como asociados. Los p-valores obtenidos para cada gen con PASCAL,

fueron, en general, menos significativos que los obtenidos con MAGMA. Además, el número de genes estadísticamente significativos tras la corrección por Bonferroni fue también menor cuando se usó PASCAL. Todos estos resultados sugieren que PASCAL podría ser un método de GBA más conservador que MAGMA. En base a lo expuesto, en el presente estudio se propone que PASCAL se utilize como una herramienta GBA complementaria a otras existentes, pues ha resultado eficaz a la hora de reportar nuevas asociaciones de genes localizados en la misma región de LD en la cual se localizan otros genes ya previamente asociados por MAGMA.

A continuación, se profundizará en algunos aspectos biológicos de los siguientes *loci* que resultaron asociados: *NKX2-2*, *NKX2-4*, *CRHR1-IT1*, *LOC644172*, y *C8orf74*.

NKX2-2 y *NKX2-4* son miembros de la familia de factores de transcripción *homeobox*. *NKX2-2* codifica para un factor de transcripción implicado en la morfogénesis del sistema nervioso central y su papel es esencial durante la diferenciación, en fases muy tempranas del desarrollo, de poblaciones neuronales que se localizan en el romboencéfalo y la médula espinal[178]. *NKX2-4* (*homeobox protein nkx-2.4*) posee también un papel clave durante el desarrollo del cerebro dado que su inhibición en el cerebro anterior promueve la proliferación de progenitores neurales al mismo tiempo que inhibe su diferenciación, resultando todo ello en una neurogénesis deficiente[179]. Los *heatmaps* de expresión construidos a partir de los datos de *GTEx* y *BrainSpan*, revelaron dos clústeres de expresión en los cuales está incluido el gen *NKX2-2*. Los genes del primer clúster mostraron una infraexpresión en estados prenatales y una sobreexpresión en el periodo postnatal, mientras que los genes del segundo clúster mostraron una tendencia opuesta. *NKX2-2* y su interactor *OLIG-2* (*oligodendrocyte transcription factor 2*) están incluidos en el primer clúster. Este hecho sugiere la existencia de mecanismos de regulación génica diferentes para *NKX2-2* en el cerebro en desarrollo y el cerebro adulto. Así pues, es sabido que *NKX2-2* se expresa inicialmente en células precursoras de oligodendrocitos y luego su expresión disminuye. Sin embargo, se ha demostrado que su expresión vuelve a incrementar en etapas posteriores del neurodesarrollo, lo cual permite el mantenimiento de las estructuras

mielínicas[180]. Esto hecho es particularmente interesante dado que estudios previos han relacionado la expresión aberrante de genes en oligodendrocitos con la patogénesis de los TEA[181,182]. Además, *NKX2-2* y *OLIG-2* estaban enriquecidos en términos biológicos relacionados con la unión a proteínas, cromatina y factores de transcripción, todos ellos procesos biológicos subyacentes a la patogénesis de los TEA. Sin embargo, cuando se llevó a cabo el metaanálisis de expresión génica diferencial, debe considerarse que ninguno de estos genes mostró diferencias en el perfil de expresión en cerebros de individuos con TEA respecto a cerebros control. Estos resultados señalan la necesidad de realizar estudios funcionales *in vivo* con estos genes para poder caracterizar completamente sus funciones biológicas y determinar en detalle su posible implicación en la etiología de los TEA.

Un segundo clúster de expresión, de acuerdo al *heatmap* construido con datos de *BrainSpan*, incluyó a los genes *XRN2,* sus interactores y *CRHR1-IT1*. En primer lugar, debe señalarse que FunCoup calculó un número mayor de interactores para *XRN2* que para el resto de genes. Esto puede implicar un sesgo en los análisis posteriores que se llevaron a cabo, pero, al mismo tiempo, señala la importancia de este gen si se considera la extensa red de tipo PPI formada en torno a él. Además, *XRN2* y una gran parte de sus interactores (*CDKN2AIP, ILF3, GNL3 ADAR, LYAR, SNW1, RBM39, SYNCRIP*, y *PTBP1*) mostraron asociación en el metaanálisis de expresión. Esto demuestra la expresión diferencial de estos genes en el cerebro de individuos con TEA en comparación con cerebros control y, por tanto, su potencial papel en la fisiopatología de los TEA[138,176,183]. Además, el análisis de enriquecimiento en términos biológicos mostró que *XRN2* y sus interactores participan en funciones relacionadas con el espliceosoma, el transporte del ARN y la unión a ácidos nucleicos. Todos ellos, constituyen procesos biológicos esenciales en cualquier organismo y son especialmente relevantes en etapas tempranas del desarrollo. Se ha demostrado, además, que *XRN2* podría tener un papel en la finalización de la transcripción a través de la degradación del extremo 3´UTR[184]. *CRHR1-IT1* codifica para un *long intergenic non-protein coding RNA* que se ha asociado recientemente a susceptibilidad al comportamiento antisocial[185]. *CRHR1-IT1*, comparte parte de su secuencia con *CRHR1*,

el cual codifica para el receptor 1 de la hormona liberadora de corticotropina[48]. *CRHR1*, constituye el principal componente de la vía hipotalámica-pituitaria-adrenal y se ha demostrado su asociación, en repetidas ocasiones, con la respuesta psicopatológica ante el estrés[186]. Aunque la función de *CRHR1-IT1* permanece aún sin caracterizar, se cree que podría tener un papel en la regulación de la expresión de *CRHR1*. De esa manera, *CRHR1* y *CRHR1-IT1* podrían estar involucrados en la modulación del comportamiento y la cognición, ambas características importantes en los TEA[175,185,187,188]. Sin embargo, el metaanálisis de expresión diferencial entre tejido cerebral de individuos con TEA y tejido cerebral control no reveló ninguna asociación significativa para ninguno de los dos genes.

Por último, otros dos genes, *C8orf74* y *LOC644172*, mostraron asociación en el GBA realizado con PASCAL. Se debe señalar que *LOC644172*, localizado *upstream* con respecto a *CRHR1* y *CRHR1-IT1*, no fue reconocido ni por FunCoup ni por FUMA, al igual que *C8orf74*, que forma parte de un marco de lectura abierto o ORF (de sus siglas en inglés *Open Reading Frame*). Esto se debe a que ambos genes no están caracterizados funcionalmente y, por tanto, no se encuentran incluidos en bases de datos genéticas y/o funcionales. Sin embargo, el metaanálisis del perfil de expresión reveló que *C8orf74* mostró una expresión diferencial en cerebros de individuos con TEA en comparación con cerebros control, lo cual indica la necesidad de una mejor caracterización funcional de este gen para conocer cuáles son los mecanismos biológicos en los que está implicado.

4.5 CONCLUSIONES

Se ha llevado a cabo un GBA, mediante PASCAL, que ha usado como *input* los *summary statistics* del último metaánalisis de GWAS de TEA. Muchos de los *loci* que resultaron asociados habían sido previamente reportados por MAGMA. Sin embargo, PASCAL también ha identificado nuevos *loci* asociados en aquellas regiones de LD en las que se localizan muchos de los *loci* previamente asociados mediante MAGMA. Estos resultados sugieren que PASCAL se puede utilizar como una aproximación GBA complementaria a la hora de extraer

información adicional (nuevos genes asociados) a partir de los resultados de un GWAS o de un metaanálisis de los mismos.

La segunda parte del estudio se ha centrado fundamentalmente en la caracterización biológica de los *loci* asociados. Así pues, se llevó a cabo la construcción de una red génica y la anotación funcional de dichos *loci*, incluyendo la construcción de *heatmaps* de expresión génica y un análisis de expresión diferencial de *sets* de genes. Ambas aproximaciones han servido para caracterizar el contexto biológico de los genes asociados a TEA y seleccionar así aquellos genes candidatos más adecuados para futuros estudios funcionales.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

5 CAPÍTULO 2

5.1 OBJETIVO

En el presente estudio se han detectado mutaciones *de novo* (germinales y PZMs) en una cohorte española de TEA formada por 360 tríos (probando afecto y ambos progenitores sanos). Las mutaciones *de novo* detectadas en esta cohorte, se han combinado con un listado de mutaciones *de novo* previamente publicado por el ASC[116]. El objetivo principal de este trabajo ha sido explorar si las mutaciones germinales o PZMs tienden a acumularse en diferentes genes de riesgo en los TEA. Además, mediante herramientas bioinformáticas se ha determinado el impacto de estas mutaciones en diferentes jerarquías biológicas, desde el nivel gen, pasando por términos GO y tipos celulares y a través de los periodos del neurodesarrollo, hasta llegar al nivel de áreas cerebrales.

5.2 MÉTODOS

5.2.1 Sujetos de estudio

La extracción de ADN a partir de sangre periférica se realizó con el kit *Chemagic DNA Blood 100 Kit (PerkinElmer Inc, Massachusetts, USA)* siguiendo las indicaciones del fabricante. Los sujetos de Santiago (N = 136) fueron reclutados en el Complejo Hospitalario Universitario de Santiago de Compostela y en entidades gallegas que trabajan con individuos con TEA (ASPANAES, BATA, MENELA y ASPERGA). Los sujetos de Madrid (N = 224), fueron reclutados como parte del programa AMITEA en el servicio de Psiquiatría del Niño y del Adolescente del Hospital Gregorio Marañón. Solo se incluyeron individuos de 3 años o más. Todos los participantes habían sido diagnosticados previamente de TEA por un neurológo pediátrico o un psiquiatra de acuerdo a los criterios diagnósticos del DSM-IV o el DSM-5. Las pruebas diagnósticas ADOS (de sus siglas en inglés, *Autism Diagnostic Observation Schedule*) y ADI-R (de sus siglas en

inglés, *Autism Diagnostic Interview-Revised*) también se administraron en los casos en los que fue necesario.

El proyecto fue aprobado por el Comité Ético de Investigación Clínica de Galicia (Código 2012/098; Anexo 1). Los progenitores, o en su defecto, los tutores legales, fueron debidamente informados de la naturaleza y el objetivo del estudio, y se requirió la firma del consentimiento informado para su participación en él (Anexos 2 y 3).

5.2.2 Control de calidad de las muestras y detección de mutaciones *de novo*

5.2.2.1 Procesamiento de datos y anotación de variantes

La secuenciación del exoma completo de cada uno de los tríos que forma parte de la cohorte española (N = 360) fue realizada por el ASC (https://genome.emory.edu/ASC/)[104]. Este consorcio proporcionó un único archivo VCF con los datos de las variaciones en la secuencia exónica en crudo para todos los individuos de la cohorte. La herramienta *bcftools* se usó para obtener archivos individuales de cada uno de los sujetos de estudio. Estos archivos individuales contienen las variantes en regiones codificantes identificadas para cada individuo y anotadas con la herramienta SnpEff (*Genomic variant annotations and functional effect prediction toolbox*) versión 4.3T (http://snpeff.sourceforge.net/).

5.2.2.2 Control de calidad específico para las muestras

El grado de parentesco familiar en la cohorte española (360 tríos) se obtuvo calculando el número de errores mendelianos en cada trío. Para ello se usó la función "--mendel", disponible en la herramienta VCFtools (http://vcftools.sourceforge.net/). Aquellas muestras con una media de errores mendelianos con una desviación significativa de la media no se consideraron en análisis posteriores.

Para identificar discrepancias entre el sexo nominal y el sexo determinado genéticamente se empleó la opción "--sexcheck" de la herramienta PLINK que permite inferir el sexo de cada una de las muestras a partir de genotipos en los cromosomas X e Y.

Finalmente, para identificar muestras *outlier* se empleó el comando de PLINK "pseq i-stats". Las muestras en las cuales, alguno de los

siguientes parámetros: *NALT (*de sus siglas en inglés, *number of non-reference genotypes), NMIN (*de sus siglas en inglés, *number of genotypes with a minor allele*), *NHET (*de sus siglas en inglés, *number of heterozygous genotypes for individual), NVAR (*de sus siglas en inglés, *total number of called variants for indivdual)* y *RATE* (de sus siglas en inglés, *genotyping rate for individual)* se desvió más de 4 DE de la media, fueron eliminadas. Consecuentemente, el trío entero se eliminó si alguno de sus miembros era un *outlier* siguiendo los parámetros anteriormente descritos.

Las muestras de la cohorte que pasaron todos los controles de calidad (360 tríos) son las mismas que se han incluido en la publicación de Satterstrom *et al.*,[104].

5.2.2.3 Detección de mutaciones *de novo*

La detección de variantes *de novo* en la muestra española (360 tríos), definidas como aquellas variantes presentes en el probando y ausentes en los progenitores, se realizó empleando las opciones de filtrado descritas por Lim *et al.*,[116]. En este estudio, las variantes que fueron clasificadas como PZMs fueron resecuenciadas mediante tres tecnologías de secuenciación diferentes alcanzando un *ratio* de validación muy elevado (87%-97%).

Así pues, se definieron como mutaciones *de novo* aquellas variantes con genotipos 1/0 o 1/1 en el probando y 0/0 en los progenitores. Solo se consideraron aquellas variantes con GQ ≥ 20 (calidad del genotipo asignado) y una profundidad de lectura del alelo alternativo ≥ 7. Las variantes con uno o más alelos presentes en la base de datos ExAC (http://exac.broadinstitute.org/) también fueron eliminadas. Además, se eliminaron las variantes separadas por menos de 20 pb y aquellas con un RVIS > 75% (de sus siglas en inglés, *Residual Variation Intolerance Score*), obtenido de ExAC.

SnpEff fue empleado para clasificar las variantes exónicas de acuerdo a la predicción de su impacto en el tránscrito canónico: alto, moderado o bajo. Las variantes de bajo impacto incluyeron variantes sinónimas, las variantes de impacto moderado incluyeron variantes *missense*, y las variantes de alto impacto incluyeron variantes *nonsense*.

En este estudio solo se consideraron sustituciones de una sola base, excluyendo, por lo tanto, variantes *frameshift*.

Para clasificar las variantes *missense* se usaron dos predictores *in silico* de patogenicidad (CADD y SIFT). Se definieron como variantes probablemente patogénicas, aquellas variantes con un CADD score > 20 y aquellas variantes que SIFT clasificó como patogénicas.

Finalmente, las variantes *de novo* se clasificaron como germinales o PZM según su AAF (número de *reads* del alelo alternativo / (número de *reads* del alelo de referencia + número de *reads* del alelo alternativo)). De ese modo, las variantes con AAF ≥ 0.40 se categorizaron como germinales y las variantes con AAF < 0.40, como PZMs[116].

90 tríos de la cohorte española ya habían sido analizados previamente por Lim *et al.*,[116]. Las variantes *de novo* detectadas en estos tríos fueron usadas como controles positivos para comprobar que la detección de PZMs en el presente estudio era correcta. Así pues, se demostró que la mayoría de las variantes *de novo* eran detectadas y clasificadas correctamente como germinales o PZMs. (Tablas Sumplementarias 1 y 2 de Alonso-González *et al.*,[189]*).*

En gran parte de los análisis que se llevaron a cabo se usó un *dataset* llamado "cohorte combinada" (N = 2171) que incluye todas las variantes no sinónimas detectadas en los 360 tríos de la cohorte española y las variantes no sinónimas identificadas en individuos con TEA secuenciados por el ASC y previamente publicadas. Las variantes duplicadas en ambas cohortes fueron eliminadas. (Tabla Suplementaria 3 de Lim *et al.*,[116] y Tablas Suplementarias 1 y 3 de Alonso-González *et al.*,[189]*).*

Para algunos de los análisis también se creó una cohorte control que incluyó a 288 hermanos no afectos de individuos con TEA secuenciados por el ASC bajo las mismas condiciones que la cohorte española[116] (Tabla Suplementaria 4 de Alonso-González *et al.*,[189]).

5.2.3 Test Transmission and De novo Association (TADA-Denovo)

El *software* TADA-Denovo se utilizó para identificar y priorizar genes de riesgo en los TEA utilizando las mutaciones *de novo*

(germinales y PZMs) detectadas en la cohorte española (N = 360) y en la cohorte combinada (N = 2171). (Tabla Suplementaria 3 de Alonso-González *et al*.,[189]). El análisis realizado con TADA considera la carga mutacional de cada uno de los genes, así como el impacto funcional de las diferentes clases de mutaciones[102]. TADA se usó de manera independiente en los dos *sets* de genes en la cohorte española (genes con PZMs (N = 105); genes con mutaciones germinales (N = 181)) y en la cohorte combinada (genes con PZMs (N = 362); genes con mutaciones germinales (N = 1210)). (Tablas Suplementarias 5, 6, 7 y 8 de Alonso González *et al*.,[189]). Se utilizó una cohorte control (N = 288) para ajustar y estimar los parámetros requeridos por TADA[116]. (Tabla Suplementaria 4 de Alonso-González *et al.,*[189]). En el análisis se incluyeron dos tipos de mutaciones *de novo:* LoF y mutaciones *missense* probablemente patogénicas. Para ajustar el *ratio* mutacional para cada clase de mutación se usaron los *ratios* mutacionales por gen calculados por Samocha *et al*.,[190] y sobre ellos se emplearon las siguientes fórmulas para calibrarlos: para mutaciones LoF; (nonsense+splice) x (sin_{obs}/sin_{esp}) y para mutaciones *missense* probablemente patogénicas; missense x ($N_{Prob.patogénica}$ /$N_{missensetotal}$) x (sin_{obs}/sin_{esp}). Sin_{obs} es el número de mutaciones sinónimas *de novo* observado en la cohorte control (N = 119)[116] y sin_{esp} es el número de mutaciones sinónimas *de novo* esperado en la misma cohorte calculado a partir de la suma del *ratio* de mutaciónes *de novo* sinónimas por gen ($2*n*\mu$) (N = 79.04). $N_{Prob.patogénica}$ es el número de variantes *missense* probablemente patogénicas en la cohorte control (N = 212). $N_{missensetotal}$ es el número total de mutaciones *missense* en la cohorte control (N = 296). Para estimar el riesgo relativo (γ) de cada clase de mutación, se calculó la carga de mutaciones de cada clase en los casos de la cohorte española con respecto a los controles (LoF = 2.21; *missense* probablemente patogénicas = 1.36). A continuación, se aplicó la siguiente fórmula para calcular el riesgo relativo: $\gamma = 1+(\lambda-1) / \pi$, donde π, la fracción de genes riesgo, se ajustó a 0.05 (el parámetro por defecto). TADA utilizó los datos según los parámetros anteriormente descritos y los p-valores no corregidos para cada gen se calcularon obteniendo distribuciones nulas (N repeticiones = 10000). Para determinar el umbral adecuado que permite identificar a un gen como

significativo, TADA usa una aproximación bayesiana de FDR para controlar el *ratio* de falsos positivos, obtenindo q-valores para cada gen. El paquete de R, qqman, se usó para representar en *Manhattan plots* los p-valores (-log10) para cada gen en la cohorte combinada[191].

Los genes con FDR < 0.1 (genes con mutaciones germinales) y FDR < 0.3 (genes con PZMs) fueron clasificados de acuerdo al *score* propuesto por SFARI (https://gene.sfari.org/database/gene-scoring/). Además, también se consultó la base de datos OMIM para comprobar si alguno de esos genes se relacionaba con alguna enfermedad mendeliana.

5.2.4 Análisis de enriquecimiento en *sets* de genes de mutaciones germinales y PZMs

El análisis de enriquecimiento en *sets* de genes de mutaciones *de novo missense* y *nonsense* (germinales y PZMs) se llevó a cabo con la herramienta DNENRICH[192]. DNENRICH determina si existe un enriquecimiento de mutaciones *de novo* en grupos de genes definidos *a priori* teniendo en cuenta su tamaño, el contexto trinucleotídico y el efecto funcional predicho para cada tipo de mutación. Las mutaciones *de novo* que se incluyeron en el análisis de la cohorte española fueron germinales y PZMs (mutaciones germinales = 236; PZMs = 164) (Tablas Suplementarias 9 y 10 de Alonso-González *et al.*,[189]). Para el análisis de la cohorte combinada, en cambio, se creó un *subset* de PZMs con el objetivo de asegurar que estas mutaciones estén contribuyendo al fenotipo (PZMs = 676). (Tabla Suplementaria 11 de Alonso-González *et al.*,[189]). Para ese propósito, los individuos con mutaciones germinales en genes de riesgo (genes con *score* de SFARI 1 o 2) fueron eliminados del *dataset* de PZMs. Así pues, en el análisis de la cohorte combinada se usaron mutaciones germinales y el *subset* de PZMs (mutaciones germinales = 780; PZMs = 239). (Tablas Suplementarias 12 y 13 de Alonso-González *et al.*,[189]). El análisis también se realizó en la cohorte control usando ambos tipos de mutaciones *de novo* (mutaciones germinales = 780; PZMs = 239) (Tabla Suplementaria 13 de Alonso-González *et al.*,[189]).

Como *inputs* para el análisis se usaron los nombres oficiales de todos los genes del genoma y todos sus alias y la matriz de tamaños de

genes que proporciona DNENRICH, así como lo siguientes *sets* de genes: 1) genes diana de *FMRP* identificados por Darnel *et al.*,[193] y descargados de *Genebook* (http://zzz.bwh.harvard.edu/genebook/) (N = 788); 2) genes incluidos en la ontología GO:0006325, relacionada con la organización de la cromatina (N = 723) (http://www.geneontology.org/); 3) genes sinápticos (N = 903)[194]; 4) ortólogos humanos de genes esenciales en ratón (N = 2472)[195]; 5) genes diana de *CHD8* expresados en cerebro fetal humano (N = 2725)[196]; 6) lista de genes SFARI (N = 990) (https://gene.sfari.org/autdb/HG_Home); 7) genes intolerantes a mutaciones LoF (pLi > 0.9) (N = 3230)[197]; 8) genes diana de *RBFOX* (N = 587)[198] ; 9) genes diana de miR-137 (N = 428)[199]; 10) genes diana de *CELF-4* (N = 954)[200]; 11) genes con expresión monoalélica en neuronas en diferenciación (N = 802)[201]; 12) genes conocidos de DI (N = 1547)[202]; 13) *enhancers* intergénicos e intrónicos expresados en cerebro (N = 673)[203]; 14) posibles *enhancers* de genes con expresión en el telencéfalo (N = 79)[118]; y 15) genes diana de miR-138 (N = 255)[204]. Se obtuvieron los p-valores para cada *set* de genes llevando a cabo un millón de permutaciones.

5.2.5 Análisis de enriquecimiento de ontologías génicas

La herramienta *Enrichr* se utilizó para realizar un análisis de enriquecimiento de ontologías génicas que permite explorar si los genes con mutaciones germinales y los genes con PZMs están involucrados en diferentes funciones biológicas. Para ello, Enrichr obtiene p-valores corregidos por el método Benjamini-Hochberg p[Pbh] para cada término GO. Para este análisis se utilizaron los listados de mutaciones germinales *missense* y *nonsense* y el *subset* de PZMs de la cohorte combinada (genes con mutaciones germinales = 1972; genes con PZMs = 624) (Tabla Suplementaria 15 de Alonso-González *et al.*,[189]).

La herramienta REViGO (http://revigo.irb.hr/) se utilizó para visualizar los términos GO en *scatterplots*. Esta herramienta permite, mediante el *score* de semejanza semántica *SimRel*, prescindir de términos GO redundantes. Los 30 términos GO más enriquecidos en cada grupo de genes (germinal y PZM) se visualizaron usando una

modificación del *script* de R que proporciona la herramienta online REViGO.

Los 50 términos GO más enriquecidos en cada grupo de genes también se visualizaron en una red usando el *plugin* para la visualización de enriquecimiento funcional de *Cytoscape* (v.3.6.1), *Enrichement Map*[205]. Cada nodo de la red representa un *set* de genes pertenecientes a un término GO, siendo el tamaño del nodo proporcional al número de genes que participa en esa ontología (coeficiente de solapamiento). Así, cada pareja de nodos se consideró conectada si el coeficiente de solapamiento entre ambos era ≥ 0.7, siendo el grosor de la arista proporcional a este solapamiento. Al mismo tiempo el grosor del borde de cada nodo representa el p-valor correspondiente a cada término GO.

5.2.6 Análisis de enriquecimiento por tipo celular y análisis de expresión en regiones cerebrales a lo largo del desarrollo.

El paquete de R, EWCE (de sus siglas en inglés, *Expression Weighted Cell-type Enrichment)* (https://github.com/NathanSkene/EWCE) se empleó para realizar un análisis de expresión génica diferencial en los diferentes tipos de neuronas, de los genes con mutaciones germinales (N = 1972) y los genes con PZMs (N = 624) (Tabla Suplementaria 15 de Alonso-González *et al*.,[189]). EWCE permite identificar en qué tipos celulares neuronales se produce una sobreexpresión de genes con mutaciones germinales o PZMs.

Para este análisis se usaron los datos de transcriptoma de célula única de cerebro de ratón (ctd_allKI) proporcionados por el instituto Karolinska. Los datos de expresión de este *dataset* proceden de las siguientes regiones cerebrales: neocórtex, hipocampo, hipotálamo, estriado y mesencéfalo, así como muestras de tejido cerebral enriquecido en oligodendrocitos, neuronas dopaminérgicas e interneuronas parvalbúmina (células totales = 9970). Para realizar el análisis de enriquecimiento se utilizaron como *background* todos los genes humanos con ortólogo de ratón. La distribución de probabilidad para los listados de genes de interés, se calculó creando 100000 listas

de genes aleatorias a partir del *set* de genes *background*, ajustando la longitud del tránscrito y el contenido GC, y determinando el nivel de expresión para cada una de ellas. Este método de muestreo *bootstrapping* se utilizó en el nivel de anotación 1, que se refiere al tipo celular.

El paquete, de R pSI (*Specificity Index Statistic*) se utilizó para caracterizar el patrón de expresión de los genes con mutaciones germinales y los genes con PZMs en diferentes regiones cerebrales y a lo largo de diferentes periodos del desarrollo[206]. Las listas de genes cuyo patrón de expresión es conocido (human.rda) se obtuvieron a partir de datos de *BrainSpan* (*set* de genes para 6 regiones cerebrales y *set* de genes para 10 periodos del desarrollo). El test de iteración de Fisher, que incluye el propio paquete, se usó para analizar si los dos grupos de genes (genes con mutaciones germinales y genes con PZMs) estaban significativamente sobrerrepresentados en alguno de los *sets*. Las regiones cerebrales que estaban significativamente enriquecidas para alguno de los grupos de genes en algún periodo del neurodesarrollo (p-valor ajustado < 0.05) se representaron en una matriz.

5.3 RESULTADOS

5.3.1 Test Transmission and De novo Association (TADA-Denovo)

El test *Transmission And De novo Association* (TADA) identifica los genes de riesgo para TEA en función de varios parámetros: el *ratio* mutacional del gen, la recurrencia de mutaciones *de novo* y el impacto funcional de cada tipo de mutación[102]. Así pues, el análisis TADA-Denovo se realizó de manera independiente en ambos grupos de genes (genes con mutaciones germinales y genes con PZMs). Su principal objetivo es identificar posibles genes de riesgo en los TEA que puedan estar involucrados diferencialmente en su etiología dependiendo de si portan mutaciones germinales o PZMs. El análisis se limitó a mutaciones patogénicas (mutaciones LoF y mutaciones *missense* probablemente patogénicas) para incrementar la probabilidad de encontrar genes candidatos. En primer lugar, se analizó el *set* de genes de la cohorte española (360 tríos) (genes con mutaciones germinales = 181; genes con PZMs = 105). Así, el análisis de genes con mutaciones

germinales identificó 12 genes con FDR < 0.3, incluyendo 3 genes (*SCN2A*, *ARID1B* y *CHD8*) con FDR < 0.1 (Tabla 30). (Tabla Suplementaria 16 de Alonso-González *et al*.,[189]). El análisis de genes con PZMs identificó 13 genes con FDR < 0.3 (Tabla 31) de los cuales 4 (*KMT2C*, *FRG1*, *GRIN2B* y *MAP2K3*) tenían un FDR < 0.1. (Tabla Suplementaria 17 de Alonso-González *et al*.,[189])

Genes	q-valor	p-valor
SCN2A	0.004	2.76×10^{-7}
ARID1B	0.050	2.76×10^{-6}
CHD8	0.066	3.31×10^{-6}
FIG4	0.103	9.39×10^{-5}
RBM15	0.126	1.16×10^{-5}
HUWE1	0.148	3.54×10^{-5}
KIAA1107	0.188	5.08×10^{-5}
VWAS5B1	0.218	5.08×10^{-5}
EMCN	0.242	6.13×10^{-5}
SH2B2	0.262	7.90×10^{-5}
ASMT	0.277	9.34×10^{-5}
MYLK4	0.291	9.67×10^{-5}

Tabla 30. Genes de riesgo en los TEA con mutaciones germinales en la cohorte española. Los p-valores y q-valores se obtuvieron tras realizar el análisis TADA-Denovo considerando las mutaciones germinales identificadas en la cohorte de TEA española (360 tríos). Solo se muestran los genes con FDR < 0.3.

Genes	q-valor	p-valor
KMT2C	0.001	4.76×10^{-7}
FRG1	0.015	4.46×10^{-7}
GRIN2B	0.040	4.76×10^{-7}
MAP2K3	0.08	6.67×10^{-6}
SRGAP2	0.106	7.62×10^{-6}
MBD6	0.124	7.62×10^{-6}
POTEB2	0.168	2.57×10^{-5}
CALML6	0.200	2.67×10^{-5}
PRDX6	0.226	3.05×10^{-5}
SSR2	0.245	3.14×10^{-5}
VEGFA	0.264	4×10^{-5}
CANX	0.278	0.0001
ZNF276	0.290	0.00012

Tabla 31. Genes de riesgo en los TEA con PZMs en la cohorte española. Los p-valores y q-valores se obtuvieron tras realizar el análisis TADA-Denovo considerando las PZMs identificadas en la cohorte de TEA española (360 tríos). Solo se muestran los genes con FDR < 0.3.

En la cohorte combinada, el análisis con TADA de genes con mutaciones germinales (N = 1210) identificó 34 genes con FDR < 0.1 (Tabla 32 y Figura 24) y 102 genes con FDR < 0.3. (Tabla Suplementaria 18 de Alonso-González *et al*.,[189]). Tres de los genes (*SCN2A, ARID1B, CHD8*) con mutaciones germinales fueron priorizados por TADA tanto en la cohorte española como la combinada.

El análisis de genes con PZMs (N = 362) en la cohorte combinada identificó tres genes (*FRG1, KMT2C* y *NFIA*) con FDR < 0.1 y 14 con FDR < 0.3 (Tabla 33 y Figura 25). (Tabla Suplementaria 19 de Alonso-González *et al*.,[189]). Solo dos genes, *KMT2C* y *FRG1*, mostraron asociación tras la corrección por FDR (< 0.1) tanto en la cohorte combinada como en la cohorte española.

Genes	q-valor	p-valor
SCN2A	5.04×10^{-12}	4.13×10^{-8}
CHD8	2.40×10^{-5}	4.13×10^{-8}
ARID1B	5.36×10^{-5}	4.13×10^{-8}
SLC6A1	0.00015	2.48×10^{-7}
SYNGAP1	0.0005	6.61×10^{-7}
KDM5B	0.0008	8.26×10^{-7}
SUV420H1	0.002	5.37×10^{-6}
TRIP12	0.003	5.79×10^{-6}
PTEN	0.004	1.14×10^{-5}
KATNAL2	0.008	5.01×10^{-5}
NRXN1	0.012	5.79×10^{-5}
CREBBP	0.02	5.92×10^{-5}
CELF4	0.02	6.09×10^{-5}
STXBP1	0.02	6.48×10^{-5}
DYRK1A	0.02	7.09×10^{-5}
CHD2	0.03	0.0001
ANK2	0.03	0.0001
WDFY3	0.03	0.0001
UNC80	0.04	0.0002
CLASP1	0.04	0.0002
TMEM39B	0.05	0.0002
PRKAR1B	0.05	0.0002
USP45	0.05	0.0003
NUAK1	0.06	0.0004
NAA15	0.06	0.0004
FOXP1	0.07	0.0004
ZC3H11A	0.07	0.0004
DPP3	0.07	0.0005
PRKDC	0.08	0.0005
ATP1A1	0.08	0.0005
LRP5	0.09	0.0005
SLC12A3	0.09	0.0006
FBXO18	0.096	0.0006
PTK7	0.0999	0.0007

Tabla 32. Genes de riesgo en los TEA con mutaciones germinales en la cohorte combinada. Los p-valores y q-valores se obtuvieron tras realizar el análisis TADA-Denovo considerando las mutaciones germinales identificadas en la cohorte de TEA combinada (2103 tríos). Solo se muestran los genes con FDR < 0.1.

Genes	q-valor	p-valor
FRG1	0.04	4.14×10^{-3}
KMT2C	0.07	0.00018
NFIA	0.09	0.00028
SMARCA4	0.12	0.00052
PRKDC	0.13	0.00055
KLF16	0.15	0.00064
GRIN2B	0.17	0.00095
MAP2K3	0.18	0.00098
HNRNPU	0.21	0.0019
POTEB2	0.23	0.002
RNPC3	0.25	0.002
FAM177A1	0.27	0.002
CALML6	0.28	0.002
CMPK2	0.3	0.003

Tabla 33. Genes de riesgo en los TEA con PZMs en la cohorte combinada. Los p-valores y q-valores se obtuvieron tras realizar el análisis TADA-Denovo considerando PZMs identificadas en la cohorte de TEA combinada (2103 tríos). Solo se muestran los genes con FDR < 0.3.

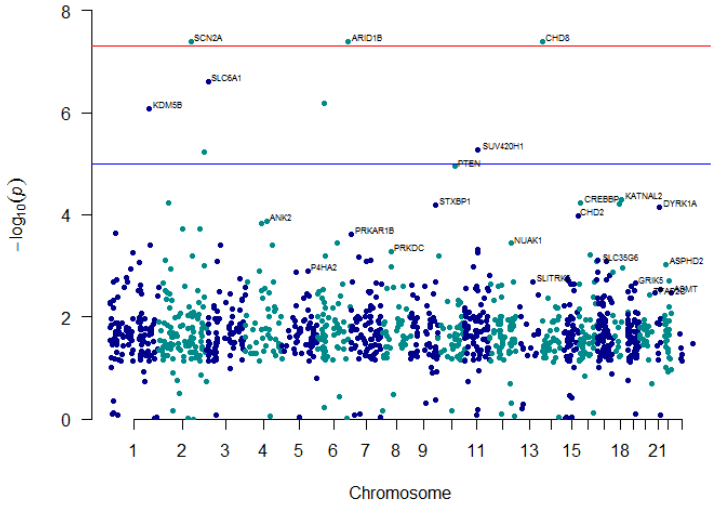

Figura 25 *Manhattan plot* mostrando los genes asociados (de riesgo) en los TEA en el análisis de priorización realizado con TADA-Denovo (en el eje x e y se representan cromosoma y log10 del p-valor para cada gen). Se muestran los p-valores obtenidos en el análisis de mutaciones germinales en la cohorte combinada.

