FACULTADE DE MEDICINA
E ODONTOLOXÍA

Traballo de fin de grao

Descubrimento, caracterización fenotípica e funcional *in silico* de mutacións *de novo* en *PKD1* nunha cohorte de exoma de TEA.

Descubrimiento, caracterización fenotípica y funcional *in silico* de mutaciones *de novo* en *PKD1* en una cohorte de exoma de TEA.

In silico discovery, phenotypic and functional characterization of *de novo* mutations in *PKD1* in an ASD exome cohort.

Autor: Fernández Fernández, Pedro
Titor: Carracedo Álvarez, Ángel María
Cotitoras: Rodríguez Fontenla, María Cristina
Domínguez Alonso, Sara
Departamento: Departamento de Ciencias Forenses, Anatomía Patológica, Ginecología y Obstetricia, y Pediatría

Xuño de 2024

Traballo de Fin de Grao presentado na Facultade de Medicina e Odontoloxía da Universidade de Santiago de Compostela para a obtención do Grao en Medicina.

AGRADECIMIENTOS

Me gustaría expresar mi más sincero agradecimiento a mis cotutoras, Cristina Rodríguez y Sara Domínguez. Este trabajo fue posible gracias a su dedicación, orientación y paciencia. También quisiera extender mi gratitud a mi tutor Ángel Carracedo, por brindarme la oportunidad de llevar a cabo este proyecto, orientarme en el proceso y servir de inspiración como científico y persona. A mi familia, por su constante apoyo y comprensión durante este largo camino que fue la carrera de medicina.

“Knowledge speaks, but wisdom listens.”
Jimi Hendrix

Glosario de Abreviaturas

AA: Aminoácidos.
ADPKD: *Autosomal Dominant Polycystic Kidney Disease* (Enfermedad Renal Poliquística Autosómica Dominante).
ANNOVAR: *ANNOtate VARiation* (Herramienta bioinformática para la anotación de variantes).
ASC: *Autism Sequencing Consortium* (Consorcio de secuenciación de autismo).
BAP: *Broad Autism Phenotype* (Fenotipo amplio del autismo).
CNVs: *Copy Number Variations* (Variaciones en el número de copias).
dbSNP: *Database for Nonsynonymous SNPs' Functional Predictions* (Base de datos para las predicciones funcionales de SNPs no sinónimos).
DI: discapacidad intelectual.
DNMs: *De Novo Mutations* (Mutaciones de novo).
DSM: *Diagnostic and Statistical Manual of Mental Disorders* (Manual diagnóstico y estadístico de los trastornos mentales).
ExAC: *Exome Aggregation Consortium* (Consorcio de Agregación del Exoma).
GWAS: *Genome-Wide Association Study* (Estudio de asociación del genoma completo).
GTEx: *Genotype-Tissue Expression Project* (Proyecto de expresión de genotipo-tejido).
IA: Inteligencia Artificial.
NCBI: *National Center for Biotechnology Information* (Centro Nacional para la Información Biotecnológica)
NGS: *Next-Generation Sequencing* (Secuenciación de nueva generación)
PC1: Policistina-1
PB: pares de bases.
PKD1: *Polycystic Kidney Disease 1* (gen de la enfermedad renal poliquística 1).
SNVs: *Single Nucleotide Variants* (Variantes de nucleótido único).
TEA: Trastornos del Espectro Autista
TND: Trastornos del Neurodesarrollo
UCSC: *University of California, Santa Cruz* (Universidad de California en Santa Cruz)
VEP: *Variant Effect Predictor* (Predictor de efecto de variantes)
WES: *Whole-exome Sequencing* (Secuenciación del exoma completo)

ÍNDICE

RESUMEN

Antecedentes: Los trastornos del espectro autista (TEA) cuentan con una elevada prevalencia (hasta un 1%), son de los trastornos del neurodesarrollo (TND) más heredables que existen (80% de heredabilidad). Hasta la fecha, el gen *Polycystic Kidney Disease* 1 (*PKD1)*, implicado en la enfermedad renal poliquística autosómica dominante (ADPKD), nunca ha sido descrito como un gen de riesgo para los TEA.

Objetivo: Describir y estudiar 10 mutaciones *de novo* en *PKD1* halladas en individuos con TEA para determinar si este gen está implicado en estos TND. Caracterizar los probandos, anotar las variantes con herramientas bioinformáticas y predecir su impacto funcional con inteligencia artificial (IA) y *software* específicos.

Metodología: Se hallaron 10 mutaciones en *PKD1* (4 en una cohorte de exoma de 360 tríos y 6 pertenecientes a una cohorte del *Autism Sequencing Consortium).* Se llevó a cabo un análisis del fenotipo y genotipo de algunos probandos. Se estudió la expresión de *PKD1*. Se analizó la prevalencia de las variantes y se anotaron con la herramienta ANNOVAR para examinar su naturaleza y características. Se estudiaron las mutaciones a nivel funcional con tres herramientas que usan *machine learning*, dos de ellas (AlphaMissense y SpliceAI) son sistemas de IA.

Resultados: El análisis del fenotipo y genotipo de los probandos sitúa a las mutaciones en *PKD1* como mejores candidatas para explicar los TEA. Se objetivó que este gen se expresa 5,44 veces más en el cerebro que en el tejido renal. La anotación de las variantes reveló que una mutación se encuentra en un intrón, una en una zona de empalme o *splicing* y 8 en exones. De estas 8, 4 son silenciosas y 4 *missense*. La frecuencia poblacional de todas las mutaciones es mínima. La predicción del efecto funcional de las mutaciones *missense* fue distinta en Polyphen-2 y AlphaMissense. Polyphen-2 clasificó como probablemente benigna una variante, una como posiblemente maligna y dos como probablemente malignas. Las probabilidades de malignidad fueron 0.007, 0.925, 0.998 y 0.994 respectivamente. AlphaMissense consideró todas las variantes posiblemente benignas. SpliceAI indicó que la mutación en la zona de empalme tenía una probabilidad del 93% de ser maligna.

Conclusiones: El estudio de las variantes en *PKD1* sugiere que algunas mutaciones tienen alto potencial patogénico en los TEA (probabilidades de malignidad superiores al 92% en 4 casos). Si bien algunos resultados de las herramientas predictoras de patogenicidad son contradictorios, presentamos por primera vez *PKD1* como probable gen de riesgo en los TEA.

Palabras clave: Trastornos del espectro autista (TEA), *Polycystic Kidney Disease 1* (*PKD1)*, Policistina-1 (PC1), enfermedad renal poliquística autosómica dominante (ADPKD), Inteligencia artificial (IA), Polyphen-2, AlphaMissense, SpliceAI.

RESUMO

Antecedentes: Os trastornos do espectro autista (TEA) contan cunha elevada frecuencia (ate un 1%), son dos trastornos no neurodesenvolvemento máis herdables que existen (80% de heredabilidade). Ate a data, o xen *Polycystic Kidney Disease* 1 (*PKD1*), implicado na enfermidade renal poliquística autosómica dominante (ADPKD), nunca foi descrito como un xen de risco para os TEA.

Obxectivo: Describir e estudar 10 mutacións *de novo* achadas en *PKD1* nunha en individuos TEA para determinar se este xen está implicado nos TEA. Caracterizar fenotípicamente os probandos, anotar as variantes con ferramentas bioinformáticas e predicir o impacto funcional das variantes con intelixencia artifical (IA) e *software* específicos.

Metodoloxía: Atopáronse 10 mutacións en *PKD1* (4 nunha cohorte de exoma de 360 tríos e 6 pertencentes a unha cohorte do *Autism Sequencing Consortium*) . Levouse a cabo unha análise do fenotipo e xenotipo de algúns probandos. Estudouse a expresión de PKD1 nos tecidos. Analizouse a frecuencia das variantes e anotáronse coa ferramenta ANNOVAR para examinar a natureza e características das mutacións. O estudo das variantes a nivel funcional fíxose con tres ferramentas que usan machine learning, dúas delas (AlphaMissense e SpliceAI) son sistemas de IA.

Resultados: A análise do fenotipo e xenotipo dos probandos sitúa ás mutacións en *PKD1* como as mellores candidatas para explicar os TEA. O estudo de expresión amosou que este xen exprésase 5,44 veces máis no cerebro que no tecido renal. A anotación das variantes revelou que unha atópase nun intrón, unha nunha zoa de empalme e 8 en exóns. Destas 8, 4 son silenciosas e 4 *missense*. A frecuencia poblacional de todas as mutacións é mínima. A predicción do efecto funcional das mutacións *missense* foi distinta en Polyphen-2 e AlphaMissense. Polyphen-2 clasificou como posiblemente benigna unha variante, unha como posiblemente maligna e dúas como probablemente malignas. As probabilidades de malignidade foron 0.007, 0.925, 0.998 e 0.994 respectivamente. AlphaMissense considerou todas as variantes posibelmente benignas. SpliceAI indicou que a mutación na zoa de empalme tiña una probabilidade do 93% de ser maligna.

Conclusións: O estudo das variantes en *PKD1* suxire que algunhas mutación teñen alto potencial patoxénico nos TEA (probabilidades de malignidade superiores ao 92% en 4 ocasións). Aínda que algúns resultados das ferramentas predictoras de patoxenicidade son contraditorios, presentamos como primeira vez *PKD1* como probable xen de risco nos TEA.

Palabras chave: Trastornos do espectro autista (TEA), *Polycystic Kidney Disease 1* (*PKD1)*, Policistina-1 (PC1), enfermidade renal poliquística autosómica dominante (ADPKD), Intelixencia artificial (IA), Polyphen-2, AlphaMissense, SpliceAI.

ABSTRACT

Background: Autism spectrum disorders (ASD) have a high prevalence (up to 1%) and are among the most heritable neurodevelopmental disorders (80% of heritability). To date, Polycystic Kidney disease (*PKD1)*, which causes autosomal dominant polycystic kidney disease (ADPKD), has never been described as a risk gene for ASD.

Objective: To describe and study 10 *de novo* mutations in *PKD1* found in individuals with ASD to determine if this gene is implicated in ASD. To characterize the probands, annotate the variants with bioinformatics tools and predict the functional effect using artificial intelligence (AI).

Methodology: 10 mutations were identified in *PKD1* (4 in the exome cohort of 360 trios and the other 6 from the cohort of the Autism Sequencing Consortium). Phenotype and genotype analyses were performed on some probands. PKD1 expression was studied. The prevalence of the variants was analyzed and annotated using the ANNOVAR tool to examine their characteristics. Functional analysis of the mutations was performed using 3 machine learning tools, two of which (AlphaMissense and SpliceAI) are AI systems.

Results: Phenotype and genotype analysis of the probands suggests that mutations in PKD1 are the strongest candidates to explain ASD in the probands. It was observed that this gene is expressed 5.44 times more in the brain than in renal tissue. Variants annotation revealed that one mutation is in an intron, one in a splicing site, and 8 in exons (4 of which are silent and 4 missense). The frequency of all mutations is really low. The prediction of the functional effect of the missense mutations varied between Polyphen-2 and AlphaMissense. Polyphen-2 classified one variant as probably benign, one as possibly pathogenic, and two as probably pathogenic, with malignancy probabilities of 0.007, 0.925, 0.998, and 0.994, respectively. AlphaMissense considered all variants possibly benign. SpliceAI indicated that the splice site mutation had a 93% probability of being harmful.

Conclusions: The variants study in *PKD1* suggests that some mutations have high pathogenic potential in ASD (the malignancy probabilities are above 92% in 4 cases). Although some results from pathogenicity prediction tools are contradictory, we present PKD1 for the first time as a potential risk gene in ASD.

Keywords: Autism Spectrum Disorders (ASD), Polycystic Kidney Disease 1 (*PKD1*), Polycystin-1 (PC1), Autosomal Dominant Polycystic Kidney Disease (ADPKD), Artificial Intelligence (AI), Polyphen-2, AlphaMissense, SpliceAI.

1. INTRODUCCIÓN

1.1. Trastornos del espectro autista

1.1.1. Definición

La palabra "**autismo**" deriva del término griego *autós* 'uno mismo', hace referencia a estar centrado en uno mismo o abstraído (1). Pese a tratarse de una palabra presente en el vocabulario desde hace milenios, hasta 1911 no aparece en el ámbito científico. El médico **Paul Bleuler** la usa para describir un síntoma de la esquizofrenia severa que hace que los pacientes se alejen de la realidad y se aíslen del mundo que les rodea. Alegaba que los individuos evitaban situaciones insatisfactorias y las reemplazaban con fantasías y alucinaciones (2).
La primera vez que se considera el autismo como un trastorno independiente es en 1943, cuando el psiquiatra **Leo Kanner** describe 11 casos de niños y niñas que compartían una serie de patrones de comportamiento distintivos. Estos niños y niñas presentaban falta de interacción social, dificultad en la comunicación, preferencia por la soledad y resistencia al cambio (3).
Hoy en día se prefiere el término **Trastornos del Espectro Autista (TEA)** ya que se trata de un conjunto de alteraciones del neurodesarrollo bastante heterogéneas. Estas comparten una instauración temprana de problemas en las habilidades de comunicación social y comportamientos repetitivos y rígidos (4).
Es importante recalcar la diversidad de trastornos que se incluyen dentro de esta categoría. Es complicado discernir entre los distintos subgrupos y establecer los límites entre los individuos afectos y no afectos continúa siendo un desafío. Todo esto ha provocado que el diagnóstico y clasificación de los TEA se haya ido modificando a lo largo de la historia.
Destaca la contribución de **Lorna Wing** en la década de los 80. Insiste en la necesidad de considerar el autismo como un grupo grande de condiciones con ciertas características en común. También acuña el término "**síndrome de Asperger**", una forma de TEA de alto grado de funcionamiento. Se basa en una publicación de Hans Asperger de 1944 para describir este fenotipo de pacientes.
Los trabajos de Wing hicieron que otros científicos comenzasen a investigar las diversas formas de presentación de los TEA y su asociación con otros trastornos del neurodesarrollo (TND).
En el campo de la psiquiatría, el "*The Diagnostic and Statistal Manual of Mental Disorders"* o **DSM** es mundialmente reconocido como manual de referencia para clasificación y diagnóstico de diversos trastornos. El **síndrome de Asperger** no se incluye en el DSM hasta su cuarta edición (1994), considerándose un trastorno independiente de los TEA. Más tarde, en 2013, con el DSM-V, este síndrome entra a formar parte de la misma categoría que los TEA.
También cabe destacar la aparición de nuevos conceptos como **BAP** *(Broad Austism Phenothype)*. El BAP incluye a individuos, normalmente familiares de personas diagnosticadas con TEA, que no cumplen todos los criterios diagnósticos, pero sí comparten ciertos síntomas o rasgos de personalidad (5).
Todo esto es una muestra de la gran plasticidad del concepto TEA a lo largo de la historia. Se trata de un término en constante revisión y de difícil caracterización.

1.1.2. Comorbilidades

Los TEA se **asocian a una gran cantidad de patologías y enfermedades genéticas**. Estas comorbilidades, sumadas a la discapacidad intelectual, explican mejor el aumento del riesgo de mortalidad que el trastorno en sí mismo.Muchas de ellas son **neurológicas**. Se ha descrito, por ejemplo, su asociación con hidrocefalia, macrocefalia, epilepsia, parálisis cerebral y migraña.

Hasta un 60% de los niños con TEA tienen un electroencefalograma anormal y alrededor de un 10-30% tienen epilepsia (6,7).
La **discapacidad intelectual** (DI) y los TEA guardan una estrecha correlación. Se observa que alrededor del 70% de los individuos diagnosticados con TEA presentan alguna forma de DI. A su vez, hasta el 40% de las personas con DI pueden ser diagnosticadas con TEA(6).
Los individuos con TEA también sufren de numerosas **enfermedades psiquiátricas** entre las que se encuentran trastorno bipolar, depresión, trastorno obsesivo-compulsivo (TOC), ansiedad y trastorno por déficit de atención e hiperactividad (TDAH).
Otra comorbilidad es la **disfunción del sistema nervioso autónomo**. Suele haber una hiperfunción del sistema nervioso simpático y una disminución de la actividad parasimpática.
Existe una dificultad para el control de esfínteres, que se manifiesta más frecuentemente como retraso en el control de la micción durante la niñez. Hecho que provoca mayor riesgo de vergüenza pública y pérdida de autoestima.
Además, entre el 46% y el 84% de los individuos con TEA padecen **patologías gastrointestinales**. Algunas de ellas son estreñimiento, diarrea, reflujo gastroesofágico, enfermedad inflamatoria intestinal e incluso alergias alimentarias.
Las **alteraciones del sueño** disminuyen la calidad de vida de los sujetos y llegan a estar presentes hasta en el 80% de los individuos con TEA. Entre ellas la más común es el insomnio de conciliación, pero también puede haber despertares nocturnos y sonambulismo (8).
Hasta un 96% de los individuos con TEA padecen **alteraciones sensoriales,** que se pueden manifestar como hipo o hipersensibilidad ante estímulos auditivos, táctiles y visuales. Destaca la hipersensibilidad auditiva, que se caracteriza por una reacción exagerada ante sonidos cotidianos que la mayor parte de personas pueden tolerar fácilmente. También está alterada la capacidad para discernir entre sonidos relevantes del ruido de fondo.
Todo esto empobrece la calidad de vida de los individuos y puede justificar el retraso en la adquisición del lenguaje (por el procesamiento auditivo), la dificultad para leer emociones en los rostros (procesamiento visual) e incluso los problemas de incontinencia de esfínteres (9,10).
En conclusión, los TEA presentan numerosas y diversas comorbilidades. Es crucial conocerlas y tratarlas ya que afectan al buen pronóstico del trastorno.

1.1.3. Diagnóstico

Actualmente, para el diagnóstico de los TEA, el pilar fundamental continúa siendo la evaluación clínica y utilización de los criterios diagnósticos de los manuales DSM-V y CIE-10 (4). Un resumen del DSM-V se recoge en la siguiente tabla.

