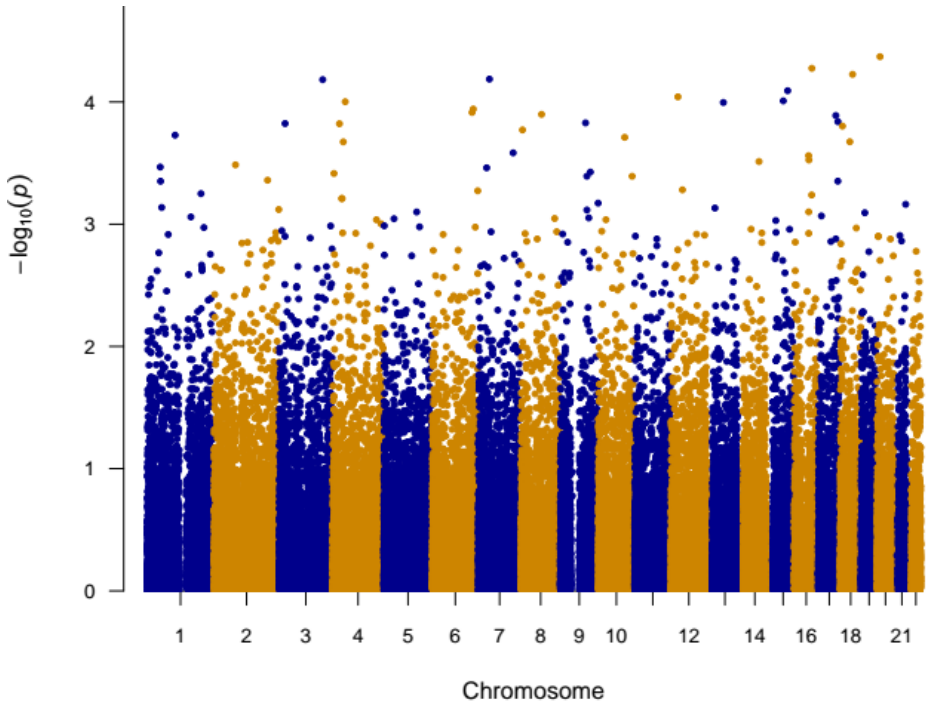
IDENTIFICACIÓN DE VARIANTES GÉNICAS RELACIONADAS CON LA EDAD DE INICIO DE LA ESQUIZOFRENIA

Carla Montagud Almarcha

TRABAJO FINAL DE GRADO DE BIOTECNOLOGÍA



Tutor académico: Gerard Pujadas Anguiano, Titular de Universidad, Departamento de Bioquímica y Biotecnología, gerard.pujadas@urv.cat

En cooperación con: Institut Pere Mata, Institut d'Investigació Sanitària Pere Virgili

Supervisor: Gerard Muntané Medina, Investigador Área de Recerca en Neurociència Aplicada, muntaneg@peremata.com, gerard.muntane@urv.cat

10 Junio 2020

Yo, Carla Montagud Almarcha, con DNI 20087263Y, soy conocedora de la guía de prevención del plagio en la URV *Prevenció, detecció i tratament del plagi en la docència: guía per a estudiants* (aprobada en julio de 2017) (http://www.urv.cat/ca/vida-campus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) y afirmo que este TFG no constituye ninguna de las conductas consideradas como plagio por la URV.

Tarragona, 10 de julio de 2020.

Carla Montagud Almarcha

ÍNDICE

DATOS DEL CENTRO

Este trabajo se basa en las tareas realizadas en mi estancia de prácticas curriculares, que se han llevado a cabo en el grupo de investigación Neurociència Aplicada. Este grupo pertenece al Institut Pere Mata (IPM), que a su vez presenta un convenio con el Institut d'Investigació Sanitària Pere Virgili (IISPV) y la Universitat Rovira i Virgili (URV). Además, entra dentro del área de Neurociencias y Salud Mental del IISPV y se centra en la investigación clínica y básica sobre factores genéticos y ambientales involucrados en las enfermedades psiquiátricas(*Institut d'Investigació Sanitària Pere Virgili - IISPV*, n.d.; Mata et al., 2013). Más concretamente, este grupo presenta cuatro líneas de investigación principales:

- El gen DDR1 y la melanina en la psicosis.
- Herencia materna y genoma mitocondrial en la psicosis.
- Epidemiología de la psicosis funcional.
- Psicosis confusional o *delirium*.

En mi caso, junto con mi tutor profesional, nos hemos centrado en identificar variantes génicas relacionadas con la edad de inicio de la esquizofrenia. Este estudio se ha llevado a cabo gracias a dos consorcios que presenta el grupo de investigación con otras instituciones.

En primer lugar, el Centro de Investigación Biomédica de Red de Salud Mental (CIBERSAM) nos ha cedido una muestra de 3.888 individuos genotipados, habiendo entre ellos 1.867 pacientes de esquizofrenia. El CIBERSAM tiene como objetivo llevar a cabo investigaciones centradas en la prevención de los trastornos mentales, en mejorar la calidad de vida de las personas implicadas y los tratamientos para estas enfermedades, entre las que destacan la esquizofrenia, el trastorno bipolar y la depresión. (*Cibersam*, n.d.) Por otro lado, este centro está asociado al Psychiatric Genomics Consortium (PGC), un consorcio que busca la colaboración mundial para aportar conocimiento desde el punto de vista biológico, clínico y terapéutico de once trastornos psiquiátricos. (*What Is the PGC?* | *Psychiatric Genomics Consortium*, n.d.)

En segundo lugar, el grupo alemán Klinische Forschergruppe 241 (KFO) se asoció al grupo de Neurociència Aplicada y aportó muestras genotipadas de 1.448 individuos que padecen esquizofrenia. El KFO busca comprender las bases biológicas del desarrollo de la psicosis, investigando el genoma de individuos que sufren esquizofrenia o trastorno bipolar(*Clinical Research Group 241*, n.d.).

RESUMEN

Este trabajo recoge los resultados obtenidos durante mi estancia de prácticas en el grupo de investigación Neurociència Aplicada del Institut d'Investigació Sanitaria Pere Virgili (IISPV). Esta se centra en identificar variantes génicas relacionadas con la edad de inicio de la esquizofrenia. Para ello es necesario llevar a cabo estudios de asociación del genoma completo (GWAS), que se basan en encontrar polimorfismos de un único nucleótido (SNPs) que estén asociados a la aparición de un fenotipo en concreto, en nuestro caso la edad a la que aparece el primer episodio esquizofrénico. Se buscan estas variantes ya que se espera que la genética pueda jugar un papel importante en la edad de inicio de esta enfermedad, además, se han identificado diversos SNPs relacionados con la esquizofrenia llevando a cabo GWAS. Nos interesamos en la edad de inicio debido a que la mayoría de las personas que padecen esta enfermedad la desarrollan entre el final de la adolescencia y el principio de la juventud, pero se han visto casos de individuos que han sufrido el primer brote psicótico al principio de la adolescencia o incluso al principio de la madurez. Los resultados de este estudio muestran como significativos una serie de variantes génicas (rs6847217, rs9997352 y rs4371589) que se encuentran en el cromosoma 4. Aunque estos SNPs no tengan ningún fenotipo asociado hoy en día, se sabe que se localizan en el intrón del gen CORIN y que este gen se traduce a una serin-proteasa, una proteína transmembrana de tipo II que puede funcionar como un péptido natriurético cerebral de tipo convertasa.

PALABRAS CLAVE: esquizofrenia, SNP, GWAS, edad de inicio, asociación, psicosis.

INTRODUCCIÓN

1. ESQUIZOFRENIA

A pesar de los diversos modelos diseñados para intentar explicar las causas de la aparición de la esquizofrenia, se sabe que es una enfermedad cerebral hereditaria, crónica, grave e incapacitante del desarrollo neurológico, con antecedentes genéticos y neurobiológicos heterogéneos.(Perkovic et al., 2017) Se trata de un desorden psiquiátrico complejo, multifactorial y poligénico que afecta al 1% de la población mundial, además de ser una de las principales causas de discapacidad a nivel mundial. Sin embargo, este síndrome no afecta a todas las personas que lo poseen por igual, por lo que cada paciente puede presentar un diferente grado de capacidad a la hora de desenvolverse en las diversas funciones de una vida cotidiana (Marder & Cannon, 2019).

Aun así, esta enfermedad provoca una disminución de unos 15 o 20 años en la esperanza de vida media en comparación con el resto de la sociedad. Se cree que esto se debe a un estilo de vida poco saludable y al aumento de la tasa de suicidio, que es 12 veces mayor en personas que sufren esquizofrenia. (Valton et al., 2017) Además, se ha comprobado que esta enfermedad genera una carga económica, tanto para la sociedad como para los familiares. A parte de los costes de tratamientos, también influyen factores indirectos como la pérdida de empleo, que a su vez está influenciado en gran parte por los déficits cognitivos y por la sintomatología negativa que presentan las personas que padecen esquizofrenia.

1.1. SINTOMATOLOGÍA

Los síntomas de esta enfermedad se clasifican en diversas categorías: positivos, negativos, cognitivos y del estado de ánimo. Los síntomas positivos, como las alucinaciones o los delirios, hacen referencia a aquellos que no aportan deficiencias ni distanciamiento social. Las alucinaciones engloban diversas percepciones vívidas de estímulos complejos en ausencia de un estímulo externo, como oír voces o ver personas u objetos que no se encuentran presentes. En cambio, los delirios abarcan falsas creencias que no se ajustan a las normas socioculturales del individuo y, además, persisten a pesar de ser contradichas por la realidad o por pruebas racionales. Por otro lado, los síntomas negativos se componen de todos aquellos que están involucrados en el distanciamiento social, como la apatía, la pérdida de interés, la ausencia de afectividad, la disminución de la motivación y de la expresividad, etc. Sin embargo, los déficits cognitivos abarcan la disminución del rendimiento de la memoria y la velocidad

de procesamiento mental, el déficit de atención y de racionamiento y la pobreza y desorganización del habla, esto suele estar asociado a una disminución del coeficiente intelectual de unos 10 puntos después de la aparición de la enfermedad. (Valton et al., 2017) Finalmente, los síntomas de estados de ánimo comprenden los cambios a la hora de expresar depresión, alegría o tristeza. (Perkovic et al., 2017)

1.2. DIAGNÓSTICO

Los primeros indicios de esquizofrenia suelen aparecer entre finales de la adolescencia y principios de la veintena. Aun así, existen casos del desarrollo de la esquizofrenia al principio de la adolescencia o incluso al final de la juventud o al inicio de la edad adulta. Antes del inicio de la psicosis, los afectados suelen presentar un periodo llamado pródromo de la psicosis. Este suele durar meses o incluso años y se caracteriza por cambios sutiles en el comportamiento y un declive en las funciones psicomotrices. (Marder & Cannon, 2019)

Para poder diagnosticar a un paciente con esquizofrenia, este tiene que cumplir una serie de criterios establecidos en el “Diagnostic and Statistical Manual for Mental Disorders” (DSM) (*Anexo I*) (American Psychiatric Association, 2013), además de descartar otros estados psicóticos similares. (Marder & Cannon, 2019) Por otro lado, existe otro método de diagnóstico basado en los criterios de la “International Classification of Disorders” (ICD) de la Organización Mundial de la Salud (OMS). Ambos métodos presentan enfoques comunes, como la omisión de los subtipos clínicos de la esquizofrenia (paranoica o desorganizada), debido a que no tienen relevancia a la hora del diagnóstico o el tratamiento. Sin embargo, a la hora de diagnosticar, el DSM exige la presencia de deficiencias funcionales, mientras que para el ICD no es obligatorio que las haya. (Gaebel & Zielasek, 2015)

1.3. CARACTERES GENÉTICOS Y FENOTÍPICOS

Al igual que en otros trastornos mentales complejos, se ha visto que el desarrollo de la esquizofrenia se asocia con interacciones complejas entre diversos factores de riesgo genéticos, factores ambientales y la exposición a experiencias traumáticas tempranas, que pueden provocar cambios importantes e irreversibles en el desarrollo neurológico del sistema nervioso central. (Perkovic et al., 2017)

Por un lado, se ha observado que factores ambientales, como complicaciones durante el parto, adversidades durante la infancia o residir en zonas urbanas durante la misma, pueden afectar en un 20% al riesgo de sufrir esquizofrenia. (Marder & Cannon, 2019) Otros estudios afirman que los pacientes desarrollan esta enfermedad debido a que no

son capaces de hacer frente a situaciones de estrés (como el abuso físico, el abuso sexual, el maltrato o la intimidación) y no logran resultados positivos cuando se enfrentan a las adversidades o a los factores estresantes. (Perkovic et al., 2017) Otro ejemplo sería el uso indebido de drogas como el cannabis, la anfetamina o la cocaína.

Por otro lado, estudios sobre la familia, los gemelos y la adopción estiman que un 80% del riesgo de padecer esquizofrenia se debe a factores genéticos hereditarios. En este caso, puede haber diversas variantes implicadas, desde polimorfismos de un solo nucleótido (SNP) comunes que tienen una implicación pequeña, hasta mutaciones más grandes que afectan en mayor medida, pero presentan una menor frecuencia de incidencia.

La mayoría de los genes implicados han sido identificados en estudios de asociación del genoma completo (GWAS) y en estudios de expresión génica, y se ha observado que están relacionados con el sistema inmunitario, el desarrollo del citoesqueleto, la plasticidad y la función sináptica. (Marder & Cannon, 2019)

A lo largo de los años se han realizado diversas hipótesis sobre genes implicados en el desarrollo de la esquizofrenia, como DISC1, DNTBP1, NRG1 y COMT. Pero, debido a la ausencia de descubrimientos significativos, su influencia sigue siendo objeto de debate. La ausencia de resultados significativos puede deberse a varias razones, como las dificultades a la hora de llevar a cabo replicaciones de los hallazgos positivos, un poder estadístico insuficiente o un conocimiento limitado de los hipotéticos genes que se piensa que están involucrados en la esquizofrenia. (Henriksen et al., 2017)

En cuanto a los caracteres fenotípicos, se han detectado diferencias neuroatómicas relacionadas con pacientes que sufren esquizofrenia. En primer lugar, se desarrolló la hipótesis de la dopamina al observar los resultados tras un tratamiento con fármacos antipsicóticos que bloquean los receptores de dopamina D2 (D2r). En estos pacientes se observó una elevada señalización dopaminérgica, una elevada síntesis y liberación de dopamina tanto en la fase prodrómica como en la psicótica y un aumento de densidad de los D2r estriatales.(Valton et al., 2017) Debido a que se asocia la dopamina con comportamientos que predicen una recompensa posterior, algunos estudios relacionan el aumento de este neurotransmisor con un aumento de la importancia de otros estímulos, anteriormente inofensivos. Esto podría explicar la paranoia y la disociación de la realidad características de la esquizofrenia.(Marder & Cannon, 2019) Por otro lado, la hipótesis del glutamato surgió debido a la aparición de psicosis en sujetos sanos cuando se exponían a drogas psicoactivas, como la ketamina y la fenciclidina. Estas drogas bloquean los sitios de unión del glutamato de los receptores del ácido N-metil-

D-aspártico (NMDAr) que participan en la regulación del potencial postsináptico, en la plasticidad neuronal, el aprendizaje y la memoria. Además, existe la hipótesis gabaérgica, esta se desarrolló después de observar una reducción del ácido γ-aminobutírico (GABA) cortical, una actividad disfuncional y una reducción de los marcadores de las interneuronas inhibidoras situadas en las zonas prefrontales de los pacientes. Finalmente, debido a diversas observaciones fenotípicas, como la disminución del volumen cortical, un pliegue cortical prefrontal anormal, los ventrículos agrandados y una conectividad sináptica anormal, se desarrolló la hipótesis de la desconexión. Este nombre se debe a la creencia de que estos fenotipos son debidos a una sincronía o desconexión reducida entre las diferentes áreas corticales, lo que conlleva un mayor esfuerzo a la hora de realizar tareas cognitivas. (Valton et al., 2017)

Por otro lado, en estudios realizados *post mortem* se ha observado una disminución en el volumen de materia gris en la región prefrontal y parahipocampal del cerebro de personas con esquizofrenia, además de una disminución del número de dendritas y espinas dendríticas. La pérdida de materia gris se ha asociado a elevados niveles del factor de necrosis tumoral α (TNF-α), que se encarga de activar las microglías cerebrales (células encargadas de la defensa del sistema nervioso central). Además, se ha asociado el aumento del TNF-α a la alteración de las comunicaciones entre la corteza prefrontal, la corteza temporal, el tálamo, el hipocampo y el cerebelo. (Marder & Cannon, 2019)

1.4. TRATAMIENTO

Por desgracia, hoy en día no existe un tratamiento definitivo para esta enfermedad, en mayor parte debido a la falta de comprensión de las causas y los procesos que se llevan a cabo en este desorden.

Lo único que se puede hacer es intentar tratar los síntomas positivos con medicamentos antipsicóticos y tratamientos psicosociales. Con esto se busca reducir lo máximo posible estos síntomas (alucinaciones o delirios que podrían conducir a producir daños), los riesgos para el paciente y su entorno y evitar la recaía de la psicosis. Valton *et. al* estiman que, de la totalidad de afectados que

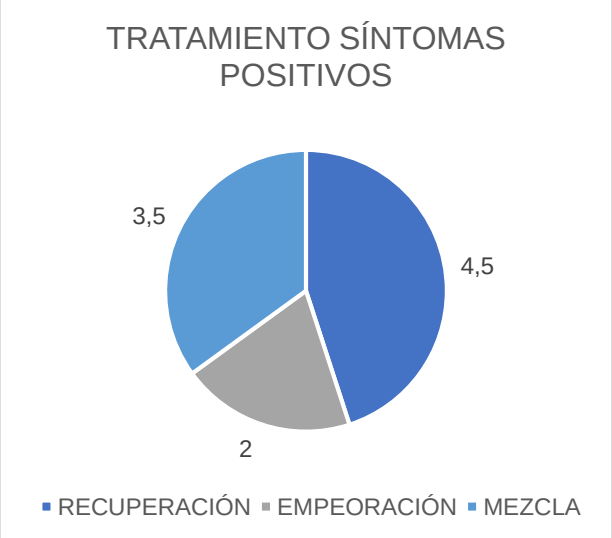


Figura 1. Representación poblacional de los resultados del tratamiento de la esquizofrenia con medicamentos antipsicóticos y tratamientos psicosociales.

siguen este tipo de tratamiento, un 45% se recupera después de un par de episodios, el 20% muestra un empeoramiento de los síntomas y un 35% muestran una mezcla de recuperación y empeoramiento (*Figura 1)*.