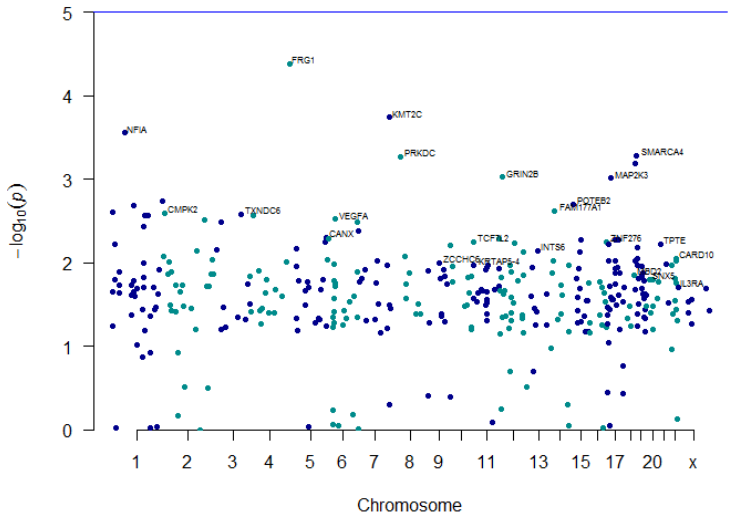

Figura 24. *Manhattan plot* mostrando los genes asociados (de riesgo) en los TEA en el análisis de priorización realizado con TADA-Denovo. (en el eje x e y se representan cromosoma y log10 del p-valor para cada gen). Se muestran los p-valores obtenidos en el análisis de PZMs en la cohorte combinada.

Un total de 17 genes (50%) de entre los genes con mutaciones germinales de la cohorte combinada (34 genes, FDR < 0.1), se identificaron como genes candidatos según los criterios de SFARI (*scores* 1, 2, 1s y 2s). Además, 11 de esos genes (64.70%) mostraron FDR < 0.1 en análisis TADA previos[105]. Del resto de genes identificados por TADA (FDR < 0.1) 10 estaban presentes en la base de datos de SFARI (*score* 3, 4 y 5), y 5 de ellos se habían relacionado previamente con otra enfermedad distinta a TEA en la base de datos OMIM (Anexo 4).

El análisis de PZMs mostró asociación de *KMT2C* (*score* SFARI s2) así como de otros 3 genes con FDR < 0.1 (*NFIA* y *FRG1*). *NFIA* se había reportado previamente como un posible gen candidato en TEA (*score* SFARI 4). Sin embargo, *FRG1*, mostró asociación por primera vez en este análisis. De los restantes genes con FDR < 0.3, *SMARCA4*, *PRKDC, KLF16*, y *HNRNPU* (*score* SFARI 3 y 4) son genes que habían sido previamente identificados como genes candidatos con valores de FDR entre 0.1 y 0.3 en análisis TADA previos, y *GRIN2B* había sido reportado en SFARI como un gen de riesgo de alta probabilidad (*score* 1). (Anexo 5)

5.3.2 Análisis de enriquecimiento en *sets* de genes de mutaciones germinales y PZMs

DNENRICH se empleó para realizar un análisis de enriquecimiento de mutaciones germinales y PZMs en *sets* de genes (ver métodos) previamente involucrados en la etiología de los TEA y otros TND. En este análisis solo se consideraron mutaciones *nonsense* y *missense* y las mutaciones sinónimas fueron eliminadas debido a que contribuyen con poca probabilidad al fenotipo autista.

En primer lugar, para el análisis de enriquecimiento se utilizó el listado de genes con mutaciones *de novo* (germinales y PZMs) de la cohorte española (Tablas Suplementarias 9 y 10 de Alonso-González *et al.*,[189]). Los resultados indicaron que los genes con mutaciones germinales (genes con mutaciones germinales = 228; mutaciones germinales = 236) estaban enriquecidos para varios *sets* de genes: genes diana de *FMRP* (p-valor = 0.00013); genes conocidos de DI (p-valor = 0.0073); genes intolerantes a mutaciones LoF (p-valor = 0.002); genes SFARI

(p-valor = 1 x 10^{-6}); y genes relacionados con la organización de la cromatina (p-valor = 0.00018) (Tabla 34). Sin embargo, el listado de genes con PZMs (genes con PZMs = 155; PZMs = 164) solo mostró asociación con genes intolerantes a mutaciones LoF (Tabla 35)

***Sets* de genes**	**p-valor**	**Mutaciones observadas**	**Mutaciones esperadas**
Genes con expresión monoalélica en neuronas en diferenciación	0.6	12	12.502
Genes diana de *CELF4*	1	0	0.05078
Genes esenciales	0.0709	49	39.968
Genes diana de *FMRP*	0.00013	39	20.9788
***Enhancers* candidatos de genes con expresión en el telencéfalo**	0.5386	1	0.771873
Gene diana de *CHD8*	0.226	39	34.4983
Genes conocidos de DI	0.0073	40	27.0286
Genes intolerantes a mutaciones LoF	0.0018	79	58.9156
Gene SFARI	1 x 10^{-6}	49	20.9601
Genes sinápticos	0.3592	14	12.4002
Genes relacionados con la organización de la cromatina	0.00018	24	10.7377
Genes diana de miR-137	0.3251	8	6.49354
Genes diana de miR-128	1	0	0.022676
Genes diana de *RBFOX*	1	0	0.055409

Tabla 34. Resultados del análisis de enriquecimiento en *sets* de genes para los genes con mutaciones germinales (cohorte española).

Sets de genes	p-valor	Mutaciones observadas	Mutaciones esperadas
Genes con expresión monoalélica en neuronas en diferenciación	0.9727	4	8.48746
Genes diana de *CELF4*	1	0	0.034487
Genes esenciales	0.2340	31	27.1352
Genes diana de *FMRP*	0.0764	20	14.2422
***Enhancers* candidatos de genes con expresión en el telencéfalo**	0.0977	2	0.524631
Gene diana de *CHD8*	0.0884	30	23.4115
Genes conocidos de DI	0.7554	16	18.3456
Genes intolerantes a mutaciones LoF	0.0283	51	39.9973
Gene SFARI	0.0764	20	14.2361
Genes sinápticos	0.3339	10	8.41702
Genes relacionados con la organización de la cromatina	0.0625	12	7.28793
Genes diana de miR-137	0.8198	3	4.40692
Genes diana de miR-128	1	0	0.015316
Genes diana de *RBFOX*	1	0	0.037786

Tabla 35. Resultados del análisis de enriquecimiento en *sets* de genes para los genes con PZMs (cohorte española).

El análisis DNENRICH en la cohorte combinada mostró un enriquecimiento de los genes con mutaciones germinales y PZMs para diferentes *sets* de genes: organización de la cromatina, genes SFARI, genes intolerantes a mutaciones LoF, genes diana de *CHD8* y genes esenciales. Además, los genes con mutaciones germinales (genes con mutaciones germinales = 1972; mutaciones germinales = 2270) mostraron también enriquecimiento para genes diana de *FMRP* (p-valor = 1×10^{-6}), genes conocidos de DI (p-valor = 1×10^{-6}) y genes sinápticos (p-valor = 4×10^{-6}) (Tabla 36 y Figura 26). Los genes con PZMs (genes con PZMs = 624; PZMs = 676) solo mostraron asociación para genes diana de miR-137 (p-valor = 0.0019) (Tabla 37 y Figura 26).

Sets de genes	p-valor	Mutaciones observadas	Mutaciones esperadas
Genes con expresión monoalélica en neuronas en diferenciación	0.6999	121	126.324
Genes diana de *CELF4*	1	0	0.51357
Genes esenciales	$1.x\ 10^{-6}$	533	404.162
Genes diana de *FMRP*	$1.x\ 10^{-6}$	331	212.218
Enhancers candidatos de genes con expresión en el telencéfalo	0.0550	13	7.82029
Gene diana de *CHD8*	$3.5\ x10^{-5}$	420	348.544
Genes conocidos de DI	$1.x\ 10^{-6}$	373	273.213
Genes intolerantes a mutaciones LoF	$1.x\ 10^{-6}$	733	595.59
Gene SFARI	$1.x\ 10^{-6}$	434	211.973
Genes sinápticos	$4.x\ 10^{-6}$	180	125.408
Genes relacionados con la organización de la cromatina	$1.x\ 10^{-6}$	168	108.449
Genes diana de miR-137	0.1368	75	65.7234
Genes diana de miR-128	1	0	0.22838
Genes diana de *RBFOX*	0.4299	1	0.562709

Tabla 36. Resultados del análisis de enriquecimiento en *sets* genes para los genes con mutaciones germinales (cohorte combinada).

Sets de genes	p-valor	Mutaciones observadas	Mutaciones esperadas
Genes con expresión monoalélica en neuronas en diferenciación	0.9744	26	36.4938
Genes diana de *CELF4*	1	0	0.147914
Genes esenciales	0.0118	140	116.895
Genes diana de *FMRP*	0.0425	75	61.4294
Enhancers candidatos de genes en el telencéfalo	0.0796	5	2.2672
Genes diana de *CHD8*	0.0228	120	100.751
Genes conocidos de DI	0.1840	87	79.0127
Genes intolerantes a mutaciones LoF	0.0003	212	172.214
Gene SFARI	1×10^{-7}	127	61.2987
Genes sinápticos	0.6109	35	36.2813
Genes relacionados con la organización de la cromatina	0.0106	45	31.3424
Genes diana de miR-137	0.0019	33	19.0234
Genes diana de miR-128	1	0	0.066195
Genes diana de *RBFOX*	1	0	0.162842

Tabla 37. Resultados del análisis de enriquecimiento en *sets* de genes para los genes con PZMs (cohorte combinada).

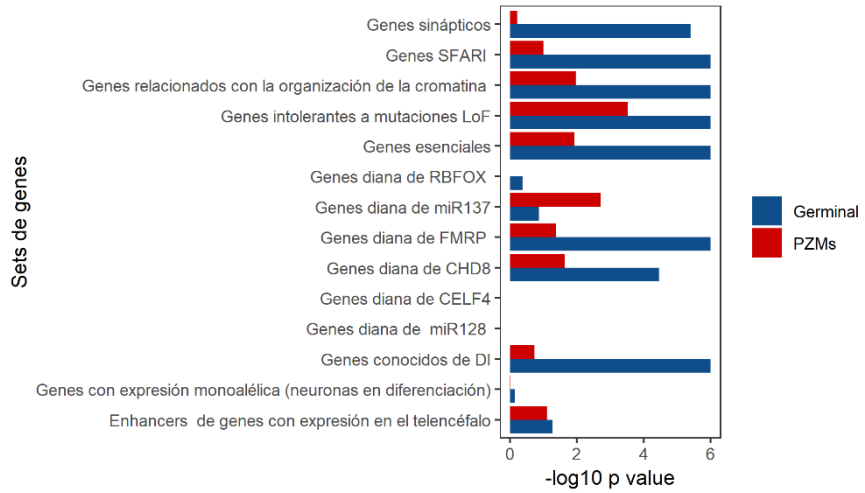

Figura 26. Análisis de enriquecimiento en s*ets* de genes usando mutaciones *de novo* germinales y PZMs de la cohorte combinada. Se muestra el -log10 p-valor obtenido para cada set de genes con cada tipo de mutación (germinal o PZMs).

El mismo análisis se llevó a cabo utilizando los listados de genes con mutaciones germinales y PZMs de los hermanos no afectos (genes con mutaciones germinales = 744; mutaciones germinales = 780; genes PZMs = 237; PZMs = 239) (Tabla Suplementaria 14 de Alonso-González *et al.*,[189]), mostrando solo asociación para genes diana de *FMRP*.

5.3.3 Análisis de enriquecimiento de ontologías génicas

El análisis de enriquecimiento de ontologías génicas reveló importantes diferencias entre los genes con mutaciones germinales y los genes con PZMs de la cohorte combinada. (Tabla Suplementaria 15 de Alonso-González *et al.*,[189])

Los genes con mutaciones germinales mostraron un enriquecimiento estadísticamente significativo para diferentes términos GO relacionados con la función sináptica y la regulación de la transcripción. En concreto, merece la pena destacar la asociación para términos GO relacionados con el transporte iónico: GO:0006814, p[Pbh] = 0.005; GO0035725, p[Pbh] = 0.005 and GO0006816, p[Pbh]

= 0.006 (Figura 27a). (Tabla Suplementaria 22 de Alonso-González *et al*.,[189])

El *subset* de genes con PZMs estaba enriquecido para términos GO relacionados con la regulación de la expresión génica, biosíntesis, diferenciación y migración celular: GO0010629, p[Pbh] = 0.074; GO:2000113, p[Pbh] = 0.092; GO:0045652, p[Pbh] = 0.092; GO:0030336, p[Pbh] = 0.0995. (Figura 27b) (Tabla Suplementaria 23 de Alonso-González *et al*.,[189]).

Los resultados del análisis de enriquecimiento de ontologías también se representaron agrupando semánticamente los 50 términos GO más significativos en cada grupo de genes. Así, el análisis de genes con mutaciones germinales mostró 4 clústeres; tres bien diferenciados (desarrollo y diferenciación neuronal, funciones sinápticas, y organización de la cromatina) y un cuarto clúster más heterogéneo relacionado con el desarrollo embrionario (Figura 28a).

En el caso de los genes con PZMs, todos los clústeres resultantes mostraron una interrelación parcial. Sin embargo, fue posible identificar un clúster que incluyó términos relacionados con la regulación de procesos esenciales como son la fosforilación de proteínas, regulación del crecimiento, regulación negativa de procesos de biosíntesis celular y regulación positiva de la transcripción del ADN. Además, también se identificaron términos GO relacionados con desarrollo neuronal y el desarrollo embrionario (Figura 28b).

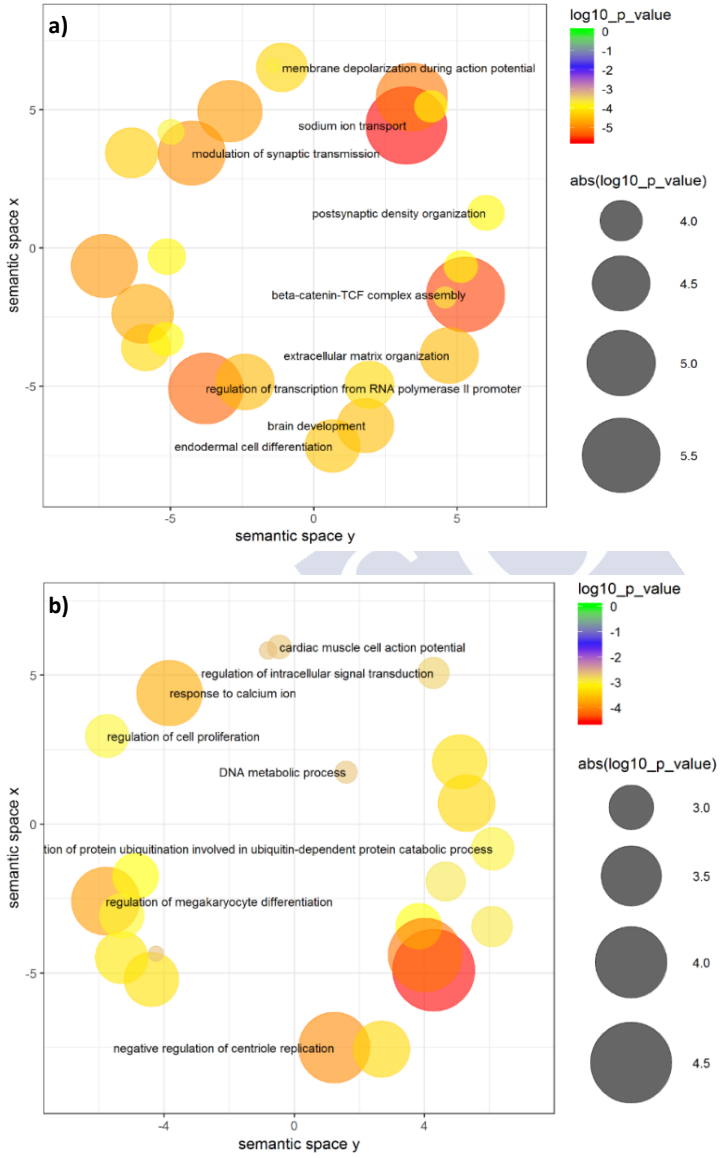

Figura 27. *Scatterplots* que representan los 30 procesos biológicos más significativos de la cohorte combinada. a) Los 30 procesos biológicos enriquecidos en genes con mutaciones germinales. b) Los 30 procesos biológicos enriquecidos en genes con mutaciones PZMs.

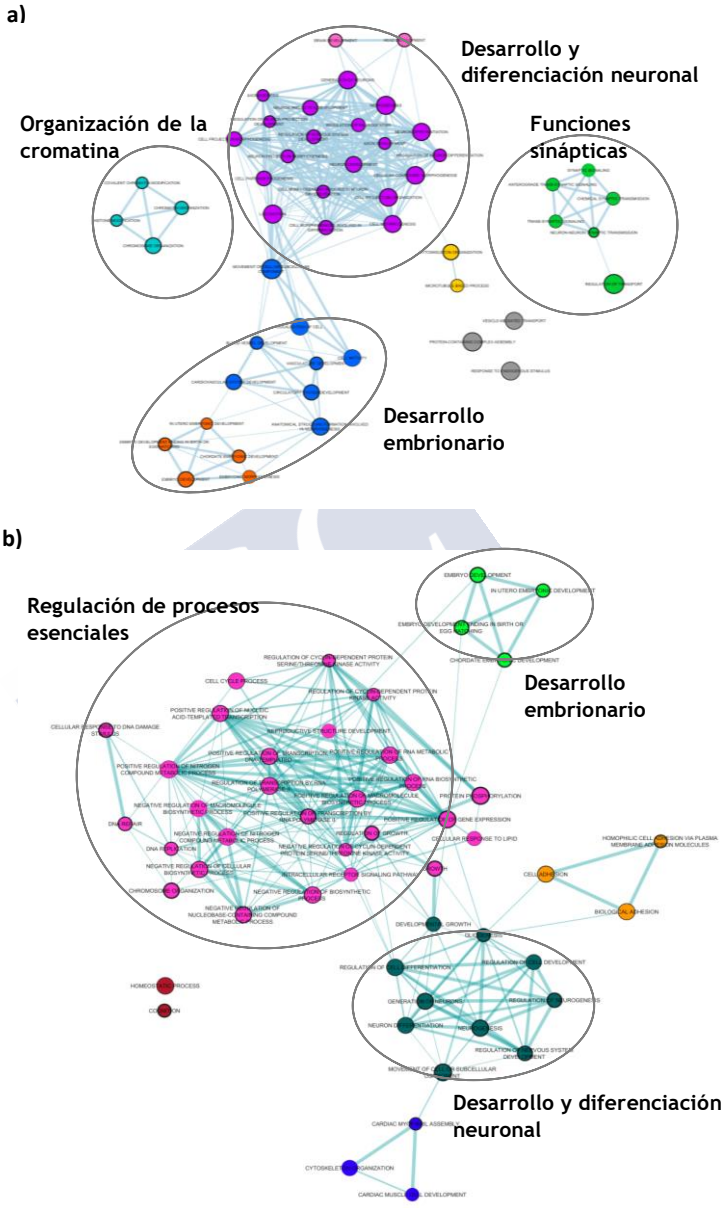

Figura 28. Visualización de los términos GO más significativos en los genes con mutaciones *de novo* de la cohorte combinada, agrupados por funciones biológicas. a) Clústeres de términos GO de los genes con mutaciones germinales. b) Clústeres de términos GO de los genes con PZMs.

5.3.4 Análisis de enriquecimiento por tipo celular y análisis de expresión en regiones cerebrales a lo largo del desarrollo

En primer lugar, se examinó si los genes con mutaciones germinales y genes con PZMs estaban diferencialmente expresados en el *dataset* de transcriptoma correspondiente al nivel 1 de anotación de tipos celulares.

Tal y como se esperaba, los genes con mutaciones germinales mostraron un enriquecimiento significativo en varios tipos celulares (Tabla 38 y Figura 29). Los tipos celulares que resultaron más significativos fueron aquellos relacionados con la transmisión neuronal (neuroblastos dopaminérgicos, p-valor < 0.0001; neuronas dopaminérgicas embrionarias, p-valor < 0.0001; neuronas GABAérgicas embrionarias, p-valor < 0.0001; neuronas serotoninérgicas, p-valor < 0.0001).

Tipo celular	p-valor	Cambio de expresión sobre la media	DE de la media
Neuroblasto dopaminérgico	<0.0001	1.17053835	4.27167674
Neurona dopaminérgica embrionaria	<0.0001	1.175277135	5.113211893
Neurona GABAérgica embrionaria	<0.0001	1.11078582	4.626262187
Neuroblasto	<0.0001	1.187690135	5.49208347
Neurona piramidal CA1	<0.0001	1.132632722	5.365027275
Neurona piramidal SS	<0.0001	1.12556771	5.157499425
Neurona serotoninérgica	<0.0001	1.170544581	5.491769105
Progenitor neural	0.0001	1.15212965	
Células de glía radiales	0.0003	1.124574836	3.589901456
Neuronas de núcleos del telencéfalo embrionarias	0.0004	1.104053784	3.513053618

Tabla 38. Niveles de asociación de los 10 tipos celulares neuronales en genes con mutaciones germinales de la cohorte combinada

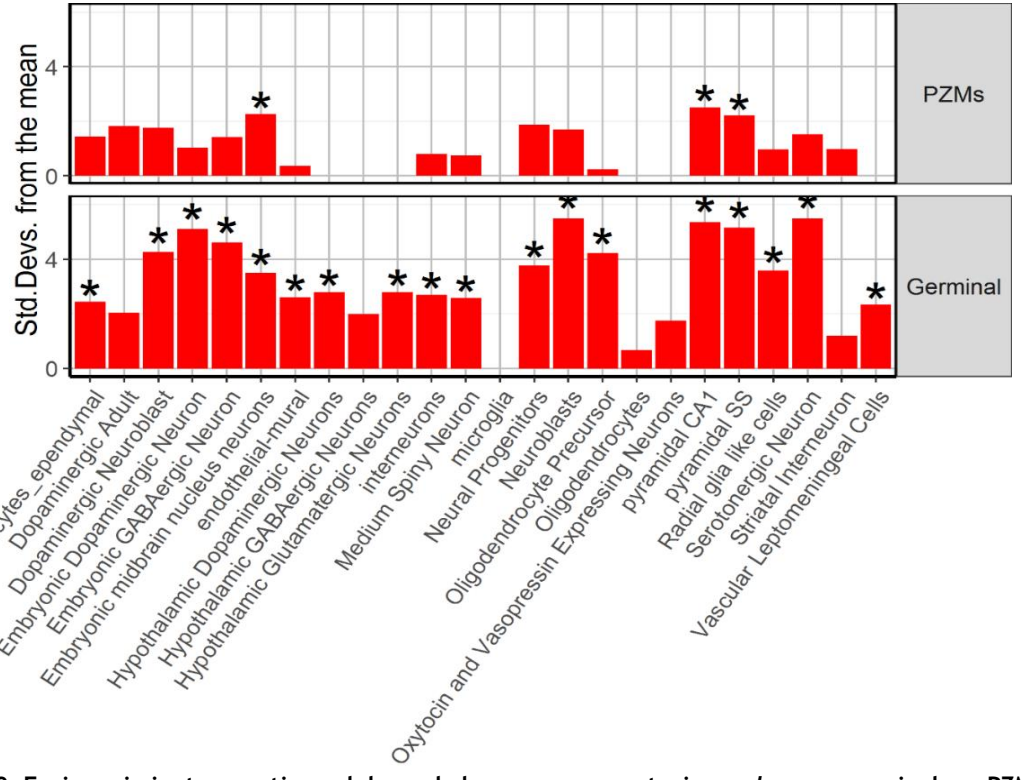

Figura 29. Enriquecimiento para tipo celulares de los genes con mutaciones *de novo* germinales y PZMs en la cohorte combinada (Análisis EWCE)

Los genes con PZMs mostraron enriquecimiento para tres tipos celulares diferentes: neuronas piramidales CA1, p-valor = 0.0066; neuronas piramidales SS, p-valor = 0.016; y neuronas de núcleos del telencéfalo embrionarias, p-valor = 0.02 (Tabla 39 y Figura 29).

Tipo celular	p-valor	Cambio de expresión sobre la media	DE de la media
Neuroblasto dopaminérgico	0.0066	1.116362092	2.509881972
Neurona dopaminérgica embrionaria	0.0152	1.09832511	2.209046983
Neurona GABAérgica embrionaria	0.0185	1.125338403	2.263702976
Neuroblasto	0.0336	1.141154419	1.880862217
Neurona piramidal CA1	0.0404	1.096578007	1.823257602
Neurona piramidal SS	0.0474	1.130737104	1.762033343
Neurona serotoninérgica	0.0572	1.108135618	1.686197427
Progenitor neural	0.067	1.090682473	1.525134485
Células de glía radiales	0.0802	1.123951233	1.432463223
Neuronas de núcleos del telencéfalo embrionarias	0.0827	1.062881294	1.417776316

Tabla 39. Niveles de asociación de los 10 tipos celulares neuronales en genes con PZMs de la cohorte combinada.

Para conocer el patrón de expresión espacio-temporal se analizó la expresión de genes con mutaciones germinales y genes con PZMs en diferentes regiones cerebrales y a lo largo de diferentes periodos del neurodesarrollo.

Los genes con mutaciones *de novo* germinales, mostraron una expresión significativa en córtex, estriado, cerebelo y amígdala en periodos prenatales (temprano, medio y tardío) (Tabla 40 y Figura 30a y 30b).

Los genes con PZMs mostraron una expresión significativa en córtex durante el periodo fetal temprano-medio. Aunque no se encontró asociación en otras áreas cerebrales u otros periodos del neurodesarrollo, sí se obtuvieron p-valores marginalmente significativos en córtex (periodo fetal temprano medio y tardío) y

amígdala (periodo fetal temprano y medio) (Tabla 41 y Figura 31a y 31b).

Región cerebral	Periodo del desarrollo	p-valor
Córtex	Fetal temprano	0.003
Córtex	Fetal temprano medio	0.003
Córtex	Edad adulta	0.003
Estriado	Fetal temprano	0.008
Cerebelo	Fetal tardío medio	0.027
Amígdala	Fetal temprano medio	0.046

Tabla 40. Análisis de expresión de genes con mutaciones germinales de la cohorte combinada en diferentes regiones cerebrales y diferentes periodos del neurodesarrollo.

Región cerebral	Periodo del desarrollo	p-valor
Córtex	Fetal temprano medio	0.00115023
Córtex	Fetal temprano	0.05214531
Córtex	Fetal tardío medio	0.05214531
Amígdala	Fetal temprano medio	0.06882706

Tabla 41. Análisis de expresión de genes con PZMs de la cohorte combinada en diferentes regiones cerebrales y diferentes periodos del neudesarrollo.

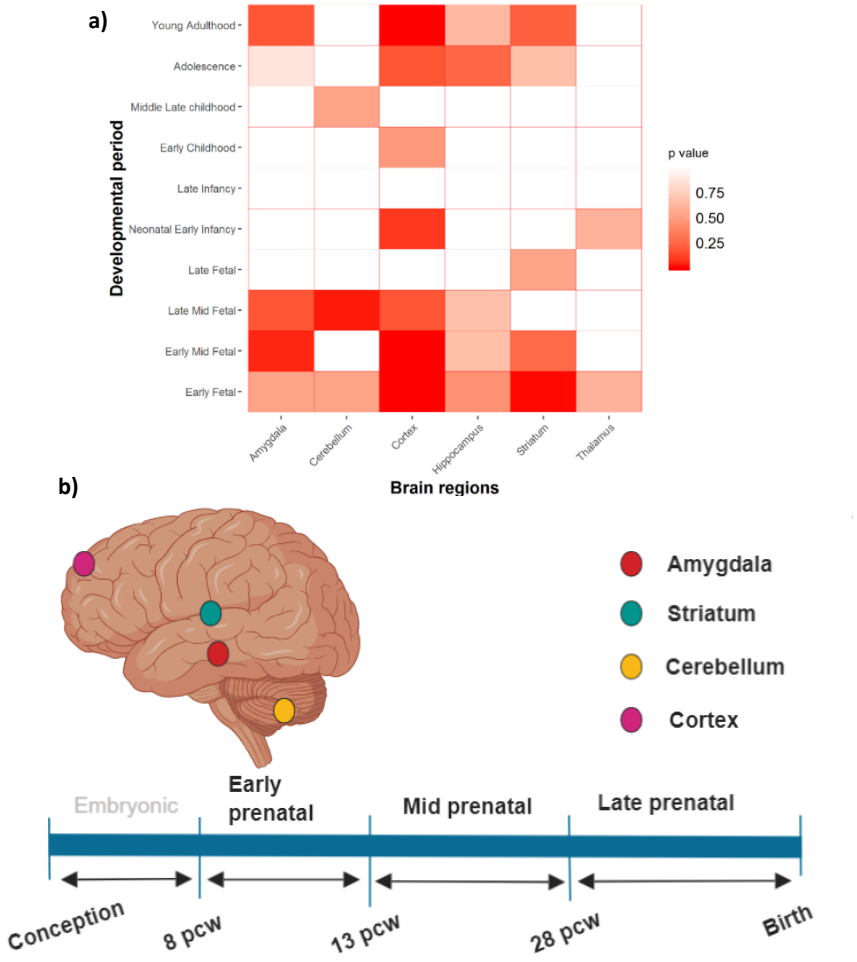

Figura 30. Análisis de expresión de genes con mutaciones germinales en regiones cerebrales a lo largo del neurodesarrollo. a) Se muestra una matriz de expresión donde se ha marcado en color rojo aquellas áreas cerebrales y periodos del neurodesarrollo donde se ha detectado una expresión de genes con mutaciones germinales. b) Se simplifica la información de la imagen anterior en una figura donde solo se marcan las áreas y periodos del neurodesarrollo donde la expresión de los genes es significativa.

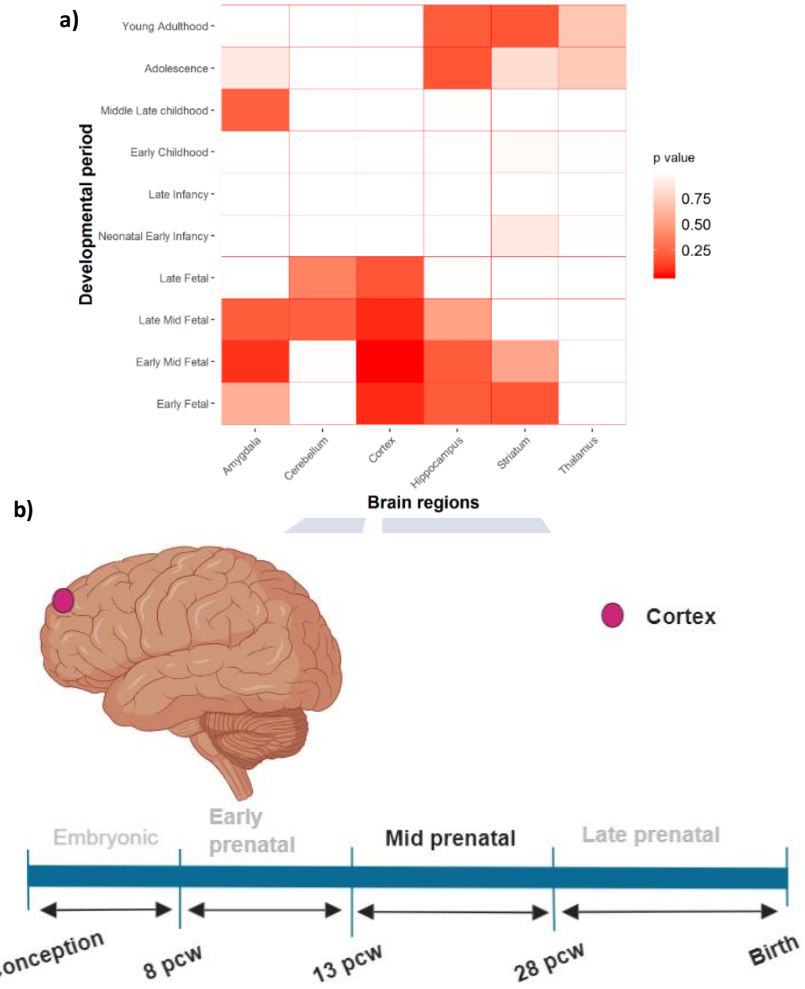

Figura 31. Análisis de expresión de genes con PZMs en regiones cerebrales a lo largo del neurodesarrollo. a) Se muestra una matriz de expresión donde se ha marcado en color rojo aquellas áreas cerebrales y periodos del neurodesarrollo donde se ha detectado una expresión de genes con PZMs. b) Se simplifica la información de la imagen anterior en una figura donde solo se marcan las áreas y periodos del neurodesarrollo donde la expresión de los genes es significativa.

5.4 DISCUSIÓN

Los TEA son TND complejos, caracterizados por una elevada heterogeneidad clínica y genética. Se calcula que aproximadamente 1000 genes pueden estar involucrados en su patogénesis, pero solo un pequeño número de ellos se ha caracterizado funcionalmente. Por ello, se considera que la gran mayoría (o una gran parte) de los factores genéticos implicados en los TEA todavía no se han identificado[105].

Las mutaciones *de novo* son una causa genética conocida en los TEA, aunque, generalmente, se suelen identificar solo aquellas mutaciones que son germinales, obviando las que son postcigóticas. Así pues, tanto el impacto en el diagnóstico de las PZMs como su efecto fenotípico, aún no se ha caracterizado en su totalidad. Sin embargo, estudios recientes en los TEA, sugieren que haya genes de riesgo en los cuales sea más frecuente identificar mutaciones en mosaico que en otros[115,116,207,208]. Asimismo, un estudio reciente, mostró que se pueden observar malformaciones cerebrales cuando solo un 10% de las células de sangre periférica presenta una mutación[109]. Además, las mutaciones en mosaico pueden actuar modificando el fenotipo, lo cual resulta en pacientes con sintomatología más leve[209].

Nuestro análisis de PZMs con TADA ha identificado 3 genes asociados con TEA (q-valor < 0.1; *FRG1, KMT2C* y *NFIA*). *FRG1* no había sido asociado nunca con TND, mientras que *KMT2C* y *NFIA* si habían sido previamente reportados como genes implicados en la etiología de los TEA/DI[210,211]. Sin embargo, ninguno de estos genes resultó asociado cuando el análisis se llevó a cabo en genes con mutaciones germinales. Estos resultados, sugieren pues, la existencia de genes de riesgo en los TEA que pueden acumular diferencialmente un tipo u otro de mutación.

FRG1 (*Facioscapulohumeral muscular dystrophy (FSHD) region gene 1*) codifica para una proteína citoplasmástica importante para el desarrollo muscular y vascular. *FRG1* se localiza a 100 pb de las unidades de repetición en el cromosoma 4q35, cuya deleción se relaciona con el síndrome de Distrofia Muscular Fascioescapulohumeral (FSHD OMIM #158900)[212]. Aunque la patogénesis de este síndrome se desconoce, lo cierto es que esta enfermedad a menudo muestra comorbilidad con epilepsia y DI[213]. Se

cree que *FRG1* actúa como un regulador de *splicing* y recientemente se ha publicado que *Rbfox1* disminuye su expresión cuando *FRG1* incrementa la suya[214]. *RBFOX1* tiene un papel fundamental en la migración neuronal y en el desarrollo de las sinapsis durante la corticogénesis y además, mutaciones en este gen se han relacionado con la patogénesis de los TEA[215]. Sin embargo, a pesar de esta información, se debe tener en cuenta que *FRG1* es tolerante para mutaciones LoF (pLI = 0) y para mutaciones *missense* (Z = 0.18), y, por tanto, se debe manejar este hallazgo con cautela. Parece entonces más probable, que la causa etiológica de los TND relacionados con este gen, sea la desregulación de genes diana de *FRG1* en lugar de su haploinsuficiencia.