CRITERIOS DIAGNÓSTICOS DE LOS TEA SEGÚN DSM-V	
A	Deficiencias constantes en la comunicación y en la interacción social manifestado por lo siguiente, en el momento actual o por los antecedentes.
	1- Deficiencias en la reciprocidad socioemocional (acercamiento social anormal, fracaso en la conversación…).
	2- Deficiencias en las conductas comunicativas no verbales (poca integración verbal/no verbal, anomalías del contacto visual).
	3- Deficiencias en el desarrollo, mantenimiento y comprensión de las relaciones sociales (dificultades para ajustar el comportamiento al contexto, dificultad para compartir juegos o hacer amigos, ausencia de interés por otras personas…).

B	Patrones restrictivos y repetitivos de comportamiento, actividades o intereses, que se manifiestan en 2 o más de los siguientes puntos, actualmente o por los antecedentes.
	1- Movimientos, uso de objetos o habla estereotipados y repetitivos (alineación de los juguetes, ecolalia, frases idiosincrásicas).
	2- Persistencia en la rutina, rigidez excesiva, patrones de comportamiento verbal y no verbal ritualizado.
	3- Intereses extremadamente limitados y fijos que son anormales en cuanto a su identidad o foco de interés.
	4- Hiper- o hiporreactividad ante estímulos sensoriales o interés inhabitual en aspectos sensoriales del entorno.
C	Los síntomas deben estar presentes desde las primeras fases del desarrollo (pueden no manifestarse o estar enmascarados hasta etapas posteriores de la vida).
D	Los síntomas causan un deterioro clínicamente significativo en lo social, laboral u otras áreas importantes.
E	Las alteraciones anteriores no se explican mejor por la discapacidad intelectual o por el retraso global del desarrollo. (Al coincidir habitualmente TEA y DI, para diagnosticarlos como comorbilidad, la comunicación social debe estar por debajo de lo esperable para el nivel general de desarrollo.)

Tabla 1: Criterios diagnósticos de los TEA según el manual DSM-V.

Todos los anteriores criterios diagnósticos se deben cumplir. Además, según el manual, se debe **especificar también el nivel de gravedad** y los siguientes **elementos**: Con o sin déficit intelectual acompañante; con o sin deterioro del lenguaje acompañante; si se asocia a una afección médica, genética o a un factor ambiental conocido; si se asocia a otro TND, mental o del comportamiento; si se acompaña de catatonía.

Los **niveles de gravedad** se establecen en base a los criterios A y B (capacidad de comunicación social y patrones de comportamiento). Son 3 niveles para cada dominio e indican el grado de asistencia requerida, siendo el nivel 1 "necesita ayuda", el 2 "necesita ayuda notable" y el 3 "necesita ayuda muy notable".

Asimismo, es necesario establecer un **diagnóstico diferencial** con patologías con características similares como el síndrome de Rett, mutismo selectivo y DI sin TEA (4).

1.1.4. Herramientas de detección

Una **detección temprana** es clave para una implementación precoz de terapias y tratamientos. Aunque cada vez es menos frecuente, todavía es común un diagnóstico tardío, a partir de los 3-4 años. Los síntomas pueden estar presentes incluso antes de los 12 meses. La pérdida del contacto ocular, el no reconocimiento del nombre y la poca interacción con los padres son algunos de los signos más precoces. Actualmente, se recomienda realizar un cribado entre los 18 y los 24 meses, sobre todo en aquellos niños en los que exista sospecha o riesgo elevado.

Existen **numerosos cuestionarios** para realizar el cribado u orientar el diagnóstico. Algunos pueden ser aplicados por los padres o cuidadores, uno de los más utilizados es el *Modified checklist for autism in toddlers* (**M-CHAT**), se puede usar entre los meses 16 y 30 y dura entre 5 y 10 minutos. Otro cuestionario simular es el *Social communication questionnaire* (**SCQ**), con la diferencia de que solo se puede realizar a partir de los 4 años de edad. En cuanto a las pruebas aplicadas por profesionales, destaca *The autism diagnostic interview-revised* (**ADI-R**). Esta prueba requiere entrenamiento por parte del profesional y lleva aproximadamente 2 horas, pues se trata de una **entrevista estructurada**. Otro cuestionario muy usado es *Childhood autism rating scale, first or second edition* (**CARS**, **CARS-2**), que también precisa experiencia por parte del evaluador, pero requiere menos tiempo (30 minutos) (3).

Aunque se han descrito alteraciones cerebrales, las **pruebas de neuroimagen** y electroencefalografía no se recomiendan de rutina, únicamente pueden ser útiles en niños con antecedentes de microcefalia, macrocefalia, regresión, epilepsia o exploración neurológica anormal. Todos los niños requieren una **historia médica detallada** y una **exploración física exhaustiva**. El análisis metabólico puede ser de utilidad en ciertos casos. Es crucial realizar pruebas genéticas. Debe incluir cariotipo, estudio de *FMR1* (para descartar X frágil) y un estudio de microarrays. Otro gen que se puede incluir es *MECP2*, con el fin de descartar síndrome de Rett. En aquellos casos en los que las pruebas anteriores sean negativas se puede hacer una secuenciación de exoma completo (3,11).

1.1.5. Epidemiología

Aunque la OMS estima que, de cada 160 niños, 1 tiene TEA, en las 2 últimas décadas se está observando un aumento en la prevalencia, que ya **se acerca al 1%**. Las mejoras en las técnicas diagnósticas, la mayor concienciación por parte de la población y los cambios en los criterios diagnósticos son algunos de los motivos que explican este aumento en la prevalencia. El número de casos de TEA varía enormemente según el país, pudiendo llegar al 2% en algunas zonas de Asia, Europa y América del Norte (12).

Sobre la **proporción de género** en los TEA, resulta destacable que el ratio de diagnóstico hombre-mujer es 4 a 1, estudios recientes apuntan incluso a una proporción menor. Existe un sesgo diagnóstico en niñas que se puede deber a un fenotipo autista sutil o a una capacidad para camuflar los síntomas (11).

1.1.6. Evolución y tratamiento

El **riesgo de mortalidad es 2.8 veces mayor** que el de aquellas personas no afectas, en gran medida esto se debe a las comorbilidades anteriormente explicadas. Entre el 58-78% de los adultos con TEA tienen resultados deficientes en términos de calidad de vida, educación y empleo. Una inteligencia infantil superior y mejores habilidades comunicativas predicen una mejor integración en el futuro.

El **tratamiento de los TEA** debe ser individualizado y adaptado a cada individuo. Es crucial un enfoque multidisciplinar y multidimensional. Los objetivos principales son conseguir la mayor calidad de vida posible, mejorar las habilidades comunicativas y fomentar la independencia y desarrollo. Las intervenciones más efectivas se constituyen de terapias conductuales y educativas. Los **fármacos tienen un papel secundario**. Los antipsicóticos (como risperidona y aripiprazol) han demostrado reducir los comportamientos rígidos repetitivos en niños. Muchos medicamentos continúan a estudio como la oxitocina o los inhibidores selectivos de la recaptación de serotonina. Existe evidencia de un beneficio en tratar el TDAH que suele acompañar los TEA con fármacos como el metilfenidato. Por último, existen ciertos suplementos seguros, pero sin beneficio demostrado como vitaminas, melatonina, dieta libre de gluten y omega-3 (3).

1.1.7. Etiología de los tea

Los **TEA son de los TND más heredables que existen hoy en día.** Los gemelos homocigotos tienen tasas de concordancia de hasta 3 veces superiores a aquellos heterocigotos. Asimismo, algunos estudios apuntan a una heredabilidad del 80%. Tan solo un 10% de los pacientes no tienen antecedentes familiares de TEA en la familia, estos casos se conocen como individuos

simplex. Por el contrario, las familias *multiplex* (las más comunes)son aquellas en las que hay 2 o más casos de TEA (13,14).

Es necesario mencionar que, aunque la genética desempeña un papel fundamental en la etiología del trastorno, las causas ambientales no deben ser obviadas. Incluso se ha propuesto que ciertos cambios epigenéticos pueden estar implicados (15).

Los **factores ambientales** respaldados por una mayor evidencia son la edad parental avanzada; consumo de medicación durante el embarazo (sobre todo ácido valproico, pero también talidomida y misoprostol); complicaciones durante el parto como traumatismos, isquemia o hipoxia; bajo peso al nacer y nacimiento prematuro. La fuerza de asociación es menor con la obesidad y diabetes maternal. Por otra parte, no se ha visto que la vacunación o el hábito tabáquico durante el embarazo aumenten el riesgo de padecer TEA (16).

Para comprender la **base genética** de los TEA es necesario hacer un pequeño repaso histórico sobre el genoma humano.

A lo largo de la historia, diversos estudios han cambiado nuestra **percepción acerca del ADN**. El **Proyecto Genoma Humano**, consiguió secuenciar por primera vez la mayor parte del genoma humano. Fue una investigación ambiciosa que contó con la colaboración de numerosos países y que se demoró 13 años (1990-2003). Aun así, quedaron muchos interrogantes abiertos que otras investigaciones, como el Proyecto *Encyclopedia of DNA Elements* (**ENCODE)**, intentaron responder (17). Inicialmente, ENCODE se centró en estudiar la parte codificante del genoma, que representa solamente alrededor del 1% del mismo. Se consideraba que la mayoría del ADN era *junk DNA* o **ADN basura** (18). En fases posteriores se investigó el 99% restante y concluyó, entre otras cosas, que el 80,4% del genoma humano participa, al menos, en un evento bioquímico en al menos un tipo celular. En la actualidad se considera el concepto de ADN basura obsoleto (17,19).

Existen diversas formas de clasificar las mutaciones en el ADN (20).

- Según **la célula dónde se produzcan** pueden ser germinales (afectan a los gametos) o somáticas (afectan a las células no sexuales).
- Según **si se heredan o no**, existen mutaciones *de novo* y heredadas. Las heredadas están presentes en alguno de los progenitores mientras que las mutaciones *de novo* son aquellas que ocurren por primera vez de forma espontánea en un individuo.
- **Atendiendo a la estructura**, se puede hablar de mutaciones cromosómicas, SNVs e indels.
    - Las mutaciones cromosómicas son cambios en el número (aneuploidías o euploidías) o estructura (deleciones, duplicaciones, inversiones) de los cromosomas. Incluimos aquí los copy number variation (**CNV**), que son alteraciones en el número de copias de una región del ADN que pueden ir desde unos cientos de pares de bases (pb) hasta miles de pb.
    - Los *Single Nucleotide Variants* (**SNV**) son mutaciones por sustitución en las que una base es reemplazada por otra (por ejemplo "G>A", una adenina es reemplazada por una guanina). Aquellos SNVs con una frecuencia poblacional superior al 1% se denominan SNPs (*Single-nucleotide polymorphism*).
    - Los indels son mutaciones por inserción o deleción de bases nitrogenadas.
- Según la **localización de la variante** en una región codificante para proteína o no, existen mutaciones codificantes y no codificantes. Las mutaciones codificantes se

encuentran en los exones y las no codificantes pueden estar en intrones, regiones reguladoras o segmentos intergénicos.

- Según la **frecuencia poblacional** de las mutaciones existen variantes comunes (con una frecuencia mayor o igual al 1% en la población de estudio) y variantes raras (con frecuencias inferiores al 1%).
- Existen también otros tipos de variantes como las mutaciones en **zonas de empalme o *splicing***. Son aquellas que ocurren en regiones implicadas en el proceso de eliminación de las zonas no codificantes (intrones) y unión o empalme de aquellas zonas sí codificantes (exones) durante la maduración.

Las mutaciones también se subdividen según el **efecto que tienen sobre la proteína**. En la siguiente tabla (tabla 2) se recogen los tipos de mutaciones, su efecto en la proteína y un ejemplo.

Tipo de mutación	Efecto sobre la proteína	Ejemplo
Mutación silenciosa, sinónima o *silent*	La mutación produce un triplete que codifica el mismo aminoácido. No se afecta la función de la proteína.	Si en codón AAA se sustituye la tercera base, AAG, el mismo aminoácido "lisina" es incorporado.
Mutación de cambio de sentido, o *missense*	La mutación produce un triplete que codifica un aminoácido distinto.	Si en codón AAA se sustituye la tercera base, AAC, un aminoácido distinto al original es incorporado, "asparagina".
Mutación sin sentido o *non-sense*	La mutación produce un triplete que provoca una parada o *STOP*. La proteína resultante está acortada.	Si en codón UAC se sustituye la tercera base, UAA, se para la traducción.
Mutación del marco de lectura o *frameshift*	Adición o deleción de uno o varios pares nucleotídicos (nunca múltiplos de 3) que desplazan la pauta de lectura.	Una inserción o eliminación de un par de bases en una secuencia.

Tabla 2: tipos de mutaciones según su efecto en la proteína. El término "mutación no sinónima" engloba aquellas que no son silenciosas o sinónimas. Sin embargo, algunos programas bioinformáticos usados en este trabajo utilizan "mutación no sinónima" para referirse a las variantes *missense*.

La **base genética de los TEA** es compleja. Las alteraciones mejor descritas son las variantes comunes, raras y mutaciones *de novo*.

Las **variantes comunes** confieren un menor riesgo individual de padecer la enfermedad, por ejemplo, un único SNP no es capaz de producir por sí mismo el trastorno.
Las **variantes raras**, con una frecuencia alélica menor en la población, suponen un riesgo superior.
Las **mutaciones *de novo* (DNMs)** son de las que predisponen a un mayor riesgo. Tan solo una mutación *de novo* en heterocigosis puede ser suficiente para provocar la enfermedad. Esto está respaldado por el hecho de que las "grandes" CNV ocurridas *de novo* son más frecuentes en las familias *simplex* que en las *multiplex*. Además, las CNV se asocian con fenotipos severos. De todas formas, no es incompatible que, en un mismo individuo, tanto las variantes comunes como las raras contribuyan en el desarrollo de los TEA (3,21).

La **gran heterogeneidad de *loci* en los TEA** fue evidente desde los primeros estudios. Algunos apuntan a que puede haber hasta 800 genes implicados, aunque no todos tienen la misma relevancia (22).

1.1.7.1. Estudios de GWAS y TEA

Los estudios de asociación del genoma completo (**GWAS**) supusieron un gran cambio en el campo de la genética permitiendo numerosos descubrimientos científicos. Su particular utilidad es la capacidad para buscar asociaciones entre SNPs y enfermedades frecuentes como la diabetes, enfermedad coronaria o patologías psiquiátricas. Se fundamentan en la premisa de que la acumulación de diversas variantes comunes (SNPs) tiene el potencial de producir la enfermedad a estudio (23).

Los primeros GWAS publicados sobre los TEA no fueron muy fructíferos ya que una de las limitaciones que tienen estos estudios es la necesidad de cohortes muy grandes. Posteriormente este problema fue solucionado y este tipo de estudios resultaron ser exitosos.

- ***Estudio de Grove et al. (2019)***

Unos de los GWAS más importantes es un metaanálisis realizado en 2019 por el *Psychiatric Genetics Consortium* (**PGC**) que incluye más de 18000 casos y 27000 controles y reveló 5 *loci* de riesgo asociados con los TEA de forma exclusiva (tabla 3). Estas posiciones genéticas son una muestra de la ya conocida heterogeneidad de *loci* de los TEA, pues se encontraban en los cromosomas 1, 7, 8 y 20.

SNP	Cromosoma	Posición	Valor p	Alelos	Genes próximos
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*
rs201910565	1	96561801	2.48 × 10−8	A /AT	*LOC102723661, PTBP2*
rs71190156	20	14836243	2.75 × 10−8	GTTTTTTT/G	*MACROD2*
rs111931861	7	104744219	3.53 × 10−8	A/G	*KMT2E, SRPK2*

Tabla 3: *Loci* de riesgo asociados con los TEA identificados por Grove et al. (2019).

Este estudio también encontró **otros 7 *loci* compartidos** con otras patologías psiquiátricas con arquitectura genética similar (esquizofrenia, depresión mayor y bajo nivel educativo) (24).

1.1.7.2. Estudios de búsqueda de mutaciones *de novo*

Pese a que las **mutaciones *de novo* (DNMs)** tan solo explican menos del 10% de los casos de TEA existen muchos estudios que se han centrado en investigar los genes implicados (25). Las DNMs son aquellas que están presentes únicamente en el paciente y no en los progenitores (padre y madre). Este grupo de individuos se denomina **trío genético**. Mediante el estudio de varios tríos es posible encontrar nuevos genes implicados en este trastorno (26).

Las DNMs pueden ser tanto CNVs (*Copy number variations*) como SNVs (Single Nucleotide Variants). Las investigaciones iniciales buscaban CNVs mediante el uso de microarrays de alta resolución. Se detectaron 277 CNV presentes en el 44% de las familias con TEA y ausentes en el grupo control, lo que indica que las anomalías cromosómicas desempeñan un papel fundamental en la etiología de los TEA. En este momento incluso se sugiere que las variantes estructurales están presentes en una frecuencia lo suficientemente elevada en individuos con

TEA como para recomendar análisis citogenéticos y de microarrays de rutina en la práctica clínica diaria (27).

El desarrollo de la ***next-generation sequencing*** **(NGS)** y la reducción de los precios de secuenciación del genoma y exoma han permitido que se realicen cada vez más estudios de secuenciación del exoma completo o **WES (*Whole-exome Sequencing*)**. Los resultados obtenidos son cada vez más rápidos, baratos y precisos (28). De todas maneras, los WES siguen siendo menos costo-efectivos que los métodos tradicionales y cuentan con diversas limitaciones. Por ejemplo, esta tecnología no es capaz de detectar CNVs o mutaciones en zonas reguladoras o regiones intergénicas.

Posteriormente se estudian **SNVs**, que resultan especialmente interesantes porque el riesgo que conllevan a nivel individual es mayor que el de los SNPs. Como resultado, es más probable que los genes en los que se encuentran los SNVs tengan un papel más relevante en los TEA (29).

- ***Estudio de Satterstrom et al. (2020)***

Se trata del estudio **WES** más grande realizado hasta la fecha. El estudio cuenta con 35584 muestras, de las cuales 11986 presentan un diagnóstico de TEA.