Se ha visto que los medicamentos antipsicóticos son eficaces a la hora de reducir los síntomas psicóticos en episodios agudos de esquizofrenia. Aunque no todos los medicamentos afectan de la misma manera a cada paciente, la selección de estos debe basarse en la respuesta del paciente a esta misma. Estos medicamentos suelen administrarse por vía oral, pero pueden administrarse por vía intramuscular a aquellos pacientes que estén agitados. Los medicamentos antipsicóticos, como la clozapina, la risperidona o la olanzapina, son eficaces cuando los niveles del sistema nervioso central son suficientes para ocupar alrededor del 70% de los receptores D2. Aunque se ha visto que un aumento de la cantidad de este medicamento no implica un aumento de la ocupación de los D2r, ya que los pacientes que presentan su primer episodio de esquizofrenia suelen responder a dosis más bajas. Aun así, son más vulnerables a los efectos secundarios de estos medicamentos, como el aumento de peso, los efectos secundarios sexuales y efectos extrapiramidales. (Marder & Cannon, 2019)

El medicamento antipsicótico más eficaz para aquellas personas con síntomas psicóticos persistentes que viven en comunidad y cuyos síntomas influyen en su calidad de vida es la clozapina. Aun así, este medicamento también presenta efectos secundarios, ya que aproximadamente el 1% de los pacientes que lo toman presentan una agranulocitosis potencialmente mortal. Por esto hay que controlar regularmente la cantidad de glóbulos blancos y neutrófilos. (Marder & Cannon, 2019) Para estos pacientes, es efectivo llevar a cabo la Terapia Cognitiva Conductual, en la que se intenta ayudar al paciente a considerar explicaciones alternativas a sus delirios.

2. GENOME-WIDE ASSOCIATION STUDY (GWAS)

A diferencia de estudios genéticos realizados anteriormente, que se basan en encontrar marcadores genéticos basándose en una hipótesis y normalmente no obtienen ningún resultado significativo, los GWAS analizan el genoma sin basarse en ningún gen candidato y buscan asociaciones entre las variantes comunes y el trastorno.

Según el National Institutes of Health, un GWAS se define como un estudio de la variación genética a través de todo el genoma humano, diseñado para identificar las asociaciones genéticas que presentan rasgos observables. (Pearson & Manolio, 2008) Se trata de una comparación entre dos grandes grupos de individuos, uno formado por personas sanas y otro formado por personas que padecen la enfermedad que se quiere estudiar. Estos estudios se pueden llevar a cabo gracias a los recientes avances

tecnológicos, como microarrays y chips, que permiten llevar a cabo un escaneo rápido y económico de más de un millón de SNPs de todo el genoma. (Henriksen et al., 2017)

Un GWAS se divide en cuatro partes (Pearson & Manolio, 2008):

1. Selección de una gran muestra de individuos que presenten la enfermedad o el rasgo de interés y un grupo de comparación adecuado.
2. Aislamiento del DNA, genotipado y análisis de los datos (control de calidad) para confirmar que se ha llevado a cabo una correcta genotipación.
3. Análisis estadísticos de las asociaciones entre las variantes génicas, que han superado los filtros del control de calidad, y el fenotipo.
4. Replicación de la asociación en una población independiente para comprobar que se ha realizado correctamente el estudio de asociación.

El control de calidad se lleva a cabo con la finalidad de descartar tanto variantes génicas como individuos que presenten algún error de genotipado. Es necesario llevar a cabo este paso debido a que las muestras iniciales suelen ser imperfectas, por lo que si no se lleva a cabo no se obtendrían resultados significativos. Existen diversas razones por las que los datos pueden presentar errores, como una mala calidad de las muestras de DNA, una mala hibridación del DNA con la matriz, un mal funcionamiento de las sondas de genotipado o que se contaminen las muestras. Un control de calidad estándar está compuesto por diversos pasos y filtros, que pueden combinarse de diferentes formas dependiendo del estudio que se quiera llevar a cabo (Marees et al., 2018). En este caso, las variantes génicas y los individuos se filtran en base a:

1. **Missingness**. Indica tanto la cantidad de SNPs que falta en cada individuo específico, como el número de individuos para los que falta información de una variante génica en específico. Un nivel elevado de Missingness se debe a una mala calidad del DNA o que han surgido problemas técnicos durante la selección y el genotipado del DNA.
2. **Sex discrepancies**. Se trata de inconsistencias entre el sexo asignado y el sexo genético de los sujetos. Estas discrepancias suelen deberse a confusiones de las muestras en el laboratorio. Solo se puede llevar a cabo este análisis cuando hay SNPs que pertenecen a cromosomas sexuales.
3. **Minor Allele Frecuency (MAF).** Indica la frecuencia del alelo menos frecuente en un lugar específico del genoma. Las variantes que presenten un MAF muy bajo deben ser excluidas, ya que no serían detectadas en el estudio de asociación.

4. **Hardy-Weinberg Equilibrium (HWE).** Se analiza la relación entre las frecuencias de los alelos y del genotipo. La desviación del HWE indica que las frecuencias del genotipo son significativamente diferentes a las frecuencias esperadas. Cuando esto ocurre suele deberse a un error de genotipado.
5. **Heterozygosity**. La heterocigosidad hace referencia al transporte de dos alelos diferentes de una variante génica específica. Una elevada proporción de genotipos heterocigotos dentro de un mismo individuo puede deberse a una mala calidad de la muestra, mientras que un bajo nivel de heterocigosidad puede indicar endogamia. En ambos casos, el individuo debería ser eliminado.
6. **Relatedness.** Indica el grado de relación genética entre un par de individuos, es decir, el nivel de parentesco. A no ser que se quiera llevar a cabo un estudio familiar, normalmente al realizar un GWAS se asume que los individuos no están emparentados, ya que esto podría conducir a estimaciones erróneas, por lo que solo se conservaría uno de los dos individuos emparentados.

Cuando se obtienen las muestras correctamente genotipadas se lleva a cabo un estudio de población, en el cual primero hay que llevar a cabo un análisis de componentes principales (PCA), donde se convierten los datos a analizar en variables no correlacionadas para así poder determinar que los individuos de nuestra muestra se corresponden con la población asignada. Este estudio se lleva a cabo debido a que las frecuencias de los alelos pueden ser diferentes dependiendo de la población a la que pertenezca el individuo. Si no se llevase a cabo este análisis, al presentar individuos que no corresponden a la población indicada, podríamos obtener asociaciones positivas falsas o incluso no encontrar asociaciones significativas.

Finalmente, se llevaría a cabo el estudio de asociación, que sirve para encontrar variantes relacionadas al fenotipo que se está estudiando, realizándose una regresión lineal y una corrección múltiple. Para ello hay que mirar el valor *P*, este indica la probabilidad de obtener una diferencia mayor a la observada y al mismo tiempo que no haya una diferencia real al resto de la muestra. Es decir, encontrar un SNP lo suficientemente diferente al resto de SNPs para ser significativo pero que no sea tan diferente que no tenga nada en común con ellos y con el fenotipo. Para determinar que una variante es significativa, debe presentar un valor *P* menor a 5e-08. Este valor se obtiene al aplicar la corrección de Bonferroni, esta indica que, para reducir la tasa de falsos positivos, se debe dividir el valor *P* convencional por el número de muestras. Por lo que el valor *P* significativo para un análisis de 1 millón de variantes génicas sería inferior a 5e-08 (0.05/1e06). (Pearson & Manolio, 2008)

En el ámbito de la esquizofrenia, los GWAS han identificado diversas variantes génicas relacionadas con esta enfermedad, pero que presentan un efecto individual muy pequeño. Se identificaron exitosamente 128 asociaciones, las cuales abarcaban 108 loci importantes dentro del genoma. (Henriksen et al., 2017) Se encontraron diversos genes, como el ZNF804A, NRGN, RELN, TCF4, el receptor de dopamina D2 y varios genes que intervienen en la neurotransmisión, la plasticidad sináptica y en funciones inmunitarias centrales. (Wang et al., 2011)

3. PUBLICACIONES ANTERIORES

Como he mencionado antes, normalmente, el primer episodio de psicosis ocurre al final de la adolescencia o al principio de la juventud, pero se han visto casos en los que este episodio ocurre al principio de la adolescencia o en la edad adulta. Se han llevado a cabo algunos GWAS que buscan una variante génica relacionada con este fenotipo y así poder encontrar una explicación para las diferentes edades de inicio de esta enfermedad. De los pocos estudios que se han realizado, en ninguno de ellos se ha encontrado algún SNP significativo.

Wang et al. llevaron a cabo un GWAS con 1.172 pacientes de esquizofrenia y 1.379 controles, todos ellos pertenecientes a una muestra europea-americana, y se analizaron 729,454 SNPs genotipados con el Affimetrix Genome-wide human SNP Array 6.0. En este estudio los SNPs fueron filtrados según el MAF (<1%) y el HWE (<0.00001), los individuos fueron sometidos a un estudio de población y finalmente se llevó a cabo la asociación. En ella se obtuvieron 104 SNPs que presentaron un valor P < e-04. De entre ellos, con un valor P = 3.1e-07, el rs7819815 fue el SNP más significativo, pero no llegó a presentar un valor P suficiente para considerarse realmente significativo. Esta variante génica se encontraba cerca del gen ZFAT que codifica una proteína que probablemente se une al DNA y funciona como un regulador transcripcional involucrado en la apoptosis y la supervivencia celular. Además, este gen está relacionado con un mayor riesgo de sufrir tiroiditis autoinmune, esclerosis múltiple e influye en la estatura de un adulto. (Wang et al., 2011)

Posteriormente, Woolston et al. llevaron a cabo un estudio similar, pero en este caso los individuos estaban emparentados. Para ello utilizaron una muestra de 1.207 pacientes de esquizofrenia y 1.035 controles, todos ellos participantes del Taiwan Schizophrenia Linkage Study (TSLS). En este estudio se analizaron 642.832 SNPs que fueron genotipados con el Affymetrix Axiom Genome-wide CHB 1 Array Plate. Estas variantes fueron filtradas según el MAF (<5%) y el HWE (<0.001), los individuos fueron sometidos a un estudio poblacional y posteriormente se llevó a cabo el estudio de asociación. En

este caso, se obtuvieron solo 17 SNPs que presentaron un valor *P* < e-04, siendo el rs6900852 el más significativo de todos ellos con un valor *P* = 1.0e-04. Este SNP se encuentra dentro del gen NCOA7 que está asociado con la enfermedad de Párkinson y su expresión es regulada en el cerebro. En este estudio se afirma que cada vez hay más pruebas que relacionan la esquizofrenia con el Párkinson, ya que tienen disfunciones dopaminérgicas opuestas, síntomas psicóticos similares y comparten genes implicados. (Woolston et al., 2017)

Finalmente, el estudio realizado por Bergen et al. fue el más similar al nuestro. En este se buscaron variantes génicas relacionadas a la edad de inicio, a la gravedad de la enfermedad, al sexo y a la historia familiar. Utilizaron la información de 2.762 pacientes de esquizofrenia y 3.187 controles, todos ellos pertenecientes al International Schizophrenia Consortium (ISC) y con una ascendencia europea. Los SNPs fueron genotipados usando los arrays Affymetrix 5.0 y 6.0, fueron filtrados por HWE y MAF (<1%) y posteriormente fueron imputados utilizando Beagle como software y HapMap2 como panel de referencia. La imputación se utiliza para aumentar el tamaño de la muestra de variantes génicas. Este proceso se basa en la comparación de tramos cortos de un genoma individual con tramos de genomas previamente caracterizados que se encuentran en el panel de referencia. Finalmente se llevó a cabo el estudio de asociación, en el que la variante más significativa fue el rs11999864, con un valor *P* de 1.52e-07. Este SNP abarca dos genes, el OR2K2 que sintetiza un receptor olfativo y el KIAA0368 que sintetiza una *scaffolding protein*. (Bergen et al., 2014)

Lo que diferencia este estudio de los realizados anteriormente es que:

- Tenemos muestras de dos consorcios diferentes (3.315 pacientes de esquizofrenia y 2.021 controles), pero todos los individuos presentan una ascendencia europea.
- Cada consorcio utiliza un array diferente, por lo que genotipan diferentes SNPs.
- Llevamos a cabo un control de calidad completo, variando entre los filtros para poder obtener los mejores resultados posibles.
- Llevamos a cabo dos imputaciones con dos paneles de referencia diferentes para poder aumentar el tamaño de las variantes génicas y así obtener resultados significativos.
- Tenemos un mayor número de variantes génicas de partida.

HIPÓTESIS DEL TRABAJO Y OBJETIVOS

La iniciativa de llevar a cabo este estudio parte de la necesidad de encontrar una explicación de por qué unas personas desarrollan la esquizofrenia a los 15 años, otras sufren el primer episodio esquizofrénico a los 23 años y otras a los 35 años (*Figura 2*). Como se ha descubierto que los factores genéticos contribuyen sustancialmente a la probabilidad de padecer esquizofrenia, con una heredabilidad estimada del 80%, se quiso llevar a cabo un estudio de los polimorfismos de un solo nucleótido (SNP). Se eligieron los SNPs debido a que anteriores estudios de asociación del genoma completo (GWAS) habían descubierto más de un centenar de estas variantes génicas implicadas en el aumento del riesgo de padecer esquizofrenia.

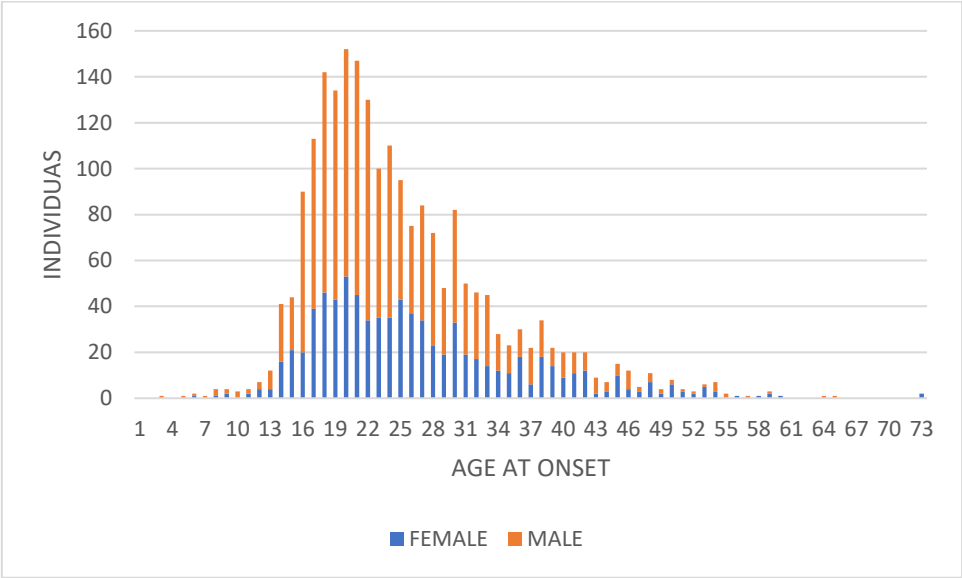


Figura 2. Agrupación de individuos, según el sexo, que presentan la misma edad de inicio de la esquizofrenia.

Con este estudio pretendemos encontrar alguna variante génica que esté relacionada de forma significativa (P < 5e-08) con la edad de inicio de la esquizofrenia. Para ello llevaremos un GWAS completo de las muestras del CIBERSAM, de las muestras del KFO y de una muestra que comprenda las dos anteriores. Además, llevaremos a cabo un proceso de imputación en cada caso, para amentar las probabilidades de obtener algún resultado significativo.

Finalmente, en el caso de que encontremos alguna variante génica significativa, intentaremos averiguar de qué forma está involucrada en la esquizofrenia y si presenta algún fenotipo asociado utilizando las bases de datos dbSNP y ENSEMBL y un software llamado FUMA que analiza los resultados del estudio de asociación.

METODOLOGÍA

1. DATOS INICIALES

Partimos de datos proporcionados por dos consorcios diferentes. CIBERSAM nos cede 603.132 SNPs procedentes de 3.888 individuos (1.867 pacientes de esquizofrenia y 2.021 controles, aunque estos no los utilizamos). Estas variantes han sido genotipadas con el Infinium PsychArray-24 BeadChip de Illumina, un microarray de alta densidad y bajo costo desarrollado en colaboración con el Psychiatric Genomics Consortium. Este array fue diseñado para utilizarse en estudios genéticos a gran escala centrados en la predisposición y el riesgo de enfermedades psiquiátricas como la esquizofrenia, el trastorno bipolar o los trastornos del espectro autista. (*Infinium PsychArray-24 Kit | Psychiatric Predisposition Microarray*, n.d.)

Además, KFO nos proporciona una muestra de 276.154 SNPs procedentes de 1.448 individuos que sufren esquizofrenia. Estas variantes fueron genotipadas con el Infinium OmniExpressExome-8 BeadChip de Illumina, un array asequible que proporciona un contenido genómico completo. Este microarray fue diseñado especialmente para llevar a cabo GWAS, ya que presenta la mayor cantidad de SNPs comunes seleccionados del Proyecto Internacional HapMap, para así descubrir nuevas asociaciones con rasgos y enfermedades.(*Infinium OmniExpressExome-8 Kit*, n.d.)

2. GENOME-WIDE ASSOCIATION STUDY

Este estudio ha sido realizado en su totalidad en un ambiente Linux, ya que es el sistema operativo en el cual funcionan los programas utilizados. Para llevar a cabo la mayoría de los análisis estadísticos de este estudio he utilizado un software llamado PLINK.(Chang et al., 2015; *PLINK 1.9*, n.d.) Además, las representaciones gráficas asociadas a los resultados estadísticos se han obtenido utilizando el paquete de recursos R.(R Core Team, 2020)

Utilizando diversos tutoriales(Harvard University, 2017; *Main - GPA: GWAS Analysis*, n.d.), se diseñaron una serie de scripts para llevar a cabo las etapas del GWAS, un ejemplo de estos se encuentra en el *Anexo II*.