KMT2C (lysine (K)-specific methyltransferase 2C) codifica para una metiltranferasa encargada de la regulación de la transcripción génica. Mutaciones *de novo* LoF en este gen se han identificado en pacientes con síndrome de Kleefstra 2 (OMIM #617768). El síndrome de Kleefstra está causado por la haploinsuficiencia de *EHMT1 (euchromatin histone methyltransferase 1)* y se caracteriza por retraso en el desarrollo psicomotor, DI, rasgos dismórficos leves y TEA. Sin embargo, existe una interacción molecular entre *KMT2C* y *EHMT1*, ambos involucrados en la regulación de la plasticidad sináptica en el cerebro adulto, que puede dar lugar a los síntomas mencionados[210]. Además, mutaciones en mosaico en *KMT2C* se han detectado en estudios de secuenciacón de exoma completo donde se han empleado *pipelines* adecuados para su detección[207]. Estos hallazgos resaltan el posible papel causal de las PZMs cuando éstas se localizan en genes de riesgo conocidos para TEA.

NFIA (Nuclear Factor I A) codifica para un miembro de la familia de factores de transcripción NF1 *(nuclear factor 1)* cuyo papel es fundamental en la regulación de la gliogénesis y otros procesos neuronales[216]. Aunque se han identificado mutaciones *de novo* en este gen en pacientes con TEA, no hay evidencia suficiente para considerarlo un gen de riesgo en este TND[92]. Sin embargo, sí se ha descrito un síndrome relacionado con este gen, causado por mutaciones tanto *de novo* como heredadas. Este síndrome se caracteriza por malformaciones cerebrales y puede manifestar también defectos del

tracto urinario (OMIM #613735)[217]. Su presentación clínica es muy variable y pocas veces todos los síntomas descritos se encuentran en un individuo afecto[218]. Hasta donde se sabe, esta es la primera vez que *NFIA* es señalado como un posible gen candidato en los TEA. Esto ha sido posible gracias al análisis de PZMs que hemos llevado a cabo, lo que significa que, de otra manera, no lo habríamos detectado.

El hecho de que haya genes en nuestro genoma que portan diferentes tipos de mutaciones *de novo* (germinales y/o PZMs) puede deberse a que, las mutaciones germinales, pero no todas las mutaciones en mosaico, sean letales en algunos genes. Este es el caso, por ejemplo, del síndrome de Rett, en el que las mutaciones en *MECP2* son letales en hombres y dominantes en mujeres, pero se han descrito casos compatibles con la vida de mutaciones mosaico en hombres[218]. Sin embargo, teniendo en cuenta los resultados de este estudio, otra hipótesis razonable sería considerar que las PZMs en algunos genes causen manifestaciones clínicas diferentes a las que se esperarían si las mutaciones fueran germinales. Así, aunque se ha descrito que el síndrome de Kleefstra está causado por mutaciones *de novo* LoF en *KMT2C*, PZMs en el mismo gen causan una presentación clínica más leve[210]. Este hecho explicaría que los pacientes con síntomas menos severos y síntomas nucleares de TEA están sobrerrepresentados en la cohorte de estudio, facilitando la asociación de *KMT2C* en el análisis de PZMs. Por el contrario, es probable que los pacientes con mutaciones germinales en el mismo gen hayan sido descartados por no cumplir criterios de inclusión y presentar una clínica muy sindrómica. Por la misma razón, *NFIA* podría estar asociado con TEA solo cuando las mutaciones en él son mosaico.

Desde el punto de vista biológico, el análisis de enriquecimiento realizado con genes con mutaciones *de novo* germinales y genes con PZMs ha mostrado una asociación significativa con *sets* de genes previamente implicados en la patogénesis de los TEA: genes diana de *FMRP*, genes relacionados con la organización de la cromatina, y genes candidatos de SFARI[105]. Sorprendentemente, solo los genes con PZMs han mostrado una asociación para genes diana de miR-137. miR-137 es un ARN no codificante, conocido por su papel esencial durante el neurodesarrollo[219]. Su expresión es crucial para mantener un balance

correcto entre la diferenciación y la proliferación neuronal. También está involucrado en la maduración neuronal, el desarrollo de las dendritas y la sinaptogénesis[220–222]. Cabe resaltar que variantes comunes identificadas en el gen que codifica para miR-137 se han asociación con TEA y esquizofrenia[223,224]. Además, en modelos de ratón, la pérdida completa de miR-137 es letal, pero su pérdida parcial resulta en un fenotipo que se asemeja mucho a los síntomas nucleares de los TEA como el comportamiento repetitivo y el comportamiento social deficiente[225].

El análisis de enriquecimiento en ontologías génicas también apunta a que ambos tipos de genes están involucrados en funciones biológicas diferentes[116]. Así, los genes con mutaciones germinales están fundamentalmente asociados con términos GO relacionados con el transporte iónico y la regulación de las funciones sinápticas. Por el contrario, los genes con PZMs están enriquecidos en términos GO relacionados con la regulación negativa de la expresión génica.

El análisis de enriquecimiento para tipos celulares neuronales y el análisis del patrón de expresión espacio-temporal también ha ayudado a comprender el impacto funcional que tienen las mutaciones germinales y las PZMs. Así pues, los genes con mutaciones germinales han mostrado un enriquecimiento significativo en neuronas tanto excitatorias como inhibitorias, mientras que los genes con PZMs han mostrado asociación fundamentalmente con neuronas piramidales. Además, los resultados del análisis indican que los genes con mutaciones germinales se expresan tanto en neuronas fetales como adultas y en diferentes áreas cerebrales (córtex, estriado, cerebelo y amígdala). Por el contrario, la expresión de los genes con PZMs está restringida al córtex durante el periodo fetal medio. Cabe destacar en este punto, que el estudio llevado a cabo por Lim *et al*., resaltó la amígdala como un área cerebral en la cual los genes con PZMs se expresaban de manera significativa. Sin embargo, este análisis solo incluyó PZMs localizados en exones "críticos" (exones que no acumulan mutaciones deletéreas en individuos sanos), en comparación con el presente estudio, en el cual se han incluido todos los genes en los cuales se detectó alguna PZMs no sinónima[116].

Los resultados obtenidos en este trabajo, en conjunto, sugieren que los genes con mutaciones germinales y los genes con PZMs están involucrados en diferentes mecanismos biológicos de susceptibilidad para los TEA. Así, las mutaciones germinales estarían ligadas fundamentalmente a un déficit en la comunicación neuronal, lo cual respalda la teoría de que la pérdida del balance excitatorio/inhibitorio se relaciona con la etiología de los TEA[104,226]. Además, el hecho de que distintas áreas cerebrales, a lo largo de diferentes periodos del desarrollo, se vean afectadas por las mutaciones germinales, explicaría la heterogeneidad clínica que caracteriza a los TEA, así como su alta comorbilidad con otros trastornos como epilepsia y DI. Por el contrario, aunque se necesita una replicación en cohortes mayores, los resultados de este trabajo sugieren que en presencia de PZMs, es más común la interrupción de procesos biológicos que ocurren muy temprano durante el neurodesarrollo, como la neurogénesis, la migración neuronal o la diferenciación. Sin embargo, esta disrupción solo tendría lugar en algunas células del cerebro, mientras que el resto mantendría su funcionamiento normal. En línea con esta hipótesis, cerebros de niños con TEA han mostrado parches de organización laminar anómala que podría ser el resultado, solo en algunas células, de una alteración de la migración neuronal a su destino final[227].

Así pues, los hallazgos obtenidos en este estudio señalan a un periodo crítico durante el desarrollo fetal medio en el cual se produce una sobreexpresión de genes asociados a los TEA. Es posible entonces, que mutaciones en mosaico en genes involucrados en procesos esenciales, puedan ser causa suficiente de manifestaciones clínicas propias de los TEA.

5.5 CONCLUSIONES

El análisis de mutaciones *de novo* llevado a cabo en este trabajo proporciona información sobre el papel de las PZMs en la etiología de los TEA y respalda evidencias previas sobre su papel patogénico y su contribución al riesgo en este TND.

Los resultados de este estudio sugieren que las PZMs puedan tener un impacto funcional en procesos biológicos y periodos del neurodesarrollo diferentes a los de las mutaciones germinales. Además,

el análisis realizado ha revelado que los genes de riesgo en TEA pueden portar mutaciones germinales o PZMs de manera diferencial y este hallazgo resalta la importancia de llevar a cabo una detección precisa de ambos tipos de mutaciones en los estudios de exoma y genoma completo.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

6 CAPÍTULO 3

6.1 OBJETIVO

En el presente estudio se han analizado los exomas completos de 125 tríos (probando afecto y progenitores sanos) pertenecientes a una cohorte de TEA, en busca de variantes genéticas clínicamente relevantes que sirvan para proporcionar un diagnóstico genético en los probandos. Para ello, en cada probando, se ha revisado manualmente un archivo que contiene todas las variantes con MAF < 0.01%, exónicas o de *splicing*, no sinónimas, localizadas en genes asociados a TND. También se han analizado todas las variantes *de novo* localizadas en todos los genes del genoma.

El objetivo principal de este trabajo ha sido calcular el rendimiento diagnóstico de la secuenciación de exoma completo en una cohorte con diagnóstico clínico de TEA, usando una aproximación de trío completo. Esto servirá para averiguar qué probandos podrían beneficiarse de la secuenciación de exoma completo como primera herramienta diagnóstica y mejorar así el algoritmo diagnóstico que se emplea en los TEA en la actualidad.

6.2 MÉTODOS

6.2.1 Pacientes y diagnóstico clínico

En este estudio se analizaron los exomas de 125 individuos con un diagnóstico clínico de TEA, y sus dos progenitores, del Complejo Hospitalario de Santiago de Compostela y de entidades gallegas que trabajan con individuos con TEA (ASPANAES, BATA, MENELA y ASPERGA). Estos 125 individuos formaban parte de una cohorte mayor (N = 136 tríos), en la cual se había llevado a cabo un análisis de CNVs mediante *microarrays* cromosómicos (Affymetrix Cytoscan HD Array). Así pues, se excluyeron de este estudio 10 individuos en los cuales se encontró una CNV clínicamente significativa y un individuo con un Síndrome de X frágil.

El diagnóstico clínico de TEA fue realizado por un neurólogo pediátrico o un psiquiatra de acuerdo a los criterios diagnósticos del DSM-IV o DSM-5 (ver Introducción). También se solicitó información clínica adicional de cada paciente y esta se hizo llegar a través de informes emitidos por neurólogos, psicólogos o psiquiatras. Solo se incluyeron pacientes con 3 años o más. Además, aquellos niños con un trastorno genético conocido previo, que pudiese explicar el diagnóstico clínico de TEA, fueron excluidos del estudio.

El proyecto fue aprobado por el Comité Ético de Investigación Clínica de Galicia (Código 2015/098; Anexo 1), Los progenitores, o en su defecto, los tutores legales, fueron debidamente informados de la naturaleza y el objetivo del estudio, y se requirió la firma de un consentimiento informado para su participación en él (Anexos 3 y 4).

6.2.2 Extracción de ADN y secuenciación de exoma completo

El ADN de cada uno de los sujetos que componen el trío fue extraído a partir de sangre periférica usando el kit *Chemagic DNA Blood 100 Kit (PerkinElmer Inc, Massachusetts, USA),* siguiendo las indicaciones del fabricante.

La secuenciación del exoma completo de cada uno de los tríos (probando y progenitores sanos) fue realizada por el ASC (https://genome.emory.edu/ASC/)[104].

6.2.3 Selección de genes asociados a TND

Se creó una lista de 261 genes asociados a TND (Anexo 6). La selección de estos genes se llevó a cabo en base al nivel de evidencia de asociación con TEA de acuerdo a los criterios SFARI (https://gene.sfari.org), y en base a su relación causal con encefalopatías epilépticas, DI, metabolopatías y trastornos sindrómicos que cursan con TEA y/o DI según OMIM y la literatura científica (Figura 32).

SFARI Gene asigna a cada uno de los genes incluidos en su base de datos un *score* que refleja el nivel de evidencia con el que se asocia el gen a TEA. Los criterios de SFARI para asignar este *score* han sido modificados recientemente (19 de diciembre de 2019). Sin embargo, el listado de genes de este estudio fue creado con anterioridad, de manera que los criterios de selección usados responden a la versión previa, la

cual puede ser encontrada en SFARI Gene Archive. Así pues, se seleccionaron genes pertenecientes a las categorías 1 (nivel de evidencia muy alto), 2 (nivel de evidencia moderado) o sindrómicos (genes que predisponen a TEA en el contexto de un trastorno genético sindrómico). Este listado inicial se amplió tras revisar la base de datos OMIM y la literatura científica, y se incluyeron nuevos genes asociados a encefalopatías epilépticas, DI, metabolopatías y trastornos sindrómicos que cursan con TEA y/o DI. Los genes incluidos debían de cumplir uno, o dos de los siguientes criterios:

1. Que estudios funcionales demuestren que existe una relación causal entre el gen y la enfermedad.
2. Que múltiples individuos no relacionados entre sí y con clínica similar presenten mutaciones en el gen, siendo la distribución de estas variantes significativamente superior en casos con respecto a controles.

6.2.4 Análisis clínico-genómico

6.2.4.1 Anotación, filtrado y clasificación de variantes

El ASC proporcionó un único archivo VCF con los datos de las variaciones en la secuencia exónica en crudo para todos los individuos de la cohorte. La herramienta *bcftools* se empleó para obtener archivos individuales de cada uno de los sujetos de estudio, que incluían todas las variantes en regiones codificantes identificadas en cada individuo. Así pues, cada VCF se anotó usando la herramienta Annovar (*ANNOtate VARiation*) versión 4.3T. Se generaron así ficheros con extensión .csv que fueron abiertos y guardados como archivos Excel con extensión .xlsx. Sobre estos archivos se realizó el filtrado de variantes genéticas.

Para la selección de variantes candidatas para un diagnóstico genético se empleó un *pipeline* automatizado que permitió su filtrado. (http://github.com/xbello/dffiltering). En primer lugar, se eliminaron las variantes con MAF > 0.01 por ser consideradas probablemente benignas, teniendo en cuenta la prevalencia de los TEA en la población (aproximadamente 1%). A continuación, se seleccionaron las variantes no sinónimas localizadas en exones o regiones flanqueantes (variantes que pueden afectar al *splicing*) situadas en alguno de los genes de

nuestro listado, para una revisión manual. Debido a que también estaban disponibles los datos genómicos de los progenitores, también se tuvo en cuenta el patrón de herencia de cada variante (variante *de novo* o heredada) en el proceso de priorización.

Así pues, para evaluar la relación de cada variante candidata con la caracterización fenotípica de cada paciente, se usó el software *Alamut® Visual* (*Interactive Biosoftware*) y la plataforma Varsome (*The Human Genomics Community*) (https://varsome.com/). También se tuvieron en cuenta los predictores de patogenicidad *in silico* integrados en el *software Alamut® Visual*. Para variantes *missense* se usaron Align-GVGD, SIFT y Mutation Taster, y para variantes de *splicing* se usaron SSF *(SpliceSiteFinder*), MES *(MaxEntScan)* y NNS (*NNSPLICE*). Además, también se consultaron bases de datos clínicas y genómicas como OMIM, SFARI Gene, GeneReviews, (https://www.ncbi.nlm.nih.gov/books/NBK1116/), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), DECIPHER (https://decipher.sanger.ac.uk/) y la literatura científica. Finalmente, las variantes seleccionadas fueron clasificadas de acuerdo a las guías propuestas por el ACMG[173] (ver Introducción) de la siguiente manera: 1) patogénicas, 2) probablemente patogénicas, 3) significado incierto, 4) probablemente benignas y 5) benignas (Figura 32).

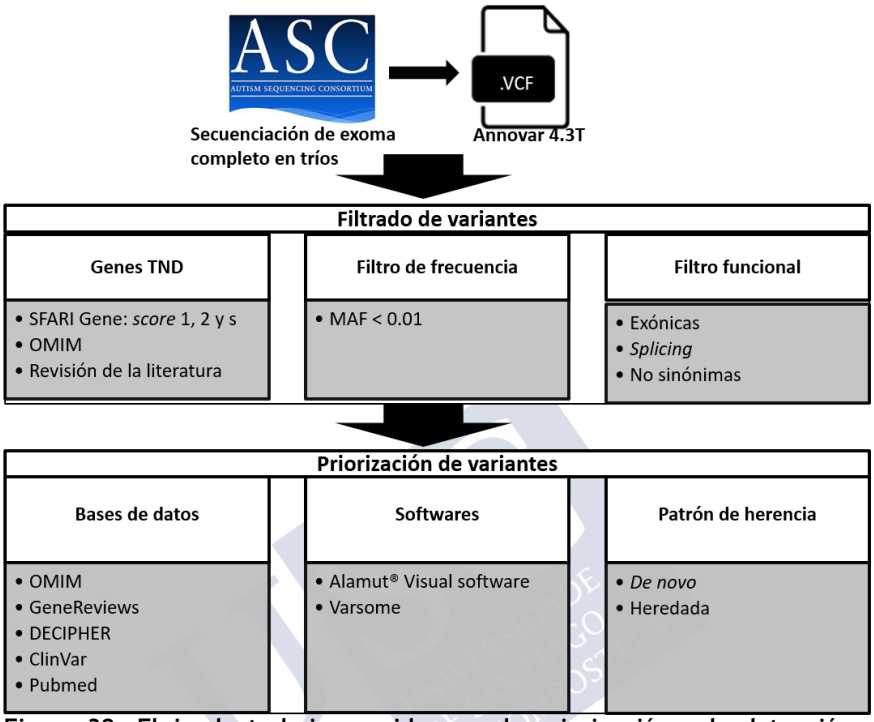

Figura 32. Flujo de trabajo seguido para la priorización y la detección de variantes clínicamente relevantes.

6.2.4.2 Secuenciación Sanger para la validación de variantes genéticas clínicamente relevantes

Todas las variantes que fueron clasificadas como patogénicas, probablemente patogénicas se confirmaron por secuenciación Sanger. En primer lugar, se diseñaron *primers* específicos para cada región de interés usando el *software* Primer 3 (http://bioinfo.ut.ee/primer3-0.4.0). A continuación, se amplificaron las regiones de interés por PCR y posteriormente se purificaron los productos de cada reacción con *ExoSAP-IT™ (Thermo Fisher Scientific, Massachusetts, USA).* Los productos de la PCR se secuenciaron mediante Sanger bidireccionalmente usando el kit *BigDye™ Terminator v3.1 Cycle Sequencing (Thermo Fisher Scientific, Massachusetts, USA)* y los productos de la reacción de secuenciación fueron purificados con el kit *Optima DTR™ 96-Well Plate Kit (Edge Bio, Maryland, USA).*

Finalmente, se empleó el *ABI 3730XL DNA Analyzer (Thermo Fisher Scientific, Massachusetts, USA)* para separar los productos de secuenciación por electroforesis capilar y los electroferogramas resultantes se analizaron usando el *software Staden Package* (http://staden.sourceforge.net).

6.2.5 Detección de variantes *de novo*

Con el objetivo de identificar nuevos genes candidatos asociados a la etiología de los TEA, se identificaron variantes *de novo* localizadas en todos los genes del genoma. Para ello, se aplicaron los mismos criterios de filtrado que se han descrito en el Capítulo 2.

Así pues, las mutaciones *de novo* se definieron como aquellas variantes cuyos genotipos fueron 1/0 o 1/1 en el probando y 0/0 en los progenitores. Solo se consideraron variantes con GQ ≥ 20 y una profundidad de lectura del alelo alternativo ≥ 7. Las variantes que presentaban un alelo o más en la base de datos ExAC (http://exac.broadinstitute.org/) fueron eliminadas. También se eliminaron las variantes separadas por menos de 20 pb y aquellas con un RVIS > 75%, obtenido de ExAC. Es preciso señalar que los datos de ExAC se encuentran disponibles en la actualidad en la base de datos ampliada gnomAD (*Genome Aggregation Database*), sin embargo, cuando comenzó este estudio la base de datos disponible era ExAC. Finalmente, solo se analizaron variantes no sinónimas exónicas y variantes de *splicing*.

Para evaluar si un gen con una mutación *de novo* estaba relacionado con el fenotipo del probando, se llevó a cabo una revisión manual de la literatura científica, así como la revisión de las bases de datos OMIM y SFARI Gene. Así pues, los genes fueron seleccionados como posibles genes candidatos en base a si cumplían uno, o los dos siguientes criterios:

1) Tener un papel demostrado en el neurodesarrollo a través de estudios experimentales.
2) Conferir riesgo para otros TND o trastornos psiquiátricos. (Figura 33).

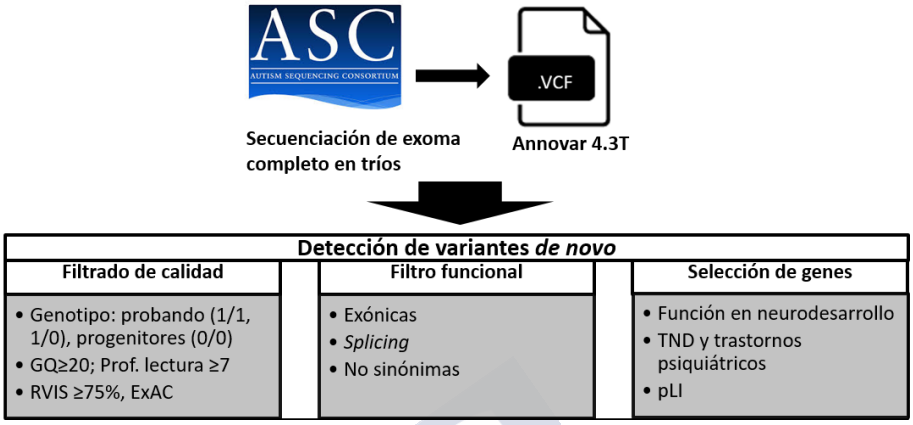

Figura 33. Flujo de trabajo seguido para la detección de variantes *de novo* en genes candidatos.

6.3 RESULTADOS

6.3.1 Descripción de la cohorte

La cohorte gallega de TEA está constituida por 125 tríos, y todos sus probandos tienen un diagnóstico clínico confirmado de TEA. En la mayoría de los casos los probandos procedían de asociaciones de autismo de la Comunidad Autónoma de Galicia.

El número de varones afectos (76.8%) fue superior al de mujeres afectas (*ratio* 3:1). En 120 de los 125 tríos, el TEA fue esporádico, lo que significa que no existía una historia familiar previa de TEA. Sin embargo, 5 familias tenían, además, otro hermano afecto, aunque este no fue secuenciado ni incluido en el estudio. En un 21.6% de los casos el probando tuvo un diagnóstico clínico de TEA de alto funcionamiento. Las manifestaciones clínicas comórbidas más frecuentes en la muestra fueron DI (39.2%), epilepsia (18.4%), trastornos del comportamiento (34.4%), trastornos psiquiátricos (29.6%) y retraso psicomotor (24%). El resto de características clínicas de la muestra se resumen en la Tabla 42.

Todos los individuos de la cohorte se habían sometido previamente a un test de cribado de primera línea (*microarray* para la detección de CNVs y cribado para el Síndrome del X frágil). Sin embargo, no se

detectó en ninguno de ellos una CNV clínicamente significativa ni ninguna variante LoF en *FMR1* (Tabla 42).

	Número de probandos
Total	125
Mujeres	29 (23.2%)
Hombres	96 (76.8%)
Autismo de alto funcionamiento	27 (21.6%)
Autismo esporádico	120 (96%)
Autismo familiar	5 (4%)

Características clínicas

Condición clínica	**Pacientes totales (n=125)**	**Individuos con variants clínicamente relevantes**	**Individuos con variantes de significado incierto**	**Individuos sin hallazgos**
Epilepsia	23 (18.4%)	6 (26.1%)	3 (13%)	14 (60.9%)
Discapacidad intelectual	49 (39.2 %)	15 (30.6%)	8 (16.3%)	26 (53.1%)
Rasgos dismórficos	23 (18.4%)	9 (39.1%)	3 (13%)	11 (47.8%)
Trastorno psiquiátrico	37 (29.6%)	2 (5.4%)	5 (13.5%)	30 (81.1%)
Trastorno del comprotamiento	43 (34.4%)	6 (14%)	6 (14%)	31 (72.1%)
Hipoacusia	14 (11.2%)	1 (7.1%)	4 (28.6%)	9 (64.3%)
Hiperacusia	5 (4%)	0	1 (20%)	4 (80%)
Macrocefalia	10 (8%)	3 (30%)	2 (20%)	5 (50%)
Microcefalia	2 (1.6%)	0	0	2 (100%)
Ausencia de lenguaje	32 (25.6%)	9 (28.1%)	5 (15.6%)	17 (53.1%)
Trastorno del sueño	8 (6.4%)	1 (12.5%)	0	7 (87.5%)
Tortícolis congénita	3 (2.4%)	0	1 (33.3%)	2 (66.7%)
Trastorno dermatológico	19 (15.2%)	5 (26.3%)	6 (31.6%)	8 (42.1%)

Trastorno psicomotor	30 (24%)	8 (26.7%)	8 (26.7%)	17 (56.7%)
Trastorno neurológico	9 (7.2%)	3 (33.3%)	2 (22.2%)	4 (44.4%)
Trastorno endocrino	15 (12%)	3 (20%)	2 (13.3%)	10 (66.7%)
Trastorno otorrinolaringológico	7 (5.6%)	2 (28.6%)	1 (14.3%)	5 (71.4%)
Trastorno esquelético	20 (16%)	7 (35%)	3 (15%)	10 (50%)
Trastorno cardiovascular	3 (2.4%)	2 (66.7%)	0	1 (33.3%)
Trastorno renal	6 (4.8%)	1 16.7%)	1 (16.7%)	4 (66.7%)
Trastornogenitourinario	3 (2.4%)	1 (33.3%)	0	2 (66.7%)
Trastorno gastrointestinal	17 (13.6%)	2 (11.8%)	8 (47.1%)	7 (41.2%)
Trastorno oftalmológico	7 (5.6%)	2 (28.6%)	1 (14.3%)	4 (57.1%)
Test genéticos previos				
Array	125 (100%)			
X-Frágil	125 (100%)			
Otros test	9 (7.2%)			

Tabla 42. Principales características clínicas de la cohorte gallega de TEA (N = 125). Se muestra el número de individuos que padece alguna condición clínica comórbida a los TEA en la muestra total y en los 3 grupos de pacientes: pacientes en los que se hallaron variantes clínicamente relevantes, pacientes en los que se hallaron variantes de significado incierto y pacientes sin hallazgos significativos. También se indica el número de mujeres, hombres, tipo de autismo (familiar, esporádico o de alto funcionamiento) y el número de individuos que se había sometido previamente a algún test genético.

6.3.2 Descripción del listado de genes asociados a TND

De los 261 genes seleccionados (Anexo 6), 100 estaban asociados a trastornos de herencia autosómica dominante (AD), 71 estaban asociados a trastornos de herencia autosómica recesiva (AR), 14 estaban asociados a trastornos tanto de herencia AD como AR, 48 estaban asociados a trastornos de herencia ligada al cromosoma X, y en 28 genes el patrón de herencia del trastorno era desconocido.

De acuerdo al fenotipo principal asociado a cada gen, 9 genes conferían susceptibilidad a TEA, 91 estaban relacionados con síndromes genéticos bien caracterizados, 64 estaban relacionados con DI, 33 estaban asociados a errores en el metabolismo y 19 lo estaban con algún tipo de encefalopatía epiléptica. El resto de los genes que no pertenecían a alguna de estas categorías fueron seleccionados por

presentar altos niveles de evidencia de asociación con TEA según la base de datos SFARI.

6.3.3 Clasificación de variantes y rendimiento diagnóstico

6.3.3.1 Variantes clínicamente relevantes

Se identificaron un total de 18 variantes clínicamente relevantes en 16 genes diferentes. Así, en dos genes, *SCN2A* y *KMT2C*, se identificó más de una variante. 2 de las variantes clínicamente relevantes (11.2%) se encontraron en genes ligados al cromosoma X (*IQSEC2* y *ARHGEF9*) y 16 (88.89%) en genes asociados a trastornos AD.

De las 18 variantes clínicamente relevantes, 8 se clasificaron como variantes patogénicas y 8 como probablemente patogénicas según los criterios de la ACMG. 2 variantes, p.Ser492Phe en *SIK1* y, p.R2Q en *ARHGEF9*, se clasificaron como variantes de significado incierto, posiblemente patogénicas, por presentar criterios de patogenicidad y benignidad contradictorios. La primera, presentó criterios para ser clasificada como benigna, pero era una variante *de novo*. La segunda, se identificó en un gen asociado a un trastorno AR ligado al cromosoma X. Aunque el probando era una mujer, los síntomas solapaban con aquellos descritos en la literatura por lo cual se consideró posible que el fenotipo se explicara por un patrón de inactivación del cromosoma X aleatorio. Dado que esta posibilidad no se pudo explorar experimentalmente, la variante fue clasificada por ese motivo como de significado incierto, posiblemente patogénica (Tablas 43 y 44).

Variante	*Herencia*	*Regla*	*Patogenicidad*	*Explicación*
SLC1A2:c.2T>C	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (pérdida del codon de inicio) que afecta a *SLC1A2*. Este es un mecanismo etiológico conocido de Encefalopatía epiléptica temprana infantil, tipo 41.
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		BP4	Benigna (sugestivo)	Variante benigna dada la clasificación como benigna de 5 predictores *in silico* de patogenicidad (DANN, DEOGEN2, EIGEN, MVP y REVEL) frente a 4 que lo hacen como patogénica (FATHMM-MKL, M-CAP, MutationTaster, SIFT y GERP)
SCN2A:c.4204A>T	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (*nonsense*) que afecta a *SCN2A*. Este es un mecanismo etiológico conocido de Encefalopatía epiléptica temprana infantil tipo 11 y Epilepsia familiar benigna infantil tipo 3.
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 4 predictores *in silico* de patogenicidad (DANN, EIGEN, FATHMM-MKL y MutationTaster) mientras que ninguno la clasifica como benigna.
BRAF:c.1459G>A	Heredada	PM1	Patogénica (probabilidad moderada)	Variante localizada en una región de 61 pb donde se han identificado 10 variantes patogénicas y ninguna benigna (patogenicidad = 100%)
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD

		PM5	Patogénica (probabilidad moderada)	Variante alternativa chr7:140477848 A⇒C (Val487Gly) está clasificada en ClinVar como patogénica
		PP2	Patogénica (sugestivo)	216 de las 232 variantes *missense* que no son de significado incierto en BRAF son patogénicas (93.1%), lo cual está por encima del umbral (51%) y 227 de las 417 reportadas en el gen son patogénicas (54.4%) lo cual está por encima del umbral (12%)
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 8 predictores *in silico* (DANN, DEOGEN2, FATHMM-MKL, M-CAP, MVP, MutationTaster, PrimateAI y REVEL) frente a 2 que lo hacen como benigna (EIGEN y MutationAssessor)
SMAD4:c.1486C>T	*De novo*	PM1	Patogénica (probabilidad moderada)	Variante localizada en una región de 61 pb donde se han identificado 8 variantes patogénicas y 9 benignas (patogenicidad = 47.1%)
		PM2	Patogénica (probabilidad moderada)	En GnomAD se han contado 2 alelos lo cual es inferior a 5, que sería el umbral para el gen dominante *SMAD4*
		PP2	Patogénica (sugestivo)	56 de las 60 variantes *missense* que no son de significado incierto en *SMAD4* son patogénicas (93.3%) lo cual está por encima del umbral (51%) y 180 de las 716 reportadas en el gen son patogénicas (25.1%) lo cual está por encima del umbral (12%)
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 10 predictores *in silico* (DANN, DEOGEN2, EIGEN, FATHMM-MKL, M-CAP, MVP, MutationAssessor, MutationTaster, REVEL y SIFT) frente a 1 que lo hace como benigna (PrimateAI.)
		PP5	Patogénica (probabilidad alta)	ClinVar la clasifica como patogénica con 2 estrellas, con 7 reportes, 15 publicaciones y ningún conflicto.

ARID1B:c.5431G>T	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (*nonsense*) que afecta a *ARID1B*. Este es un mecanismo etiológico conocido del Síndrome de Coffin-Siris 1
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 4 predictores *in silico* (DANN, EIGEN, FATHMM-MKL y MutationTaster) mientras que ninguno lo hace como benigna
TBL1XR1:c.442dup	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (*frameshift*) que afecta a *TBL1XR1*. Este es un mecanismo etiológico conocido de Discapacidad intelectual AD tipo 41 y Síndrome Pierpont
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 1 predictor *in silico* (GERP) mientras que ninguno lo hace como benigna
CTNNB1:c.1930del	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (*frameshift*) que afecta a *CTNNB1*. Este es un mecanismo etiológico conocido de Vitroretinopatía exudativa 7 y Trastorno del neurodesarrollo con diplejía espástica y defectos visuales
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 1 predictor *in silico* (GERP) mientras que ninguno lo hace como benigna

KMT2C:c.6617del	*De novo*	PVS1	Patogénica (probabilidad alta)	Variante nula (*frameshift*) que afecta a *KMT2C*. Este es un mecanismo etiológico conocido del Síndrome de Kleefstra 2.
		PM2	Patogénica (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 1 predictor *in silico* (GERP) mientras que ninguno lo hace como benigna

Tabla 43. Variantes patogénicas detectadas en la cohorte gallega de TEA. Se detallan los criterios proporcionados por la plataforma Varsome que sigue las guías de la ACMG para la clasificación de variantes. PVS1: *pathogenic very strong*; PVS1-4: *pathogenic strong* 1-4; PM1-M6: *pathogenic moderate* 1-6; PP1-5: *pathogenic supporting* 1-5; BP1-7: *benign supporting* 1-7.