Se introduce un nuevo marco analítico bayesiano, denominado TADA, que incorpora puntuaciones a nivel del gen dependiendo del número, tipo y localización de las variantes en cada gen. Esto permitió descubrir 102 genes asociados a los TEA. Se observa que las variantes *de novo* presentes en estos genes se hallan con mayor frecuencia en cohortes de trastornos de neurodesarrollo y de TEA. Se aprecian también diferencias fenotípicas, pues 49 de estos genes presentan frecuencias más elevadas de DNMs disruptivas en individuos con retraso del neurodesarrollo severo en comparación con otros 53 genes, que presentan frecuencias mayores en individuos con TEA y sin DI. Algunos de estos genes son: *SCN2A, FOXP, CHD8, SHANK3, SLC6A1 y ADNP*.

Es relevante mencionar que también se estudia la función celular de estos genes relacionados con los TEA. Se aprecia que su expresión está aumentada en neuronas excitadoras e inhibidoras tanto maduras como en maduración desde mediados del desarrollo fetal en adelante. Por último, se confirma su papel en la comunicación neuronal y regulación de la expresión génica (30).

1.1.8. TEA sindrómico

Los TEA se relacionan estrechamente con varios trastornos genéticos, dando lugar a lo que anteriormente se conocía como autismo sindrómico. Por ejemplo, entre un 20-50% de los individuos con X frágil son diagnosticados de TEA. El porcentaje es un poco superior (24-60%) en la esclerosis tuberosa. Otros trastornos asociados son el síndrome de Down (5-39%), fenilcetonuria (5-20%), síndrome de Angelman (50-81%) y síndrome de Charge (15-50%). Muchos de estos trastornos llevan asociada una DI, por lo que el diagnóstico de TEA es en muchas ocasiones obviado (3).

1.2. *PKD1*

1.2.1. Estructura de *PKD1*

PKD1 (polycystic kidney disease-1) es un gen largo, formado por 14,148 pares de bases (pb) distribuidas a lo largo de 46 exones. Se localiza en el cromosoma 16p13.3, es decir, en el brazo

corto (p) del cromosoma 16, en la región 13.3. Codifica la información para la proteína **policistina-1 (PC1)**, una larga proteína transmembrana formada por 40303 aminoácidos (aa) y múltiples dominios. Las mutaciones en este gen están implicadas en el 85% de los casos de enfermedad renal poliquística autosómica dominante o **ADPKD** (Autosomal Dominant Polycystic Kidney Disease) (31,32).

El gen *PKD1* posee una serie de características que lo convierten en un gen particular y que dificultan su estudio.

1- Su gran tamaño: los aproximadamente 14.000 pb hacen que el empleo de herramientas bioinformáticas sea más complejo.
2- Elevada proporción de bases GC (Guanina-Citosina) y un gran número de dinucleótidos CpG (Citosina-Guanina).
3- Contiene un tramo de polipirimidina de 2,5 kilobases en el intrón 21: este intrón es a su vez el más largo de todo el genoma humano.
4- El 70% de *PKD1* se encuentra duplicado a lo largo del cromosoma 16. Existen 16 **repeticiones *PKD*** (o *PKD-repeats*) en total.
5- Como consecuencia de las características anteriores, a lo largo del cromosoma 16 hay varios *PKD1-like loci* o pseudogenes homólogos: estas copias transcripcionalmente activas son difíciles de distinguir del tránscrito del *loci* de *PKD1* (31, 33).

1.2.2. Estructura de la proteína PC1

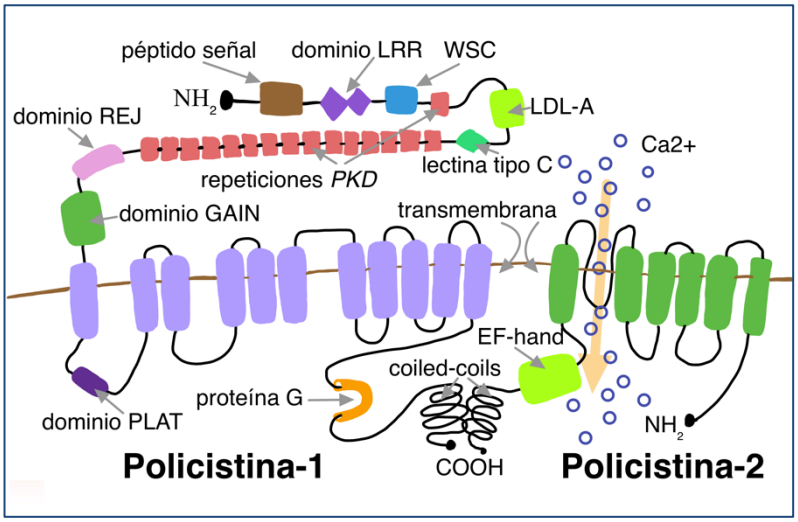

Figura 1: representación del complejo PC1/PC2. Imagen de elaboración propia basada en el trabajo de Cordido et al. (2017) (33).

PC-1 es una proteína de gran tamaño que funciona como un receptor acoplado con proteína G atípico. En cuanto a su estructura, está formada por unos 40300 aa; atraviesa la bicapa lipídica 11 veces; tiene un extremo N-terminal extracelular y un extremo intracelular C-terminal. El extremo N-terminal tiene gran tamaño (más de 3000 aa) y es rico en sitios de unión. Por este motivo, se considera que actúa como un sensor mecánico y receptor de ligandos. Aunque la forma en la que los estímulos actúan sobre PC-1 y su significado funcional no están claros. Forma el **complejo PC-1/PC-2** con la policistina-2 PC-2 (codificada por el gen *PKD2*). Esta proteína es miembro de la familia de proteínas receptoras transitorias o TRP (*transient receptor protein*) (34).

En cuanto a **su función**, este complejo proteico es un mecanorreceptor y sensor químico, también tiene un papel clave en la señalización de calcio y proteína G. Además, recientemente se ha descrito su implicación en la adhesión celular (35,36).

PC-1 y PC-2 son capaces de señalar, bien a través de mecanismos independientes entre sí, o bien a través de mecanismos interdependientes. Diversos experimentos, que incluían el cultivo de líneas celulares renales, demostraron la existencia del complejo PC-1/PC-2. Se han descrito zonas de ambas proteínas que pueden interactuar entre sí físicamente a través de sus dominios en espiral. Esta teoría del complejo proteico e interdependencia es respaldada por el hecho de que las mutaciones tanto en *PKD-1* como en *PKD-2* son capaces de producir la misma enfermedad. A su vez, estas proteínas pueden actuar de forma independiente. Se ha demostrado experimentalmente que los canales de PC-2 no requieren a PC-1 para desempeñar su función y viceversa (34).

1.2.3. La enfermedad renal poliquística autosómica dominante

La **enfermedad renal poliquística autosómica dominante (ADPKD)** es una patología caracterizada, como su nombre indica, por la formación de múltiples quistes epiteliales en los túbulos renales. Como consecuencia se produce un deterioro de la función renal (36).

La ADPKD es causada por mutaciones tanto en el gen *PKD1* como en *PKD2,* que codifican las ya mencionadas proteínas PC-1 y PC-2 respectivamente. Aproximadamente el **85%** de los **casos** son explicados por **mutaciones en *PKD1***, justificando los casos restantes las alteraciones en *PKD2*. También existen mutaciones en otras proteínas asociadas a los cilios que son reconocidas como factores patogénicos importantes. Se ha visto, por ejemplo, que las mutaciones en el gen *GANAB* están presentes en el 0,3% de los pacientes. Es importante señalar que, por lo general, las mutaciones en *PKD2* y *GANAB* producen una enfermedad más leve con una clínica más larvada (33,37).

Como se mencionó anteriormente, la principal función de PC-1 y PC-2 es la señalización, pero también se ha descrito la participación de estas proteínas en la función celular (contribuye en la migración celular, la formación del citoesqueleto actínico y en la polaridad celular planar). En cuanto a su implicación en la señalización, las vías más destacables son: cMET (receptor del factor de crecimiento de los hepatocitos), STAT3 (factor de transcripción y activador de la señal 3), mTOR (diana de rapamicina en mamíferos), la vía de señalización Wnt, PI3K/Akt (fosfoinosítida 3cinasa), CFTR (regulador de la conductancia transmembrana de fibrosis quística) y EGFR (receptor del factor de crecimiento epidérmico).

La **hipótesis más defendida** sobre la patogenia de la enfermedad justifica que, una vez PC-1 y PC-2 pierden la función ciliar, ocurren una serie de eventos que acaban produciendo la formación y crecimiento de los quistes. La primera consecuencia de un funcionamiento erróneo de estas proteínas es la reducción de la señalización de calcio y, como resultado, un aumento de la actividad de la adenilil ciclasa. Esto produce una disminución de la actividad de la fosfodiesterasa y un incremento del AMP cíclico celular. A su vez, este incremento promueve la actividad de la proteína cinasa A. La proteína cinasa A se encuentra en las células que recubren los quistes renales y actúa a través de los canales de cloro y acuaporina promoviendo la formación y aumento de tamaño de los mismos a través de la secreción de líquido (36).

1.2.3.1. Clínica

Se trata de una patología que puede **cursar de forma asintomática** hasta la cuarta o quinta década de la vida, momento en el cual los pacientes ya tienen entre cientos y miles de quistes renales. En este momento los riñones pueden llegar a pesar 20 veces más de lo habitual y cuadruplicar su tamaño habitual. La manifestación clínica más habitual es el **dolor lumbar**, presente en más de la mitad de los pacientes. También son motivos de consulta comunes la hematuria (por la ruptura de un quiste) y las infecciones urinarias (en forma de pielonefritis o infección del quiste), estas últimas son de hecho la segunda causa de muerte. La comorbilidad más destacable es la hipertensión arterial, esta predispone a complicaciones cardiovasculares, que constituyen la primera causa de mortalidad. Muchas veces la ADPKD se diagnostica debido a la hipertensión arterial, que precede a la disminución de la tasa de filtrado glomerular. Otros trastornos asociados son la presencia de quistes hepáticos y cálculos renales. También es necesario señalar que los aneurismas intracraneales son hasta 5 veces más comunes que en la población general y son una fuente importante de mortalidad (36, 38).

1.2.3.2. Diagnóstico:

El diagnóstico se basa en la presencia de quistes renales y de antecedentes familiares. La ecografía es la principal herramienta de detección, aunque se pueden usar otras pruebas. El número de quistes necesarios para el diagnóstico varía según la edad, incrementándose a medida que aumenta la edad (36).

1.2.4. Papel de PKD1 en otras enfermedades

En la base de datos **MalaCards** (https://www.malacards.org) (que es complementaria a GeneCards) se puede encontrar información muy variada acerca de *PKD1* como su papel en distintas enfermedades. Se ha propuesto que este gen podría estar relacionado con muchas patologías además de la ADPKD, aunque en muchos casos la evidencia es escasa (39).

Recientemente se ha publicado un estudio que relaciona mutaciones recesivas en *PKD1* con las **convulsiones febriles y epilepsia**. Tras la secuenciación de exoma completo de 314 tríos, se encontraron **8 mutaciones *missense*** en heterocigosis en *PKD1*. Presentaron una frecuencia poblacional superior a la esperada en población con convulsiones. Además, propone un mecanismo fisiopatológico argumentando que este gen es clave a la hora del desarrollo neuronal. Se expresa 1,79 veces más en el cerebro que en el riñón y su papel en el desarrollo neuronal es clave. Este hecho puede ser demostrado a través de ratones *knockout* (es decir, con *PKD1* inactivado), se evidenció que la falta de este gen causa defectos en el tubo neural letales. El estudio propone este gen como **posible causante de epilepsia** con una significación estadística elevada y una buena correlación genotípica-fenotípica. A mayores, aporta una sólida explicación neuroanatómica a través de estudio de expresión y registros electroencefalográficos (40).

2. JUSTIFICACIÓN Y OBJETIVOS

Los TEA son una serie de TND con una **prevalencia realmente elevada** en población general (1-2%) Si además se tienen en cuenta sus comorbilidades y cómo afectan en gran medida a la calidad de vida de los pacientes, se hace evidente la fuerte implicación sociosanitaria que tienen estas patologías.

Los criterios diagnósticos son inespecíficos y están sujetos a numerosos cambios y modificaciones. Por este motivo, cada vez cobran más importancia los **test genéticos**, pues cabe recordar que en los TEA la genética desempeña un papel fundamental. Es necesario conocer todas las mutaciones implicadas en la enfermedad por su potencial en el desarrollo de nuevas herramientas diagnósticas y terapéuticas.

Hasta la fecha, ***PKD1* no ha sido descrito como un posible gen de riesgo** para los TEA.

En este contexto, el **objetivo principal** de este trabajo es la descripción y estudio exhaustivo de 10 mutaciones *de novo* en *PKD1* encontradas en dos cohortes de TEA, para así poder determinar si *PKD1* está implicado en los TEA.

Objetivos específicos:

1- Caracterizar fenotípicamente las variantes halladas en *PKD1* en individuos diagnosticados de TEA.
2- Describir detalladamente las mutaciones *de novo* en *PKD1* a nivel molecular y funcional *in silico.*
    a. Realizar la anotación funcional de cada variante con herramientas bioinformáticas lo cual permitirá una exploración minuciosa de sus posibles efectos a nivel biológico.
    b. Comparar las mutaciones encontradas con aquellas causantes de ADPKD. Con el fin de encontrar similitudes que pueden reforzar la teoría de la implicación de *PKD1* en los TEA.
    c. Estimar la frecuencia de las variantes en la población general.
3- Realizar un estudio de expresión de los exones de *PKD1* y de sus isoformas.
4- Hallar en qué dominios se encuentran las variantes y estudiar su función.
5- Evaluar el impacto de cada mutación en la función y estructura de la proteína mediante el uso de herramientas bioinformáticas basadas en Inteligencia Artificial.

3. METODOLOGÍA

3.1. Bases de datos

Los avances en el campo de la genética y genómica han permitido la obtención de una vasta cantidad de datos relacionados con diversos aspectos del ADN como pueden ser secuencias, mutaciones y expresión de genes, entre otros. Como consecuencia, la necesidad de contar con herramientas bioinformáticas y bases de datos especializadas se ha hecho evidente. Hoy en día existen un sinfín de recursos de biología computacional que permiten el estudio de los genes (41).

Para la realización de este trabajo fue necesario recurrir a muchas de estas bases de datos, que a continuación se explicarán brevemente.

- **ExAC (Exome Aggregation Consortium):** esta base datos alberga la secuenciación de alta calidad del exoma de 60706 individuos. Contiene alrededor de una variante por cada 8 bases de exoma recogiendo más de 10 millones de mutaciones (42).
- **dbSNP (Database for Nonsynonymous SNPs' Functional Predictions):** contiene las anotaciones funcionales de más de 87 millones de SNPs no sinónimos caracterizados según los efectos basados en variantes y genes.
- **ClinVar:** proporciona un informe entre asociaciones entre variantes y enfermedades clínicamente relevantes. Alberga aproximadamente 110000 variantes.
- **GTEx (Genotype-Tissue Expression Project):** contiene información acerca de la expresión de genes, exones, isoformas y otros elementos genómicos en los diferentes tejidos. Esto es posible a través de la relación y comparación de muestras procedentes de autopsias con los genotipos de sus respectivos individuos (41).

Las siguientes bases de datos funcionan como colecciones de varias fuentes de información.

- **UCSC Genome Browser:** este recurso, servido por la Universidad de California, Santa Cruz (UCSC), funciona como buscador de genes, visor gráfico y prestador de diversos servicios bioinformáticos (43).
- **Ensembl:** ofrece diversas herramientas, destaca su VEP (Variant Effect Predictor), capaz de estimar el efecto de una mutación, aunque con ciertas limitaciones.
- **NCBI (National Center for Biotechnology Information):** recopila enlaces a publicaciones, bases de datos y recursos varios.
- **GeneCards:** Resumen de la nomenclatura de un gen, su función, expresión de ARNm, entre otros.
- **Uniprot:** contiene la información básica acerca de una proteína, se puede saber, por ejemplo, su secuencia de aminoácidos, estructura y dominios. También posee enlaces de interés a publicaciones sobre la proteína a estudio (41).
- **GnomAD:** permite acceder de forma intuitiva y rápida a la información disponible sobre variantes genéticas. Es la colección de variación poblacional más grande y ampliamente utilizada. Entre los datos disponibles se encuentra la prevalencia de las mutaciones conocidas, pudiendo filtrar por grupos étnicos o médicos (población sin cáncer, sin enfermedades neurológicas…) (44).

3.2. Estudio de exoma completo en una cohorte de tea

Parte de las mutaciones (4 de 10) que se usaron para realizar este trabajo proceden de un estudio previo realizado por la **Universidade de Santiago de Compostela** (USC) en el año 2020 (45, 46). Las otras 6 mutaciones proceden del *Autism Sequencing Consortium* **(ASC)**, formado por cohortes de varias partes del mundo, destaca la cohorte del ya explicado estudio de referencia, realizado por **Satterstrom et al**.

En el trabajo de la USC se secuenció de forma completa el exoma de 360 individuos diagnosticados previamente de TEA en busca de variantes de *novo.* Para ello, también se analizó el exoma de los progenitores, comprobando que las variantes encontradas en los individuos afectos no estuviesen presentes ni en el padre ni en la madre. Se trabajó por lo tanto con **tríos genéticos**, secuenciando en total 1080 exomas. El objetivo de dicho estudio era analizar la contribución de la variación rara al riesgo de los TEA.

Se realizó de la siguiente manera: primero se obtuvo la cohorte de pacientes a partir de casos con diagnóstico clínico de TEA (según criterios DSM-IV y DSM-V). Los diagnósticos fueron realizados por neurólogos pediátricos o psiquiatras.

La cohorte completa o **cohorte española (N= 360)** estaba formada a su vez por dos de menos tamaño.

- Una constituida por pacientes del Complejo Hospitalario de Santiago de Compostela o de entidades gallegas colaboradoras con individuos con TEA (N=136).
- Otra formada por participantes del servicio de Psiquiatría del Niño y del Adolescente del Hospital Gregorio Marañón, en Madrid (N=224).

Se **excluyeron** aquellos individuos menores de 3 años o con un trastorno genético que pudiese explicar el TEA. Además, se obtuvo información clínica adicional de los pacientes de la cohorte santiaguesa.

El **ASC** fue el encargado de realizar la secuenciación completa de todos los exomas. Generó un archivo VCF o Variant Call Format (Formato de Llamada de Variantes, en español) con todos los datos de las variaciones en la secuencia exónica. Posteriormente, con las herramientas BCFtools y SnpEff se obtienen los archivos individuales de cada sujeto y se anotan las variantes encontradas. Se realizó una **búsqueda de DNMs**, es decir, aquellas en las que el genotipo del probando afecto tenía un alelo o ambos (1-0 o 1-1) mientras los dos progenitores carecían del alelo (0-0). Para la detección de las DNMs se usaron herramientas de control de calidad y filtrado estrictas que permitieron eliminar errores de secuenciación en este paso previo y no considerarlos como variantes *de novo*. Se excluyeron, por ejemplo, variantes con una calidad de genotipo asignado inferior a 20, con uno o varios alelos recogidos en la base de datos Exome Aggregation Consortium (ExAC) y aquellas con una separación inferior a 20 pb. Las DNMs encontradas se clasificaron en germinales o postcigóticas, resecuenciando estas últimas a través de diferentes técnicas más precisas como la secuenciación Sanger.

Gracias a este estudio fue posible encontrar 4 sujetos con DNMs en el gen *PKD1* y sin otras mutaciones capaces de explicar el diagnóstico de TEA (45,46). Para la búsqueda de las otras 6 mutaciones del **ASC** se usó su portal web (https://asc.broadinstitute.org/).

3.3. Caracterización fenotípica de los probandos

A continuación, se recoge una breve caracterización clínica y fenotípica de 4 de los probandos incluidos en este trabajo. Únicamente se disponen de los datos de los individuos de la cohorte española (N=360).

Probando 1

- Varón, con 2 años y 7 meses de edad al diagnóstico de TEA e inicio de los síntomas a los 28 meses.

- Antecedentes personales:
  - Como antecedentes neuropsiquiátricos destacan una discapacidad intelectual y retraso en la adquisición del lenguaje. También coexisten un retraso en el control de los esfínteres, estereotipias y regresión evolutiva.
  - El nivel adquirido del lenguaje se describe como fluido (paciente con 17 años en el momento de la exploración).
  - El paciente no presenta antecedentes médicos de interés. Tanto la gestación como el parto fueron normales. Las edades materna y paterna eran de 32 y 28 años respectivamente.
- Antecedentes familiares:
  - Sin antecedentes familiares de interés.
  - Tiene un hermano no afecto de TEA
- Pruebas complementarias:
  - En cuanto a pruebas diagnósticas y cuestionarios tiene una puntuación de 33 puntos en el Test Car (Comportamientos Autistas Revisados) y 43 puntos el Test IDEA (Inventario de Desarrollo de Habilidades de Evaluación). Además, tiene un test de Vineland-II que refleja un nivel adaptativo general bajo (66).
- Tratamiento:
  - En el momento de la inclusión del paciente en el estudio se encontraba a tratamiento con metilfenidato.
- Genotipo:
  - Cariotipo, X frágil y array de SNPs-CNV negativos.
  - Tras la secuenciación de exoma completo se encontró una sustitución de una guanina por una adenina (G:A) en la posición 16:2161078, en el gen *PKD1*.

Probando 2

- Varón, cuya edad al diagnóstico de TEA no se encuentra especificada, pero se estima alrededor de 18 meses el momento de inicio de los síntomas.
- Antecedentes personales:
  - Entre los antecedentes neuropsiquiátricos, se encuentra un retraso en la adquisición del lenguaje y retraso psicomotor.
  - El nivel adquirido del lenguaje se describe como “frases simples”.
  - Como único antecedente médico de interés figura una macrocefalia. La gestación y el parto fueron normales. Las edades materna y paterna eran de 37 y 36 años respectivamente.
- Antecedentes familiares:
  - Tío diagnosticado de TEA.
  - Tiene una hermana no afecta de TEA.
- Pruebas complementarias:

  - Puntuación de 62 puntos en el test de Vineland-II (refleja un nivel adaptativo general bajo). Los test de ADI-R (Autism Diagnostic Interview-Revised), ADOS (Autism Diagnostic Observation Schedule) y SCQ (Social Communication Questionnaire) muestran valores compatibles con TEA.
- Tratamiento:
  - En el momento de la inclusión del paciente en el estudio no se encontraba a tratamiento farmacológico.
  - Genotipo:
    - X frágil y array de SNPs-CNV negativos.
    - Tras la secuenciación de exoma se identificó una VUS (variante de significado incierto) en el gen *NF1*, NM_00267.3(NF1):c.4750A>G, p.(Ile1584Val) en heterocigosis, heredada de la madre. La mutación en *PKD1* tenía la posición 16:2163160 y se trata de una sustitución de una adenina por una citosina (A:C).

Probando 3

- Varón, con 3 años y 2 meses de edad en el momento del diagnóstico de TEA del que se desconoce la fecha del inicio de la sintomatología.
- Antecedentes personales:
  - Presenta varios antecedentes neuropsiquiátricos: trastorno por déficit de atención e hiperactividad (TDAH), retraso en el desarrollo del lenguaje, comportamiento disruptivo. Además, manifestó retraso en el control de esfínteres, estereotipias, regresión evolutiva y problemas del sueño.
  - El nivel del lenguaje adquirido en el momento de la exploración se describe como "frases simples" (paciente con 14 años en ese instante).
  - La edad de la madre al nacimiento era 37 años y la del padre 41.
  - Sin antecedentes médicos de interés.
- Antecedentes familiares:
  - Prima con diagnóstico de síndrome de Asperger.
  - Primo diagnosticado de TEA.
- Pruebas complementarias:
  - En el test CARS obtuvo una puntuación de 35,5 puntos. El Vineland-II mostró un nivel adaptativo general bajo. Las pruebas ADI-R y M-CHAT dieron resultados compatibles con el TEA.
- Tratamiento:
  - El paciente está a tratamiento con metilfenidato y antipsicóticos no especificados.
- Genotipo:
  - La mutación en *PKD1* se encuentra en la posición 16:2153404:C:T.

Probando 4

- Varón, con 3 años al diagnóstico de TEA y 2 años y 6 meses al inicio de los síntomas.
- Antecedentes personales:
  - Los antecedentes neuropsiquiátricos más destacables son un retraso en la adquisición del lenguaje y control de esfínteres, estereotipias y regresión evolutiva.
  - El nivel del lenguaje adquirido en el momento de la exploración se describe como "frases simples" (paciente con 25 años en ese instante).
  - Las edades de los progenitores en el momento del nacimiento eran 39 para la madre y para el 47 padre.
  - En cuanto a la gestación, cabe destacar que fue concebido mediante fecundación in vitro y el embarazo fue gemelar. Se deduce gemelos dicigóticos debido a que la hermana es una mujer.

  - Como antecedentes somáticos, el paciente presenta alergia al polen, estreñimiento,
- Antecedentes familiares:
  - Tiene una hermana gemela no afecta de TEA.
  - Tiene un tío con esquizofrenia, demencia, trastornos amnésicos y otros trastornos cognoscitivos.
  - Padre con antecedentes de epilepsia en la infancia
- Pruebas complementarias:
  - Realizó los tests CARS e IDEA con puntuaciones de 34 y 51 puntos respectivamente. También realizó los test Vineland-II, SCQ, ADOS y Peabody. Todos ellos con resultados compatibles con TEA.
- Tratamiento:
  - No consta ningún tratamiento farmacológico.
- Genotipo:
  - Cariotipo y X frágil negativos.
  - Array de SNPs con una microduplicación en 7p21.1, de alrededor de 500 Kb, que afecta a los genes *AGR3* y *AHR*. Sin embargo, hay escasa información acerca de la misma y de su relación con los TEA ([hg19] 7p21.1(16.918.219-17.422.044)x3).
  - Se encontró una mutación en *PKD1* en posición 16:2159430:G:A (HG19).

3.4. Localización de las variantes

Existen diferentes **versiones del genoma humano** utilizadas como referencia para la secuenciación y el análisis genómico. Estas versiones representan al ser humano promedio y se actualizan periódicamente. A día de hoy, las más utilizadas son **hg19** (o GRCh37) y **hg38** (GRCh38), siendo esta última la más reciente (2019). Aun así, pese a ser hg19 del año 2009 continúa siendo ampliamente utilizada (47).

Una vez detectadas las 10 variantes, cuyas coordenadas iniciales usaban como genoma de referencia la versión hg19, se convirtieron a hg38. Este proceso de traducir las coordenadas de una versión a otra se conoce como ***liftOver***. Para llevarla a cabo se utilizó la herramienta en línea "LiftOver Tool" del ya mencionado recurso **UCSC Genome Browser (https://genome.ucsc.edu)**. El fin de esta conversión de coordenadasfue facilitar el trabajo posterior, ya que algunos recursos están disponibles de forma exclusiva en una u otra versión.

En esta misma web, que cuenta con un **buscador de coordenadas y visor gráfico**, se revisó manualmente en qué región del gen se encontraba cada mutación. Pudiendo especificar el exón concreto. Finalmente, con las bases de datos GTEx (https://gtexportal.org/home/)y GnomAD(https://gnomad.broadinstitute.org)se buscó la posición de cada exón y si había algún SNP de referencia. De esta forma se creó la tabla 4.

Para posteriormente poder comparar resultados, se usó la **ADPKD Variant Database (https://pkdb.mayo.edu)**. Se trata de la base de datos de variantes de la poliquistosis renal y está disponible en la web de la Fundación PKD (48).

Se filtró buscando las variantes halladas en *PKD1* y *any Pathogenic*, lo cual incluye únicamente las variantes patogénicas y probablemente patogénicas. Quedando excluidas las variantes *benign* (benignas), *likely Benign* (posiblemente benignas) y *Variant of Uncertain Significance (*variante de significado incierto).

Se usó la función "*fx*=CONTAR.SI(B2:B1226;"EX_")" para buscar cuántas variantes han sido descritas en cada uno de los exones del gen. Especificando el rango de celdas en el que se quiere buscar un criterio (B2:B1226) y el exón (EX"N.º de exón").

De esta forma se creó una lista (tabla 5) con los 46 exones del gen y el número de mutaciones patogénicas descritas en cada uno de ellos. En dicha tabla se subrayan los exones en dónde se encuentran las variantes estudiadas en este trabajo.

3.5. Estudio de expresión

Para estudiar cómo era la expresión de los exones en los tejidos se elaboró un ***heatmap* o mapa de calor** (una representación visual que usa colores para mostrar una concentración o valor numérico)**.** Se hizo lo mismo con las distintas isoformas del gen *PKD1*. Para ello, primero se creó una tabla introduciendo manualmente los datos almacenados en la base de datos **GTEx** (Genotype-Tissue Expression) (https://gtexportal.org/home/), dónde se encuentra la expresión de los diferentes elementos genómicos en diversos tejidos del organismo. Las unidades de la expresión de exones son la mediana del número de lecturas por base. En el caso de las isoformas son transcritos por millón (TPM).

Como los TEA son TND, los tejidos seleccionados para elaborar el *heatmap* fueron todos los pertenecientes al **sistema nervioso central** y **2 tejidos renales** (corteza y médula renal), debido a la implicación de *PKD1* en la ADPKD. Asimismo, se incluyó también un **tejido del tracto digestivo** como control (colon), pues se vio que la expresión del gen en este adquiere valores intermedios con respecto al resto de tejidos del organismo.

Para crear el *heatmap* se usó **RStudio** (Versión 2023.12.0), un entorno de desarrollo integrado (IDE) para el lenguaje de programación R. La función utilizada fue "**heatmap.2"**, forma parte del paquete "**gplots"**, que proporciona herramientas para la visualización gráfica de datos. Esta función permite crear mapas de calor a partir de una matriz de datos. De esta manera, se introdujeron las tablas creadas a partir de la información de GTEx en RStudio y se crearon los *heatmap* de la representación de exones (gráfico 1) e isoformas (gráfico 2).

3.6. Frecuencia de las variantes

El análisis de la variabilidad genética en las diversas poblaciones es un punto fundamental a la hora de estudiar mutaciones y sus posibles efectos en la salud humana.

La base de datos GnomAD proporciona una información muy valiosa acerca de las **frecuencias poblacionales de las mutaciones**. Además, ofrece una amplia gama de filtros que permiten comparar prevalencias entre diferentes grupos de individuos. Se puede establecer, por ejemplo, que se excluyan individuos con cáncer o enfermedades neurológicas. Además, la información se muestra separada por etnias y continentes, de manera que la validez de los datos es mayor.

Para realizar la tabla (tabla 6), se buscaron las frecuencias en **población europea** y en **población *non-neuro***. GnomAD especifica que el grupo *non-neuro* está constituido por individuos que no participaron en estudios de neurología o psiquiatría o que, si lo hicieron, fue como grupo control. Por ende, en este conjunto excluye individuos con TEA u otro TND, lo que lo convierte en un grupo representativo de la población general.

GnomAD divide la **población europea** en finlandesa y no finlandesa, se usó el grupo de población europea no finlandesa como representante del otro grupo por ser más adecuado. En este conjunto sí se incluyen individuos con TEA o trastornos neuro-psiquiátricos.

3.7. Descripción de anotación en ANNOVAR

Una vez halladas las mutaciones, es necesario anotarlas a través de un *software* específico, que en este caso fue **ANNOVAR** (del inglés, ANNOtate VARiation) (https://annovar.openbioinformatics.org/en/latest/). Se trata de una aplicación bioinformática que se ejecuta localmente en un ordenador a través de la línea de comandos. No solo permite la anotación de variantes, también filtra y prioriza según los criterios establecidos, examina la consecuencia funcional en los genes, infiere bandas citogenéticas y genera informes detallados que permiten una fácil interpretación de los resultados. Otra función interesante es la de comparar las frecuencias de cada variante en distintas bases de datos (49).

En primer lugar, se obtuvo un **archivo VCF** (del inglés, Variant Call Format) en hg38 y otro en hg19, pues algunas bases de datos que usa ANNOVAR solo están disponibles en una de las versiones. Este fichero fue generado usando una **herramienta de Ensembl** que tiene este propósito. El VCF es un formato usado comúnmente en genética para almacenar datos de secuenciación. Cada entrada proporciona información de cada variante genética, distribuida en forma de columnas, que incluye el número del cromosoma, posición, ID, alelo de referencia, alelo alternativo y otros elementos. A partir de este VCF se pudo crear un archivo ***.input*** con el que el programa puede trabajar.

ANNOVAR ofrece diferentes formas de anotar las variantes, a continuación, se detalla el uso de cada tipo de anotación junto con el significado de cada elemento.

3.7.1. GENE-BASED ANNOTATION

Permite identificar si los SNVs o CNVs causan cambios en la codificación de las proteínas y, en caso afirmativo, qué aminoácidos se afectan. El sistema ofrece diferentes bases de datos y sistemas de definición de genes como *UCSC genes, AceView genes, GENCODE genes, ENSEMBL genes y RefSeq gene*.

Cuando se ejecuta el código, el programa devuelve por defecto 2 archivos, uno con todas las variantes anotadas y otro solo con las que se encuentran en exones (ya que suelen ser las de mayor interés para el investigador). La información que aporta de cada variante es la siguiente:

- **Func:** esta es la primera columna, dice si la variante se encuentra en un exón, región intergénica, intrón, zona no codificante de ARN (ncRNA), zona de splicing…
- **Gene:** es la segunda columna, detecta si la variante está dentro de un gen y, en caso afirmativo especifica de cuál se trata.
- **GeneDetail:** aporta información a mayores sobre el gen y la mutación, pero no lo hace si la variante se encuentra en un exón o intrón.
- **ExonicFunc:** solo cubre esta columna en el caso de que la variante se encuentre en un exón. Describe si la mutación supone un cambio aminoacídico en la proteína, es decir si es **sinónima** (synonymous SNV) o **no sinónima** (nonsynonymous SNV).
- **AAChange:** únicamente se cubre esta casilla en el caso de que la mutación sea exónica. El programa detalla, en este orden: el gen implicado; el número del transcrito de ARNm; el número del exón; el cambio de secuencia en el ADN con su posición y el cambio en la secuencia de aminoácidos en la proteína con su posición en la cadena peptídica.

Para este trabajo se anotó usando diferentes bases de datos. En primer lugar, se usó **refGene**, el conjunto de datos de genes de la secuencia de referencia del NCBI (RefSeq). También se usó **ensGene** (de Ensembl) y **knownGene** (de UCSC).

A continuación, se explican resumidamente los comandos y códigos utilizados para obtener los resultados, mostrados en la tabla 7.

1- En primer lugar, es necesario descargar las bases de datos que se van a usar, lo cual se realiza con el siguiente comando: ***annotate_variation.pl -downdb -buildver hg38 refGene humandb/***.
   - **"perl"** indica el lenguaje de programación usado (Perl).
   - **"annotate_variation.pl"** corresponde con el nombre del script que se está ejecutando.
   - **"-downdb"** es una opción del script que indica que se va a descargar una base de datos.
   - **"-buildver hg38"** especifica la versión del genoma humano usada, en este caso HG38, pero en algunos casos fue necesario usar HG19.
   - **"refGene"** es el nombre de la base de datos que se va a usar, en este caso refGene.
   - **"humandb/"** es el nombre del directorio dónde se guardará la base de datos.

   Este paso es similar para descargar cualquier banco de datos en los demás tipos de anotación. Es necesario repetirlo para cada *database.*

2- En segundo lugar, se corre el siguiente comando: ***perl table_annovar.pl sinAnotar.avinput humandb/ -buildver hg38 -out misoutputs/GeneBased -remove -protocol refGene,ensGene,knownGene -operation g,g,g -nastring . -csvout -polish.***
   - **"perl"** indica el lenguaje de programación usado (Perl).

- "**table_annovar.pl"** es el nombre del script utilizado. Se puede utilizar también el script "annotate_variation.pl" pero se decidió usar el primero para trabajar más cómodamente con los resultados en forma de tabla.
- "**sinAnotar.avinput"** es el archivo que contiene la información de las variantes previa a la anotación.
- "**humandb/"** es el directorio con las bases de datos que se van a usar.
- "**-buildver hg38"** es la versión del genoma usada.
- "**-out misoutputs/GeneBased"** especifica el directorio donde se guardarán los archivos ("misoutputs") y el prefijo del nombre de los nuevos documentos ("GeneBased").
- "**-remove"** señala que se eliminarán las variantes duplicadas.
- **"-protocol refGene,ensGene,knownGene"** indica las bases de datos que se van a usar. En este caso, la base de datos del NCBI, Ensembl y UCSC.
- **"-operation g,g,g"** especifica la operación que usará para cada protocolo o base de datos. En este caso, "g" de *gene-based annotation*.
- **"-nastring ."** indica el carácter que se usará para representar los valores faltantes en el archivo de salida. En este caso ".".
- **"-csvout"** especifica el formato que tendrá el archivo de salida (CSV).
- **"-polish"** se aplicará un proceso de pulido final a los archivos para facilitar su lectura y trabajo.

3.7.2. FILTER-BASED ANNOTATION:

Sirve para identificar variantes recogidas en bases de datos específicas. Es capaz de indicar si una mutación está registrada en dbSNP, así como determinar la frecuencia alélica en ExAC o en GnomAD.

A continuación, se explica el significado de cada una de las columnas, para la base ExAC y GnomAD, que son las que se usaron para este trabajo. GnomAD tiene 2 bancos de datos, uno solamente de exoma y otro de genoma completo, se usaron ambos debido a que hay 2 mutaciones encontradas fuera de exoma. La tabla 8 corresponde a ExAC y la tabla 9 a GnomAD.

En el caso de la base ExAC (Exome Aggregation Consortium):

- **ExAC_ALL:** (All) muestra la frecuencia alélica en todas las poblaciones.
- **ExAC_AFR:** (African) muestra la frecuencia alélica en poblaciones africanas.
- **ExAC_AMR:** muestra la frecuencia alélica en poblaciones americanas (America).
- **ExAC_EAS:** (East Asian) muestra la frecuencia alélica en poblaciones asiáticas orientales.
- **ExAC_FIN:** (Finnish) muestra la frecuencia alélica en poblaciones finlandesas.
- **ExAC_NFE:** (Non-Finnish European) muestra la frecuencia alélica en poblaciones europeas no finlandesas.
- **ExAC_OTH:** (Other) muestra la frecuencia alélica en otras poblaciones.
- **ExAC_SAS:** (South Asian) muestra la frecuencia alélica en poblaciones del sur de Asia

En el caso de la base GnomAD (tanto *exome* como *genome*):

- **ALL:** frecuencia alélica en todas las poblaciones.
- **AMR:** frecuencia alélica en población americana.
- **ASJ:** frecuencia alélica en poblaciones de judíos asquenazíes.
- **EAS:** frecuencia alélica en poblaciones asiáticas orientales.
- **FIN:** frecuencia alélica en poblaciones finlandesas.
- **NFE:** frecuencia alélica en población europeas no finlandesas
- **OTH:** frecuencia alélica en otras poblaciones
- **SAS:** frecuencia alélica en poblaciones del sur de Asia.

El comando utilizado, tras la descarga de las bases de datos necesarias, fue el siguiente***: perl table_annovar.pl sinAnotar.avinput humandb/ -buildver hg38 -out misoutputs/FilterAnno -remove -protocol exac03,gnomad_genome,gnomad_exome -operation f,f,f -nastring . -csvout -polish.***

Los elementos del código se interpretan de igual manera que los comandos anteriores. En este caso se usa "**-protocol exac03,gnomad_genome,gnomad_exome"** para indicar las 3 bases de datos usadas y "**-operation f,f,f"** para indicar el tipo de operación (*filter-based annotation)*.

3.7.3. REGION-BASED ANNOTATION:

Este modo de anotación permite identificar variantes en regiones específicas del genoma. Es especialmente útil para caracterizar mutaciones que se encuentren en zonas no codificantes. Con *region-based annotation* se puede buscar si una variante, por ejemplo, codifica para un **microRNA o snoRNA** (small nucleolar RNA), si se encuentra en una zona ligada a un **factor de transcripción** o si se halla próxima a una región de ARN genómico. Otras funciones son especificar en qué banda cromosómica se encuentra la mutación (cytoBand) o buscar si ha sido reportada en **GWAS**.

Se usaron 5 protocolos de anotación: cytoBand (busca la banda cromosómica), gwasCatalog (indica si la variante ha sido reportada en algún GWAS), wgRna (busca si se solapa con regiones de microRNAs o snoRNAs), tfbsConsSites (indica si se encuentra en una zona ligada a un factor de transcripción), targetScanS (identifica variantes que interrumpen los sitios de unión a microRNAs predichos).

Este tipo de anotación no permite obtener una tabla con todos los protocolos como en los anteriores casos. No se puede usar por tanto el *script* "table_annovar.pl", en su lugar se usó

"annotate_variation.pl". Un ejemplo de los comandos que se utilizaron para obtener los resultados de este punto es el siguiente: ***annotate_variation.pl -regionanno -dbtype cytoBand -buildver hg38 sinAnotar.avinput humandb/ -out misoutputs/RegAnno.***

Los elementos del comando que varían con respecto a los anteriores son:

- **"annotate_variation.pl"** es el nombre del *script* que va a ser utilizado.
- **"-regionanno"** indica el modo de anotación, que es *region-based annotation.*
- **"-dbtype cytoBand"** especifica la base de datos o protocolo que se va a usar. En este caso "cytoBand".

3.8. Localización de las mutaciones en la proteína

La anotación de las variantes con ANNOVAR aportó numerosa información nueva. Gracias a *gene-based annotation* se puede predecir el impacto de cada mutación a nivel de la proteína codificada (no sinónima/silenciosa), la posición del aminoácido en la cadena peptídica y, en el caso de las mutaciones *missense*, el nuevo aminoácido codificado.
Con esta información se decidió usar el paquete de RStudio **G3viz**, que permite visualizar información genética en forma de diagrama de *lollipops* o "piruletas". Crea una representación esquemática de la proteína a estudio con todos sus dominios y unas "piruletas" o marcadores que señalan dónde se encuentra cada mutación. De esta manera, no solo se puede apreciar fácilmente dónde se agrupan las variantes, sino que también se da a conocer en qué dominio se encuentra cada una. Para crear este gráfico antes es necesario hacer un archivo CSV que contenga la información sobre las variantes. Se excluyó la variante situada en un intrón y la ubicada en una zona de *splicing* debido a que no codifican para ningún aminoácido de la proteína PC1.

3.9. Predicción del efecto de las mutaciones

Existen diferentes herramientas bioinformáticas que permiten predecir *in silico* el efecto de las mutaciones no sinónimas, una de ellas es **PolyPhen-2 (*Polymorphism Phenotyping v2*)**. Se trata de un servidor web de la Universidad de Harvard capaz de predecir el efecto de estas variantes en la estabilidad y función de las proteínas. Esto ayuda a discernir si una mutación es potencialmente benigna o maligna.
Funciona extrayendo información de bases de datos como dbSNP y Ensembl sobre las características de las proteínas afectadas por las variantes. Analiza la posición de la variante en la cadena, la estructura secundaria de la proteína y qué dominios son los más importantes o interaccionan con otras proteínas. Posteriormente, hace una clasificación probabilística y genera una puntuación para cada variante predicha con técnicas de *machine learning*. La puntuación indica la probabilidad de que la variante sea dañina y se interpreta como una escala continua dónde los valores altos indican más probabilidad de que sea perjudicial (50).
Con el propósito de avanzar en el mundo de la biología molecular han salido nuevos programas bioinformáticos. La **inteligencia artificial (IA)** se ha convertido en una valiosa herramienta en este campo. Es capaz de solucionar cuestiones científicas complejas que anteriormente exigían más recursos. Uno de estos desafíos científicos consistía en predecir el plegamiento de las proteínas.
En este contexto, la empresa de Google, Deepmind, lanza en 2021 una IA llamada **AlphaFold**. Durante más de 50 años se han estudiado diversas maneras de predecir la estructura tridimensional de las proteínas, pero ninguna de ellas ha alcanzado una precisión cercana a la experimental como ha hecho AlphaFold. Para hacer esta predicción usa una **red neuronal**

profunda que construye el modelo y lo compara con otros publicados en la *Protein Data Bank* (PBD). Posteriormente, usa técnicas de refinamiento, optimización y *deep learning* para ajustarse a las restricciones físicas y químicas de las proteínas. Este método, además de ser superior a otras alternativas, se aproxima a la exactitud experimental. Tiene una precisión mediana del esqueleto de 0,96 Å. (intervalo de confianza del 95% = 0,85-1,16 Å). Como referencia para la exactitud, el ancho de un átomo de carbono es 1.4 Å (51).
Si bien esta inteligencia artificial puede predecir con bastante exactitud la estabilidad de la proteína en función de su estructura 3D, no sirve para anticipar el efecto de las mutaciones. Cuando Deepmind publicó AlphaFold, señaló que no es esperable que la IA produzca una proteína desplegada dada una secuencia que contiene una mutación puntual desestabilizadora. El programa no ha sido diseñado para predecir el efecto de mutaciones en la estructura (52).

Debido a esta limitación, un grupo de investigadores decidió crear **AlphaMissense**. Se conocen alrededor de 4 millones de variantes *missense,* pero sólo se sabe el efecto clínico de un 2% de las mismas. Usando técnicas de *deep learning* y basándose en AlphaFold, AlphaMissense logra predicciones de patogenicidad de cambio de sentido de última generación. De esta manera, se creó una base de datos con los 71 millones de variantes posibles en el proteoma humano y su posible significación clínica (53).
Para correr el código del programa, debido a la gran capacidad computacional que requiere, se puede hacer desde un *script* de Google Colab. Permite acceder al código desde un entorno de ejecución de Python con acceso a una GPU (Graphics Processing Unit) y TPU (Tensor Processing Unit) remotas con potencia informática de alto rendimiento. La TPU es una unidad de procesamiento, desarrollada por Google, especializada en acelerar tareas de inteligencia artificial y *machine learning*, por lo que es idónea para AlphaMissense.

Es importante señalar que, como se halló una variante de una zona de *splicing* o empalme, se usó otra IA, **SpliceAI**, específicamente diseñada para este propósito. Se trata de un complejo algoritmo de predicción también basado en *deep learning* y de código abierto. Permite conocer un valor delta, que indica la probabilidad de que la mutación sea patogénica. A mayores, señala a qué distancia (en pares de bases) se encuentra el punto de interés (posición de aceptor o donante de empalme) y si hay una ganancia o pérdida de bases. Da un valor delta para cada una de estas 4 opciones: pérdida de aceptor, pérdida de donante, ganancia de aceptor, ganancia de donante (54).

4. RESULTADOS

4.1. Localización de las variantes

hg19 (GRCh37)	hg38 (GRCh38)	Exón	Localización del exón	Probandos	SNP de referencia
16:2147139:G:A	chr16:2097138:G:A	*Mutación intrónica*	-	-	rs540447857
16:2152577:C:T	chr16:2102576:C:T	25	2102381-2102633	-	rs367709319
16:2153404:C:T	chr16:2103403:C:T	23	2103266-2103895	Probando 3	rs749700361
16:2154548:C:T	chr16:2104547:C:T	22	2104498-2104642	-	rs772293778
16:2154626:G:A	chr16:2104625:G:A	22	2104498-2104642	-	rs574400883
16:2159430:G:A	chr16:2109429:G:A	15	2108252-2111871	Probando 4	rs1245626836
16:2160512:G:A	chr16:2110511:G:A	15	2108252-2111871	-	rs961711184
16:2161078:G:A	chr16:2111077:G:A	15	2108252-2111871	Probando 1	rs771687427
16:2163160:A:C	chr16:2113159:A:C	a 2 pb del EX12	12:2113161-211392	Probando 2	rs1345202767
16:2164380:C:T	chr16:2114379:C:T	11	2114170-2114925	-	rs151016310

Tabla 4: localización de las variantes en el genoma. De las mutaciones referidas en la tabla anterior, las que refieren probando provienen de la secuenciación de exoma de la cohorte española (N=360), el resto de las mutaciones se han identificado en el estudio de Satterstrom et al. (referencia) en diferentes cohortes mundiales a partir de los datos del ASC.

Con la información de la tabla 4 se pueden sacar los siguientes resultados.

- Una mutación (16:2147139:G:A en hg19) se encuentra en un intrón.
- Una mutación (16:2163160:A:C) se encuentra en una zona no codificante, a 2 pares de bases del exón 12 (EX12).
- El resto de las variantes, 8, se encuentran en exones. Una en el 25, otra en el 23, 2 en el 22, 3 en el 15 y una en el 11.

A continuación, se muestra la tabla 5.

Exón	N.º de mutaciones	Exón	N.º de mutaciones	Exón	N.º de mutaciones	Exón	N.º de mutaciones
Exón 1	23	**Exón 13**	8	**Exón 25**	24	**Exón 37**	19
Exón 2	6	**Exón 14**	15	**Exón 26**	16	**Exón 38**	13
Exón 3	10	**Exón 15**	267	**Exón 27**	24	**Exón 39**	20
Exón 4	9	**Exón 16**	15	**Exón 28**	12	**Exón 40**	28
Exón 5	46	**Exón 17**	24	**Exón 29**	9	**Exón 41**	21
Exón 6	16	**Exón 18**	35	**Exón 30**	7	**Exón 42**	17
Exón 7	15	**Exón 19**	19	**Exón 31**	4	**Exón 43**	24
Exón 8	8	**Exón 20**	14	**Exón 32**	3	**Exón 44**	15
Exón 9	11	**Exón 21**	14	**Exón 33**	9	**Exón 45**	42
Exón 10	20	**Exón 22**	11	**Exón 34**	9	**Exón 46**	40

Exón 11	36	**Exón 23**	56	**Exón 35**	5		
Exón 12	13	**Exón 24**	14	**Exón 36**	22		

Tabla 5: conteo de variantes patogénicas y probablemente patógenicas por exón en la ADPKD.

La tabla con el conteo de variantes por exón en la ADPKD reflejó lo siguiente:

- En todos los exones hay variantes que pueden causar la enfermedad.
- Existe un exón, el 15, que contiene muchas más variantes patogénicas demostradas (267) que ningún otro.
- Hay 7 exones (5, 11, 15, 18, 23, 45 y 46) con más de 35 mutaciones.

Las mutaciones halladas en los individuos con TEA coinciden en los exones que más mutaciones acumulan en la ADPKD.

4.2. Estudio de expresión

Los *heatmap* usan una escala visual de colores para representar la expresión de los exones o isoformas. Indicando los colores más oscuros una mayor expresión.

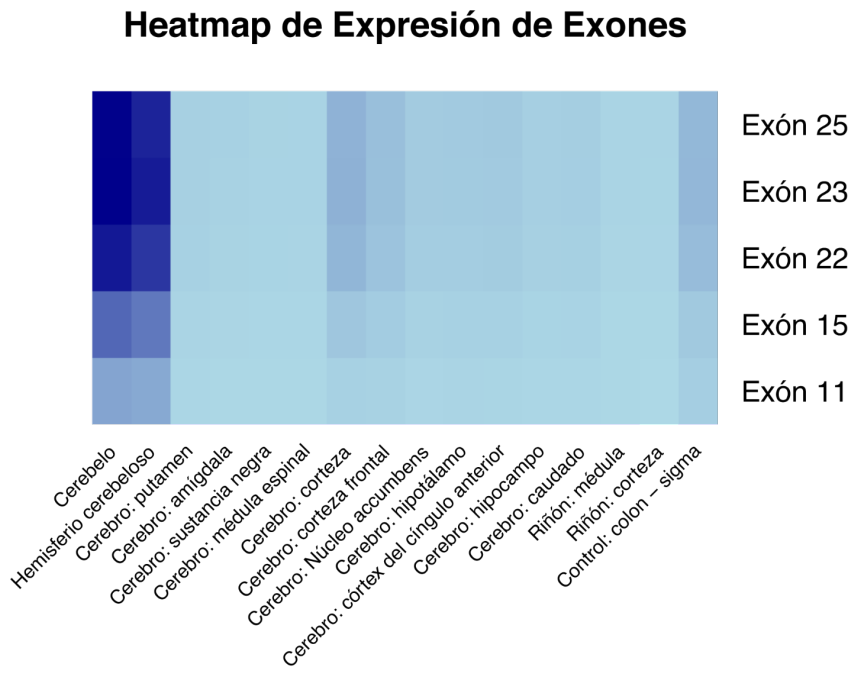

Gráfico 1: Heatmap de expresión de los exones realizado con los datos de GTEx. En el eje X se muestran los tejidos y en el Y los exones que albergan las variantes a estudio. El gradiente de colores (de azul claro a azul oscuro) refleja el nivel de expresión.

Analizando visualmente el **heatmap de exones** se aprecia una mayor expresión en el cerebelo y hemisferio cerebeloso para todos los exones que portan las mutaciones, en relación al resto de tejidos . La expresión también es elevada en el cerebro, en especial en la corteza frontal, córtex del cíngulo anterior y núcleo caudado. También se percibe que, a medida que los exones se localizan más alejados del comienzo del gen, mayor es su expresión en varios tejidos (el número 25 se transcribe notablemente más que el 11).

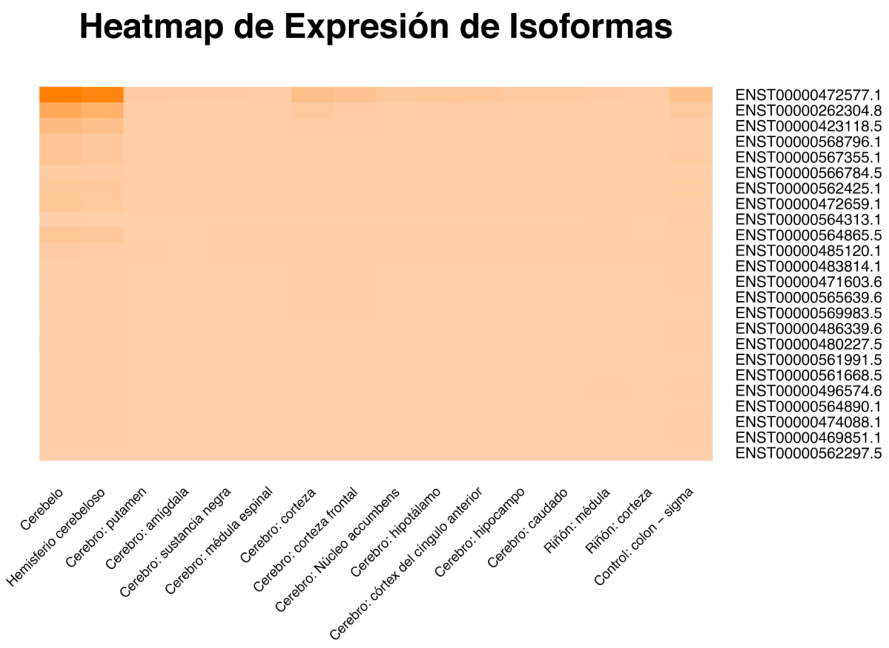

Gráfico 2. Heatmap de expresión de las isoformas realizado con los datos de GTEx. En el eje X se muestran los tejidos y en el Y las principales isoformas. El gradiente de colores (de naranja claro a naranja oscuro) refleja el nivel de expresión.

La interpretación de **mapa de calor de isoformas del gen** se realiza de igual manera. Una isoforma es una de las formas alternativas que un gen tiene para expresarse. Los diferentes sitios de *splicing* o empalme e iniciación de la traducción dan lugar a proteínas con distintas funciones biológicas y formadas por distintos aminoácidos. Pero todas ellas codificadas por el mismo gen, en este caso *PKD1* (55).

Las isoformas más presentes en el organismo son las halladas en las filas superiores (*ENST00000472577.1, ENST00000262304.8, ENST00000423118.5*). Los resultados son similares a los del *heatmap* de exones en relación a la expresión, así pues, hay también una mayor expresión de las isoformas *ENST00000472577.1, ENST00000262304.8, ENST00000423118.5* y *ENST00000568796.1* en los tejidos cerebelosos y cerebrales.

En ambos mapas de calor llama la atención la **mínima expresión existente en los tejidos renales en comparación con los tejidos del SNC.** Se evidencia incluso que la transcripción del gen en el tejido que se usa de control (colon) es mayor que en el tejido renal.

Se pueden comparar los valores de la expresión de las isoformas con los datos de la tabla que se usó para crear el heatmap. En el caso de la isoforma *ENST00000472577.1*, que es la más presente en el organismo, tiene un valor de 81,8 TPM en la corteza cerebral, 17,2 en la médula renal y 12,9 en la corteza renal. Si se calcula la media de expresión en los riñones y se dividen los TPM se observa que la isoforma *ENST00000472577.1* se expresa aproximadamente **5.44 veces más en la corteza cerebral que en el riñón.**

4.3. Frecuencia de las variantes

La tabla resultante de la estimación de la frecuencia de las variantes utilizando los datos de GnomAD es la siguiente.

variante (GRch37)	Frecuencia en población europea	Frecuencia *non-neuro*
16:2147139:G:A	5.296 × 10^(-5)	3.174 × 10^(-5)
16:2152577:C:T	1.794 × 10^(-5)	2.281 × 10^(-5)
16:2153404:C:T	1.299 × 10^(-4)	1.425 × 10^(-4)
16:2154548:C:T	1.176 × 10^(-4)	1.107 × 10^(-4)
16:2154626:G:A	5.765 × 10^(-5)	5.139 × 10^(-5)
16:2159430:G:A	6.482 × 10^(-5)	7.343 × 10^(-5)
16:2160512:G:A	0	0
16:2161078:G:A	1.576 × 10^(-5)	1.973 × 10^(-5)
16:2164380:C:T	0	0
16:2163160:A:C	0	0

Tabla 6: frecuencia de las variantes a partir de los datos de GnomAD.

Tal y como se puede apreciar, la frecuencia poblacional de las mutaciones es extremadamente baja. Incluso 3 de ellas nunca han sido reportadas. Tan solo 3 variantes (16:2147139:G:A; 16:2154548:C:T y 16:2154626:G:A) tienen una frecuencia mayor en población europea que en el grupo control (sin enfermedades neuropsiquiátricas).

4.4. Anotación en ANNOVAR

4.4.1. GENE-BASED ANNOTATION

La primera tabla (tabla 7) muestra la *gene-based annotation* usando **refGene.** Se decidió no incluir las otras 2 tablas (en las que se usa ensGene y knownGene) debido a que los resultados son idénticos.

Posición	Localización	Gen	Función gen	Gen de referencia / Cambio aminoacídico
2097138 : G:A	intronic	PKD1	.	.
2102576 : C:T	exonic	PKD1	synonymous SNV	PKD1:NM_000296:exon25:c.G9006A:p.S3002S,PKD1:NM_001009944:exon25:c.G9006A:p.S3002S
2103403 : C:T	exonic	PKD1	nonsynonymous SNV	PKD1:NM_000296:exon23:c.G8654A:p.R2885Q,PKD1:NM_001009944:exon23:c.G8654A:p.R2885Q
2104547 : C:T	exonic	PKD1	synonymous SNV	PKD1:NM_000296:exon22:c.G8112A:p.A2704A,PKD1:NM_001009944:exon22:c.G8112A:p.A2704A

2104625: G:A	exonic	PKD1	synonymous SNV	PKD1:NM_000296:exon22:c.C8034T:p.L2678L,PKD1:NM_001009944:exon22:c.C8034T:p.L2678L
2109429: G:A	exonic	PKD1	nonsynonymous SNV	PKD1:NM_000296:exon15:c.C5738T:p.A1913V,PKD1:NM_001009944:exon15:c.C5738T:p.A1913V
2110511V: G:A	exonic	PKD1	synonymous SNV	PKD1:NM_000296:exon15:c.C4656T:p.V1552V,PKD1:NM_001009944:exon15:c.C4656T:p.V1552V
2111077: G:A	exonic	PKD1	nonsynonymous SNV	PKD1:NM_000296:exon15:c.C4090T:p.R1364C,PKD1:NM_001009944:exon15:c.C4090T:p.R1364C
2113159: A:C	splicing	PKD1	.	.
2114379: C:T	exonic	PKD1	nonsynonymous SNV	PKD1:NM_000296:exon11:c.G2644A:p.V882M,PKD1:NM_001009944:exon11:c.G2644A:p.V882M

Tabla 7: anotación en *gene-based annotation* con ANNOVAR. Los nombres de las columnas del archivo original fueron traducidos al castellano. Los nombres, explicados anteriormente, eran: position, function, gene, exonicFunc y AAchange. En el caso de la mutación en la zona de *splicing*, ANNOVAR aportó otra columna (GeneDetail) con la siguiente información: NM_000296:exon12:c.2985+2T>G;NM_001009944:exon12:c.2985+2T>G.

Las 4 casillas sombreadas en un tono más oscuro muestran la información anotada por el programa. Si bien ya se sabía qué mutaciones se encontraban en exones, la columna "**Localización"** muestra que hay una variante que se halla en una zona de *splicing*.

Es especialmente relevante la columna "**Función gen"**, pues indica la implicación que tiene la mutación en la traducción de ARNm a un aminoácido. Concluye que las variantes en las posiciones 2103403, 2109429, 2111077, 2114379 son **no sinónimas**, es decir, forman un codón que codifica para un aminoácido distinto al original, con las implicaciones estructurales y funcionales que puede haber para la proteína.

La columna "**Gen de referencia / Cambio aminoacídico"** también aporta datos de interés pues informa sobre el aminoácido original, su posición y, en caso de las variantes no sinónimas, el aminoácido alternativo. Por ejemplo, en la mutación con la posición 2103403 indica p.R2885Q. Lo cual se interpreta de la siguiente manera: "p" indica que se está describiendo un cambio en la proteína; "R" representa al aminoácido original, que es la arginina; "2885" es en número de la posición del aminoácido en la secuencia proteica, "Q" es el nuevo aminoácido, en este caso la glutamina.

4.4.2. FILTER-BASED ANNOTATION:

Position	ExAC_ALL	ExAC_AFR	ExAC_AMR	ExAC_EAS	ExAC_FIN	ExAC_NFE	ExAC_OTH	ExAC_SAS
2097138: G:A	4.983 x 10^-2	0	0	1.6 x 10^-3	0	0	0	0
2102576: C:T	4 x 10^-4	0	0	0	0	3.152 x 10^-2	0	2.5 x 10^-3
2103403: C:T	8.891 x 10^-2	0	8.809 x 10^-2	1 x 10^-4	0	1 x 10^-4	0	6.121 x 10^-2
2104547: C:T	2 x 10^-4	0	0	2 x 10^-3	0	3 x 10^-4	0	0
2104625: G:A	6 x 10^-4	1.9 x 10^-3	2.7 x 10^-3	0	0	2 x 10^-4	0	8 x 10^-4
2109429: G:A	.	.	.	.	.	.	.	.
2110511V: G:A	.	.	.	.	.	.	.	.
2111077: G:A	5.927 x 10^-2	0	0	7 x 10^-4	0	1.552 x 10^-2	0	0
2113159: A:C	.	.	.	.	.	.	.	.
2114379: C:T	1 x 10^-4	1.9 x 10^-3	0	0	0	0	0	0

Tabla 8: frecuencia poblacional de las variantes anotadas con ANNOVAR usando la base de datos ExAC.

Position	gnomAD_genome_ALL	gnomAD_genome_AFR	gnomAD_genome_AMR	gnomAD_genome_ASJ	gnomAD_genome_EAS	gnomAD_genome_FIN	gnomAD_genome_NFE	gnomAD_genome_OTH
2097138: G:A	1.00x10^-4	0	0	0	1.90x10^-3	0	6.70x10^-2	0
2102576: C:T	.	.	.	.	.	.	.	.
2103403: C:T	1.00x10^-4	2.00x10^-4	0	0	0	0	1.00x10^-4	0
2104547: C:T	3.34x10^-2	0	0	0	0	0	6.83x10^-2	0
2104625: G:A	3.52x10^-2	0	0	0	0	0	7.19x10^-2	0
2109429: G:A	3.23x10^-2	0	0	0	0	0	0	1.00x10^-3
2110511V: G:A	.	.	.	.	.	.	.	.
2111077: G:A	3.23x10^-2	0	0	0	0	0	6.68x10^-2	0
2113159: A:C	0	0	0	0	0	0	0	0
2114379: C:T	4.00x10^-4	1.50x10^-3	0	0	0	0	0	0

Tabla 9: frecuencia poblacional de las variantes anotadas con ANNOVAR usando las bases de datos de GnomAD exome y GnomAD genome.

4.4.3. REGION-BASED ANNOTATION

- La anotación en **"cytoBand"** reveló que todas las variantes se encontraban en la misma banda cromosómica, 16p13.3*,* que es dónde se encuentra *PKD1.*
- La anotación en el protocolo **"gwasCatalog"** otorgó resultados negativos, lo cual significa que ninguna variante ha sido descrita en un GWAS.
- Los otros 3 protocolos también devolvieron un resultado negativo, se puede descartar por tanto que las mutaciones se encuentren en zonas de microRNA, snoRNA, zonas ligadas a factores de transcripción o zonas que interrumpen los sitios de unión a microRNAs predichos.

4.5. Localización de las mutaciones en la proteína

A continuación, se muestra el gráfico de localización en el que se puede apreciar visualmente dónde se sitúa cada mutación en la proteína final con sus diferentes dominios. El gráfico usa un código de colores. Amarillo para las variantes *missense* y rojo para las silenciosas. 3 variantes missense caen en dominios *PKD* de repetición y la restante entre el dominio *REJ* y el dominio *PLAT*.

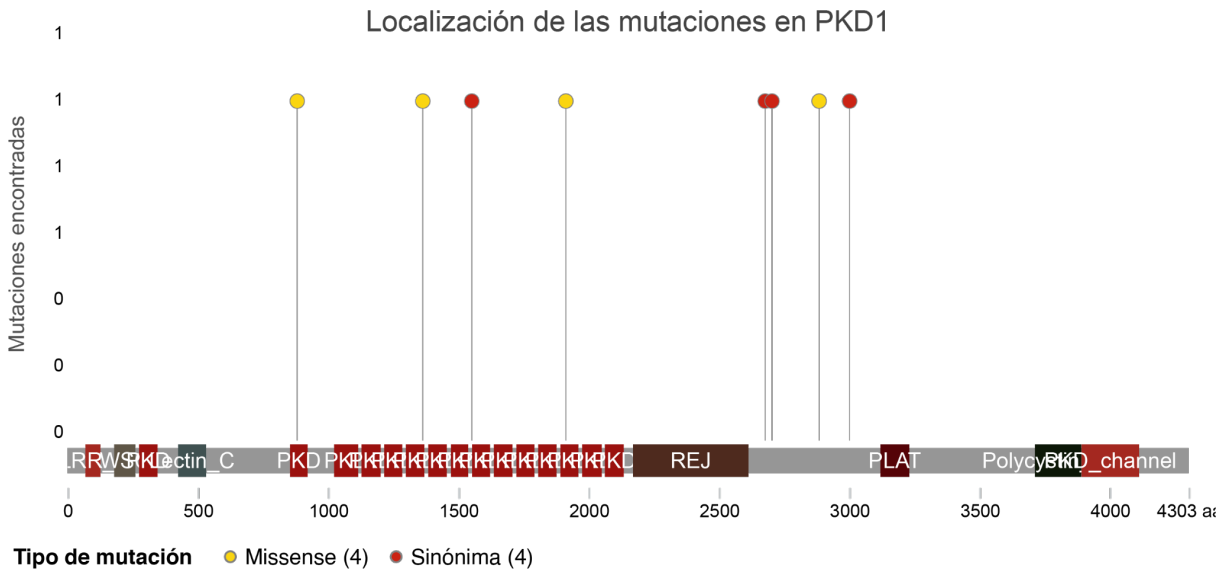

Gráfico 3: localización de las variantes a lo largo de la proteína. Se muestra cada mutación (en amarillo las *missense* o no sinónimas, en rojo las sinónimas). Los números del eje X representan la posición del aa en la cadena y las letras corresponden con los nombres de los dominios.

4.6. Predicción del efecto de las mutaciones

Los resultados de **Polyphen-2** y **AlphaMissense** para las **mutaciones *missense*** se recogen en la siguiente tabla.

Posición	Cambio de aminoácido	Resultados de Polyphen-2	Sensibilidad/ Especificidad	Resultados AlphaFold
2103403	p.R2885Q	BENIGN with a score of 0.007	S: 0.96; E: 0.75	0.0684 likely_benign
2109429	p.A1913V	POSSIBLY DAMAGING with a score of 0.925	S: 0.81; E: 0.94	0.1238 likely_benign
2111077	p.R1364C	PROBABLY DAMAGING with a score of 0.998	S: 0.27; E: 0.99	0.1289 likely_benign
2114379	p.V882M	PROBABLY DAMAGING with a score of 0.994	S: 0.69; E: 0.97	0.1147 likely_benign

Tabla 10: resultados de Polyphen-2 y AlphaMissense. La columna de "sensibilidad/especificidad" fue aportada por Polyphen-2 y por tanto únicamente es aplicable a sus resultados.

El análisis de la **mutación de empalme**, chr16-2113159-A-C (hg38), con **SpliceAI** reveló lo siguiente:

1. Una puntuación de Δ de **0.01** para pérdida de aceptor a una distancia de 216pb.
2. Una puntuación de Δ de **0.93** para pérdida de donante a una distancia de -2pb.
3. Una puntuación de Δ de **0.01** para ganancia de aceptor a una distancia de -100pb.
4. Una puntuación de Δ de **0.16** para ganancia de donante a una distancia de -2pb.

Tal y como se puede ver, **la probabilidad de que la variante sea patogénica es muy alta** ya que la puntuación para pérdida de donante es de 0.93.

5. DISCUSIÓN

Debido a que los TEA son de los TND más heredables que existen, se han descrito numerosos genes implicados en su patogenia. Sin embargo, *PKD1* nunca ha sido definido como un gen de riesgo.

Sobre la **caracterización fenotípica y genotípica** de aquellos probandos de los que hay información, cabe destacar lo siguiente. El análisis completo del exoma de todos ellos permite descartar que exista alguna otra alteración genética capaz de explicar los fenotipos. Situando a las variantes en *PKD1* como mejores candidatas para justificar los TEA. Además, en la mayor parte de los probandos se comprobó que fuesen negativos específicamente para X frágil. **Casi todos los probandos (1, 2 y 4) proceden de familias *simplex***, sin más individuos afectos. Estos 3 individuos tienen hermanos o hermanas no afectas de TEA. Por el contrario, el probando 3, sin hermanos, tiene un primo diagnosticado de TEA y una prima con síndrome de Asperger, por lo que se podría considerar una familia *multiplex*, con 2 o más individuos afectos. El hecho de que la mayor parte de probandos sean de familias *simplex* es positivo, pues la probabilidad de que una variante señale un gen implicado en la enfermedad es mayor que en las familias *multiplex* (56). El estudio de la caracterización fenotípica reveló también las edades de los progenitores en el momento del nacimiento de los probandos. La edad paterna y materna juega un papel esencial en la etiología de los TEA. Por cada 10 años que aumenta la edad materna el riesgo de TEA aumenta un 18%. Este incremento es mayor en el caso de la edad paterna, que supone un aumento de riesgo del 21% (16). Tan solo los progenitores del probando 1 tenían menos de 36 años en el momento de su nacimiento. Las edades paternas de los probandos 3 y 4 fueron 41 y 49 años respectivamente. Se puede observar que, en general, las edades de los progenitores son avanzadas. La variedad de cuestionarios realizados por los probandos impide la comparación directa del grado de TEA entre ellos. Todos tienen en común un retraso en la adquisición del lenguaje, aunque la habilidad de expresión verbal no es la misma en todos ellos. El probando 1 habla de manera fluida mientras que los demás solo emiten frases simples.

Al utilizar *UCSC Genome Browser* se apreció que *PKD1* se encontraba **muy próximo al gen *TSC2***. Ambos se ubican en la **misma banda cromosómica**, 16p13.3. *TSC2* es responsable, junto con *TSC1* de la esclerosis tuberosa, una enfermedad genética rara. En ella se altera el desarrollo de la piel, cerebro, riñones, corazón y pulmones. También hay una afectación a nivel conductual, social, intelectual, psicosocial y psiquiátrico. Se considera que los TEA están presentes hasta en el 61% de los pacientes (57). Además, los genes *TSC1* y *TSC2* se encuentran en la lista de genes del estudio previamente comentado que relacionaba TEA e IRC. Tomando los datos de UCSC, las posiciones de *PKD1* y *TSC2* son 2,088,708-2,135,898 y 2,047,985-2,089,491 respectivamente. Los **genes incluso se superponen** durante el fin de *TSC2* y el inicio de *PKD1* durante un tramo de 783 pb (2,089,491−2,088,708=783 pb).

Si se compara la **ubicación exónica** de las variantes en *PKD1* en la cohorte de TEA con las de la base de datos de ADPKD se pueden observar patrones similares. En primer lugar, todas las regiones exónicas de *PKD1* contienen mutaciones capaces de causar la ADPKD, lo que sugiere una importancia funcional de todo el gen. Uno de los hallazgos más destacados es la concentración de variantes patogénicas en el exón 15, pues contiene un total de 267 mutaciones. Esto lo convierte en un **punto crítico en la patogénesis de la ADPKD** ya que las alteraciones en este exón tienen un impacto significativo en el gen. Por tanto, es probable que las mutaciones en el exón 15 de la cohorte de TEA provoquen una alteración funcional que refleje un fenotipo compatible con este TND. En total 3 de las 10 variantes estudiadas se encontraban en el exón

15. La muestra es pequeña, pero se puede intuir que, en el caso de los TEA, el exón 15 también es el que más variantes acumula. Se identificaron 7 exones (5, 11, 15, 18, 23, 45 y 46) que contienen más de 35 mutaciones cada uno en la ADPKD. De todos ellos, los que albergan variantes en la cohorte de TEA son el 11, 23 y el ya mencionado 15. Son por tanto candidatos prometedores para futuras investigaciones. Por último, los exones 25 y 22 en los que se hallaron las variantes tan solo contienen 24 y 11 mutaciones patológicas respectivamente descritas en ADPKD. Todos estos resultados apoyan la hipótesis de la relación de *PKD1* con los TEA, pues existe una evidente correlación entre las mutaciones descritas en la ADPKD y las halladas en esta investigación.

La **elevada expresión de *PKD1*** y de su proteína, PC1, en los **tejidos del SNC** ha sido documentada tanto en estudios previos como en esta investigación (a través del estudio de expresión). Teniendo en cuenta la notable implicación de este gen en la ADPKD resultaría inusual que, siendo su expresión **5.44** veces mayor en el cerebro que en los riñones, no estuviese involucrado en los TEA o en alguna otra patología neuropsiquiátrica. También se evidenció que la expresión en el cerebelo es **25,88** veces superior que en el riñón. Es necesario comparar estos datos con la base neuroanatómica de los TEA. Si bien en un primer momento el **papel del cerebelo** podría parecer menos relevante en este tipo de patologías, existe evidencia de su fuerte implicación. Según numerosas investigaciones, el rol del cerebelo en funciones cognitivas superiores como el lenguaje, procesamiento cognitivo y regulación afectiva ha sido históricamente subestimado. Se ha demostrado a través de estudio por resonancia magnética nuclear funcional la implicación de esta estructura en la cognición social (58). De hecho, la primera alteración neuroanatómica descrita en los TEA (en 1988) fue una hipoplasia de los lóbulos centrales del vermis cerebeloso (59). Un hallazgo constante es la disminución (hasta en un 25%) del número y tamaño de las células de Purkinje en la corteza neocerebelosa posterolateral y arquicerebelosa. La implicación de la estructura nerviosa en los TEA ha sido estudiada con animales de experimentación. Tsai et al. 2012 reveló que al inhibir el gen *TSC1* específicamente en las células de Purkinje los ratones mostraban comportamientos compatibles con TEA. Sin embargo, Donovan & Basson, 2017 observaron en su laboratorio que una reducción del 50% del número de células de Purkinje no era suficiente para causar clínica de TEA. Aunque se debe profundizar en el estudio de la contribución específica de las células de Purkinje, **ambos estudios concluyen que el cerebelo tiene un papel crítico en los TEA**.

Entre las funciones superiores cognitivas de la **corteza frontal** se encuentran la toma de decisiones, comunicación, memoria, aprendizaje, control de emociones y comportamiento social. Se han objetivado anomalías en crecimiento, grosor y organización neuronal a nivel cortical. Esto convierte a la corteza frontal, junto con el cerebelo y la amígdala, en uno de los tejidos más estudiados en los TEA. Varios estudios longitudinales han revelado un crecimiento anormal y aumento de volumen de materia gris. Esta variación en el volumen se explica por un incremento del área de superficie cortical y, sobre todo, por un aumento en el grosor de la corteza (58). En el presente estudio se determinó una **expresión elevada en la corteza cerebral (81,8 TPM)** de la isoforma más abundante, *ENST00000472577.1*.

El análisis de ambos heatmaps es congruente con la base neuroanatómica de los TEA. Aunque algunos exones que albergan las variantes se expresan más que otros, todos muestran la misma relación entre los diferentes tejidos. Se observa un fenómeno similar con las isoformas.

PKD1 nunca ha sido estudiado como gen de riesgo para los TEA, luego no existe una **propuesta fisiopatológica específica**. Sin embargo, se ha relacionado con la epilepsia y, teniendo en cuenta que en ambas patologías la corteza cerebral juega un papel crucial, el fundamento

biológico se podría extrapolar a los TEA. Además, como se mencionó en el apartado de comorbilidades, la epilepsia es una de las comorbilidades asociadas a los TEA. Esto refuerza la hipótesis de la implicación de *PKD1* en estos TND. La proteína PC1 desempeña diversas funciones en la proliferación celular, apoptosis y transporte de iones, sobre todo de calcio. Adquiere un **papel importante en las neuronas simpáticas**, funcionando como un activador de canales de calcio dependientes de voltaje y de canales rectificadores de potasio intracelulares. Ambos canales son activados por proteínas G. Una deleción de PC1 causa una pérdida completa de la estabilidad iónica, con las consecuencias que eso conlleva. La correcta regulación del calcio es clave para la estabilidad y excitación neuronal. En especial unos niveles adecuados de calcio intracelular son críticos para el correcto desarrollo de las funciones neuronales (40). Estos hallazgos se usaron para justificar la actividad epileptógena de mutaciones en *PKD1*. Es posible que la pérdida de la función neuronal debida a la regulación del calcio también **explique la fisiopatología de los TEA**.

No se halló en la literatura científica ningún estudio que relacionase la ADPKD con los TEA. Sin embargo, existe un artículo que examina la asociación entre la **insuficiencia renal crónica (IRC)** y los TEA. Existen muchos genes que han sido descritos en ambas entidades como pueden ser *TSC1, TSC2, MECP2, FMR1* y *NF1*. Además, la prevalencia de IRC en la población con DI o TEA se estima en torno a un 25%. Recientemente también se estudió la frecuencia de TEA en una cohorte de 224 pacientes pediátricos con IRC de un centro médico. 24 individuos estaban diagnosticados de TEA. Si bien los datos epidemiológicos respaldan una relación entre los TEA y la IRC se desconoce el mecanismo patológico. Las mutaciones en genes comunes pueden explicar esta asociación, pero no se pueden descartar otros factores como el daño cerebral adquirido, la prematuridad y la epilepsia temprana (60).

El estudio de prevalencia mediante la **comparación de los datos de GnomAD** reveló resultados poco concluyentes. La principal limitación que existe para estudiar la presencia de estas variantes es su **baja frecuencia**. El objetivo inicial consistía en comparar las frecuencias de la población general con las de la población *non-neuro*. Teóricamente, al tener los TEA una prevalencia elevada, la frecuencia de las variantes en la población total debería ser superior a la del grupo *non-neuro* (formado únicamente por individuos sin patologías neuropsiquiátricas). Al ser las mutaciones tan infrecuentes los resultados están muy posiblemente artefactados. Sería necesario un tercer grupo formado por individuos con TEA. De esta forma se podría estimar directamente cómo de comunes son estas variantes en la propia población a estudio. De las 10 variantes solo 7 tenían una frecuencia conocida en la población europea. De estas 7, únicamente 3 de ellas eran más comunes en población europea que en el grupo de control.

La anotación a través de ANNOVAR permitió profundizar en la comprensión de cada una de las variantes. El análisis se llevó a cabo con tres enfoques diferentes: anotación basada en genes *(gene-based annotation)*, anotación basada en filtros (*filter-based annotation)* y anotación basada en regiones (*region-based annotation)*.

Para la **anotación basada en genes** *(gene-based annotation)* se usó con el protocolo *refGene*, que usa la base de datos del NCBI. También se usaron los protocolos *ensGene* (Ensembl) y *knownGene* (UCSC) pero no se incluyeron las tablas ya que los resultados fueron exactamente iguales a los de refGene. Esta **consistencia entre los resultados** de las tres bases de datos es una muestra de la robustez de los hallazgos. La tabla incluye información fundamental de cada variante, su posición precisa en el genoma, si se encuentra en zona codificante o no, si hay cambio en el aminoácido (variante sinónima o no sinónima), la posición del aa en la cadena peptídica y, en el caso de las variantes no sinónimas, el nuevo aminoácido codificado.

Se objetiva que **4 variantes son sinónimas** (*2110511V: G:A, 2104625: G:A, 2104547: C:T, 2102576: C:T*), por tanto, es muy poco probable que puedan ser patogénicas, ya que, aunque se encuentran en exones, el cambio de bases codifica exactamente el mismo aminoácido. Tradicionalmente estas mutaciones se han considerado **silenciosas o *silent*** debido al poco impacto que se presupone que tienen en la proteína. Sin embargo, cada vez existe más evidencia de que estas mutaciones pueden desempeñar un papel importante. Incluso se está abandonando el término "mutación silenciosa" y reemplazando por "mutación sinónima". Estas variantes pueden alterar la optimización de la expresión acelerando o desacelerando la elongación de la traducción. También pueden afectar en la estabilidad de las moléculas de ARNm y en la descomposición antes de su traducción. Por último, la ubiquitinación mejorada puede provocar un plegamiento erróneo de las proteínas y una degradación aumentada (57). Aunque siguen existiendo muchos interrogantes acerca del significado funcional y clínico de las mutaciones sinónimas, actualmente no se puede descartar que tengan una implicación. Por ende, con la evidencia actual, **no es posible negar** por completo que las 4 variantes sinónimas halladas tengan una significación clínica.

De las 6 restantes, **una se encuentra en un intrón** (*2097138: G:A*). Pese a encontrarse en una región no codificante, su patogenicidad no puede ser completamente descartada. Su análisis sería muy complejo y podría requerir incluso un estudio funcional, por este motivo se decidió no abordar su exploración en detalle. Otra variante (*2113159: A:C*), estaba ubicada en una **zona de empalme o *splicing***. Las zonas de *splicing* son regiones específicas en un pre-ARNm dónde ocurre el proceso de empalme durante la maduración del ARN. . Las consecuencias de las mutaciones en estas zonas del ADN pueden tener consecuencias fatales para la función de la proteína y alterar gravemente su estructura. Es necesario interpretar de forma individual los resultados que aporta ANNOVAR sobre esta variante, que son los siguientes: "*NM_000296:exon12:c.2985+2T>G;NM_001009944:exon12:c.2985+2T>G*". "*NM_000296*" y "*NM_001009944*" son los identificadores de las isoformas de ARN dónde ocurre la mutación. El resultado es igual en ambas, la variante ocurre en el exón 12 del gen (*:exon12*). "*2985+2*" significa que la mutación ocurre a 2 nucleótidos después de los últimos del exón 12. "*T>G*" indica el cambio de base.

ANNOVAR mostró que las restantes **4 variantes eran no sinónimas** (*2114379: C:T, 2103403: C:T, 2111077: G:A, 2109429: G:A*). Las mutaciones no sinónimas representan el paradigma fundamental en la genética molecular y en la patogenia de las enfermedades genéticas. Su capacidad de alterar la estructura tridimensional de las proteínas y su función biológica las convierte en el prototipo de mutación con relevancia clínica y fisiológica. El programa no indica la probabilidad de que la variante sea patogénica, simplemente indica que se trata de una mutación no sinónima y señala los aminoácidos alternativos con su posición en la cadena peptídica. Esta información es crucial, pues es necesaria para utilizar herramientas de predicción de patogenicidad.

La anotación basada en filtros *(filter-based annotation)* mostró las prevalencias de cada variante según diversas bases de datos *(exac03, GnomAD exome, GnomAD gneome)* y dividida por poblaciones. Una vez más se aprecia que la frecuencia de las mutaciones es muy baja. En muchas bases de datos nunca se han registrado por lo que es muy difícil comparar los resultados. No obstante, se puede afirmar que no hay ninguna variante claramente más prevalente que las demás. Ciertas variantes tienen frecuencias más altas en subpoblaciones específicas, esto sugiere **posibles patrones de variación geográfica**. Un ejemplo es la variante *2102576: C:T* que posee una frecuencia baja en la población general (ExAC_ALL: 0.0004) y una superior en poblaciones europeas no finlandesas (ExAC_NFE: 0.03152). Esto resalta la importancia de tener en cuenta la variabilidad genética entre poblaciones al analizar frecuencias alélicas.

Se utilizaron 5 protocolos distintos para la **anotación basada en regiones** *(region-based annotation)*. El primero, cytoBand, tenía como objetivo comprobar que todas las mutaciones se encontraban en la misma banda cromosómica (16p13.3). El segundo, gwasCatalog, tenía gran relevancia, pues indica si alguna de las variantes fue reportada en algún GWAS. La importancia de este hecho radica en que, si una variante es capaz de producir una enfermedad o alteración, es posible que también pueda producir otra distinta. Este fenómeno se conoce como **pleiotropía** y sugiere que una única mutación puede afectar a varios caracteres distintos y no relacionados entre sí. El programa indicó que **ninguna de las variantes había sido reportada nunca en un GWAS**, esto resulta coherente ya que las frecuencias de las variantes son muy bajas y los GWAS son más sensibles cuando se trabaja con variación común.
Se usaron otros 3 protocolos para la anotación basada en regiones. El objetivo era caracterizar e investigar la variante *2097138: G:A*, que no se encuentra en una zona codificante del ADN. Con los protocolos *wgRna, tfbsConsSites, targetScanS* se puede estudiar si las mutaciones están implicadas en mecanismos de regulación génica. Los resultados fueron **negativos** por lo que se puede descartar que esta variante esté involucrada en procesos de control del ADN.
ANNOVAR ha demostrado ser una **valiosa herramienta** para anotar variantes funcionalmente aportando datos clave para este estudio. Aportó información detallada sobre el tipo de mutación, su posición, implicación en la proteína, cambio aminoacídico, frecuencia poblacional y descartó que las variantes no codificantes estuviesen implicadas en procesos de regulación del ADN. Queda reflejada la importancia de este tipo de programas, que permiten contextualizar las mutaciones y orientar la investigación.

La herramienta G3viz reveló un **diagrama de la proteína PC1** con marcadores señalando las 8 mutaciones halladas en zonas codificantes. Se puede apreciar cómo 4 mutaciones no se ubican en ningún dominio específico (una de ellas *missense*), pero no esto no significa carezcan de relevancia. La función de una proteína no depende únicamente de los dominios específicos, sino también de interacciones más complejas con su entorno y de las regiones sin dominios. Las otras 4 variantes (de las cuales 3 son *missense*) se encuentran en la zona de repetición de dominios *PKD*. ANNOVAR aportó la posición exacta del aminoácido en la cadena peptídica. En *Uniprot* se puede conocer la posición de cada dominio. Esto permite complementar la información dada por el diagrama de la proteína. De esta manera, si se analizan las variantes *missense 2114379: C:T, 2111077: G:A* y *2109429: G:A* se objetiva que **los dominios exactos** en los que se encuentran son *PKD 3, PKD 8* y *PKD 15* respectivamente. La proteína PC1 consta de varios dominios extracelulares entre los que se incluye una **repetición de 16 unidades *PKD***, aquí es dónde se encuentran las 3 variantes. Estos complejos están formados por repeticiones 80-90 aa. Se desconoce la función exacta de cada uno de estos dominios. En un inicio, debido a su plegamiento similar al de una inmunoglobulina, se propuso que pudiesen tener una función inmune, pero se descartó. Las repeticiones *PKD* adquieren un plegamiento de *sándwich de lámina ß* y son todas muy similares entre sí. En total, estas 16 unidades representan el 30% de la proteína PC1, por ende, se piensa que desempeñan un papel importante (62). Investigaciones más recientes experimentaron en laboratorio con las repeticiones *PKD* hallando a través de análisis de microscopía atómica que estas estructura tienen una **elevada resistencia mecánica**. Lo más relevante de esta investigación, y aplicable a este trabajo, es que se observó que las mutaciones *missense* eran capaces de alterar esta resistencia mecánica. Teniendo en cuenta que PC1 participa en la adhesión celular es posible que las repeticiones de *PKD* sean un elemento clave en esta función proteica (35). En síntesis, las 3 mutaciones *missense* encontradas en las repeticiones *PKD* es muy posible que afecten en la resistencia mecánica que aportan estos dominios en PC1.

La **predicción del efecto de las 4 mutaciones *missense*** se realizó a través de dos programas, uno se trata de una IA (AlphaMissense) y otro no (Polyphen-2). Los dos modelos proporcionan una puntuación de patogenicidad para cada variante, pero no funcionan de igual manera. **Polyphen-2** es una herramienta bioinformática que usa modelos estructurales y datos evolutivos para predecir el impacto funcional propiamente dicho. Desde su lanzamiento original en 2010 se ha ido mejorando y perfeccionando para ser cada vez más preciso. Entre las técnicas que utiliza se encuentra el *machine learning,* pero es importante señalar que no se trata de una IA. Hoy en día es considerada una de las herramientas de referencia para la predicción del efecto de las mutaciones. En contraste, **AlphaMissense**, además de *machine learning*, utiliza **técnicas de *deep learning*.** Mediante el uso de complejas redes neuronales profundas, puede predecir la estructura tridimensional de las proteínas y anticipar cómo los cambios en aminoácidos pueden afectar a su estructura plegada. De esta forma se estima la implicación funcional de la mutación.
En cuanto a los resultados aportados por cada una de las herramientas se observan grandes discrepancias en muchas mutaciones. Polyphen-2 clasifica la **primera variante (2103403)** como benigna, con una probabilidad de malignidad de 0,007 y una sensibilidad realmente alta (0,96). AlphaMissense también considera que es muy poco probable que sea dañina (0,0684). En esta primera mutación ambas herramientas están de acuerdo, por lo que parece improbable que pueda estar implicada en la enfermedad. Para la **segunda variante (2109429)** Polyphen-2 la clasifica como posiblemente dañina, con una probabilidad de 0,925 y una sensibilidad y especificidad muy altas (0.81 y 0.94 respectivamente). Por el contrario, AlphaMissense aporta una probabilidad de 0,1238, lo cual equivale a mutación posiblemente benigna. Para la **tercera variante (2111077)** Polyphen-2 establece una probabilidad de malignidad en 0,998, se aprecia que la **probabilidad de malignidad es de prácticamente el 100%** según la herramienta. En este caso la sensibilidad no es tan alta (0.27) pero la especificidad sí (0.99). AlphaMissense sugiere que esta variante es posiblemente benigna (probabilidad de 0.1289). La **cuarta variante (2114379)** es **probablemente maligna** según Polyphen-2 (0.994) con elevada especificidad (0.97), AlphaMissense indica que es posiblemente benigna (0.1147).
Salvo por la primera variante, los resultados son dispares entre ambas herramientas. **Polyphen-2 únicamente considera benigna una variante**, indica que otra es posiblemente maligna y que las dos últimas son probablemente malignas. Si analizamos los números, llama la atención como la tercera variante (2111077) y la cuarta (2114379) tienen unas probabilidades de malignidad extremadamente altas, de un 99,8% y 99,4% respectivamente. Mientras que estos mismos datos aportados por AlphaMissense son 12,38% y 11,47%.
Es necesario tener en cuenta que AlphaMissense es una herramienta muy reciente y tiene limitaciones. Se lanzó en septiembre de 2023 y, en el momento de la elaboración de este trabajo, apenas hay publicaciones que la hayan usado. Además, no existen estudios que la comparen con herramientas tradicionales como Polyphen-2. Sus creadores defienden la exactitud de AlphaMissense en que se basa en AlphaFold, cuya precisión está demostrada por numerosos estudios independientes y se lanzó 2 años antes. También es necesario considerar una de las mayores limitaciones que tiene AlphaFold, su capacidad para trabajar con proteínas muy grandes. Este es el caso de PC1, que cuenta con más de 4000 aa. Es preciso entender cómo funciona AlphaFold para poder comprender el porqué de la inexactitud de AlphaMissense. Cuando predice el plegamiento de una proteína la IA realiza complejas operaciones que combina con plegamientos conocidos de proteínas similares o de la propia proteína. En el caso de PC1 gran parte de su estructura tridimensional es desconocida. Además, no todas las regiones de la proteína se predicen con la misma confianza. Es posible que las mutaciones introducidas en AlphaMissense se ubiquen en zonas dónde la predicción del plegamiento de AlphaFold es poco fiable.

Aunque Polyphen-2 no use IA, es una herramienta compleja y validada. Es interesante que aporta también la sensibilidad y especificidad. En el caso de las mutaciones de este trabajo las predicciones de Polyphen-2 semejan altamente fiables. Aquellas mutaciones que clasifica como patológicas lo hace con una alta probabilidad (siempre superior al 92%) y con especificidades superiores al 94%.
En este contexto, **parecen más fiables los resultados de Polyphen-2 que los de AlphaMissense** y podemos asumir que 3 de las 4 variantes *missense* es altamente probable que sean patogénicas (>92% de probabilidad de malignidad).
En cuanto a la **variante hallada en la zona de *splicing*** se usó otra IA, **SpliceAI**, que considera claramente la mutación como patogénica, con una probabilidad de 0,93 para pérdida de donante a una distancia de -2 pb. Sobre esta IA es necesario comentar que existe desde hace más años que AlphaMissense, está respaldada por más evidencia y sus resultados son más fiables.

Este trabajo contó con **diversas limitaciones.** Las propias características del gen y de su proteína dificultan su estudio. El gran tamaño de ambos limita el uso y efectividad de las herramientas bioinformáticas como AlphaMissense. El desconocimiento de la estructura terciaria de PC1, de la función de sus dominios e incluso de la función de la proteína evidencian la necesidad de un estudio exhaustivo de la misma. Debido a su alta expresión en el SNC y a que se ha descrito en enfermedades neurológicas (epilepsia) resultaría **relevante un modelo de validación funcional** en este tejido. En muchas bases de datos las mutaciones tenían una frecuencia 0. Esto se debe a que son muy poco comunes, pero también a que el número de genomas o exomas que conforman las bases de datos es menor al ideal. Esto evidencia la necesidad de contar con bases de datos extensas con información de muchos individuos de diferentes regiones del mundo. La NGS y las nuevas tecnologías han disminuido los costes de secuenciación por lo que es esperable que en los próximos años surjan bases de datos cada vez mayores.
En lo respectivo a los TEA, **conocer al detalle la fisiopatología subyacente** de este trastorno sería de gran utilidad. Se deben realizar investigaciones acerca de su desarrollo e, idealmente, se deberían conocer todos los genes implicados. De esta manera se podrán mejorar y unificar las herramientas de diagnóstico. También sería posible desarrollar tratamientos y terapias específicas para este trastorno con una prevalencia tan elevada.
En este escenario, este trabajo de investigación ha intentado descubrir a la comunidad científica un **nuevo gen de riesgo para los TEA, *PKD1***. Estudios posteriores deben profundizar en la comprensión de este gen en los TEA, idealmente, en un modelo de validación funcional y no *in silico*.

6. CONCLUSIONES

El objetivo principal consistía en la descripción y estudio exhaustivo de las 10 mutaciones halladas en *PKD1*. A lo largo de este trabajo se ha profundizado y detallado cada mutación. Los hallazgos más relevantes son los siguientes:

- **Conclusión 1:** el estudio fenotípico y genotípico de los probandos reveló que las variantes en *PKD1* son las mejores candidatas para explicar los TEA. Todas son SNVs y DNMs.
- **Conclusión 2:** las ubicación de las variantes a nivel molecular es la siguiente: Una se encuentra en un intrón (*chr16:2097138:G:A*); otra en una zona de *splicing* o empalme (*chr16:2113159:A:C*); las 8 restantes en exones (*chr16:2102576:C:T*, *chr16:2103403:C:T, chr16:2104547:C:T, chr16:2104625:G:A, chr16:2109429:G:A, chr16:2110511:G:A, chr16:2111077:G:A, chr16:2114379:C:T*).
- **Conclusión 3:** la anotación con ANNOVAR clasificó las mutaciones exónicas en sinónimas *(2110511V: G:A, 2104625: G:A, 2104547: C:T, 2102576: C:T*) y no sinónimas o *missense* (*2114379: C:T, 2111077: G:A, 2109429: G:A, 2103403: C:T*). Es posible que las 4 mutaciones sinónimas no tengan una implicación en los TEA. Se descartó que la variante intrónica estuviera implicada en mecanismos de regulación del ADN. Es esperable que sea benigna.
- **Conclusión 3:** la comparación de la ubicación de las variantes exónicas con los exones implicados en la ADPKD reveló grandes similitudes. Varios de los exones que contienen las variantes de este trabajo (11, 15 y 23) se encuentran entre los que más mutaciones acumulan en la ADPKD.
- **Conclusión 4:** mediante el estudio de prevalencia y la anotación con ANNOVAR se objetivó que la frecuencia poblacional de todas las variantes es realmente baja.
- **Conclusión 5:** se objetivó una elevada expresión de *PKD1* en aquellos tejidos más relevantes en los TEA, es decir, cerebelo y corteza cerebral (25.88 y 5.44 veces más que en el tejido renal respectivamente).
- **Conclusión 6:** los dominios proteicos donde se hallan 3 mutaciones *missense* (repeticiones PKD) juegan un papel importante en la resistencia mecánica de la proteína.
- **Conclusión 7:** las herramientas bioinformáticas que estudiaron el efecto funcional de las mutaciones aportaron lo siguiente:
    - Polyphen-2 clasificó como **benigna una variante** (*2103403: C:T*), como **posiblemente maligna una** (*2109429: G:A*) y como **probablemente malignas dos** (*2114379: C:T, 2111077: G:A*). Con unas probabilidades de patogenicidad respectivas de 0.007, 0.925, 0.998 y 0.994. La especificidad fue alta.
    - AlphaMissense clasificó todas las mutaciones como posiblemente benignas. La herramienta es muy reciente y apenas existe experiencia usándola, es necesario tener en cuenta la probable inexactitud de esta IA.
    - La IA SpliceAI reveló que la variante *chr16-2113159-A-C* es muy **probablemente patogénica** con una probabilidad de 0.93.
- **Conclusión 8:** Debido a lo anteriormente expuesto y pese a que son necesarios más estudios, se sugiere que *PKD1* podría ser un gen de riesgo en los TEA.

7. BIBLIOGRAFÍA

1. Real Academia Española. Diccionario de la lengua española [Internet]. 23.ª ed. versión 23.7 en línea. Madrid: Real Academia Española; [consultado el 3 de enero de 2024]. Disponible en: https://dle.rae.es
2. Evans B. How autism became autism. Hist Hum Sci. julio de 2013;26(3):3-31.
3. Lai MC, Lombardo MV, Baron-Cohen S. Autism. The Lancet. 2014;383(9920):896-910.
4. American Psychiatric Association, American Psychiatric Association, editores. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Washington, D.C: American Psychiatric Association; 2013. 947 p.
5. Sucksmith E, Roth I, Hoekstra RA. Autistic traits below the clinical threshold: re-examining the broader autism phenotype in the 21st century. Neuropsychol Rev. 2011;21(4):360-89.
6. Hossain MM, Khan N, Sultana A, Ma P, McKyer ELJ, Ahmed HU, et al. Prevalence of comorbid psychiatric disorders among people with autism spectrum disorder: An umbrella review of systematic reviews and meta-analyses. Psychiatry Res. 2020;287:112922.
7. Chiurazzi P, Kiani AK, Miertus J, Paolacci S, Barati S, Manara E, et al. Genetic analysis of intellectual disability and autism. Acta Bio-Medica Atenei Parm. 2020;91(13-S):e2020003.
8. Al-Beltagi M. Autism medical comorbidities. World J Clin Pediatr. 2021;10(3):15-28.
9. Rotschafer SE. Auditory Discrimination in Autism Spectrum Disorder. Front Neurosci. 2021;15:651209.
10. Marco EJ, Hinkley LBN, Hill SS, Nagarajan SS. Sensory Processing in Autism: A Review of Neurophysiologic Findings. Pediatr Res. 2011;69(5 Pt 2):48R-54R.
11. Singhi P, Malhi P. Early Diagnosis of Autism Spectrum Disorder: What the Pediatricians Should Know. Indian J Pediatr. 2023;90(4):364-8.
12. Reviriego E, Bayón JC, Gutiérrez A, Galnares-Cordero L. Trastornos del Espectro Autista: evidencia científica sobre la detección, el diagnóstico y el tratamiento [Internet]. GuíaSalud; 2022 [citado 14 de enero de 2024]. Disponible en: https://portal.guiasalud.es/opbe/tea-deteccion-diagnostico-tratamiento/
13. Genovese A, Butler MG. The Autism Spectrum: Behavioral, Psychiatric and Genetic Associations. Genes. 2023;14(3):677.
14. Bai D, Yip BHK, Windham GC, Sourander A, Francis R, Yoffe R, et al. Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort. JAMA Psychiatry. 2019;76(10):1035-43.
15. Diaz-Anzaldua A, Díaz-Martínez A. Genetic, environmental, and epigenetic contribution to the susceptibility to autism spectrum disorders. Rev Neurol. 2013;57:556-68.
16. Modabbernia A, Velthorst E, Reichenberg A. Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. Mol Autism. 2017;8:13.
17. Moraes F, Góes A. A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. Biochem Mol Biol Educ. 2016;44(3):215-23.
18. Fagundes NJR, Bisso-Machado R, Figueiredo PICC, Varal M, Zani ALS. What We Talk About When We Talk About "Junk DNA". Genome Biol Evol. 2022;14(5):evac055.
19. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012;489(7414):57-74.
20. Charlotte A. Spencer , Michael R. Cummings y William S. Klug. CONCEPTOS DE GENÉTICA. 8.ª ed. Pearson; 2006.

21. Manoli DS, State MW. Autism Spectrum Disorder Genetics and the Search for Pathological Mechanisms. Am J Psychiatry. 2021;178(1):30-8.
22. Butler MG, Rafi SK, Manzardo AM. High-resolution chromosome ideogram representation of currently recognized genes for autism spectrum disorders. Int J Mol Sci. 2015;16(3):6464-95.
23. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. Am J Hum Genet. 2012;90(1):7-24.
24. Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, Grove J, Ripke S, et al. Identification of common genetic risk variants for autism spectrum disorder. 2019;51(3):431-44.
25. Huguet G, Benabou M, Bourgeron T. The Genetics of Autism Spectrum Disorders. 2016. p. 101-29.
26. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-21.
27. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, et al. Structural Variation of Chromosomes in Autism Spectrum Disorder. Am J Hum Genet. 2008;82(2):477-88.
28. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017;18:14.
29. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-33.
30. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020;180(3):568-584.e23.
31. Hughes J, Ward CJ, Peral B, Aspinwall R, Clark K, San Millán JL, et al. The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. Nat Genet. 1995;10(2):151-60.
32. Pletnev V, Huether R, Habegger L, Schultz W, Duax W. Rational proteomics of PKD1. I. Modeling the three dimensional structure and ligand specificity of the C_lectin binding domain of Polycystin-1. J Mol Model. 2007;13(8):891-6.
33. Cordido A, Besada-Cerecedo L, García-González MA. The Genetic and Cellular Basis of Autosomal Dominant Polycystic Kidney Disease—A Primer for Clinicians. Front Pediatr. 2017;5:279.
34. MacKay CE, Floen M, Leo MD, Hasan R, Garrud TA, Fernández-Peña C, et al. A plasma membrane-localized polycystin-1/polycystin-2 complex in endothelial cells elicits vasodilation. eLife. 2022;11:e74765.
35. Maser RL, Calvet JP, Parnell SC. The GPCR properties of polycystin-1- A new paradigm. Front Mol Biosci [Internet]. 4 de noviembre de 2022 [citado 8 de mayo de 2024];9. Disponible en: https://www.frontiersin.org/articles/10.3389/fmolb.2022.1035507
36. Loscalzo J, Fauci AS, Kasper DL, Hauser S, Longo D, Jameson JL, editores. Harrison. Principios de Medicina Interna. 21.ª ed. New York: McGraw-Hill Education; 2022.
37. Kim DY, Park JH. Genetic Mechanisms of ADPKD. Adv Exp Med Biol. 2016;933:13-22.
38. Chapman AB, Devuyst O, Eckardt KU, Gansevoort RT, Harris T, Horie S, et al. Autosomal Dominant Polycystic Kidney Disease (ADPKD): Executive Summary from a

Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2015;88(1):17-27.
39. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(Database issue):D877-87.
40. Wang JY, Wang J, Lu XG, Song W, Luo S, Zou DF, et al. Recessive PKD1 Mutations Are Associated With Febrile Seizures and Epilepsy With Antecedent Febrile Seizures and the Genotype-Phenotype Correlation. Front Mol Neurosci. 2022;15:861159.
41. Chen J, Coppola G. Chapter 7 - Bioinformatics and genomic databases. En: Geschwind DH, Paulson HL, Klein C, editores. Handbook of Clinical Neurology [Internet]. Elsevier; 2018 [citado 16 de abril de 2024]. p. 75-92. (Neurogenetics, Part I; vol. 147). Disponible en: https://www.sciencedirect.com/science/article/pii/B9780444632333000075
42. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
43. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. 2022;51(D1):D1188-95.
44. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022;43(8):1012-30.
45. Alonso-Gonzalez A, Calaza M, Amigo J, González-Peñas J, Martínez-Regueiro R, Fernández-Prieto M, et al. Exploring the biological role of postzygotic and germinal de novo mutations in ASD. Sci Rep. 2021;11(1):319.
46. Aitana Alonso González. Bases genéticas de los trastornos del espectro autista: estudio de la variación común y rara. [Santiago de Compostela]: Universidade de Santiago de Compostela; 2020.
47. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics. 2017;109(2):83-90.
48. ADPKD Variant Database [Internet]. [citado 23 de febrero de 2024]. Disponible en: https://pkdb.mayo.edu/variants
49. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
50. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20.
51. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.
52. Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. PLOS ONE. 2023;18(3):e0282689.
53. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
54. de Sainte Agathe JM, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Hum Genomics. 2023;17:7.
55. Protein Isoform - an overview | ScienceDirect Topics [Internet]. [citado 23 de abril de 2024]. Disponible en: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/protein-isoform

56. Oerlemans AM, Hartman CA, Franke B, Buitelaar JK, Rommelse NNJ. Does the cognitive architecture of simplex and multiplex ASD families differ? J Autism Dev Disord. 2016;46:489-501.
57. Vignoli A, La Briola F, Peron A, Turner K, Vannicola C, Saccani M, et al. Autism spectrum disorder in tuberous sclerosis complex: searching for risk markers. Orphanet J Rare Dis. 2015;10:154.
58. Donovan APA, Basson MA. The neuroanatomy of autism – a developmental perspective. J Anat. 2017;230(1):4-15.
59. Courchesne E, Yeung-Courchesne R, Press GA, Hesselink JR, Jernigan TL. Hypoplasia of cerebellar vermal lobules VI and VII in autism. N Engl J Med. 1988;318(21):1349-54.
60. Clothier J, Absoud M. Autism spectrum disorder and kidney disease. Pediatr Nephrol Berl Ger. 2021;36(10):2987-95.
61. Oelschlaeger P. Molecular Mechanisms and the Significance of Synonymous Mutations. Biomolecules 2024;14(1):132.
62. Bycroft M, Bateman A, Clarke J, Hamill SJ, Sandford R, Thomas RL, et al. The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. EMBO J. 1999;18(2):297-305.