2.1. CONTROL DE CALIDAD

Se pueden realizar los diversos pasos del control de calidad utilizando varias combinaciones de comandos del PLINK (*Tabla 1*).

Tabla 1. Comandos básicos utilizados en PLINK para llevar a cabo el control de calidad.

PASO	*COMANDO*	*FUNCIÓN*
MISSINGNESS	--missing	Examina la falta de genotipado de SNPs y de individuos

	--geno	Excluye SNPs que no se encuentran en la mayoría de los sujetos según el parámetro indicado
	--mind	Excluye los individuos que presentan un alto rango de pérdida de genotipado según el parámetro indicado
SEX DISCREPANCIES	--check-sex	Analiza la discrepancia de sexo
MAF	--freq	Calcula la frecuencia de MAF para realizar un histograma
	--maf	Incluye solo los SNPs que presenten un MAF superior al umbral asignado
HWE	--hardy	Calcula las distribuciones de HWE para hacer un histograma
	--hwe	Excluye SNPs que presenten un HWE desviado al asignado
HETEROZYGOSITY	--indep-pairwise	Calcula la tasa de heterocigosidad según los parámetros asignados
RELATEDNESS	--rel-cutoff	Excluye un miembro de cada par de muestras con una relación genómica mayor a la asignada

2.2. ESTUDIO DE POBLACIÓN

Posteriormente, para llevar a cabo el estudio de población, utilizamos los datos del "100 Genome Project". Seleccionamos las variantes de nuestras muestras que coincidan con las variantes del proyecto y fusionamos los archivos. A continuación hay que llevar a cabo un análisis de componentes principales mediante un software llamado EIGENSOFT(Patterson et al., 2006; Price et al., 2006). Luego, descargamos la información de población de los datos del proyecto, convertimos los códigos de población en códigos de superpoblación y añadimos nuestros datos etiquetados con CIBER o KFO dependiendo de su origen. Después, mediante R, obtenemos el gráfico que indique la ascendencia de nuestros individuos y seleccionamos aquellos que se alejen de la zona de población europea. Finalmente, con PLINK, eliminaremos los individuos seleccionados.

2.3. ESTUDIO DE ASOCIACIÓN

Cuando estamos seguros de que nuestra población presenta una ascendencia europea, llevamos a cabo el estudio de asociación. En primer lugar, se lleva a cabo un análisis de componentes principales de nuestra muestra y, utilizando PLINK, se cogen los 10 primeros componentes y se realiza una regresión lineal y una corrección múltiple utilizando como fenotipo la edad de inicio de la esquizofrenia de nuestros individuos.

La regresión lineal nos proporciona un fichero con las variantes génicas y el valor *P* de cada una de ellas. Este fichero lo utilizaremos para, mediante R, obtener un gráfico de Manhattan que nos indicará los SNPs más significativos. Por otro lado, la corrección múltiple nos proporcionará una serie de valores estadísticos que utilizaremos para,

mediante R, obtener un gráfico Q-Q que sirve para ver la desviación de los datos obtenidos frente a los datos esperados.

3. IMPUTACIÓN

Debido a que no presentamos una cantidad de muestra suficiente para obtener resultados significativos, se lleva a cabo el proceso de imputación. La imputación genotípica combina las variantes génicas del GWAS junto con un panel de referencia de haplotipos conocidos. En nuestro caso, hemos utilizados dos paneles diferentes.

Antes de llevar a cabo este proceso, tenemos que preparar nuestros archivos y convertirlos en el formato que requieren los softwares de imputación. En primer lugar, mediante PLINK, partiendo de nuestros archivos que han pasado el control de calidad, dividimos las muestras por cromosomas y las trasformamos para que presenten una extensión .vcf. A continuación, mediante el software BCFtools(*Bcftools by Samtools*, n.d.), transformamos los archivos .vcf en .vcf.gz, que es la extensión que se necesita para llevar a cabo la imputación.

Finalmente, con las muestras imputadas, se volvería a realizar el GWAS, teniendo en esta ocasión más probabilidades de obtener variantes génicas significativas.

3.1. HAPLOTYPE REFERENCE CONSORTIUM (HRC)

Este panel de referencia está compuesto por 32.740 muestras y 64.940 haplotipos, que presentan una ascendencia mayormente europea, además de incluir los datos del “100 Genome Project”. La imputación con este panel se lleva a cabo mediante un servidor llamado “Michigan Imputation Server”(Das et al., 2016) que utiliza Minimac 4 (Fuchsberger et al., 2015).

3.2. TRANS-OMICS FOR PRECITION MEDICINE (TOPMed)

Por otro lado, este panel(Taliun et al., 2019) incluye la información de 97.256 genomas profundamente secuenciados y 194.512 haplotipos procedentes de individuos que presentan ascendencia de diverso origen . Con este panel se utiliza un servidor llamado “TOPMed Imputation Server” que también utiliza Minimac4 para llevar a cabo la imputación.

4. FUMA GWAS

Finalmente, utilizamos un programa llamado FUMA GWAS(Watanabe et al., n.d.), que se encarga de analizar los resultados del GWAS y nos indica a que genes pertenecen nuestros resultados y en que tejido se expresan mayoritariamente estos genes. Para obtener estos datos es necesario introducir en este software el fichero obtenido de la regresión lineal producida en PLINK.

RESULTADOS Y DISCUSIÓN

Para poder plasmar los resultados del estudio y que sea entendible, voy a dividir los resultados en primer lugar según el panel de referencia utilizado en la imputación y en segundo lugar según el origen de las muestras.

1. RESULTADOS ANTES DE LA IMPUTACIÓN

Antes de llevar a cabo la imputación, para poder obtener una muestra lo mejor genotipada posible, aplicamos el control de calidad y el estudio de población. Además, aunque sabíamos que no obtendríamos ningún resultado significativo, realizamos la asociación.

1.1. CIBERSAM

Inicialmente, estas muestras están compuestas por 603.132 SNPs y 3.888 individuos. De estos individuos, nos quedamos con 1.867 (1.164 M y 703 F) que son los que padecen de esquizofrenia. Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo III*.

1.1.1. CONTROL DE CALIDAD

En el primer paso del control de calidad filtramos los SNPs y los individuos y nos quedamos con aquellos que presenten un *Missingness* menor al 2% (559.466 SNPs y 1.840 individuos). Luego, se eliminan 186 individuos que presentan una discrepancia en el sexo asignado. Posteriormente, utilizando un filtro de *MAF* < 0.05 nos quedamos con 250.800 SNPs y con el filtro de *HWE* > 1e-06 se eliminan 465 SNPs. Después se eliminan 84 individuos que presentan heterocigosidad. Finalmente, nos quedamos con 1.536 individuos que no presentan parentesco entre ellos (*Tabla 2*).

Tabla 2. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*		*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	603.132	1.867	1.164 M 703 F	-	-
MISSINGNES (< 0.02)	559.466	1.840	1.146 M 694 F	43.666	27
SEX DISCREPANCIES	559.466	1.654	1.095 M 559 F	-	186
MAF (< 0.05)	250.800	1.654	1.095 M 559 F	308.666	-
HWE (> 1e-06)	250.335	1.654	1.095 M 559 F	465	-
HETEROZYGOSITY	250.335	1.616	1.067 M 549 F	-	38
RELATEDNESS	250.335	1.570	1.041 M 529 F	-	46

1.1.2. ESTUDIO POBLACIONAL

Después de realizar el control de calidad llevamos a cabo el estudio de población (*Figura 3*), en este se eliminan 34 individuos que se alejan de la zona de población europea y nos quedamos con 250.335 SNPs y 1.536 individuos (1.021 M y 515 F).

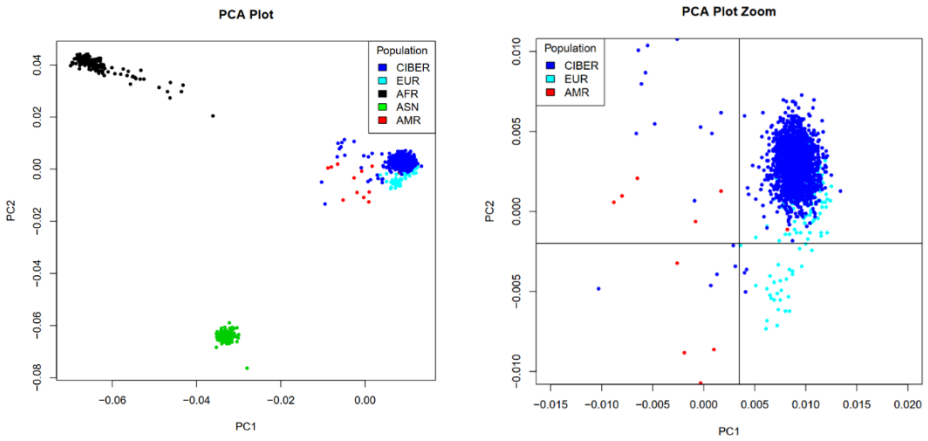


Figura 3. Gráfico y gráfico aumentado del estudio poblacional de las muestras del CIBERSAM.

1.1.3. ESTUDIO DE ASOCIACIÓN

Finalmente, se lleva a cabo el estudio de asociación (*Figura 4*). Los SNPs más significativos que se obtuvieron presentaron un valor *P* de orden 1e-06 (rs1044088 y rs6812381) (*Tabla 3*), por lo que no llegan a ser lo suficientemente significativos para relacionarlos con la edad de inicio de la esquizofrenia.

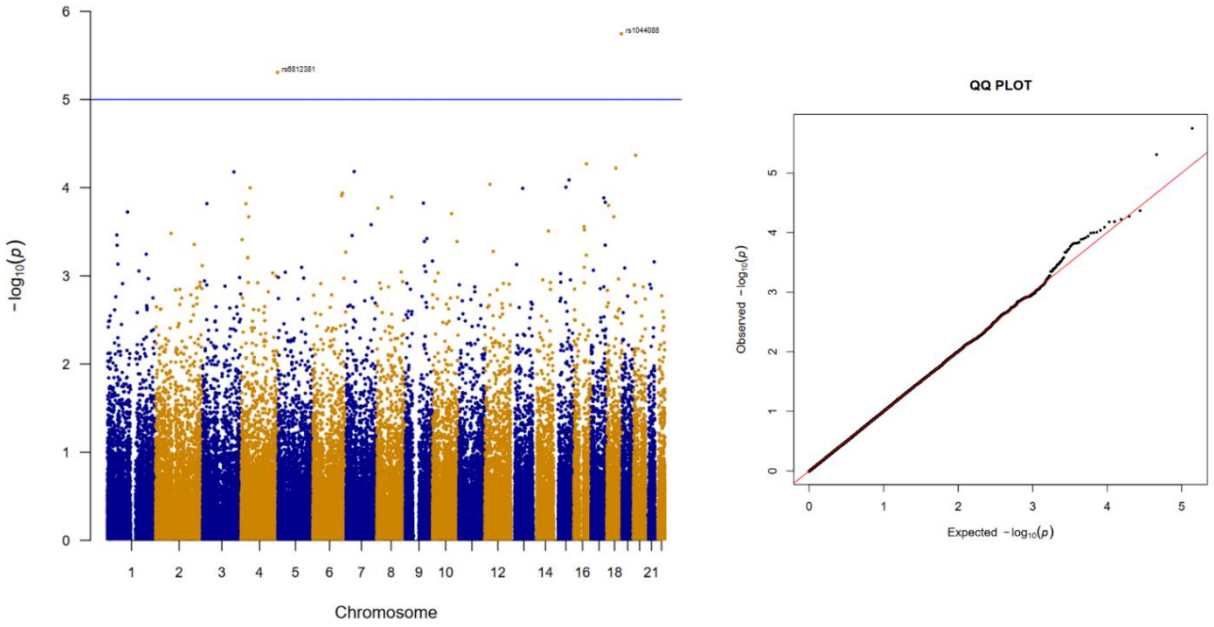


Figura 4. Gráfico de Manhattan y gráfico Q-Q resultado del análisis de asociación de los datos del CIBERSAM.

Tabla 3. Resultados más significativos del estudio de asociación de los datos del CIBERSAM

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*
rs1044088	18	74682214	ZNF236	C	T	1.774e-06
rs6812381	4	188493082	LINC02492 LOC105377604	A	G	4.88e-06

1.2. KFO

En este caso partimos con 276.154 SNPs y 1.448 individuos (751 M y 697F). Estos datos ya han sido filtrados parcialmente por el Klinische Forschergruppe 241 pero igualmente los volveremos a filtrar según los parámetros utilizados en los datos del CIBERSAM. Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo IV*.

1.2.1. CONTROL DE CALIDAD

Analizando la pérdida de genotipado se eliminaron 12 individuos. Posteriormente, aplicando el filtro del *MAF* < 0.05 nos quedamos con 248.053 SNPs. En el caso de la heterocigosidad, eliminamos 19 individuos. Finalmente, nos quedamos con 1.408 individuos que no presentan parentesco entre ellos (*Tabla 4*).

Tabla 4. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*	*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	276.154	1.448 751 M 697 F	-	-
MISSINGNES (< 0.02)	276.154	1.436 745 M 691 F	-	12
MAF (< 0.05)	248.053	1.436 745 M 691 F	28.101	-
HETEROZYGOSITY	248.053	1.417 737 M 680 F	-	19
RELATEDNESS	248.053	1.408 734 M 674 F	-	9

1.2.2. ESTUDIO POBLACIONAL

Después de realizar el control de calidad llevamos a cabo el estudio poblacional (*Figura 5*), en este se eliminan 5 individuos que se alejan de la zona de población europea y nos quedamos con 248.053 SNPs y 1.403 individuos (732 M y 671 F).

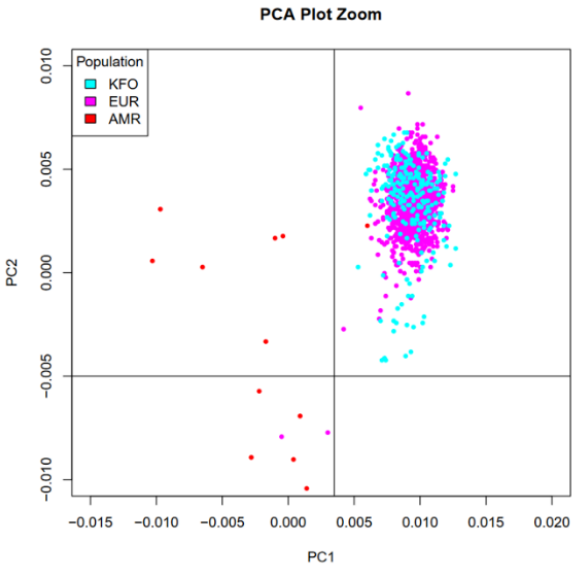


Figura 5. Gráfico aumentado del estudio poblacional de las muestras del KFO.

1.2.3. ESTUDIO DE ASOCIACIÓN

Finalmente, se lleva a cabo el estudio de asociación (*Figura 6*). Los SNPs más significativos que se obtuvieron presentaron un valor *P* de orden 1e-05 (rs2937550, rs5015755, rs3949904, rs39033, rs11730375, rs79434212, rs13411730, rs10123153,

rs2031373) (*Tabla 5*), por lo que no llegan a ser lo suficientemente significativos para relacionarlos con la edad de inicio de la esquizofrenia.

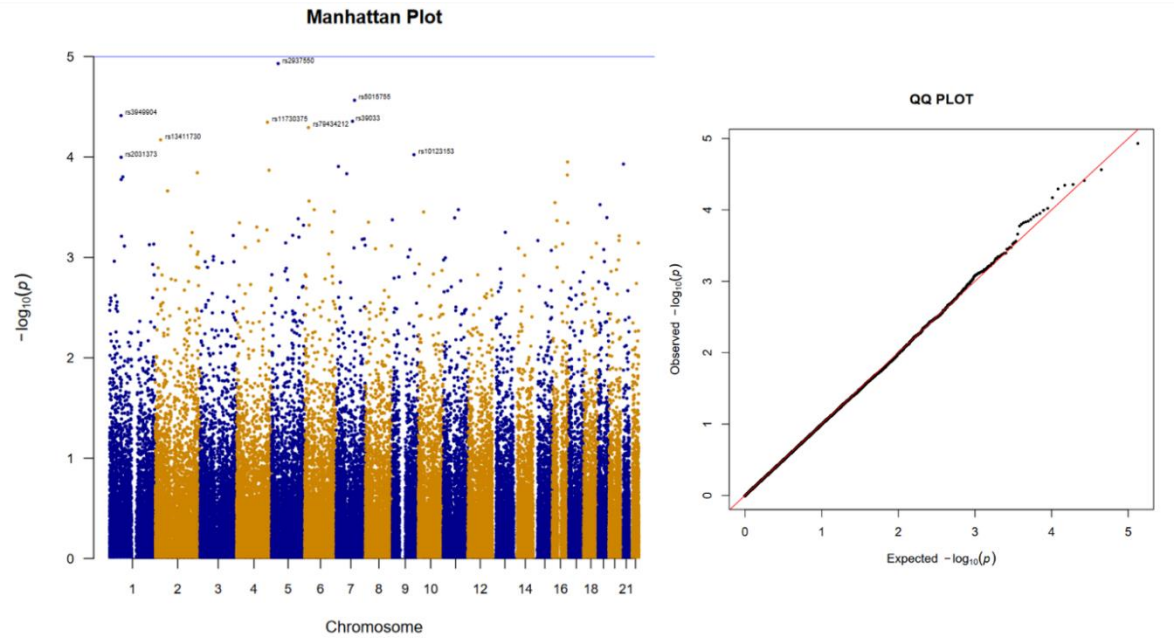


Figura 6. Gráfico de Manhattan y gráfico Q-Q resultado de la asociación de los datos del KFO.

Tabla 5. Resultados más significativos del análisis de asociación de los datos del KFO.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*
rs2937550	5	36473237	-	C	T	1.166e-05
rs5015755	7	100013402	ZCWPW1	G	T	2.707e-05
rs3949904	1	62848090	-	G	A	3.845e-05
rs39033	7	89099843	LOC105375387	T	C	4.378e-05
rs11730375	4	168626439	-	C	T	4.478e-05
rs79434212	6	20137457	MBOAT1	T	G	5.053e-05
rs13411730	2	28605588	LOC105374383	C	T	6.608e-05
rs10123153	9	117728895	DELEC1	T	C	9.427e-05
rs2031373	1	63083289	DOCK7	C	A	9.995e-05

2. IMPUTACIÓN CON EL PANEL DE REFERENCIA HRC

Debido a que, con los datos proporcionados por los consorcios, no hemos encontrado ninguna variante significativa, se lleva a cabo un proceso de imputación con el panel HRC para aumentar el número de las variantes génicas, pero mantener el tamaño del genoma. Posteriormente, luego de filtrar los SNPs imputados según la calidad de imputación (INFO>.8), llevamos a cabo otro control de calidad y otro análisis poblacional y finalmente el estudio de asociación.

2.1. CIBERSAM

Partiendo de los 250.335 SNPs obtenidos del control de calidad anterior, después de la imputación con el panel de referencia HRC, obtenemos 14.471.945 SNPS. De estos solo 1.707.254 superan el filtro de calidad (INFO>.8). Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo V*.

2.1.1. CONTROL DE CALIDAD

Todos los SNPs e individuos presentan un *Missingness* inferior a 0.02, por lo que no se elimina ninguno. No hay ningún individuo con discrepancias de sexo. Posteriormente, utilizando el filtro de *HWE* > 1e-06, se eliminan 3 SNPs. Después se eliminan 25 individuos que presentan heterocigosidad. Como no hay ningún individuo que presente parentesco con otro, nos quedamos con 1.511 individuos. Finalmente, aplicamos el filtro de *MAF* > 0.01 y nos quedamos con 99.828 SNPs (*Tabla 6*).

Tabla 6. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*		*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	1.707.254	1.536	1.021 M 515 F	-	-
HWE (> 1e-06)	1.707.251	1.536	1.021 M 515 F	3	-
HETEROZYGOSITY	1.707.251	1.511	1.004 M 507 F	-	25
MAF (< 0.01)	99.828	1.511	1.004 M 507 F	1.607.423	-

2.1.2. ESTUDIO POBLACIONAL

En este caso, cuando se ha llevado a cabo el estudio de población (*Figura 7*), no se ha encontrado ningún individuo que estuviese fuera de la zona europea.

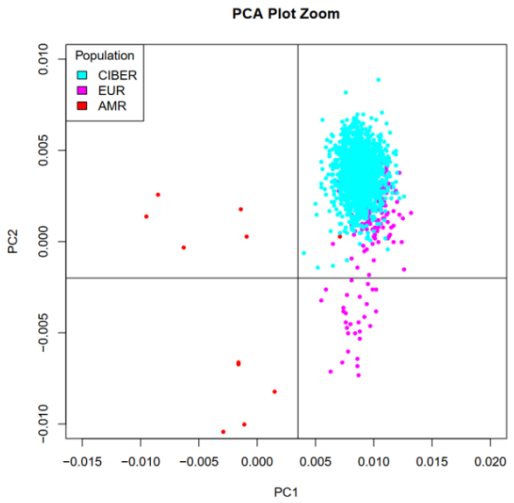


Figura 7. Gráfico aumentado del estudio poblacional de las muestras del CIBERSAM imputado con el panel HRC.

2.1.3. ESTUDIO DE ASOCIACIÓN

Por último, se lleva a cabo el estudio de asociación (*Figura 8*). El SNP más significativo presenta un valor *P* del orden de 1e-06 (rs6812381) (*Tabla 7*), por lo que no es suficientemente significativo como para estar asociado a la edad de inicio de la esquizofrenia.

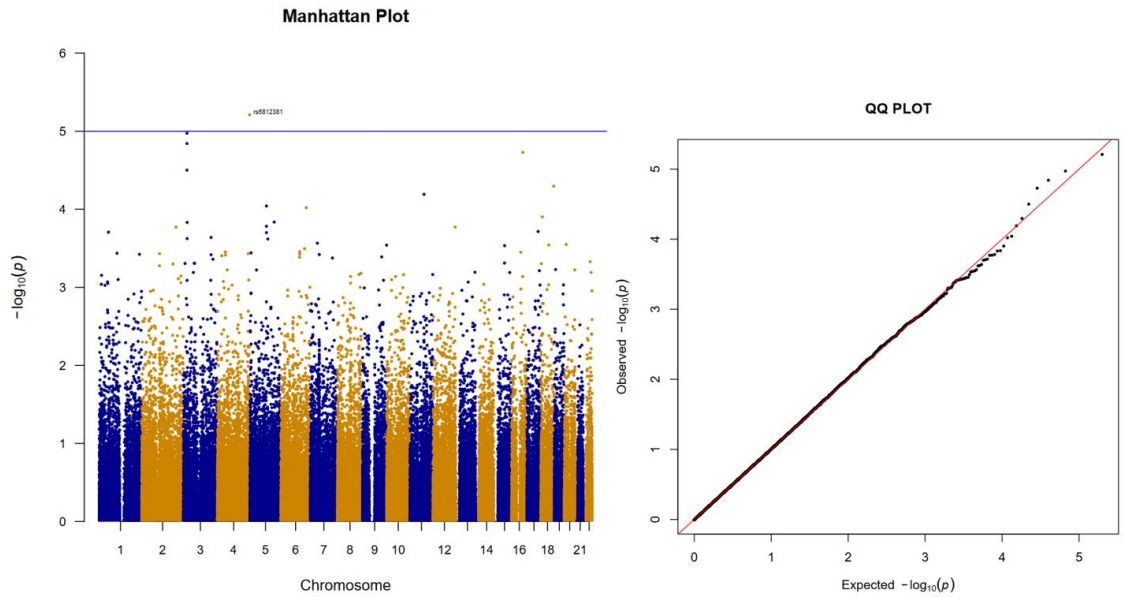


Figura 8. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del CIBERSAM imputados con el panel HRC.

Tabla 7. Resultado más significativo del estudio de asociación de los datos del CIBERSAM imputados con el panel HRC.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO*
rs6812381	4	188493082	LINC02492	A	G	6.117e-06	0.98512

2.2. KFO

Partiendo de los 248.053 SNPs obtenidos del control de calidad anterior, después de la imputación con el panel de referencia HRC, obtenemos 14.0.77.398 SNPS. De estos solo 2.499.673 superan el filtro de calidad (INFO>.8). Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo VI*.

2.2.1. CONTROL DE CALIDAD

Todos los SNPs e individuos presentan un *Missingness* inferior a 0.02, por lo que no se elimina ninguno. No hay ningún individuo con discrepancias de sexo. Posteriormente, utilizando el filtro de *HWE* > 1e-06, se eliminan 9 SNPs. Después se eliminan 14 individuos que presentan heterocigosidad. Como no hay ningún individuo que presente parentesco con otro, nos quedamos con 1.389 individuos. Finalmente, aplicamos el filtro de *MAF* > 0.01 y nos quedamos con 146.235 SNPs (*Tabla 8*).

Tabla 8. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*	*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	2.499.673	1.403 732 M 671 F	-	-
HWE (> 1e-06)	2.499.664	1.403 732 M 671 F	9	-

HETEROZYGOSITY	2.499.664	1.389	728 M 661 F	-	14
MAF (< 0.01)	146.235	1.389	728 M 661 F	2.353.429	-

2.2.2. ESTUDIO POBLACIONAL

En este caso, cuando se ha llevado a cabo el estudio de población (*Figura 9*), no se ha encontrado ningún individuo que estuviese fuera de la zona europea.

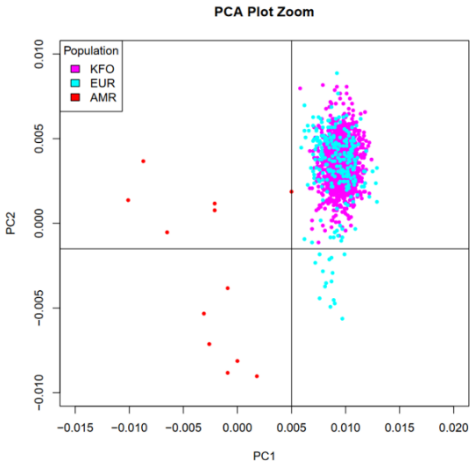


Figura 9. Gráfico aumentado del estudio poblacional de las muestras del KFO imputado con el panel HRC.

2.2.3. ESTUDIO DE ASOCIACIÓN

Por último, se lleva a cabo el estudio de asociación (*Figura 10*). Los SNPs más significativos presentan un valor *P* del orden de 1e-06 y 1e-07 (rs140595382, rs143958850, rs39033, rs4664680, rs2271034) (*Tabla 9*), por lo que no son suficientemente significativos como para estar asociados a la edad de inicio de la esquizofrenia.

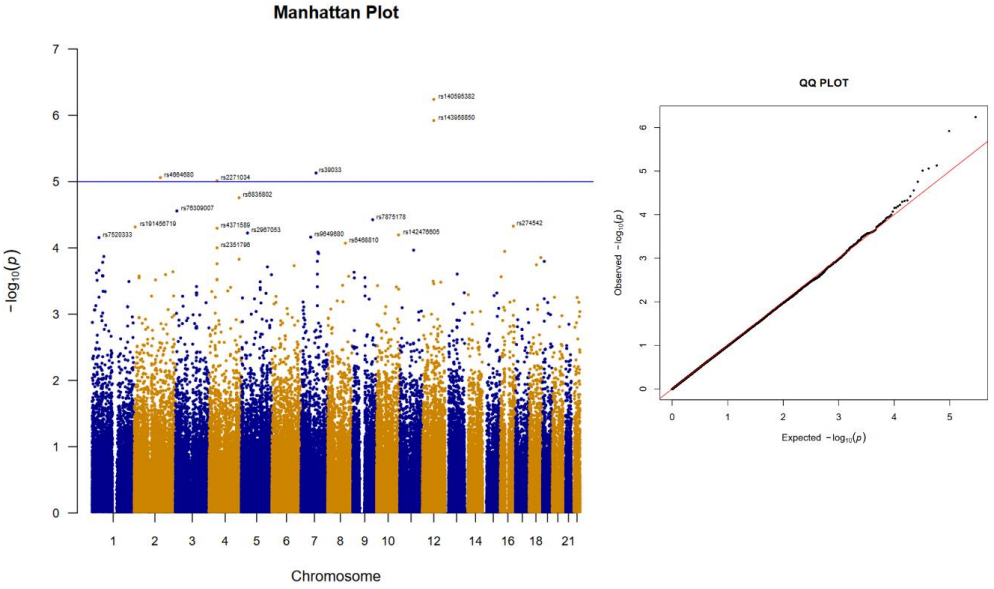


Figura 10. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del KFO imputados con el panel HRC.

Tabla 9. Resultado más significativo del estudio de asociación de los datos del KFO imputados con el panel HRC.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO*
rs140595382	12	67000435	GRIP1	A	T	5.693e-07	0.87896
rs143958850	12	67037767	GRIP1	G	T	1.187e-06	0.89617
rs39033	7	89099843	LOC105375387	T	G	7.341e-06	0.98950

rs4664670	2	154298534	LINC01850	G	T	8.618e-06	0.84392
rs2271034	4	47588848	ATP10D	T	C	9.36e-06	0.88103

2.3. CIBERSAM-KFO

Juntamos los individuos y las variantes filtradas (INFO>.8) y obtenidas después de cada imputación y obtenemos una muestra con 3.256.923 SNPs y 2.939 individuos. De estos SNPs, solo 950.004 son comunes a ambas muestras. Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo VII*.

2.3.1. CONTROL DE CALIDAD

Solo los 950.004 SNPs presentes en ambas muestras superan el filtrado por *Missingness* < 0.02. Además, no se encuentran discrepancias en el sexo asignado de ningún individuo. En cuanto a la frecuencia de *HWE*, se eliminan 17 variantes. Posteriormente, se eliminan 29 individuos que presentan heterocigosidad y 13 individuos que presentan parentesco entre ellos. Finalmente, aplicando el filtro *MAF (<0.01*), obtenemos 68.943 SNPs (*Tabla 10*).

Tabla 10. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*	*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	3.256.923	2.939 1.753 M 1.186 F	-	-
MISSINGNES (< 0.02)	950.004	2.939 1.753 M 1.186 F	2.306.919	0
HWE (> 1e-06)	949.987	2.939 1.753 M 1.186 F	17	-
HETEROZYGOSITY	949.987	2.910 1.732 M 1.178 F	-	29
RELATEDNESS	949.987	2.897 1.723 M 1.174 F	-	13
MAF (< 0.01)	68.943	2.897 1.723 M 1.174 F	881.044	-

2.3.2. ESTUDIO POBLACIONAL

Llevamos a cabo el estudio poblacional (*Figura 11*) con las muestras de los dos consorcios juntos y eliminamos un individuo que se encuentra alejado de la zona europea.

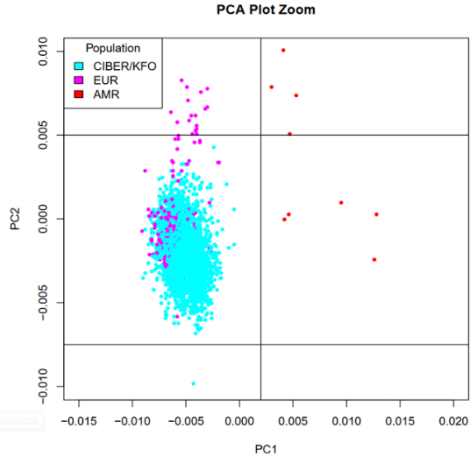


Figura 11. Gráfico aumentado del estudio poblacional de las muestras del CIBERSAM y el KFO juntas e imputadas con el panel HRC.

2.3.3. ESTUDIO DE ASOCIACIÓN

Finalmente, llevamos a cabo el estudio de asociación (*Figura 12*). En este encontramos 3 SNPs (rs6847217, rs9997352 y rs4371589) (*Tabla 11*) que presentan un valor *P* del orden de 1e-08. Estos son lo suficientemente significativos como para estar involucrados en la aparición de la esquizofrenia. Posteriormente buscaremos a que gen pertenece y si presenta algún fenotipo asociado anteriormente.

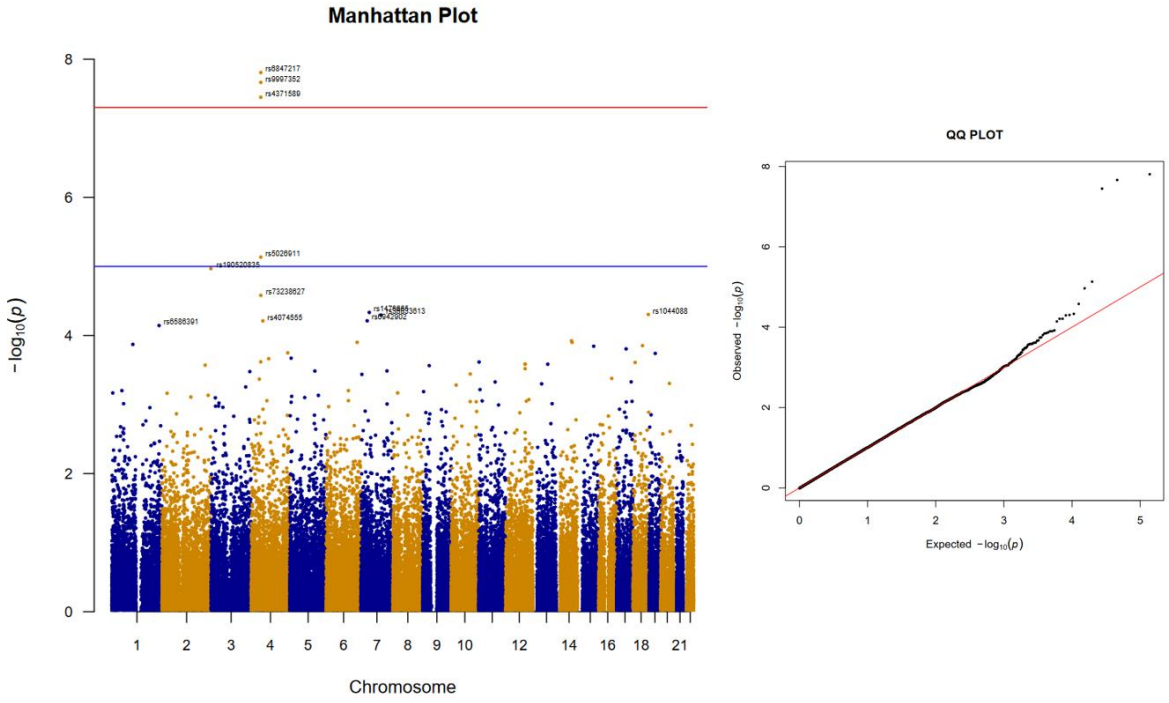


Figura 12. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del CIBERSAM y de KFO agrupados e imputados con el panel HRC.

Tabla 11. Resultado más significativo del estudio de asociación de los datos del CIBERSAM y del KFO agrupados e imputados con el panel HRC.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO CIBER*	*INFO KFO*
rs6847217	4	47606137	CORIN	G	A	1.541e-08	0.91476	0.93116
rs9997352	4	47620949	CORIN	C	G	2.143e-08	0.91791	0.93419
rs4371589	4	47634479	CORIN LOC105374444	C	T	3.51e-08	0.87178	0.90361

3. IMPUTACIÓN CON EL PANEL DE REFERENCIA TOPMED

Utilizamos el panel de referencia TOPMed debido a qué se publicó mientras estábamos llevando a cabo el estudio. Además, este panel presenta 3 veces más información que el panel HRC. Aunque, al contrario que el HRC, la información pertenece a personas de diversa ascendencia. Igual que en el caso anterior, después de la imputación llevaremos a cabo un proceso de filtrado de los SNPs (INFO>.8), el control de calidad, el estudio poblacional y, finalmente, el estudio de asociación.

3.1. CIBERSAM

Partiendo de los 248.053 SNPs obtenidos del control de calidad anterior, después de la imputación con el panel de referencia TOPMed, obtenemos 27.426.300 SNPS. De estos solo 7.233.599 superan el filtro de calidad (INFO>.8). Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo VIII*.

3.1.1. CONTROL DE CALIDAD

Todos los SNPs e individuos presentan un *Missingness* inferior a 0.02, por lo que no se elimina ninguno. No hay ningún individuo con discrepancias de sexo. Posteriormente, utilizando el filtro de *HWE* > 1e-06, se eliminan 4 SNPs. Después se eliminan 14 individuos que presentan heterocigosidad. Como no hay ningún individuo que presente parentesco con otro, nos quedamos con 1.522 individuos. Finalmente, aplicamos el filtro de *MAF* < 0.01 y nos quedamos con 320.987 SNPs (*Tabla 12*).

Tabla 12. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*		*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	7.233.599	1.536	1.021 M 515 F	-	-
HWE (> 1e-06)	7.233.595	1.536	1.021 M 515 F	4	-
HETEROZYGOSITY	7.233.595	1.522	1.021 M 510 F	-	14
MAF (< 0.01)	320.987	1.522	1.021 M 510 F	6.912.608	-

3.1.2. ESTUDIO POBLACIONAL

En este caso, cuando se ha llevado a cabo el estudio de población, no se ha encontrado ningún individuo que estuviese fuera de la zona europea.

3.1.3. ESTUDIO DE ASOCIACIÓN

Por último, se lleva a cabo el estudio de asociación (*Figura 13*). Los SNPs más significativos presentan un valor *P* del orden de 1e-06 (rs190213379, rs142960304 y rs143001143) (*Tabla 13*), por lo que no son suficientemente significativos como para estar asociados a la edad de inicio de la esquizofrenia.

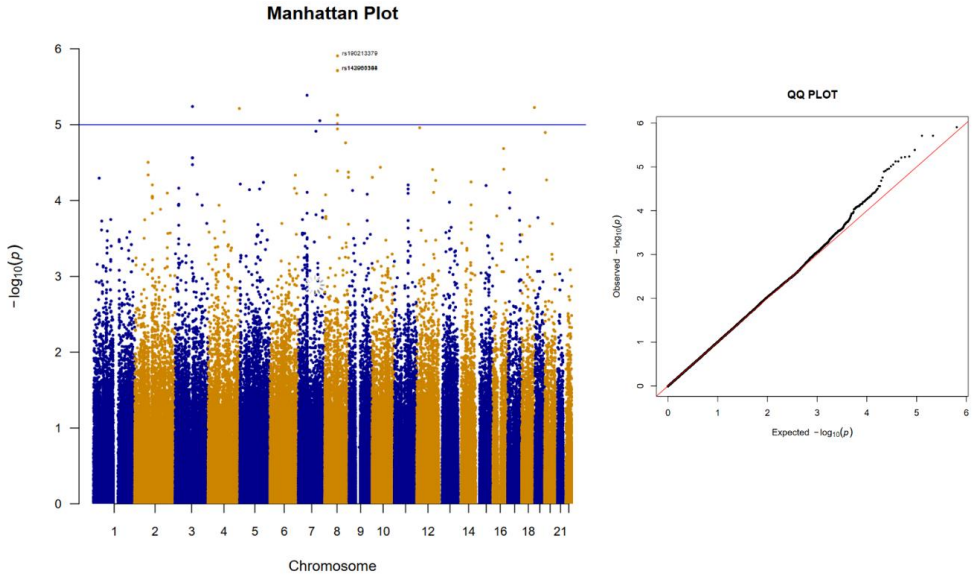


Figura 13. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del CIBERSAM imputados con el panel TOPMed.

Tabla 13. Resultado más significativo del estudio de asociación de los datos del CIBERSAM imputados con el panel TOPMed.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO*
rs190213379	8	75789106	-	G	T	1.228e-06	0.82981
rs142960304	8	75512699	HNF4G	C	T	1.921e-06	0.89284
rs143001143	8	75481075	HNF4G	G	A	1.921e-06	0.85982

3.2. KFO

Partiendo de los 248.053 SNPs obtenidos del control de calidad anterior, después de la imputación con el panel de referencia TOPMed, obtenemos 23.922.693 SNPS. De estos solo 7.537.335 superan el filtro de calidad (INFO>.8). Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo IX*.

3.2.1. CONTROL DE CALIDAD

Todos los SNPs e individuos presentan un *Missingness* inferior a 0.02, por lo que no se elimina ninguno. No hay ningún individuo con discrepancias de sexo. Posteriormente, utilizando el filtro de *HWE* > 1e-06, se eliminan 2 SNPs. Después se eliminan 8 individuos que presentan heterocigosidad. Como no hay ningún individuo que presente parentesco con otro, nos quedamos con 1.395 individuos. Finalmente, aplicamos el filtro de *MAF* < 0.01 y nos quedamos con 436.147 SNPs (*Tabla 14*).

Tabla 14. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*		*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	7.537.335	1.403	732 M 671 F	-	-
HWE (> 1e-06)	7.537.333	1.403	732 M 671 F	2	-

HETEROZYGOSITY	7.537.333	1.395	729 M 666 F	-	8
MAF (< 0.01)	436.147	1.389	729 M 666 F	7.101.186	-

3.2.2. ESTUDIO POBLACIONAL

En este caso, cuando se ha llevado a cabo el estudio de población, no se ha encontrado ningún individuo que estuviese fuera de la zona europea

3.2.3. ESTUDIO DE ASOCIACIÓN

Por último, se lleva a cabo el estudio de asociación (*Figura 14*). Los SNPs más significativos presentan un valor *P* del orden de 1e-07 (rs11744742 y rs80323751) (*Tabla 15*), por lo que no son suficientemente significativos como para estar asociados a la edad de inicio de la esquizofrenia.

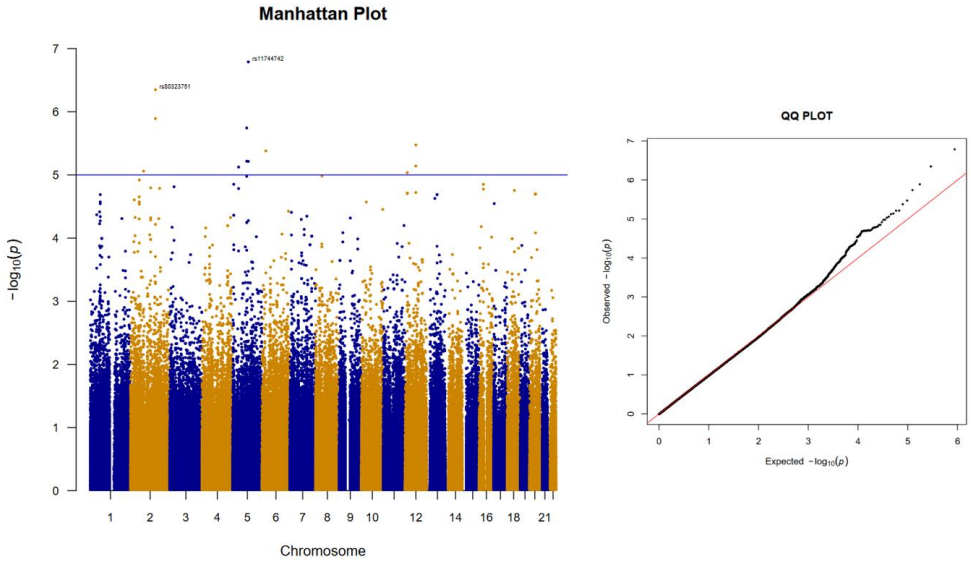


Figura 14. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del KFO imputados con el panel TOPMed.

Tabla 15. Resultado más significativo del estudio de asociación de los datos del KFO imputados con el panel TOPMed.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO*
rs11744742	5	96558701	LOC101929710	T	C	1.608e-07	0.80074
rs80323751	2	153651076	-	C	T	4.431e-07	0.86716

3.3. CIBERSAM-KFO

Juntamos los individuos y las variantes filtradas (INFO>.8) y obtenidas después de cada imputación y obtenemos una muestra con 12.517.765 SNPs y 2.939 individuos. De estos SNPs, solo 2.253.169 son comunes a ambas muestras. Los gráficos asociados a cada paso del control de calidad se encuentran en el *Anexo X*.

3.3.1. CONTROL DE CALIDAD

Solo los 2.253.169 SNPs presentes en ambas muestras superan el filtrado por *Missingness* < 0.02. Además, no se encuentran discrepancias en el sexo asignado de ningún individuo. En cuanto a la frecuencia de *HWE*, se eliminan 3 variantes. Posteriormente, se eliminan 28 individuos que presentan heterocigosidad y 7 individuos que presentan parentesco entre ellos. Finalmente, aplicando el filtro *MAF (<0.01*), obtenemos 156.885 SNPs (*Tabla 16*).

Tabla 16. Proceso de filtrado de SNPs y de individuos según los parámetros del control de calidad.

ETAPAS	*SNPs*	*INDIVIDUOS*		*SNPs ELIMINADOS*	*INDIVIDUOS ELIMINADOS*
DATOS INICIALES	12.517.765	2.939	1.753 M 1.186 F	-	-
MISSINGNES (< 0.02)	2.253.169	2.939	1.753 M 1.186 F	10.264.596	0
HWE (> 1e-06)	2.253.166	2.939	1.753 M 1.186 F	3	-
HETEROZYGOSITY	2.253.166	2.911	1.736 M 1.175 F	-	28
RELATEDNESS	2.253.166	2.904	1.732 M 1.172 F	-	7
MAF (< 0.01)	156.885	2.904	1.732 M 1.172 F	2.096.281	-

3.3.2. ESTUDIO POBLACIONAL

En este caso, cuando se ha llevado a cabo el estudio de población, no se ha encontrado ningún individuo que estuviese fuera de la zona europea

3.3.3. ESTUDIO DE ASOCIACIÓN

Por último, se lleva a cabo el estudio de asociación (*Figura 15*). Los SNPs más significativos presentan un valor *P* del orden de 1e-07 (rs684721, rs9997352, rs5858082 y rs4371589) (*Tabla 17*), por lo que no son suficientemente significativos como para estar asociados a la edad de inicio de la esquizofrenia.

Tabla 17. Resultado más significativo del estudio de asociación de los datos del CIBERSAM y del KFO agrupados e imputados con el panel TOPMed.

SNP	*CHR*	*POSITION (bp)*	*GENE*	*REF*	*ALT*	*P-VALUE*	*INFO CIBER*	*INFO KFO*
rs6847217	4	47606137	CORIN	G	A	1.178e-07	0.92307	0.92592
rs9997352	4	47620949	CORIN	C	G	2.029e-07	0.92604	0.92843
rs5858082	4	47639154	CORIN LOC105374443	TAC	T	3.636e-07	0.91841	0.90942
rs4371589	4	47634479	CORIN LOC105374444	C	T	6.216e-07	0.89501	0.90420

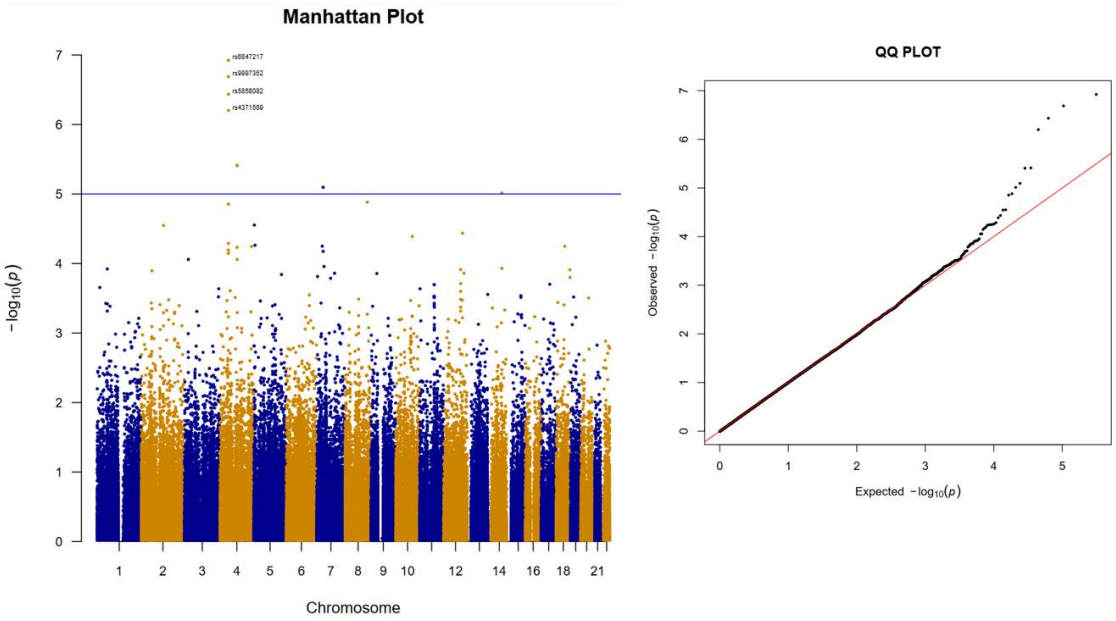


Figura 15. Gráfico de Manhattan y gráfico Q-Q resultado del estudio de asociación de los datos del CIBERSAM y del KFO agrupados e imputados con el panel TOPMed.

DISCUSIÓN

Con todos los resultados obtenidos, en el estudio llevado a cabo con los datos del CIBERSAM y del KFO agrupados e imputados con el panel de referencia HRC, se pueden observar 3 variantes génicas (rs6847217, rs9997352 y rs4371589) significativas con un valor *P* de nivel de 1e-08. Estas variantes también se encuentran presentes en el estudio llevado a cabo con el panel de imputación TOPMed, pero en este caso ninguna presenta un valor *P* significativo.

Utilizando las bases de datos dbSNP y ENSEMBL encontramos que estos SNPs se encuentran en el cromosoma 4 (4p12) y forman parte del intrón de un gen denominado CORIN. Este gen se traduce en una serin-proteasa transmembrana de tipo II de la familia de la tripsina y puede funcionar como una convertasa del péptido natriurético pro-cerebral. Las serin-proteasas o serin-peptidasas son unas enzimas capaces de degradar enlaces peptídicos, y se denominan de esta forma porque presentan una serina en su centro activo.

Como no hemos encontrado ninguna relación entre el gen CORIN y la esquizofrenia, hemos recurrido a un software llamado FUMA GWAS (Functional Mapping and Annotation of Genome-Wide Association Studies) (Watanabe et al., n.d.). Introducimos los datos de la regresión lineal, resultado del estudio de asociación, en este software y obtenemos diversos gráficos. En primer lugar, obtenemos un gráfico de Manhattan (*Figura 16)* que confirma que el gen CORIN es significativo en este estudio. En segundo

lugar, obtenemos un análisis de expresión tisular (*Figura 17)* que indica que la piel es el tejido donde más se expresa el gen CORIN.

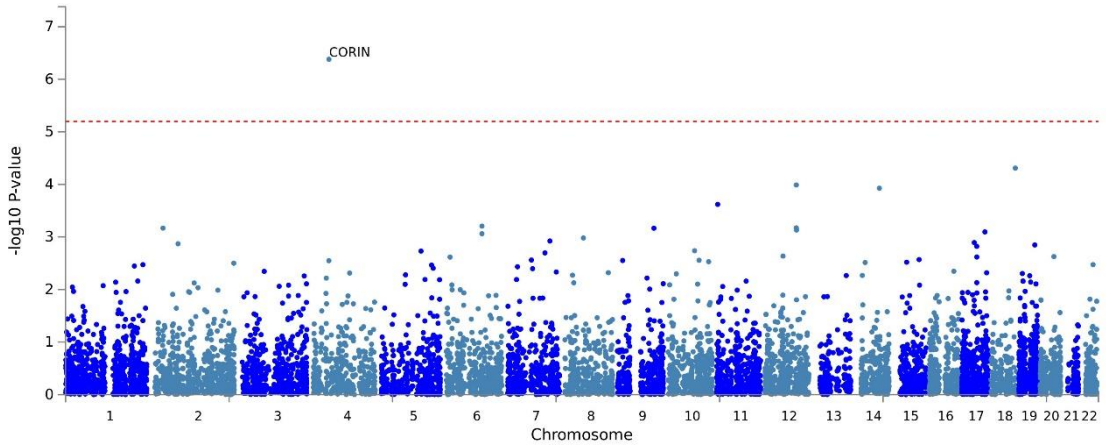


Figura 16. Gráfico de Manhattan que indica el gen más significativo de este estudio.

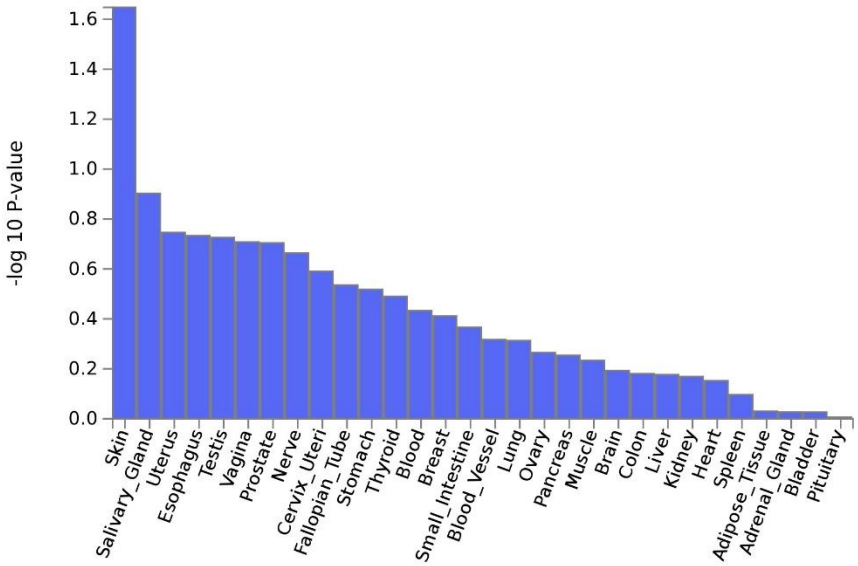


Figura 17. Gráfico MAGMA que indica el tejido en el que más se expresa el gen significativo.

Finalmente, buscamos si existía alguna asociación entre cualquier serin-proteasa y la esquizofrenia y encontramos un estudio(Duchatel et al., n.d.) que relacionaba la disminución parcial de la reelina, en ratones reeler heterocigóticos (rl+/-), con la aparición de efectos similares a los observados en esta enfermedad. La reelina es segregada por las células Cajal-Retzius durante el desarrollo de la zona marginal y se encarga de guiar la migración neuronal y su posicionamiento durante el desarrollo del cerebro. Duchatel et al. asociaron esta proteína al aumento de densidad de las neuronas intersticiales de la materia blanca, un fenotipo característico de la esquizofrenia.

CONCLUSIÓN

Finalmente, después de llevar a cabo este estudio, podemos confirmar la eficacia de los estudios de asociación del genoma completo a la hora de identificar variantes génicas relacionadas con el fenotipo asociado.

Hemos identificado una serie de SNPs (rs6847217, rs9997352 y rs4371589) presentes en el cromosoma 4 y que, según la regresión lineal, están asociados a la edad de inicio de la esquizofrenia. Esta combinación de SNPs se localizan en el intrón del gen CORIN, también conocido como CRN, ATC2 Lrp4, PEE5 o TMPRSS10. Este gen codifica una serin-proteasa de tipo II que pertenece a la superfamilia de la tripsina, además, existen diferentes isoformas para la proteína codificada. Una de ellas se encarga de convertir el péptido natriurético pro-atrial en péptido natriurético auricular biológicamente activo, que es una hormona cardíaca que regula el volumen de sangre y la presión. Por otro lado, esta proteína también se encarga de convertir el péptido natriurético pro-cerebral. Además, se ha observado que la piel es el tejido donde más se expresa este gen.

Cabe destacar, que este estudio es el principio de un proyecto que continuará partiendo de estos resultados como base.

BIBLIOGRAFÍA

Bcftools by samtools. (n.d.). Retrieved July 8, 2020, from http://samtools.github.io/bcftools/

Bergen, S. E., O'dushlaine, C. T., Lee, P. H., Fanous, A. H., Ruderfer, D. M., Ripke, S., Sullivan, P. F., Smoller, J. W., Purcell, S. M., & Corvin, A. (2014). Genetic modifiers and subtypes in schizophrenia: Investigations of age at onset, severity, sex and family history HHS Public Access. *Schizophr Res*, *154*(0), 48–53. https://doi.org/10.1016/j.schres.2014.01.030

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

Cibersam. (n.d.). Retrieved June 28, 2020, from https://www.cibersam.es/

Clinical Research Group 241. (n.d.). Retrieved June 28, 2020, from http://www.kfo241.de/index_en.php

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Diagnostic and statistical manual of mental disorders, 5th ed. Washington, DC: American Psychiatric Association, 2013.

Duchatel, R. J., Weickert, C. S., & Tooney, P. A. (n.d.). *White matter neuron biology and neuropathology in schizophrenia*. https://doi.org/10.1038/s41537-019-0078-8

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). Minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. https://doi.org/10.1093/bioinformatics/btu704

Gaebel, W., & Zielasek, J. (2015). Schizophrenia in 2020: Trends in diagnosis and therapy. *Psychiatry and Clinical Neurosciences*, *69*(11), 661–673. https://doi.org/10.1111/pcn.12322

Hardvard University. (2017). *PLINK: Whole genome data analysis toolset*. American Journal of Human Genetics. http://zzz.bwh.harvard.edu/plink/tutorial.shtml

Henriksen, M. G., Nordgaard, J., & Jansson, L. B. (2017). Genetics of schizophrenia: Overview of methods, findings and limitations. In *Frontiers in Human Neuroscience* (Vol. 11). Frontiers Media S. A. https://doi.org/10.3389/fnhum.2017.00322

Infinium OmniExpressExome-8 Kit. (n.d.). Retrieved July 8, 2020, from https://emea.illumina.com/products/by-type/microarray-kits/infinium-omni-express-exome.html?langsel=/es/

Infinium PsychArray-24 Kit | Psychiatric predisposition microarray. (n.d.). Retrieved July 8, 2020, from https://www.illumina.com/products/by-type/microarray-kits/infinium-psycharray.html

Institut d'Investigació Sanitària Pere Virgili - IISPV. (n.d.). Retrieved June 24, 2020, from http://www.iispv.cat/es_index.html

Main - GPA: GWAS Analysis. (n.d.). Retrieved June 28, 2020, from https://farre-xavi.gitbook.io/gpa-gwas-analysis/

Marder, S. R., & Cannon, T. D. (2019). Schizophrenia. In *New England Journal of Medicine* (Vol. 381, Issue 18, pp. 1753–1761). https://doi.org/10.1056/NEJMra1808803

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

Mata, P., Del, D., & Investigació, P. D. (2013). *Institut pere mata*. 1–9.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA - Journal of the American Medical Association*, *299*(11), 1335–1344. https://doi.org/10.1001/jama.299.11.1335

Perkovic, M. N., Erjavec, G. N., Strac, D. S., Uzun, S., Kozumplik, O., & Pivac, N. (2017). Theranostic biomarkers for schizophrenia. In *International Journal of Molecular Sciences* (Vol. 18, Issue 4). https://doi.org/10.3390/ijms18040733

PLINK 1.9. (n.d.). Retrieved June 29, 2020, from https://www.cog-genomics.org/plink/1.9/general_usage#cite

Price, Alkes L., et al. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38.8 (2006): 904-909

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*, 563866. https://doi.org/10.1101/563866

Valton, V., Romaniuk, L., Douglas Steele, J., Lawrie, S., & Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. In *Neuroscience and Biobehavioral Reviews* (Vol. 83, pp. 631–646). https://doi.org/10.1016/j.neubiorev.2017.08.022

Wang, K. S., Liu, X., Zhang, Q., Aragam, N., & Pan, Y. (2011). Genome-wide association analysis of age at onset in schizophrenia in a European-American sample. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *156*(6), 671–680. https://doi.org/10.1002/ajmg.b.31209

Watanabe, K., Taskesen, E., Van Bochoven, A., & Posthuma, D. (n.d.). *Functional mapping and annotation of genetic associations with FUMA*. https://doi.org/10.1038/s41467-017-01261-5

What is the PGC? | Psychiatric Genomics Consortium. (n.d.). Retrieved June 28, 2020, from https://www.med.unc.edu/pgc/

Woolston, A. L., Hsiao, P. C., Kuo, P. H., Wang, S. H., Lien, Y. J., Liu, C. M., Hwu, H. G., Lu, T. P., Chuang, E. Y., Chang, L. C., Chen, C. H., Wu, J. Y., Tsuang, M. T., & Chen, W. J. (2017). Genetic loci associated with an earlier age at onset in multiplex schizophrenia. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-06795-8

AUTOEVALUACIÓN

Llevar a cabo este trabajo, asociado a las prácticas externas, fue una buena experiencia para poder comprender como sería trabajar en un grupo de investigación. Sinceramente, al empezar este trabajo pensé que con los conocimientos básicos de bioinformática que había adquirido durante el grado tendría suficiente, pero pronto me di cuenta de mi error. Incluso después de la exhaustiva búsqueda de información, he llegado a la conclusión de que me queda mucho por aprender en este ámbito, por lo que me gustaría enfocar mi futuro académico en esta dirección.

Aun así, he aumentado mi conocimiento del idioma bash y he aprendido a realizar diversos análisis estadísticos, utilizando el software PLINK para obtenerlos y el paquete de recursos R para representarlos. Bien es cierto, que al principio del grado no entendí muy bien porque realizamos la asignatura Estadística, pero a lo largo de este, y sobre todo en las prácticas externas, me he dado cuenta de que es una herramienta imprescindible en cualquier trabajo que implique una gran cantidad de datos, ya que sin esta sería imposible analizarlos y obtener resultados.

ANEXOS

ANEXO I. CRITERIOS PARA DIAGNÓSTICAR ESQUIZOFRENIA SEGÚN EL "DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS, FIFTH EDITION (DSM-5)"

Los criterios específicos del DSM-5 para la esquizofrenia son los siguientes:

- La presencia de al menos dos de los siguientes cinco elementos, cada uno de los cuales está presente durante una porción de tiempo clínicamente significativa durante un período de un mes (o menos si se trata con éxito), siendo al menos uno de ellos los elementos 1, 2 o 3: delirios, alucinaciones, habla desorganizada, comportamiento groseramente desorganizado o catatónico y síntomas negativos (por ejemplo, disminución de la motivación y disminución de la expresividad).
- Durante una parte clínicamente significativa del tiempo transcurrido desde el inicio de la perturbación, el nivel de funcionamiento en una o más áreas importantes (por ejemplo, el trabajo, las relaciones interpersonales o el cuidado de sí mismo) está notablemente por debajo del nivel alcanzado antes del inicio; cuando el inicio se produce en la niñez o la adolescencia, no se alcanza el nivel esperado de funcionamiento interpersonal, académico u ocupacional.
- Los signos continuos de la perturbación persisten durante un período de por lo menos 6 meses, que debe incluir por lo menos un mes de síntomas (o menos si se trata con éxito); los síntomas prodrómicos suelen preceder a la fase activa, y pueden seguirle síntomas residuales, caracterizados por formas leves de alucinaciones o delirios.
- Se ha descartado el trastorno esquizoafectivo y el trastorno depresivo o bipolar con síntomas psicóticos porque no se han producido episodios depresivos, maníacos o mixtos importantes simultáneamente con los síntomas de la fase activa, o bien los episodios de humor que se han producido durante los síntomas de la fase activa han estado presentes durante una minoría de la duración total de los períodos activo y secundario de la enfermedad.
- La perturbación no es atribuible a los efectos fisiológicos de una sustancia (por ejemplo, una droga de uso indebido o un medicamento) o a otra afección médica.
- Si hay antecedentes de trastorno del espectro autista o de un trastorno de la comunicación de aparición en la infancia, el diagnostico adicional de esquizofrenia se hace solo si hay también delirios o alucinaciones prominentes, además de los demás síntomas requeridos o de la esquizofrenia, durante por lo menos un mes (o menos si se trata con éxito).
- Además de las áreas de dominio de los síntomas identificadas en el primer criterio de diagnóstico, la evaluación de los dominios de los síntomas de la cognición, la depresión y la manía es vital para distinguir entre la esquizofrenia y otros trastornos psicóticos.

ANEXO II. EJEMPLO SIMPLIFICADO DE SCRIPTS UTILIZADOS PARA LLEVAR A CABO ESTE ESTUDIO

SCRIPT .sh

- CONTROL DE CALIDAD
    - MISSINGNESS

plink --bfile Cibersam --missing --out missing (fichero para histograma en R)

plink --bfile Cibersam --geno 0.02 --make-bed --out Cibersam_1

plink --bfile Cibersam_1 --mind 0.02 --make-bed --out Cibersam_2

    - SEX DISCREPANCIES

plink --bfile Cibersam_2 --check-sex --out sex (fichero para histograma en R)

plink --bfile Cibersam_2 --remove sex_discrepancies.txt --make-bed --out Cibersam_3

    - MINOR ALLELE FRECUENCY

plink --bfile Cibersam_3 --freq --out maf_dist (fichero para histograma en R)

plink --bfile Cibersam_3 --maf 0.05 --make-bed --out Cibersam_4

    - HARDY-WEINBERG EQUILIBRIUM

plink --bfile Cibersam_4 --hardy --out hwe (fichero para histograma en R)

plink --bfile Cibersam_4 --hwe 1e-6 --make-bed --out Cibersam_5

    - HETEROZYGOSITY

plink --bfile Cibersam_5 --range --indep-pairwise 50 5 0.2 --out indepSNP

plink --bfile Cibersam_5 --extract indepSNP.prune.in --het --out het_check (fichero para histograma en R)

plink --bfile Cibersam_5 --remove fail_het_ind.txt --make-bed --out Cibersam_6

    - RELATEDNESS

plink --bfile Cibersam_6 --extract indepSNP.prune.in --rel-cutoff 0.125 --make-bed --out Cibersam_7

- ESTUDIO POBLACIONAL

plink --bfile Cibersam_7 --bmerge 1KG.bed 1KG.bim 1KG.fam --make-bed --out Cibersam_1KG

plink --bfile Cibersam_1KG --recode --out Cibersam_1KG

convetf -p ConvertPED.par (convierte los ficheros .ped y .map en ficheros que pueda utilizar EIGENSOFT)

smartpca -i Genotypefile.eigenstratgeno -a Snpfile.snp -b Indvfile.ind -o PCA_ciber -p Plot_ciber -e Eigenvalues_ciber -l Smartpca_ciber.log

#Se utiliza el fichero PCA_ciber.evec para hacer el gráfico en R

plink --bfile Cibersam_7 --keep simples_to_keep.txt --make-bed --out Cibersam_8

- ESTUDIO DE ASOCIACIÓN

plink --bfile Cibersam_8 --recode Cibersam_8

convertf -p ConvertPED_assoc.par

smartpca -i Genotype_assoc.eigenstratgeno -a Snpfile_assoc.snp -b Indvfile_assoc.ind -o PCA_assoc -p Plot_assoc -e Eigenvalues_assoc -k 10 -l Smartpca_assoc.log

plink --bfile Cibersam_8 --linear hide-covar --pheno fenotipo.txt --covar PCA_assoc.evec --out CIBER (fichero para gráfico Manhattan en R)

plink --bfile Cibersam_8 --linear hide-covar --adjust --pheno fenotipo.txt --covar PCA_assoc.evec --out CIBER (fichero para gráfico Q-Q en R)

- IMPUTACIÓN

plink --bfile Cibersam_8 --chr 1 --make-bed --out Cibersam_chr1

plink --bfile Cibersam_chr1 --recode --out Cibersam_chr1

plink --ped Cibersam_chr1.ped --map Cibersam_chr1.map --recode vcf --out Cibersam_chr1

bftools sort Cibersam_chr1.vcf -Oz -o Cibersam_chr1.vcf.gz

SCRIPT .r

- CONTROL DE CALIDAD
    - MISSINGNESS

lmiss <- read.table("missing.lmiss", header=T)

hist(x = lmiss$F_MISS, main = "SNPs missingness Histogram", xlab = "SNPs missingness", ylab = "Frecuency", col = "blue")

dev.copy(pdf, "SNPs_Missingness_Histogram.pdf")

dev.off()

imiss <- read.table("/missing.imiss", header=T)

hist(x = imiss$F_MISS, main = "Individual missingness Histogram", xlab = "Individual missingness", ylab = "Frecuency", col = "deepskyblue")

dev.copy(pdf, "Individual_Missingness_Histogram.pdf")

    - SEX DISCREPANCIES

sex <- read.table("sex.sexcheck", header=T)

hist(sex$F[which(sex$PEDSEX==1)], main = "Man Sex Check", xlab = "F", ylab = "Frecuency", col = "seagreen")

dev.copy(pdf, "Man_SexCheck_Histogram.pdf")

dev.off()

hist(sex$F[which(sex$PEDSEX==2)], main = "Fem Sex Check", xlab = "F", ylab = "Frecuency", col = "seagreen3")

dev.copy(pdf, "Fem_SexCheck_Histogram.pdf")

dev.off()

    - MINOR ALLELE FREQUENCY

maf <- read.table("maf_dist.frq", header=T)

hist(maf$MAF, main = "MAF DISTIBUTION HISTOGRAM", xlab = "MAF", ylab = "Frecuency", col = "firebrick3")

dev.copy(pdf, "MAF_Distribution_Histograma.pdf")

dev.off()

    - HARDY-WEINBERG EQUILIBRIUM

hwe <- read.table("hwe.hwe", header=T)

hist(hwe$P, main = "HWE Distribution Histogram", xlab = "HWE P", ylab = "Frecuency", col = "darkorange")

dev.copy(pdf, "HWEdistributionHistograma.pdf")

dev.off()

    - HETEROZYGOSITY

het <- read.table("het_check.het", header=T)

het$RATE <- (het$N.NM.-het$O.HOM.)/het$N.NM.

```
plot(het$RATE, main = "Heterozygosity Plot", xlab = "Frecuency", ylab = "HET RATE")

mean(het$RATE)

mean(het$RATE)+3*sd(het$RATE)

mean(het$RATE)-3*sd(het$RATE)

abline(h=mean(het$RATE)-3*sd(het$RATE), col="red")

abline(h=mean(het$RATE)+3*sd(het$RATE), col="red")

dev.copy(pdf, "HET_plot.pdf")

dev.off()

which(het$RATE< mean(het$RATE)-3*sd(het$RATE) | het$RATE> mean(het$RATE)+3*sd(het$RATE))

het[which(het$RATE< mean(het$RATE)-3*sd(het$RATE) | het$RATE>
mean(het$RATE)+3*sd(het$RATE)),][,c(1,2)]

HET <- het[which(het$RATE< mean(het$RATE)-3*sd(het$RATE) | het$RATE>
mean(het$RATE)+3*sd(het$RATE)),][,c(1,2)]

capture.output(HET, file = "/home/ubuntu/Desktop/SCZ_AAO/CIBER/QC/Fail_Het_ind.txt")

hist(het$RATE[which(het$RATE> mean(het$RATE)-3*sd(het$RATE) & het$RATE<
mean(het$RATE)+3*sd(het$RATE))], main = "HET without Overthreshold Histogram", xlab = "HET RATE",
ylab = "Frecuency", col = "palevioletred1")

dev.copy(pdf, "hist_HET.pdf")

dev.off()
```

- ESTUDIO DE POBLACIÓN

```
pca <- read.table("PCA_ciber.eigenvec", header=F)

colnames(pca) <- c("FID","IID", paste("PC",1:20, sep=""))

plot(pca$PC1, pca$PC2, col=as.factor(pcat$POPULATION), pch=19, main = "PCA Plot", xlab = "PC1", ylab
= "PC2")

legend(x = "topright", legend = c("CIBER", "EUR", "AFR", "ASN", "AMR"), fill = c("blue", "cyan", "black",
"green", "red"), title = "Population")

dev.copy(pdf, "plot_PCA.pdf")

dev.off()
```

- ESTUDIO DE ASOCIACIÓN
    - GRÁFICO DE MANHATTAN

```
man <- read.table("CIBER.assoc.linear", header=T)

manhattan(man, main = "Manhattan Plot", cex = 0.8, cex.axis = 0.8, col = c("blue4", "orange3"), annotatePval
= 0.00001, annotateTop = FALSE)

dev.copy(pdf, "Manhattan_plot.pdf")

dev.off()
```

    - GRÁFICO Q-Q

```
unadj <- read.table("CIBER.assoc.linear.adjusted", header=T)

qq(unadj$UNADJ, main = "QQ PLOT")

dev.copy(pdf, "QQ_plot.pdf")

dev.off()
```

ANEXO III. GRÁFICOS CONTROL DE CALIDAD DE LOS DATOS DEL CIBERSAM

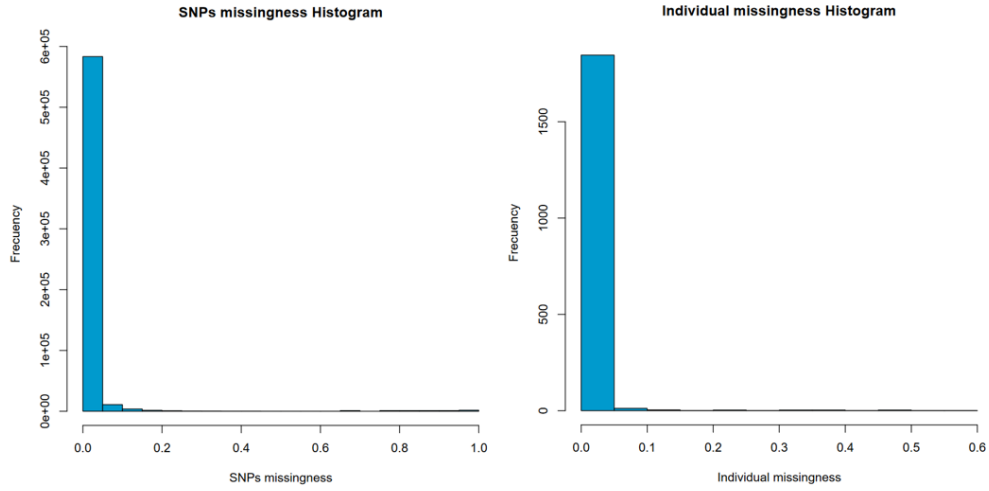


Figura 18. Histograma de pérdida de genotipado en SNPs y en individuos de los datos del CIBERSAM.

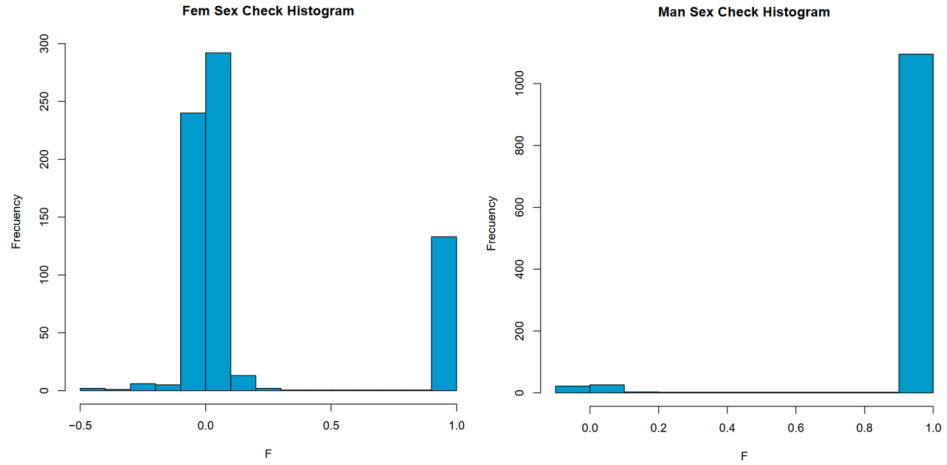


Figura 19. Histograma discrepancias en el sexo femenino y masculino en los datos del CIBERSAM.

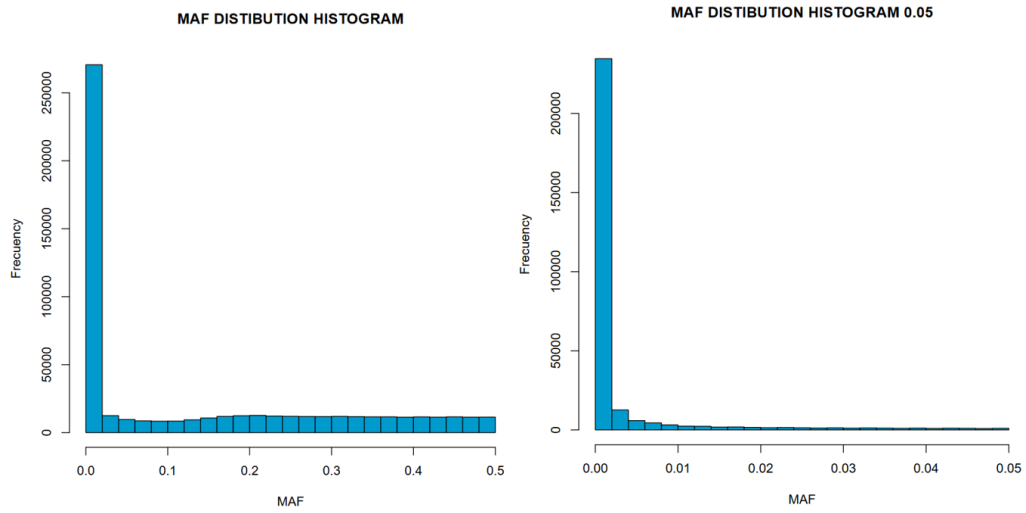


Figura 20. Histograma distribución MAF antes y después de aplicar el filtro < 0.05 en los datos del CIBERSAM

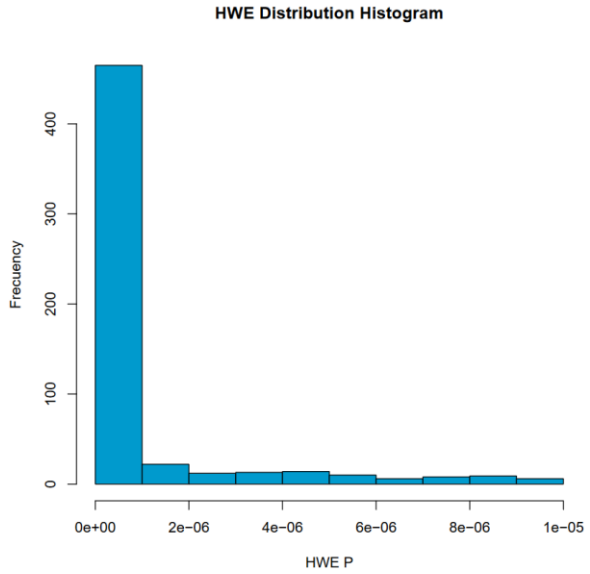


Figura 21. Histograma de distribución del HWE de los datos del CIBERSAM.

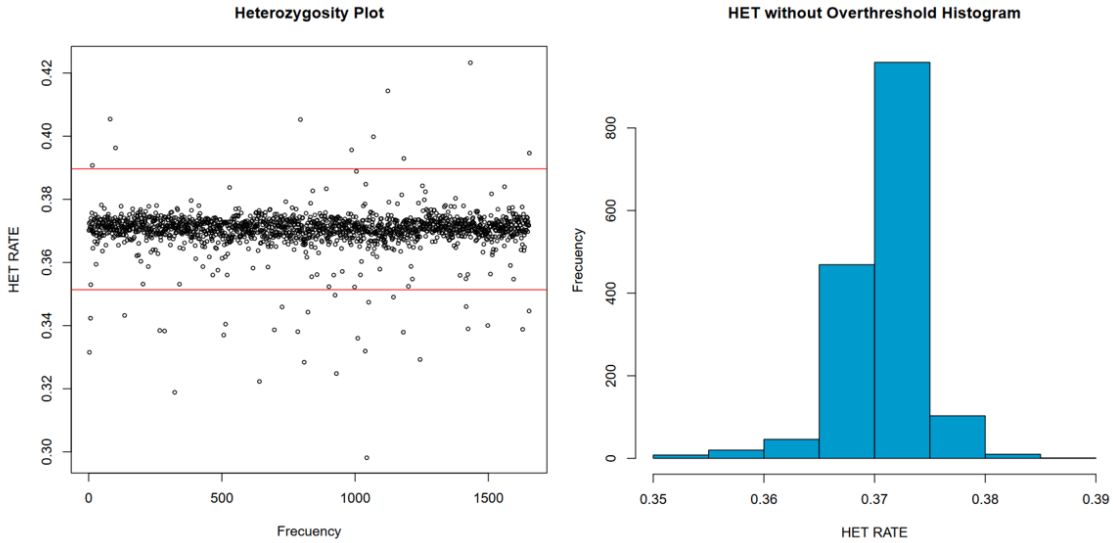


Figura 22. Gráfico e histograma de la heterocigosidad de los datos del CIBERSAM.

ANEXO IV. GRÁFICOS CONTROL DE CALIDAD DE LOS DATOS DEL KFO

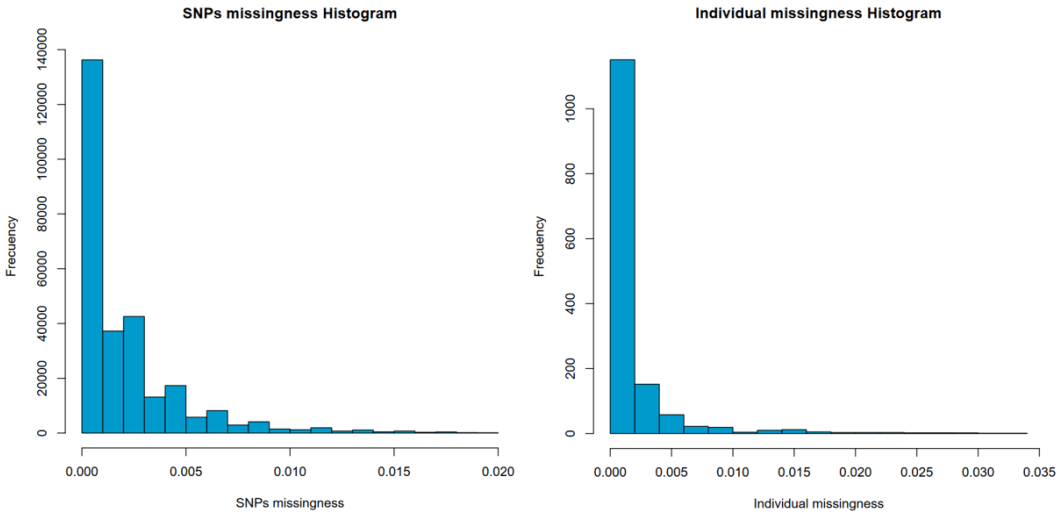


Figura 23. Histograma de pérdida de genotipado en SNPs y en individuos de los datos del KFO.

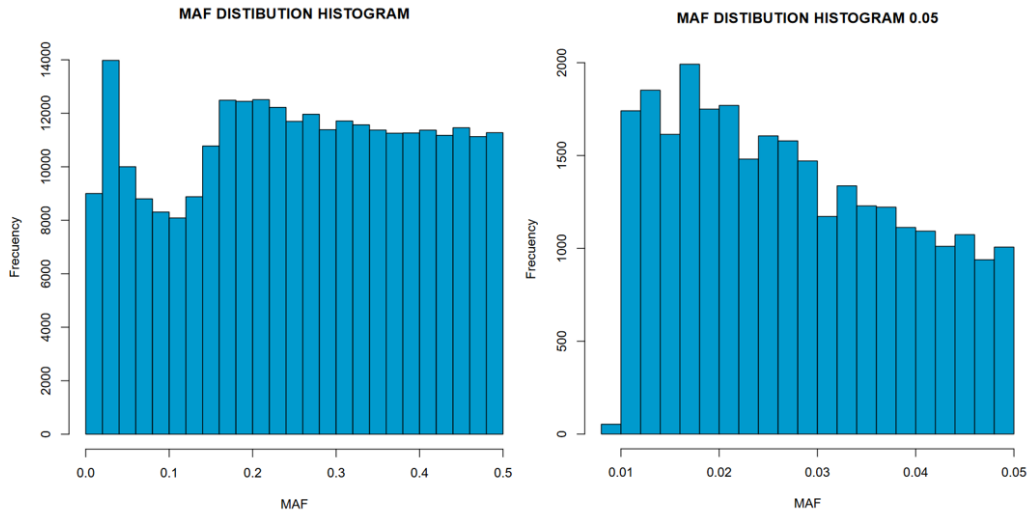


Figura 24. Histograma distribución MAF antes y después de aplicar el filtro < 0.05 en los datos del KFO.

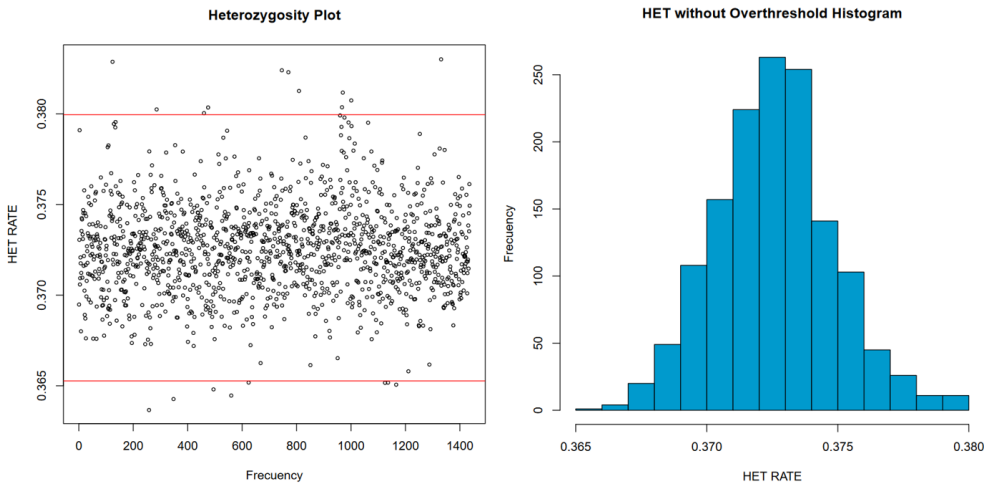


Figura 25. Gráfico e histograma de la heterocigosidad de los datos del KFO.

ANEXO V. GRÁFICOS CONTROL DE CALIDAD DEL CIBERSAM IMPUTADO CON EL HRC

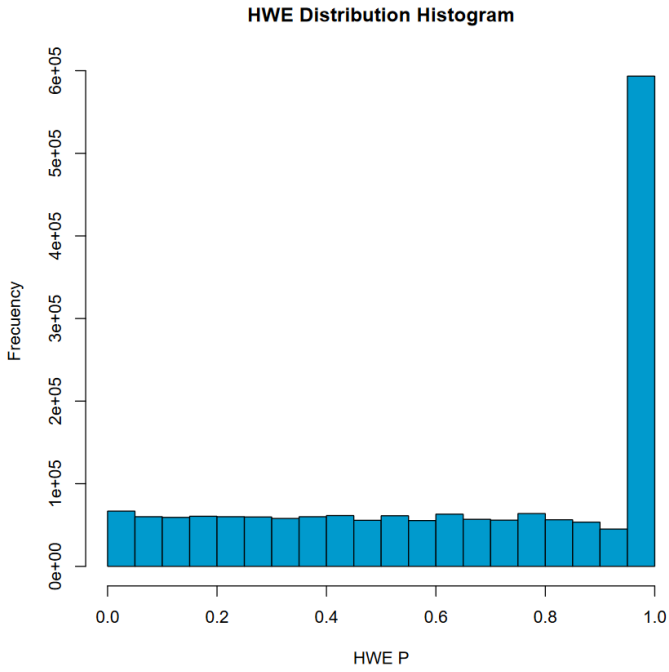


Figura 26. Histograma de distribución del HWE de los datos del CIBERSAM imputados con el panel HRC.

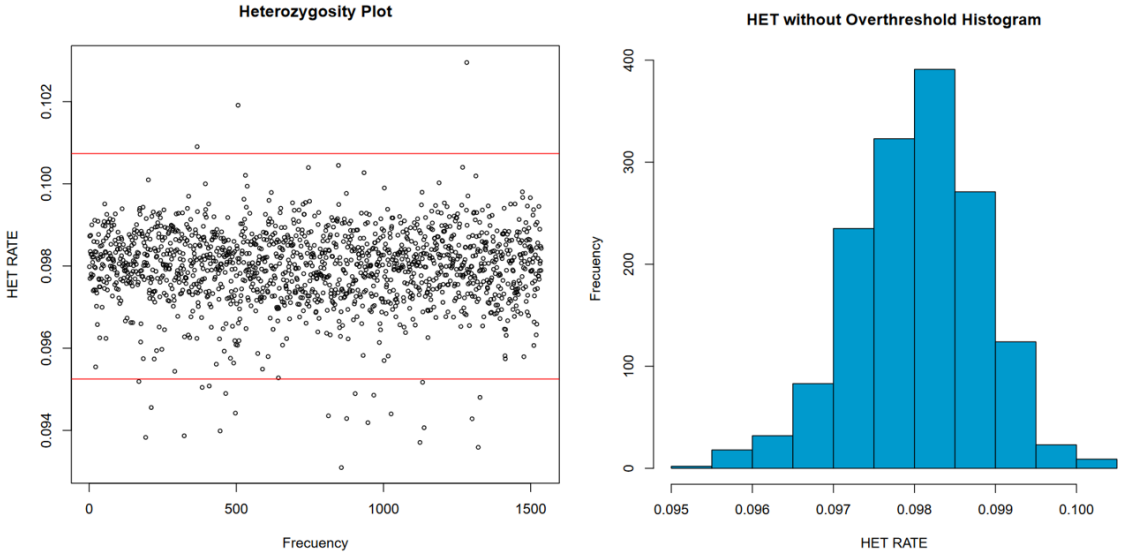


Figura 27. Gráfico e histograma de la heterocigosidad de los datos del CIBERSAM imputados con el panel HRC.

ANEXO VI. GRÁFICOS CONTROL DE CALIDAD DEL KFO IMPUTADO CON EL PANEL HRC

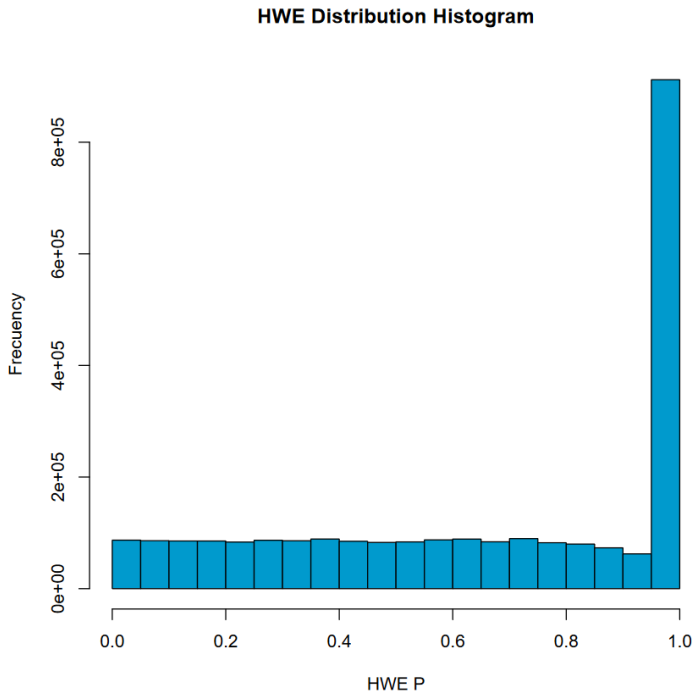


Figura 28. Histograma de distribución del HWE de los datos del KFO imputados con el panel HRC.

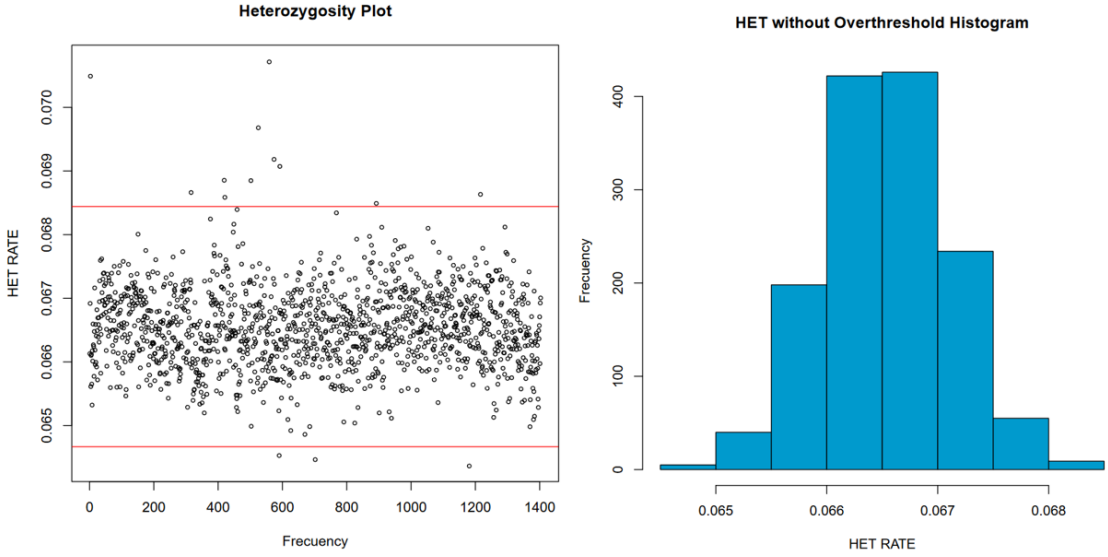


Figura 29. Gráfico e histograma de la heterocigosidad de los datos del KFO imputados con el panel HRC.

ANEXO VII. GRÁFICOS CONTROL DE CALIDAD DEL CIBERSAM Y EL KFO AGRUPADOS E IMPUTADOS CON EL PANEL HRC

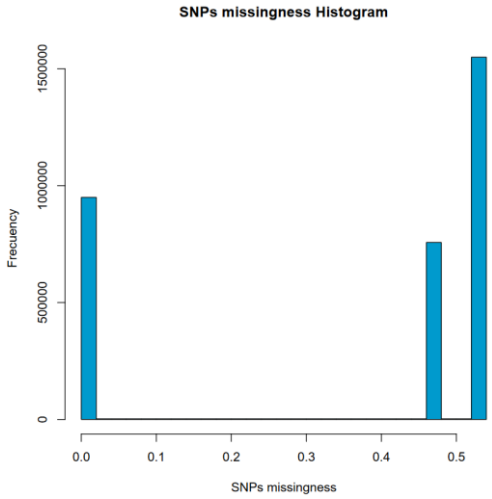


Figura 30. Histograma de pérdida de genotipado en SNPs de los datos del CIBERSAM y del KFO agrupados e imputados con el panerl HRC.

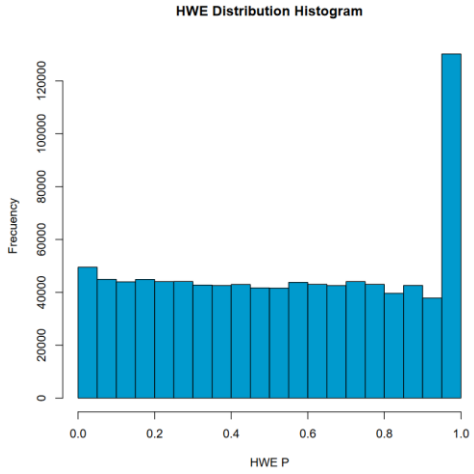


Figura 31. Histograma de distribución del HWE de los datos del CIBERSAM y del KFO agrupados e imputados con el panel HRC.

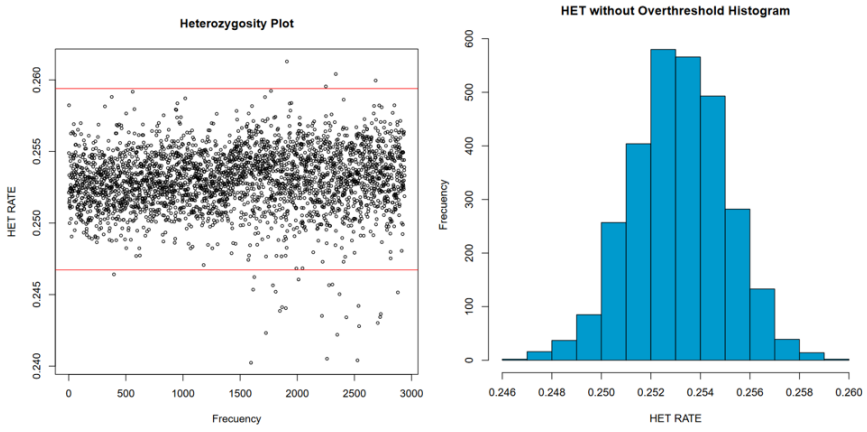


Figura 32. Gráfico e histograma de la heterocigosidad de los datos del CIBERSAM y del KFO agrupados e imputados con el panel HRC.

ANEXO VIII. GRÁFICOS CONTROL DE CALIDAD DEL CIBERSAM IMPUTADO CON EL TOPMED

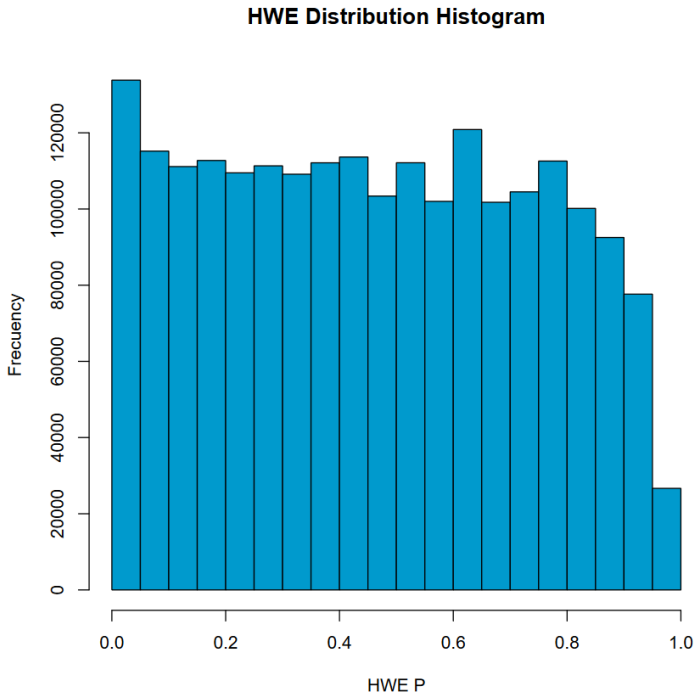


Figura 33. Histograma de distribución del HWE de los datos del CIBERSAM imputados con el panel TOPMed.

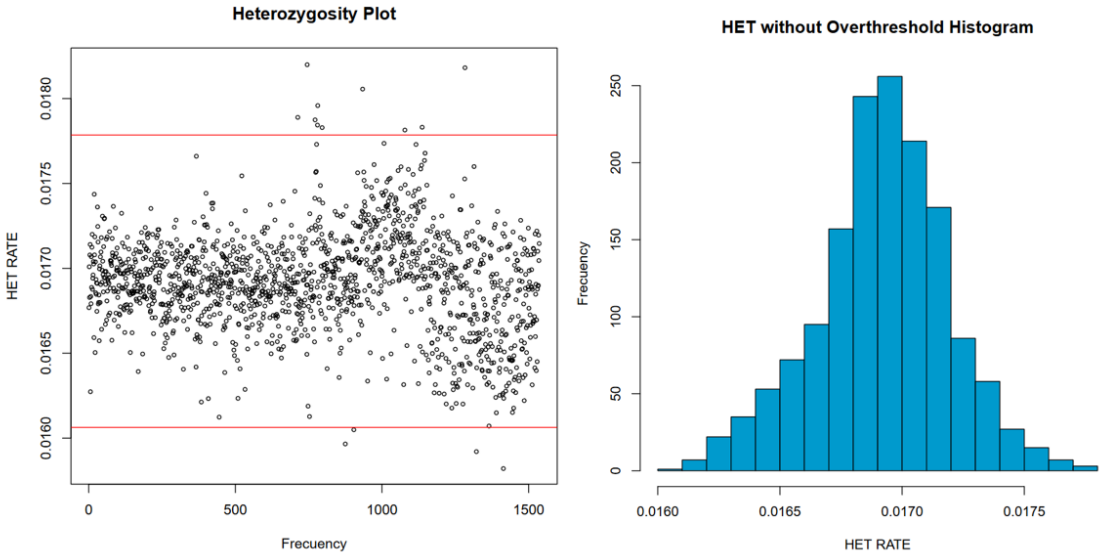


Figura 34. Gráfico e histograma de la heterocigosidad de los datos del CIBERSAM imputados con el panel TOPMed.

ANEXO IX. GRÁFICOS CONTROL DE CALIDAD DEL KFO IMPUTADO CON EL TOPMED

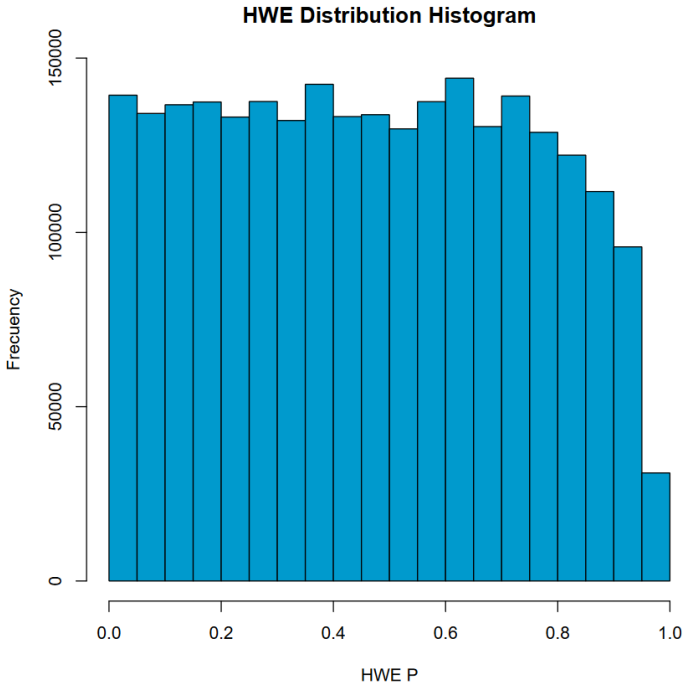


Figura 35. Histograma de distribución del HWE de los datos del KFO imputados con el panel TOPMed.

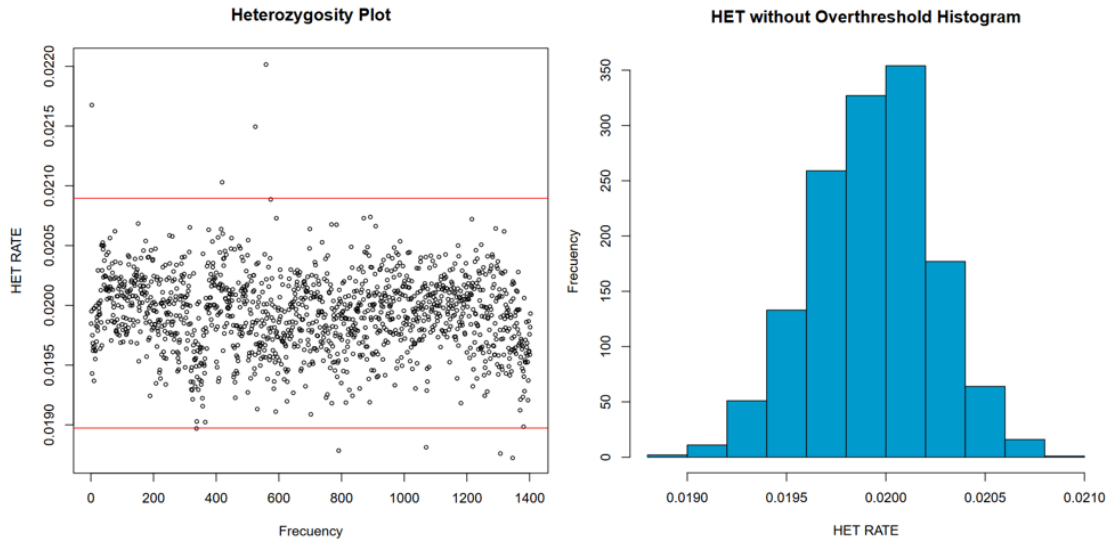


Figura 36. Gráfico e histograma de la heterocigosidad de los datos del KFO imputados con el panel TOPMed.

ANEXO X. GRÁFICOS CONTROL DE CALIDAD DEL CIBERSAM Y DEL KFO IMPUTADOS CON EL TOPMED

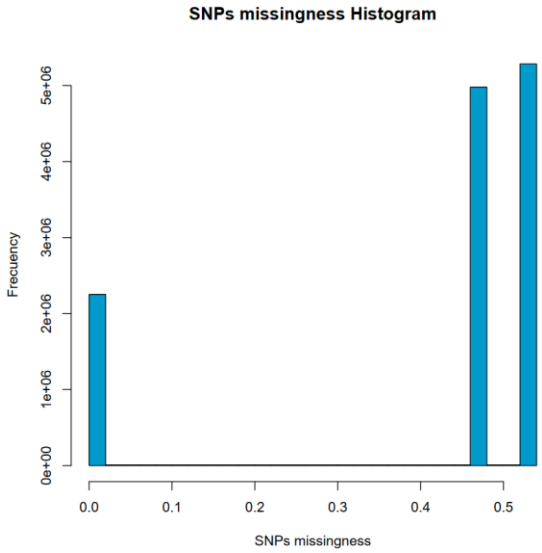


Figura 37. Histograma de pérdida de genotipado en SNPs de los datos del CIBERSAM y del KFO agrupados e imputados con el panel TOPMed.

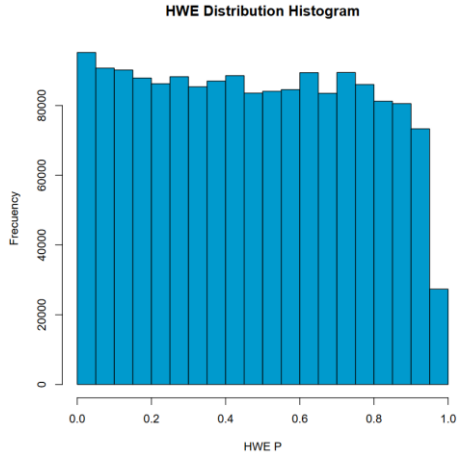


Figura 38. Histograma de distribución del HWE de los datos del CIBERSAM y del KFO agrupados e imputados con el panel TOPMed.

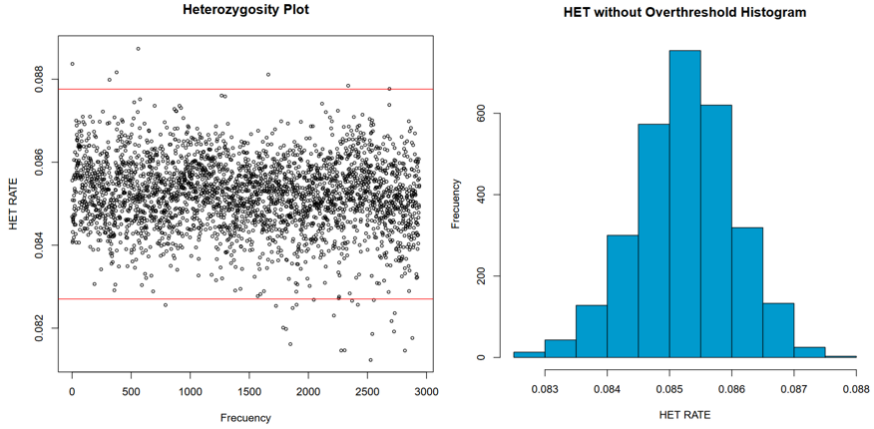


Figura 39. Gráfico e histograma de la heterocigosidad de los datos del CIBERSAM y del KFO agrupados e imputados con el panel TOPMed.