Variante	*Herencia*	*Regla*	*Patogenicidad*	*Explicación*
SIK1:c.1475C>T	*De novo*	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		BP1	Benigna (sugestivo)	32 de las 32 variantes *missense* que no son de significado incierto en *SIK1* son benignas (100.0%) lo cual está por encima del umbral (51%) y 138 de las 213 variantes reportadas en el gen son benignas (64.8%) lo cual está por encima del umbral (24%)
		BP4	Benigna (sugestivo)	Variante benigna dada la clasificación como benigna de 7 predictores *in silico* (DANN, DEOGEN2, EIGEN, MVP, MutationTaster, PrimateAI y REVEL) frente a 5 que lo hacen como patogénica (FATHMM-MKL, M-CAP, MutationAssessor, SIFT, GERP)
TRIP12:c.1863C>G	*De novo*	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP2	Patogénica (sugestivo)	1 de 1 variante *missense* que no es de significado incierto en *TRIP12* es patogénica (100%) lo cual está por encima del umbral (51%). 15 de las 24 variantes reportadas en el gen son patogénicas (62.5%) lo cual está por encima del umbral (12%)
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 7 predictores *in silico* (DANN, EIGEN, FATHMM-MKL, M-CAP, MutationAssessor, MutationTaster and SIFT) frente a 4 que lo hacen como benigna (DEOGEN2, MVP, PrimateAI y REVEL)

SCN2A:c.4446+1A>G	*De novo*	PVS1	Pathogenic (probabilidad alta)	Variante nula (*splicing*) que afecta a *SCN2A*. Este es un mecanismo etiológico conocido de Encefalopatía epiléptica temprana infantil tipo 11 y epilepsia familiar infantil tipo 3
		PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 4 predictores *in silico* (DANN, EIGEN, FATHMM-MKL y MutationTaster) mientras que ninguno lo hace como benigna
IQSEC2:c.1663G>A	*De novo*	PM1	Patogénico (probabilidad moderada)	12 de las 24 variantes que no son de significado incierto localizadas en el dominio de la proteína IQSEQ2 son patogénicas (50%) lo cual está por encima del umbral (17.2%)
		PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 6 predictores *in silico* (DANN, FATHMM-MKL, M-CAP, MutationTaster, REVEL y SIFT) frente a 1 que lo hace como benigna (MVP)
		BP1	Benigna (sugestivo)	17 de las 33 variantes *missense* que no son de significado incierto en *IQSEC2* son benignas (51.5%) lo cual está por encima del umbral (51%) y 133 de las 300 variantes reportadas en el gen son benignas (44.3%) lo cual está por encima del umbral (24%)
SMARCA2:c.3292G>C	*De novo*	PM1	Patogénico (probabilidad moderada)	30 de las 34 variantes que no son de significado incierto localizadas en el dominio de la proteína SMARCA2 son patogénicas (88.2%) lo cual está por encima del umbral (17.2%)

		PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP2	Patogénica (sugestivo)	78 de las 94 variantes *missense* que no son de significado incierto en *SMARCA2* son patogénicas (83%) lo cual está por encima del umbral (51%) y 78 de las 199 variantes reportadas en el gen son patogénicas (39.2%) lo cual está por encima del umbral (12%)
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 11 predictores *in silico* (DANN, DEOGEN2, EIGEN, FATHMM-MKL, M-CAP, MVP, MutationAssessor, MutationTaster, PrimateAI, REVEL y SIFT) mientras que ninguno lo hace como benigna
EP300:c.1646G>T	Heredada	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		BP1	Benigna (sugestivo)	48 de las 60 variantes *missense* que no son de significado incierto en *EP300* son benignas (80%) lo cual está por encima del umbral (51%) y 147 de las 319 variantes reportadas en el gen son benignas (46.1%) lo cual está por encima del umbral (24%)
		BP4	Benigna (sugestivo)	Variante benigna dada la clasificación como benigna de 7 predictores *in silico* (EIGEN, MVP, MutationAssessor, MutationTaster, PrimateAI, REVEL y SIFT) frente a 5 que lo hacen como patogénica (DANN, DEOGEN2, FATHMM-MKL, M-CAP y GERP)
KMT2C:c.13289T>G	*De novo*	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD

		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 9 predictores *in silico* (DEOGEN2, EIGEN, FATHMM-MKL, M-CAP, MutationAssessor, MutationTaster, PrimateAI, REVEL y SIFT) frente a 2 que lo hacen como benigna (DANN y MVP)
		BP1	Benigna (sugestivo)	24 de las 28 variantes *missense* que no son de significado incierto en *KMT2C* son benignas (85.7%) lo cual está por encima del umbral (51%) y 51 de las 107 variantes reportadas en el gen son benignas (47.7%) lo cual está por encima del umbral (24%)
CHD8:c.2025-1G>T		PVS1	Patogénica (alta probabilidad)	Variante nula (*splicing*) que afecta a *CHD8*. Este es un mecanismo etiológico conocido de susceptibilidad a TEA 18
		PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 4 predictores *in silico* (DANN, EIGEN, FATHMM-MKL y MutationTaster) mientras que ninguno la clasifica como benigna
CACNA1E:c.934A>G	*De novo*	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PP2	Patogénica (sugestivo)	9 de las 14 variantes *missense* que no con de significado incierto en *CACNA1E* son patogénicas (64.3%), lo cual está por encima del umbral (51%) y 11 de las 42 reportadas en el gen son patogénicas (26.2%) lo cual está por encima del umbral (12%)

		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 11 predictores *in silico* DANN, DEOGEN2, EIGEN, FATHMM-MKL, M-CAP, MVP, MutationAssessor, MutationTaster, PrimateAI, REVEL y SIFT) mientras que ninguno lo hizo como benigna.
ARHGEF9:c.5G>A	*De novo*	PM2	Patogénico (probabilidad moderada)	Variante no encontrada en GnomAD
		PM5	Patogénica (sugestivo)	Variante alternativa chrX:62926209 G⇒A (Arg104Trp) está clasificada en ClinVar como probablemente patogénica
		PP3	Patogénica (sugestivo)	Variante patogénica dada la clasificación como patogénica de 9 predictores *in silico*(DANN, DEOGEN2, FATHMM-MKL, M-CAP, MVP, MutationAssessor, MutationTaster, REVELy SIFT) mientras que ninguno lo hizo como benigna
		PP5	Patogénica (sugestivo)	ClinVar la clasifica como patogénica con una 1 sola estrella

Tabla 44. Variantes probablemente patogénicas y variantes de significado incierto posiblemente patogénicas identificadas en la cohorte gallega de TEA. Se detallan los criterios proporcionados por la plataforma Varsome que sigue las guías de la ACMG para la clasificación de variantes. PVS1: pathogenic very strong; PVS1-4: pathogenic strong 1-4; PM1-M6: pathogenic moderate 1-6; PP1-5: pathogenic supporting 1-5; BP1-7: benign supporting 1-7

6.3.3.2 Rendimiento diagnóstico y clasificación de pacientes

El rendimiento diagnóstico de la cohorte se estimó en un 14,4% al identificarse variantes cínicamente relevantes (patogénicas, probablemente patogénicas o de significado incierto, posiblemente patogénicas) en 18 pacientes. (Figura 34 y Tablas 43 y 44).

En 26 pacientes (20.8%) se identificaron variantes de significado incierto (Figura 34). De ellos, 5 pacientes eran portadores de 2 variantes de este tipo. En todos los casos, la correlación del fenotipo del paciente con las características clínicas asociadas al gen no era completa. Además, en dos pacientes, la información clínica no estaba disponible por lo que fue imposible clasificar correctamente la variante. Todas las variantes fueron heredadas de un progenitor aparentemente sano y en la mayoría de los casos las variantes fueron transmitidas desde una madre sana al probando, indistintamente de su sexo (70.0% vs 20%; test binomial p = 0.029). 8 de las variantes se localizaron en genes asociados a susceptibilidad a TEA: *CHD8, KATNAL2, CNTNAP2, PTCHD1, BRWD3, NLGN3, SHANK2*, y *SLC9A9*.

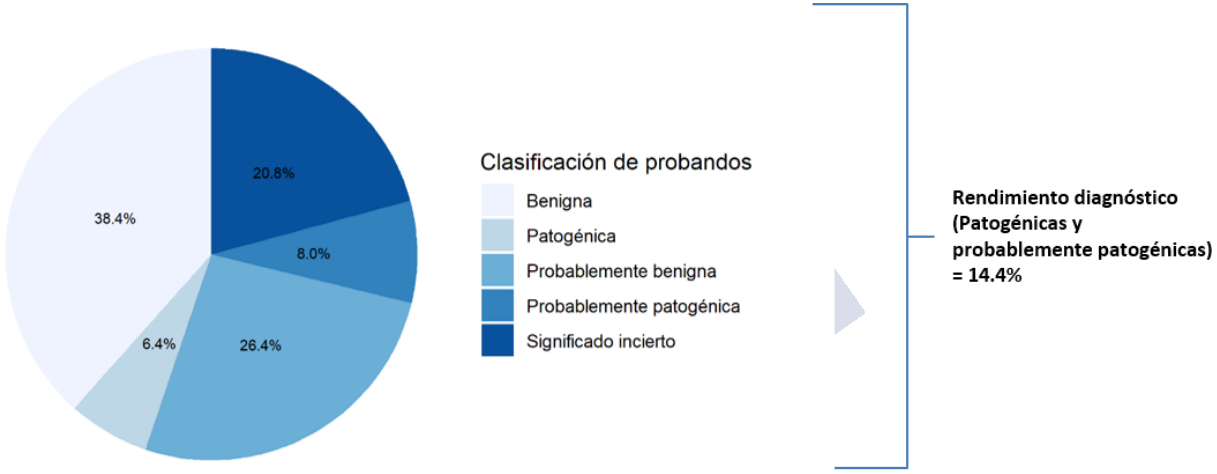

Figura 34. Rendimiento diagnóstico de la secuenciación de exoma completo en una cohorte gallega de TEA (N = 125). La clasificación de los probandos se ha realizado según el tipo de variantes que se identificó en ellos. El rendimiento diagnóstico se ha calculado teniendo en cuenta las variantes patogénicas y probablemente patogénica según los criterios del ACMG.

6.3.4 Correlación genotipo-fenotipo

Entre las variantes con relevancia clínica, 7 de ellas (38.8%) se encontraron en genes asociados a trastornos genéticos sindrómicos que cursan con DI y/o TEA entre otros, 5 variantes (27.8%) se encontraron en genes asociados a encefalopatía epiléptica, 3 variantes (16.7%) se localizaron en genes asociados a DI y solo una se localizó en un *loci* de riesgo para TEA (Tabla 45 y Figura 35).

Los pacientes con mutaciones en genes asociados a trastornos sindrómicos estaban, por lo general, más afectos que el resto de pacientes. Así pues, todos presentaron DI, 4 (57.2%) presentaron retraso o ausencia total de lenguaje, 4 (57.2%) retraso psicomotor, 3 desarrollaron epilepsia y 2 presentaron algún defecto congénito cardiaco.

Solo en un caso, ASD_94, se confirmó molecularmente la sospecha previa de Síndrome de Coffin-Siris. Un caso parecido ocurrió en ASD_64 que había sido diagnosticado de Hipomelanosis de Ito, una condición genética rara caracterizada por parches de piel hipopigmentada, dificultades para el aprendizaje, epilepsia, escoliosis o estrabismo. En ese paciente, se encontró una mutación en *BRAF*, un gen asociado a un conjunto de síndromes de base genética cuyas manifestaciones clínicas son compatibles con la Hipomelanosis de Ito.

En otros casos, algunos de los síntomas asociados al trastorno genético no estaban presentes en el momento de la evaluación o bien las características clínicas que manifestaban los pacientes eran atípicas. Por ejemplo, en ASD_81, se identificó una mutación patogénica en *SMAD4*, que se había relacionado con el Síndrome de Myhre. Sin embargo, el probando no presentaba ni hipoacusia ni defectos en el tracto gastrointestinal, ambos hallazgos clínicos comunes en el síndrome. En el caso de ASD_105, en el cual se detectó una mutación en *KMT2C*, asociado al Síndrome de Kleefstra, las manifestaciones clínicas que presentó el paciente eran más leves que las reportadas en la literatura. Por el contrario, ASD_125, con una mutación en el mismo gen, sí que presentó síntomas más severos que solapan mejor con aquellos que se han descrito en la literatura. Otro caso que llamó la atención fue ASD_84, cuya manifestación clínica más grave fue la presencia de crisis epilépticas refractarias a tratamiento.

Consecuentemente, el paciente había sido sometido a diversos test genéticos a lo largo de su vida, aunque solo se habían examinado genes asociados a encefalopatías epilépticas. Sin embargo, en este proceso diagnóstico no se tuvieron en cuenta otras condiciones clínicas presentes en el individuo, como DI, rasgos dismórficos, retraso psicomotor o ausencia del lenguaje que podrían haber conducido más rápidamente a un diagnóstico clínico. Así pues, el estudio de su exoma reveló una mutación patogénica en *SMARCA2*, asociada al Síndrome de Nicolaides-Baraitser.

Entre los pacientes con mutaciones patogénicas o probablemente patogénicas en genes asociados a encefalopatías epilépticas, llamó la atención que algunos de ellos no hubiesen presentado, hasta la fecha, ningún episodio epiléptico, lo cual explica que ningún clínico los hubiese diagnosticado antes. Este fue el caso de ASD_09, con una mutación en *SIK1*, ASD_58, con una mutación en *CACNA1E*, y ASD_32, con una mutación en *SCN2A*. Sin embargo, todos ellos presentaron DI y el resto de manifestaciones clínicas asociadas a cada gen.

Todos los pacientes que presentaron mutaciones en genes asociados a DI manifestaron DI moderada o grave además de ausencia de lenguaje. El resto de manifestaciones clínicas que presentó cada paciente se correspondían con aquellas asociadas al gen mutado en cada caso.

Probando	Sexo	Fenotipo	Fenotipo OMIM	Variante	Herencia	Clasificación
ASD_09	Mujer	TEA, retraso del lenguaje, TDAH	Encefalopatía epiléptica temprana infantil, tipo 30	NM_173354.3 (SIK1): c.1475C>T p.Ser492Phe	AD, *De novo*	Significado incierto, posiblemente patogénica
ASD_19	Varón	TEA, epilepsia, hipoacuasia	Encefalopatía epiléptica temprana infantil, tipo 41	NM_004171.3 (SLC1A2): c.2T>C p.(Met1?)	AD, *De novo*	Patogénica
ASD_32	Mujer	TEA, DI, ausencia de lenguaje	Encefalopatía epiléptica temprana infantil, tipo 11, Epilepsia benigna familiar infantil, tipo 3	NM_001040142.1 (SCN2A):c.4204A>T p.(Lys1402*)	AD, *De novo*	Patogénica
ASD_40	Mujer	TEA, DI, obesidad	Discapacidad intelectual AD, tipo 49	NM_001348325.1 (TRIP12): c.1863C>G p.(Phe621Leu)	AD, *De novo*	Probablemente patogénica
ASD_64	Varón	TEA, DI, hipomelanosis de Ito	Síndrome cardiofaciocutáneo, Síndrome Leopard 3, Síndrome de Noonan 7	NM_004333.4 (BRAF): c.1459G>A p.(Val487Met)	AD, heredada	Patogénica
ASD_67	Mujer	TEA, DI, TDAH, epilepsia,	Encefalopatía epiléptica temprana infantil, tipo 11,	NM_001040142.1 (SCN2A): c.4446+1A>G p.?	AD, *De novo*	Probablemente patogénica

		ausencia de lenguaje	Epilepsia benigna familiar infantil, tipo 3			
ASD_78	Varón	TEA, DI, epilepsia, TDAH	Discapacidad intelectual ligada al X, 1/78	NM_015075.1 (IQSEC2):c.1663G>A, p.(Gly555Ser)	XLD, *De novo*	Probablemente patogénica
ASD_81	Mujer	TEA, DI, epilepsia, acrodisostosis ausencia de lenguaje, defecto cardiaco congénito	Síndrome de Myhre	NM_005359.3 (SMAD4):c.1486C>T, p.(Arg496Cys)	AD, *De novo*	Patogénica
ASD_84	Varón	TEA, DI, epilepsia, rasgos dismórficos, retraso psicomotor, ausencia de lenguaje	Síndrome Nicolaides-Baraitser	NM_139045.3 (SMARCA2):c.3292G>C, p.(Gly1098Arg)	AD, *De novo*	Probablemente patogénica
ASD_94	Mujer	TEA, DI, epilepsia, rasgos dismórficos, retraso psicomotor, ausencia de lenguaje, defecto	Síndrome Coffin-Siris	NM_020732.3 (ARID1B): c.5431G>T, p.(Glu1811*)	AD, *De novo*	Patogénica

		cardiaco congénito				
ASD_100	Varón	TEA, DI, TDAH, macrocefalia, polidactilia	Síndrome Rubinstein-Taybi tipo 2	NM_001429.3 (EP300):c.1646G>T, p.(Ser549Ile)	AD, heredada	Probablemente patogénica
ASD_105	Mujer	TEA, retraso del lenguaje, rasgos dismórficos	Síndrome de Kleefstra	NM_170606.2 (KMT2C):c.13289T>G, p.(Leu4430Trp)	AD, *De novo*	Probablemente patogénica
ASD_109	Varón	TEA, DI, retraso psicomotor, hipotonía, ausencia de lenguaje	Discapacidad intelectual AD, tipo 41, Síndrome de Pierpont	NM_001321193.1 (TBL1XR1):c.442dup, p.(Met148Asnfs*9)	AD, *De novo*	Patogénica
ASD_120	Varón	TEA, RGD	Susceptibilidad a TEA, tipo 18	NM_001170629.1 (CHD8):c.2025-1G>T, p.?	AD, *De novo*	Probablemente patogénica
ASD_123	Mujer	TEA, DI, microcefalia, diplejía espástica, ausencia de lenguaje	Trastorno del neurodesarrollo con diplejía espástica y defectos visuales	NM_T001098209.1 (CTNNB1):c.1930del, p.(Leu644Phefs*35)	AD, *De novo*	Patogénica
ASD_125	Mujer	TEA, DI, macrocefalia, rasgos dismórficos, retraso psicomotor,	Síndrome de Kleefstra	NM_170606.2 (KMT2C):c.6617del, p.(Pro2206Leufs*33)	AD, *De novo*	Patogénica

		sindactilia, ausencia de lenguaje				
ASD_58	Mujer	TEA, rasgos dismórficos	Encefalopatía epiléptica temprana infantil, tipo 69	NM_000721 (CACNA1E) exon6: c.934A>G p.T312A	AD, *De novo*	Probablemente patogénica
ASD_59	Mujer	TEA, DI, epilepsia, ausencia de lenguaje	Encefalopatía epiléptica temprana infantile, tipo 8	NM_001173480 (ARHGEF9) exon2:c.5G>A p.R2Q	XLR, *De novo*	Significado incierto, posiblemente patogénica

Tabla 45. Pacientes con variantes clínicamente relevantes. Se muestra las características clínicas de los individuos con variantes patogénicas o probablemente patogénicas. También se muestra el fenotipo OMIM asociado a cada a gen, el patrón de herencia de cada trastorno y la clasificación de las variantes según las guías de la ACMG. AD: Autosomal dominant; XLD: X-linked dominant; XLR: X-linked recessive.

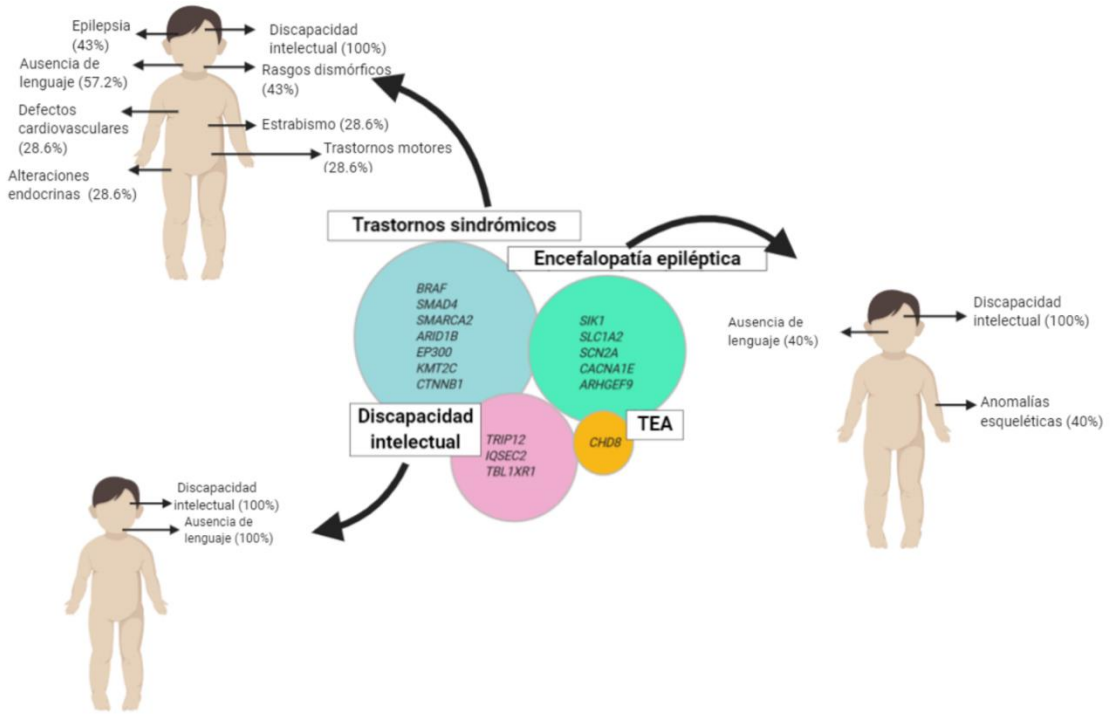

Figura 35. Genes con mutaciones patogénicas o probablemente patogénicas en la cohorte gallega de TEA. Los genes se han clasificado de acuerdo al fenotipo proporcionado por OMIM. Se detallan también las principales características clínicas que comparten los pacientes pertenecientes a cada grupo.

6.3.5 Detección de mutaciones *de novo* en nuevos genes candidatos

Se identificaron 15 mutaciones *de novo* en posibles genes candidatos para TEA en 13 pacientes diferentes (10.4%) (Tabla 46). En ninguno de estos pacientes se había identificado variantes patogénicas o probablemente patogénicas. La selección de genes se realizó principalmente revisando la literatura científica y se tuvieron en cuenta dos criterios principales: la participación del gen en algún proceso esencial durante el neurodesarrollo y/o su asociación previa con algún TND o algún trastorno psiquiátrico. Sin embargo, no existen, por el momento, estudios funcionales o un número suficiente de individuos no emparentados con clínica similar y mutaciones en el mismo gen, que prueben la vinculación de estos genes con TEA. Por esa razón, no se pudieron considerar las mutaciones detectadas en ellos como patogénicas, y, por tanto, ninguna de ellas se tuvo en cuenta a la hora de estimar el rendimiento diagnóstico en la cohorte.

Solo uno de los genes, *KDM5B*, estaba incluido en la base de datos OMIM, pues existe evidencia de que las mutaciones homocigotas o heterocigotas compuestas en él causan Discapacidad Intelectual de herencia AR, tipo 65. El gen fue seleccionado como candidato, pues se han descrito también mutaciones *missens*e y *nonsense de novo* heterocigotas en pacientes con TEA. 6 de los genes seleccionados (*DLGAP1*, *TAOK1*, *CSMD1*, *CHRNA7*, *KDM5B* y *SSPO*) estaban incluido en la base de datos SFARI Gene. Sin embargo, solo *TAOK1* tenía un *score* SFARI de 1, lo que significa que existe un nivel de evidencia muy alto como para considerarlo gen de riesgo en los TEA. Este gen se añadió recientemente a la base de datos de SFARI lo que explica que no estuviera en nuestro listado inicial de genes. En 6 genes se identificaron mutaciones *nonsense* o de *splicing*. Se asume pues, que esas variantes tendrían un alto impacto funcional, al provocar una ausencia total del producto génico. No obstante, solo *TAOK1* tiene un pLI = 1, lo que significa que las variantes LoF son un mecanismo conocido de enfermedad en ese gen. Para el resto de los genes no había información sobre su pLI, salvo para *KDM5B*, cuyo pLI es igual a 0.

Probando	Gen	Variante	Evidencias	Función	Enfermedad
ASD_01	*KIAA1107*	NM_015237: exon6:c.1271C>G p.S424X	pLI = NA OMIM = NA SFARI = no	*KIAA1107* codifica para la proteína *APache* (*AP2-interacting chathrin endocytosis*), la cual se expresa en axones y en terminales presinápticas. Se requiere su correcto funcionamiento para mantener el reciclaje de las vesículas sinápticas en las neuronas maduras. Su papel parece fundamental para que las neuronas *in vitro* e *in vivo* se desarrollen correctamente[228]	
ASD_02	*HCN2*	NM_001194: exon2:c.1025G>C p.R342P	pLI = 0.83 OMIM = NA SFARI = no	*HCN2* codifica para la isoforma 2 del canal HCN (*hiperpolarization-activated cyclic nucleotide-gated*). *HCN2* se expresa principalmente en cerebro y contribuye a mantener rítmica la actividad eléctrica espontánea[229]	La disrupción del canal HCN2 da lugar a hiperexcitabilidad y a un potencial de acción incontrolado en la célula, lo cual predispone a epilepsia. Tanto las mutaciones LoF, como de ganancia de función, se han descrito como mecanismos que causan epilepsia espontanea generalizada[230,231]

ASD_10	*NPAS3*	NM_001164749: exon8:c.886G>A: p.V296M	pLI = 0.98 OMIM = NA SFARI = no	*NPASS* codifica para un factor de transcripción de la familia *Helix-PAS* que se expresa en el sistema nervioso en desarrollo. Se encarga de regular la expresión de genes que interactúan con ARNT (*Aryl hydrocarbon receptor nuclear translocator)*[232]. Este gen parece haber contribuido a la evolución del cerebro humano. Así pues, un metaanálisis reciente reportó que *NPAS3* contiene el mayor *clúster* de regiones aceleradas no codificantes del genoma humano[233]	Estudios GWAS han identificado variantes en este gen asociadas a esquizofrenia y trastorno bipolar [234]. Modelos de ratón *knockout* para el gen manifiestan anomalías cerebrales y alteraciones en su comportamiento[235]
ASD_11	*DLGAP1*	NM_001242765: exon9:c.1820T>C p.V607A	pLI = NA OMIM = NA SFARI = *score* 2	*DLGAP1* codifica para una proteína de anclaje del complejo de densidad postsináptico de neuronas glutamatérgica. Se requiere para mantener la correcta organización estructural del complejo, así como para la interacción con otras proteínas de anclaje[236]	Variantes raras en *DLGAP1* se han asociado a TEA y a TOC[181,237]. Modelos de raton *knockout* para el gen muestran déficits en el comportamiento social[238]
ASD_31	*ARHGEF5*	NM_005435: exon2:c.1381C>T p.Q461X	pLI = NA OMIM = NA; SFARI = no	*ARHGEF5* codifica para el factor RHO 5 de intercambio de nucleótidos de guanina. Así pues, promueve el intercambio	Su asociación con el neurodesarollo no se ha estudiado en detalle, pero

				GDP/GTP en proteínas de la familia Rho, permitiendo así su activación. Por lo tanto, juega un papel fundamental en multitud de procesos celulares, controlando la activación de esta familia de proteínas[239]	mutaciones homocigotas en miembros de la misma familia (*ARHGEF2*) causan DI y malformaciones cerebrales[240].
ASD_35	*ELFN1*	NM_001128636: exon2:c.570C>A p.N190K	pLI = 0.7 OMIM = NA SFARI = no	*ELFN1* codifica para una proteína postsináptica (*extracellular leucine-rich repeat fibronectin containing 1*) que actúa en el hipocampo regulando el reclutamiento de interneuronas[241]	Modelos de ratón *knockout* para el gen manifiestan crisis epilépticas, hiperlocomoción e hiperactividad[242]. Mutaciones raras *missense* se han identificado en pacientes con TEA y TDAH[243].
ASD_35	*TAOK1*	NM_020791: exon3:c.136C>T p.R46X	pLI = 1 OMIM = NA SFARI = *score* 1	*TAOK1* codifica para la proteína *Serine/threonine kinase* que se encarga de mantener la estabilidad de los microtúbulos y su correcto anclaje, para asegurar la correcta segregación cromosómica tanto en la mitosis como durante la interfase celular[244]	Variantes *de novo* en *TAOK1* se han asociado recientemente a TND[245]
ASD_44	*CSMD1*	NM_033225: exon21:c.3161G>T p.S1054I	pLI = NA; OMIM = NA; SFARI = *score* 3	*CSMD1* codifica para la proteína *CUB And Sushi Multiple Domains 1*, un regulador del complemento de activación e	Variantes comunes en *CSMD1* se han asociado con funciones

				inflamación en el sistema nervioso central[246]	cognitivas y con esquizofrenia[247,248]
ASD_56	*FURIN*	NM_002569: exon3:c.208C>T p.R70X	pLI = 1 OMIM = NA SFARI = no	*FURIN* codifica para la enzima *Paired Basic Amino Acid Residue-Cleaving*, una convertasa que regula la activación de multitud de proteínas pertenecientes a la ruta constitutiva exocítica y endocítica. Su homólogo en *c.elegans* se necesita para promover la formación de neuronas PVD, además de regular la formación de otros tipos neuronales[249]	Se han encontrado variantes eQTLs en el gen asociadas a esquizofrenia. En modelos de pez cebra, la disminución de su expresión se asocia con una reducción del tamaño de la cabeza del 24%. El *knockout* en progenitores neurales humanos provoca una migración anómala[250]
ASD_61	*SP2*	NM_003110: exon7:c.1750C>T p.R584C	pLI = NA OMIM = NA SFARI = no	*SP2* codifica para el factor de transcripción *zinc-finger transcription factor specificity protein 2*, que regula el ciclo celular en células madre neurales y en progenitores neurales intermedios. Su papel es esencial durante la neurogénesis en periodos embrionarios y postnatales[251].	
ASD_87	*CHRNA7*	NM_001190455: exon1:c.127C>T p.P43S	pLI = NA OMIM =NA SFARI = *score* 2	*CHRNA7* codifica para la subunidad α7 del receptor nicotínico de acetilcolina. Regula la liberación tanto del neurotransmisor inhibitorio	Los fenotipos neuropsiquiátricos asociados a CNVs en la región 15q13.3 se explican

				GABA, como el neurotransmisor excitatorio glutamato, en el hipocampo[252]. También media la movilización extracelular de Ca2+, teniendo un papel crítico en la plasticidad sináptica[253]	probablemente por cambios en la dosis génica de *CHRNA7*[254,255]
ASD_114	*ATRNL1*	NM_207303: exon28: c.4006C>G p.P1336A	pLI = NA OMIM = NA SFARI = no	*ATRNL1* codifica para la proteína *Attractin Like 1*. Su función no se conoce en detalle pero podría jugar un papel importante en la regulación de la homeostasis energética participando en rutas de señalización de la melanocortina[256]	Una deleción *de novo* afectando al gen *ATRNL1* fue identificada en un paciente con alteraciones cognitivas, rasgos autistas, ataxia, defectos cardiacos y rasgos dismórficos[257]
ASD_114	*KDM5B*	NM_001314042: exon14:c.1816C>T p.R606X	pLI = 0 OMIM = #618109 SFARI = *score* 2	*KDM5B* codifica para la enzima demetilasa 5B específica de lisina. Es una demetilasa de histonas H3K4me2 y H3K4me3, y favorece la inhibición de la transcripción. También participa en el mantenimiento de la estabilidad del genoma y en la reparación del ADN[257]	Las mutaciones homocigotas y heterocigotas compuestas causan un síndrome caracterizado por retraso en el desarrollo, rasgos faciales dismórficos y camptodactilia. Sin embargo también se han descrito variantes *de novo missense* y *nonsense* heterocigotas

					en pacientes con TEA[258]
ASD_118	*SSPO*	NM_198455: exon50:c.7451A>T p.D2484V	pLI = 0 OMIM =NA; SFARI = *score* 3	*SSPO* codifica para la proteína *Spondin* la cual tiene un papel en el desarrollo del sistema nervioso central. Incrementa la longitud de las dendritas y favorece su extensión de manera dosis dependiente[259]	Variantes *de novo missense* se han identificado en pacientes con TEA[92]
ASD_121	*PANX2*	NM_001160300 exon1:c.181G>A p.V61M	pLI = 0.01; OMIM = NA SFARI = no	*PANX2* codifica para la proteína *Pannexin 2*. Esta proteína juega un papel importante en el desarrollo neural durante periodos embrionarios, postnatales y adultos. Su expresión modula el tiempo de diferenciación neuronal en el hipocampo[260]	

Tabla 46. Variantes *de novo* detectadas en genes candidatos. El pLI para cada variante se calculó a partir de la base de datos de ExAC. La base de datos SFARI se consultó para averiguar si alguno de los genes había sido descrito previamente en pacientes con TEA**.** Se consultó la literatura científica para determinar el papel en el neurodesarrollo de cada gen y su implicación previa con algún TND y/o trastornos psiquiátricos.

6.4 DISCUSIÓN

El rendimiento diagnóstico de la secuenciación de exoma completo en una cohorte gallega de TEA, formada por 125 tríos, ha sido de 14.4%. En la literatura científica se ha reportado gran variabilidad con respecto a los valores estimados de rendimiento diagnóstico entre diferentes estudios de secuenciación de exoma completo (8.4-25%)[169,261]. El valor obtenido en este estudio, se sitúa, por tanto, dentro del rango definido por estudios previos. Esta enorme variabilidad puede deberse principalmente a dos razones: la heterogeneidad clínica existente entre las cohortes de TEA lo que inevitablemente conlleva una carga genética diferente entre estudios y a la utilización de diferentes estrategias de secuenciación (secuenciación del probando o secuenciación del trío completo).

En el presente estudio se ha seguido la estrategia basada en la secuenciación del exoma del trío al completo (probando y ambos progenitores sanos). Esta técnica presenta un rendimiento diagnóstico superior que la secuenciación del probando solo, debido a que permite detectar con eficacia y rapidez variantes *de novo* y variantes heterocigotas compuestas[262]. Las variantes *de novo* confieren un alto riesgo a nivel individual, ya que normalmente tiene un efecto deletéreo en el gen. Además, se ha demostrado que tienen un papel importante en la etiología de los TEA esporádicos[90–92,94]. Un 96% de los individuos incluidos en nuestra cohorte presentaban TEA esporádico. Consecuentemente, la mayoría de las mutaciones que fueron clasificadas como patogénicas resultaron ser *de novo* (88.9%) y se identificaron en genes relacionados con trastornos de herencia AD. Por otro lado, muchas variantes heterocigotas muy raras (MAF < 0.01) fueron clasificadas como variantes de significado incierto o probablemente benignas al conocer que fueron heredadas de progenitores aparentemente sanos. Así pues, aunque esta estrategia conlleva el coste adicional que supone secuenciar también a los progenitores, resulta útil a la hora de alcanzar un diagnóstico con rapidez[94].

En lo que se refiere a las características clínicas de la muestra, se ha reportado que la probabilidad de recibir un diagnóstico genético en TEA incrementa en aquellos individuos que están más gravemente

afectos[169,261,263]. En un metaanálisis de estudios de exoma completo se estimó un rendimiento diagnóstico del 31% para aquellos TND que se manifestaron sin otra comorbilidad asociada. Sin embargo, el rendimiento diagnóstico para aquellos casos que también manifestaron otras condiciones clínicas neurológicas o sistémicas fue del 53%. Además, también se comprobó que el rendimiento diagnóstico para TEA era del 16% y del 37% para casos complejos con DI y/o TEA[170]. En línea con estas publicaciones, la mayoría de los pacientes que recibieron un diagnóstico en esta cohorte estaban gravemente afectados por trastornos genéticos sindrómicos (38.8%), encefalopatías epilépticas (27.8%) o síndromes que cursan con DI (16.7%). Solo en un individuo se identificó una mutación en un gen estrictamente asociado a TEA (*CHD8)*. En particular, los pacientes con mutaciones patogénicas en genes asociados a síndromes genéticos (*BRAF, SMAD4, SMARCA2, ARID1B, EP300, KMT2C* y *CTNNB1*) se caracterizaron por presentar manifestaciones clínicas muy complejas, lo que supone una dificultad para los clínicos a la hora de proponer un diagnóstico de sospecha en base a las características clínicas que observan. Aunque en todos ellos se presuponía algún trastorno de base genética, solo en un caso la secuenciación del exoma completo confirmó molecularmente una sospecha previa de Síndrome de Coffin-Siris. En el resto de casos, la falta de especificidad entre síndromes, el hecho de que algunas manifestaciones clínicas no estuvieran presenten en el momento de la evaluación y la presencia de características clínicas anómalas, explican que los pacientes no se hubieran diagnosticado antes.

Entre los pacientes con variantes patogénicas en genes asociados a encefalopatías epilépticas (*SIK1, SLC1A2, SCNA2, CACNA1E,* y *ARHGEF9),* 4 de ellos no habían presentado ningún episodio epiléptico hasta la fecha, siendo el TEA el principal diagnóstico en todos ellos. Sin embargo, las variantes fueron clasificadas como probablemente patogénicas si el fenotipo del paciente solapaba solo parcialmente con el fenotipo aportado por la base de datos OMIM para ese gen. Así pues, se tuvo en cuenta que se necesita un número elevado de casos con mutaciones en un gen, correctamente fenotipados, para concluir una descripción adecuada de un trastorno asociado a un gen. Por ejemplo, en ASD_32 se identificó una variante *de novo nonsense* en *SCN2A* que

codifica para la subunidad alpha tipo II del canal de sodio dependiente de voltaje. Este gen se asocia con encefalopatía epiléptica temprana infantil tipo 11 en la base de datos OMIM. No obstante, existe también una alta evidencia de asociación con TEA, detectándose de manera recurrente mutaciones *de novo* en este gen en individuos pertenecientes a cohortes de TEA[104,105]. En estas cohortes, sin embargo, un porcentaje muy bajo de individuos presenta también epilepsia. Se ha reportado entonces, que las variantes *missense*, que causan una ganancia de función, son frecuentemente la causa de la epilepsia mientras que las variantes *nonsense*, que interrumpen la función del canal, son factor de riesgo para TEA[264]. En la misma línea, en los pacientes ASD_09 y ASD_58 se identificaron mutaciones de *novo missense* en los genes *SIK1* y *CACNA1E* respectivamente. Ambos genes se han relacionado, muy recientemente, con encefalopatías epilépticas (OMIM #616341, #618285), lo que explica que solo un pequeño número de casos se hayan descrito, y, por tanto, las características clínicas de ambos trastornos no están claras todavía. Tampoco se han descubierto aún, los mecanismos biológicos que explican su patogénesis[265,266]. En conclusión, estos resultados sugieren que, a medida que se van identificando y caracterizando clínicamente nuevos casos, se van ampliando las características clínicas asociadas a un síndrome.

La presentación clínica tan compleja que presentan, en general, los pacientes con TEA, justifica el análisis de genes implicados en la etiología de diferentes trastornos o enfermedades. Por ese motivo, en el presente estudio se han seleccionado genes asociados a TND incluso si el autismo no estaba descrito entre sus características principales. Sin embargo, sí que ha habido un estricto proceso de selección para incluir genes cuya asociación con la enfermedad estaba demostrada. Así pues, la selección de genes no se ha realizado en base a la información de los últimos estudios genéticos de TEA, de carácter académico, en los cuáles se han reportado numerosos genes asociados[104]. En su lugar, se ha optado por la inclusión de genes para la cuales existen estudios funcionales que evidencian su asociación con una enfermedad o bien existe un número de casos suficiente, con clínica similar, en los cuales se han encontrado mutaciones en el mismo gen. Esta manera de proceder, reduce enormemente el número de hallazgos incidentales o el

número de variantes de significado incierto detectadas en la muestra[165]. Merece la pena mencionar en este punto, que el análisis de variantes *de novo* localizadas en todos los genes del genoma, no detectó ninguna variante clínicamente significativa adicional en la muestra, lo que demuestra que la selección de genes fue adecuada para las características de la cohorte.

Aunque en efecto, el análisis de variantes *de novo* no incrementó el rendimiento diagnóstico, sí que ha proporcionado hallazgos interesantes, reportando nuevos genes candidatos como *TAOK1*, cuya asociación con TEA aún no está clara. En 8 individuos no emparentados entre sí y con características clínicas muy diversas entre ellos se han identificado variantes raras *de novo* en *TAOK1*[245]. Sin embargo, el número de casos es aún insuficiente para definir un fenotipo asociado al gen, y probablemente ese sea el motivo por el cual, este gen no se asocia todavía a ninguna enfermedad en OMIM. Este hallazgo, por tanto, requiere ser replicado en futuros estudios genéticos y ser validado funcionalmente para determinar firmemente su asociación a los TEA u otros TND. Además, ninguna de las mutaciones *de novo* en genes candidatos se detectó en individuos con mutaciones patogénicas o probablemente patogénicas, lo que sugiere que estas puedan tener un efecto deletéreo y no ser compatibles con la vida en presencia de un segundo evento genético altamente penetrante. Se espera, por tanto, que alguna de estas variantes pueda ser clasificada como patogénica en un futuro próximo a medida que incrementa la información existente en bases de datos. Así pues, se calcula que cada año se añaden aproximadamente 266 nuevos trastornos a la base de datos OMIM y que el número de variantes patogénicas en genes causales de enfermedad se incrementa en 9210 variantes y 241 genes cada año en la base de datos HGMD (de sus siglas en inglés, *Human Gene Mutation Database*)[267]. Este hecho, unido a la mejora de las técnicas de análisis, puede mejorar el rendimiento diagnóstico entre un 10-15%, si los datos se reanalizan cada 2 o 3 años[267–269]. Por ejemplo, el estudio *The Deciphering Developmental Disorders (DDD)* fue capaz de diagnosticar a 182 nuevos pacientes cuando reanalizaron los datos. El DDD, estimó pues, que un 35% de los nuevos diagnósticos se debieron a la identificación de nuevos genes asociados a enfermedad, un 23% a

las mejoras en el análisis y un 8% a la inclusión de nuevos métodos analíticos que no se incluyeron en el estudio previo[267].

El uso de técnicas de NGS supone aún un reto en la práctica clínica pues las guías actuales de la ACMG todavía recomiendan el *array* como herramienta de cribado de primera línea en pacientes con TND[154]. Sin embargo, en el estudio de Srivastava *et al.*, se recomendó la secuenciación del exoma completo como herramienta de primera línea seguida del *array* si el paciente no era diagnosticado, tras realizar una comparativa del rendimiento diagnóstico de ambas técnicas[170]. En este estudio también se ha demostrado que la detección de variantes patogénicas o probablemente patogénicas es superior con técnicas de NGS en comparación con arrays cromosómicos (14.4% vs 8%). Estudios de costo-eficacia también han apoyado la implementación de la secuenciacón de exoma completo en la neurología pediátrica, al comprobar que otros test de rutina, como el *array*, el cribado de X frágil o la resonancia magnética, se omitirían si el diagnóstico se alcanza de manera precoz[270]. Además, en la rutina de los laboratorios de diagnóstico genético se comienzan a utilizar análisis que incluyen algoritmos específicos para la detección conjunta de variantes estructurales y SNVs a partir de los datos de secuenciación de exoma completo. La realización de una sola técnica de diagnóstico (exoma completo) frente a la estrategia de exoma seguida del *array*, supone una reducción importante del coste invertido[271,272]. Un 94.4% de los pacientes de esta cohorte tenían hecho algún test genético de cribado previo. Esto significa que habían permanecido sin diagnóstico durante mucho tiempo a pesar de haberse sometido a evaluaciones genéticas de primera línea.

En este estudio se ha secuenciado el exoma completo de una cohorte de TEA compuesta por 125 tríos con el objetivo de obtener un diagnóstico genético en los probandos. Los resultados obtenidos confirman la utilidad de esta aproximación. en casos de TEA y otras comorbilidades clínicas asociadas, en los que el diagnóstico clínico no está claro, pero existe una sospecha de un trastorno genético de base. Se podría, esperar, incluso, que el rendimiento diagnóstico fuera superior al que se ha obtenido en la cohorte (14.4%) si la secuenciación de exoma se hubiese considerado como la primera opción en estos

casos. Además, conllevaría una reducción del tiempo para obtener un diagnóstico, limitando así la inversión económica que conllevan otras pruebas diagnósticas. Hacen falta, sin embargo, estudios clínicos adicionales en los cuales se recojan en detalle datos clínicos y se incluyan datos de tipo coste-eficacia para definir correctamente la estrategia a seguir en casos de TEA, en la práctica clínica.

6.5 CONCLUSIONES

El rendimiento diagnóstico de la secuenciación de exoma completo, usando una aproximación de trío completo en una cohorte gallega de TEA ha sido del 14.4 %, siendo muy superior al rendimiento de los *arrays* cromosómicos (7.5%). De acuerdo a los resultados de este estudio, parece aconsejable el uso de la secuenciación de exoma completo como primera herramienta diagnóstica, seguida del *array* en aquellos casos que no reciban un diagnóstico genético.

La mayoría de los pacientes que recibieron un diagnóstico molecular presentaban un trastorno genético sindrómico, una encefalopatía epiléptica o un síndrome que cursa con DI.

La identificación de mutaciones *de novo* en todos los genes del genoma, no ha incrementado el rendimiento diagnóstico de la técnica, pero si ha reportado nuevos genes candidatos, como *TAOK1*, que en un futuro podrían ser considerados genes de riesgo para TEA.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

7 DISCUSIÓN GENERAL

Los TEA, son un grupo de TND, que se definen por déficits en la comunicación social y por la presencia de comportamientos restringidos y repetitivos. Aunque en efecto, la comunicación y el comportamiento, son las principales áreas afectadas, los TEA son muy heterogéneos fenotípicamente entre sí. Esta complejidad fenotípica, viene dada por la intensidad de los síntomas principales, el sexo, el nivel de funcionamiento cognitivo y las comorbilidades asociadas[2].

La heterogeneidad fenotípica que caracteriza a los TEA, se acompaña, a su vez, por una elevada heterogeneidad genética. Así pues, se estima que cientos de genes estén implicados en su etiología[104]. Además, se han identificado toda clase de variantes genéticas, tanto raras como comunes, que confieren riesgo para este TND. Por ese motivo, el estudio de su arquitectura genética en la últimas décadas ha estado marcado por el extenso debate de qué hipótesis genética se ajusta mejor: "enfermedad común-variante común" o "enfermedad común-variante rara"[273]. La primera, defiende que el riesgo para este TND se debe a múltiples alelos de bajo riesgo, comunes en la población. La segunda, por el contrario, atribuye ese riesgo a un amplio número de variantes raras, que tienen un alto impacto funcional en el individuo. Hoy en día, se sabe que ambas hipótesis encajan solo parcialmente. Así pues, el modelo genético aceptado en este TND es aquel en el que, tanto las variantes comunes como las variantes raras, están implicadas en su etiología. Ese es el motivo por el cual en este manuscrito se ha querido abordar el estudio de ambos tipos de variantes en los capítulos 1 y 2 respectivamente.

La variación común tiene un papel muy importante en la etiología de los TEA, pues, tal y como se ha explicado anteriormente en la Introducción, explica aproximadamente un 50% de su heredabilidad[25]. Sin embargo, muchos de los estudios GWAS llevados a cabo en este TND han fracasado a la hora de obtener señales significativas, con la

excepción del último trabajo del PGC, en el cual se identificaron 93 SNPs estadísticamente significativos, de los cuales 53 fueron replicados[48]. La dificultad del estudio de la variación común en los TEA, se explica por la existencia de cientos de miles de SNPs asociados, siendo el efecto de cada uno de ellos, a nivel individual, muy pequeño. Se requieren, por tanto, cohortes muy grandes, que alcancen la potencia estadística suficiente, para poder detectar estas señales. Por otro lado, los resultados son difíciles de interpretar en un contexto biológico, puesto que la mayoría de las señales que han resultado asociadas se localizan en regiones no codificantes.

Las herramientas GBA se utilizan de manera complementaria a los GWAS y facilitan la interpretación de sus resultados, permitiendo identificar genes asociados a la enfermedad, en lugar de SNPs. Para ello, los GBA obtienen un estadístico para cada uno de los genes del genoma al considerar todos los SNPs que cubren el gen. De esa manera, esta estrategia tiene en cuenta, no solo los SNPs significativos, sino también aquellos de efecto moderado cuyo p-valor se acerca al umbral de significación estadística sin superarlo. Gracias a que las herramientas GBA consideran el gen como la unidad básica, posteriormente, se pueden estudiar las rutas biológicas en la que estos genes intervienen.

En el capítulo 1 de esta tesis se ha comprobado como el uso de un algoritmo de GBA, facilita la interpretación de los resultados de un GWAS, gracias a la identificación de genes de riesgo en los TEA. Sin embargo, llama la atención que en este, y otros trabajos, los genes que resultan asociados apenas solaban con los genes que se han identificados en estudios de secuenciación de exoma y genoma completo[46,48]. Este hallazgo, pone de manifiesto que el estudio de la variación común y la variación rara identifica genes diferentes que probablemente están involucrados en rutas biológicas independientes.

Recientemente, en el campo de la genética de las enfermedades complejas, se ha propuesto un nuevo modelo genético, llamado omnigénico, que, podría explicar, al menos en parte, este fenómeno. Según este modelo, todos los genes que se expresan en un tejido que se relaciona con una enfermedad, intervienen en la patogénesis de la misma. No obstante, el modelo no otorga el mismo peso a todos los genes y diferencia dos tipos: centrales y periféricos. Los genes centrales

constituyen un pequeño número de genes con un papel biológico interpretable en la enfermedad. Así pues, el producto génico de los genes centrales (proteína o ARN en caso de un gen no codificante) interviene de manera directa en procesos celulares, que, de estar alterados, producen un fenotipo particular. Los genes periféricos, en cambio, constituyen la mayor parte de los genes asociados a una enfermedad, pero no tienen una vinculación directa con ella. Se trata de genes que ejercen una función meramente reguladora sobre los genes principales en aquellos tejidos y periodos del desarrollo que se asocian con la enfermedad[274].

En el caso de los TEA, la distinción entre genes centrales y periféricos debe hacerse con cautela, pues existen algunas discrepancias con el modelo si se tiene en cuenta nuestro conocimiento actual sobre los TEA. Así pues, el modelo onmigénico, acepta que los genes que portan variación rara y los genes que se encuentran cerca de los SNPs que resultan asociados en un GWAS, podrían ser genes centrales. Sin embargo, esto no parece ocurrir así en los TEA, pues, de ser cierto, esperaríamos un solapamiento mayor entre los genes identificados en estudios GWAS y los genes identificados en estudios de secuenciación de exoma, donde se ha priorizado el estudio de la variación rara. Además, se ha reportado consistentemente que la variación común asociada a los TEA se localiza fundamentalmente en regiones reguladoras activas transcripcionalmente en cerebro fetal. Por todo ello, consideramos que es más aceptable considerar los genes cercanos a las señales significativas de un GWAS como periféricos, en lugar de centrales. No obstante, la definición de gen periférico en los TEA es también controvertida, dado que una de las rutas biológicas clave que interviene en la fisiopatología de los TEA es la regulación, a través de mecanismos pretranscripcionales y postranscripcionales, del neurodesarrollo[104,105]. Por tanto, los llamados genes periféricos tendrían mayor relevancia en los TEA en comparación con otras enfermedades de base genética conocida. En favor de esta teoría, el propio modelo omnigénico, reconoce, que en algunas enfermedades puede haber genes periféricos de mayor efecto, que regulan de manera coordinada muchos genes centrales (*peripheral master regulator*). Dichos genes se podrían corresponder también con las señales significativas de un GWAS[275].

Los genes que han resultado asociados en el estudio de GBA del capítulo 1, por tanto, podrían ser interpretados como genes periféricos vinculados a la patogénesis de los TEA. Este listado inicial, fue ampliado al añadir sus principales interactores, empleando la herramienta FunCoup. Esta estrategia, cobra aún más sentido teniendo en cuenta el modelo omnigénico, pues se presupone que los genes que interactúan entre sí se expresan en el mismo tejido y participan conjuntamente en rutas biológicas, de tal manera que estos nuevos genes también podrían ser considerados genes periféricos relacionados con los TEA. Así pues, con esta aproximación, no solo se han identificado posibles genes de riesgo adicionales, sino que la anotación funcional que se llevó a cabo de todos ellos permitió caracterizarlos biológicamente e indagar en posibles mecanismos biológicos relacionados con la patogénesis de los TEA.

Cabe destacar, sin embargo, una limitación importante del algoritmo GBA empleado en el estudio. Para obtener el estadístico de cada gen, PASCAL asigna los SNPs a su gen correspondiente en base a su localización física, obviando su papel regulador. Aunque esta estrategia es acertada para SNPs localizados en regiones codificantes, se debe tener en cuenta que los SNPs en regiones no codificantes pueden interactuar con genes localizados muy lejos físicamente. Las interacciones que se establecen entre ellos, además, son específicas tanto del tejido como de cada periodo del desarrollo[276]. El hecho de que el mapeo de los SNPs se haya realizado, de manera exclusiva, en función de la distancia física, podría haber dificultado la identificación de otros genes que también podrían ser relevantes en la patogénesis de los TEA.

Posteriormente a la finalización del estudio que corresponde al capítulo 1 de esta tesis, fue diseñada una herramienta, H-MAGMA, cuyo objetivo es solventar esta limitación. La principal ventaja que ofrece H-MAGMA, con respecto a los algoritmos convencionales como MAGMA, VEGAS o PASCAL, es la capacidad que tiene de mapear los SNPs en genes según evidencias funcionales genómicas. En concreto, la asignación de cada SNP localizado en regiones no codificantes al gen correspondiente, se realiza teniendo en cuenta la interacción existente entre ambos medida por Hi-C, en tejidos y etapas

del desarrollo específicos, y solo los SNPs exónicos se asignan según la distancia física. El empleo de este nuevo algoritmo usando como *input* los *summary statistics* de diferentes GWAS, reveló que esta nueva manera de mapear los SNPs introduce cambios relevantes con respecto a los resultados obtenidos con algoritmos convencionales. La razón es que se estima que muy pocos SNPs en regiones no codificantes interactúan con el gen más próximo (los SNPs intrónicos solo lo hacen en un 20% de los casos y los intergénicos tan solo en un 5%)[277].

Al emplear H-MAGMA, usando como *input*, los *summary statistics* que fueron empleados tanto en el capítulo 1 de esta tesis como en el trabajo de Grove *et al.*, se identificaron nuevos genes asociados. De ellos, algunos se identificaron también con PASCAL, lo que permitió confirmar que son genes candidatos para TEA[277]. Uno de estos genes fue *XRN2*, que además llamó especialmente la atención por la cantidad de interactores que identificó FunCoup. Aunque esto pueda implicar introducir un sesgo en los estudios de enriquecimiento en rutas biológicas que se hicieron a continuación, el hallazgo se interpretó como positivo, puesto que ponía de manifiesto la importancia del gen y su posible papel regulador[184]. Además, gran parte de los interactores de *XRN2* mostraron una expresión diferencial en el cerebro de individuos con TEA en comparación con cerebros de individuos control, lo que sugiere que *XRN2* sea un gen periférico involucrado en una extensa red en la que podrían estar participando diversos genes implicados en la fisiopatología de los TEA. Otro gen que resultó asociado por ambos algoritmos, H-MAGMA y PASCAL, fue *C8orf74*. *C8orf74* forma parte de un marco de lectura abierta y se desconoce aún su función biológica. Se convierte, por tanto, en un candidato para futuros estudios funcionales.

Si bien en el capítulo 1 se ha estudiado en mayor detalle la variación común, el capítulo 2 ha profundizado en el estudio de la variación rara y en concreto, en el estudio de las mutaciones *de novo*. Las mutaciones *de novo* surgen espontáneamente en un individuo, y de tener un impacto funcional en la proteína, es muy poco probable que se propaguen en la población. Por ese motivo, se consideran variantes especialmente deletéreas, y su identificación en individuos afectos nos

permite detectar genes directamente implicados en la etiología de una enfermedad, o genes centrales, si atendemos al modelo omnigénico que se ha explicado con anterioridad[274].

El estudio de las variantes *de novo* ha sido de vital importancia en los TEA, permitiendo identificar genes de riesgo como *CHD8, SCN2A o ARID1B*.[91,95,275]. Sin embargo, las mutaciones *de novo* se han detectado clásicamente en estado germinal, lo que significa que durante años se ha obviado el papel de las PZMs en la patogénesis de los TEA. Los resultados del capítulo 2 de esta tesis, prueban, sin embargo, que la correcta detección de PZMs y su posterior análisis permite identificar nuevos genes de riesgo en TEA y explorar rutas biológicas alternativas a aquellas en las cuales intervienen genes con mutaciones germinales.

En efecto, el análisis de PZMs llevado a cabo en el capítulo 2 mostró al menos 2 nuevos genes asociados a los TEA (*FRG1* y *NFIA*) Para explicar el hecho de que haya genes que acumulan diferencialmente un tipo u otro de mutación *de novo*, en este manuscrito se han propuesto dos hipótesis.

La primera de ellas señala la posibilidad de que haya genes en los que solo se pueden detectar mutaciones somáticas. Es decir, mutaciones que en un estado germinal producen síndromes incompatibles con la vida y que, por tanto, solo es posible identificarlas en estado mosaico. Este fenómeno ya se ha descrito en algunos síndromes, como el Síndrome de Proteus, causado por una mutación en mosaico en *AKT1* o el síndrome Sturge Weber causado por una mutación en mosaico en *GNAQ*[278,279].

La segunda hipótesis sugiere que los individuos portadores de PZMs manifiestan síntomas atípicos o más leves, de los que se esperaría en presencia de una mutación germinal y que ello repercuta en su correcto diagnóstico. Por ejemplo, mutaciones germinales en algunos genes, como *KMT2C*, se relacionan con síndromes genéticos muy severos que generalmente enmascaran características propias de los TEA. Sin embargo, si la mutación es postcigótica, el fenotipo es más leve, y se pueden identificar más fácilmente criterios diagnósticos para TEA. Puede ser, entonces, que algunos estudios genéticos de TEA se hayan llevado a cabo en cohortes que han reclutado pacientes siguiendo

criterios diagnósticos que favorecen la inclusión de individuos con PZMs y la exclusión de individuos con mutaciones germinales.

En el capítulo 2, también se ha señalado que las PZMs se localizan preferentemente en genes que intervienen en procesos biológicos esenciales durante el neurodesarrollo, como son la neurogénesis, la migración o la diferenciación neuronal. En línea con los argumentos mencionados anteriormente, esto podría deberse a que la interrupción de esos procesos es compatible con la vida solo si una pequeña población de células se ve afectada por ello. Así pues, existen numerosos trastornos caracterizados por una migración neuronal deficiente en los cuales se ha identificado una mutación en mosaico como causa principal. Este es el caso, por ejemplo, de la Heterotopia Nodular Periventricular, en la cual, las neuronas no migran hasta la corteza cerebral, y, en cambio, forman conglomerados cerca de los ventrículos. Este trastorno, ligado al cromosoma X y causado por mutaciones en *FLN1*, se observa generalmente en mujeres, siendo letal en hombres[280]. Sin embargo, si las mutaciones son en mosaico, se han descrito formas leves de este trastorno tanto en hombres como en mujeres[281,282]. Otro ejemplo es la lisencefalia, un trastorno de la migración neuronal que se caracteriza por la ausencia o reducción de las circunvoluciones cerebrales. La causa de este trastorno se debe a mutaciones germinales en *LIS1* en hombres y mujeres y mutaciones germinales en el gen ligado al cromosoma X, *DCX*, en hombres[283]. La severidad de sus síntomas se relaciona con la frecuencia del alelo alternativo de la mutación causal, lo que significa que el porcentaje de células afectas está directamente implicado con el fenotipo resultante[109,284].

Pese a que los resultados de esta tesis, y los de otros trabajos recientes[115,116,207,208], sugieren que las PZMs tienen un papel clave en la etiología de los TEA y otros TND, su detección es aún un campo de estudio emergente y muchas de las cuestiones vinculadas a estas variantes están aún lejos de ser resueltas. Sin ir más lejos, en el capítulo 2 se detallan algunas limitaciones que también están presentes en otros estudios genéticos de PZMs y TEA. En primer lugar, la detección de PZMs se ha realizado a partir de datos de secuenciación de exoma de ADN extraído de sangre periférica. Esto significa, que solo se han

podido detectar variantes que han ocurrido en etapas muy tempranas del neurodesarrollo, concretamente antes de la gastrulación. Esta aproximación impide, por tanto, detectar mutaciones que hayan ocurrido en etapas más tardías del neurodesarrollo y que solo estén presentes en el cerebro. Por otro lado, solo se han detectado mutaciones de una sola base (SNVs) localizadas en regiones exónicas. Se desconoce, por tanto, el papel de las PZMs en regiones no codificantes o el papel de las otras clases de mutaciones en estado mosaico (CNVs, *indels*, etc), en la etiología de los TEA.

Para estudiar variantes que aparecen en etapas más tardías del desarrollo, y que solo están presentes en un pequeño porcentaje de células, se utilizan dos aproximaciones: la secuenciación de célula única y las tecnologías de secuenciación de nueva generación en ADN extraído de tejido cerebral. La secuenciación de célula única permite detectar mutaciones que afectan a una sola neurona postmitótica. Sin embargo, se debe alcanzar un número mínimo de células afectadas en un tejido para que las mutaciones tengan un impacto funcional en el individuo[285]. Esta técnica, por tanto, no tiene una utilidad clínica, aunque su uso ha permitido estudiar diferentes linajes neuronales durante el desarrollo normal del cerebro[286]. La secuenciación de ADN extraído de tejido cerebral o de grupos celulares, sí que permite detectar mutaciones probablemente patogénicas. El éxito para identificar estas variantes, depende del número de células que porta la mutación y de la profundidad con la cual se secuencie cada base nucleotídica. La secuenciación de ADN extraído directamente de grupos celulares permite detectar variantes restringidas a un solo tipo celular y con un valor de AAF muy bajo. Sin embargo, el aislamiento de un tipo celular concreto es un proceso técnicamente complejo[285]. La secuenciación de ADN extraído de tejido, es, por el contrario, un proceso más simple. Además, empleando una profundidad de lectura adecuada, (300X-1000X), se pueden detectar mutaciones somáticas presentes en aproximadamente un 5% de las células[287]. Esta técnica se ha empleado con éxito en tejido cerebral *postmortem* de individuos con TEA, sin embargo, tiene dos claras limitaciones[287,288]. La primera, que una profundidad de lectura muy elevada implica un alto coste económico. La segunda, y más evidente, es que se trata de una técnica insostenible

en la práctica clínica, teniendo en cuenta que no se puede acceder a tejido cerebral de manera rutinaria y sin el empleo de técnicas invasivas. Con respecto a la segunda cuestión, sin embargo, diversos estudios sugieren que la mayoría de las mutaciones somáticas asociadas a una enfermedad tienen lugar preferentemente en las primeras divisiones del cigoto, y, por tanto, pueden ser detectadas en diversos tejidos, incluido sangre periférica[286,288]. Así, pues, la detección de PZMs a partir de ADN extraído de sangre periférica se podría implementar en la práctica clínica sin que ello supusiese la pérdida significativa de variantes PZMs posiblemente causales.

Los principales estudios llevados a cabo en los TEA, se han centrado principalmente en la detección de PZMs que son SNVs o *indels*[115,116,207,208,287,288]. Sin embargo, las neuronas del cerebro acumulan toda clase de mutaciones en mosaico, incluyendo expansiones trinucleotídicas, inserciones de retrotransposones L1 y CNVs[289–291]. La razón para obviar el papel de las diferentes clases de mutaciones probablemente se deba a que el ratio mutacional de las SNVs es significativamente superior al de las demás, lo que significa que son, con mayor probabilidad, causa de enfermedad[286]. En efecto, la frecuencia estimada de CNVs mosaico en individuos con TND es realmente baja (0.18-0.59%), aunque se ha de tener en cuenta que los estudios llevados a cabo hasta el momento identificaron CNVs a una resolución muy baja (aproximadamente 2 MB), con lo cual estos trabajos podrían no reflejar la carga real de estas mutaciones en individuos afectos[292,293]. Por otro lado, todos los estudios de PZMs se han llevado a cabo a partir de datos de secuenciación de exoma. Por tanto, se desconoce también el impacto de estas variantes en regiones no codificantes.

Los estudios de secuenciación de genoma completo serán, en un futuro próximo, la clave para entender el papel de las PZMs durante el neurodesarrollo, gracias a su capacidad para detectar todo el espectro mutacional, así como cubrir regiones no codificantes. De hecho, un estudio preliminar ya se ha llevado a cabo en TEA. En él se ha secuenciado a gran profundidad (250X) el genoma completo de ADN extraído directamente de tejido cerebral. Aunque el trabajo no se focalizó en la detección de variantes estructurales, los resultados

sugieren que las SNVs mosaico localizadas en regiones reguladoras también contribuyen al riesgo de los TEA. A medida que el coste de los estudios de secuenciación disminuya, se podrá incrementar el tamaño muestral de la cohorte de estudio lo que permitirá cuantificar en detalle el ratio mutacional y las características de las diferentes clases de mutaciones en mosaico en los TEA[288].

Los capítulos 1 y 2 de esta tesis han estudiado, de manera independiente, la variación común y la variación rara asociada a los TEA. En el capítulo 1 se ha realizado un GBA, con el cual hemos extraído información biológica de un GWAS, identificando genes asociados a TEA y caracterizándolos funcionalmente. En el capítulo 2 se han identificado mutaciones *de novo* (germinales y postcigóticas) y se ha estudiado su impacto funcional en las diferentes jerarquías biológicas. Se han usado pues, diferentes aproximaciones en cada uno de ellos, pero en ambos trabajos se han identificado genes y rutas biológicas involucrados en su patogénesis. El reto actual de los estudios genéticos en los TEA, es englobar todo este conocimiento de una manera coherente y generar un modelo genético adecuado que se ajuste a la heterogeneidad genética y fenotípica observada en los TEA. Probablemente el éxito de los futuros estudios resida en comprender mejor cómo interactúan las variantes comunes y raras entre sí.

En la arquitectura genética de los TEA se distinguen dos extremos bien definidos. En uno de ellos, se encuentran los trastornos sindrómicos de herencia mendeliana cuya causa se atribuye a una única CNV o SNV altamente penetrante. Este sería el caso, por ejemplo, del Síndrome de Smith-Magenis o de Sotos[294]. En el otro extremo, se pueden encontrar individuos con TEA de altas capacidades cognitivas y sin comorbilidades asociadas, en los cuales se espera una contribución mayor de la variación común. Entre ambos extremos, existe un amplio espectro alélico de diferentes frecuencias poblacionales que interactúan entre sí[295]. Para explicar cómo estas interacciones resultan en fenotipos muy diversos existen dos modelos que se explicarán a continuación.

De acuerdo a un modelo oligogénico o del segundo evento, las manifestaciones clínicas de una variante potencialmente patogénica dependen de la presencia de otras variantes genéticas raras. La hipótesis

del segundo evento surgió hace ya una década con el estudio de la microdeleción 16p12.1 en individuos con déficits cognitivos graves. Así pues, llamó la atención que dicha CNV era frecuentemente heredada de progenitores que no manifestaban los mismos síntomas que el probando. El estudio concluyó que dicha deleción era un factor de riesgo para el desarrollo de TND, pero que era necesario un segundo evento genético durante el neurodesarrollo para que aparecieran manifestaciones clínicas más severas[296]. Este modelo se extendió rápidamente a otros síndromes genéticos, incluyendo 1q21.1, 7q11.23 y 16p11.2, donde una presentación clínica más severa se correlacionó con el número de variantes raras secundarias[294]. Recientemente, se ha comprobado que la hipótesis del segundo evento no solo se atribuye a CNVs, sino también a SNVs. Así pues, el fenotipo de individuos portadores de SNVs patogénicas en genes de riesgo para TEA se relaciona con el número de SNVs raras adicionales presentes en su genoma[297]. Todos estos resultados prueban, que al menos en algunos casos, el riesgo para TEA y otros TND depende de la contribución conjunta de múltiples variantes raras.

El siguiente modelo, por otro lado, atribuye las manifestaciones clínicas de los TEA al resultado de la interacción de variantes raras de efecto moderado con múltiples alelos de bajo riesgo. Diversos estudios han señalado que, al menos en algunos subtipos de TEA, caracterizados por una alta capacidad cognitiva, el efecto aditivo de la variación común podría ser causa suficiente para manifestar el fenotipo[48]. Sin embargo, lo cierto es que un estudio reciente concluyó, que esta contribución es independiente del CI o de la presencia de una mutación deletérea *de novo*[298]. No obstante, se ha observado que la contribución del riesgo poligénico en individuos portadores de CNVs de riesgo es inversamente proporcional al efecto de la CNV[299].

Los dos modelos mencionados son compatibles con la hipótesis de que los factores de riesgo genético para TEA se encuentran distribuidos en la población general y la interacción de todos ellos resulta en el amplio espectro fenotípico que caracteriza a los TEA. Así pues, existiría un umbral, a partir del cual un individuo cumple criterios diagnósticos para TEA en función de su carga genética[300]. Prueba de ello, es que existe una enorme variabilidad entre individuos a la hora de

comunicarse e interaccionar socialmente y que algunos de los factores de riesgo genético conocidos para los TEA se han identificado en la población general. Este es el caso, por ejemplo, de la deleción 16p11.2, que se ha identificado en individuos que no alcanzan los criterios diagnósticos para TEA, o la identificación de CNVs de riesgo para TEA y esquizofrenia en controles con afectación cognitiva leve[294,301]. Recientemente, también se ha encontrado una correlación entre variantes comunes asociadas a los TEA y dificultades en la comunicación y el comportamiento social en la población general[302].

Todo este conocimiento sobre la arquitectura genética de los TEA se transfiere, sin embargo, muy lentamente a la práctica clínica. El capítulo 3 de esta tesis, refleja claramente la problemática actual en el diagnóstico de los TEA aunque se debe remarcar que el estudio se llevó a cabo en una cohorte muy pequeña. Así pues, solo se logró un diagnóstico genético en pacientes con mutaciones altamente penetrantes, cuya sola presencia es causa suficiente para causar el trastorno. De hecho, en ninguno de estos pacientes se identificó una variante de *novo* adicional afectando a alguno de los genes que definimos como "candidatos". Este hallazgo, ya se identificó en algunos trastornos sindrómicos asociados a CNVs como el síndrome de Smith-Magenis, la deleción 17q21.31 y la deleción 9q34. En estos trastornos, al contrario que en otros de expresividad variable, la frecuencia de eventos genéticos secundarios es muy baja, indicando que la existencia de una variante rara adicional es incompatible con la vida[294].

Los resultados de nuestro estudio indican, pues, que los algoritmos diagnósticos actuales favorecen el diagnóstico de individuos con trastornos sindrómicos de herencia mendeliana, pero no sirven para diagnosticar al resto, cuyo fenotipo es el resultado de la interacción de variantes raras y comunes. Estos individuos, tal y como indican los modelos genéticos anteriormente descritos, se beneficiarían de un diagnóstico si a la hora de interpretar una variante se tuviese en cuenta su *background* genético.

Para conocer la carga genética asociada a TEA que tiene un individuo, un método sencillo y poco costoso consiste en estudiar en detalle a sus familiares de primer grado. Se ha reportado de manera consistente que a menudo, éstos manifiestan una serie de características

cognitivas o de conducta propias de los TEA, cuya severidad no es suficiente como para que sean consideradas criterios diagnósticos, pero sí que se acercan al umbral diagnóstico. El conjunto de estas características constituye el fenotipo ampliado del autismo o BAP (de sus siglas en inglés, *Broad Autism Phenotype*)[303]. Al igual que ocurre en los TEA, los estudios en gemelos demuestran que estas características tienen un origen genético, puesto que la concordancia del BAP en gemelos MC (90%) es muy superior con respecto a la de gemelos DC (10%)[304]. Estos datos sugieren que el riesgo genético para los TEA es superior en individuos con BAP que en la población general, y que dicho riesgo es heredable. El BAP, por tanto, podría ser considerado como una medida indirecta del *background* genético existente en una familia.

Los estudios familiares de carácter clínico y epidemiológico han encontrado una estrecha relación entre las puntuaciones BAP en los progenitores y la severidad de los síntomas de TEA en probandos, o el tipo de autismo que manifiestan[305,306]. Así pues, en familias con varios miembros afectos es más común encontrar un mayor porcentaje de individuos con BAP que en familias con autismo esporádico[307]. Este hallazgo tiene sentido teniendo cuenta que los estudios genéticos señalan mecanismos etiológicos diferentes en ambos tipos de TEA. Así, los TEA esporádicos estarían más frecuentemente vinculados a variantes causales *de novo* y en los TEA familiar, la herencia de factores genéticos de riesgo sería el mecanismo etiológico más probable. En el reciente estudio de Pizzo *et al.*, también se encontró una relación entre la historia familiar y las manifestaciones clínicas del probando. Así, aunque no se midieron características clínicas de BAP, sí se observó que los individuos con TEA con una historia familiar de trastornos psiquiátricos o del neurodesarrollo manifestaban una clínica más severa y compleja que los que pertenecían a familias sin antecedentes similares. En estos individuos se observó, en efecto, una mayor transmisión de variantes genéticas raras que podrían ser la causa de la severidad de sus síntomas[297].

Estos resultados sugieren que, las medidas de BAP y una correcta historia clínica familiar, podrían ayudar a identificar diferentes subgrupos de TEA y así orientar mejor el diagnóstico en cada caso.

Otra medida de riesgo genético es la puntuación de riesgo poligénico o PRS (de sus siglas en inglés *Poligenic Risk Score*). El PRS es un *score* que predice el riesgo de un individuo para desarrollar una enfermedad teniendo en cuenta el efecto de la variación común. Para calcularlo, se suma los efectos de los diferentes alelos de riesgo cuya aportación se calcula a partir de los tamaños de efecto de los GWAS[308].

El PRS en una medida que empieza a ser empleada en trastornos comunes como son la diabetes tipo 2, la enfermedad de las arterias coronarias o el Alzheimer para la toma de decisiones clínicas en función de la predicción del riesgo. Por ejemplo, en base a estas puntuaciones, se puede decidir qué intervenciones terapeúticas son más recomendables en cada caso, diseñar programas de *screening* adaptados y orientar hábitos de vida personalizados[309]. Así pues, el PRS no están pensado para realizar un diagnóstico, pero permite estratificar la población y diseñar intervenciones más personalizadas.

Para hacer el cálculo del PRS, hace falta una muestra, que se llama "de descubrimiento", sobre la cual se estiman los alelos de riesgo asociados a una enfermedad y el tamaño de efecto para cada uno ellos. Con esta información, el modelo de riesgo poligénico, puede calcular el PRS en individuos que hayan sido genotipados[308]. El problema en los TEA, es que, hasta hace relativamente poco, ningún GWAS había obtenido señales significativas. Como se ha comentado en diversas ocasiones a lo largo de este manuscrito, el trabajo de Grove *et al*., ha solventado esta limitación permitiendo estimar a partir de sus datos puntuaciones de riesgo poligénico. En este trabajo la estimación del PRS sugiere la existencia de una arquitectura genética diferente en los diferentes subtipos de TEA según las categorías diagnósticas del ICD-10 (*International Classification of Diseases, 10th revisión*). El hallazgo más relevante, es la confirmación de que los subtipos con altas capacidades cognitivas (Asperger y Autismo de la Infancia) están más relacionados con la variación común que el resto, en los que se espera una contribución mayor de las variantes raras[48]. Aunque no dejan de ser resultados preliminares, estos hallazgos constituyen la primera evidencia de que existen diferentes subtipos de TEA, cada uno con una arquitectura genética propia y que, por tanto, su diagnóstico genético se debería orientar de manera diferente.

Otra medida que favorecería el diagnóstico de los pacientes con TEA es la implementación de algoritmos diagnósticos adecuados para la detección de variantes PZMs de manera rutinaria. Los estudios genéticos llevados a cabo hasta el momento calculan que hasta un 4% de estas variantes contribuyen al riesgo de los TEA[115,116], lo que significa que un amplio porcentaje de individuos se podría beneficiar de un diagnóstico en el caso de que se llevase a cabo un correcto análisis. Estos individuos podrían manifestar fenotipos leves o atípicos que podrían estar pasando desapercibidos. Además, la ausencia de herramientas adecuadas para detectarlas, puede estar conduciendo a diagnósticos erróneos en algunos casos. Por ejemplo, existen PZMs, cuya AAF es lo suficientemente alta, como para que puedan ser falsamente identificadas como mutaciones germinales. El reanálisis de la SSC reveló, en efecto, que hasta un 22% de las variantes que habían sido clasificadas como germinales, eran, en realidad, PZMs[115]. Este dato erróneo puede repercutir en el diagnóstico de un probando que herede una de estas variantes de alguno de sus progenitores. Si la variante es, en realidad, patogénica, el progenitor podría estar enmascarando un fenotipo leve o atípico que no alcanza criterios diagnósticos para TEA. En el probando, la misma variante en estado germinal, podría ser la causa de su patología, pero ser incorrectamente clasificada como benigna por ser heredada de un progenitor aparentemente sano. En la cohorte del DDD, por ejemplo, la correcta identificación de mosaicismos tanto en probandos como en progenitores logró un diagnóstico adicional del 1% en la cohorte. De lo individuos diagnosticados, en un 2% de los casos, la variante clasificada como patogénica había sido heredada de un progenitor sano, donde la variante se encontraba en estado mosaico[115]. Se trata, pues, de un fenómeno recurrente que habría de ser tenido en cuenta a la hora de llevar a cabo la clasificación de variantes en la práctica clínica.

Las variantes PZMs, podrían, además, actuar como moduladores de la expresión o de la penetrancia de otras mutaciones germinales deletéreas[287]. Incluso, podrían enmascarar trastornos autosómicos recesivos, en los cuales la segunda variante podría estar en estado mosaico y restringida a tejido cerebral, sin que pueda ser detectada por esa razón.

Así pues, los hallazgos de los estudios genéticos llevados a cabo en los últimos años, sugieren que las PZMs son una causa etiológica conocida en los TEA y como tal, deberían ser considerados en la práctica clínica. No solo se requiere la implementación de un algoritmo adecuado para detectarlas, sino alcanzar un consenso a la hora de interpretarlas y clasificarlas.

El objetivo final de los estudios genéticos llevados a cabo en los TEA, debería ser su traslación, lo más pronto posible, a la práctica clínica. Sin embargo, tal y como se ha expuesto con anterioridad, solo un pequeño porcentaje de individuos se beneficia de un diagnóstico. Así pues, si sumamos el rendimiento diagnóstico del exoma y el *array* solo se diagnóstica aproximadamente un 25% de los casos[169]. La mayor limitación en los TEA es la falta de guías clínicas específicas para este TND que incorporen la información adquirida en estudios genéticos de carácter académico. Las últimas recomendaciones al respecto de la ACMG en los TEA datan de 2013[156], y, por lo tanto, se encuentran completamente desactualizadas. Además, para la interpretación de variantes se usan guías que han sido diseñadas para trastornos de herencia mendeliana, siendo la experiencia del genetista crucial para discriminar entre variantes benignas y patogénicas[173].

Los últimos estudios genéticos comienzan a evidenciar la existencia de diferentes subtipos de TEA que deberían ser abordados de manera diferente en la práctica clínica[173]. Así pues, es probable que, en los TEA con altas capacidades cognitivas, el efecto de la variación común o de las PZMs sea superior. Por ello, este tipo de paciente podría beneficiarse, en un futuro próximo de la utilización rutinaria del cálculo del PRS y la detección de PZMs en el diagnóstico genético. En pacientes con una mayor afectación cognitiva, es probable que la contribución de la variación rara sea superior. Este es el caso de trastornos sindrómicos de herencia mendeliana, pero también de individuos cuyo fenotipo es el resultado de la interacción de múltiples variantes de efecto moderado. Para el diagnóstico de cualquiera de los subtipos, sin embargo, es de especial importancia fenotipar correctamente a los individuos y obtener una historia familia detallada donde se describa si existen antecedentes de trastornos psiquiátricos o

del neurodesarrollo. Para ello es necesario fomentar un correcto diálogo entre clínicos y genetistas.

En conclusión, aunque nuestro conocimiento sobre la arquitectura genética de los TEA es cada vez más detallado, hacen falta más estudios que corroboren estos hallazgos a un nivel clínico. En los próximos años, los estudios de secuenciación de genoma completo revolucionarán el campo de la genómica, tal y como lo han hecho durante la última década los estudios de secuenciación de exoma completo. Esta aproximación permitirá conocer el impacto funcional de las diferentes clases de mutaciones en los TEA, así como el papel regulador de variantes no codificantes. Será crucial para entonces, saber transferir correctamente todo este conocimiento a la práctica clínica, y que éste repercuta directamente en la calidad de vida de los pacientes con TEA.

.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

8 CONCLUSIONES

Primera: Se ha realizado un GBA, mediante PASCAL, usando como *input* los *summary statistics* del último metaánalisis de GWAS de TEA. El GBA ha identificado 8 *loci* asociados ($p < 2.26 \times 10^{-6}$), de los cuales, 4 (*NKX2-2, NKX2-4, CRHR1-IT1, C8orf74* y *LOC644172)*, eran hallazgos nuevos que no habían sido previamente reportados por MAGMA.

Segunda: La anotación funcional empleando los datos de expresión de *BrainSpan* (29 periodos del desarrollo) y *GTEx* (53 tejidos), ha permitido llevar a cabo la caracterización biológica de los genes asociados en el GBA y sus principales interactores. Se han distinguido dos clústeres de expresión, lo cual indica la existencia de mecanismos de regulación diferentes entre cerebro fetal y adulto. *NKX-2* y su interactor *OLIG-2* pertenecen al primer clúster, y mostraron infraexpresión en el periodo prenatal y sobreexpresión en el periodo postnatal. Al segundo clúster pertenecen *XRN2* y la mayoría de sus interactores, que mostraron la tendencia opuesta.

Tercera: La correcta detección de PZMs en un estudio de secuenciación de exoma completo ha permitido identificar nuevos genes de riesgo en TEA, que acumulan un porcentaje mayor de PZMs que de mutaciones germinales. Este ha sido el caso de *FRG1* y *NFIA* que no habían sido descritos previamente en TEA.

Cuarta: El empleo de diferentes herramientas bioinformáticas para caracterizar los procesos biológicos en los que intervienen genes con mutaciones germinales y genes con PZMs, ha permitido determinar que ambos tipos de genes intervienen en mecanismos biológicos de susceptibilidad a TEA diferentes. Los genes con mutaciones germinales están relacionados con un déficit generalizado de la comunicación neuronal durante todo el periodo prenatal, mientras que los genes con PZMs se asocian con la interrupción de procesos biológicos como la

neurogénesis, la migración o la diferenciación neuronal, solo en algunas células, y durante el periodo prenatal medio.

Quinta: La secuenciación de exoma completo de una cohorte gallega de TEA, compuesta por 125 tríos, ha permitido establecer un diagnóstico genético en 18 probandos, lo que significa un rendimiento diagnóstico de la técnica de un 14.4%.

Sexta: La mayoría de los pacientes que recibieron un diagnóstico genético presentaba un trastorno sindrómico, una encefalopatía epiléptica o una discapacidad intelectual. Esto sugiere que los individuos con TEA y con otras comorbilidades asociadas en los cuales haya una sospecha previa de trastorno genético podrían beneficiarse del empleo rutinario de la secuenciación de exoma completo como primera herramienta diagnóstica, seguida del *array* en caso de no ser diagnosticados mediante exoma.

First: *A GBA has been performed with PASCAL, using as input the summary statistics of the last ASD GWAS meta-analysis. The GBA has identified 8 associated loci (p < 2.26 x 10-6), of which, 4 (NKX2-2, NKX2-4, CRHR1-IT1, C8orf74 and LOC644172), were new findings that had not been previously reported by MAGMA.*

Second: *Functional annotation using BrainSpan (29 periods of development) and GTEx (53 tissues) expression data, has allowed to carry out the biological characterization of the associated genes in the GBA and their main interactors. Two expression clusters have been distinguished, indicating the existence of different regulatory mechanisms between fetal and adult brain. NKX-2 and its OLIG-2 interactor belong to the first cluster, and showed under-expression in the prenatal period and over-expression in the postnatal period. XRN-2 and most of its interactors belong to the second cluster, which showed the opposite trend*

Third: *Accurate detection of PZMs in a whole exome sequencing study has allowed the identification of new risk genes in ASD, which accumulate a higher percentage of PZMs, than germinal mutations. This has been the case for FRG1 and NFIA, that had not been previously described in ASD.*

Fourth: *A characterization of the biological processes underlying genes with germinal and genes with PZMs was performed using bioinformatic tools. This study has made it possible to determine that both types of genes are involved in different biological mechanisms of susceptibility to ASD. Genes with germline mutations are associated with impairments in neuronal communication during the prenatal period, while genes with PZMs are linked to disruption of biological processes such as neurogenesis, migration or neuronal differentiation, only in a few cells and during the mid-prenatal period.*

Fifth: *Whole exome sequencing in a Galician cohort of ASD composed of 125 trios, has allowed to establish a genetic diagnosis in 18 probands, which means a diagnostic yield of 14.4%.*

Sixth: *Most patients who received a genetic diagnosis had a syndromic disorder, epileptic encephalopathy, or intellectual disability. This suggests that individuals with ASD and other associated comorbidities, in whom there is a prior suspicion of a genetic disorder, may benefit from the use of whole exome sequencing as first tier test, followed by array if they are not diagnosed by exome.*

.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

9 BIBLIOGRAFÍA

1. Kanner, L. Autistic disturbances of affective contact. *Nerv. Child* **2**, 217–250 (1943).
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (Arlington, VA: American Psychiatric Publishing)*. (2013).
3. Lotter, V. Epidemiology of Autistic Conditions in Young Children. 124–137 (1964).
4. Fombonne, E. Epidemiology of pervasive developmental disorders. *Pediatr. Res.* **65**, 591–598 (2009).
5. Christensen, D. L. *et al.* Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 4 Years — Early Autism and Developmental Disabilities Monitoring Network, Seven Sites, United States, 2010, 2012, and 2014. *MMWR. Surveillance Summaries* **68**, 1–19 (2019).
6. Elsabbagh, M. *et al.* Global Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Res.* **5**, 160–179 (2012).
7. Loomes, R., Hull, L. & Mandy, W. P. L. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *J. Am. Acad. Child Adolesc. Psychiatry* **56**, 466–474 (2017).
8. Lai, M.-C., V. Lombardo, M. & Baron-Cohen, S. Autism. *Lancet* **383**, 896–910 (2014).
9. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder Analysis method B. *Jama* **318**, 1182–1184 (2017).
10. Grabrucker, A. M. Environmental factors in autism. *Front. Psychiatry* **3**, 1–13 (2013).
11. Mandy, W. & Lai, M. C. Annual Research Review: The role of the environment in the developmental psychopathology of autism spectrum condition. *J. Child Psychol. Psychiatry Allied Discip.* **57**, 271–292 (2016).
12. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* **9**, 255–

266 (2008).
13. Tenesa, A. & Haley, C. S. The heritability of human disease: Estimation, uses and abuses. *Nat. Rev. Genet.* **14**, 139–149 (2013).
14. Bailey, A. *et al.* Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychol. Med.* **25**, 63–77 (1995).
15. Rosenberg, R. E., Law, J. K., Yenokyan, G. & Mcgready, J. Characteristics and Concordance of Autism Spectrum Disorders Among 277 Twin Pairs. (2009). doi:10.1001/archpediatrics.2009.98
16. Rutter, M. Heritability of autism spectrum disorders : a meta-analysis of twin studies. **5**, 585–595 (2016).
17. Sandin, S., Lichtenstein, P., Larsson, H., Cm, H. & Reichenberg, A. The familial risk of autism. **311**, 24794370 (2014).
18. Hansen, S. N. *et al.* Recurrence Risk of Autism in Siblings and Cousins: A Multinational, Population-Based Study. *J. Am. Acad. Child Adolesc. Psychiatry* **58**, 866–875 (2019).
19. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
20. International HapMap Consortium. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
21. Belmont, J. W. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
22. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
23. Geschwind, D. H. & State, M. W. Autism 1 Gene hunting in autism spectrum disorder : on the path to precision medicine. 1109–1120 (2015). doi:10.1016/S1474-4422(15)00044-7
24. Timpson, N. J., Greenwood, C. M. T., Soranzo, N. & Lawson, D. J. Genetic architecture : the shape of the genetic contribution to human traits and disease. *Nat. Publ. Gr.* **19**, 110–124 (2017).
25. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
26. Teare, M. D. & Barrett, J. H. Genetic linkage studies. *An Introd. to Genet. Epidemiol.* 39–59 (2011). doi:10.2307/j.ctt1t895v2.8
27. Freitag, C. M. The genetics of autistic disorders and its clinical relevance: A review of the literature. *Mol. Psychiatry* **12**, 2–22 (2007).
28. Weiss, L. A. *et al.* A genome-wide linkage and association scan reveals

novel loci for autism. *Nature* **461**, 802–808 (2009).
29. Werling, D. M., Lowe, J. K., Luo, R., Cantor, R. M. & Geschwind, D. H. Replication of linkage at chromosome 20p13 and identification of suggestive sex-differential risk loci for autism spectrum disorder. *Mol. Autism* **5**, 1–16 (2014).
30. Consortium, T. A. G. P. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**, 319 (2007).
31. Cantor, R. M. *et al.* Replication of Autism Linkage : Fine-Mapping Peak at 17q21. 1050–1056 (2005).
32. Stone, J. L. *et al.* Evidence for Sex-Specific Risk Alleles in Autism Spectrum Disorder. 1117–1123 (2004).
33. Mittal, R., Aggarwal, A. & Srivastava, G. The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *Int. J. Dermatol.* **44**, 1031–1034 (2005).
34. Alarcón, M. *et al.* Linkage, Association, and Gene-Expression Analyses Identify CNTNAP2 as an Autism-Susceptibility Gene. *Am. J. Hum. Genet.* **82**, 150–159 (2008).
35. Cordell, H. J. & Clayton, D. G. Genetic Epidemiology 3 Genetic association studies. 1121–1131
36. Holt, R. *et al.* Linkage and candidate gene studies of autism spectrum disorders in European populations. *Eur. J. Hum. Genet.* **18**, 1013–1020 (2010).
37. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.* **9**, 341–355 (2008).
38. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 1–13 (2012).
39. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology* **8**, (2012).
40. Anney, R. *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072–4082 (2010).
41. Ma, D. Q. *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet.* **73**, 263–273 (2010).
42. Jones, R. M. *et al.* Genome-Wide Association Study of Autistic-Like Traits in a General Population Study of Young Adults. *Front. Hum. Neurosci.* **7**, 1–10 (2013).
43. Kerin, T. *et al.* A Noncoding RNA Antisense to Moesin at 5p14.1 in

Autism. *Sci. Transl. Med.* **4**, 128ra40 LP-128ra40 (2012).
44. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
45. Torrico, B. *et al.* Lack of replication of previous autism spectrum disorder GWAS hits in European populations. *Autism Res.* **10**, 202–211 (2017).
46. Consortium, T. A. S. D. W. G. of T. P. G. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 21 (2017).
47. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* 1–4 (2019). doi:10.1093/bioinformatics/btz633
48. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics* **23**, 22 (2019).
49. Huang, H., Chanda, P., Alonso, A., Bader, J. S. & Arking, D. E. Gene-Based tests of association. *PLoS Genet.* **7**, (2011).
50. Neale, B. M. & Sham, P. C. The future of association studies: Gene-based analysis and replication. *Am. J. Hum. Genet.* **75**, 353–362 (2004).
51. Mishra, A. & Macgregor, S. VEGAS2 : Gene-based test software using 1000 Genomes reference. (2014).
52. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology* **11**, (2015).
53. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput. Biol.* **12**, (2016).
54. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
55. Koboldt, D. C., Larson, D. E., Chen, K., Ding, L. & Wilson, R. K. Genomic Structural Variants. *Methods Mol Biol* **838**, 1–14 (2012).
56. Poultney, C. S. *et al.* Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *American Journal of Human Genetics* **93**, 607–619 (2013).
57. Brandler, W. M. *et al.* Frequency and Complexity of de Novo Structural Mutation in Autism. *Am. J. Hum. Genet.* **98**, 667–679 (2016).
58. Zhou, B. *et al.* Whole-genome sequencing analysis of CNV using low-

coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J. Med. Genet.* **55**, 735–743 (2018).

59. Tan, R. *et al.* An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Hum. Mutat.* **35**, 899–907 (2014).
60. Jacquemont, M. *et al.* Array-based comparative genomic hybridisation identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. 843–850 (2006). doi:10.1136/jmg.2006.043166
61. Sebat, J. *et al.* Strong Association of De Novo Copy Number Mutations with Autism. *Science (80-.).* **316**, 445–449 (2007).
62. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *J. Hum. Genet.* 477–488 (2008). doi:10.1016/j.ajhg.2007.12.009.
63. Pinto, D. *et al.* Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. *Nature* **466**, 368–372 (2010).
64. Levy, D. *et al.* Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron* **70**, 886–897 (2011).
65. Sanders, S. J. *et al.* Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron* **70**, 863–885 (2011).
66. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
67. Jacquemont, S. *et al.* A higher mutational burden in females supports a 'female protective model' in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
68. Leppa, V. M. M. *et al.* Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *Am. J. Hum. Genet.* **99**, 540–554 (2016).
69. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American Journal of Human Genetics* **94**, 677–694 (2014).
70. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92**, 221–237 (2013).
71. Vissers, L. E. and P. S. Microdeletion and microduplication syndromes. *Methods Mol Biol* **838**, 29–75 (2012).

72. Watson, C. T., Marques-Bonet, T., Sharp, A. J. & Mefford, H. C. The Genetics of Microdeletion and Microduplication Syndromes: An Update. *Annu. Rev. Genomics Hum. Genet.* 1–30 (2014). doi:10.1146/annurev-genom-091212-153408
73. Rosenfeld, J. a, Coe, B. P., Eichler, E. E., Cuckle, H. & Shaffer, L. G. Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet. Med.* **15**, 478–81 (2013).
74. Malhotra, D. & Sebat, J. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
75. Dijk, E. L. Van, Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **34**, 666–681 (2018).
76. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
77. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
78. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. **2012**, (2012).
79. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
80. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
81. Jaszczyszyn, Y., Thermes, C. & Dijk, E. L. Van. Ten years of next-generation sequencing technology. **30**, (2014).
82. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).
83. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State. *Nat. Rev. Genet.* **10**, 57 (2009).
84. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
85. Fischbach, G. D. & Lord, C. NeuroView The Simons Simplex Collection : A Resource for Identification of Autism Genetic Risk Factors NeuroView. *Neuron* **68**, 192–195 (2010).
86. Feliciano, P. *et al.* SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).

87. D Buxbaum, J. *et al.* The Autism Sequencing Consortium: Large scale, high throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2013).
88. Yuen, R. K. C. *et al.* Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–191 (2015).
89. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
90. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
91. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-U124 (2012).
92. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285–299 (2012).
93. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
94. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
95. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
96. Dong, S. *et al.* De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9**, 16–23 (2014).
97. Ronemus, M., Iossifov, I., Levy, D. & Wigler, M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nature Reviews Genetics* **15**, 133–141 (2014).
98. Lim, E. T. *et al.* Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. *Neuron* **77**, 235–242 (2013).
99. Doan, R. N. *et al.* Recessive gene disruptions in autism spectrum disorder. *Nat. Genet.* **51**, (2019).
100. Yu, T. W. *et al.* Using Whole-Exome Sequencing to Identify Inherited Causes of Autism. *Neuron* **77**, 259–273 (2013).
101. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
102. He, X. *et al.* Integrated Model of De Novo and Inherited Genetic

Variants Yields Greater Power to Identify Risk Genes. *PLoS Genet.* **9**, (2013).
103. Alonso-Gonzalez, A., Rodriguez-Fontenla, C. & Carracedo, A. De novo mutations (DNMs) in autism spectrum disorder (ASD): Pathway and network analysis. *Frontiers in Genetics* **9**, (2018).
104. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* (2020). doi:10.1016/j.cell.2019.12.036
105. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–15 (2014).
106. Kosmicki, J. A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
107. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *www.biorsiv.org* 148353 (2017). doi:10.1101/148353
108. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–20 (2013).
109. Jamuar, S. S. *et al.* Somatic Mutations in Cerebral Cortical Malformations. *N. Engl. J. Med.* **371**, 733–743 (2014).
110. Lee, J. H. *et al.* De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat Genet* **44**, 941–945 (2012).
111. Poduri, A. *et al.* Somatic Activation of AKT3 Causes Hemispheric Developmental Brain Malformations. *Neuron* **74**, 41–48 (2012).
112. Pieras, J. I. *et al.* Somatic mosaicism for Y120X mutation in the MECP2 gene causes atypical Rett syndrome in a male. *Brain Dev.* **33**, 608–611 (2012).
113. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
114. Pagnamenta, A. T. *et al.* Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. *J. Hum. Genet.* 70–72 (2012). doi:10.1038/jhg.2011.128
115. Krupp, D. R. *et al.* Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am. J. Hum. Genet.* **101**, 369–390 (2017).
116. Lim, E. T. *et al.* Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Publ. Gr.* (2017).

doi:10.1038/nn.4598
117. Collins, R. L. *et al.* Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 1–21 (2017).
118. Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895–908 (2013).
119. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
120. Takata, A. Estimating contribution of rare non-coding variants to neuropsychiatric disorders. *Psychiatry Clin. Neurosci.* **73**, 2–10 (2019).
121. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
122. Zhou, J. *et al.* Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
123. Werling, D. M. *et al.* An analytical framework for whole genome sequence association studies and its implications for autism spectrum disorder. *Nature Genetics* **50**, 727–736 (2018).
124. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722.e12 (2017).
125. Brandler, W. M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. **331**, 327–331 (2018).
126. Ruzzo, E. K. *et al.* Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks Article Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* **178**, 850–866 (2019).
127. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
128. Miller, J. Transcriptional Landscape of the Prenatal Human Brain. *Nature* **508**, 199–206 (2014).
129. Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. **45**, (2013).
130. Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat. Publ. Gr.* (2015). doi:10.1038/nrg3934
131. Graaf, A. Van Der & Franke, L. Gene co-expression analysis for functional classification and gene – disease predictions. **19**, 575–592 (2018).
132. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression

measures : mutual information , correlation , and model based indices. (2012).
133. Mcdowall, M. D., Scott, M. S. & Barton, G. J. PIPs : human protein – protein interaction prediction database. **37**, 651–656 (2009).
134. Lehne, B. & Schlitt, T. Protein – protein interaction databases : Keeping up with growing interactomes. **3**, 291–297 (2009).
135. Gilman, S. R. *et al.* Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat. Neurosci.* **15**, 1723–1728 (2012).
136. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* **25**, 142–154 (2015).
137. Gilman, S. R. *et al.* Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses. *Neuron* **70**, 898–907 (2011).
138. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
139. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications* **5**, 5748 (2014).
140. Parikshak, N. N. *et al.* Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).
141. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, (2013).
142. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* **19**, 1454–1462 (2016).
143. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nat. Neurosci.* **18**, 191–198 (2014).
144. Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
145. Rubenstein, J. L. R. & Merzenich, M. M. Model of autism : increased ratio of excitation / inhibition in key neural systems. 255–267 (2003). doi:10.1046/j.1601-183X.2003.00037.x

146. Yong, J. & Claudianos, C. Neuroscience and Biobehavioral Reviews Genetic heterogeneity in autism : From single gene to a pathway perspective. *Neurosci. Biobehav. Rev.* **68**, 442–453 (2016).
147. Devlin, B. & Scherer, S. W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229–237 (2012).
148. Persico, A. M. & Napolioni, V. Autism genetics. *Behav. Brain Res.* (2013). doi:10.1016/j.bbr.2013.06.012
149. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research* **1380**, 42–77 (2011).
150. Sztainberg, Y. & Zoghbi, H. Y. Lessons learned from studying syndromic autism spectrum disorders. *Nature Neuroscience* **19**, 1408–1417 (2016).
151. Miles, J. H. Genetics in Medicine Autism spectrum disorders — A genetics review. **13**, (2011).
152. Zarrei, M. *et al.* A large data resource of genomic copy number variation across neurodevelopmental disorders. *npj Genomic Med.* **4**, (2019).
153. Fernandez, B. A. & Scherer, S. W. Syndromic autism spectrum disorders: moving from a clinically defined to a molecularly defined approach. *Dialogues in clinical neuroscience* **19**, 353–371 (2017).
154. Miller, D. T. *et al.* Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
155. Waggoner, D. *et al.* Yield of additional genetic testing after chromosomal microarray for diagnosis of neurodevelopmental disability and congenital anomalies: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **20**, 1105–1113 (2018).
156. Schaefer, G. B. & Mendelsohn, N. J. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. *Genet. Med.* **15**, 399–407 (2013).
157. Contractor, A., Klyachko, V. A. & Portera-Cailliau, C. Altered Neuronal and Circuit Excitability in Fragile X Syndrome. *Neuron* **87**, 699–715 (2015).
158. Miles, J. H. Autism spectrum disorders—A genetics review. *Genet. Med.* **13**, 278–294 (2011).

159. Clifford, S. *et al.* Autism Spectrum Phenotype in Males and Females with Fragile X Full Mutation and Premutation. *J. Autism Dev. Disord.* **37**, 738–747 (2007).
160. Tassone, F. *et al.* Clinical involvement and protein expression in individuals with the FMR1 premutation. *Am. J. Med. Genet.* **91**, 144–152 (2000).
161. Hundscheid, R. D. L., Smits, A. P. T., Thomas, C. M. G., Kiemeney, L. A. L. M. & Braat, D. D. M. Female carriers of fragile X premutations have no increased risk for additional diseases other than premature ovarian failure. *Am. J. Med. Genet. A* **117A**, 6–9 (2003).
162. Neul, J. L. *et al.* Rett syndrome: revised diagnostic criteria and nomenclature. *Ann. Neurol.* **68**, 944–950 (2010).
163. Zahorakova, D. *et al.* MECP2 mutations in Czech patients with Rett syndrome and Rett-like phenotypes: novel mutations, genotype-phenotype correlations and validation of high-resolution melting analysis for mutation scanning. *J. Hum. Genet.* **61**, 617–625 (2016).
164. Mester, J. L., Tilot, A. K., Rybicki, L. A., Frazier 2nd, T. W. & Eng, C. Analysis of prevalence and degree of macrocephaly in patients with germline PTEN mutations and of brain weight in Pten knock-in murine model. *Eur. J. Hum. Genet.* **19**, 763–768 (2011).
165. Sun, Y. *et al.* Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Hum. Mutat.* **36**, 648–655 (2015).
166. Hoang, N., Buchanan, J. A. & Scherer, S. W. Heterogeneity in clinical sequencing tests marketed for autism spectrum disorders. *npj Genomic Medicine* **3**, (2018).
167. O'Donnell-Luria, A. H. & Miller, D. T. A Clinician's perspective on clinical exome sequencing. *Human Genetics* **135**, 643–654 (2016).
168. Points to consider in the clinical application of genomic sequencing. *Genet. Med.* **14**, 759–761 (2012).
169. Tammimies, K. *et al.* Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA - J. Am. Med. Assoc.* **314**, 595–903 (2015).
170. Srivastava, S. *et al.* Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet. Med.* (2019). doi:10.1038/s41436-019-0554-6
171. Szego, M. J. & Zawati, M. H. Whole Genome Sequencing as a Genetic

Test for Autism Spectrum Disorder: From Bench to Bedside and then Back Again. *J. Can. Acad. Child Adolesc. Psychiatry* **25**, 116–121 (2016).

172. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score : variant. *Nature Publishing Group* (2017). doi:10.1038/nrg.2017.52
173. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
174. Zhang, S. *et al.* dbMDEGA: A database for meta-analysis of differentially expressed genes in autism spectrum disorder. *BMC Bioinformatics* **18**, (2017).
175. Zhong, X. *et al.* Effect of corticotropin-releasing hormone receptor1 gene variation on psychosocial stress reaction via the dorsal anterior cingulate cortex in healthy adults. *Brain Res.* **1707**, 1–7 (2018).
176. Ginsberg, M. R., Rubin, R. A., Falcone, T., Ting, A. H. & Natowicz, M. R. Brain Transcriptional and Epigenetic Associations with Autism. *PLoS ONE* **7**, (2012).
177. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics* **50**, 229–237 (2018).
178. Briscoe, J. & Ericson, J. The specification of neuronal identity by graded Sonic Hedgehog signalling. *Semin Cell Dev Biol* **10**, 353–362 (1999).
179. Shen, T. *et al.* Brain-specific deletion of histone variant H2A.z results in cortical neurogenesis defects and neurodevelopmental disorder. *Nucleic Acids Research* **46**, 2290–2307 (2018).
180. Cai, J. *et al.* Co-localization of Nkx6.2 and Nkx2.2 homeodomain proteins in differentiated myelinating oligodendrocytes. *Glia* **58**, 458–468 (2010).
181. Li, J. *et al.* Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Molecular Systems Biology* **10**, 774–774 (2014).
182. Zeidán-Chuliá, F. *et al.* Up-Regulation of Oligodendrocyte Lineage Markers in the Cerebellum of Autistic Patients: Evidence from Network Analysis of Gene Expression. *Mol. Neurobiol.* **53**, 4019–4025 (2016).
183. Chow, M. L. *et al.* Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes

at young versus mature ages. *PLoS Genet.* **8**, e1002592 (2012).
184. Eaton, J. D. *et al.* Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes Dev.* **32**, 127–139 (2018).
185. Day, F. R., Ong, K. K. & Perry, J. R. B. Elucidating the genetic basis of social interaction and isolation. *Nature Communications* **9**, (2018).
186. Binder, E. B. & Nemeroff, C. B. The CRF system, stress, depression and anxietyinsights from human genetic studies. *Molecular Psychiatry* **15**, 574–588 (2010).
187. Smith, D. J. *et al.* Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry* **21**, 749–757 (2016).
188. Grimm, S. *et al.* The interaction of corticotropin-releasing hormone receptor gene and early life stress on emotional empathy. *Behav. Brain Res.* **329**, 180–185 (2017).
189. Alonso-González, A. *et al.* Postzygotic and germinal de novo mutations in ASD: exploring their biological role. *bioRxiv* 2020.05.21.107987 (2020). doi:10.1101/2020.05.21.107987
190. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
191. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software* **3**, 731 (2018).
192. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
193. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
194. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry* **17**, 142–153 (2012).
195. Georgi, B., Voight, B. F. & Bućan, M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genetics* **9**, (2013).
196. Cotney, J. *et al.* The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6**, 6404 (2015).
197. M. Lek, Goveas, J. S. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **27**, 94–102 (2017).

198. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Rep.* **6**, 1139–1152 (2014).
199. Collins, A. L. *et al.* Transcriptional targets of the schizophrenia risk gene MIR137. *Translational Psychiatry* **4**, (2014).
200. Wagnon, J. L. *et al.* CELF4 Regulates Translation and Local Abundance of a Vast Set of mRNAs, Including Genes Associated with Regulation of Synaptic Function. *PLoS Genetics* **8**, (2012).
201. Lin, M. *et al.* Allele-Biased Expression in Differentiating Human Neurons: Implications for Neuropsychiatric Disorders. *PLoS ONE* **7**, (2012).
202. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 5–10 (2016).
203. Bartonicek, N. *et al.* Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol.* **18**, (2017).
204. Ching, A.-S. & Ahmad-Annuar, A. A Perspective on the Role of microRNA-128 Regulation in Mental and Behavioral Disorders. *Front. Cell. Neurosci.* **9**, (2015).
205. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* **5**, e13984 (2010).
206. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).
207. Freed, D. & Pevsner, J. The Contribution of Mosaic Variants to Autism Spectrum Disorder. *PLoS Genet.* **12**, 1–20 (2016).
208. Dou, Y. *et al.* Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Human Mutation* **38**, 1002–1013 (2017).
209. de Lange, I. M. *et al.* Mosaicism of de novo pathogenic SCN1A variants in epilepsy is a frequent phenomenon that correlates with variable phenotypes. *Epilepsia* **59**, 690–703 (2018).
210. Koemans, T. S. *et al.* Functional convergence of histone methyltransferases EHMT1 and KMT2C involved in intellectual disability and autism spectrum disorder. *PLoS Genet.* **13**, e1006864 (2017).
211. Schanze, I. *et al.* NFIB Haploinsufficiency Is Associated with

Intellectual Disability and Macrocephaly. *American Journal of Human Genetics* **103**, 752–768 (2018).
212. Hanel, M. L. *et al.* Facioscapulohumeral muscular dystrophy (FSHD) region gene 1 (FRG1) is a dynamic nuclear and sarcomeric protein. *Differentiation* **81**, 107–118 (2011).
213. Saito, Y. *et al.* Facioscapulohumeral muscular dystrophy with severe mental retardation and epilepsy. *Brain and Development* **29**, 231–233 (2007).
214. Pistoni, M. *et al.* FRG1 Downregulation and Altered 3 Splicing by in a Mouse Model of Facioscapulohumeral Muscular Dystrophy (FSHD). *PLoS Genet.* **9**, e1003186 (2013).
215. Hamada, N. *et al.* Essential role of the nuclear isoform of RBFOX1, a candidate gene for autism spectrum disorders, in the brain development. *Sci. Rep.* **6**, 30805 (2016).
216. Glasgow, S. M. *et al.* Glia-specific enhancers and chromatin structure regulate NFIA expression and glioma tumorigenesis. *Nature Neuroscience* **20**, 1520–1528 (2017).
217. Lu, W. *et al.* NFIA haploinsufficiency is associated with a CNS malformation syndrome and urinary tract defects. *PLoS Genetics* **3**, 830–843 (2007).
218. Revah-Politi, A. *et al.* Loss-of-function variants in NFIA provide further support that NFIA is a critical gene in 1p32-p31 deletion syndrome: A four patient series. *American Journal of Medical Genetics, Part A* **173**, 3158–3164 (2017).
219. Mahmoudi, E. & Cairns, M. J. MiR-137: An important player in neural development and neoplastic transformation. *Molecular Psychiatry* **22**, 44–55 (2017).
220. Szulwach, K. E. *et al.* Cross talk between microRNA and epigenetic regulation in adult neurogenesis. *J. Cell Biol.* **189**, 127–141 (2010).
221. Smrt, R. D. *et al.* MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase mind bomb-1. *Stem Cells* **28**, 1060–1070 (2010).
222. He, E. *et al.* MIR137 schizophrenia-associated locus controls synaptic function by regulating synaptogenesis, synapse maturation and synaptic transmission. *Human Molecular Genetics* **27**, 1879–1891 (2018).
223. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).

224. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
225. Cheng, Y. *et al.* Partial loss of psychiatric risk gene Mir137 in mice causes repetitive behavior and impairs sociability and learning via increased Pde10a. *Nature Neuroscience* (2018). doi:10.1038/s41593-018-0261-7
226. Nelson, S. B. & Valakh, V. Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders. *Neuron* **87**, 684–698 (2015).
227. Stoner, R. *et al.* Patches of disorganization in the neocortex of children with Autism. *N. Engl. J. Med.* **370**, 1209–1219 (2014).
228. 1-s2.0-S2211124717317382-main.pdf.
229. A family of hyperpolarization- activated mammalian cation channels.
230. Li, M. *et al.* Gain-of-function HCN2 variants in genetic epilepsy. *Human Mutation* **39**, 202–209 (2018).
231. DiFrancesco, J. C. *et al.* HCN ion channels and accessory proteins in epilepsy: genetic analysis of a large cohort of patients and review of the literature. *Epilepsy Res.* **153**, 49–58 (2019).
232. Luoma, L. M. & Berry, F. B. Molecular analysis of NPAS3 functional domains and variants. *BMC Molecular Biology* **19**, (2018).
233. Kamm, G. B., Pisciottano, F., Kliger, R. & Franchini, L. F. The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. *Mol. Biol. Evol.* **30**, 1088–1102 (2013).
234. Macintyre, G. *et al.* Association of NPAS3 exonic variation with schizophrenia. *Schizophr. Res.* **120**, 143–149 (2010).
235. Brunskill, E. W. *et al.* Abnormal neurodevelopment, neurosignaling and behaviour in Npas3-deficient mice. *Eur. J. Neurosci.* **22**, 1265–1276 (2005).
236. Rasmussen, A. H., Rasmussen, H. B. & Silahtaroglu, A. The DLGAP family: Neuronal expression, function and role in brain disorders. *Molecular Brain* **10**, (2017).
237. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
238. Coba, M. P. *et al.* Dlgap1 knockout mice exhibit alterations of the postsynaptic density and selective reductions in sociability. *Scientific Reports* **8**, (2018).

239. Jaiswal, M., Dvorsky, R. & Ahmadian, M. R. Deciphering the molecular and functional basis of Dbl family proteins: A novel systematic approach toward classification of selective activation of the Rho family proteins. *J. Biol. Chem.* **288**, 4486–4500 (2013).
240. Ravindran, E. *et al.* Homozygous ARHGEF2 mutation causes intellectual disability and midbrain-hindbrain malformation. *PLoS Genetics* **13**, (2017).
241. Sylwestrak, E. L. & Ghosh, A. Elfn1 regulates target-specific release probability at CA1-interneuron synapses. *Science (80-.).* **338**, 536–540 (2012).
242. Dolan, J. & Mitchell, K. J. Mutation of Elfn1 in mice causes seizures and hyperactivity. *PLoS ONE* **8**, (2013).
243. Tomioka, N. H. *et al.* Elfn1 recruits presynaptic mGluR7 in trans and its loss results in seizures. *Nature Communications* **5**, 1–16 (2014).
244. rsob-4-130108.pdf.
245. Dulovic-Mahlow, M. *et al.* De Novo Variants in TAOK1 Cause Neurodevelopmental Disorders. *Am. J. Hum. Genet.* **105**, 213–220 (2019).
246. Kraus, D. M. *et al.* CSMD1 Is a Novel Multiple Domain Complement-Regulatory Protein Highly Expressed in the Central Nervous System and Epithelial Tissues. *J. Immunol.* **176**, 4419–4430 (2014).
247. Donohoe, G. *et al.* Neuropsychological effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253. *Genes, Brain and Behavior* **12**, 203–209 (2013).
248. Athanasiu, L. *et al.* A genetic association study of CSMD1 and CSMD2 with cognitive function. *Brain. Behav. Immun.* **61**, 209–216 (2017).
249. Salzberg, Y., Ramirez-Suarez, N. J. & Bülow, H. E. The Proprotein Convertase KPC-1/Furin Controls Branching and Self-avoidance of Sensory Dendrites in Caenorhabditis elegans. *PLoS Genetics* **10**, (2014).
250. Menachem Fromer1, 2*, Panos Roussos1, 2, 3, 4*, Solveig K Sieberts5*, Jessica S Johnson1, David H Kavanagh1, 2, Thanneer M Perumal5, Douglas M Ruderfer1, 2, Edwin C Oh6, 7, Aaron Topol1, Hardik R Shah2, Lambertus L Klei8, Robin Kramer9, Dalila Pinto1, 2, 3^. Gene Expression Elucidates Functional Impact of Polygenic Risk for Schizophrenia. **8**, 1–33 (2012).
251. Liang, H. *et al.* Neural development is dependent on the function of specificity protein 2 in cell cycle progression. *Development* **140**, 552

LP – 561 (2013).
252. Schaaf, C. P. Nicotinic acetylcholine receptors in human genetic disease. *Genet. Med.* **16**, 649–656 (2014).
253. Dickinson, J. A., Hanrott, K. E., Mok, M. H. S., Kew, J. N. C. & Wonnacott, S. Differential coupling of α7 and non-α7 nicotinic acetylcholine receptors to calcium-induced calcium release and voltage-operated calcium channels in PC12 cells. *J. Neurochem.* **100**, 1089–1096 (2007).
254. Gillentine, M. A. & Schaaf, C. P. The human clinical phenotypes of altered CHRNA7 copy number. *Biochemical Pharmacology* **97**, 352–362 (2015).
255. M.A., G. *et al.* Functional Consequences of CHRNA7 Copy-Number Alterations in Induced Pluripotent Stem Cells and Neural Progenitor Cells. *Am. J. Hum. Genet.* **101**, 874–887 (2017).
256. HAQQ, A. M. *et al.* Characterization of a novel binding partner of the melanocortin-4 receptor: attractin-like protein. *Biochemical Journal* **376**, 595–605 (2003).
257. 1-s2.0-S1769721210000765-main.pdf.
258. Lebrun, N. *et al.* Novel KDM5B splice variants identified in patients with developmental disorders: Functional consequences. *Gene* **679**, 305–313 (2018).
259. Gobron, S. *et al.* Subcommissural organ/Reissner's fiber complex: Characterization of SCO-spondin, a glycoprotein with potent activity on neurite outgrowth. *Glia* **32**, 177–191 (2000).
260. Swayne, L. A. & Bennett, S. A. L. Connexins and pannexins in neuronal development and adult neurogenesis. *BMC Cell Biology* **17**, (2016).
261. Rossi, M. *et al.* Outcomes of Diagnostic Exome Sequencing in Patients With Diagnosed or Suspected Autism Spectrum Disorders. *Pediatric Neurology* **70**, 34-43.e2 (2017).
262. K., R. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* **18**, 696–704 (2016).
263. Yu, Y. & Li, F. Genetic Diagnostic Evaluation of Trio-Based Whole Exome Sequencing Among Children With Diagnosed or Suspected Autism Spectrum Disorder. **9**, 1–8 (2018).
264. Ben-Shalom, R. *et al.* Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Observed in Individuals With Autism Spectrum Disorder or Infantile Seizures. *Biol. Psychiatry* **82**, 224–232 (2017).

265. J., H. *et al.* De novo mutations in SIK1 cause a spectrum of developmental epilepsies. *American Journal of Human Genetics* **96**, 682–690 (2015).
266. Helbig, K. L. *et al.* De Novo Pathogenic Variants in CACNA1E Cause Developmental and Epileptic Encephalopathy with Contractures, Macrocephaly, and Dyskinesias. *Am. J. Hum. Genet.* **103**, 666–678 (2018).
267. Wenger, A. M., Guturu, H., Bernstein, J. A. & Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genetics in Medicine* **19**, 209–214 (2017).
268. Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* **20**, 1216–1223 (2018).
269. Xiao, B. *et al.* Marked yield of re-evaluating phenotype and exome/target sequencing data in 33 individuals with intellectual disabilities. *Am. J. Med. Genet. Part A* **176**, 107–115 (2018).
270. Vissers, L. E. L. M. *et al.* A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet. Med.* **19**, 1055–1063 (2017).
271. Pfundt, R. *et al.* Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.* **19**, 667–675 (2017).
272. Retterer, K. *et al.* Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet. Med.* **17**, 623–629 (2015).
273. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
274. Boyle, E. A., Li, Y. I. & Pritchard, J. K. Perspective An Expanded View of Complex Traits : From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
275. Liu, X. *et al.* Trans Effects on Gene Expression Can Drive Omnigenic Inheritance Theory Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).
276. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

277. Sey, N. Y. A. *et al.* A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, (2020).
278. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N. Engl. J. Med.* **365**, 611–9 (2011).
279. Shirley, M. D. *et al.* Sturge–Weber Syndrome and Port-Wine Stains Caused by Somatic Mutation in. 1971–1979 (2013). doi:10.1056/NEJMoa1213507
280. Sheen, V. L. *et al.* Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. *Hum. Mol. Genet.* **10**, 1775–1783 (2001).
281. Guerrini, R. *et al.* Germline and mosaic mutations of FLN1 in men with periventricular heterotopia. *Neurology* **63**, 51–56 (2004).
282. Parrini, E., Mei, D., Wright, M., Dorn, T. & Guerrini, R. Mosaic mutations of the FLN1 gene cause a mild phenotype in patients with periventricular heterotopia. *Neurogenetics* **5**, 191–196 (2004).
283. Sicca, F. *et al.* Mosaic mutations of the LIS1 gene cause subcortical band heterotopia. *Neurology* **61**, 1042–1046 (2003).
284. Gleeson, J. G. *et al.* Somatic and germline mosaic mutations in the doublecortin gene are associated with variable phenotypes. *Am. J. Hum. Genet.* **67**, 574–581 (2000).
285. D'Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelopmental disease. *Nature Neuroscience* **21**, 1504–1514 (2018).
286. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science (80-.).* **350**, 94–98 (2015).
287. Li, M., Lam, A. N., Sestan, N. & Walsh, C. A. Targeted DNA Sequencing from Autism Spectrum Report Targeted DNA Sequencing from Autism Spectrum Disorder Brains Implicates Multiple Genetic Mechanisms. *Neuron* **88**, 910–917 (2015).
288. Rodin, R. E. *et al.* The Landscape of Mutational Mosaicism in Autistic and Normal Human Cerebral Cortex. *bioRxiv* 2020.02.11.944413 (2020). doi:10.1101/2020.02.11.944413
289. Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
290. Evrony, G. D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**,

483–496 (2012).
291. McConnell, M. J. *et al.* Mosaic Copy Number Variation in Human Neurons. *Science (80-.).* **342**, 632 LP – 637 (2013).
292. King, D. A. *et al.* Mosaic structural variation in children with developmental disorders. **24**, 2733–2745 (2015).
293. King, D. A. *et al.* Detection of structural mosaicism from targeted and whole-genome sequencing data. 1704–1714 (2017). doi:10.1101/gr.212373.116.Freely
294. Girirajan, S. *et al.* Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).
295. Hu, W. F., Chahrour, M. H. & Walsh, C. A. The Diverse Genetic Landscape of Neurodevelopmental Disorders. (2014). doi:10.1146/annurev-genom-090413-025600
296. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42**, 203–209 (2010).
297. Pizzo, L. *et al.* Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genet. Med.* **21**, 816–825 (2019).
298. These, D., We, Q. & These, Q. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. doi:10.1038/ng.3863
299. Bergen, S. E. *et al.* Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *Am. J. Psychiatry* **176**, 29–35 (2019).
300. I *et al.* The Genetic Basis of Phenotypic Diversity: Autism as an Extreme Tail of a Complex Dimensional Trait. *Intech* **i**, 13 (2012).
301. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
302. Robinson, E. B. *et al.* Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15161–15165 (2014).
303. Ingersoll, B. & Wainer, A. The Broader Autism Phenotype. in *Handbook of Autism and Pervasive Developmental Disorders, Fourth Edition* (American Cancer Society, 2014). doi:10.1002/9781118911389.hautc02
304. Genetic Aspects of the Broad Autism Phenotype. in *The Broad Autism*

Phenotype **29**, 37–63 (Emerald Group Publishing Limited, 2015).

305. Sasson, N. J., Lam, K. S. L., Parlier, M., Daniels, J. L. & Piven, J. Autism and the broad autism phenotype: Familial patterns and intergenerational transmission. *J. Neurodev. Disord.* **5**, 1–7 (2013).
306. Gerdts, J. A., Bernier, R., Dawson, G. & Estes, A. The broader autism phenotype in simplex and multiplex families. *J. Autism Dev. Disord.* **43**, 1597–1605 (2013).
307. Davidson, J. *et al.* Expression of the broad autism phenotype in simplex autism families from the simons simplex collection. *J. Autism Dev. Disord.* **44**, 2392–2399 (2014).
308. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. **9**, (2013).
309. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
310. Huguet, G., Benabou, M. & Bourgeron, T. The Genetics of Autism Spectrum Disorders. in 101–129 (2016). doi:10.1007/978-3-319-27069-2_11

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

10 ANEXOS

10.4 ANEXO 4. DESCRIPCIÓN DE LOS GENES CON MUTACIONES GERMINALES OBTENIDOS EN LA COHORTE COMBINADA.

Se muestran los genes con valores de FDR < 0.1. Se incluye la siguiente información: gen ya reconocido como un gen de riesgo de TEA, gen incluido en la base de datos SFARI, gen previamente indentificado en otros análisis TADA y enfermedad OMIM asociada al gen en inglés.

Gen	q-valor	p-valor	Gen candidato	Gen SFARI	Clasificación SFARI	TADA previo FDR <0.01	OMIM
SCN2A	5.04×10^{-12}	4.13×10^{-8}	si	si	1	si	Epileptic encephalopathy, early infantile, 11; Seizures, benign familial infantile, 3
CHD8	2.40×10^{-5}	4.13×10^{-8}	si	si	1s	si	{Autism, susceptibility to, 18}
ARID1B	5.36×10^{-5}	4.13×10^{-8}	si	si	1s	si	Coffin-siris syndrome 1
SLC6A1	0.00014991	2.48×10^{-7}	si	si	2s	si	Myoclonic-atonic epilepsy
SYNGAP1	0.000508219	6.61×10^{-7}	si	si	1s	si	Mental retardation, autosomal dominant 5
KDM5B	0.000841602	8.26×10^{-7}	si	si	2s	no	Mental retardation, autosomal recessive 65

KMT5B	0.001898533	5.37×10^{-6}	si	si	1S	si	Mental retardation, autosomal dominant 51
TRIP12	0.0027703	5.79×10^{-6}	si	si	1S	no	Mental retardation, autosomal dominant 49
PTEN	0.004131864	1.14×10^{-5}	si	si	1S	si	Cowden syndrome 1;Lhermitte-Duclos syndrome;Macrocephaly/autism syndrome;Prostate cancer, somatic;{Glioma susceptibility 2};{Meningioma}
KATNAL2	0.008224845	5.01×10^{-5}	si	si	1	si	NA
NRXN1	0.01194415	5.79×10^{-5}	si	si	2	no	Pitt-Hopkins-like syndrome 2; {Schizophrenia, susceptibility to, 17}
CREBBP	0.015098004	5.92×10^{-5}	no	si	5	no	Menke-Hennekam syndrome 1;Rubinstein-Taybi syndrome 1
CELF4	0.017826919	6.09×10^{-5}	no	si	3	no	NA
STXBP1	0.020341344	6.48×10^{-5}	no	si	3s	NO	Epileptic encephalopathy, early infantile, 4
DYRK1A	0.0227959	7.09×10^{-5}	si	si	1s	si	Mental retardation, autosomal dominant 7

CHD2	0.025896407	0.00010719	si	si	1s	no	Epileptic encephalopathy, childhood-onset
ANK2	0.029572774	0.000136694	si	si	1	si	Cardiac arrhythmia, ankyrin-B-related; Long QT syndrome 4
WDFY3	0.033145437	0.000147769	si	si	2	no	?Microcephaly 18, primary, autosomal dominant
UNC80	0.037280204	0.00018595	no	si	4	no	Hypotonia, infantile, with psychomotor retardation and characteristic facies 2
CLASP1	0.041141977	0.000192149	no	si	3	no	NA
TMEM39B	0.045179962	0.000226612	no	no	NA	no	NA
PRKAR1B	0.049032281	0.000240909	no	no	NA	no	NA
USP45	0.053777544	0.000349174	no	si	3	no	NA
NUAK1	0.058346344	0.000365372	no	si	3	no	NA
NAA15	0.062733828	0.000387769	si	si	1s	si	Mental retardation, autosomal dominant 50
FOXP1	0.066799738	0.000390331	si	si	1s	no	Mental retardation with language impairment and with or without autistic features

ZC3H11A	0.07061159	0.00039876	no	no	NA	no	NA
DPP3	0.074752544	0.000470909	no	no	NA	no	NA
PRKDC	0.079128953	0.00053124	no	si	4	no	Immunodeficiency 26, with or without neurologic abnormalities
ATP1A1	0.083428257	0.000549008	no	si	4s	no	Charcot-Marie-Tooth disease, axonal, type 2DD; Hypomagnesemia, seizures, and mental retardation 2
LRP5	0.087492397	0.000553223	no	no	NA	no	Exudative vitreoretinopathy 4; Hyperostosis, endosteal; Osteopetrosis, autosomal dominant 1; Osteoporosis-pseudoglioma syndrome; Osteosclerosis; Polycystic liver disease 4 with or without kidney cysts; van Buchem disease, type 2; [Bone mineral density variability 1]; {Osteoporosis} 166710 AD 3
SLC12A3	0.091716102	0.000608099	no	no	NA	no	Gitelman syndrome
FBXO18	0.095909862	0.000644628	no	no	NA	no	NA

PTK7	0.099893315	0.000651157	no	si	3	si	NA

10.5 ANEXO 5. DESCRIPCIÓN DE LOS GENES CON PZMS OBTENIDOS EN LA COHORTE COMBINADA.

Se muestran los genes con valores de FDR <0.1. Se incluye la siguiente información: gen ya reconocido como un gen de riesgo de TEA, gen incluido en la base de datos SFARI, gen previamente indentificado en otros análisis TADA y enfermedad OMIM asociada al gen en inglés.

Gen	q-valor	p-valor	Gen candidato	Gen SFARI	Clasificación SFARI	TADA previo FDR <0.01	OMIM
FRG1	0.035045881	4.14×10^{-5}	no	no	NA	no	Facioscapulohumeral muscular dystrophy 1
KMT2C	0.069130706	0.000182873	si	si	s2	si	Kleefstra syndrome 2
NFIA	0.091745554	0.000279834	no	si	4	no	Brain malformations with or without urinary tract defects
SMARCA4	0.118024039	0.000517127	no	si	3	no	Coffin loris

PRKDC	0.13475291	0.000545856	no	si	4	no	Immunodeficiency 26, with or without neurologic abnormalities
KLF16	0.149512294	0.00064116	no	si	4	no	NA
GRIN2B	0.169042489	0.000949171	si	si	1	si	Epileptic encephalopathy, early infantile, 27; Mental retardation, autosomal dominant 6
MAP2K3	0.184127855	0.000977624	no	no	NA	NA	NA
HNRNPU	0.209192662	0.001851381	no	si	4	no	Epileptic encephalopathy, early infantile, 54
POTEB2	0.23133466	0.002020718	no	no	NA	NA	NA

RNPC3	0.250159757	0.002077072	no	no	NA	NA	?Growth hormone deficiency, isolated, type V
FAM177A1	0.26825174	0.002417127	no	no	NA	NA	NA
CALML6	0.28374479	0.002460773	no	no	NA	NA	NA
CMPK2	0.297509845	0.002585083	no	no	NA	NA	NA

10.6 ANEXO 6. SELECCIÓN DE GENES

Se muestra la lista formada por 261 genes asociados a TND. La selección de estos genes se llevó a cabo usando las bases de datos OMIM, SFARI Gene (https://gene.sfari.org) y la literatura científica.

Gene	SFARI score	SFARI criteria	OMIM phenotype	OMIM code	Inheritance
ADNP	score 1	High Confidence Criteria 1.1	Helsmoortel-van der Aa syndrome	615873	AD
AFF2	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked, FRAXE type	309548	XLR
AHDC1	score 1	High Confidence Criteria 1.1	Xia-Gibbs syndrome	615829	AD
ANK2	score 1	High Confidence Criteria 1.1	Cardiac arrhythmia, ankyrin-B-related; Long QT syndrome 4	600919; 600919	AD, AD
ANKRD11	score 1	High Confidence Criteria 1.1	KBG syndrome	148050	AD
ARID1B	score 1	High Confidence Criteria 1.1	Coffin-Siris syndrome 1	135900	AD
ASH1L	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 52	617796	AD
ASXL3	score 1	High Confidence Criteria 1.1	Bainbridge-Ropers syndrome	615485	AD
ATRX	score 1	High Confidence Criteria 1.1	Alpha-thalassemia myelodysplasia syndrome, somatic; Alpha-thalassemia/mental retardation syndrome; Mental retardation-hypotonic facies syndrome, X-linked	300448; 301040; 309580	NA; XLD; XLR
AUTS2	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 26	615834	AD

BCKDK	score 1	High Confidence Criteria 1.1	Branched-chain ketoacid dehydrogenase kinase deficiency	614923	NA
BCL11A	score 1	High Confidence Criteria 1.1	Dias-Logan syndrome	617101	AD
CASK	score 1	High Confidence Criteria 1.1	FG syndrome 4; Mental retardation and microcephaly with pontine and cerebellar hypoplasia; Mental retardation, with or without nystagmus	300422; 300749; 300422	NA; XLD; NA
CHD2	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, childhood-onset	615369	AD
CHD8	score 1	High Confidence Criteria 1.1	{Autism, susceptibility to, 18}	615032	AD
CTCF	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 21	615502	AD
CTNNB1	score 1	High Confidence Criteria 1.1	Colorectal cancer, somatic; Exudative vitreoretinopathy 7; Hepatocellular carcinoma, somatic; Medulloblastoma, somatic; Neurodevelopmental disorder with spastic diplegia and visual defects; Ovarian cancer, somatic; Pilomatricoma, somatic	114500; 617572; 114550; 155255; 615075; 167000; 132600	NA; AD; NA; NA; AD; NA; NA
CUL3	score 1	High Confidence Criteria 1.1	Pseudohypoaldosteronism, type IIE	614496	AD
DDX3X	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked 102	300958	XLD, XLR
DSCAM	score 1	High Confidence Criteria 1.1	NA	NA	NA
DYRK1A	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 7	614104	AD
EHMT1	score 1	High Confidence Criteria 1.1	Kleefstra syndrome 1	610253	AD

EP300	score 1	High Confidence Criteria 1.1	Colorectal cancer, somatic; Menke-Hennekam syndrome 2; Rubinstein-Taybi syndrome 2	114500; 618333; 613684	NA; NA; AD
FOXP1	score 1	High Confidence Criteria 1.1	Mental retardation with language impairment and with or without autistic features	613670	AD
GIGYF2	score 1	High Confidence Criteria 1.1	{Parkinson disease 11}	607688	NA
GRIN2B	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 27; Mental retardation, autosomal dominant 6	616139; 613970	AD; AD
HRAS	score 1	High Confidence Criteria 1.1	Bladder cancer, somatic; Congenital myopathy with excess of muscle spindles; Costello syndrome; Nevus sebaceous or woolly hair nevus, somatic; Schimmelpenning-Feuerstein-Mims syndrome, somatic mosaic; Spitz nevus or nevus spilus, somatic; Thyroid carcinoma, follicular, somatic	109800; 218040; 218040; 162900; 163200; 137550; 188470	NA; AD; AD; NA; NA; NA; NA
IQSEC2	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked 1/78	309530	XLD
IRF2BPL	score 1	High Confidence Criteria 1.1	Neurodevelopmental disorder with regression, abnormal movements, loss of speech, and seizures	618088	AD
KATNAL2	score 1	High Confidence Criteria 1.1	NA	NA	NA
KMT2A	score 1	High Confidence Criteria 1.1	Leukemia, myeloid/lymphoid or mixed-lineage; Wiedemann-Steiner syndrome	159555; 605130	AD; AD
KMT2C	score 1	High Confidence Criteria 1.1	Kleefstra syndrome 2	617768	AD
KMT5B	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 51	617788	AD

MAGEL2	score 1	High Confidence Criteria 1.1	Schaaf-Yang syndrome	615547	AD
MBD5	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 1	156200	AD
MBOAT7	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal recessive 57	617188	AR
MECP2	score 1	High Confidence Criteria 1.1	{Autism susceptibility, X-linked 3}; Encephalopathy, neonatal severe; Mental retardation, X-linked syndromic, Lubs type; Mental retardation, X-linked, syndromic 13; Rett syndrome; Rett syndrome, atypical; Rett syndrome, preserved speech variant	300496; 300673; 300260; 300055; 312750; 312750	XL; XLR; XLR; XLR; XLD; XLD; XLD
MED13L	score 1	High Confidence Criteria 1.1	Mental retardation and distinctive facial features with or without cardiac defects; Transposition of the great arteries, dextro-looped 1	616789; 608808	AD; AD
MYT1L	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 39	616521	AD
NCKAP1	score 1	High Confidence Criteria 1.1	NA	NA	NA
NLGN3	score 1	High Confidence Criteria 1.1	{Asperger syndrome susceptibility, X-linked 1}; {Autism susceptibility, X-linked 1}	300494; 300425	XL; XL
NRXN1	score 1	High Confidence Criteria 1.1	{Schizophrenia, susceptibility to, 17}; Pitt-Hopkins-like syndrome 2	614332; 614325	AR
POGZ	score 1	High Confidence Criteria 1.1	White-Sutton syndrome	616364	AD
PPP2R5D	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 35	616355	AD
PTCHD1	score 1	High Confidence Criteria 1.1	{Autism, susceptibility to, X-linked 4}	300830	XLR

PTEN	score 1	High Confidence Criteria 1.1	{Glioma susceptibility 2}; {Meningioma}; Cowden syndrome 1; Lhermitte-Duclos syndrome; Macrocephaly/autism syndrome; Prostate cancer, somatic	613028; 607174; 158350; 158350; 605309; 176807	NA; AD; AD; AD; AD; NA
PTPN11	score 1	High Confidence Criteria 1.1	LEOPARD syndrome 1; Leukemia, juvenile myelomonocytic, somatic; Metachondromatosis; Noonan syndrome 1	151100; 607785; 156250; 163950	AD;NA; AD; AD
RAI1	score 1	High Confidence Criteria 1.1	Smith-Magenis syndrome	182290	AD
RELN	score 1	High Confidence Criteria 1.1	{Epilepsy, familial temporal lobe, 7}; Lissencephaly 2 (Norman-Roberts type)	616436; 257320	AD; AR
RERE	score 1	High Confidence Criteria 1.1	Neurodevelopmental disorder with or without anomalies of the brain, eye, or heart	616975	AD
RIMS1	score 1	High Confidence Criteria 1.1	Cone-rod dystrophy 7	603649	NA
SCN1A	score 1	High Confidence Criteria 1.1	Epilepsy, generalized, with febrile seizures plus, type 2; Epileptic encephalopathy, early infantile, 6 (Dravet syndrome); Febrile seizures, familial, 3A; Migraine, familial hemiplegic, 3	604403; 607208; 604403; 609634	AD; AD; AD; AD
SCN2A	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 11; Seizures, benign familial infantile, 3	613721; 607745	AD; AD
SCN8A	score 1	High Confidence Criteria 1.1	?Myoclonus, familial, 2; Cognitive impairment with or without cerebellar ataxia; Epileptic encephalopathy, early infantile, 13; Seizures, benign familial infantile, 5	618364; 614306; 614558; 617080	AD; AD; AD; AD
SETBP1	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 29; Schinzel-Giedion midface retraction syndrome	616078; 269150	AD; AD
SETD5	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 23	615761	AD

SHANK2	score 1	High Confidence Criteria 1.1	{Autism susceptibility 17}	613436	NA
SHANK3	score 1	High Confidence Criteria 1.1	{Schizophrenia 15}; Phelan-McDermid syndrome	613950; 606232	AD; AD
SIN3A	score 1	High Confidence Criteria 1.1	Witteveen-Kolk syndrome	613406	AD
SLC6A1	score 1	High Confidence Criteria 1.1	Myoclonic-atonic epilepsy	616421	AD
SPAST	score 1	High Confidence Criteria 1.1	Spastic paraplegia 4, autosomal dominant	182601	AD
SRCAP	score 1	High Confidence Criteria 1.1	Floating-Harbor syndrome	136140	AD
STXBP1	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 4	612164	AD
SYNGAP1	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 5	612621	AD
TBR1	score 1	High Confidence Criteria 1.1	Intellectual developmental disorder with autism and speech delay	606053	AD
TCF4	score 1	High Confidence Criteria 1.1	Corneal dystrophy, Fuchs endothelial, 3; Pitt-Hopkins syndrome	613267; 610954	AD; AD
TRIP12	score 1	High Confidence Criteria 1.1	Mental retardation, autosomal dominant 49	617752	AD
TSC2	score 1	High Confidence Criteria 1.1	?Focal cortical dysplasia, type II, somatic; Lymphangioleiomyomatosis, somatic; Tuberous sclerosis-2	607341; 606690; 613254	NA; NA; AD
UBE3A	score 1	High Confidence Criteria 1.1	Angelman syndrome	105830	AD
UPF3B	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked, syndromic 14	300676	XLR

WAC	score 1	High Confidence Criteria 1.1	Desanto-Shinawi syndrome	616708	AD
WDFY3	score 1	High Confidence Criteria 1.1	?Microcephaly 18, primary, autosomal dominant	617520	AD
ZBTB20	score 1	High Confidence Criteria 1.1	Primrose syndrome	259050	AD
ADSL	score 1	High Confidence Criteria 1.1	Adenylosuccinase deficiency	103050	AR
ALDH5A1	score 1	High Confidence Criteria 1.1	Succinic semialdehyde dehydrogenase deficiency	271980	AR
ARX	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 1; Hydranencephaly with abnormal genitalia; Lissencephaly, X-linked 2; Mental retardation, X-linked 29 and others; Partington syndrome; Proud syndrome	308350; 300215; 300215; 300419; 309510; 300004	XLR; XL; XLR; XLR; XL
BRAF	score 1	High Confidence Criteria 1.1	Adenocarcinoma of lung, somatic; Cardiofaciocutaneous syndrome; Colorectal cancer, somatic; LEOPARD syndrome 3; Melanoma, malignant, somatic; Nonsmall cell lung cancer, somatic; Noonan syndrome 7	211980; 115150; NA; 613707; NA; NA; 613706	NA; AD; NA; AD; NA; NA; AD
CACNA1C	score 1	High Confidence Criteria 1.1	Brugada syndrome 3; Long QT syndrome 8; Timothy syndrome	611875; 618447; 601005	NA;NA; AD
CDKL5	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 2	300672	XLD
CHD7	score 1	High Confidence Criteria 1.1	CHARGE syndrome; Hypogonadotropic hypogonadism 5 with or without anosmia	214800; 612370	AD; AD
CREBBP	score 1	High Confidence Criteria 1.1	Menke-Hennekam syndrome 1; Rubinstein-Taybi syndrome 1	618332; 180849	NA; AD

DHCR7	score 1	High Confidence Criteria 1.1	Smith-Lemli-Opitz syndrome	270400	AR
DMPK	score 1	High Confidence Criteria 1.1	Myotonic dystrophy 1	160900	AD
FMR1	score 1	High Confidence Criteria 1.1	Fragile X syndrome; Fragile X tremor/ataxia syndrome; Premature ovarian failure 1	300624; 300623; 311360	XLD; XLD; XL
FOXG1	score 1	High Confidence Criteria 1.1	Rett syndrome, congenital variant	613454	AD
KIAA2022	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked 98	300912	XLD
NF1	score 1	High Confidence Criteria 1.1	Neurofibromatosis, type 1	162200	AD
NIPBL	score 1	High Confidence Criteria 1.1	Cornelia de Lange syndrome 1	122470	AD
NSD1	score 1	High Confidence Criteria 1.1	Leukemia, acute myeloid; Sotos syndrome 1	601626; 117550	AD; AD
PACS1	score 1	High Confidence Criteria 1.1	Schuurs-Hoeijmakers syndrome	615009	AD
PCDH19	score 1	High Confidence Criteria 1.1	Epileptic encephalopathy, early infantile, 9	300088	XL
POMGNT1	score 1	High Confidence Criteria 1.1	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3; Muscular dystrophy-dystroglycanopathy (congenital with mental retardation), type B, 3; Muscular dystrophy-dystroglycanopathy (limb-girdle), type C, 3; Retinitis pigmentosa 76	253280; 613151; 613157; 617123	AR; AR; AR; AR
SLC9A6	score 1	High Confidence Criteria 1.1	Mental retardation, X-linked syndromic, Christianson type	300243	XLD

TSC1	score 1	High Confidence Criteria 1.1	Focal cortical dysplasia, type II, somatic; Lymphangioleiomyomatosis; Tuberous sclerosis-1	607341; 606690; 191100	NA; NA; AD
VPS13B	score 1	High Confidence Criteria 1.1	Cohen syndrome	216550	AR
BRSK2	score 1	High Confidence Criteria 1.1	NA	NA	NA
CACNA1D	score 2	Strong Candidate Criteria 2.1	Primary aldosteronism, seizures, and neurologic abnormalities; Sinoatrial node dysfunction and deafness	615474; 614896	AD; AR
CACNA1H	score 2	Strong Candidate Criteria 2.1	{Epilepsy, childhood absence, susceptibility to, 6}; {Epilepsy, idiopathic generalized, susceptibility to, 6}; Hyperaldosteronism, familial, type IV	611942; 611942; 617027	NA, NA, AD
CACNA2D3	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
CC2D1A	score 2	Strong Candidate Criteria 2.1	Mental retardation, autosomal recessive 3	608443	AR
CEP41	score 2	Strong Candidate Criteria 2.1	Joubert syndrome 15	614464	AR
CNTN4	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
CTNND2	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
DOCK8	score 2	Strong Candidate Criteria 2.1	Hyper-IgE recurrent infection syndrome, autosomal recessive	243700;	AR
DYNC1H1	score 2	Strong Candidate Criteria 2.1	Charcot-Marie-Tooth disease, axonal, type 20; Mental retardation, autosomal dominant 13; Spinal muscular atrophy, lower extremity-predominant 1, AD	614228; 614563; 158600	AD; AD; AD

ERBIN	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
GABRB3	score 2	Strong Candidate Criteria 2.1	{Epilepsy, childhood absence, susceptibility to, 5}; Epileptic encephalopathy, early infantile, 43	612269; 617113	NA; AD
GRIK2	score 2	Strong Candidate Criteria 2.1	Mental retardation, autosomal recessive, 6	611092	AR
GRIN2A	score 2	Strong Candidate Criteria 2.1	Epilepsy, focal, with speech disorder and with or without mental retardation	245570	AD
GRIP1	score 2	Strong Candidate Criteria 2.1	Fraser syndrome 3	617667	AR
ILF2	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
INTS6	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
KAT2B	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
KCNJ10	score 2	Strong Candidate Criteria 2.1	Enlarged vestibular aqueduct, digenic; SESAME syndrome	600791; 612780	AR; AR
KDM5B	score 2	Strong Candidate Criteria 2.1	Mental retardation, autosomal recessive 65	618109	AR
KDM5C	score 2	Strong Candidate Criteria 2.1	Mental retardation, X-linked, syndromic, Claes-Jensen type	300534	XLR
KDM6A	score 2	Strong Candidate Criteria 2.1	Kabuki syndrome 2	300867	XLD
KIRREL3	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
LAMB1	score 2	Strong Candidate Criteria 2.1	Lissencephaly 5	615191	AR

MET	score 2	Strong Candidate Criteria 2.1	?Deafness, autosomal recessive 97; {Osteofibrous dysplasia, susceptibility to}; Hepatocellular carcinoma, childhood type, somatic; Renal cell carcinoma, papillary, 1, familial and somatic	616705; 607278; 114550; 605074	AR; AD; NA, NA
NLGN4X	score 2	Strong Candidate Criteria 2.1	{Asperger syndrome susceptibility, X-linked 2}; {Autism susceptibility, X-linked 2}; Mental retardation, X-linked	300497; 300495; 300495	XL; XL; XL
OPHN1	score 2	Strong Candidate Criteria 2.1	Mental retardation, X-linked, with cerebellar hypoplasia and distinctive facial appearance	300486	XLR
PAH	score 2	Strong Candidate Criteria 2.1	[Hyperphenylalaninemia, non-PKU mild]; Phenylketonuria	261600; 261600	AR; AR
RANBP17	score 2	Strong Candidate Criteria 2.1	NA	NA	NA
SMAD4	score 2	Strong Candidate Criteria 2.1	Juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome; Myhre syndrome; Pancreatic cancer, somatic; Polyposis, juvenile intestinal	175050; 139210; 260350; 174900	AD; AD; NA; AD
TBL1XR1	score 2	Strong Candidate Criteria 2.1	Mental retardation, autosomal dominant 41; Pierpont syndrome	616944; 602342	AD; AD
ZMYND11	score 2	Strong Candidate Criteria 2.1	Mental retardation, autosomal dominant 30	616083	AD
MSNP1AS	score 2	Strong Candidate Criteria 2.1			
CNTNAP2	score 2S	Strong Candidate, Syndromic Criteria 2.1, Syndromic	{Autism susceptibility 15}; Cortical dysplasia-focal epilepsy syndrome; Pitt-Hopkins like syndrome 1	612100; 610042; 610042	NA; AR; AR
DEAF1	score 2S	Strong Candidate, Syndromic Criteria 2.1, Syndromic	?Dyskinesia, seizures, and intellectual developmental disorder; Mental retardation, autosomal dominant 24	617171; 615828	AR; AD
KAT6A	score 2S	Strong Candidate, Syndromic	Mental retardation, autosomal dominant 32	616268	AD

		Criteria 2.1, Syndromic			
PRODH	score 2S	Strong Candidate, Syndromic Criteria 2.1, Syndromic	{Schizophrenia, susceptibility to, 4}; Hyperprolinemia, type I	600850; 239500	AD; AR
USP7	score 2S	Strong Candidate, Syndromic Criteria 2.1, Syndromic	NA	NA	NA
DDC	score 3	Suggestive Evidence Criteria 3.1	Aromatic L-amino acid decarboxylase deficiency	608643	AR
DPYD	score 3	Suggestive Evidence Criteria 3.1	5-fluorouracil toxicity; Dihydropyrimidine dehydrogenase deficiency	274270; 274270	AR; AR
EIF4E	score 3	Suggestive Evidence Criteria 3.1	{Autism, susceptibility to, 19}	615091	NA
IL1RAPL1	score 3	Suggestive Evidence Criteria 3.1	Mental retardation, X-linked 21/34	300143	XLR
KANK1	score 3	Suggestive Evidence Criteria 3.1	Cerebral palsy, spastic quadriplegic, 2	612900	NA
MAOA	score 3	Suggestive Evidence Criteria 3.1	{Antisocial behavior}; Brunner syndrome	300615; 300615	XLR; XLR
MCPH1	score 3	Suggestive Evidence Criteria 3.1	Microcephaly 1, primary, autosomal recessive	251200	AR
MTHFR	score 3	Suggestive Evidence Criteria 3.1	{Neural tube defects, susceptibility to}; {Schizophrenia, susceptibility to}; {Thromboembolism, susceptibility to}; {Vascular disease, susceptibility to}; Homocystinuria due to MTHFR deficiency	601634; 181500; 188050; NA; 236250	AR; AD; AD; NA; AR
POMT1	score 3	Suggestive Evidence Criteria 3.1	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 1; Muscular dystrophy-dystroglycanopathy (congenital with mental	236670; 1 613155; 609308	AR; AR; AR

			retardation), type B, 1; Muscular dystrophy-dystroglycanopathy (limb-girdle), type C, 1		
RAB39B	score 3	Suggestive Evidence Criteria 3.1	Mental retardation, X-linked 72; Waisman syndrome	300271; 311510	XLR; XLR
SLC6A8	score 3	Suggestive Evidence Criteria 3.1	Cerebral creatine deficiency syndrome 1	300352	XLR
SLC9A9	score 3	Suggestive Evidence Criteria 3.1	{?Autism susceptibility 16}	613410	NA
TSPAN7	score 3	Suggestive Evidence Criteria 3.1	Mental retardation, X-linked 58	300210	XLR
ATP1A3	score 3S	Suggestive Evidence, SyndromicCriteria 3.1, Syndromic	Alternating hemiplegia of childhood 2; CAPOS syndrome; Dystonia-12	614820; 601338; 128235	AD; AD; AD
CEP290	score 3S	Suggestive Evidence, SyndromicCriteria 3.1, Syndromic	?Bardet-Biedl syndrome 14; Joubert syndrome 5; Leber congenital amaurosis 10; Meckel syndrome 4; Senior-Loken syndrome 6	615991; 610188; 611755; 611134; 610189	AR; AR; NA; AR; AR
MEF2C	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Mental retardation, stereotypic movements, epilepsy, and/or cerebral malformations	613443	AD
NR2F1	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Bosch-Boonstra-Schaaf optic atrophy syndrome	615722	AD
NTNG1	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	NA	NA	NA
RNF135	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	NA	NA	NA
RPS6KA3	score 3S	Suggestive Evidence, Syndromic	Coffin-Lowry syndrome; Mental retardation, X-linked 19	303600; 300844	XLD; XLD

		Criteria 3.1, Syndromic			
SATB2	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Glass syndrome	612313	AD
SMC3	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Cornelia de Lange syndrome 3	610759	AD
SYNE1	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Arthrogryposis multiplex congenita, myogenic type; Emery-Dreifuss muscular dystrophy 4, autosomal dominant; Spinocerebellar ataxia, autosomal recessive 8	618484; 612998; 610743	AR; AD; AR
HDAC4	score 3S	Suggestive Evidence, Syndromic Criteria 3.1, Syndromic	Brachydactyly mental retardation syndrome		
ALG6	score S	Syndromic Syndromic	Congenital disorder of glycosylation, type Ic	603147	AR
GAMT	score S	Syndromic Syndromic	Cerebral creatine deficiency syndrome 2	612736	AR
GATM	score S	Syndromic Syndromic	Cerebral creatine deficiency syndrome 3	612718	AR
HDAC8	score S	Syndromic Syndromic	Cornelia de Lange syndrome 5	300882	XLD
RAD21	score S	Syndromic Syndromic	?Mungan syndrome; Cornelia de Lange syndrome 4	611376; 614701	AR; AD
ACSL4			Mental retardation, X-linked 63	300387	XLD
ACY1			Aminoacylase 1 deficiency	609924	AR

ADGRG1			Polymicrogyria, bilateral frontoparietal; Polymicrogyria, bilateral perisylvian	606854; 615752	AR; NA
AHI1			Joubert syndrome 3	608629	AR
ALDH7A1			Epilepsy, pyridoxine-dependent	266100	AR
AP1S2			Mental retardation, X-linked syndromic 5	304340	XLR
ASL			Argininosuccinic aciduria	207900	AR
BBS10			Bardet-Biedl syndrome 10	615987	AR
BCKDHA			Maple syrup urine disease, type Ia	248600	AR
BCKDHB			Maple syrup urine disease, type Ib	248600	AR
BTD			Biotinidase deficiency 253260	253260	AR
C12orf57			Temtamy syndrome	218340	AR
CAMTA1			Cerebellar ataxia, nonprogressive, with mental retardation	614756	AD
CBS			Thrombosis, hyperhomocysteinemic; Homocystinuria, B6-responsive and nonresponsive types	236200; 236200	AR; AR
CC2D2A			COACH syndrome; Joubert syndrome 9; Meckel syndrome 6	216360; 612285; 612284	AR; AR; AR

CDON			Holoprosencephaly 11	614226	AD
CHKB			Muscular dystrophy, congenital, megaconial type	602541	AR
COG5			Congenital disorder of glycosylation, type Iii	613612	NA
DAG1			Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 9; Muscular dystrophy-dystroglycanopathy (limb-girdle), type C, 9	616538; 613818	AR; AR
DBT			Maple syrup urine disease, type II	248600	AR
DCHS1			Mitral valve prolapse 2; Van Maldergem syndrome 1	607829; 601390	AD; AR
DCX			Lissencephaly, X-linked; Subcortical laminal heterotopia, X-linked	300067; 300067	XL; XL
DLG3			Mental retardation, X-linked 90	300850	XLR
DPYS			Dihydropyrimidinuria	222748	AR
DYM			Dyggve-Melchior-Clausen disease; Smith-McCort dysplasia	223800; 607326	AR; AR
FGFR2			Antley-Bixler syndrome without genital anomalies or disordered steroidogenesis; Apert syndrome; Beare-Stevenson cutis gyrata syndrome; Bent bone dysplasia syndrome; Craniofacial-skeletal-dermatologic dysplasia; Craniosynostosis, nonspecific; Crouzon syndrome; Gastric cancer, somatic; Jackson-	207410; 101200; 123790; 614592; 101600; NA; 123500; 613659; 123150; 149730; 101600; 101400;	AD; AD; AD; AD; AD; NA; AD; NA; AD; AD; AD; NA; NA

			Weiss syndrome; LADD syndrome; Pfeiffer syndrome; Saethre-Chotzen syndrome; Scaphocephaly and Axenfeld-Rieger anomaly; Scaphocephaly, maxillary retrusion, and mental retardation	NA; 609579	
FH			Fumarase deficiency; Leiomyomatosis and renal cell cancer	606812; 150800	AR; AD
FOLR1			Neurodegeneration due to cerebral folate transport deficiency	613068	AR
FTSJ1			Mental retardation, X-linked 9/44	309549	XLR
GDI1			Mental retardation, X-linked 41	300849	XLD
GLYCTK			D-glyceric aciduria	220120	AR
GNS			Mucopolysaccharidosis type IIID	252940	AR
GRIA3			Mental retardation, X-linked 94	300699	XLR
GSS			Glutathione synthetase deficiency; Hemolytic anemia due to glutathione synthetase deficiency	266130; 231900	AR; AR
HCFC1			Mental retardation, X-linked 3 (methylmalonic acidemia and homocysteinemia, cblX type)	309541	XLR
HCN1			Epileptic encephalopathy, early infantile, 24; Generalized epilepsy with febrile seizures plus, type 10	615871; 618482	AD; AD
HEPACAM			Megalencephalic leukoencephalopathy with subcortical cysts 2A; Megalencephalic leukoencephalopathy with subcortical cysts 2B,	613925; 613926	AR; AD

			remitting, with or without mental retardation		
HERC2			[Skin/hair/eye pigmentation 1, blond/brown hair]; [Skin/hair/eye pigmentation 1, blue/nonblue eyes]; Mental retardation, autosomal recessive 38	227220; 227220; 615516	AR; AR; AR
HGSNAT			Mucopolysaccharidosis type IIIC (Sanfilippo C); Retinitis pigmentosa 73	252930; 616544	AR; AR
HOXA1			Athabaskan brainstem dysgenesis syndrome; Bosley-Salih-Alorainy syndrome	601536; 601536	NA; NA
HPD			Hawkinsinuria; Tyrosinemia, type III	140350; 276710	AD; AR
HUWE1			Mental retardation, X-linked syndromic, Turner type	309590	XL
KCNH1			Temple-Baraitser syndrome; Zimmermann-Laband syndrome 1	611816; 135500	AD; AD
KIAA0196			Ritscher-Schinzel syndrome 1; Spastic paraplegia 8, autosomal dominant	220210; 603563	AR; AD
KIF7			?Al-Gazali-Bakalinova syndrome; ?Hydrolethalus syndrome 2; Acrocallosal syndrome; Joubert syndrome 12	607131; 614120; 200990; 200990	AR; AR; AR; AR
KMT2D			Kabuki syndrome 1	147920	AD
KRAS			Arteriovenous malformation of the brain, somatic; Bladder cancer, somatic; Breast cancer, somatic; Cardiofaciocutaneous syndrome 2; Gastric cancer, somatic; Leukemia, acute myeloid, somatic; Lung cancer, somatic; Noonan syndrome 3; Oculoectodermal syndrome, somatic; Pancreatic	108010; 109800; 114480; 615278; 137215; 601626; 211980; 609942; 600268; 260350; 614470; 163200	NA; NA; NA; NA; NA; NA; NA; AD; NA; NA; AD; NA

			carcinoma, somatic; RAS-associated autoimmune leukoproliferative disorder; Schimmelpenning-Feuerstein-Mims syndrome, somatic mosaic		
L1CAM			Corpus callosum, partial agenesis of; CRASH syndrome; Hydrocephalus due to aqueductal stenosis; Hydrocephalus with congenital idiopathic intestinal pseudoobstruction; Hydrocephalus with Hirschsprung disease; MASA syndrome	304100; 303350; 307000; 307000; 307000; 303350	XLR; XLR; XLR; XLR; XLR; XLR
L2HGDH			L-2-hydroxyglutaric aciduria	236792	AR
MAN1B1			Mental retardation, autosomal recessive 15	614202	AR
MAP2K1			Cardiofaciocutaneous syndrome 3	615279	NA
MED12			Lujan-Fryns syndrome; Ohdo syndrome, X-linked; Opitz-Kaveggia syndrome	309520; 300895; 305450	XLR; XLR; XLR
METTL23			Mental retardation, autosomal recessive 44	615942	AR
MKKS			Bardet-Biedl syndrome 6; McKusick-Kaufman syndrome	605231; 236700	AR; AR
MLL			Leukemia, myeloid/lymphoid or mixed-lineage; Wiedemann-Steiner syndrome	159555; 605130	AD; AD
MPLKIP			Trichothiodystrophy 4, nonphotosensitive	234050	AR
NAGLU			?Charcot-Marie-Tooth disease, axonal, type 2V; Mucopolysaccharidosis type IIIB (Sanfilippo B)	616491; 252920	AD; AR

NDN			Prader-Willi syndrome	176270	AD
NDP			Exudative vitreoretinopathy 2, X-linked; Norrie disease	305390; 310600	XLD, XLR; XLR
NFIX			Marshall-Smith syndrome; Sotos syndrome 2	602535; 614753	AD; AD
NSUN2			Mental retardation, autosomal recessive 5	611091	AR
OCRL			Dent disease 2; Lowe syndrome	300555; 309000	XLR; XLR
PCCA			Propionicacidemia	606054	AR
PCCB			Propionicacidemia	606054	AR
PHF6			Borjeson-Forssman-Lehmann syndrome	301900	XLR
PHF8			Mental retardation syndrome, X-linked, Siderius type	300263	XLR
PIGV			Hyperphosphatasia with mental retardation syndrome 1	239300	AR
PIK3R2			Megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 1	603387	AD
PQBP1			Renpenning syndrome	309500	XLR
PRKD1			Congenital heart defects and ectodermal dysplasia	617364	AD
PRSS12			Mental retardation, autosomal recessive 1	249500	AR

ROGDI			Kohlschutter-Tonz syndrome	226750	AR
RPGRIP1L			COACH syndrome; Joubert syndrome 7; Meckel syndrome 5	216360; 611560; 611561	AR; AR; AR
SGSH			Mucopolysaccharidosis type IIIA (Sanfilippo A)	252900	AR
SHH			Holoprosencephaly 3; Microphthalmia with coloboma 5; Schizencephaly; Single median maxillary central incisor	142945; 611638; 269160; 147250	AD; AD; NA; AD
SLC16A2			Allan-Herndon-Dudley syndrome	300523	XL
SLC17A5			Salla disease; Sialic acid storage disorder, infantile	604369; 269920	AR; AR
SLC1A2			Epileptic encephalopathy, early infantile, 41	617105	AD
SLC35C1			Congenital disorder of glycosylation, type IIc	266265	AR
SMARCA2			Nicolaides-Baraitser syndrome	601358	AD
SMC1A			Cornelia de Lange syndrome 2	300590	XLD
SNX14			Spinocerebellar ataxia, autosomal recessive 20	616354	AR
SOX10			PCWH syndrome; Waardenburg syndrome, type 2E, with or without neurologic involvement; Waardenburg syndrome, type 4C	609136; 611584; 613266	AD; AD; AD
SOX11			Coffin-Siris syndrome 9	615866	AD

STAMBP			Microcephaly-capillary malformation syndrome	614261	AR
TAF1			Dystonia-Parkinsonism, X-linked; Mental retardation, X-linked, syndromic 33	314250; 300966	XLR; XLR
TBC1D20			Warburg micro syndrome 4	615663	AR
TCN2			Transcobalamin II deficiency	275350	AR
TUBA1A			Lissencephaly 3	611603	AD
TUBG1			Cortical dysplasia, complex, with other brain malformations 4	615412	AD
TUSC3			Mental retardation, autosomal recessive 7	611093	AR
UPB1			Beta-ureidopropionase deficiency	613161	AR
USP9X			Mental retardation, X-linked 99; Mental retardation, X-linked 99, syndromic, female-restricted	300919; 300968	XLR; XLD
ZEB2			Mowat-Wilson syndrome	235730	AD
ZNF711			Mental retardation, X-linked 97	300803	XL
FMN2			Mental retardation, autosomal recessive 47	616193	AR
DEPDC5			Epilepsy, familial focal, with variable foci 1	604364	AD
SIK1			Epileptic encephalopathy, early infantile, 30	616341	AD

SHROOM4			Stocco dos Santos X-linked mental retardation syndrome	300434	XL

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Secretaría Xeral

Comité Ético de Investigación Clínica de Galicia
Edificio Administrativo de San Lázaro
15781 SANTIAGO DE COMPOSTELA
Tlf: 881 546425 Fax: 881 541804
ceic@sergas.es

galicia

DICTAMEN DEL COMITÉ ÉTICO DE INVESTIGACIÓN CLÍNICA DE GALICIA

Paula M. López Vázquez, Secretaria del Comité Ético de Investigación Clínica de Galicia

CERTIFICA:

Que este Comité evaluó en su reunión del día 28/06/2012 el estudio:

Título: Contribución a la búsqueda de las causas genéticas de los Trastornos del espectro autista con retraso mental
Promotor: Ángel Carracedo Álvarez
Código de Registro CEIC de Galicia: 2012/098

Y que este Comité de conformidad con sus Procedimientos Normalizados de Trabajo y tomando en cuenta los requisitos éticos, metodológicos y legales exigibles a los estudios de investigación con seres humanos, sus muestras o registros, emite un **DICTAMEN FAVORABLE** al estudio propuesto y que se llevará a cabo en:

Centros	Investigadores principales
Fundación Pública Galega de Medicina Xenómica	Ángel Carracedo Álvarez

En Santiago de Compostela a 11 de julio de 2012
La Secretaria

Paula M. López Vázquez

HOJA DE INFORMACIÓN AL PARTICIPANTE EN UN ESTUDIO DE INVESTIGACIÓN

TÍTULO: Contribución a la búsqueda de las causas genéticas de los Trastornos del Espectro Autista.

INVESTIGADOR: Ángel Carracedo Álvarez, Director Ejecutivo de la Fundación Pública Galega de Medicina Xenómica

Este documento tiene por objeto ofrecerle información sobre un estudio de investigación en el que se le invita a participar. Este estudio se está realizando en la Fundación Pública Galega de Medicina Xenómica y fue aprobado por el Comité Ético de Investigación Clínica de Galicia.

Si decide participar en el mismo, debe recibir información personalizada del investigador, leer antes este documento y hacer todas las preguntas que necesite para comprender los detalles sobre el mismo. Si así lo desea, puede llevar el documento, consultarlo con otras personas, y tomarse el tiempo necesario para decidir si participar o no.

La participación en este estudio es completamente voluntaria. Vd. puede decidir no participar o, si acepta hacerlo, cambiar de opinión retirando el consentimiento en cualquier momento sin obligación de dar explicaciones. Le aseguramos que esta decisión no afectará a la relación con su médico ni a la asistencia sanitaria a la que Vd. tiene derecho.

¿Cual es el propósito del estudio?

Los Trastornos del Espectro Autista (TEA), se caracterizan por ser alteraciones en el desarrollo y funcionamiento cerebral pero las causas que los producen aún no se conocen. Se ha demostrado una alta implicación de factores genéticos en los TEA (80%, en relación con el ambiente), sin embargo en la mayoría de los casos no es posible determinar la alteración genética responsable.

Los avances en investigación genómica han cambiado el panorama de la investigación genética del autismo, posibilitando que hasta en una cuarta parte de los casos se pueda encontrar una causa genética potencialmente determinante del cuadro.

En esta investigación vamos a emplear técnicas de alta resolución molecular (como son los microarrays o chips de ADN) para analizar cientos de segmentos de DNA distribuidos por todo el genoma humano. Esto posibilitará la detección de anomalías estructurales submicroscópicas que no serían detectables con las técnicas genéticas tradicionales (cariotipo y X- frágil convencional).

Para proceder con este estudio, necesitamos analizar el material genético de personas diagnosticadas con TEA, para lo cual será necesario obtener una muestra de su sangre. También será fundamental realizar una evaluación clínica y neuropsicológica exhaustiva de los sujetos a estudio, con vistas a establecer grupos o fenotipos lo más definidos posibles y facilitar así una posible correlación genética.

¿Por qué me ofrecen participar a mí?

La selección de las personas invitadas a participar depende de unos criterios que están descritos en el protocolo de la investigación. Estos criterios sirven para seleccionar a la población en la que se responderá el interrogante de la investigación. Vd. es invitado a participar porque cumple eses criterios. Tales criterios consisten en presentar un diagnóstico de trastorno del espectro autista (TEA) y tener una edad comprendida entre los 3 y los 18 años.

El hecho de presentar discapacidad intelectual y/o estar también diagnosticados de otros problemas médicos, neurológicos y psiquiátricos no supondrá impedimento alguno para participar en este estudio.

Se espera que participen 200 personas en este estudio.

¿En qué consiste mi participación?

Si acepta participar en la presente investigación, le será requerida una muestra de unos 10 ml de sangre, que será utilizada para extraer el ADN. La muestra de ADN pasará a formar parte de una colección de muestras disponible para investigar las bases genéticas del autismo. Será conservada y podrá ser usada posteriormente en otros estudios, salvo que usted manifieste lo contrario señalándolo al final de este documento. Además, se recogerán en un formulario los datos personales y los antecedentes familiares. Esto permitirá a los científicos disponer de muestras para investigar: 1) qué genes influyen en el desarrollo de los TEA o en la protección del desarrollo de los mismos, 2) cómo están influenciados estos genes por el entorno y, 3) qué genes influyen en la eficacia/resistencia a tratamientos específicos.

Además de esta extracción de sangre, se le hará una entrevista con instrumentos de valoración clínica (entrevistas diagnósticas y cuestionarios) y cognitiva (pruebas de atención, memoria, funciones ejecutivas, capacidad intelectual,…) que completarán y darán apoyo a la investigación genética. En un primer momento, bastará con una entrevista clínica preliminar (en torno a 1 hora de duración) que confirme el diagnóstico de TEA. Únicamente aquellos sujetos que presenten hallazgos genéticos significativos serán requeridos para una valoración clínica más detallada.

Las muestras que nos proporcione estarán **codificadas** mediante un código de seguridad, que servirá para asociar los datos con Vd. en los supuestos en que esta identificación fuere necesaria para su beneficio y de conformidad con lo previsto en la Ley 14/2007, de Investigación Biomédica, y el y RD1717/2011. Dicho código estará en poder únicamente del personal responsable de la colección. Los investigadores, por tanto, no podrán conocer su identidad personal, pero sí datos como su sexo o edad, manteniendo siempre la debida confidencialidad conforme a la legislación vigente.

A partir de las muestras donadas se aislarán las células contenidas en las mismas y se extraerán los ácidos nucleicos sobre los que se realizarán los análisis genéticos; además se obtendrá suero y plasma para estudios fenotípicos.

El equipo de Investigación de la Fundación Pública Galega de Medicina Xenómica mantendrá la titularidad de las muestras. Los productos obtenidos de las muestras y los datos asociados a las mismas se archivarán y quedarán custodiados, por un periodo mínimo de 2 años, en las instalaciones de la Fundación Pública Galega de Medicina Xenómica.

Los análisis genéticos y fenotípicos realizados serán tratados estadísticamente, exclusivamente para fines de investigación biomédica de acuerdo a lo descrito anteriormente.

El promotor o el investigador pueden decidir finalizar el estudio antes de lo previsto o interrumpir su participación por aparición de nueva información relevante, por motivos de seguridad, o por incumplimiento de los procedimientos del estudio.

¿Qué riesgos o inconvenientes tiene?

La donación de sangre apenas tiene efectos secundarios; como mucho, podría aparecer un pequeño hematoma en la zona de punción, que desaparecería en unos días.

El no proporcionar su consentimiento no tendrá ninguna consecuencia en su tratamiento clínico.

¿Obtendré algún beneficio por participar?

Usted proporcionará las muestras de forma totalmente voluntaria.

Su participación no tiene ninguna compensación económica pero tampoco supone gasto alguno para usted. Su atención médica no se verá afectada por el hecho de que participe o no en esta investigación.

En el caso de hallarse información que pudiera ser clínicamente relevante para usted o su familia, se le ofrecerá la posibilidad de recibir asesoramiento genético al respecto.

Estimamos que entre un 20 y 25 % de los casos podremos encontrar una potencial causa genética (heredada o no ya que hay muchas mutaciones que surgen "de novo) y que en algunos casos sabremos su significado y en otros no.

El resultado de esta investigación puede desvelar una mutación (alteración genética) responsable del trastorno (TEA) en el caso de su familia. Este conocimiento puede ser de gran importancia para sus familiares y otras familias afectas. Los resultados de esta investigación que puedan ser de beneficio para usted o su familia le serán comunicados en un informe.

Como resultado del análisis genético podría detectarse también información inesperada relativa a otras enfermedades médicas. Únicamente será comunicada aquella información con relevancia clínica y potencialmente beneficiosa para el sujeto/sujetos en cuestión.

Si usted no desea recibir información sobre los resultados del análisis genético deberá marcar con una X las casillas correspondientes en el Consentimiento Informado. Un resultado negativo no excluye que el síndrome sea heredable, ya que podría estar causado por una mutación no analizada. Tal y como mencionamos anteriormente, como consecuencia del análisis también es posible que se identifiquen cambios genéticos de significado no claro.

De todas formas es importante tener en cuenta que al ser éste un estudio de investigación, el beneficio redundará más a nivel global y científico que personal. Sin embargo los conocimientos obtenidos gracias a los estudios llevados a cabo a partir de sus muestras y datos de las otras muchas personas que participen, supondrán una valiosa fuente de información que revertirá en un mejor conocimiento de estos trastornos posibilitando que en un futuro se desarrollen nuevos métodos de diagnóstico y tratamiento.

¿Recibiré la información que se obtenga del estudio?

Si Vd. lo desea, se le facilitará un resumen de los resultados del estudio.

También podrá recibir los resultados de las pruebas que se le practiquen si así lo solicita. Estos resultados pueden no tener aplicación clínica ni una interpretación clara, por lo que, si quiere disponer de ellos, deberían ser comentados con un miembro del equipo de investigación responsable del estudio.

¿Se publicaran los resultados de este estudio?

Los resultados de este estudio serán publicados en publicaciones científicas para su difusión, pero no se transmitirá ningún dato que pueda llevar a la identificación de los participantes.

¿Cómo se protegerá la confidencialidad de mis datos?

El tratamiento, comunicación y cesión de sus datos se hará conforme a lo dispuesto por la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. En todo momento, Vd. podrá acceder a sus datos, corregirlos o cancelarlos.

Sólo el equipo investigador y las autoridades sanitarias, que tienen deber de guardar la confidencialidad, tendrán acceso a todos los datos recogidos por el estudio. Se podrá transmitir a terceros información que no pueda ser identificada. En el caso de que alguna información sea transmitida a otros países, se realizará con un nivel de protección de los datos equivalente, como mínimo, al exigido por la normativa de nuestro país.

Los profesionales responsables de la custodia de las muestras y datos asociados a las mismas garantizarán que la identidad del donante no sea accesible a los investigadores, cuando datos o muestras sean transferidas a un investigador dentro de ese centro o de otro centro.

En algunas ocasiones podría ser de utilidad realizar el análisis genético a otros miembros de su familia. En ningún caso serán contactadas otras personas de su familia con este propósito sin su permiso.

La información será almacenada en soporte informático. Los datos registrados serán tratados estadísticamente, de forma codificada, para los fines de investigación científica que se describieron anteriormente.

¿Qué ocurrirá con las muestras obtenidas?

El responsable de la custodia de las muestras es el Dr. Ángel Carracedo Álvarez, y serán almacenadas en la Fundación Pública Galega de Medicina Xenómica, el tiempo necesario para terminar con esta línea de investigación.

Al firmar consentimiento, usted puede optar por autorizar a que se cedan las muestras biológicas a terceros con destino a otros proyectos de investigación relacionados o no con el área de investigación. En ese caso, sus muestras y los datos asociados serán guardados de forma codificada, que quiere decir que poseen un código que se puede relacionar, mediante una información, con la identificación del donante. Esta información está a cargo del investigador principal y sólo pueden acceder a ella los miembros del equipo investigador, representantes del promotor del estudio y las autoridades sanitarias en ejercicio de sus funciones.

Sólo se ceden muestras para su utilización en proyectos de investigación que hayan sido valorados positivamente por un Comité de Ética de la Investigación, y autorizados por la autoridad sanitaria.

Los datos clínicos relevantes para cruzar con la información genética también podrán ser cedidos.

Ningún dato de identificación personal será cedido.

¿Quién me puede dar más información?

Para más información puede contactar con la responsable clínica de este proyecto de investigación (Lorena Gómez Guerrero) en el teléfono de la FPGMX (981951491) o bien dirigirse a la siguiente dirección de correo electrónico: proyectogenetico@fundacionmariajosejove.org

Revocación del consentimiento.

Es usted libre de cambiar de opinión en cualquier momento y revocar el presente consentimiento, caso en el que la muestra será destruida. Esto en ningún caso conllevará ningún perjuicio para usted.

Muchas gracias por su colaboración.

Leí y comprendí el presente documento. Todas mis preguntas sobre la investigación fueron respondidas de forma satisfactoria.

D./Dña. ,
tutor / responsable legal de (NOMBRE DEL NIÑO/A:
CÓDIGO DEL NIÑO/A:) doy mi consentimiento para participar en la presente investigación.

El/la participante,	El / la representante / tutor legal,	El/la investigador/a,
Fdo.:	Fdo:	Fdo.:
Fecha:	Fecha:	Fecha:

DOCUMENTO DE CONSENTIMIENTO PARA LA PARTICIPACIÓN EN UN ESTUDIO DE INVESTIGACIÓN

TÍTULO: **Contribución a la búsqueda de las causas genéticas de los Trastornos del Espectro Autista**

Yo, (participante o, en su caso, representante / tutor legal), _____

_____

☐ He leído la hoja de información al participante del estudio arriba mencionado que se me entregó, he podido hablar con el investigador responsable y hacerle todas las preguntas sobre el estudio necesarias para comprender sus condiciones y considero que he recibido suficiente información sobre el estudio.

☐ Comprendo que mi participación es voluntaria, y que puedo retirarme del estudio cuando quiera, sin tener que dar explicaciones y sin que esto repercuta en mis cuidados médicos.

☐ Accedo a que se utilicen mis datos en las condiciones detalladas en la hoja de información al participante.

☐ Presto libremente mi conformidad para participar en el estudio.

☐ Autorizo al acceso a toda la información clínica del participante recogida en el servicio gallego de salud

Respeto a la conservación y utilización futura de los datos y/o muestras detallada en la hoja de información al participante,

☐ Accedo a que mis datos y/o muestras se conserven formando parte de una colección de ADN para la investigación en genética de los trastornos del espectro autista (TEA) en las condiciones mencionadas.

☐ Accedo a que los datos y/o muestras se conserven para usos posteriores en líneas de investigación relacionadas con la presente, y en las condiciones mencionadas.

En cuanto a los resultados de las pruebas realizadas,

☐ DESEO conocer los resultados de mis pruebas

☐ DESEO conocer los resultados de mis pruebas en caso de hallarse información médica relevante ajena a la naturaleza del estudio.

☐ REVOCO EL PRESENTE CONSENTIMIENTO

El/la participante,	El / la representante / tutor legal,	El/la investigador/a,
Fdo.:	Fdo:	**Fdo.:**
Fecha:	Fecha:	**Fecha:**

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA