ESCUELA DE DOCTORADO INTERNACIONAL DE LA USC

María Soledad
Otero Piñeiro

Tesis doctoral

"GENERACIÓN DE UN REPOSITORIO DE DATOS DE GENOTIPADO DE MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA Y DISEÑO DE UN NUEVO *ARRAY* PARA ESTUDIOS DE GWAS"

Santiago de Compostela, 2023

Programa de doctorado en Medicina Molecular

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

TESIS DOCTORAL

"GENERACIÓN DE UN REPOSITORIO DE DATOS DE GENOTIPADO DE MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA Y DISEÑO DE UN NUEVO *ARRAY* PARA ESTUDIOS DE GWAS"

Memoria para optar al grado de Doctor

Presentada por:

María Soledad Otero Piñeiro

Directores:

Dr. Ángel M. Carracedo Álvarez

ESCOLA DE DOUTORAMENTO INTERNACIONAL

PROGRAMA DE DOUTORAMENTO EN MEDICINA MOLECULAR

SANTIAGO DE COMPOSTELA, JULIO 2023

DECLARACIÓN DEL AUTOR DE LA TESIS
"GENERACIÓN DE UN REPOSITORIO DE DATOS DE GENOTIPADO DE MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA Y DISEÑO DE UN NUEVO *ARRAY* PARA ESTUDIOS DE GWAS"

Dña. Mª Soledad Otero Piñeiro

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

1) La tesis abarca los resultados de elaboración de mi trabajo.
2) En su caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3) La tesis es la versión definitiva presentada para su defensa y coincide con la versión enviada en formato electrónico.
4) Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.
5) Yo, María Soledad Otero Piñeiro, con DNI 36150620W, declaro que la presente tesis doctoral no presenta conflicto de interés alguno.

En Santiago de Compostela, 13 de julio de 2023

Fdo.: Mª Soledad Otero Piñeiro

AUTORIZACIÓN DEL DIRECTOR / TUTOR DE LA TESIS

"GENERACIÓN DE UN REPOSITORIO DE DATOS DE GENOTIPADO DE MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA Y DISEÑO DE UN NUEVO *ARRAY* PARA ESTUDIOS DE GWAS""

El Profesor Doctor D. Ángel María Carracedo Álvarez

INFORMA/N:

Que la presente tesis, se corresponde con el trabajo realizado por Dña. Mª Soledad otero Piñeiro, bajo mi dirección, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En Santiago de Compostela, 13 de julio de 2023

Fdo.: Prof. Dr. Ángel Carracedo Álvarez

A mis padres y a mi hermana

FINANCIACIÓN

Los trabajos aquí presentados han sido financiados con fondos del Centro Nacional de Genotipado - Plataforma de Recursos Biomoleculares y Bioinformáticos - IPT13/0001 - ISCIII-SGEFI / FEDER - del Instituto de Salud Carlos III.

AGRADECIMIENTOS

AGRADECIMIENTOS

"No me gusta la deshonestidad, porque es perversa y representa lo contrario de la ciencia"
Ángel Carracedo

Lo cierto es que no me ha resultado nada fácil llegar a terminar este trabajo. Incluso llegué a pensar que nunca lo lograría. Siempre quise dedicarme a la investigación, justo por este afán innato de querer hacer el bien y convertir el mundo en un lugar mejor…Finalmente puedo decir, que parte de este objetivo, que he tenido desde siempre, se ha cumplido. Y no por el hecho de obtener un título, ni por haber dedicado mi tiempo a alcanzar una meta que, sin embargo, he descubierto que no tiene fin…pero sí porque en el camino me he topado con grandes personas que han hecho que mis días fuesen, de una u otra manera, mucho más plenos y enriquecedores, ya que siempre he defendido que lo único hermoso de la vida son las relaciones que se van forjando en cada paso que damos. Por tanto, y aprovechando el espacio que este documento me permite para mostrar mi gratitud, debo decir que no existen palabras para definir todo lo que le debo a mi director y jefe durante los últimos años, Ángel Carracedo, que más allá de haberme apoyado en el trabajo, me ha ayudado en todos los aspectos que en la vida de una persona puedan suceder. GRACIAS Ángel, por permitirme pertenecer a tu gente y por tu constante lucha para sacarnos adelante…

TODA mi admiración y gratitud a grandes profesionales que SIEMPRE han estado a mi lado, haciéndome CRECER: gracias, Inés Quintela, Olalla Maroñas, Alicia de Coo…y gracias a mi queridísima Raquel Cruz…gracias por todo lo que he aprendido de vosotras y por el tiempo que me habéis dedicado. GRACIAS.

Y en todo este tiempo en el que hasta hemos estado viviendo una pandemia (mejor no profundizar en este tema…), y en el que incluso me he roto un pie (en este sí voy a profundizar…), que me obligó a ser muy dependiente durante muchos meses, he recordado con cariño a todos a los que aquí nombro y que habéis estado pendientes tanto de mi evolución como de mi estado de ánimo en

todo momento. Por tanto, gracias infinitas a la vida por haber puesto en mi camino a Miriam Álvarez, Miriam Pérez, Paula, Elena, Inés B. y a Rosanna; habéis sido un precioso e inesperado regalo.

Gracias infinitas a María Torres, por su paciencia, ayuda, por sus conocimientos, por su experiencia, por cómo me quiere de esa forma distinta e inigualable :). Gracias a Juan simplemente por ser cómo es, por su forma de transmitir paz, y sobre todo por aportar SIEMPRE soluciones a cualquier problema que pueda surgir. Gracias a Bea Sobrino por estar siempre dispuesta a ayudar en cualquier cosa. Por su presencia en mis buenos y malos momentos…

Gracias a todos los compañeros que han estado a mi lado en el día a día trabajando codo con codo en el Centro de Genotipado: Ángela, Fátima, Ana Pastoriza, Shaila, Andrea, María, Nare, Christian, Galina y Joja, compañero de cursos y viajes que me ha hecho sentir como en casa cuando no lo estaba.

Gracias a todos mis compañeros, que de una u otra manera habéis hecho que mi vida tuviese más horas libres y gracias por sacarme siempre una sonrisa.

Gracias a mis amigas "las Bárbaras", como no a Ricardo, Raquel, Chiny, Pecke…a Sandra "mi teacher", a Rubén por darme "el último empujón" y por supuesto a mis incondicionales Ricoy, Pepe y Jose, AMIGOS CON MAYÚSCULAS.

Gracias a todos los que me habéis regalado sonrisas durante todo este tiempo, el bien más preciado por esta humilde servidora, y gracias también a todos aquellos que habéis tenido que soportarme durante estos últimos años.

No me llegarían las páginas de este trabajo para nombraros a todos, así que, gracias y gracias infinitas a todos mis acompañantes de vida, y especialmente, a los que han sufrido mis ausencias…

A Javi, por estar siempre a mi lado y, sobre todo, por su apoyo en los días más difíciles…

A mi hermana Anita. Por ser como eres, por inspirarme a hacer muchas cosas, por cuidarme y amarme tanto. Por recargarme de energía y valor y ser mi ejemplo a seguir.

A todos ¡INFINITAS GRACIAS!

ÍNDICE DE CONTENIDOS

ÍNDICE DE CONTENIDOS

ABREVIATURAS, ACRÓNIMOS Y SIGLAS

ABREVIATURAS, ACRÓNIMOS Y SIGLAS

ADGC:	*Alzheimer's Disease Genetics Consortium*
ADME:	*Absorption, Distribution, Metabolism and Excretion*
ADN:	Ácido Desoxirribonucleico
ADRs:	*Adverse Drug Reactions*
AGCC:	*Software ThermoFisher GeneChip Command Console*
AIMs:	*Ancestry-Informative Markers*
ARN:	Ácido Ribonucleico
ARNm:	Ácido Ribonucleico mensajero
BCAC:	*Breast Cancer Association Consortium*
B-CAST:	*Breast CAncer Stratification*
BIC:	Criterio de Información Bayesiano
BNADN:	Banco Nacional de ADN
BOADICEA:	*Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm*
BRIDGES:	*Breast cancer RIsk after Diagnostic Gene Sequencing*
CARDIoGRAM:	*Coronary Artery Disease Genome-wide Replication and Meta-Analysis*
CCR:	Cancer colorectal
CD/CV:	*Common disease / common variant*
CDRVH:	*Rare variant / common disease hypothesis*
CeGen:	Centro Nacional de Genotipado
CEPH:	*Centre D`étude du Polymorphisme Humain*
CHR:	Cromosoma
CIBERER:	Centro de Investigación Biomédica en Red de Enfermedades Raras
CIBERSAM:	Centro de Investigaciones Biomédicas en Red en Salud Mental

CM:	Cancer mama
CNVs:	*Copy-Number Variations*
CO:	Cancer ovario
CP:	Cancer próstata
CRISPR:	*Clustered Regularly Interspaced Short Palindromic Repeats*
CSVS:	*Collaborative Spanish Variant Server*
DA:	*Discriminant Analysis*
DAPC:	*Discriminant Analysis of Principal Components*
dbGaP:	Base de Datos de Genotipo y Fenotipo
dbSNP:	*Single Nucleotide Polymorphism Database*
DFs:	Funciones discriminantes
DL:	Desequilibrio de Ligamiento
DOE:	*Departament of Energy*
DQC:	*Dish Quality Control*
EBI:	*European Bioinformatics Institute*
EC:	Enfermedad Coronaria
EC/VC:	Enfermedad común / Variante común
EDTA:	Ácido etilendiaminotetraacético
EGF:	*Ethics and Governance Framework*
EGG:	*Early Growth Genetics*
EHW:	Equilibrio *Hardy-Weinberg*
ELA:	Esclerosis Lateral Amiotrófica
EMBL-EBI:	*European Molecular Biology Laboratory-European Bioinformatics Institute*
ENCODE:	*Encyclopedia of DNA Elements*
ENoD:	Enfermedades No Diagnosticadas
EPIC:	*European Prospective Investigation in Cancer and Nutrition*
EUR:	Europa

ExAC:	*Exome Aggregation Consortium*
FGF:	Factor de Crecimiento de Fibroblastos
FIISC:	Fundación Canaria Instituto de Investigación Sanitaria de Canarias
FIMIM:	*Fundació Institut Mar d'Investigacions Mèdiques*
FPGMX:	Fundación Pública Galega de Medicina Xenómica
FUNCANIS:	Fundación Canaria de Investigación Sanitaria
GAIN:	*Genetic Association Information Network*
GATK:	*Genome Analysis Toolkit*
GCTA:	*Genome-wide Complex Trait Analysis*
Gr@ce:	*Genomic Research at Fundació ACE*
GTC:	*ThermoFisher Genotyping Console*
GTEx:	*The Genotype-Tissue Expression*
GWAS:	*Genome-Wide Association Study/ies*
$H_{0:}$	Hipótesis inicial
HGDP:	*The Human Genome Diversity Project*
HGDP-CEPH:	*Human Genome Diversity Cell Line Panel*
HGP:	*Human Genome Project*
HGVS:	*Human Genome Variation Society*
HLA:	*Human Leukocyte Antigen*
HUGO:	*Human Genome Organization*
HVID:	*HumDiv-trained*
HVP:	*Human Variome Project*
IBD:	*Identity by Descent*
IBS:	*Iberian Population in Spain*
ICD-10:	*The International Classification of Diseases*
IDC:	Miocardiopatía Idiopática
IDIBELL:	Fundació Privada Institut d'Investigació Biomèdica de Bellvitge

IDIVAL:	Fundación Instituto de Investigación Marqués de Valdecilla
IGV:	*Integrative Genomics Viewer*
IGAP:	*International Genomics of Alzheimer's Project*
IIBB:	Instituto de Investigaciones Biomédicas de Barcelona
IL:	*interleukin*
IISPV:	Institut d'Investigació Sanitària Pere Virgili
lncRNA:	*Long non-coding RNA*
InDels:	Inserciones / Deleciones
IPATIMUP:	*Institute of Molecular Pathology and Immunology of the University of Porto*
IRBLleida:	Institut de Recerca Biomèdica de Lleida
ISC:	*International Schizophrenia Consortium*
ITER:	Instituto Tecnológico y de Energías Renovables
Kb:	Kilobase
KIR:	*Killer Immunoglobulin-like Receptor*
Kbp:	Kilopares de bases
LCLs:	líneas celulares de linfoblastoides cultivados
LD:	*Linear Discriminant*
LOH:	pérdida de heterozigosidad
LSDB:	*Locus Specific Mutation Databases*
MAF:	*Minor Allele Frequency*
MC:	*GeneTitan Multi-Channel*
MDS:	*Multidimensional Scaling*
MGS:	*Molecular Genetics of Schizophrenia*
ml:	Mililitro
MLP:	*Multiple loci probes*
miARN:	microARN
mM:	milimolar
MP:	*Manhattan Plot*

MS:	Esclerosis múltiple
µl:	Microlitro
N:	Tamaño muestral
NCBI:	*National Center for Biotechnology Information*
ng:	Nanogramo
NGS:	*Next-generation sequencing*
NHGRI:	*National Human Genome Research Institute*
NHLBI:	*National Heart Lung and Blood Institute*
NIH:	*National Institute of Health*
NRF:	Factores respiratorios nucleares
OMIM:	*Online Mendelian Inheritance in Man®*
OTV:	*off-target variants*
PAMP:	Patrones Moleculares Asociados a Patógenos
Pb:	Pares de bases
PCA:	*Principal component analysis*
PCR:	*Polymerase Chain Reaction*
PCs:	Componentes principales
PEP:	Aplicación a un modelo predictivo en primeros episodios psicóticos
Polyphen-2:	*Polymorphism phenotyping*
POP:	Puntos geográficos
PKC:	Proteína quinasa C
PRB^2:	Plataforma en red de Recursos Biomoleculares y Bioinformáticos
PRS:	*Polygenic Risk Score*
QC:	*Quality control*
QTNS:	*Quantitative trail nucleotides*
RFLP:	*Restriction Fragment Length Polymorphism*
SBA:	*Spain Biobank Array*

SCOURGE:	*Spanish COalition to Unlock Research on host GEnetics on* COVID-19
SIFT:	*Sorting Tolerant from Intolerant*
SLP:	S*ingle locus probes*
SNPs:	*Single-Nucleotide Polimorphisms*
STRs:	*Short Tandem Repeats*
SWG of PGC:	*Schizophrenia Working Group of the Psychiatric Genomics Consortium*
T1D:	Diabetes tipo 1
T2D:	Diabetes tipo 2
TE:	Tris EDTA (ácido etilendiaminotetraacético)
TOC:	Trastorno Obsesivo Compulsivo
UCSC:	*University of California Santa Cruz*
UDIGEN:	Desarrollo de una Unidad de Diagnóstico Genómico
UK:	*United Kingdom*
UKBB:	*United Kingdom Biobank Array*
UNED:	Universidad Nacional de Educación a Distancia
USC:	Universidade de Santiago de Compostela
UTR:	*untraslated region*
VIH:	Virus de la Inmunodeficiencia Humana
VNTR:	*Variable Number of Tandem Repeats*
WASP:	Proteína del síndrome de *Wiskott-Aldrich*, dominio de homología-2
WTCCC:	*The Wellcome Trust Case Control Consortium*

RESUMEN

Con la finalización del Proyecto Genoma Humano en 2003 y el Proyecto Internacional HapMap en 2005, los investigadores disponemos de herramientas de investigación que permiten analizar las contribuciones genéticas a enfermedades complejas. Las herramientas incluyen bases de datos computarizadas que contienen la secuencia de referencia del genoma humano, un mapa de la variación genética humana y un conjunto de nuevas tecnologías que permiten analizar de manera rápida y precisa las variaciones genéticas a lo largo de todo el genoma para estudiar su asociación con el riesgo de padecer una determinada enfermedad. Precisamente, debido a la importancia y a la evolución de los estudios de asociación de genoma completo o GWAS (del inglés *Genome-Wide Association Study*) en los últimos años, se ha visto necesaria la disponibilidad de datos genéticos pertenecientes a personas sanas, para ser utilizados como controles y así poder comparar la frecuencia relativa de los polimorfismos o variantes genéticas en ambos grupos e identificar variantes asociadas con la enfermedad analizada. Además de permitir entender los mecanismos de enfermedad y estratificarla en grupos causales, la información derivada de los GWAS permite la estima adecuada del riesgo genético que empieza a ser incorporada a los parámetros clínicos para una determinación más precisa del riesgo de enfermedad.

Una de las posibles complicaciones de los estudios de asociación es la existencia de estratificación, diferencias genético-poblacionales entre casos y controles que pueden sesgar la interpretación de los resultados. Logros recientes de proyectos genómicos a gran escala, como el Proyecto 1000 Genomas (en adelante Proyecto 1000G), evidencian la existencia de diferencias en frecuencia alélica entre poblaciones, apuntando la necesidad de utilizar poblaciones locales como control.

Por otra parte, los GWAS tradicionalmente se han centrado en el análisis de variantes comunes. Sin embargo, los cada vez más numerosos estudios de secuenciación y *arrays* de zonas codificantes, han volcado la atención sobre las variantes raras, donde se espera resida parte de la "heredabilidad perdida", la base genética no identificada hasta el momento en los estudios de asociación de muchas enfermedades. Las diferencias en frecuencia de estas variantes raras son aún mayores, siendo muchas incluso específicas de población, por lo que es fundamental determinar las frecuencias poblacionales de estas variantes para distinguir asociaciones reales de polimorfismos específicos de población.

Con esta tesis se ha pretendido generar un repositorio de datos de genotipado de muestras de población control española que se harán accesibles a la comunidad científica para sus estudios de asociación del genoma completo. Para ello hemos genotipado 3.169 muestras procedentes del Banco Nacional de ADN (BNADN) con un panel de genoma completo y de alta densidad que incluye tanto variantes comunes como variantes de baja frecuencia y ha sido específicamente diseñado para cubrir no solo variación común, si no también variación rara funcional específica de población española.

El análisis de los datos obtenidos permitirá caracterizar en detalle la población española y analizar incluso la existencia de variabilidad local a escala microgeográfica, así como

caracterizar los diferentes patrones de estratificación poblacional en cuanto a variantes comunes y raras, parámetros que deben ser evaluados para el estudio de trastornos genéticos complejos.

Palabras clave: *Genome-Wide Association Study/ies, casos y controles, variantes comunes y raras, panel de alta densidad, estratificación poblacional*.

INTRODUCCIÓN

1 INTRODUCCIÓN

1.1 LOS ESTUDIOS DE ASOCIACIÓN DE GENOMA COMPLETO (GWAS)

1.1.1 Descripción y fundamento de los GWAS. Aspectos básicos

Un estudio de asociación de todo el genoma es un enfoque que consiste en analizar marcadores genéticos de todo el genoma de muchas personas para encontrar variaciones genéticas asociadas con una enfermedad y/o un rasgo en particular (Figura 1.1). Una vez que se identifican nuevas asociaciones genéticas, los investigadores disponen de información para desarrollar mejores estrategias para detectar, tratar y prevenir la enfermedad. Dichos estudios son particularmente útiles para encontrar variaciones genéticas que contribuyen a las enfermedades comunes y complejas, como el asma, el cáncer, la diabetes, las enfermedades cardíacas o los trastornos psiquiátricos.

Los estudios de asociación de genoma completo (en inglés *Genome-Wide Association Studies*, GWAS) utilizan tecnologías de genotipado de alto rendimiento para analizar cientos de miles de polimorfismos de un solo nucleótido (SNPs, *Single Nucleotide Polymorphisms*), biomarcadores que se describen en el apartado 1.3 de este trabajo, y estos se relacionan con las condiciones clínicas o con rasgos medibles.

Los GWAS están basados en la hipótesis enfermedad común-variante común (EC/VC), que sostiene que las variantes genéticas comunes, presentes en la población general, pueden explicar gran parte de la heredabilidad de las enfermedades, ya que su componente genético es, en su mayoría, variación común y no variación rara funcional.

La realización de estos estudios es posible gracias a los avances en las tecnologías de genotipado, al conocimiento de la secuencia completa del genoma humano, al registro de millones de SNPs en bases de datos públicas y al Proyecto Internacional HapMap (International HapMap, 2003). Este último nos permite conocer los patrones de desequilibrio de ligamiento (DL) a lo largo del genoma y, por tanto, caracterizar la diversidad haplotípica existente en las distintas poblaciones humanas (Hirschhorn & Daly, 2005). Además, nos permitió identificar los Tag-SNPs que capturan toda la variabilidad de cada bloque de ligamiento.

Se define el DL como la correlación entre dos variantes próximas, de tal forma que, los alelos de estos marcadores vecinos (observados en el mismo cromosoma) están asociados en una población con más frecuencia que la esperada en función de sus frecuencias individuales (Hirschhorn & Daly, 2005). El DL surge con la aparición de una nueva mutación, y mientras no se produzca recombinación, estará en DL con el resto de las variantes del cromosoma, formando así un nuevo haplotipo (combinación de variantes alélicas de un cromosoma). El DL puede variar dentro de una misma población o entre poblaciones como consecuencia de la variabilidad regional en los patrones de recombinación, deriva genética, selección natural, conversión génica, edad de la mutación, diversidad étnica y mezcla reciente de la población, tasa de mutación o a los patrones de emparejamiento en la población (Carlson, Eberle, Kruglyak, & Nickerson, 2004), por lo que en consecuencia va desapareciendo a lo largo del tiempo.

El DL se cuantifica mediante los parámetros D' (nivel de asociación alélica), LOD *score*[1] y R^2 (coeficiente de correlación). D' mide el nivel de heredabilidad conjunta de los dos polimorfismos y determina la región entorno al gen que se transmite haplotípicamente junto con él. R^2 mide el nivel de redundancia estadística que supone analizar dos SNPs y se utiliza para seleccionar en esa región un subconjunto de SNPs suficiente para englobar toda la variabilidad haplotípica. Los valores de ambos parámetros oscilan entre -1 y 1, siendo 0 cuando no existe asociación (Jonathan L. Haines, 2005).

Los estudios de GWA necesitan, habitualmente, un gran número de muestras para conseguir una potencia estadística suficiente para definir variantes con riesgo relativo bajo y con significancia a nivel genómico. De hecho, el valor-p propuesto para estos estudios es de $5x10^{-8}$, después de realizar la corrección de Bonferroni para comparaciones múltiples (N. Risch & Merikangas, 1996).

Lo que se pretende es que los hallazgos que proporcionan los estudios de GWA se apliquen de forma activa en la prevención y el tratamiento a través del uso de estimas de riesgo poligénico, ya implantados en proyectos piloto de muchos países. También son importantes para estratificar la enfermedad y progresar en Medicina Personalizada, para entender los mecanismos de acción, es decir la etiopatogenia de los procesos fisiológicos o patológicos y para descubrir nuevas dianas de fármacos. Además, la mayor parte de los biomarcadores de respuesta a fármacos han sido descubiertos por estudios de GWAS.

Está claro que, en la contribución genética a la enfermedad, la variación rara funcional ejerce un riesgo relativo individualmente más alto (Sudmant et al., 2015) pero en la casi totalidad de las enfermedades comunes la mayor parte de la contribución genética a la heredabilidad es variación común.

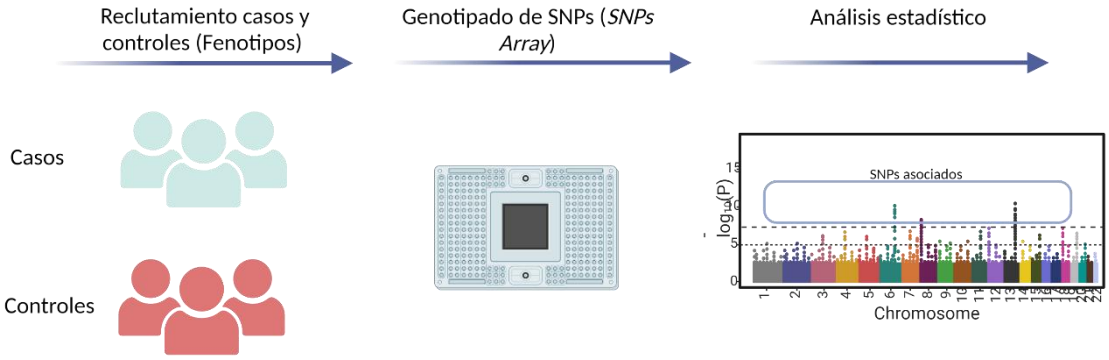


Figura 1.1. Estructura de un GWAS. Consiste en tres pasos: reclutamiento de casos y controles en función del fenotipo, genotipado de SNPs de cada individuo y análisis estadístico. Adaptado de *"Ikegawa S. A short. history of the genome-wide association study: where we were and where we are going. Genomics Inform. 2012;10(4):220-225. doi:10.5808/GI.2012.10.4.220"*. (Ikegawa, 2012). Realizado con www.Biorender.com.

[1] El término LOD *score* hace referencia al logaritmo en base 10 del cociente de verosimilitudes (en inglés, "*odds ratio*") entre la hipótesis de ligamiento de dos *loci* (para una determinada frecuencia de recombinación) y la hipótesis nula (no ligamiento). Se calcula estudiando la herencia de una enfermedad y de un marcador (o de dos marcadores) en un pedigrí y determinando el número de individuos recombinantes para los dos *loci*. En humanos, una puntuación LOD mayor a 3 indica ligamiento entre dos *loci* con una fiabilidad del 95% (https://glosarios.servidor-alicante.com/genetica/).

1.1.2 Beneficios de los GWAS

El impacto en la Medicina de los GWAS está siendo muy importante, pero lo va a ser más en el futuro. Así, se están sentando las bases para la era de la Medicina Personalizada, en la que el enfoque actual de "atención generalizada" dará paso a estrategias más personalizadas al permitir la estratificación de la enfermedad. En el futuro, y a medida que se evalúe la utilidad clínica, los profesionales de la salud podremos utilizar dichas herramientas para proporcionar a los pacientes información individualizada sobre sus riesgos de desarrollar ciertas enfermedades.

Esta información específica permitirá a los sistemas de salud adaptar los programas de prevención a la composición genética de cada persona a través de la incorporación de las estimas de riesgo poligénico al cálculo clínico del riesgo.

La comunidad científica ya ha demostrado avances usando esta nueva estrategia. Así se han reportado numerosos trabajos con éxito notable utilizando estudios de asociación genómica para identificar las variaciones genéticas que contribuyen al riesgo de todas las enfermedades. Como ejemplos se pueden mencionar la diabetes tipo 2 (T2D) y tipo 1 (T1D) (Esparza-Castro, Andrade-Ancira, Merelo-Arias, Cruz, & Valladares-Salgado, 2015; Grant & Hakonarson, 2009; Lillioja & Wilton, 2009; McCarthy & Zeggini, 2009; X. Wang et al., 2016), la enfermedad de Parkinson (Chang et al., 2017; Foo et al., 2017; Nalls et al., 2014), los trastornos cardíacos (Companioni, Rodriguez Esparragon, Fernandez-Aceituno, & Rodriguez Perez, 2011; McPherson et al., 2007; Samani et al., 2007; Wellcome Trust Case Control, 2007), la obesidad (Gonzalez et al., 2014; Saunders et al., 2007), la enfermedad de Crohn y enfermedades autoinmunes (Franke et al., 2010; H. Hakonarson & Grant, 2009; Lee et al., 2017; Lee & Parkes, 2011), el cáncer de próstata (Ahmed et al., 2016; Barnett et al., 2014; Fachal et al., 2014), también las variaciones genéticas que influyen en la respuesta a medicamentos antidepresivos (Malhotra, 2010; Niitsu, Fabbri, Bentini, & Serretti, 2013; Tansey et al., 2013), drogas (Deak et al., 2022) y ejemplos de enfermedades recientes como la COVID 19 (Cruz et al., 2022) entre otras patologías detalladas en el catálogo de estudios GWAS (MacArthur et al., 2017), (http://www.ebi.ac.uk/GWAS/).

Un ejemplo de posible traslación a la práctica clínica, entre otras muchas enfermedades, es la osteoporosis, enfermedad esquelética común que afecta aproximadamente a 200 millones de personas en todo el mundo. Como una enfermedad compleja, la osteoporosis está influenciada por muchos factores, incluida la dieta (por ejemplo, la ingesta de calcio y proteínas), la actividad física, el estado endocrino, las enfermedades coexistentes y los factores genéticos. En los últimos 14 años los GWAS y los metaanálisis han descubierto cientos de loci asociados con la densidad mineral ósea, la osteoporosis y las fracturas osteoporóticas. En este sentido el uso clínico de los hallazgos de GWAS en el campo óseo, la identificación de factores de riesgo clínicos causales, el desarrollo de dianas farmacológicas (J. Liu et al., 2023) y la predicción de enfermedades están siendo de gran utilidad en poblaciones europeas, pero se requieren más estudios genéticos en otras poblaciones para beneficiar la predicción en la población correspondiente (Zhu, Bai, & Zheng, 2021).

Los GWAS presentan la ventaja de la no necesidad de un gen candidato, es decir, no se requiere una hipótesis previa de asociación entre un gen y una enfermedad. Tuvieron gran aceptación debido a que aúnan las ventajas de los estudios de asociación, que permiten detectar pequeños efectos, y los de ligamento, que no requieren un conocimiento específico de la patogénesis (N. Risch & Merikangas, 1996; H. J. Williams, Owen, & O'Donovan, 2009).

Un efecto de gran valor de los GWAS es la creación de grandes consorcios de investigación internacionales entorno a varias enfermedades. El consorcio por excelencia en este tipo de estudio, *The Wellcome Trust Case Control Consortium* (Samani et al.), demostró que el uso de un conjunto común de controles en múltiples estudios es un enfoque sólido y eficiente, y uno de los miembros del equipo se extendió aún más, usando individuos estudiados para una

enfermedad como controles para otra. Este estudio reveló un grado de diferenciación geográfica previamente insospechado en el Reino Unido para 13 SNPs (por ejemplo, hubo una diferencia de norte a sur en la frecuencia de una variante del gen *TLR1* que podría tener un papel en la lepra y la tuberculosis entre otras enfermedades infecciosas) (Wellcome Trust Case Control, 2007). Actualmente es muy común usar como población control datos de estudios comerciales masivos de genética recreativa como *23 and me*, aunque se han planteado dudas éticas sobre el uso de estos datos en relación con la explotación comercial y la falta de consentimientos informados apropiados.

Entre otros ejemplos, el *Coronary Artery Disease Genome-wide Replication and Meta-Analysis* (CARDIoGRAM) (Preuss et al., 2010), dedicado a la enfermedad coronaria (EC), ha determinado el fenotipo y el genotipo de una población de 82.000 individuos y se han replicado los resultados de más de 40.000 (http://www.cardiogramplusc4d.org/data-downloads/). Existen otros consorcios relevantes en materia de salud mental, entre los que se encuentran el *International Schizophrenia Consortium* (N. J. Risch), *Molecular Genetics of Schizophrenia* (MGS), y el *Schizophrenia Working Group of the Psychiatric Genomics Consortium* (SWG of PGC) (https://pgc.unc.edu/).

Dentro del entorno del PGC uno de los GWAS más grandes realizados es sobre adiciones, y comprende más de un millón de personas. Estos estudios de gran tamaño consiguen detectar riesgos relativos muy bajos, por debajo incluso de 1,05, y ayudan a ver el espectro completo de la contribución genética a la enfermedad (https://pgc.unc.edu/).

Actualmente el "*Alzheimer's Disease Genetics Consortium* (ADGC)" y el "*International Genomics of Alzheimer's Project* (IGAP)" constituyen otro ejemplo de estudios de gran escala para descubrir *loci* de riesgo para la enfermedad. Su metaanálisis implica las proteínas β-amiloide y Tau, así como la vía inmunológica y el procesamiento de lípidos (Kunkle et al., 2018).

Hay que mencionar sin duda el consorcio BCAC (*The Breast Cancer Association Consortium*) y todos los esfuerzos de los proyectos derivados, como *Confluence Project*, B-CAST (*Breast CAncer STratification*) y BRIDGES (*Breast cancer RIsk after Diagnostic Gene Sequencing*), todos ellos implicados en el estudio del riesgo hereditario de cáncer de mama. Estos y otros grupos están realizando estudios con el objetivo de identificar genes que puedan estar relacionados con el riesgo de padecer cáncer de mama. El objetivo del consorcio es combinar datos de muchos estudios y proporcionar una evaluación fiable de los riesgos asociados con estos genes. BCAC es un consorcio multidisciplinario internacional, formado en abril de 2005. Cada estudio derivado de estas investigaciones ha presentado información sobre sus sujetos de estudio, incluidos datos demográficos, datos clínicos y factores de riesgo epidemiológicos clave.

Gracias a estos estudios el *Centre for Cancer Genetic Epidemiology* de la Universidad de Cambridge ha desarrollado un *software* denominado BOADICEA (*Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm*) que puede calcular el riesgo de padecer estas enfermedades en función de los antecedentes familiares incluyendo variación común y variación rara funcional. La versión que se está desarrollando ahora mismo de BOADICEA, versión 6, es el primer modelo de predicción del riesgo de cáncer de mama que utiliza información sobre variantes genéticas raras en genes de susceptibilidad al cáncer de riesgo alto y moderado, puntuación de riesgo poligénico, antecedentes familiares y otros factores de riesgo, como hormonales o de estilo de vida (https://ccge.medschl.cam.ac.uk/boadicea/), (Pal Choudhury et al., 2021), (X. Yang et al., 2022).

Existen otros consorcios, y cada vez más, como por ejemplo el *Early Growth Genetics* (EGG), que representa un esfuerzo de colaboración para combinar datos de múltiples estudios

de asociación de todo el genoma. Su objetivo es identificar *loci* que tienen un impacto en una variedad de rasgos relacionados con deficiencias en el crecimiento temprano.

Como resultado de la actividad de estos grandes consorcios, se espera que la mayor parte de las variantes genéticas que predisponen a enfermedades frecuentes como la EC y el cáncer puedan encontrarse. Hay que destacar también que algunos de estos consorcios permiten el acceso público a los datos resultantes. Estos datos abiertos pueden ser utilizados en posteriores análisis, y gracias a ellos se desarrollan *arrays* de genotipado específicos para interrogar variantes relacionadas con enfermedades concretas. Como ejemplo el *Infinium PsychArray-24 BeadChip de Illumina*, desarrollado en colaboración con el *Psychiatric Genomics Consortium*, para estudios genéticos focalizados a evaluar la predisposición y riesgo a padecer enfermedades psiquiátricas.

Muchos de estos estudios se beneficiaron de *arrays* de nuevo diseño, que junto con la captura de los SNPs de variación común contienen variantes raras funcionales. Podemos enmarcar aquí el *array* diseñado en este trabajo. Nuestro *array* ha sido construido con variación funcional en exomas incluyendo variación rara para enfermedades genéticamente complejas y de respuesta a fármacos partiendo de datos de secuenciación de exomas de población española.

El establecimiento de diferentes biobancos, como veremos en el apartado 1.5.1.6 de este trabajo, ha supuesto también un avance para la ejecución de este tipo de estudios en los que son necesarias grandes cantidades de muestras, ya que a través de ellos resulta más sencillo el acceso a muestras biológicas humanas.

En la siguiente tabla (Tabla 1.1) podemos ver el papel de los *arrays* de SNPs para GWAS en los descubrimientos de la Genética Humana (Visscher et al., 2017).

Tabla 1.1. Papel de los *arrays* de SNPs para GWAS en los descubrimientos de la Genética Humana. Adaptado de "*A plethora of pleiotropy across complex traits*" (Visscher et al., 2017).

Análisis	Objetivo	Hallazgos
GWAS	Detectar asociaciones entre SNPs y rasgos	~10.000 asociaciones robustas entre enfermedades, rasgos cuantitativos y genómicos
Análisis de CNVs	Detectar asociaciones entre CNVs y rasgos	Cientos de asociaciones con enfermedades y desórdenes
Evaluación del DL	Cuantificar la arquitectura genética	Gran variación en DL en el genoma
Estimación de la heredabilidad de los SNPs	Desentrañar la arquitectura genética	Gran proporción de variación genética capturada por marcadores comunes (SNPs)
Estimación de la correlación genética	Detectar y cuantificar pleitropía	La pleitropía es ubicua
Puntuación de riesgo poligénico	Detectar pleiotropía y validar descubrimientos de GWAS	Detección de nuevas asociaciones
Aleatorización Mendeliana	Testar las relaciones causales	Replicación de relaciones causales conocidas; evidencia empírica de asociaciones no causales
Diferentes frecuencias alélicas según las poblaciones analizadas	Reconstruir la historia de las poblaciones humanas; detectar selección	La estructura genética "imita" a la estructura geográfica; evidencia de la selección natural
Rasgos GWAS con -omics GWAS	Mapeo fino; detectar función de genes diana	2/3 de los *loci* asociados en un GWAS implican un gen que no es el más cercano a la mayoría de los SNPs asociados

Teri A. Manolio, experta en este tipo de estudios por su gran cantidad de trabajos publicados, apuesta en un comentario publicado en la revista *Nature* que "los GWAS seguirán siendo útiles, al menos durante un tiempo". Para optimizar su aportación al conocimiento científico, la investigadora señala que "deberán realizarse más GWAS en poblaciones poco analizadas, además de estudiar el efecto o repercusión de algunas variantes muy significativas encontradas en regiones poco descritas del genoma, variantes raras y de baja frecuencia" (Manolio, 2017).

Desde entonces hasta la actualidad se han llevado a cabo numerosos estudios de estas características y relacionados con todo tipo de enfermedades.

1.1.3 Limitaciones de los GWAS

A pesar de que los GWAS representan una ventaja enorme frente a los estudios de asociación basados en genes candidatos, también tienen ciertas limitaciones.

Una de las mayores limitaciones es lo que se conoce en estadística como "el problema de los tests múltiples". En un GWAS se calcula un test estadístico para cada variante, y el número de marcadores a analizar suele ser de varios millones. Esto hace que la probabilidad de encontrar falsos positivos (es decir, variantes que no están realmente asociadas a la enfermedad, aunque el resultado del test estadístico nos diga que sí) sea muy alta. Para disminuir la probabilidad de encontrar falsos positivos, se requiere un número enorme de individuos en el estudio, es decir, tenemos que incluir muchos casos y controles y corregir para comparaciones múltiples usando un test tan exigente como el test de Bonferroni. Afortunadamente, gracias al abaratamiento de los costes de los *arrays* de SNPs, hoy en día se realizan GWAS con miles de sujetos, lo cual ha permitido que en los últimos 7-8 años se hayan descubierto más variantes y genes asociados a enfermedades que en las últimas 3 o 4 décadas.

Hasta hace poco, los esfuerzos para comprender los mecanismos biológicos a través de los cuales estas diversas variantes de riesgo tenían relación directa con el fenotipo, se han visto frustrados por limitaciones en la capacidad de realizar una evaluación a gran escala del impacto funcional. Así, la mayoría de los *loci* carecen de una comprensión mecanicista de cómo influyen en los rasgos. Los estudios de la expresión génica se han convertido en una herramienta clave para vincular la variación de la secuencia de ADN a los fenotipos, por lo que también hay que tener en cuenta las limitaciones que puedan existir en los estudios sobre la naturaleza molecular de las variantes reguladoras y su influencia en el transcriptoma y el proteoma, que expliquen la vinculación de polimorfismos individuales a cambios en la expresión génica, que a su vez dan como resultado cambios fisiológicos y, en última instancia, riesgo de enfermedad (Albert & Kruglyak, 2015).

Los mapas de anotaciones y conexiones en tejidos relevantes para la enfermedad, generados por proyectos como ENCODE (*Encyclopedia of DNA Elements*) (E. P. Consortium, 2012), Epigenome RoadMap (Roadmap Epigenomics et al., 2015), y GTEx (*The Genotype-Tissue Expression*) (Keen & Moore, 2015), han sido cruciales para la interpretación de las variantes no codificantes que representan la mayoría de los alelos de riesgo identificados por GWAS.

Para dilucidar la arquitectura genética resulta de muchísimo interés la imputación. La imputación de genotipos infiere los genotipos que faltan con métodos *in sílico* utilizando información de haplotipos de muestras de referencia a partir de genotipos de *arrays* de genotipado más densos o a partir de datos de secuenciación. Este enfoque mejora el poder estadístico para detectar asociaciones al reducir la cantidad de genotipos que faltan, simplifica la armonización de datos para metaanálisis al mejorar la superposición de variantes genómicas entre conjuntos de muestras con genotipos diferentes y puede aumentar el número total y la densidad de variantes genómicas disponibles para pruebas de asociación (Naj, 2019).

Entre las limitaciones de estos estudios también hemos de considerar la dificultad para detectar *loci* con pequeños efectos y la elección preferente de SNPs, por lo que otros polimorfismos, como variantes de número de copias (CNVs, del inglés *Copy-Number Variations*) y microsatélites no se analizan habitualmente, y no se evalúa de forma rutinaria la contribución de SNPs de baja frecuencia. Otras limitaciones importantes incluyen la escasez de datos en poblaciones no europeas, que la imputación se realiza en poblaciones de origen británico en su gran mayoría y errores de genotipado (T. A. Pearson & Manolio, 2008; Tam et al., 2019). Para abordar la falta de estudios de GWAS que se centren en CNVs, se ha desarrollado

un algoritmo utilizando el *software* PennCNV (Glessner et al., 2023), que permite el análisis de grandes cohortes (>100.000 muestras). Este *software* está disponible de forma gratuita en https://github.com/CAG-CNV/ParseCNV2.

A pesar del éxito obtenido por los GWAS en la detección de *loci* asociados a enfermedades (Donnelly, 2008), hubo cierto escepticismo respecto a si los GWAS facilitarán un avance en el manejo de las enfermedades, sobre todo entre los años 2005 y 2015 (Goldstein, 2009; Manolio et al., 2009; McCarthy et al., 2008; Song, Hao, & Storey, 2015). Sin embargo, a lo largo de estos años se ha puesto de manifiesto en numerosos estudios que a mayor tamaño de muestra se producían más evidencias y descubrimientos, y esto es justo lo que ha ocurrido en las últimas décadas (Torgerson et al., 2011; Visscher, Brown, McCarthy, & Yang, 2012) (Tim Beck, Rowlands, Shorter, & Brookes, 2022). Ya en 2014 el número de SNPs asociados con enfermedades o rasgos era de unos 700, como veremos más adelante, en el punto 1.1.6, actualmente se han descubierto asociaciones entre miles de SNPs y distintas afecciones (Tam et al., 2019).

La preocupación por las limitaciones de los GWAS, ya desde el principio, era que no explicaba la mayor parte de la predisposición genética esperada (Goldstein, 2009; McClellan & King, 2010). La causa principal era la falta de potencia para encontrar riesgos relativos bajos, inferiores a 1,1 o 1,05 por ejemplo. Otra, que ya se intuía en los inicios de los GWAS (Cohen et al., 2004), es la existencia de variantes muy poco comunes de las que podemos esperar un efecto muy superior, que aún están por identificar, ya que hasta la fecha, la mayoría de las variantes genéticas que han sido interrogadas a través de GWAS son comunes en la población, presentando una frecuencia del alelo menor (MAF) >1% (Visscher et al., 2017). Se podría decir que es necesario que los chips utilizados para GWAS dispongan de la sensibilidad necesaria para detectar esas variantes raras con una MAF <1%, ya que los GWAS tradicionalmente se han centrado en el análisis de variantes comunes. Sin embargo, los cada vez más numerosos estudios de secuenciación y *arrays* de zonas codificantes, han llamado la atención sobre las variantes raras (Abdellaoui et al., 2013; Heath et al., 2008), donde se espera que resida parte de la "heredabilidad perdida", la base genética no identificada hasta el momento en los estudios de asociación de muchas enfermedades (Bycroft et al., 2019). Las diferencias existentes en las frecuencias de las variantes raras son elevadas, siendo muchas incluso específicas de población (Dopazo et al., 2016). Dopazo et al. (2016) observaron, en población española, un exceso de variantes de codificación no sinónimas de baja frecuencia, la mayoría de ellas heterocigotas, confirmando así las observaciones hechas en otras poblaciones (Coventry et al., 2010), (Y. Li et al., 2010), (Marth et al., 2011), (Keinan & Clark, 2012), (Tennessen et al., 2012).

Una conclusión inequívoca de los GWAS es que para casi cualquier rasgo complejo que se haya estudiado, muchos *loci* contribuyen a la variación genética. En otras palabras, para la mayoría de los rasgos y enfermedades estudiadas, los polimorfismos presentes en muchos genes contribuyen a la variación genética en la población (Hakon Hakonarson & Grant, 2011; Visscher et al., 2017).

Por otro lado, a pesar de que los consorcios ofrecen la ventaja de recopilar un elevado tamaño muestral, uno de los problemas es precisamente el origen de estas muestras, haciendo que uno de los puntos críticos de estos estudios sea la estratificación poblacional. Esta hace referencia a la posibilidad de obtener resultados espurios debido a las diferencias alélicas ancestrales entre las sub-poblaciones de estudio, en vez de estar asociadas al rasgo estudiado (Bergen & Petryshen, 2012), ya que existe mezcla reciente y la varianza muestral puede conducir a asociaciones erróneas entre un fenotipo y un marcador, o incluso pueden enmascarar asociaciones verdaderas. La estratificación de la población puede ocurrir si los casos y los controles tienen diferentes frecuencias en distintas poblaciones o en poblaciones mezcladas, diferentes fracciones de ascendencia, y cuando los fenotipos de interés, como la enfermedad, la

respuesta al fármaco o el metabolismo de los medicamentos también difieren entre los grupos étnicos. Aunque la mayoría de las variaciones genéticas son interindividuales, también hay una variación interpoblacional significativa (Taylor, Law, Hutchinson, Dennison, & Usher-Smith, 2023).

La identificación de la estructura poblacional permite el estudio de la historia reciente de la población e identifica los puntos débiles en los estudios de asociación, particularmente cuando se prueban variantes raras, a menudo surgidas recientemente (Bycroft et al., 2019).

Para evitar este problema se corrige habitualmente por componentes principales en los estudios de GWAS, y en los estudios de asociación por genes candidatos se presta especial atención al emparejamiento entre casos y controles según su origen geográfico e incluso se recurre al genotipado de un tipo de SNPs especiales llamados AIMs, por sus siglas en inglés *Ancestry Informative Markers*. Este tipo de marcadores informan sobre la ascendencia, ya que exhiben frecuencias sustancialmente distintas entre diferentes poblaciones (Pennisi, 2007).

En los GWAS, existen métodos que permiten detectar y corregir el problema de la estratificación poblacional a partir de los datos de genotipado obtenidos. Las dos primeras estrategias utilizadas fueron el control genómico (Devlin, Roeder, & Wasserman, 2001) y la asociación estructurada (Pritchard & Donnelly, 2001). El control genómico calcula el valor por el que se incrementa el estadístico empleado para identificar la asociación como consecuencia de la estratificación; la asociación estructurada intenta identificar cuántas sub-poblaciones existen en la muestra, de manera que asigna cada uno de los individuos de dicha muestra a una sub-población antes de contrastar las frecuencias de los marcadores.

Actualmente existen otros métodos para evaluar la estratificación poblacional, que serán analizados en detalle en el capítulo 1.5 de este trabajo, y aplicados para la obtención de nuestros resultados: se trata del Análisis de Componentes Principales (del inglés *Principal Component Analysis*, PCA) y el Escalado Multidimensional (del inglés *Multidimensional Scaling*, MDS) entre otros. El PCA identifica los componentes principales y estos son utilizados como covariables en el análisis de la asociación. Esta metodología podría no ajustar adecuadamente para estratificación poblacional si esta es debida a la presencia de varias sub-poblaciones discretas, ya que el PCA usa los vectores propios identificados como covariables continuas. Además, si hay *outliers*, los resultados basados en este tipo de ajuste podrían ser erróneos (Price et al., 2006), por eso es necesario que sean identificados previamente y eliminados del análisis. Por otro lado, el MDS ajusta para estratificación poblacional debido tanto a estructura poblacional discreta como continua; además de funcionar bien tanto para GWAS de gran tamaño como para estudios de pequeño tamaño aunque presenta el mismo problema que el PCA cuando hay presencia de *outliers* (Q. Li & Yu, 2008).

Hoy en día la aceptación del diseño experimental, que se describirá en el apartado 1.1.5 de este trabajo, es mucho mayor, ya que se han demostrado resultados robustos (Visscher et al., 2017), (Rao, Yao, & Bauer, 2021).

1.1.4 Evolución histórica de los GWAS

Los GWAS serán probablemente uno de los hitos de la ciencia del siglo XXI, tal como lo fue el Proyecto Genoma Humano en el siglo XX.

El Proyecto Genoma Humano ha revolucionado por completo la genética humana. La generación de una secuencia consenso, las mejoras en tecnologías masivas de secuenciación y genotipado y el establecimiento de consorcios con un gran número de muestras, como ya se ha mencionado, han cambiado radicalmente el conocimiento de los genes que influyen en la enfermedad humana. Se han mapeado miles de rasgos mendelianos simples, y en muchos casos se han encontrado los genes responsables de las enfermedades mendelianas, permitiendo una

mejora notable en el diagnóstico genético y se ha avanzado enormemente en el conocimiento de la enfermedad compleja.

Desde una perspectiva histórica, la evolución de las técnicas de mapeo comenzó con Morton en 1955 con su trabajo "*Sequential Tests for the Detection of Linkage*". Este método era aplicable únicamente a hermanos, aunque con cálculos más complejos también podrían analizarse grandes familias. El siguiente hito ocurrió en 1971, cuando Elston y Stewart publicaron su eficiente algoritmo para determinar la probabilidad de un pedigrí, que se convirtió en la base del programa informático LIPED[2] escrito por Jurg Ott. El eslabón débil para el éxito de estos métodos fue la cantidad de marcadores disponibles, del orden de 30-40 entre grupos sanguíneos y proteínas séricas, que resultaban tediosos para el análisis y en general no muy polimórficos. La publicación de 1980 de Botstein y colaboradores, que abogó por el uso de polimorfismos de longitud de fragmentos de restricción (RFLPs) como biomarcadores para descubrir correlaciones de genes y enfermedad, marcó el comienzo de una nueva era y el ritmo de mapeo de genes humanos fue aumentando exponencialmente. Aunque los RFLPs fueron importantes para avanzar en la genotipificación, fueron eclipsados por el descubrimiento de los microsatélites por Weber y May en 1989, que son abundantes y altamente polimórficos. Todos estos marcadores fueron sustituidos por el coste y eficacia del genotipado de alto rendimiento de SNPs, que nos dan una visión abrumadora de la variación en el genoma de un individuo.

Los genetistas, durante muchos años, eran conscientes de que la mayoría de los trastornos comunes que afectan a los seres humanos tienen un componente genético importante y sintieron que podían, en teoría al menos, desentrañar los genes responsables como un componente principal de estos trastornos. Pronto se hizo evidente la necesidad de nuevos enfoques para que el mapeo de genes implicados en trastornos complejos tuviese éxito. Un número de genetistas teóricos de poblaciones pronto se unieron para la tarea, lo que condujo a una serie de nuevos e innovadores enfoques de mapeo de genes relacionados con enfermedades humanas complejas.

Hasta hace poco, la técnica utilizada para identificar la asociación de una variante o variantes genéticas a un rasgo observable eran los estudios de asociación con genes candidatos. Estos estudios se basaban en seleccionar SNPs de un grupo de genes candidatos de estar implicados en la enfermedad, en personas enfermas y personas sanas, e intentar encontrar qué variante aparecía más frecuentemente entre las personas enfermas.

Usando estos estudios de asociación genética con genes candidatos, se han identificado los genes (y las variantes dentro de esos genes) causantes de diversas enfermedades. Sin embargo, este tipo de estudios tiene también sus limitaciones, como por ejemplo la identificación de los genes candidatos a estudiar, lo que requiere de experimentos previos para tener al menos la sospecha de que el gen que vamos a analizar puede estar implicado en la enfermedad. En un principio la selección de estos genes se llevaba a cabo según sus características funcionales (**genes candidato funcionales**), es decir, genes que, según conocimientos moleculares previos, podrían estar implicados en rutas metabólicas relacionadas con la enfermedad, llevar a cabo una función biológica clave en la patología, expresarse en regiones implicadas en el trastorno, etc.; en definitiva: las proteínas codificadas por estos genes podrían desempeñar un papel importante en la patofisiología de la enfermedad.

La alternativa a estos estudios era analizar genes situados en regiones con elevada probabilidad de estar relacionados en la etiología de la enfermedad, como pueden ser genes

[2] El programa LIPED, para LIkelihoods (probabilidades) en PEDigrís, estima la fracción de recombinación calculando las probabilidades de pedigrí para varios valores asumidos de la fracción de recombinación.
http://www.jurgott.org/linkage/liped.html.

situados en regiones cromosómicas previamente asociadas con estudios de ligamento o genes que están alterados por aberraciones cromosómicas implicadas en la patología. Son los llamados **genes candidato posicionales**.

Los estudios de asociación por genes candidatos tenían muchas limitaciones y pronto fueron sustituidos por los GWAS, que no requieren un conocimiento previo de la etiopatogenia y tienen mucho más poder.

El primer estudio GWA exitoso fue publicado en el año 2005 e investigó la degeneración macular relacionada con la edad, una de las principales causas de ceguera (Klein et al., 2005). Se encontraron dos SNPs que habían alterado significativamente la frecuencia de los alelos cuando se compararon con la misma frecuencia de los alelos de los controles sanos (Hageman et al., 2005). Desde 2005 a 2008, se identificaron y replicaron casi 100 *loci* para hasta 40 enfermedades y rasgos comunes en los estudios de GWA. Estos *loci* se disponían en genes que previamente no se sospechaba que tuviesen un papel en la enfermedad en estudio, e incluso algunos en regiones genómicas que no contienen genes conocidos (T. A. Pearson & Manolio, 2008), dando lugar a asociaciones inesperadas. Una de las grandes sorpresas iniciales de los hallazgos de GWAS, por ejemplo, fue que menos del 10% de las asociaciones genéticas con enfermedades se encuentran en las regiones codificantes del genoma (Hindorff et al., 2009).

De 2008 a 2012 se identificaron 3.600 SNPs para enfermedades o rasgos comunes. En general se aprecia que las enfermedades tienen múltiples alelos de susceptibilidad, cada uno con pequeños tamaños de efecto.

A partir de 2011 se han llevado a cabo miles de GWAS en humanos examinando prácticamente todas las enfermedades, hallando miles de asociaciones entre SNPs y distintas patologías. Varios estudios de GWA han recibido críticas por omitir importantes pasos de control de calidad, haciendo que los resultados no fuesen válidos, pero las publicaciones modernas son más estrictas en estos aspectos. Se pueden consultar los SNPs identificados en el catálogo publicado en el *European Molecular Biology Laboratory-European Bioinformatics Institute* (EMBL-EBI) en el siguiente enlace: http://www.ebi.ac.uk/GWAS/.

El Centro Nacional para la Información Biotecnológica (NCBI, del inglés *National Center for Biotechnology Information*), que constituye una parte de la Biblioteca Nacional de Medicina del *National Human Genome Research Institute* (Studies et al.), está desarrollando bases de datos para uso de la comunidad investigadora. Desde el 2007 (Mailman et al., 2007) se puede acceder a un archivo de datos de estudios de asociación de todo el genoma sobre una variedad de enfermedades y afecciones a través del sitio web del NCBI, llamado Base de Datos de Genotipo y Fenotipo (dbGaP), en constante actualización y ubicado en https://www.ncbi.nlm.nih.gov/gap.

El NIH (*National Institute of Health*), *Pfizer Global Research & Development* y otros, han formado una asociación público-privada: la Red de Información de Asociación Genética (GAIN, del inglés *Genetic Association Information Network*), para financiar estudios de asociación de todo el genoma. Después de la revisión por pares de las aplicaciones, GAIN anunció su primera ronda de estudios en octubre de 2006. Los estudios iniciales incluyeron el trastorno bipolar, la depresión mayor, la enfermedad renal en la T1D, el trastorno por déficit de atención con hiperactividad, la esquizofrenia y la psoriasis. Se pueden encontrar los datos resultantes en la base de datos de Genotipos y Fenotipos (dbGaP) mencionada en el párrafo anterior.

Poco a poco surgieron nuevos estudios. Por ejemplo, el *National Heart Lung and Blood Institute* (NHLBI) lanzó el estudio de investigación genética de Framingham, el *Framingham Heart Study* (Fradin & Fallin, 2009), en colaboración con la Facultad de Medicina de la Universidad de Boston. En ese estudio 9.000 participantes se sometieron a estudios de

asociación genómica a largo plazo para identificar los genes subyacentes a las enfermedades cardiovasculares y otras enfermedades crónicas, como la osteoporosis y la diabetes. Se puede encontrar más información sobre ese estudio en el siguiente enlace: https://www.framinghamheartstudy.org/.

Otros esfuerzos del NHLBI en esta área incluyen estudios de asociación genómica, como el Estudio de Salud de la Mujer, la iniciativa de Salud de la Mujer y el Recurso de la Asociación de genes candidatos, que reúnen muestras de ADN recogidas de múltiples estudios de cohortes. El NHLBI junto con el Instituto Nacional de Ciencias Médicas Generales de Estados Unidos, también son contribuyentes importantes a la Red de Investigación *PharmacoGenetics*. Junto con muchas otras herramientas y tecnologías, esta red está utilizando estudios de asociación de todo el genoma para explorar los efectos de los genes en las diferentes respuestas de los individuos a los medicamentos. Los datos derivados de estos trabajos se recogen en la base de datos dbGaP del NCBI.

Los ejemplos mencionados hasta el momento en este apartado hacen referencia a estudios llevados a cabo en población estadounidense. En Europa también hay una grandísima cantidad de investigaciones publicadas, quedando de manifiesto que la estructura genética poblacional es muy específica, por lo que es necesario llevar a cabo GWAS con poblaciones concretas y bien definidas. Entre otros muchos existen estudios publicados recientemente, como el trastorno por consumo de opioides en población europea con ascendencia africana, en el que se han descubierto 19 *loci* independientes de riesgo (Deak et al., 2022), GWAS que han demostrado asociaciones genéticas con fenotipo severo de *Coronavirus disease* 2019 en individuos de Italia, España, Noruega, Alemania y Austria, como el trabajo de Degenhardt et al, "*Detailed stratified GWAS analysis for severe COVID-19 in four European populations*" y las investigaciones del Consorcio SCOURGE: *Spanish COalition to Unlock Research on host GEnetics on* COVID-19 (Cruz et al., 2022).

Tras la revisión de los estudios de GWAS realizados hasta la fecha, es razonable predecir que en los próximos años continúe aumentando tanto el tamaño de la muestra para llevar a cabo estas investigaciones como el número de variantes asociadas a enfermedades, que explicarán de manera acumulativa una gran proporción de la heredabilidad (Visscher et al., 2017). En junio de 2020 ya se habían revelado 487.213 asociaciones en 6.263 estudios reflejados en 55.244 publicaciones. (https://www.ebi.ac.uk/; https://www.ebi.ac.uk/gwas/diagram); Figura 1.2.

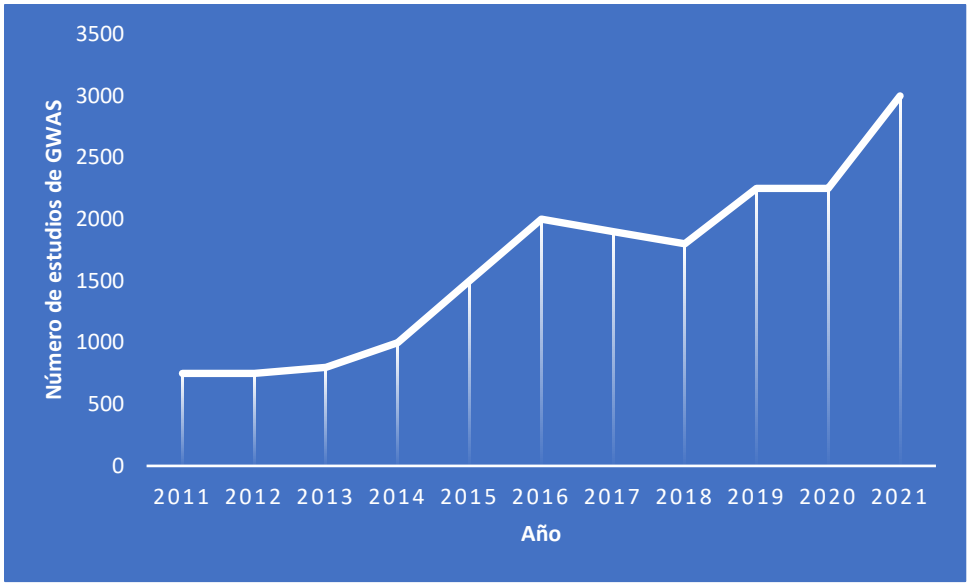


Figura 1.2. Número de GWAS llevados a cabo desde 2011 a 2021. Fuente propia. Datos extraídos de "*Examining Barriers and Opportunities of Conducting Genome-Wide Association Studies in Developing Countries*" (Dumancas, Rachal, Zamora, & de Castro, 2022).

1.1.5 Metodología de este tipo de estudio

El protocolo de diseño de los GWAS incluye todos los procesos clínicos, moleculares, bioinformáticos y componentes analíticos. La población de estudio debe definirse como una selección de individuos de esa población que exhiban el rasgo de elección. Se debe determinar la tecnología molecular a utilizar, así como establecer el sistema electrónico para almacenar y recuperar los datos clínicos y moleculares. También los métodos de análisis que se utilizarán. Finalmente, se debe garantizar que se aborden todos los problemas éticos, legales y sociales. En la Figura 1.3 se muestra el fundamento de un estudio GWA.

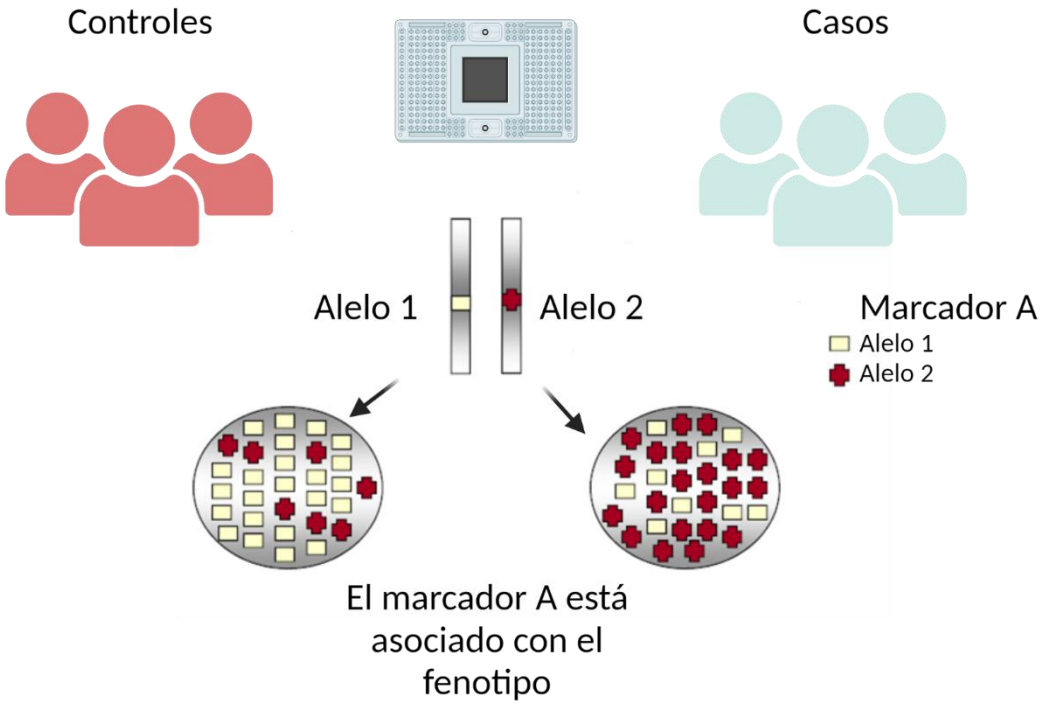


Figura 1.3. Fundamento de un Estudio de Asociación de Genoma Completo. Fuente propia. Realizado con www.Biorender.com.

La idea de que las enfermedades comunes tienen diferente arquitectura genética subyacente que los trastornos poco comunes, junto con el descubrimiento de variantes de susceptibilidad para la enfermedad común con alta frecuencia de alelos menores, condujo al desarrollo de la hipótesis enfermedad común/variante común, ya mencionada anteriormente (D. E. Reich & Lander, 2001). Debemos hacer varias consideraciones al respecto:

Primero: si las variantes genéticas comunes influyen en la enfermedad, el tamaño del efecto (o penetrancia) para cualquier variante debe ser pequeño en relación con el que se encuentra en trastornos raros. Por ejemplo, si un SNP con 40% de frecuencia en la población causa una sustitución de aminoácidos altamente nociva que conduce directamente a un fenotipo de enfermedad, casi el 40% de la población tendría ese fenotipo. Por lo tanto, la frecuencia de alelos y la prevalencia en la población están completamente correlacionadas. Sin embargo, si ese mismo SNP causó un pequeño cambio en el gen y su expresión influyendo ligeramente en el riesgo de una enfermedad, la prevalencia de la enfermedad y el alelo influyente estarían solo ligeramente correlacionados. Como tal, variantes comunes casi por definición no pueden tener alta penetrancia;

Segundo: si los alelos comunes tienen pequeños efectos genéticos (baja penetrancia), pero los trastornos comunes muestran heredabilidad en familias, múltiples alelos comunes deben influir en la susceptibilidad a la enfermedad.

Estos dos puntos sugieren que la genética tradicional basada en los estudios de ligamiento en familias no son exitosos para enfermedades complejas, lo que provocó un cambio hacia estudios basados en la población (Bush & Moore, 2012).

La frecuencia con que un alelo ocurre en la población y el riesgo incurrido por ese alelo para enfermedades complejas, son componentes clave a considerar cuando se planifica un estudio genético para poder escoger la tecnología y tamaño de muestra necesarios.

Un ejemplo del espectro de los efectos del potencial genético se puede visualizar en la figura 1.4, y se dividen por tamaño del efecto y frecuencia alélica.

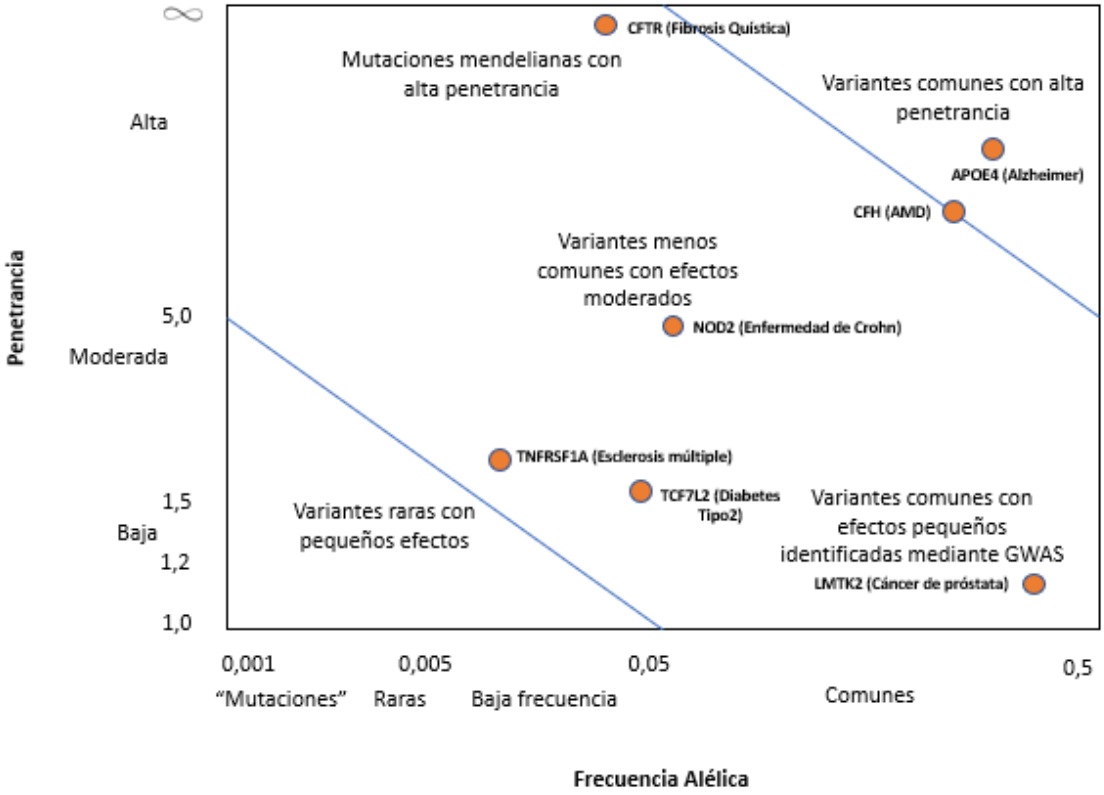


Figura 1.4. Las variantes comunes son capturadas por estudios GWAS, las variantes raras funcionales por métodos de estudios de secuenciación de nueva generación. Adaptado de "*Ikegawa S. A short history of the genome-wide association study: where we were and where we are going*. (Bush & Moore, 2012)".

Como ya se ha mencionado, la disponibilidad de muestras control para usar en comparación con las muestras de pacientes es clave en los estudios de GWA. Independientemente de la relación de la muestra control con la muestra del paciente, se debe garantizar que los controles se determinan a partir de la misma población de estudio que los pacientes**.** Además, los controles deben coincidir con los casos en los factores de confusión (cualquier factor que pueda influir en la asociación entre la enfermedad y el genotipo), tales como edad, sexo, origen étnico y ubicación geográfica.

Hay dos enfoques para hacer coincidir controles y casos (Jonathan L. Haines, 2005):

los controles pueden ser seleccionados de tal manera que la distribución general de casos y controles sea comparable con respecto a la frecuencia de factores de confusión (por ejemplo, para un estudio de trastorno autista, tanto los casos como los controles tienen una proporción de sexos de 3:1 hombres a mujeres). Esto se conoce como **frecuencia o categoría pareo**.

Alternativamente, uno o más individuos control pueden ser seleccionados para relacionar cada caso en función de las características de confusión (por ejemplo, el caso y el control son

mujeres afroamericanas, de 8 años de edad y residentes en el Santiago de Compostela, Galicia). Este enfoque se llama **emparejamiento individual**.

La selección incorrecta de controles puede llevar a conclusiones erróneas. Por ejemplo, si los casos y controles no concuerdan en el origen étnico y la frecuencia de los alelos para el marcador genético difieren según la etnia, un estudio de asociación puede ser un fracaso. Uno puede concluir falsamente una asociación entre un marcador genético y la enfermedad si el alelo del marcador "de riesgo" es más prevalente en la etnia predominante de los casos *versus* los controles. En la figura 1.5 se describe el diseño general de un estudio para identificación de genes implicados en rasgos complejos.

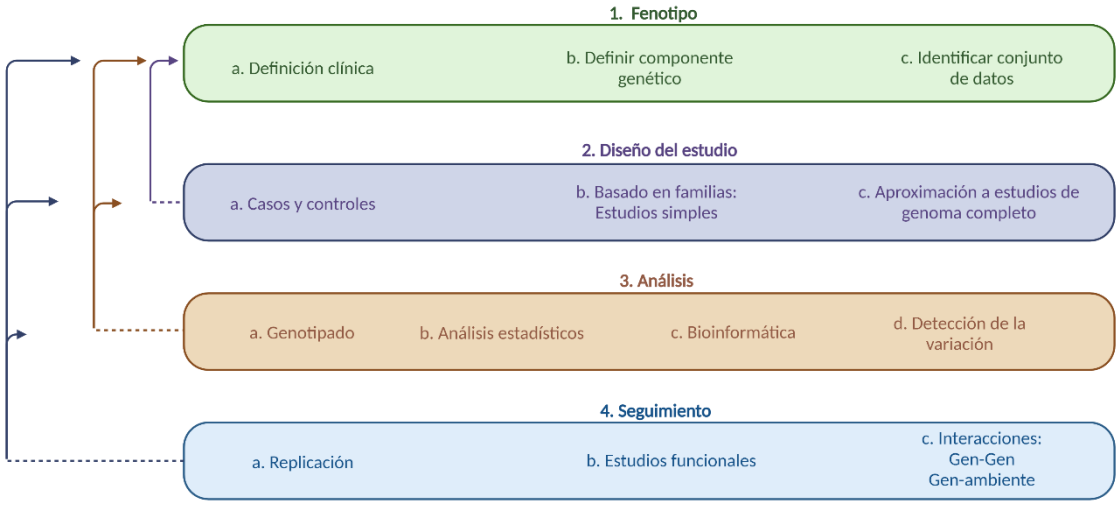


Figura 1.5. Diseño de un estudio para identificación de genes implicados en rasgos complejos. Adaptado de *"Genetic Analysis of Complex Diseases* (Jonathan L. Haines, 2005)". Realizado con www.Biorender.com.

Hay muchos tipos de métodos estadísticos (descritos en los puntos 1.5 y 3.2 de este trabajo) para analizar los datos en un GWAS. En el caso de los estudios de casos y controles el test más sencillo es un simple Chí-cuadrado para comparar la frecuencia de las variantes entre casos y controles. Se lleva a cabo el contraste de hipótesis o prueba de significación, en la que se plantea una hipótesis inicial (H_0) de igualdad de frecuencia entre casos y controles, que es aceptada si el valor-p es >0,05 y es rechazada si es menor.

Así, se analiza cada variante por separado, y si la frecuencia de una de ellas es significativamente más alta en los casos que en los controles, en principio y sin una réplica adecuada del estudio, significaría que esa variante está asociada a la enfermedad.

A continuación, se enumeran una serie de criterios a tener en cuenta en caso de estudios de asociación genotipo-fenotipo evaluados por enfoques genómicos o de genes candidatos (Studies et al., 2007):

• Los análisis estadísticos que demuestren el nivel de significación estadística de un hallazgo deben ser publicados o al menos estar disponibles para que otros pueden intentar reproducir los resultados.

• Se debe proporcionar información explícita sobre el poder del estudio para detectar una gama de efectos.

• El estudio debe ser epidemiológicamente óptimo, con una cuidadosa atención en posibles sesgos en cuanto a la selección de sujetos, caracterización de fenotipos, comparabilidad de exposiciones ambientales (cuando sea posible) y estructura de la población en casos y controles.

• Los fenotipos deben evaluarse de acuerdo con definiciones estándar proporcionadas en el informe.

• Las asociaciones deben ser consistentes (dentro del rango de fluctuación estadística esperada) y se debe informar sobre los mismos fenotipos a través de subgrupos de estudio.

• No se deben alterar los métodos de control de calidad, ya que esto podría modificar la inclusión o exclusión de un gran número de muestras o *loci*.

• Para evaluar la calidad de los datos de genotipado se deben incluir los resultados del estudio junto con ejemplos duplicados. Alternativamente las muestras deben estar disponibles al público.

• Los resultados concordantes deben ser revelados junto con las tasas de error. Un subconjunto notable de SNPs debe ser evaluado con una segunda tecnología que verifique el mismo resultado con concordancia. Ninguna tecnología está libre de errores.

• Deben ser informadas asociaciones con SNPs cercanos al supuestamente asociado a la enfermedad o rasgo si existe desequilibrio de ligamiento.

• Los resultados de los estudios de replicación deben ser reportados incluso si los resultados no son significativos.

• Se deben informar las diferencias en la estructura poblacional de los casos y controles.

• Deben ser informadas todas las pruebas estadísticas examinadas. Del mismo modo, para enfoques bayesianos, debe ser descrita previamente la elección de probabilidades.

Los GWAS han demostrado que las mismas variantes genéticas pueden asociarse significativamente con múltiples enfermedades y rasgos cuando los fenotipos se miden en diferentes individuos (siempre que no existan asociaciones medioambientales que afecten a los resultados) (Bulik-Sullivan et al., 2015; Pickrell et al., 2016; Sivakumaran et al., 2011) (T. Beck, Rowlands, Shorter, & Brookes, 2023).

Algo también muy importante que se desprende de estos estudios es la pleitropía (Masotti, Guo, & Wu, 2019): los métodos analíticos que estiman las correlaciones genéticas de los datos de GWAS han proporcionado evidencia de este fenómeno, en el que un gen es responsable de distintos fenotipos. La verdadera naturaleza de la pleiotropía es actualmente desconocida, pero en algunos casos, podría implicar un impacto de las variantes en diferentes tejidos y/o a diferentes edades. (Bulik-Sullivan et al., 2015; Cross-Disorder Group of the Psychiatric Genomics et al., 2013; Ellinghaus et al., 2016; Y. R. Li et al., 2015; Parkes, Cortes, van Heel, & Brown, 2013; Pickrell et al., 2016; Sivakumaran et al., 2011). Recientemente ha habido un considerable interés en identificar variantes particulares con efectos pleiotrópicos en diferentes rasgos (Cotsapas et al., 2011; Pickrell et al., 2016), así como en la identificación de rasgos con efectos genéticos correlacionados (Bulik-Sullivan et al., 2015). Sin embargo, la observación de que las señales genéticas se distribuyen ampliamente en todo el genoma implica que la pleiotropía puede estar omnipresente (Visscher & Yang, 2016).

Otro componente crítico en la disección de la implicación genética en una enfermedad compleja es una comprensión de las posibles interacciones entre el gen o genes que subyacen al rasgo y entre genes y otros factores de riesgo (generalmente ambientales). Este es quizás el paso menos desarrollado, ya que solo ahora es posible identificar y examinar más de un gen (y/o factor de riesgo) para enfermedades complejas. Este paso también requiere la integración de las técnicas utilizadas en genética y epidemiología, un proceso que, obviamente, continúa en desarrollo. Tales efectos solo serán identificados por el uso de métodos que consideran los efectos simultáneos de múltiples factores genéticos y ambientales. Una combinación de enfoques genéticos y epidemiológicos en el estudio de interacciones genéticas ofrece potencial

de éxito en la investigación debido a la naturaleza multifactorial de muchas enfermedades comunes y complejas.

Hay varias formas en que los genes y el entorno pueden interactuar para influir en el desarrollo de las distintas patologías. La susceptibilidad genética puede influir en el riesgo de la enfermedad en sí misma, exacerbar el efecto de un factor de riesgo ambiental o el factor de riesgo ambiental puede intensificar el efecto genético (Ottman, 1990). Por lo tanto, es necesario el estudio de diseños que consideran factores genéticos y ambientales para evaluar las interacciones entre el genotipo o la historia familiar y las influencias ambientales, mejorando así la capacidad para descubrir influencias genéticas en la enfermedad (Ottman, 1990). Dado que la exposición a algunos factores ambientales que influyen en los riesgos genéticos es modificable, el descubrimiento de tales relaciones tiene una importante trascendencia en la salud pública.

Los rasgos complejos son, por tanto, multifactoriales y están influenciados por múltiples genes y/o factores ambientales, por lo que la expresión de fenotipos complejos está influenciada por la heterogeneidad genética, la interacción gen-gen y sus efectos modificando genes, así como por la interacción gen-ambiente. La consideración de interacciones complejas es esencial en el diseño e implementación de estudios exitosos de rasgos multifactoriales.

Este modelo clásico poligénico, en principio aceptable, ha sido cuestionado en el nuevo trabajo de Jonathan Pritchard (Boyle, Li, & Pritchard, 2017), profesor de genética en la Universidad de Stanford, y sus colaboradores, en el que plantean una nueva idea para explicar las bases genéticas de las enfermedades complejas. Los investigadores consideran que los datos obtenidos de los GWAS no apoyan el modelo clásico por el que las variantes genéticas que causan las enfermedades genéticas se concentran en unos pocos genes y rutas moleculares. Por el contrario, señalan que los factores genéticos identificados en los GWAS se encuentran distribuidos por todo el genoma y su enriquecimiento en ciertas rutas moleculares es limitado.

Tras analizar el efecto de los genes sobre un rasgo complejo concreto, la altura, y encontrar qué variación genética distribuida a lo largo de todo el genoma influía en esta característica, los investigadores se replantearon la forma de considerar la contribución de los genes a los rasgos complejos. “Gradualmente comencé a darme cuenta de que los datos no encajaban realmente en el modelo poligénico,” afirma Pritchard.

Pritchard y colaboradores plantean un modelo “omnigénico”, según el cual los genes que no tienen una función directa en las rutas de la enfermedad podrían, en conjunto, tener un mayor papel en la enfermedad que los genes centrales de estas rutas, siempre que estén activos en los tejidos relevantes para la enfermedad. Así, la suma de muchos “impactos menores” sobre la enfermedad sería mayor que la de unos pocos “impactos mayores” (Thomson, Pritchard, Shen, Oefner, & Feldman, 2000). Esto explica la observación de que las variantes genéticas que se localizan cerca de los genes relacionados con funciones importantes para una enfermedad compleja representan solo una pequeña parte de la heredabilidad de dicha enfermedad, dejando mucha variación por explicar (Boyle et al., 2017).

En definitiva, el análisis exitoso de los datos del genoma analizados en *arrays* de genotipado depende de una exploración cuidadosa de los datos, así como de la preparación previa de las muestras (Sale, Mychaleckyj, & Chen, 2009). Los fallos de las muestras no contabilizadas, los errores de genotipado y la estructura de la población pueden introducir señales engañosas que imitan una asociación genuina. La interpretación cuidadosa de los datos estadísticos y las visualizaciones de datos gráficos pueden minimizar las asociaciones falsas, que deben seguirse en los experimentos de replicación. Es por tanto de suma importancia un buen diseño de las técnicas metodológicas y estadísticas para una posterior interpretación sensata de los datos genéticos generados (Teo, 2008).

Se ha determinado que el potencial éxito de un GWAS para un rasgo o enfermedad particular depende de los siguientes factores (Visscher et al., 2017):

- Cómo muchos *loci* afectan a la segregación del rasgo en la población.
- La distribución conjunta del tamaño del efecto y la frecuencia de los alelos en esos *loci* ("arquitectura genética")
- El tamaño de muestra experimental.
- El panel de variantes del genoma que se usan en el GWAS.
- Cómo de heterogéneo es el rasgo o enfermedad que se estudia. Esto hace referencia tanto a la biología del rasgo como a la capacidad para diagnosticarlo o medirlo con precisión.
- Todas las posibles fuentes de sesgo en el conjunto de datos deben ser cuidadosamente consideradas.

Como se ha resaltado en distintos puntos anteriores del presente trabajo, es muy importante tener en cuenta que la asociación alélica puede ser específica de población y que los niveles de asociación alélica entre los alelos en dos *loci* pueden diferir entre las poblaciones, ya que existe una gran variabilidad según la historia de la población, a nivel regional e incluso local. Por tanto, tras un estudio de estas características es precisa la replicación posterior de los resultados obtenidos.

En la literatura científica hay muchos informes iniciales de asociaciones alélicas o genotípicas que no se pueden replicar en absoluto o se replican solo en una pequeña minoría de estudios de seguimiento. Especialmente para efectos genéticos que pueden ser relativamente modestos, el análisis de referencia actual es observar el mismo efecto en un segundo conjunto de datos independiente: por tanto, ¿qué constituye la replicación de una asociación genotipo-fenotipo y cuál es la mejor forma de lograrlo?: el propósito de un estudio de replicación es evaluar un resultado positivo de un estudio anterior, para proporcionar credibilidad al hallazgo inicial y corroborar su validez. La replicación es esencial para establecer la credibilidad de una asociación genotipo-fenotipo, ya sea derivada de estudios de gen candidato o estudios de asociación de todo el genoma.

Sin embargo, no existió consenso sobre lo que constituye un hallazgo que merece ser replicado. Investigadores y editores de revistas han ofrecido directrices sobre cómo abordar este problema (Associating., 1999; Clark, Boerwinkle, Hixson, & Sing, 2005; Freimer & Sabatti, 2007; Neale & Sham, 2004; Todd, 2006), pero estos esfuerzos iniciales se vieron obstaculizados por la experiencia limitada y conflictos empíricos en los datos. Los estudios de replicación de trabajos de asociaciones genotipo-fenotipo pueden considerarse fiables, aunque en muchos casos los hallazgos iniciales no se han reproducido en los estudios de seguimiento debido a problemas, o bien en el estudio inicial o en el intento de replicación (Colhoun, McKeigue, & Davey Smith, 2003; Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002; Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001; Ioannidis, Trikalinos, & Khoury, 2006; Lohmueller, Pearce, Pike, Lander, & Hirschhorn, 2003). El principal problema, ya mencionado y que cabe recalcar, hace referencia al pequeño tamaño muestral o falta de comparabilidad entre casos y controles.

En el trabajo del *Nci-Nhgri Working Group* (Studies et al., 2007) se presentan las conclusiones sobre la replicación de las asociaciones genotipo-fenotipo, bien identificadas en genoma completo o en estudios de gen candidato. El grupo estaba compuesto por expertos de diversas disciplinas, incluidas bioestadística, medicina clínica, epidemiología, genética y publicación científica. El propósito era revisar el estado del campo y proponer mejores prácticas para el diseño, desarrollo y publicación de estudios de replicación que apuntan a seguir

hallazgos notables, particularmente en estudios de asociación de genoma completo. El grupo abordó tres temas. Primero, la evaluación de la validez y las limitaciones de cualquier estudio de asociación genética. En segundo lugar, los criterios para establecer la replicación de los GWAS. En tercer lugar, los puntos a considerar para la publicación de informes de alta calidad derivada de estas investigaciones. A continuación, se detallan estos criterios, que están destinados a servir como una guía para autores y editores de revistas, con el objeto de permitir una interpretación clara e inequívoca de los datos y resultados de los estudios de asociación de genoma completo y otros estudios de asociación genotipo-fenotipo:

Información del estudio

- Una descripción detallada del diseño del estudio y su implementación.
- La fuente de casos y controles (o miembros de la cohorte, si se basa en el diseño de cohorte), incluyendo el período de tiempo y la(s) ubicación(es) de los sujetos reclutados.
- Métodos para determinar y validar el fenotipo y reproducibilidad de la clasificación.
- Tasas de participación para casos, controles o miembros de la cohorte.
- Presentación de la selección de casos y controles en un diagrama de flujo, incluidos los puntos de exclusión para datos perdidos y erróneos
- Tabla inicial que compare características relevantes (como la demografía, factores de riesgo y exposiciones) de casos y controles.
- Tasa de éxito para la adquisición de ADN, incluidas comparaciones de los procedentes o no de colecciones, fallas de extracción y exclusiones debido a datos inconsistentes.

Problemas de datos

- Declaración sobre disponibilidad de resultados y datos para que, en la medida de lo posible, otros investigadores puedan analizarlos independientemente.
- Enlaces a recursos suplementarios en línea y acceso a bases de datos. Técnica de genotipado y procedimientos de control de calidad.
- Métodos de trazabilidad de la muestra, como códigos de barras, para garantizar la precisión del análisis.
- Descripción de los ensayos de genotipado y protocolos, particularmente cuando son nuevos o se aplican de un modo no estándar.
- Descripción del algoritmo de asignación de genotipos.
- Diseño de control de calidad del genotipado para las muestras, incluyendo números, ubicaciones de placas y criterios de selección para:
  -Muestras control externo aceptadas de sets estándar (como por ejemplo HapMap).
  -Muestras de control interno (muestras duplicadas; se debe especificar si estas proceden de la misma o diferente colección de ADN, extracción o alícuota).
- Calidad de los ensayos y del ADN por locus, muestra, placa o lote ("*batch*").
- Tasas de genotipado.

- Tasas de error promedio estimadas por duplicados internos o muestras externas.
- Reproducibilidad del ensayo: concordancia para rendimientos de extracción, alicuotado (muestras de control interno) y reproducibilidad del ensayo.
- Concordancia con lo publicado o con los genotipos generados previamente.
- Comprobación de coherencia mendeliana.
- Detección de relaciones inconsistentes en los sujetos de estudio.
- Evaluación de las desviaciones de Hardy-Weinberg para detectar ensayos fallidos o estratificación a gran escala por separado en casos y controles.
- Evaluación de la heterogeneidad de la población.
- Valor promedio de Chí-cuadrado y distribución completa.
- Gráficos Q-Q (gráficos que incluyen cuartiles) de análisis de Chí-cuadrado y valores-p (con una descripción específica del tipo de prueba utilizada para generar los valores).
- Validación de la mayoría de los resultados críticos en una plataforma de genotipado independiente.

Resultados

- Métodos de análisis con suficiente detalle para reconstruir el enfoque analítico y reproducir todos los resultados informados.
- Descripción de cualquier esquema de pre-análisis para seleccionar variantes para la replicación.
- Análisis de asociación sencillo de locus único y multi marcador (haplotipo).
- Modelos genéticos probados (sin restricciones de efectos de genotipo dominante, aditivo o multiplicativo).
- Visualización gráfica de la agrupación de genotipos para ensayos de alto interés.
- Verificación de resultados en *loci* altamente correlacionados.
- Discusión de la elección del umbral de significancia y la base estadística para cualquier ajuste para múltiples pruebas y la relación con el poder general del estudio.
- Importancia de cualquier "control positivo" conocido (como *loci* significativos en estudios de asociación genética anteriores).
- Consistencia de los resultados antes y después de la aplicación de filtros de control de calidad.

Estudios de replicación

- Descripción de las muestras replicadas, incluyendo fuente, verificación y comparabilidad con la muestra inicial.
- Discusión de la elección del umbral de significancia y la base estadística de cualquier ajuste para múltiples pruebas y la relación con el poder general del estudio.
- Resumen de intentos de replicación y análisis por los autores.

- Resumen de todos los intentos de replicación conocidos llevados a cabo por terceros, incluidas las no repeticiones.

Datos de genotipado y especificaciones para su deposición en bases de datos estándar

- Disponibilidad de datos de genotipos "en bruto" en la tecnología y formato del proveedor, consistente con los requisitos o restricciones impuestas por las fuentes de financiación o consentimientos informados.
- Protocolos de extracción y procesamiento de datos.
  Procedimientos, normalización y transformación de datos, así como parámetros de selección.

Puntos a considerar por revisores y autores con respecto a la prioridad a la hora de la publicación

- Poder de la observación.
- Tamaño de muestra adecuadamente grande.
- Criterios suficientemente rigurosos para la significancia (pequeños valores-p).
- Diseño de estudio de alta calidad, lo que incluye selección de población de estudio, fiabilidad de los fenotipos, medición y ajuste para posibles factores de confusión.
- Discusión y conclusiones proporcionales al tamaño de muestra, potencia, valor-p y calidad epidemiológica del diseño del estudio.
- Estándares de control de calidad utilizados, incluida la evaluación de la calidad del genotipado.
- Utilidad de las observaciones para investigaciones posteriores.
- Valor de la hipótesis inicial descrita.
- Breve presentación de las implicaciones, especialmente las que se relacionan con un mayor seguimiento de los marcadores genéticos y para investigar la plausibilidad de estudios corroborativos.
- Explicaciones de propuestas alternativas apropiadas brevemente discutidas.
- Explicaciones biológicas o funcionales basadas firmemente en los datos disponibles.

En la medida de lo posible, deben ser considerados criterios igualmente rigurosos para la evaluación de los estudios de asociación genotipo-fenotipo con disponibilidad limitada o nula de sujetos para una posterior replicación, como en estudios de enfermedades raras o toxicidad severa debido a la terapia o a las exposiciones al medio ambiente. En estas circunstancias es necesaria información adicional obtenida de técnicas de laboratorio y herramientas bioinformáticas (Studies et al., 2007).

En el primer intento de replicar un hallazgo, las poblaciones comparables deberían ser analizadas no solo por el efecto principal, sino también para protegerse contra la estratificación, presente bien en los estudios iniciales o en la replicación (Price et al., 2006; Wacholder, Rothman, & Caporaso, 2000). Existen muchos estudios iniciales que han sido replicados en poblaciones de ascendencia europea, y esto debe ser extendido a otras poblaciones.

Ya se ha demostrado en varios estudios en una población, que muchas variantes que tienen una asociación significativa con la enfermedad pueden no tener necesariamente la misma

asociación en una población diferente; como ejemplo la variante rs7903146 de *TCF7L2*[3], variante de riesgo en Europa y África Occidental pero no en Asia Oriental (Chandak et al., 2007; Helgason et al., 2007; Horikoshi et al., 2007). En algunas circunstancias podría ser imposible realizar estudios de seguimiento debido a la singularidad de la población de estudio o la falta de disponibilidad de sujetos para la replicación.

Generalmente se espera poder llevar a cabo la evaluación de una asociación en poblaciones de diferente ascendencia de la inicial, ya que la variación genómica es mayor cuando se compara a través de las poblaciones y debería aumentar la confianza en el hallazgo. Por el contrario, la falta de réplica del estudio en una población diferente a la inicial no necesariamente invalida el hallazgo original. En algunos casos, las diferencias en regiones de desequilibrio de ligamiento a través de las poblaciones se pueden utilizar para reducir la región de interés para estudios genéticos posteriores y un posible análisis funcional.

Si bien la mayoría de los esfuerzos de descubrimiento de genes relacionados con enfermedades han basado su éxito en el hallazgo de variantes raras en el gen, esto no es una evidencia suficiente y suelen ser necesarios estudios funcionales con modelos celulares y animales.

Aunque los modelos animales en ratón o pez cebra siguen siendo importantes, la combinación de la ingeniería genética mediante sistemas de edición genómica CRISPR-Cas en hiPSCS junto con la generación de organoides *in vitro* proporciona una oportunidad sin precedentes para imitar el efecto de las variantes genómicas en la etiología de la enfermedad.

Encontrar formas de mejorar los métodos de GWAS es un área activa de investigación, y, así, se están desarrollando y aproximando nuevos enfoques estadísticos con el fin de aumentar la potencia para detectar variantes individuales, teniendo en cuenta incluso baja frecuencia y variantes raras y que prueban que variantes raras están involucradas en varias enfermedades (Bodmer & Bonilla, 2008; Fearnhead, Winney, & Bodmer, 2005; Gorlov et al., 2007; Manolio et al., 2009; McCarthy et al., 2008) (Figura 1.4), que generalmente confieren un mayor riesgo de enfermedad que las variantes comunes (Babron, de Tayrac, Rutledge, Zeggini, & Genin, 2012).

En el mismo sentido está ganando importancia el estudio de las regiones genómicas codificantes, enriquecido en supuestamente variantes funcionales raras y posibles candidatos para explicar la heredabilidad en enfermedades complejas (Abdellaoui et al., 2013; Heath et al., 2008).

1.1.6 Aplicaciones

La identificación de variantes patogénicas en el genoma de los pacientes constituye una poderosa herramienta para los sistemas de salud. Se está demostrando que estas investigaciones mejoran la predicción del riesgo de enfermedad, proporcionan la identificación de dianas terapéuticas, permiten el estudio de la relación causal biomarcador / validación de dianas además de avances importantes en el campo de la farmacogenómica como se ha reiterado en este trabajo por ser una parte fundamental de los estudios genéticos (Klempner, Janjigian, & Wainberg, 2023). Debemos recordar también que puede haber descubrimientos que surjan de la mejora del conocimiento de la base molecular de la enfermedad, aparte de los ya esperados.

En la Figura 1.6 se muestra un resumen de las aplicaciones de los GWAS (Y. M. Zhang, Jia, & Dunwell, 2019).

[3] Se trata del marcador genético más significativo asociado con diabetes mellitus Tipo 2 hasta la fecha. (Vaquero et al., 2012).

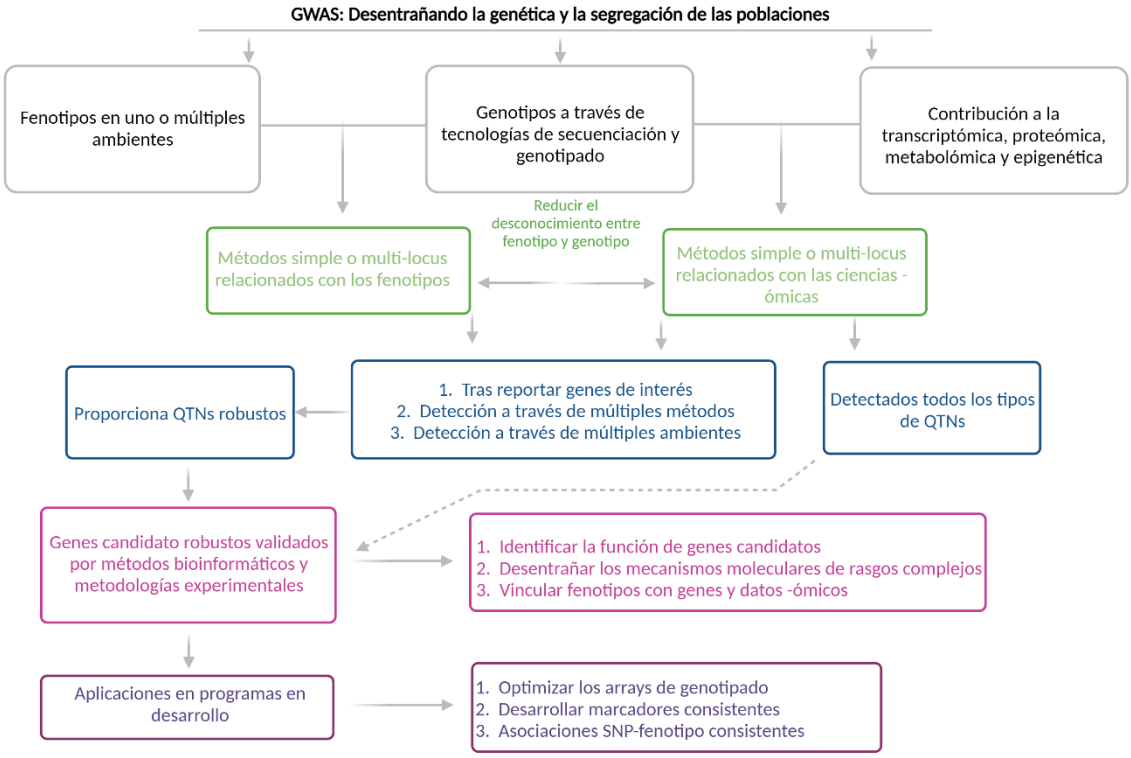


Figura 1.6. Aplicaciones de los GWAS. Adaptada de "*The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits* (Y. M. Zhang et al., 2019)". Realizado con www.Biorender.com.

La identificación de las variantes de susceptibilidad que subyacen a la etiopatogenia de la enfermedad ha aumentado la confianza de que esta información puede traducirse en mejoras clínicamente beneficiosas en la atención al paciente. Hay dos rutas principales a través de las cuales dicha traducción podría verse afectada. En la primera, la identificación de nuevas vías causales proporciona nuevas oportunidades para avances clínicos de beneficio genérico para todos aquellos que padecen (o corren el riesgo de padecer) la enfermedad en cuestión. La identificación de dianas terapéuticas puede conducir a nuevos agentes terapéuticos. La identificación de biomarcadores mejora la predicción de la enfermedad y el seguimiento de la progresión de la enfermedad y la respuesta al tratamiento. La identificación de factores ambientales que contribuyen a la enfermedad permite medidas de prevención. Incluso las asociaciones moderadas de genotipo-fenotipo pueden ofrecer nuevas oportunidades a través de la identificación de nuevas vías modificables. La segunda ruta traslacional radica en el uso del conocimiento de los patrones individuales de predisposición a la enfermedad para desarrollar enfoques más personalizados para el manejo de la enfermedad, incluido el uso de diagnósticos y pronósticos personalizados para mejorar la optimización terapéutica.

Podemos afirmar que los estudios de asociación de genoma completo resultan muy útiles en la evaluación del riesgo para determinadas enfermedades. En 2007, un estudio del WTCCC resultó muy relevante por el gran tamaño de muestra estudiado y la búsqueda de variantes genéticas implicadas en siete enfermedades muy diferentes: 14.000 casos en total (2.000 para cada enfermedad) de trastorno bipolar, enfermedad coronaria, enfermedad de Crohn, presión arterial alta, artritis reumatoide y diabetes tipo 1 y 2 (The Wellcome Trust Case Control, 2007), en comparación con un conjunto compartido de 3.000 controles. Además, el proyecto involucró a más de 50 grupos de investigación en todo el Reino Unido.

Al estudiar simultáneamente enfermedades con diferentes etiologías y contribuciones genéticas, el Consorcio esperaba obtener una idea no solo de la arquitectura genética específica de cada enfermedad (el número de genes que contribuyen y el tamaño de sus efectos), sino también las diferencias entre ellas.

El estudio reveló 24 asociaciones estadísticamente significativas entre las enfermedades y los polimorfismos específicos de un solo nucleótido. Además, identificaron una serie de señales con niveles de significancia menor que posteriormente se demostró que albergaban asociaciones reproducibles en estudios más grandes. La única enfermedad para la que no se encontraron asociaciones fue la presión arterial alta, pero esto se explicó más tarde por el descubrimiento de que la arquitectura genética de esta enfermedad difiere de la de las otras seis enfermedades analizadas, que implican muchas variantes que tienen un pequeño efecto. Tales variantes son detectables en GWAS de mayor tamaño, y desde entonces se han identificado más de 100 regiones asociadas con la hipertensión arterial (Warren et al., 2017).

En reconocimiento al éxito del WTCCC y para capitalizar el logro del enfoque GWA, se financió una nueva ronda de estudios GWA en abril de 2008 (Donnelly, 2008). Estos incluyeron 15 estudios del WTCCC en colaboración con otras instituciones y 12 estudios independientes, con un total de aproximadamente 12.0000 muestras. Muchos de los estudios representan las principales redes internacionales de colaboración que juntas han reunido grandes colecciones de muestras.

El WTCCC2 llevó a cabo estudios de asociación de genoma completo para el análisis de 13 enfermedades: espondilitis anquilosante (The Australo-Anglo-American Spondyloarthritis et al., 2011), esófago de Barrett y adenocarcinoma esofágico (The Esophageal Adenocarcinoma Genetics et al., 2012), glaucoma (The Blue Mountains Eye & The Wellcome Trust Case Control, 2013), accidente cerebrovascular isquémico (The International Stroke Genetics et al., 2012), esclerosis múltiple (The International Multiple Sclerosis Genetics & The Wellcome Trust Case Control, 2011), preeclampsia (Bauer et al., 2018; Morgan et al., 2014), enfermedad de Parkinson (U. K. P. s. D. Consortium et al., 2011), endofenotipos de psicosis (Psychosis Endophenotypes International et al., 2014), psoriasis (Genetic Analysis of Psoriasis & the Wellcome Trust Case Control, 2010), esquizofrenia (Irish Schizophrenia Genomics & the Wellcome Trust Case Control, 2012), colitis ulcerosa y leishmaniasis visceral (Leish et al., 2013).

El WTCCC2 también investiga la genética de las habilidades de lectura y matemáticas en los niños (Davis et al., 2014) y la farmacogenómica de la respuesta a las estatinas (The Go, Group, & The Wellcome Trust Case Control, 2010). Se analizaron más de 60.000 muestras utilizando chips de Affymetrix (actualmente ThermoFisher) e Illumina. Para estos estudios dentro del WTCCC y también para otros estudios independientes, el WTCCC genotipó primero 6.000 controles comunes: 3.000 procedentes de la Cohorte Británica de Nacimientos de 1958 y 3.000 de la Colección de Donantes de Sangre del Reino Unido, cuyos datos se pusieron a disposición de la comunidad científica de manera inmediata. Los datos generados a partir de las muestras de casos con las enfermedades analizadas en el WTCCC2 también estuvieron disponibles una vez publicados (http://www.well.ox.ac.uk/home); por lo tanto, el WTCCC también ayudó a propulsar una revolución en la distribución de datos. El estudio fue uno de los primeros GWAS en proporcionar información sobre el genotipo de cada participante y los rasgos asociados para su uso por parte de la comunidad científica. Aunque el acceso a estos datos se controló posteriormente para garantizar la confidencialidad de los participantes (Homer et al., 2008), la tradición de compartir datos de acceso abierto y la colaboración iniciada por el WTCCC ha continuado.

Tal y como indica el investigador Peter Donnelly (Donnelly, 2008), cuyo grupo, en colaboración con otros ha sido clave y pionero en este tipo de estudios desarrollando y

aplicando métodos estadísticos de vanguardia que utilizan datos de variación de todo el genoma para detectar y caracterizar la estructura y la mezcla de la población humana, podría decirse que uno de los hallazgos más interesantes de los estudios de GWA es una región de 120 kilobases del cromosoma 9 que está asociada con enfermedad de la arteria coronaria (Helgadottir et al., 2007; McPherson et al., 2007; The Wellcome Trust Case Control, 2007). El mecanismo por el cual esta región contribuye a la enfermedad es todavía desconocido, pero los dos genes que mapean más cerca a la señal de asociación, *CDKN2A*, que codifica p16 y también conocido como *INK4A* y *CDKN2B*, que codifica p15, también conocido como *INK4B*, están involucrados en la regulación del ciclo celular. Estos genes no se habían sugerido anteriormente como candidatos para la susceptibilidad a enfermedad cardiovascular, pero se sabe que tienen un papel en varios tipos de cáncer. Otras variantes en esta región del cromosoma 9 están asociadas con la diabetes tipo 2 (T2D) (Diabetes Genetics Initiative of Broad Institute of et al., 2007; Scott et al., 2007; Zeggini et al., 2007) y melanoma (Manolio, Brooks, & Collins, 2008).

Otra superposición implica dos *loci* asociados con T2D y cáncer de próstata (Y. Liu et al., 2022). Un locus contiene el gen *TCF2* (factor de transcripción 2, también conocido como *HNF1B*), que presenta una variante que confiere riesgo de desarrollar cáncer de próstata, pero brinda protección contra T2D. Por el contrario, el otro locus, que contiene *JAZF1* (codifica una proteína *zing-finger*), presenta variantes que se asocian con ambas enfermedades (J. Gudmundsson et al., 2007; Y. Liu et al., 2022; Thomas et al., 2008; Zeggini et al., 2008).

La enfermedad coronaria (EC) continuó siendo estudiada en numerosos trabajos. En 2010 se planteó si las variantes de riesgo interaccionan entre sí produciendo un efecto superior al de la simple suma del riesgo de cada variante individual en determinadas enfermedades, ya que los modelos con interacciones (epistasis) también son consistentes con datos observables (Carlborg & Haley, 2004; Zuk, Hechter, Sunyaev, & Lander, 2012). Carla Lluís-Ganella y colaboradores llevaron a cabo un análisis in sílico de una muestra de 7.368 casos de una base de datos del WTCCC, formada por 1.988 casos de EC y 5.380 controles. A partir de los datos publicados en la literatura médica, se seleccionaron 9 variantes con una asociación probada con el aumento de riesgo de EC. Dado que cada individuo posee dos copias de cada variante, puede ser heterocigoto u homocigoto o no tener ninguno de los dos alelos que sea de riesgo. Así pues, el número total de posibles alelos de riesgo en un determinado individuo sería 18. Las 9 variantes seleccionadas manifiestan su riesgo de EC de manera independiente de los factores de riesgo clásicos, como colesterol, diabetes mellitus o hipertensión. Los investigadores presentan un análisis original e interesante que pone de manifiesto que el número de variantes de riesgo por individuo oscilaba entre 1 y 13 (con una mediana de 7). En el caso de la EC los resultados del análisis de este trabajo indicaron claramente que el riesgo es mayor cuanto mayor es el número de alelos de riesgo. El riesgo era aditivo, con una relación de tipo lineal, que descarta toda interacción entre genes e indica que el riesgo acumulativo es simplemente el total del riesgo que comporta cada una de las variantes individuales (Lluis-Ganella et al., 2010).

También se han aplicado estos estudios en campos como la ginecología (Rahmani et al., 2013) y neurología (Feulner et al., 2010) entre muchos otros, confirmando que los GWAS son una valiosa herramienta para la identificación de variantes genéticas asociadas con las diferentes enfermedades.

Como ya se ha apuntado, otros estudios de GWA a tener en cuenta son los llevados a cabo por el Consorcio de Cáncer de Mama (BCAC, del inglés *Breast Cancer Association Consortium*) (Ku, Loy, Pawitan, & Chia, 2010). Los estudios de BCAC son los primeros GWAS publicados resultantes de un esfuerzo a gran escala y colaboración de múltiples países para investigar la base genética del cáncer de mama. El estudio se llevó a cabo en tres etapas con un tamaño de muestra total superior a 50.000 y permitió identificar varios *loci* de susceptibilidad

nuevos para cáncer de mama. Uno de ellos contiene el gen *FGFR2*, particularmente interesante, ya que codifica un receptor de tirosina quinasa que se sobreexpresa en el cáncer de mama (Easton et al., 2007). Al mismo tiempo, la asociación *FGFR2* también fue descubierta por otro GWAS para cáncer de mama (Hunter et al., 2007). A partir de este momento, en estudios posteriores (Kramer et al., 2020), se fueron identificando nuevos *loci* asociados a esta enfermedad. Estos descubrimientos proporcionan una mayor comprensión de la susceptibilidad genética al cáncer de mama y mejorarán la utilidad de las puntuaciones de riesgo genético para la detección y la prevención individualizadas (Ghoussaini et al., 2012; Michailidou et al., 2017; Qin et al., 2013).

Los resultados de GWAS para cáncer colorrectal también merecen ser mencionados (Ku et al., 2010), ya que estos estudios proporcionaron la primera evidencia que muestra que el locus 8q24 estaba asociado con más de un cáncer (Tomlinson et al., 2007; Zanke et al., 2007). Los dos estudios citados no identificaron otras asociaciones de SNPs para cáncer colorrectal excepto el locus 8q24, que se encontró previamente para cáncer de próstata (Julius Gudmundsson et al., 2007; Yeager et al., 2007). De hecho, estudios posteriores han demostrado la asociación de 8q24 y 10q24 con múltiples cánceres (Jordahl et al., 2022).

A menudo, debido a que los trastornos complejos son de importancia común y de salud pública, se pueden encontrar numerosos trabajos de investigación acerca de la misma patología con el fin de encontrar pruebas suficientes para apoyar la participación de la genética en la etiología del trastorno bajo investigación.

Otro campo que también se aborda mediante estudios de asociación de genoma completo es la epidemiología genética, que estudia la interacción entre los factores genéticos y ambientales que dan origen a las enfermedades del ser humano. Valiéndose de marcadores genéticos desarrollados a través de la biología molecular, de complejos algoritmos informáticos y de amplias bases de datos, la epidemiología genética se ha desarrollado notablemente durante los últimos 10 años.

La aplicación de métodos epidemiológicos para el estudio de la genética en enfermedades complejas ha ido en aumento. Diseños poblacionales o en familias pueden usarse para probar modelos que relacionen susceptibilidad genética y factores de riesgo ambientales. En particular, el estudio de casos y controles se emplea como enfoque para abordar el papel de los factores genéticos y sus interacciones con otros genes y factores ambientales (Bishop, 1994).

En estudios de epidemiología genética, los casos y controles se clasifican según la presencia o ausencia del genotipo, así como la presencia o ausencia de un factor de riesgo como un factor ambiental. En estos trabajos se analiza si la exposición al factor de riesgo modifica el genotipo. De todos modos, para entender la interacción entre los factores de riesgo ambientales y factores genéticos de riesgo, son necesarios estudios prospectivos de cohortes y también deben llevarse a cabo ensayos clínicos (utilizando grandes cantidades de datos de genotipado) para poder evaluar cómo la información genética se debe usar a la hora de elegir tratamientos (Donnelly, 2008).

Una oportunidad excepcional radica en estudios de reacciones adversas a medicamentos (ADRs, del inglés *Adverse Drug Reactions*) u otros tratamientos, en los que los tamaños del efecto a menudo son grandes y pueden ser directamente relevantes para la atención clínica (Chan, Jin, Loh, & Brunham, 2015) y los estudios de GWAS han contribuido enormemente a la farmacogenética y farmacogenómica (de With et al., 2023).

Las ADRs son muy importantes ya que son una causa evitable de morbilidad y mortalidad. La variación genómica de la línea germinal contribuye a las diferencias interindividuales en la respuesta a fármacos y al riesgo de ADRs. Existen reseñas previas de GWAS en farmacogenómica, cuando solo un pequeño número de *loci* había alcanzado niveles de

importancia estadística (Daly, 2010). Desde ese momento, ha habido un rápido aumento en el número de GWAS realizados para el estudio de ADRs. Sin embargo, estos, todavía representan una pequeña fracción del número total de GWAS reportados. En cualquier caso, se ha observado un aumento en la proporción de GWAS para ADRs realizados en poblaciones de países no europeos, lo que ha llevado al reconocimiento de marcadores específicos de población, destacando la naturaleza compleja de la genética de los ADRs.

Una vez más, en este tipo de estudios se ha puesto de manifiesto la existencia de abundantes variantes genéticas raras que contribuyen al riesgo de enfermedades complejas y que están geográficamente localizadas. Estas dianas son muy importantes para el desarrollo de la farmacogenómica, ya que muchas son perjudiciales y tienen relevancia para comprender el riesgo de enfermedad (Nelson et al., 2012).

En definitiva, el objetivo último y común en este campo radica en extender los beneficios de la investigación farmacogenómica a todas las poblaciones mundiales para mejorar la eficacia de los medicamentos y reducir los ADRs, para lo que es necesario comprender la contribución genética a la enfermedad humana, que requiere el conocimiento de la abundancia y distribución de la diversidad genética funcional dentro y entre las poblaciones.

1.2 LOS GWAS Y LA GENÉTICA

La genética poblacional tuvo un importante desarrollo en los años 30 del siglo pasado, principalmente por los trabajos de Ronald Fisher y Sewall Wright, incluso antes de conocerse el ADN como la unidad última de herencia. Ambos unieron los conceptos de selección natural de caracteres fenotípicos con los de la herencia mendeliana demostrando que alelos discretos podían estar fundamentados en rasgos continuos. Con el tiempo, la descripción de diversidad genotípica a nivel molecular y el hecho de que la selección no sirviera como el único proceso capaz de explicar los niveles de polimorfismos, exigieron desarrollos importantes en la teoría de la genética poblacional. Hoy en día sabemos que los cambios de frecuencias alélicas a lo largo del tiempo son las pistas que nos permiten investigar los procesos evolutivos. Entendiendo los mecanismos por los que las fuerzas de la evolución actúan sobre estas frecuencias, se pueden generar modelos matemáticos que se aproximen a la realidad, necesarios para entender la sutil interconexión entre dichas fuerzas, así como para permitirnos inferir procesos pasados a partir de la diversidad actual. La genética de poblaciones es la rama de la genética cuyo objetivo es describir la variación y distribución de la frecuencia alélica para explicar los fenómenos evolutivos. Para ello, la genética de poblaciones define a una población como un grupo de individuos de la misma especie que están aislados reproductivamente de otros grupos afines, un grupo de organismos que comparten el mismo hábitat y se reproducen entre ellos. Estas poblaciones, están sujetas a cambios evolutivos en los que subyacen cambios genéticos, los que a su vez están influidos por factores como la selección natural, la deriva genética, el flujo genético, la mutación y la recombinación genética. Así, la genética de poblaciones es un elemento esencial de la síntesis evolutiva moderna.

Los estudios genéticos de las poblaciones, como hemos visto, son la base de los GWAS. Su importancia radica en el análisis de la distribución de la variabilidad genética de los individuos vivos, caracterizando diferencias en la composición genética de las poblaciones y permitiendo evaluar la posible existencia de diferentes patrones de estratificación poblacional para las variantes genéticas comunes y raras. Así los GWAS y el estudio genético de las poblaciones están íntimamente ligados, ya bien sea para analizar la estructura genética de las poblaciones humanas modernas como para evaluar el efecto que las variantes de baja frecuencia pueden tener sobre la subestructura genética total.

Mediante este tipo de estudios y cuanto más exhaustivos sean estos, será posible llevar a cabo la caracterización de la estructura poblacional y el efecto de variantes comunes y raras en la patogénesis de la enfermedad. Son estas variantes las que explican gran parte de la diversidad genética en nuestra especie, una consecuencia del corto período evolutivo y la ascendencia compartida de la población humana.

Proyectos como la generación del mapa de haplotipos del genoma humano (HapMap) han permitido guiar el diseño y análisis de los estudios de asociación genética, arrojar luz sobre la variación estructural y la recombinación e identificar *loci* que pueden haber estado sujetos a selección natural durante la evolución humana (International HapMap, 2005).

A pesar de existir datos de diversas poblaciones, la mayoría de los GWAS se han realizado en poblaciones europeas para diversas enfermedades y rasgos, tal como hemos destacado. A partir del descubrimiento de la implicación del gen *KCNQ1* en la diabetes tipo 2 (T2D) en dos GWAS llevados a cabo en la población japonesa y replicada la asociación en otras poblaciones asiáticas y europeas (Unoki et al., 2008; Yasuda et al., 2008), se hizo evidente que podría ser posible el descubrimiento de más variantes si se realizasen más GWAS y otros estudios poblacionales. Estos estudios han subrayado, por tanto, la importancia y el valor de extender GWAS a diferentes poblaciones (Ku et al., 2010), siendo, estas investigaciones, cada vez más numerosas (Deak et al., 2022).

Curiosamente, el gen *KCNQ1* no reveló asociación para T2D en GWAS previos europeos. Esto es debido a una marcada diferencia en la frecuencia alélica, lo que resultó en una menor potencia estadística para detectar asociación en poblaciones europeas. De hecho un estudio ha demostrado una amplia variación en las frecuencias alélicas en diferentes poblaciones para los SNPs identificados por GWAS para varias enfermedades complejas y rasgos (Adeyemo & Rotimi, 2010). También se identificó un nuevo alelo de riesgo para el cáncer de mama por un GWAS llevado a cabo en una población china y esto fue replicado en mujeres de ascendencia europea. Como en el caso de T2D, el alelo de riesgo no fue detectado por varios GWAS europeos de cáncer de mama (Zheng et al., 2009). Resultados de un GWAS de lupus eritematoso sistémico también apoya la presencia de heterogeneidad genética para la susceptibilidad a esta enfermedad entre población china Han y población europea (J. W. Han et al., 2009).

Las poblaciones ancestralmente diversas y no europeas han sido infraestudiadas (Popejoy & Fullerton, 2016). Los GWAS notables en estas poblaciones incluyen estudios de afecciones cardíacas en afroamericanos (Evans et al., 2016) y apnea del sueño en hispanos y latinoamericanos (Cade et al., 2016). Uno de los siguientes pasos será identificar asociaciones en poblaciones poco estudiadas, como las de África y América Latina, y en pueblos aislados e indígenas, como los del Ártico y las islas del Pacífico. Peter Donnelly y colaboradores han demostrado en numerosos trabajos que es necesario evaluar el impacto de la estructura de la población en los estudios de asociación (Pritchard & Donnelly, 2001).

Uno de los mejores ejemplos conocidos trata sobre un estudio de asociación de T2D en tribus Pima y Papago (nativos americanos), que sufren de una tasa extremadamente alta de diabetes (Knowler, Williams, Pettitt, & Steinberg, 1988). Los datos indicaron una fuerte asociación negativa entre diabetes y un haplotipo en el locus de la inmunoglobulina G. Sin embargo, muchas de las personas incluidas en la muestra tenían datos de ascendencia europea reciente, y se encontró que el promedio fue más alto en los controles con ascendencia europea que en los individuos afectados. El haplotipo en cuestión se mostró con una frecuencia mucho más alta en los europeos en general, independientemente del fenotipo, y los autores demostraron que el efecto protector de este haplotipo desaparecía si el análisis era estratificado en función de la ascendencia informada.

El enfoque principal en el grupo Donnelly es el desarrollo y la aplicación de métodos estadísticos para comprender la variación genética, y su asociación con la variación fenotípica y la susceptibilidad a la enfermedad. Estos métodos generalmente combinan enfoques estadísticos modernos de cálculo intensivo con ideas de modelos de genética de poblaciones, y apuntan a obtener la mayor cantidad de información posible de los grandes conjuntos de datos actualmente generados por técnicas experimentales de alto rendimiento mediante estudios de asociación de todo el genoma.

Hoy en día, los estudios transétnicos son muy utilizados para añadir valor a los hallazgos y son casi un requisito de los grandes GWAS actuales.

Muchos estudios han comenzado a determinar la abundancia, la distribución y los efectos fenotípicos de variantes raras y su papel en enfermedades complejas (Coventry et al., 2010; Gravel et al., 2011). Se ha observado que la variación rara es específica de la población y que se pueden apreciar diferencias incluso dentro de Europa, muy probablemente debido a la historia demográfica. Nelson y colaboradores (Nelson et al., 2012) observaron un gradiente norte-sur en la abundancia de variantes raras en Europa, con un mayor número de variantes raras en el sur de Europa y un número muy pequeño de variantes entre los finlandeses, lo que es consistente con los gradientes observados en la diversidad haplotípica (Lao et al., 2008) y un cuello de botella ancestral finlandés (Salmela et al., 2008). Una vez más se confirma que la historia de la población es importante para los patrones de variación genética.

No olvidar que la premisa de todo estudio de asociación de genoma completo, tal y como hemos estado viendo hasta ahora, es el especial cuidado en la selección de los casos y controles, que deben ser tomados de poblaciones que sean generalmente comparables, tanto en términos de antecedentes genéticos como exposiciones ambientales (Manolio, Bailey-Wilson, & Collins, 2006)**.** También es muy importante el análisis de la estratificación de la población. Estos dos factores constituyen la base fundamental de nuestro estudio. La identificación de dicha estructura permite el estudio de la historia reciente de la población e identifica los puntos débiles en los estudios de asociación, particularmente cuando se prueban variantes raras, a menudo surgidas recientemente (Bycroft et al., 2019).

1.3 MARCADORES GENÉTICOS

Hace casi un siglo, los antropólogos comenzaron a analizar la variación humana mediante polimorfismos genéticos, también llamados marcadores genéticos moleculares. El término de **polimorfismo** fue definido por Ford (1940) como: "la aparición conjunta en un lugar de dos o más formas discontinuas de una especie, de tal modo que la más rara de ellas no se puede mantener simplemente a través de la mutación genética". En la práctica, para que un locus sea considerado polimórfico el alelo más común para ese locus debe tener una frecuencia poblacional menor del 99%. De acuerdo con la ley de Hardy-Weinberg, al menos un 2 % de la población debe ser heterocigoto para ese locus.

Muchos otros bioquímicos describieron y usaron marcadores que detectan la variación a nivel del producto del gen (nivel de aminoácidos), pero la secuencia aminoacídica, aunque relativamente variable entre poblaciones. El esudio en polimorfismos proteicos y enzimáticos supusouna enorme revolución en genética de poblaciones pero eran limitados: mostraban la variación a nivel de la expresión génica, pero no en el propio gen, lo que causó una importante pérdida de información debido a los efectos de la selección natural y a la existencia de variantes neutras y sinónimas.

Hubo tres grandes avances teóricos y tecnológicos que permitieron el uso de los polimorfismos de ADN como fuente de información más conveniente con respecto a la variabilidad en la población y en el individuo.

En primer lugar, en 1952, el ADN se confirmó como la base del material genético y la heredabilidad (Hershey & Chase, 1952) a pesar de que su existencia fue descubierta ya en 1869 por Friedrich Miescher quien encontró la molécula al inspeccionar el esperma de salmón y el pus de heridas abiertas y lo llamó nucleína. Poco después, en 1953, la estructura molecular del ADN se describió como una doble hélice (Watson & Crick, 1995). Finalmente, en 1980, Frederick Sanger desarrolló un método para la secuenciación del ADN (Sanger, Coulson, Barrell, Smith, & Roe, 1980). Este método de secuenciación todavía sigue siendo el estándar de secuencia.

Las variaciones a nivel de ADN estuvieron de repente disponibles para la investigación; inserciones, deleciones, duplicaciones e inversiones, largas modificaciones moleculares de secuencias de ADN que eran mucho más variables y estables que las proteínas, aumentando conjuntamente la resolución de los estudios de población. Después de estos eventos se produjeron cuatro revoluciones tecnológicas que modificaron nuestra comprensión de la genética humana y de la estructura de la población; estos son los RFLPs (Polimorfismos en la Longitud de Fragmentos de Restricción), la PCR (del inglés *Polymerase Chain Reaction*), el genotipado automático, el decubrimiento de los microsatélites y la NGS (de inglés *Next-generation sequencing*) que permitió descurbrir SNPs, InDels y otras variaciones más pequeñas.

Los RFLPs fueron utilizados por primera vez por Wyman y White en 1980. La técnica consiste en el uso de enzimas de restricción que son capaces de cortar el ADN exactamente a ambos lados de las regiones de interés sujetas a estudio. Pronto, en 1985, esta técnica permitió el descubrimiento de marcadores VNTR (de inglés *Variable Number of Tandem Repeats*) descritos por Hill y Jeffreys (Hill & Jeffreys, 1985); (Figura 1.7), es decir el polimrofismo de las regiones minisatélites.

Figura 1.7. Estructura de los VNTRs. Adaptado de http://www.biologia.arizona.edu/human/problem_sets/DNA_forensics_1/05t.html.

Los VNTRs son regiones hipervariables del genoma, que tienen el potencial de otorgar la explicación de la herencia genética y la diferenciación individual y, de acuerdo con el tamaño

de su motivo de repetición, podrían clasificarse como minisatélites (10-60 pb) o microsatélites (2-5 pb, también llamados *Short Tandem Repeats* o STRs y muy similares a los VNTRs).

En su trabajo, Jeffreys y colaboradores diseñaron sondas multi*loci* que permitieron la identificación simultánea de muchas regiones con polimirfismo de minisatélites, de tal forma que se podían describir patrones individuales y personalizados, lo que hoy en día se conoce como huella digital de ADN.

Las STRs son secuencias repetidas en tándem (una al lado de la otra) de entre 2 y 5 nucleótidos (Figura 1.8), característica que los diferencia de los minisatélites. En la identificación forense se usan, generalmente, las STRs de 4 o 5 nucleótidos. El tamaño total de un STR es más pequeño que el de los minisatélite, siendo inferior de 1 kb (1kb = 1000 bases). La secuencia de las regiones flanqueantes a ambos lados del STR es igual en todos los sujetos, por lo que son polimorfismos de longitud.

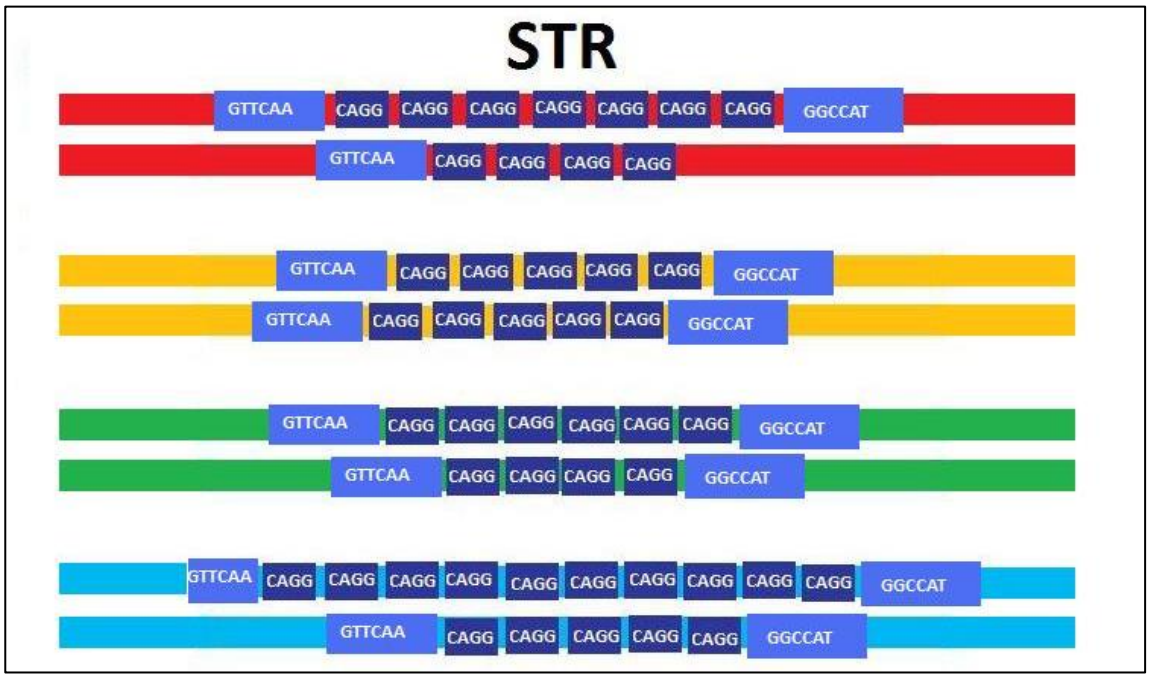


Figura 1.8. Estructura de las STRs. Esta imagen muestra de una forma muy sencilla qué son los STRs. Esquematiza, en color azul oscuro, secuencias de cuatro nucleótidos (CAGG, en este ejemplo) repetida en diferentes cromosomas un número diferente de veces (STR), siendo para el par de cromosomas rojos 7 y 4 veces, para los amarillos 5 veces cada uno, para el par de cromosomas verdes 6 y 4 veces y para el par de color cian, 10 y 5 veces. En azul claro tenemos representadas las secuencias flanqueantes de las STR. Adaptado de https://forensemolecular.es.tl/STRs.htm.

Los minisatélites demostraron ser bastante informativos, representando tanto polimorfismos de longitud como de secuencia, pero son técnicamente muy difíciles de usar y tenían probelmas de reproducibilidad. Esto fue superado mediante el uso de sondas de locus únicas (SLP) en lugar de sondas para múltiples *loci* (MLP) (Y. Nakamura, Carlson, Krapcho, Kanamori, & White, 1988). Sin embargo, la concentración extremadamente alta de ADN necesaria para el funcionamiento de los RFLPs (> 50 ng) y su dependencia de la existencia de cadenas de ADN bien conservadas (> 12 kpb) impidió que esta técnica se estandarizase en estudios de población.

Fue la introducción de la PCR (Mullis & Faloona, 1987) lo que permitió la estandarización del uso de algunos minisatélites de tamaño pequeño y de los microsatélites. La PCR faculta la producción de una gran cantidad de copias de las regiones de ADN de interés, lo que permite

la posterior detección y análisis a partir de una pequeña cantidad de ADN inicial. Al principio, la PCR se utilizaba para analizar, de forma rápida y sencilla, tanto minisatélites como microsatélites, pero hoy en día los polimorfismos bialélicos (SNPs, que serán descritos a continuación) también se pueden analizar con esta técnica.

El genotipado automático del ADN se introdujo a principios de los años 90, lo que dio luz verde a estudios a gran escala en campos como la investigación clínica y forense, genética evolutiva y de poblaciones.

Hoy en día, el genotipado automático ha llegado a su culminación con el desarrollo de la genotipificación de alto rendimiento. Se trata más de un salto de carácter cuantitativo que cualitativo, ya que la secuenciación de Sanger sigue siendo el estándar en términos de calidad de genotipado, pero desde el genotipado de solo unas pocas variantes en algunas muestras ya se pueden procesar miles de muestras y posiciones genómicas simultáneamente. Empresas como Illumina y ThermoFisher Scientific (https://www.illumina.com/, https://www.thermofisher.com/es/es/home/life-science/microarray-analysis.html), proveen cada vez de más chips que contienen cientos de miles de marcadores que cubren un espectro muy amplio de posiciones genómicas, algunos de ellos específicamente orientados a regiones exónicas, genoma completo y enfocados a la clínica. En cualquier caso, todos ellos tratan de determinar el contenido alélico de las diferentes regiones o posiciones del genoma y se diseñan específicamente a priori.

Debido a que el 99,7 % del genoma es idéntico en todos los humanos, es importante identificar y conocer qué secuencias del ADN pueden diferenciar a los individuos entre sí. Para ello, los marcadores genéticos más utilizados actualmente son los polimorfismos de un solo nucleótido o SNPs (Figura 1.9), que son, con mucha diferencia, la forma más abundante de variación genética en el genoma humano. Los SNPs son cambios de un solo par de bases en la secuencia de ADN y ocurren con alta frecuencia en el genoma humano (Genomes Project et al., 2010). Constituyen los más simples y más comunes polimorfismos genéticos a lo largo del genoma humano. Generalmente tienen dos alelos, es decir, dentro de una población hay dos posibilidades comunes de pares de bases para una ubicación de un SNP. La razón por la cual la mayoría de los SNPs son solo bialélicos es porque se trata de marcadores muy estables, con tasas de mutación muy bajas, en torno a 10^{-9}, mucho más bajas que los microsatélites (Thomson et al., 2000). Por lo tanto, la probabilidad de que una sola posición mute dos veces de forma independiente es extremadamente baja; esto, junto con la baja probabilidad de que estas dos mutaciones produzcan dos alelos diferentes, hacen aún más difícil que un SNP se vuelva multialélico. A pesar de la baja probabilidad mencionada, se han demostrado muchos cambios de nucleótidos, dando lugar a tres o incluso a cuatro alelos.

Otra característica interesante de estos marcadores es que, debido a su carácter bialélico y baja tasa de mutación, pueden exhibir una subestructura de población más acentuada que los STRs (Chakraborty, Stivers, Su, Zhong, & Budowle, 1999). Además, los SNPs no sinónimos, que están bajo selección natural, pueden originar una subestructura de población aún más acentuada.

Por su amplio espectro, se encuentra uno cada 300 - 400 pb a lo largo de todo el genoma, lo que ofrece la posibilidad de utilizarlos para obtener perfiles detallados de los genes involucrados en las enfermedades. La frecuencia de un SNP se da en términos de la frecuencia de los alelos menores o la frecuencia del alelo menos común. Por ejemplo, un SNP con una frecuencia de alelo menor (G) de 0,40 implica que el 40% de una población tiene el alelo G frente a los más comunes (el alelo principal), que se encuentra en el 60% de la población (Bush & Moore, 2012). Los SNPs son utilizados como marcadores de una región genómica con fines de estudios genéticos debido a estas características peculiares.

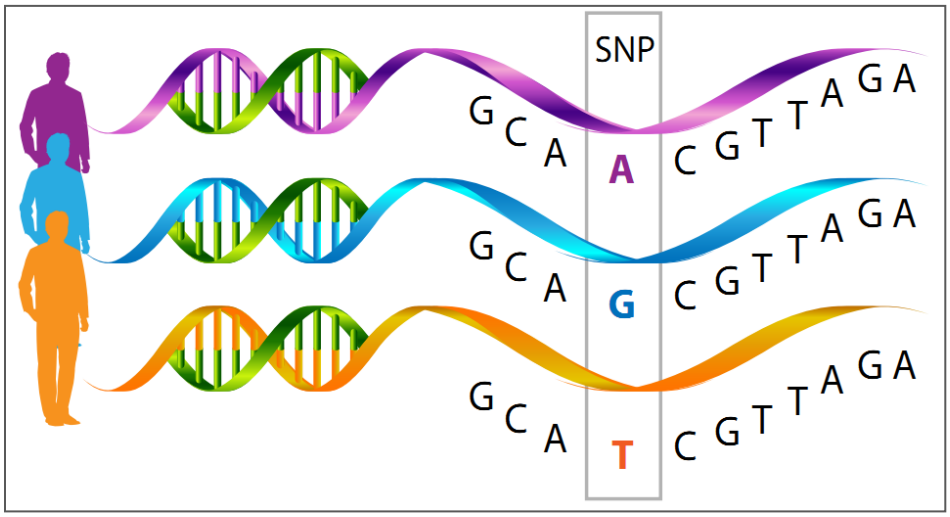


Figura 1.9. Representación de un SNP. Fuente propia. Realizado con www.Biorender.com.

Algunos SNPs pueden tener consecuencias funcionales, ya que ocasionan cambios aminoacídicos, cambios en la transcripción y estabilidad del ARNm (Ácido Ribonucleico mensajero) así como cambios en la transcripción y afinidad de unión a factores (Griffith et al., 2008). Muchos de ellos están presentes en una gran proporción de poblaciones humanas (International HapMap et al., 2010), siendo, estas variantes, específicas de cada población.

Por el momento, y debido a la miniaturización y automatización de los métodos de detección para la variabilidad molecular del ADN, los polimorfismos binarios, específicamente los SNPs, están adquiriendo un papel muy importante en los estudios genéticos.

También existen los polimorfismos que se basan en la eliminación o adición de una sola base. Estos marcadores bialélicos que comparten todas las características y aplicaciones con los SNPs se denominan InDels, inserciones / deleciones pequeñas (típicamente de 1 a 50 pb) de secuencias de ADN.

Inicialmente, se estimó que una de cada 500 o 1.000 pb mostraría variabilidad, es decir, 36 millones de SNPs en todo el genoma (Sherry, Ward, & Sirotkin, 1999). Hoy en día, después de aumentar enormemente la cantidad de datos disponibles de ADN humano debido a todos los proyectos de secuenciación que se están llevando a cabo, ahora sabemos que la estimación fue, en gran medida, corta: en marzo de 2023, dbSNP (*Single Nucleotide Polymorphism Database*) enumeró 1.053.623.523 de SNPs en humanos (NCBI dbSNP build 155 for human: https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi). En el apartado 1.5.1.3 se describe el Proyecto Internacional HapMap, creado para desarrollar un mapa de haplotipos del genoma humano en el que poder catalogar las regiones genéticas similares y diferentes entre individuos.

Debido a estos hallazgos se dispone de muchos SNPs para elegir, facilitando su selección en términos de conveniencia, informatividad, adecuación de sus regiones flanqueantes o frecuencia alélica, evitando o aprovechando el desequilibrio de ligamiento u otros factores de interés.

Es importante recalcar que, aunque el concepto "SNP" se definió inicialmente como una variante de nucleótidos con una MAF poblacional de al menos 5%, hoy en día las variantes de baja frecuencia comienzan a ser consideradas también para estudios genéticos.

Otra ventaja que presentan los SNPs frente a los microsatélites es su simplicidad, por lo que su detección y análisis es mucho más fácil de automatizar. Además, se producen fragmentos más cortos durante su amplificación por PCR, lo que resulta muy útil cuando se va a utilizar ADN de mala calidad (como es el ADN altamente degradado en la ciencia forense) o cuando muchos SNPs se analizan simultáneamente (GWAS).

La variabilidad de SNPs no codificantes identifica subestructura de la población más que cualquier otro tipo de marcador.

Concretamente, los SNPs son de suma importancia para los estudios de genética de poblaciones. Debido a su baja tasa de mutación, son los marcadores elegidos para analizar los acontecimientos evolutivos antiguos (Jobling & Tyler-Smith, 1995). El ADN acumula lentamente mutaciones tipo SNP en todas las poblaciones, y por lo tanto más poblaciones interrelacionadas compartirán más SNPs que aquellas poblaciones que divergieron a principios de la historia. Del mismo modo, de la variabilidad de SNPs de una población se desprenderá la historia demográfica de sus subpoblaciones.

Hay SNPs especiales llamados AIMs, que exhiben frecuencias sustancialmente diferentes entre poblaciones de distintas regiones geográficas, se utilizan comúnmente para inferir de manera eficiente y económica el origen geográfico y las proporciones de ancestralidad en pobalciones mezcladas (C. Phillips et al., 2007).

Por último, pero no menos importante, los SNPs son los primeros marcadores de elección para el análisis de los rasgos fenotípicos. Esta característica de los SNPs es de especial interés en múltiples disciplinas, como ciencia forense, clínica, epidemiología genética o genética de poblaciones. A continuación, se exponen motivos concretos por los que estos marcadores genéticos son fundamentales en los campos mencionados;

- En la ciencia forense, investigar las características físicas como el color de los ojos o la piel es altamente importante para reducir la cantidad de personas sospechosas cuando no hay muestras de referencia para comparar (Maronas et al., 2014; Ruiz et al., 2013).

- Inferir los fenotipos es posible porque los alelos de los genes están constituidos por una combinación concreta de SNPs, formando un genotipo particular, por lo que al inferir los alelos de los SNPs también inferimos fenotipos. Por otro lado, inferir rasgos físicos como la pigmentación humana es muy complicado, debido al hecho de que estos caracteres son complejos, multifactoriales y poligénicos, y además tienen lugar las interacciones gen-ambiente (Pulker, Lareu, Phillips, & Carracedo, 2007). Hay otras características físicas de aplicación forense que también utilizan SNPs y constituyen parte de lo que se denomina “Fenotipado forense por ADN”.

- En epidemiología genética, los SNPs continúan siendo la elección principal en estudios de genes candidatos y GWAS. Como hemos visto en apartados anteriores, este tipo de análisis se basa en el examen de muchas variantes genéticas comunes en diferentes individuos mientras se busca cualquier asociación entre un rasgo fenotípico y cualquiera de los marcadores analizados. Los GWAS generalmente enfocan los SNPs como primeras opciones para buscar la base genética de las principales enfermedades.

Como se ha descrito a lo largo de la introducción, en estos estudios se comparan genéticamente dos grupos de personas: casos (que expresan la genética del trastorno) y controles (individuos sanos). Cientos de miles de variantes genéticas dispersas a lo largo de todo el genoma se analizan en ambos grupos, normalmente mediante el uso de *arrays* de SNPs (los más utilizados y adecuados se describen en el punto 1.4). Si un alelo de cualquiera de las variantes analizadas es significativamente más frecuente entre las personas con la enfermedad, entonces se dice que el SNP está "asociado" con la enfermedad (Manolio, 2010; T. A. Pearson & Manolio, 2008), (National Human Genome Research Institute) (Welter et al., 2014),

https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/.

Por lo tanto, la asociación de SNPs a rasgos fenotípicos es cada vez más importante en estudios de genética clínica.

Por tanto, en la actualidad se dispone de numerosas herramientas para la investigación genética debido a los importantes y rápidos avances científicos acaecidos; la secuenciación del genoma humano ha pasado de ser un sueño a ser una realidad y con la secuenciación de nueva generación se ha pasado de analizar genes individuales a exomas o a genomas completos. La genotipificación ha pasado de depender principalmente de marcadores microsatélites a utilizar polimorfismos de un solo nucleótido (SNPs). La capacidad para generar datos genotípicos ha saltado desde cientos de genotipos por día hasta cientos de miles por día. Las herramientas para caracterizar la expresión de genes del genoma completo a través de *microarrays* ha surgido junto con una capacidad creciente para caracterizar las proteínas resultantes de una manera sistemática.

Todas estas variantes pueden ser detectadas y evaluadas a lo largo del genoma humano, y permiten estudiar la variabilidad de dicho genoma tanto dentro del individuo como en un contexto poblacional. Las herramientas bioinformáticas también han cambiado radicalmente, siendo actualizadas para mantenerse al día con la gran cantidad de datos que ahora podemos generar.

Debido a estos avances, es más importante que nunca determinar la aplicación de todos estos datos y métodos para la disección de los rasgos humanos.

1.4 *ARRAYS* DE GENOTIPADO UTILIZADOS EN GWAS

Los estudios de asociación de genoma completo fueron posibles gracias a la tecnología de *microarrays* para analizar millones de SNPs. Se han utilizado fundamentalmente dos plataformas de genotipado, que incluyen productos de Illumina (San Diego, California; https://www.illumina.com/) y ThermoFisher, (Santa Clara, California; http://www.thermofisher.com/br/en/home/life-science/microarray-analysis.html). Comercialmente introducidas hace más de 20 años (Dong et al., 2001; Shen et al., 2005) estas dos tecnologías han sido revisadas (Distefano & Taverna, 2011) y ofrecen diferentes enfoques para medir variación de SNPs (Schillert & Ziegler, 2012; Teo, 2012). Se han llevado a cabo numerosos estudios para el descubrimiento y la validación de SNPs incluidos en los *arrays* de estas plataformas (Dong et al., 2001; Sachidanandam et al., 2001).

La tecnología *Axiom Genotyping* es el sistema de genotipado más actual de ThermoFisher para estudios de asociación del genoma completo, replicación y estudios de asociación mediante estrategia de genes candidatos. Permite genotipar entre 800 y 2.600.000 SNPs, y ofrece secuencias cortas de ADN gracias a un lugar en el chip que reconoce una región SNP alelo específica. Los alelos se detectan por hibridación diferencial de la muestra de ADN. Esta tecnología también permite la detección de otro tipo de variantes, como son las CNVs. Concretamente el panel *Axiom Spain BioBank Array Plate*, generado en este trabajo, se ha basado en esta tecnología. Tanto su diseño como la selección de marcadores son descritos en el apartado 3.1.2.1.

La tecnología Axiom ha demostrado ser robusta y ofrece gran especificidad a nivel poblacional, gracias a la selección concreta de variantes para el diseño de sus paneles (Garcia-Etxebarria et al., 2015; Hoffmann et al., 2011; M. Nakamura et al., 2012), lo que permite incluso el mapeo de variantes en poblaciones que presentan mezclas, eficiente para identificar las bases

genéticas de aquellas enfermedades complejas con importantes disparidades étnicas (Kawai et al., 2015; X. Zhang, Mu, Liu, & Zhang, 2014).

También se han ido desarrollando distintos algoritmos con el fin de introducir mejoras en la conversión de las mediciones de intensidad de las muestras en datos brutos, que pueden influir en la calidad de las medidas finales y así generar genotipos fiables y precisos (Carvalho, Bengtsson, Speed, & Irizarry, 2007; Huang et al., 2016; Xiao, Segal, Yang, & Yeh, 2007; H.-C. Yang et al., 2008). En cuanto a los controles de calidad el uso de la metodología estadística moderna también mejora sustancialmente la exactitud y la precisión de los resultados, en relación con los procedimientos *ad hoc* introducidos por los diseñadores y fabricantes de la tecnología (Kvale et al., 2015; Reme et al., 2008).

Por otro lado, la tecnología Illumina permite el genotipado de 700.000 a 5.000.000 de SNPs. El sistema iScan de Illumina® es un revolucionario escáner de alta precisión que consta de un láser de alto rendimiento, óptica y sistemas de detección de alta resolución. Este sistema es compatible con una amplia gama de aplicaciones tanto para estudios de detección de nuevos biomarcadores como estudios de validación. La tecnologia Infinium también es óptima para el análisis de CNVs y análisis de metilación del ADN. La técnica consiste en la extensión de un oligonucleótido alelo específico de un color o una extensión de una base en dos colores. No son necesarios pasos de PCR ni de ligación. Las dianas interrogadas son capturadas y detectadas en *arrays Beadchip* de alta densidad.

Ambas plataformas tienen capacidad para albergar un elevado número de muestras de ADN (J. B. Fan et al., 2006; Peiffer & Gunderson, 2009), y se trabaja generalmente con ADN purificado en solución acuosa para la mayoría de GWAS.

Se han desarrollado gran cantidad de trabajos con la tecnología Infinium que certifican la robustez, el rendimiento y la precisión del ensayo, a la vez que permite el control del sistema de gestión de la información obtenida en el laboratorio y la trazabilidad de las muestras (Gunderson, 2009; Gunderson et al., 2006). Además proporciona acceso a prácticamente cualquier SNP en el genoma y ofrece una alta calidad de datos así como flexibilidad en el diseño del contenido de las matrices (Oliphant, Barker, Stuelpnagel, & Chee, 2002; Peiffer & Gunderson, 2009; Steemers & Gunderson, 2007).

Ambas tecnologías han contribuido inmensamente a nuestra comprensión de los patrones de variación en el genoma humano (International HapMap, 2005; The International HapMap, 2007) y han allanado el camino para numerosos estudios de asociación de genoma completo, cuyo objetivo, tal como hemos visto en los apartados 1.1.6 (Aplicaciones de los GWAS) y 1.2 (Los GWAS y la genética), es el descubrimiento de variantes implicadas en diversas enfermedades definiendo la base genética de muchas y diversas patologías (Manolio et al., 2008).

Una ventaja de estas matrices de SNPs sobre otras técnicas es la medición simultánea de CNVs y la asignación de genotipos. Además esta característica permite a los investigadores identificar regiones en el genoma caracterizadas por la pérdida de heterocigosidad (LOH) (Bacolod et al., 2009).

Además de la tecnología, otra consideración importante son los SNPs que cada plataforma selecciona para los ensayos. Esto puede ser fundamental dependiendo de la población humana específica en estudio. Por ejemplo, es importante usar un chip que tenga más SNPs con mejor cobertura genómica general para un estudio de africanos que para estudios de europeos. Esto se debe a que los genomas africanos han tenido más tiempo para recombinarse y por lo tanto tienen menos DL entre alelos en diferentes SNPs.

Actualmente, uno de los objetivos principales es desarrollar multiplexes robustas para la amplificación de muchos marcadores en una sola reacción. La tecnología para analizar la variación genómica está en constante cambio y con desarrollos continuos.

1.5 GENÉTICA Y BIOINFORMÁTICA

Los estudios de asociación, en particular, se están desarrollando y ampliando, tal y como hemos visto, en respuesta a la gran cantidad de SNPs que se han identificado en los últimos años.

La genómica computacional aborda el uso del análisis de datos computacionales y estadísticos para descifrar la biología del genoma y los datos relacionados, incluidos tanto las secuencias de ARN como de ADN, así como otros datos "postgenómicos". Estos datos se almacenan en bases de datos masivas. La genómica computacional se utiliza en combinación con diferentes enfoques estadísticos y computacionales para entender las complejidades de los genomas. Como tal, la genómica computacional, a la que también se hace referencia como genética estadística y computacional, puede considerarse como un subconjunto de bioinformática y biología computacional que se centra en el uso del genoma completo (en lugar de genes individuales) para comprender los principios de cómo el ADN de una especie controla su biología molecular (Koonin, 2001). La bioinformática, como parte clave en la genómica computacional, a menudo se describe como un campo interdisciplinario que desarrolla tanto *software* como herramientas metodológicas para la comprensión de datos biológicos, combinando informática, estadística, matemáticas e ingeniería para procesar y analizar datos biológicos. Utiliza la programación informática como parte de su metodología y es particularmente útil en campos como la genética y la genómica. Usos comunes de la bioinformática incluyen la identificación de genes y nucleótidos candidatos (SNPs) con el objetivo de mejorar los estudios de poblaciones y la genética evolutiva, la base de las enfermedades o la farmacogenómica.

Tanto la bioinformática como la genómica computacional comparten las mismas raíces. En primer lugar, durante los años 60, Margaret Dayhoff y colaboradores (Hunt, 1983) crearon una matriz de puntuaciones (*scores*) que se usó para evaluar la probabilidad de que una proteína esté relacionada con otra. Luego, en la década de 1980, los científicos comenzaron a grabar secuencias en diferentes bases de datos, desafiando la forma de búsqueda, comparación y análisis de la información genética y partir de ahí el desarrollo fue ya imparable.

En cuanto a las herramientas bioinformáticas ha habido importantes avances en los últimos años. Es críticamente importante que las cantidades cada vez mayores de datos clínicos, de historia familiar y genotípicos se almacenen en sitios bien diseñados y en bases de datos con un mantenimiento adecuado. Además, la cantidad masiva de datos de genética molecular, secuencias, rutas genómicas, bioquímica comparativa y otros datos ahora disponibles y de dominio público, requiere familiaridad con numerosas y diferentes herramientas bioinformáticas para consultar las bases de datos públicas y poder analizar estos datos.

Cuando los genomas completos comenzaron a estar disponibles, durante la década de 1990, la abundancia de conjuntos masivos de datos biológicos hicieron que los estudios computacionales se convirtiesen en uno de los medios más importantes de la ciencia biológica, introduciendo nuevos conceptos como la “minería de datos”, cuyo objetivo es la extracción de patrones y conocimiento de gran cantidad de datos (J. K. Han, M, 2001).

Además, el desarrollo de las herramientas de matemáticas asistidas por computadora como "Matlab" o "R" ayudaron a matemáticos, biólogos e informáticos para acceder a colecciones públicas de bases de datos para el análisis de la variación, análisis de la expresión génica, comparaciones de genomas completos, investigaciones de patrones sutiles en secuencias

genómicas, propuestas de mecanismos de evolución de genomas, así como la medición de la ve*loci*dad evolutiva en diferentes regiones genómicas, construcción de redes de señalización celular, predicción de genes o regiones genómicas conservadas, etc. (Cristianini, 2006).

1.5.1 Esfuerzos internacionales de investigación: bases de datos genómicas públicas

Debido a la gran cantidad de datos genómicos generados, se vio la necesidad de disponer de nuevas herramientas para compartir datos entre grupos de trabajo, y en los últimos tiempos se han desarrollado, exponencialmente, bases de datos públicas masivas aumentando la cantidad de datos disponibles. El rápido desarrollo de la genómica se ha visto determinado por el proporcional crecimiento de las bases de datos públicas.

Una serie de grandes proyectos, con la participación tanto de corporaciones públicas internacionales como de empresas privadas, no solo se han enfrentado a un esfuerzo tecnológico pionero en ese momento, sino que también han hecho esfuerzos titánicos para que sus hallazgos estén disponibles para la comunidad científica. Los recursos bioinformáticos físicos y humanos se han dedicado a la producción, el análisis, almacenamiento y publicación de resultados científicos para generar repositorios de alta calidad en todo el mundo.

Algunos de estos proyectos son:

1.5.1.1 Proyecto Genoma Humano (HGP)

Lanzado formalmente en 1990, fue un esfuerzo internacional para identificar y describir todos los genes de la especie humana, así como la determinación de la secuencia de los 3 billones de pb contenidos en el genoma humano. Los objetivos iniciales y sus ambiciones se definieron en su concepción, estudiando, por ejemplo, el contenido genético de diversos organismos como las conocidas bacterias intestinales humanas (*Escherichia coli*), la mosca de la fruta (*Drosophila melanogaster*) o el ratón de laboratorio (*Mus musculus*), almacenando toda esa información en bases de datos, recopilando información de bases de datos antiguas, creando y mejorando herramientas de análisis de datos existentes, llevando a cabo la transferencia de tecnología relacionada desde y hacia el sector privado y discutiendo y haciendo recomendaciones sobre las preocupaciones éticas y sociales que surgieron durante el proyecto (https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/).

Bajo la coordinación del DOE, Departament of Energy (Waage et al.) y el National Institute of Health (NIH) de los Estados Unidos, inicialmente se planeó una duración del proyecto de 15 años. Sin embargo, fue en 2000 cuando se publicó una secuencia preliminar del genoma humano, coincidiendo con el mismo anuncio hecho por Celera Genomics, el esfuerzo privado paralelo. Tanto los trabajos públicos como los privados fueron publicados simultáneamente en Nature y Science, respectivamente (Lander et al., 2001; Venter et al., 2001). (Figura 1.10).

Más tarde, en 2003, estaba secuenciado el 92% del genoma humano, y esta secuencia fue nombrada como NCBI36 por el Centro Nacional de Información Biotecnológica y hg18 por la University of California Santa Cruz, UCSC (IHGSC, 2004).

Posteriormente, en el año 2005 y como parte del HGP, se detalló la secuencia del cromosoma X aportando nuevos conocimientos sobre la evolución de los cromosomas sexuales y las diferencias biológicas entre hombres y mujeres.

A partir de aquí se fue evolucionando y fueron surgiendo distintas iniciativas, como *The Cancer Genome Atlas*, en el 2006 (https://www.genome.gov/19518624/2006-release-nih-announces-two-integral-components-of-the-cancer-genome), *The Human Microbiome Project* en 2007 (https://www.genome.gov/26524200/2007-release-nih-launches-human-microbiome-project) y numerosos consorcios implicados en el descubrimiento de variantes relacionadas con cáncer y otras enfermedades así como la secuenciación del genoma de otras especies.

Figura 1.10. Primeras publicaciones del genoma humano en revistas científicas, en ediciones especiales en las revistas *Nature* y *Science* en febrero de 2001. El artículo de *Nature* se centró en los esfuerzos del Proyecto Genoma Humano, mientras que *Science* lo hizo en los resultados de la compañía *Celera Genomics*. Adaptado de https://www.nature.com/ y http://www.sciencemag.org/.

En 2013 ya se comenzaron a explorar genomas de recién nacidos mediante secuenciación y en 2016 un conjunto de centros de investigación financiados por el NIH centró su investigación en comprender las bases genéticas de las enfermedades comunes (enfermedades cardíacas, derrames cerebrales, autismo…) y raras, típicamente heredadas, como la fibrosis quística y la distrofia muscular. Se creó así "*The Atlas of Human Malformation Syndromes in Diverse Populations*".

En 2016 surge el Proyecto Genoma Humano-Escrito (*Human Genome Project - Write*). Se trata de una extensión del Proyecto Genoma Humano para sintetizar el genoma humano. Este nuevo Proyecto HGP-*Write* será manejado por el Centro de Excelencia en Biología Ingeniería, una nueva organización sin ánimo de lucro. Los investigadores esperan que la capacidad de sintetizar grandes tramos del genoma humano podría dar lugar a muchos avances científicos y médicos.

En 2017, con la celebración del vigésimo aniversario del NIH se destacó la transición del centro como Centro Nacional de Investigación del Genoma Humano a un Instituto completo, en el que se alcanzaron gran cantidad de logros, desde la finalización del Proyecto del Genoma Humano, hasta el desarrollo de la tecnología de secuenciación de ADN y trasladar la medicina genómica a la clínica. El NHGRI presentó un plan estratégico en octubre de 2020 para conmemorar el trigésimo aniversario del lanzamiento del Proyecto Genoma Humano, que involucró a distintos expertos y diversas comunidades públicas para identificar nuevas áreas de la genómica que permitan aplicaciones novedosas en la enfermedad humana (https://www.genome.gov/about-nhgri/Brief-History-Timeline). La última referencia del genoma humano (GRCh38/hg38) había sido producida por el "*Genome Reference Consortium*" en diciembre del año 2013 (https://conogasi.org/diccionario/genoma-de-referencia/), hasta que en abril de 2022 el consorcio T2T (Telómero a Telómero), con más de cien investigadores liderados por Adam Phillippy, del Instituto Nacional de Investigación del Genoma Humano (NHGRI) y Karen Miga, de la Universidad de California-Santa Cruz, en EE.UU, revelaron regiones del genoma hasta entonces desconocidas.

1.5.1.2 *Centre D`étude du Polymorphisme Humain (CEPH) y The Human Genome Diversity Project* (HGDP)

Fundado en 1984 por el profesor Jean Dausset, el panel CEPH (http://www.cephb.fr/en/hgdp_panel.php) fue diseñado para facilitar la distribución de referencias de ADN procedentes de 40 familias y coordinar una colaboración internacional para la construcción del primer mapa del genoma humano. Diez años más tarde, en 1993, la Fundación Jean Dausset - CEPH se constituyó como un instituto de investigación sin ánimo de lucro que podría financiar nuevas líneas de investigación.

The Human Genome Diversity Project (HGDP) fue promovido por el Instituto Morrison de la Universidad de Stanford en estrecha colaboración con científicos de todo el mundo coordinados por Cavalli-Sforza. El objetivo del proyecto fue grabar los perfiles genéticos de las poblaciones endógenas, ya que las poblaciones aisladas son las mejores para comprender nuestro pasado lejano. Conociendo la relación entre poblaciones similares podríamos inferir el viaje de la humanidad fuera de África (Cann et al., 2002; J. Z. Li et al., 2008). En 1991, Cavalli-Sforza y varios colegas escribieron una carta a la revista científica *Genomics*, señalando la necesidad de un estudio sistemático de toda la gama de diversidad genética humana en el contexto del Proyecto del Genoma Humano. Cavalli-Sforza argumentaba que el Proyecto Genoma Humano había sido eurocéntrico, ya que las muestras tomadas, de las cuales los científicos ensamblarían la secuencia del genoma humano, provenían de personas de origen europeo.

El proyecto tenía como objetivos científicos principales rastrear la evolución y migración de diferentes poblaciones humanas, con la esperanza de crear un árbol genealógico de poblaciones humanas e identificar los genes que confieren resistencia y vulnerabilidad a las enfermedades, y usarlos para desarrollar tratamientos y pruebas médicas.

Con este objetivo, en 2002, una colaboración entre el CEPH y el HGDP dio lugar a una nueva fuente biológica, el HGDP-CEPH (*Human Genome Diversity Cell Line Panel*) (http://www.cephb.fr/en/hgdp_panel.php), una fuente de 1.063 líneas celulares de linfoblastoides cultivados (LCLs) procedentes de 1.050 individuos de 52 poblaciones mundiales (Figura 1.11), almacenadas en la Fundación Jean Dausset-CEPH en París (Cann et al., 2002).

Los datos del HGDP-CEPH han sido utilizados frecuentemente en investigación genética en campos como la antropología, genética de poblaciones y enfermedades hereditarias. Actualmente el grado de resolución alcanzado ha permitido tener en cuenta muestras duplicadas o estrechamente emparentadas (Rosenberg, 2006).

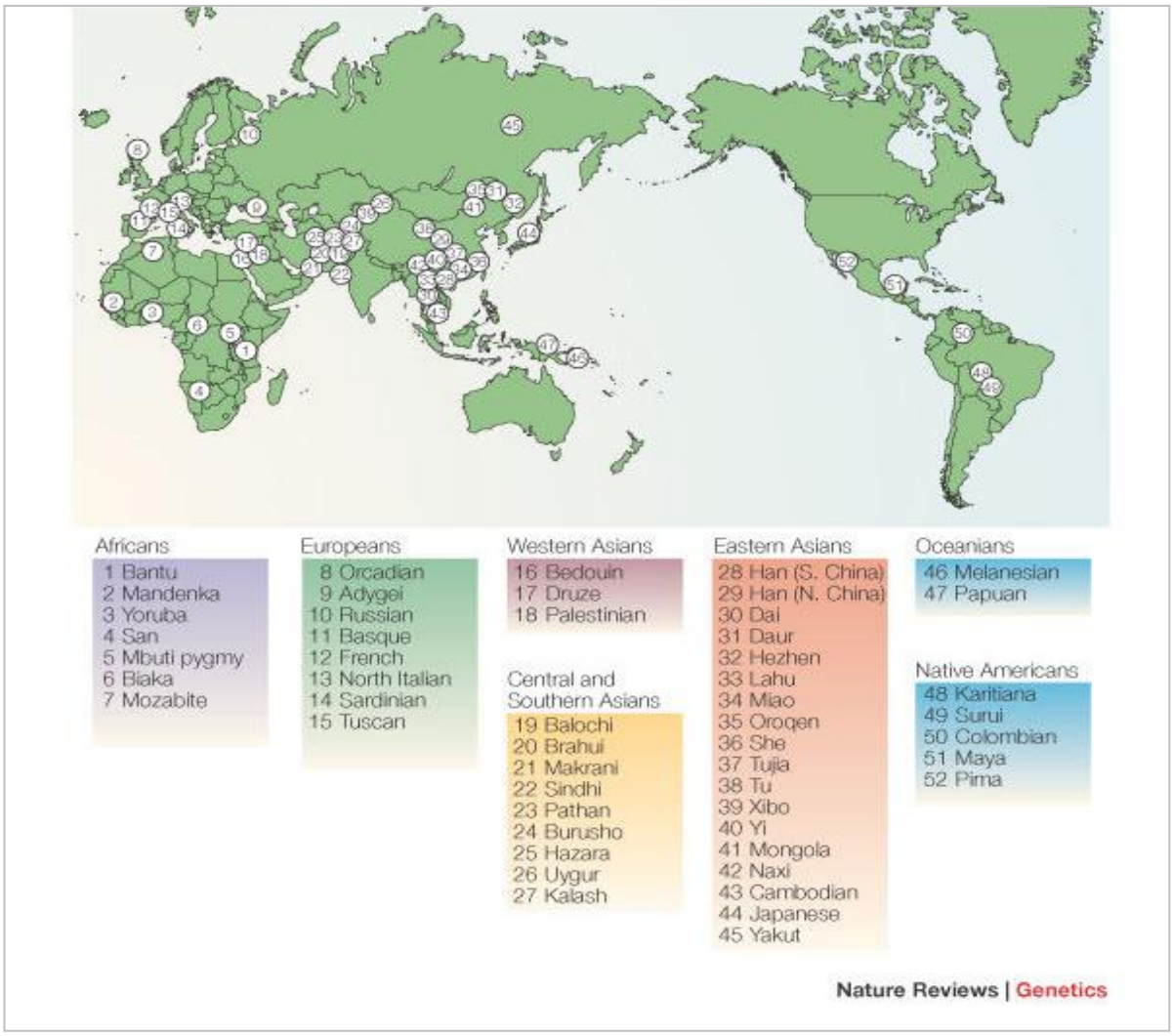


Figura 1.11. Lista de poblaciones incluidas en el panel CEPH y sus localizaciones geográficas. (Cavalli-Sforza, 2005).

Continuando con el panel de líneas celulares, muchos grupos en todo el mundo decidieron utilizar la tecnología de *microarrays* de ADN. Consecuentemente, en 2007 fue utilizado el chip de ADN más denso disponible en aquel momento (*BeadStation of Illumina*) para procesar el panel CEPH en las Universidades de Stanford y Michigan. De aquí se publicaron gran cantidad de repositorios de datos de genotipado de estas muestras accesibles a los investigadores (Jakobsson et al., 2008; J. Z. Li et al., 2008). Esto originó los datos de "*The Human Genome Diversity Project*" (HGDP) (https://www.hagsc.org/hgdp/), que comprenden unos 600.000 marcadores tipados en 1.000 muestras de más de 50 poblaciones.

1.5.1.3 Proyecto HapMap

Este proyecto internacional fue formalizado en el año 2002 con el objetivo de desarrollar un mapa haplotípico del genoma humano y así poder identificar variaciones genéticas y genes que afecten a la salud. Este proyecto consistió en la creación de un catálogo de variantes genéticas humanas comunes (; Cardon & Abecasis, 2003; International HapMap, 2003), (https://www.ncbi.nlm.nih.gov/nlmcatalog/101200656), con las posiciones que ocupan y sus frecuencias dentro de las poblaciones. Inicialmente fue concebido con la intención de compartir

con otros investigadores información relevante e importante que relaciona las distintas variantes con susceptibilidad a enfermedades, con la idea final de la prevención, diagnosis y tratamiento de la enfermedad.

Tan pronto como el genoma completo estuvo disponible, surgió la idea del proyecto HapMap. Tras la secuenciación del genoma humano el siguiente paso lógico fue el estudio de su variabilidad entre poblaciones e individuos. En concreto, la idea fue tomar ventaja en las posibles combinaciones de alelos en la misma molécula de ADN, llamados haplotipos, con la finalidad de reducir los posibles marcadores a analizar. Esto se basó en la asunción de que, mientras los alelos de diferentes cromosomas segregan al azar durante la meiosis, alelos cercanos en un cromosoma no se someten a recombinación tan frecuentemente. Por lo tanto, y a nivel de población, la recombinación podría ser estudiada para investigar si existe una asociación no debida al azar entre alelos de diferentes *loci*: como ya hemos visto, el desequilibrio de ligamiento (DL).

Durante las etapas tempranas del proyecto HapMap, las mediciones del DL fueron recopiladas a partir de datos de genotipos, y se describieron patrones comunes de variación genética en humanos, incluyendo las regiones cromosómicas con grupos de SNPs fuertemente asociados, los haplotipos en esas regiones y los SNPs objetivo dentro de ellos. También se anotaron las regiones cromosómicas donde las asociaciones entre SNPs eran débiles. Consecuentemente, si las variantes bajo desequilibrio de ligamiento eran conocidas y se puede considerar su transmisión en forma de bloques haplotípicos, el número de SNPs necesarios para estudiar se reduce a SNPs objetivo que se definen como *Tag* SNPs.

Durante la Fase I del análisis 10 centros de genotipado diferentes produjeron más de 1 millón de SNPs en 200 muestras de 4 poblaciones: residentes de Utah con ascendentes del norte y oeste de Europa, Chinos Han de Pekín, población japonesa de Tokio y Yorubas de Ibadán. En una segunda fase (Fase II) se produjeron más del triple de las variantes conocidas hasta ese momento, hasta 3 millones de SNPs en las mismas muestras, pero priorizando SNPs no sinónimos en regiones codificantes. En una larga Fase III se resecuenciaron 7 nuevas poblaciones: Afro-Americanos, Chinos residentes en Estados Unidos, Indios Gujarati, Keniatas Webuye y Maasai, Mexicanos en Estados Unidos y Toscanos (Italianos), priorizando variantes raras y genotipando 1,4 millones de SNPs (Tabla 1.2). (https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/).

Tabla 1.2. Detalles del proyecto HapMap. SNPs totales genotipados para las tres fases, SNPs objetivo y lista de poblaciones incluidas en el análisis (Buchanan, Torstenson, Bush, & Ritchie, 2012).

Fase	Número de SNPs genotipados	SNPs objetivo	Población de estudio
I	1 millón	Prioridad: SNPs codificantes para alcanzar 1 SNP cada 5kb	CEU, YRI, CHB, JPT
II	3 millones	Prioridad: SNPs no sinónimos en regiones codificantes	CEU, YRI, CHB, JPT
III	1,4 millones	Prioridad: variantes raras	CEU, YRI, CHB, JPT, ASW, CHB, GIH, LWK, MXL, MKK, TSI

1.5.1.4 *Human Variome Project* (HVP)

Fue una iniciativa del año 2006 y surgió con el objetivo de obtener y recolectar todas las variaciones del genoma humano que afectan la salud humana en una sola base de datos. Su objetivo final era crear un catálogo de variantes con frecuencias de 1% o menos en la población humana (http://www.humanvariomeproject.org/).

Ya existen muchas bases de datos sobre la variación humana, pero ninguna de ellas está interrelacionada, lo que hace que sea casi imposible administrar de manera efectiva toda esa información (R. G. H. K. J. Cotton, Haig H. , 2005; Horaitis & Cotton, 2004). Cada vez se encuentran más variantes, y se descubre más información clínica para cada una de ellas, y esta información es raramente incorporada a bases de datos antiguas. Por esa razón, el HVP ha sido descrito como un elemento esencial complementario para el HGP, constituyendo una fuente actualizada y supervisada de información que conecta las diferentes variantes genéticas con las enfermedades correspondientes. Intenta recopilar todas las variantes que afectan a la salud humana, aprovechando la información ya publicada en otras bases de datos, como resultado de proyectos anteriores: HGP, HapMap o 1000G.

El proyecto está dirigido a la sistematización e identificación de genes, sus mutaciones y variantes que los asocian a la variabilidad fenotípica de la enfermedad humana. Se basa en un sistema de nomenclatura estandarizada y un *software* que intercambia información entre bases de datos con información genética, poblacional o clínica (R. G. Cotton, Auerbach, et al., 2007).

Para que este proyecto tenga éxito, sigue siendo importante la participación de laboratorios clínicos, médicos e investigadores con el fin de mejorar la información ya disponible en línea en otras bases de datos públicas, tales como OMIM (*Online Mendelian Inheritance in Man*®) (https://www.omim.org), GenBank (https://www.ncbi.nlm.nih.gov/genbank/), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), dbGAP (https://www.ncbi.nlm.nih.gov/gap), HapMap (https://www.genome.gov/es/genetics-glossary/HapMap-mapa-de-haplotipos), NCBI (http://www.ncbi.nlm.nih.gov/), HUGO (del inglés *Human Genome Organization*), HGVS (del inglés *Human Genome Variation Society*; http://varnomen.hgvs.org/), LSDB (del inglés *Locus Specific Mutation Databases*; https://grenada.lumc.nl/LSDB_list/lsdbs/DMD), EBI (del inglés *European Bioinformatics Institute*; https://www.ebi.ac.uk/eva/), PharmGKB (https://www.pharmgkb.org/), GeneTests (https://www.ncbi.nlm.nih.gov/gtr/), Ensembl (https://www.ensembl.org/index.html) y UCSC (*University of California Santa Cruz*; https://genome.ucsc.edu/). El proyecto también evalúa los tratamientos e incluye variantes comunes y raras.

El HVP fomenta la estandarización de metodologías, tratando con preocupaciones éticas, fomentando el automatismo y promoviendo la participación de los países en desarrollo (R. G. Cotton et al., 2008; R. G. Cotton, Human Variome, et al., 2007).

1.5.1.5 Proyecto 1000 Genomas (1000G)

Iniciado en 2008, fue diseñado para ser un nuevo referente de la variabilidad humana, secuenciando hasta 1.000 genomas diferentes mediante el uso de las tecnologías de secuenciación de alto rendimiento desarrolladas en aquel momento y aprovechando la drástica reducción de los costes. Estaba dirigido a conocer y organizar todas las variantes humanas que ayudan a discernir entre dos individuos (Siva, 2008), (http://www.1000genomes.org).

Su objetivo principal es encontrar variantes genéticas con una frecuencia inferior al 1%. Estas variantes raras son un gran complemento para la información ya disponible, tanto desde el punto de vista clínico como poblacional. Desde la perspectiva clínica, la información biomédica sobre las diferentes variantes encontradas en el genoma ahora es accesible para los investigadores en una única base de datos. Gracias a este proyecto, se describieron más de cien regiones genómicas que contienen variantes relacionadas con enfermedades comunes tales como diabetes, cardiopatías coronarias, cáncer de próstata y de mama, artritis reumatoide, síndrome del colon irritable, degeneración macular, etcétera, tal como hemos visto en el apartado 1.1.6 de este trabajo.

Por otro lado, tener acceso a una base de datos de millones de variantes raras en muchas poblaciones es potencialmente útil en estudios de genética de poblaciones, ayudando a caracterizar la contribución de variantes raras a rasgos complejos y enfermedades, pero que también pueden ayudar a probar o refutar las diferencias en la estructura de la población. Al menos, esta información adicional sería un suplemento para la variabilidad común conocida.

Una comparación de los conjuntos de datos piloto del Proyecto HapMap y 1000G mostró que aproximadamente el 72% de los SNPs de HapMap también se encontraron en los datos piloto del Proyecto 1000 Genomas. Después de filtrar las variantes de HapMap con una MAF <5% (por separado para cada población), el 99% de los SNPs de HapMap se encontraron en los datos de 1000G (Buchanan et al., 2012). Un año después, el catálogo contenía 600 individuos y hoy en día han sido completamente secuenciados 2.500 individuos de 26 poblaciones diferentes (Tabla 1.3). La cantidad aumentada en millones de variantes está ayudando a localizar con mayor precisión las regiones asociadas con enfermedades.

Tabla 1.3. Códigos de poblaciones y continentes de cada una de las 26 poblaciones incluidas en el proyecto 1000G. Estas poblaciones han sido divididas en 5 súper poblaciones (continentes): AFR, *African*; AMR, *Ad Mixed American*; EAS, *East Asian*; EUR, *European*; SAS, *South Asian*.

Código de la población	Población	Continentes
JCHB	*Han Chinese in Beijing, China*	EAS
JPT	*Japanese in Tokyo, Japan*	EAS
CHS	*Southern Han Chinese*	EAS
CDX	*Chinese Dai in Xishuangbanna, China*	EAS
KHV	*Kinh in Ho Chi Minh City, Vietnam*	EAS
CEU	*Utah Residents (CEPH) with Northern and Western European Ancestry*	EUR
TSI	*Toscani in Italia*	EUR
FIN	*Finnish in Finland*	EUR
GBR	*British in England and Scotland*	EUR
IBS	*Iberian Population in Spain*	EUR
YRI	*Yoruba in Ibadan, Nigeria*	AFR
LWK	*Luhya in Webuye, Kenya*	AFR
GWD	*Gambian in Western Divisions in the Gambia*	AFR
MSL	*Mende in Sierra Leon*	AFR
ESN	*Esan in Nigeria*	AFR
ASW	*Americans of African Ancestry in SW USA*	AFR
ACB	*African Caribbeans in Barbados*	AFR
MXL	*Mexican Ancestry from Los Angeles USA*	AMR
PUR	*Puerto Ricans from Puerto Rico*	AMR
CLM	*Colombians from Medellin, Colombia*	AMR
PEL	*Peruvians from Lima, Peru*	AMR
GIH	*Gujarati Indian from Houston, Texas*	SAS
PJL	*Punjabi from Lahore, Pakistan*	SAS
SEB	*Bengali from Bangladesh*	SAS
STU	*Sri Lankan Tamil from the UK*	SAS
ITU	*Indian Telugu from the UK*	SAS

1.5.1.6 Nuevos proyectos de Biobancos a gran escala

En los últimos tiempos, muchos gobiernos en todo el mundo han comenzado a impulsar el desarrollo de diferentes Biobancos. Se trata de un biorepositorio que está destinado al almacenamiento, a gran escala, de muestras biológicas humanas, para dar acceso a investigación y a datos de decenas o cientos de miles de personas. Por lo tanto, los investigadores podríamos recurrir a estos biobancos sin la necesidad absoluta de reclutar nuevas y suficientes muestras, de manera que los estudios de asociación de todo el genoma pueden ser más factibles. Desgraciadamente no los hay en muchas poblaciones ni países ni para todos los grupos de enfermedad.

En 2005 se diseñó el Proyecto Biobanco del Reino Unido, un estudio a largo plazo en Inglaterra, Gales y Escocia, con el objetivo de investigar la interacción entre el medio ambiente y los genes y su contribución al desarrollo de la enfermedad (http://www.ukbiobank.ac.uk/).

En los primeros años del proyecto, más de medio millón de voluntarios del Reino Unido, de entre 40 y 69 años, se inscribieron en el estudio. Estos participantes tendrán seguimiento durante al menos 25 años, incluyendo información sobre hábitos alimenticios, estilo de vida, historial médico, etc. Además, también se van teniendo en cuenta muchos factores ambientales básicos y variables físicas (como la altura sentado y de pie, el índice de masa corporal, el peso, la presión arterial, exposición a radiación y radón, memoria etc...). También fueron tomadas muestras de sangre, saliva y orina. El plan es rastrear a todos estos voluntarios durante los próximos años, registrando todas las enfermedades, recetas de medicamentos, fallecimientos...lo que es posible a través del Servicio Nacional de Salud centralizado del Reino Unido.

En marzo de 2012, la base de datos estaba en línea para que los investigadores solicitaran su uso, utilizando la cantidad masiva de muestras para comparar individuos de casos / controles para una enfermedad en particular, para intentar medir los beneficios o perjuicios en cualquier interacción entre genética, medio ambiente y medicamentos. Este procedimiento siempre estuvo bajo la vigilancia del *Ethics and Governance Framework* (EGF): (Biobanco del Reino Unido, 2007: http://www.ukbiobank.ac.uk/resources/). Este está destinado a proporcionar una base de confianza mediante la especificación pública de los estándares a través de los que se administra el Biobanco del Reino Unido. El *Framework* se publicó por primera vez en 2003 y la tercera versión está disponible en el sitio web del Biobanco del Reino Unido. La responsabilidad del *Framework* recae en el Biobanco del Reino Unido, incluidas las revisiones que serán necesarias para reflejar los cambios en el ámbito tecnológico, legal o ético.

Se ha realizado un estudio genómico completo con el array UK Biobank Array que incluye más de 800.000 SNPs, con la contribución del Wellcome Trust.

Como se dijo anteriormente, hay más proyectos como este en curso en todo el mundo. Entre otros, el EPIC (*European Prospective Investigation in Cancer and Nutrition*) comenzó en 1992 implicando a medio millón de participantes de diez países europeos (Alemania, Dinamarca, España, Francia, Grecia, Países Bajos, Italia, Noruega, Reino Unido y Suecia), estudiando el papel de los genes y dieta en el desarrollo del cáncer (Bingham & Riboli, 2004). Otro proyecto similar a los británicos se planeó en los Estados Unidos en 2006 y en China, el *Kadoorie Biobank,* que recogió y probó más de medio millón de muestras humanas con el objetivo de investigar enfermedades crónicas, incluyendo una nueva prueba cada año (http://www.ckbiobank.org/site/). En Islandia, una empresa privada china ya secuenció a la mitad de la población del país, y está planeando terminar la secuencia de toda la población (270.000 individuos). Este proyecto incluye datos de salud y genealógicos. Un último ejemplo sería el Proyecto del Genoma de Estonia, iniciado en el año 2000 y destinado a recopilar datos

de genealogía, genoma y salud del 5% de la población de Estonia (Frank, 2000; Leitsalu et al., 2015), (http://www.genomenewsnetwork.org/).

España, a través de la infraestructura IMPaCT está compilando una cohorte poblacional de 200.000 con información epidemiológica, clínica y genómica, que junto con los datos que se obtienen del programa IMPaCT Genómica es la contribución española al esfuerzo 1M Genomes (1 millón de genomas), una gran iniciativa europea que se complementa con la acción B1M Genomes (más de 1 millón de genomas), uno de cuyos subprogramas es el Genoma de Europa. España tiene un papel muy destacado en esta iniciativa.

1.5.2 Bioinformática computacional y estadística: herramientas para el análisis de datos multivariantes

Con el aumento repentino en la cantidad de datos biológicos que se ha producido en las últimas décadas debido al progreso enorme en genómica (Figura 1.12), la bioinformática y las herramientas estadísticas se están convirtiendo en algo esencial de forma rutinaria para los biólogos. Esto significa que se están aplicando algoritmos y tecnología informática para la gestión, el filtrado, el análisis y el almacenamiento de grandes conjuntos de datos de información biológica particular, con el objetivo final de comprender los hechos biológicos a partir de la enorme cantidad de datos, que sin dichas herramientas sería imposible.

El gran aumento en la producción de datos generados por los métodos desarrollados recientemente en cuanto a secuenciación de alto rendimiento se logró como resultado de la contribución combinada de secuenciación automatizada y tecnología de *microarrays*, supercomputación e internet. Todo comenzó en la década de 1980, cuando los científicos luchaban con el almacenamiento y la indexación de la información biológica que estaba empezando a producirse a grandes escalas (Attwood, 1999).

Las tareas de secuenciación de nueva generación, genotipado, almacenamiento masivo de datos, extracción de datos, interpretación y análisis ahora son llevados a cabo completamente gracias a la informática, que hace posible que esto se haga de forma eficiente e inteligente. Deben ser almacenadas enormes cantidades de datos, pero estos deben ser de fácil acceso y también fiable. Así, cualquier centro de altas capacidades de genotipado y secuenciación necesita disponer de capacidad de almacenamiento de datos y computación suficiente.

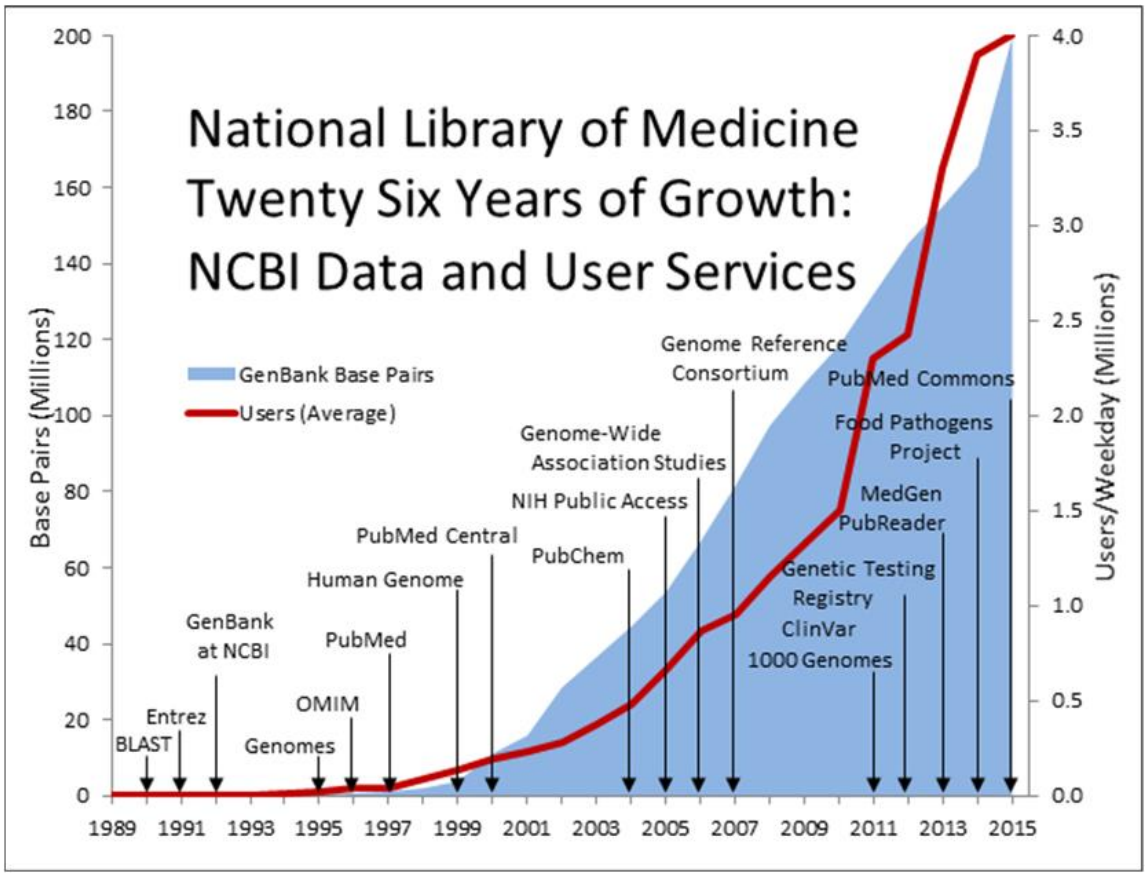


Figura 1.12. Evolución en el tiempo del contenido en NCBI en pb. El incremento sobre la información biológica disponible se ofrece representada en números de pares de bases (en azul) de las secuencias disponibles. Fuente "*Department of Health and Human Services National Institutes of Health National Library of Medicine*": https://www.nlm.nih.gov/about/2017CJ.html.

Después del almacenamiento, la extracción de la información útil es la siguiente tarea computacional. Los datos por sí solos no tienen sentido antes de ser filtrados y analizados, tareas impracticables debido al gran tamaño de las bases de datos.

Aparte de la aplicación en el diagnóstico, otra de las áreas en las que las herramientas bioinformáticas han sido utilizadas es en los estudios comparativos, analizando y comparando el material genético de diferentes especies para comprender las funciones de los genes, los mecanismos de la herencia o la evolución de las especies.

Además de la genómica humana, la bioinformática también puede ayudar al desarrollo de la genómica en ganadería, la agricultura y microorganismos. Los microorganismos están presentes de manera ubicua en el medio ambiente y nuestros cuerpos, y el conocimiento de los genomas completos de esos organismos aumentaría sus aplicaciones en ambiente, salud, energía e industria. La agricultura y el cultivo también se verían beneficiados por herramientas bioinformáticas que encuentran genes específicos que luego podrían usarse para producir variedades más fuertes, resistentes y productivas.

Finalmente, las posibilidades en investigación genética se han incrementado drásticamente en los últimos tiempos, debido, tanto a la disponibilidad masiva de datos genómicos como a los desarrollos en la ciencia de la computación. De considerable importancia en este sentido, y como ya se ha mencionado, son los esfuerzos realizados por el *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov/) de EE. UU., El Instituto *Wellcome*

Trust Sanger del Reino Unido (https://www.sanger.ac.uk/), el Instituto Europeo de Bioinformática (http://www.ebi.ac.uk/), el Laboratorio Europeo de Biología Molecular (http://www.embl.org/) y el Instituto Nacional de Bioinformática (España; https://inb-elixir.es/). El subprograma de IMPaCT Datos dentro de la infraestructura IMPaCT representa un esfuerzo sin precedentes a nivel nacional para la compilación de datos genómicos y clínicos y su obtención fácil gracias al desarrollo de *beacons* (https://www.isciii.es/QueHacemos/Financiacion/Documents/IMPaCT%20Web/PLAN_ESTRATEGICO_IMPaCT.pdf).

Existen herramientas bioinformáticas de todo tipo que están disponibles de forma gratuita en las páginas web de estos Institutos para consulta y procesamiento de información relacionada con la biología molecular, muchos de ellos fuertemente relacionados con la inferencia de la estructura de la población.

La mayoría de estos enfoques bioinformáticos son ahora una parte esencial del diseño del análisis, ejecución y publicación de resultados: R (https://cran.r-project.org/), BioPerl (http://bioperl.org/), BioPython (http://biopython.org/), SPSS (IBM, 2011), etc...

El uso de algunas de estas herramientas se ha ampliado internacionalmente: GCTA (J. Yang, Lee, Goddard, & Visscher, 2011), Plink (Purcell et al., 2007), STRUCTURE (Pritchard, Stephens, & Donnelly, 2000), fastPhase (Scheet & Stephens, 2006) etc..., mientras que otras han comenzado a destacar recientemente por sus enfoques innovadores para la caracterización sutil de la subestructura de la población, como por ejemplo ChromoPainter, fineSTRUCTURE (Lawson, Hellenthal, Myers, & Falush, 2012) y Globetrotter (Hellenthal et al., 2014) entre otras, o llevando a cabo análisis de la estructura espacial de datos genéticos, como SPA (W. Y. Yang, Novembre, Eskin, & Halperin, 2012) y Loco-LD (Baran, Quintela, Carracedo, Pasaniuc, & Halperin, 2013). Algunas de estas herramientas bioinformáticas, popularmente consideradas útiles para investigar la estructura de la población utilizando datos genéticos, han sido de especial interés para el progreso de esta tesis y se explicarán a continuación, pero antes se ofrecerá una introducción general a los enfoques de análisis de estructura de población.

Tradicionalmente, dos de los tipos más populares de enfoques para esta tarea son los enfoques no paramétricos y el enfoque del modelo bayesiano (paramétrico).

Procedimientos estadísticos paramétricos tales como STRUCTURE (Pritchard et al., 2000) se basan en suposiciones acerca de la forma de la distribución (por ejemplo, asumir una distribución normal de los datos) en la población subyacente y también sobre la forma o los parámetros (como las desviaciones estándar o medias) de la distribución supuesta. Esos parámetros son necesarios para explicar la estructura de la población. Los métodos basados en modelos bayesianos intentan reconstruir eventos históricos de una manera más directa que los procedimientos no paramétricos. Una vez que se asume un número discreto o K poblaciones (número de poblaciones definido por nosotros en el estudio), luego las frecuencias de alelos para cada grupo y para cada miembro individual de la población se estiman simultáneamente a partir de los datos a través del modelo bayesiano.

Enfoques contemporáneos como el DAPC (del inglés *Discriminant Analysis of Principal Components*), que se explica más adelante (T. Jombart, Devillard, & Balloux, 2010), tienen una efectividad mayor al poder ampliar este método, permitiendo asumir que los individuos presentan una mezcla, es decir, descienden de más de una de las poblaciones K inicialmente descritas. Estos enfoques son muy poderosos, pero dependen mucho de la capacidad computacional. Además, la determinación del número de poblaciones (K) es computacionalmente costosa incluso para el *software* más rápido: ADMIXTURE (Alexander, Novembre, & Lange, 2009). Finalmente, las relaciones entre los clústeres inferidos pueden no

explicarse de ninguna manera, y solo modificar el valor de K y verificar las variaciones de los resultados pueden arrojar algo de luz sobre eso (T. Jombart et al., 2010).

Por otro lado, los procedimientos estadísticos no paramétricos como el PCA **(**del inglés *Principal Components Analysis*) se basan en el análisis de una matriz con entradas que cuantifican la similitud genética entre todos los individuos por parejas (Menozzi, Piazza, & Cavalli-Sforza, 1978). Estos enfoques se basan en ninguna o pocas suposiciones sobre la forma de los parámetros o la distribución de la población.

Normalmente, tanto los enfoques paramétricos como los no paramétricos se pueden aplicar al mismo conjunto de datos para proporcionar un resumen útil de toda la información contenida.

1.5.2.1 PCA (*Principal Component Analysis*)

La teoría básica para el análisis de componentes principales (PCs) se inventó en 1901 (K. Pearson, 1901) y se ha desarrollado independientemente y nombrado en la década de 1930 (H. Hotelling, 1933; H. Hotelling, 1936). PCA usa una transformación ortogonal para convertir un conjunto de observaciones de variables que pueden correlacionarse en un conjunto de valores que se llaman "componentes principales", que siempre están no correlacionados linealmente.

La cantidad de componentes principales siempre es menor o igual a la cantidad variables originales. Por lo tanto, la ventaja de usar componentes principales en lugar de las variables originales es que a través de esta transformación el primer componente principal obtenido siempre será para la mayor variabilidad posible en los datos, y cada componente sucesivo a su vez tiene la mayor varianza a la izquierda, con la única condición de que cada nuevo componente sea ortogonal al precedente, sin correlación. Por lo tanto, los componentes principales son los vectores propios de la matriz de covarianza, que es simétrica. A menudo, se puede pensar que la operación de PCA revela la estructura interna de los datos, proporcionando al usuario una imagen de los datos con menor punto de vista informativo. La dimensionalidad reducida se logra utilizando solo los primeros componentes (Jolliffe, 2002).

Por lo tanto, los componentes principales aplicados en datos genéticos representan direcciones que maximizan explicar el patrón observado de similitud genética. Al trazar los PCs sucesivos, se puede lograr la representación visual de distintos patrones de estructura en los datos. En estas parcelas, grupos de individuos pueden entenderse como poblaciones genéticas, mientras que la mezcla de dos poblaciones se representa como grupos de individuos que se encuentran a lo largo de una línea (D. Reich, Thangaraj, Patterson, Price, & Singh, 2009). Es importante destacar que otros eventos históricos también pueden generar señales de PCs coincidentes (McVean, 2009) pero también pueden darse otras interpretaciones de los PCs (Novembre et al., 2008).

En el caso de plataformas grandes, como ThermoFisher e Illumina, ya se han publicado subpaneles específicos de diferentes matrices para evaluar y corregir el efecto de la mezcla en ancestralidad.

El PCA es también la técnica a la que muchos apuntaban para caracterizar las diferencias genéticas entre las poblaciones de distintos puntos geográficos, y además minimiza las asociaciones falsas al tiempo que maximiza el poder para detectar asociaciones verdaderas.

PCA, en el que se basa el método EIGENSTRAT para la corrección de la estratificación de la población en estudios de asociación genética (Price et al., 2006), se ha implementado en el paquete EIGENSOFT (https://www.hsph.harvard.edu/alkes-price/software/) pero también se ha incluido en otro *software* como GCTA (J. Yang et al., 2011), que incluye muchos otros análisis para comprender mejor la arquitectura genética de los rasgos complejos.

1.5.2.2 DAPC (*Discriminant Analysis of Principal Component*)

Se implementa el análisis discriminante de los componentes principales (DAPC) (T. Jombart et al., 2010) en el paquete "adegenet" para R, y creado como respuesta a los desafíos en el análisis de gran cantidad de datos.

Este enfoque se debió esencialmente al hecho de que los algoritmos Bayesianos basados en modelos predefinidos de población como el *software* STRUCTURE están demasiado limitados por los costos computacionales y no son adecuados para analizar la gran cantidad de datos que se producen hoy en día, ya que sería ineficiente. En consecuencia, era necesario utilizar métodos menos intensivos en cuanto a informática se refiere. Por otro lado, los análisis de multivariantes están específicamente dedicados a extraer información de grandes conjuntos de datos sin ser demasiado exigentes computacionalmente. Desgraciadamente, en ese momento, no había métodos multivariantes específicamente diseñados para estudiar la estructura genética de las poblaciones naturales. Por ejemplo, ya los métodos multivariantes existentes como PCA carecían de la capacidad de proporcionar una evaluación grupal, requiriendo por tanto que los clústeres se definieran a priori. Además, intentan resumir la variabilidad general entre los individuos, incluida la divergencia entre los grupos y dentro de ellos. Por lo tanto, un método adecuado debe descuidar la variabilidad dentro del grupo mientras se enfoca en la variación entre grupos.

DAPC es una técnica multivariante diseñada para descubrir y representar grupos de individuos que están genéticamente relacionados. Además, cuando los clústeres no se pueden definir a priori, DAPC puede usar *sequential K-means* y modelos Bayesianos para inferir clústeres genéticos. Cuando esto sucede, DAPC amplía efectivamente el modelo STRUCTURE para permitir el análisis de individuos que presenten mezcla, es decir, que desciendan de más de una población. A los individuos se les asignan vectores de ascendencia que representan la proporción de sus ancestros que provienen de cada una de las poblaciones K (T. Jombart et al., 2010).

DAPC también asigna individuos a grupos y permite la evaluación visual de la diferenciación entre poblaciones, así como la contribución alélica a la estructura de la población. Concretamente, DAPC funciona mejor que STRUCTURE en la caracterización de la estructura de la población (T. Jombart et al., 2010), y desentraña de manera más rápida y eficiente la subestructura de la población compleja.

1.5.2.3 fineSTRUCTURE

Como se ha estado afirmando, en los últimos tiempos, el genotipado de alta densidad y la secuenciación de alto rendimiento han hecho posible la producción de una gran cantidad de datos útiles para explorar la subestructura de la población. La mayor cantidad de datos podría otorgar un nuevo nivel de resolución en análisis de poblaciones, dibujando nuevos patrones de propagación y apareamiento representando los antiguos y nuevos eventos históricos. Sin embargo, también se ha dicho antes que este aumento en la disponibilidad de datos también implica nuevos desafíos derivados de las limitaciones computacionales, incluso si consideramos recientes avances computacionales y experimentales (Browning & Browning, 2007; H. C. Fan, Wang, Potanina, & Quake, 2011; Kitzman et al., 2011; Niu, 2004; Scheet & Stephens, 2006).

De hecho, tanto enfoques tradicionales no paramétricos (similar a PCA) como paramétricos (basados en modelos Bayesianos o similares a STRUCTURE) analizan los SNPs independientemente entre sí. Por lo tanto, todos estos enfoques ignoran la información adicional contenida en las posiciones relativas de estas mutaciones individuales a lo largo del genoma. Por el contrario, el uso de esta información permitiría nuevas oportunidades para la inferencia de ascendencia aprovechando los patrones de variación correlacionada. Los marcadores

estrechamente posicionados, no afectados por la recombinación, se heredan juntos, lo que resulta en un desequilibrio de ligamiento a nivel de la población.

Por lo tanto, el análisis basado en haplotipos, que explota este tipo de información, tiene el potencial para ser utilizado para el análisis de la estructura de la población. Sin embargo siguen siendo escasas las metodologías propuestas para aprovechar la información del DL contenida en conjuntos de marcadores estrechamente posicionados (Browning & Weir, 2010; Conrad et al., 2006; Gattepaille & Jakobsson, 2012; Hellenthal, Auton, & Falush, 2008; Jakobsson et al., 2008; Lawson et al., 2012).

El enfoque fineSTRUCTURE permite aplicar metodologías paramétricas y no paramétricas, análogamente a PCA y STRUCTURE, aprovechando la información contenida en la estructura de haplotipos. Inicialmente, se obtiene una "matriz de coancestría" a partir de los datos por medio del *software* ChromoPainter (Lawson et al., 2012). Esta simple y única "matriz de coancestría" encierra cada bit de la información que luego será explorada por PCA y enfoques tipo STRUCTURE, reduciendo así la dimensionalidad de los datos mediante la estimación de las relaciones entre todos los pares de individuos a través de esta "representación cromosómica" (N. Li & Stephens, 2003). Esta matriz calcula aproximadamente la cantidad de "fragmentos" discretos del genoma de cualquier individuo que están más estrechamente relacionados con la fracción análoga del genoma de cualquier otro. Después de esto, la matriz podría usarse para aprender sobre la subestructura de la población, identificando grupos de individuos con ancestros históricos similares, correspondientes a poblaciones genéticamente relacionadas, revelando también características detalladas de interacciones históricas. Por consiguiente, la composición de la población y las relaciones entre las poblaciones están determinadas por las diferentes características y probabilidades de compartir fragmentos de ADN comparables entre individuos dentro o fuera de la misma población (Lawson et al., 2012).

Cuando están disponibles conjuntos densos de marcadores, normalmente producidos por *arrays* de genotipado de alto rendimiento, este nuevo enfoque es sustancialmente más sensible que todos los métodos publicados anteriormente, que solo tratan a los marcadores de forma independiente (Eigenstrat, STRUCTURE, ADMIXTURE, etc.) infiriendo simultáneamente hasta 100 poblaciones con separación mejorada, revelando nuevas diferencias en la genética ancestral entre individuos de la misma población (Lawson et al., 2012).

1.5.2.4 Otras herramientas Bioinformáticas

Existen otras herramientas bioinformáticas, softwares extremadamente útiles para los estudios científicos en general y particularmente para inferir estructura poblacional. A continuación, se mencionan algunas de ellas:

a) Arlequin:

Un *software* integrado para el análisis de datos de genética de poblaciones. Su objetivo es proporcionar al usuario en genética de poblaciones un conjunto bastante grande de métodos básicos y pruebas estadísticas, con el fin de extraer información sobre las características genéticas y demográficas de una colección de muestras de una población. Las pruebas estadísticas implementadas en Arlequin han sido elegidas para minimizar el ocultamiento y las suposiciones y para ser lo más poderoso posible, lo que permite el análisis de SNPs y STRs del cromosoma Y que no se pueden hacer con Plink (Excoffier & Lischer, 2010), (http://cmpg.unibe.ch/software/arlequin35/).

b) Network:

Software libre para estudios de red filogenética que genera árboles evolutivos y redes de datos genéticos, lingüísticos y otros datos. La red puede proporcionar estimaciones de edad para cualquier antepasado en el árbol (http://www.fluxus-engineering.com/sharenet.htm).

c) BioPerl:

Conjunto de módulos para ayudar a escribir *scripts* Perl bioinformáticos, con algunos guiones funcionales también. Perl es un lenguaje de programación rico en características con una interfaz muy fácil de ejecutar. Es muy potente, pero también muy flexible, lo que dificulta compartir fácilmente los *scripts*. (http://bioperl.org/).

d) Biopython:

De forma análoga a BioPerl, es un conjunto de herramientas disponibles para computación biológica escrita en lenguaje de programación Python. Python es el lenguaje de programación mejor orientado para las tareas más complejas de manipulación de datos, mantenimiento y uso compartido con otros durante el trabajo en equipo. (http://biopython.org/).

e) R:

El lenguaje R y una gran cantidad de paquetes de genética estadística que se ejecutan en R están disponibles en línea (https://cran.r-project.org/). Este lenguaje y *software* son los más ricos en gráficos, técnicas, modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, agrupamiento, etc. R deriva del lenguaje S, convirtiéndose en un instrumento fácil de usar, orientado a objetos, interpretación de idioma, bueno para la presentación de datos finales, pero menos recomendable que Perl o Python para grandes cantidades de datos y edición. Además, muchas bibliotecas y paquetes están continuamente accesibles para usuarios, y contienen un código específico para muchas tareas científicas, incluido el análisis del genoma, como "adegenet" (mencionado anteriormente) o GenABEL (Aulchenko, Ripke, Isaacs, & van Duijn, 2007).

f) fGCTA:

Se trata de una herramienta para el análisis de rasgos complejos del genoma (GCTA, del inglés *Genome-wide Complex Trait Analysis*). Fue originalmente diseñada para estimar la proporción de varianza fenotípica explicada por SNPs genómicos para rasgos complejos (el método GREML), y posteriormente se ha ampliado para muchos otros análisis tales como PCA y para comprender mejor la arquitectura genética de los rasgos complejos (J. Yang et al., 2011). Ahora este *software* está ya incorporado en la versión 1.9 de Plink (http://cnsgenomics.com/software/gcta/#Overview).

g) Shapeit:

Es un método rápido y preciso para la estimación de haplotipos (también conocido como *phasing*) a partir de genotipos o datos de secuenciación (Delaneau, Marchini, & Zagury, 2011), (http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html).

h) fastPHASE:

Software para reconstrucción de haplotipos y estimación de genotipos faltantes de datos de población (https://bioinformaticshome.com/tools/imputation/descriptions/fastPHASE.html#gsc.tab=0), (Scheet & Stephens, 2006).

i) Matlab:

Conocido paquete MATLAB para informática técnica (https://es.mathworks.com/). Es un lenguaje y un entorno interactivo que permite la exploración y visualización de datos.

j) Plink:

Se trata de un conjunto de herramientas gratuitas de análisis de asociación de genoma completo de código abierto, diseñado para realizar una variedad de análisis básicos a gran escala de una manera computacionalmente eficiente (https://www.cog-genomics.org/plink2), (Purcell et al., 2007). Este *software* ha sido utilizado para los análisis de este trabajo, permitiendo la identificación de *outliers*, el cálculo del IBD (*Identity by Descent*) y del DL.

k) skatMeta:

Paquete que calcula las estadísticas necesarias de las variantes estudiadas en cada cohorte individual y luego realiza el metaanálisis. (https://www.rdocumentation.org/packages/skatMeta/versions/1.4.3/topics/skatMeta).

l) SNPassoc:

Este paquete realiza el análisis más común cuando se llevan a cabo estudios de asociación de genoma completo. Estos análisis incluyen estadísticas descriptivas y análisis exploratorio de valores perdidos, cálculo del equilibrio de Hardy-Weinberg, análisis de asociación basado en modelos lineales generalizados (para rasgos cuantitativos o binarios) y análisis de múltiples SNPs (análisis de haplotipos y epistasis). También se pueden estimar distribuciones exactas de *scores* de alelos de riesgo genético. (https://cran.r-project.org/web/packages/SNPassoc/index.html).

m) IMPUTE2:

Se trata de un programa de imputación de genotipos y haplotipos basado en las ideas de Howie y colaboradores (Howie, Donnelly, & Marchini, 2009), (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html).

n) SNPTEST:

Es un programa para análisis de asociación simple de SNPs en GWAS. (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html).

o) METAL:

METAL es una herramienta para escaneos de metaanálisis de GWAS. Puede combinar tanto tests estadísticos y errores estándar como valores-p entre los estudios (teniendo en cuenta el tamaño de la muestra y la dirección del efecto).

El uso de METAL es una buena alternativa para un análisis directo de datos combinados de múltiples estudios. Es especialmente apropiado cuando los datos de los estudios individuales no pueden analizarse juntos debido a las diferencias en origen étnico, distribución de fenotipos, género o restricciones en el intercambio de datos impuestos a nivel individual. El resultado ofrece poca o ninguna pérdida de eficiencia en comparación con el análisis de un conjunto combinado de datos que incluya datos de todos los estudios individuales. (https://genome.sph.umich.edu/wiki/METAL_Documentation#Brief_Description).

p) GWAMA:

Software flexible y de código abierto para el metaanálisis de estudios de asociación de genoma completo. El *software* incorpora una variedad de facilidades para detección de errores y proporciona un rango de estadísticas del metaanálisis. Se presenta con *scripts* de formato simple de archivos que contienen los resultados de cada estudio de asociación y generan resúmenes gráficos de los resultados del metaanálisis del genoma completo (Magi & Morris, 2010), (https://www.geenivaramu.ee/en/tools/gwama).

Con todas las herramientas disponibles y evaluadas podemos llevar a cabo el análisis de nuestros datos y los objetivos de esta tesis, para lo que fundamentalmente hemos utilizado R, Plink y SPSS, evaluando la subestructura poblacional de nuestras muestras con PCA y DAPC.

1.6 EL CENTRO NACIONAL DE GENOTIPADO: HISTORIA, COMPOSICIÓN Y FUNCIONES

El Centro Nacional de Genotipado (CeGen) es una unidad coordinada y creada por el Profesor Ángel Carracedo en el año 2003. Desde el año 2021 ha pasado a formar parte de la Fundación Púbica Galega de Medicina Xenómica (CeGen-FPGMX, nodo de Santiago de Compostela), donde se ha llevado a cabo este trabajo, habiendo sido en los años anteriores parte del Programa de Trabajo de la Plataforma de Recursos Biomoleculares y Bioinformáticos (PRB2 y PRB3) entre otros organismos.

Como Unidad, el CeGen-FPGMX trabaja de forma estable y consolidada ofreciendo servicios de genotipado y complementarios al mismo, como extracción, cuantificación, normalización, purificación y testado de la integridad de ácidos nucleicos, así como servicios de análisis de la metilación del ADN.

Los integrantes de la Unidad conforman un equipo con amplia trayectoria y experiencia en tecnologías de genotipado, bioinformática, bioestadística, diagnóstico genético y farmacogenético, análisis de datos genómicos y también en secuenciación masiva. El equipo humano del CeGen-FPGMX posee un conocimiento y experiencia que posibilita que preste un soporte de alto nivel científico, técnico y tecnológico a aquellos centros de I+D+I, en su mayor medida en el área de Ciencias y Tecnologías de la Salud, que así lo requieren. La Unidad se compone de 3 áreas: la Científico-Técnica, la de Soporte bioinformático, metodológico y estadístico y la Técnológica. Ésta última engloba a la Sección de Plataformas y de Calidad.

La Unidad CeGen-FPGMX, sita en la planta -2 del edificio de consultas externas del Complexo Hospitalario Universitario de Santiago de Compostela, cuenta en sus instalaciones con una sala de recepción de muestras y extracción de ADN, laboratorios de pre- y post-PCR, una sala de equipos de detección, zona de despachos y áreas de uso común con el resto de los servicios de la FPGMX.

Las instalaciones del CeGen están equipadas con las más avanzadas tecnologías necesarias para la prestación de un servicio de calidad, incluyendo el genotipado de SNPs, InDels y CNVs y análisis de metilación del ADN en proyectos de pequeña, mediana y gran escala (en relación al número de muestras y número y tipo de variantes genéticas) así como servicios complementarios, además de tareas de control de calidad, asesoramiento científico-técnico y soporte en el análisis de datos genómicos.

Entre el equipamiento, destacan:

Plataformas de genotipado:

1) Plataforma de genotipado de baja-media capacidad: MassArray de Agena Bioscience, para genotipado de SNPs, InDels y CNVs seleccionados por el usuario y análisis de metilación del ADN. Esta tecnología cuenta con el *Agena Bioscience software: MassArray Typer 4.0, Typer Analyzer v4.0.26 y Epityper v1.2*.

La flexibilidad en el diseño y la posibilidad de usar muestras de menor calidad, como ADN extraído de tejidos en parafina o salivas, la convierte en una tecnología muy demandada.

En relación a los paneles comerciales la cartera de servicios incluye varios relacionados con detección de mutaciones en rutas tumorales: OncoCarta, OncoFOCUS y LungCarta y con aplicación en farmacogenética, diseñados siguiendo las recomendaciones del *PharmaADME working group*: iPLEX PGx Pro (192 SNPs y CNVs en 36 genes), iPLEX PGx 74 (69 SNPs en 20 genes), iPLEX *CYP2D6* (35 SNPs y CNVs en *CYP2D6*), iPLEX *CYP2C19* (31 SNPs en *CYP2C19*) y el panel iPLEX *CYP2C9* / *VKORC1* (51 SNPs en 4 genes), análisis de mutaciones de baja frecuencia en cáncer colorrectal y de pulmón: iPLEX HS Colon Panel (genes *KRAS*, *NRAS*, *BRAF*, *EGFR*, y *PIK3CA*) e iPLEX HS Lung Panel (genes *BRAF*, *EGFR*, *ERBB2*, *KRAS* y *PIK3CA*)

Como apoyo al diagnóstico con esta Plataforma se está analizando, de forma rutinaria, un panel no comercial de diagnóstico y cribado de mutaciones seleccionadas en el gen *CFTR*.

2) Plataformas de genotipado de media-alta capacidad: GeneChip y GeneTitan de ThermoFisher para genotipado de SNPs, InDels y CNVs del genoma completo y pre-seleccionados. Cuenta con el ThermoFisher *software: Command Console software; OncoScan Console software; Chromosome Analysis Suite; Genotyping Console software; Axiom Analysis Suite, Axiom HLA analysis software* y *Axiom CNV summary tools software.*

Con la tecnología GeneChip destacan los paneles *CytoScan High Density array* ((con aproximadamente 2,7 millones de marcadores distribuidos a lo largo de todo el genoma para permitir la detección de anomalías cromosómicas, pérdida de heterocigosidad (LOH) y disomías uniparentales y se utiliza, fundamentalmente, como herramienta diagnóstica de trastornos del neurodesarrollo y anomalías congénitas múltiples)), *el OncoScan FFPE array* (permite detectar, a partir de muestras de tumores sólidos incluidas en parafina, CNVs y LOHs a lo largo de todo el genoma y mutaciones puntuales en genes de cáncer), el *CytoExon array*, que es un panel diseñado para proporcionar cobertura a nivel de exones de genes en todo el genoma y pensado, principalmente, para el descubrimiento de CNVs exónicas y para la validación de pequeñas CNVs detectadas por secuenciación.

Esta Plataforma proporciona herramientas para el diagnóstico genético de los trastornos del neurodesarrollo y/o anomalías congénitas

La Plataforma GeneTitan se utiliza, fundamentalmente, para estudios de asociación en genoma completo (GWAS), replicación y estudios de asociación mediante estrategia de genes candidato, farmacogenética/farmacogenómica y medicina personalizada de precisión. Entre los paneles diseñados "a la carta", destaca el *Axiom Spain Biobank array* presentado en este trabajo y otros con interés para centros de I+D+I en Ciencias y Tecnologías de la Salud, como el *Axiom Precision Medicine Research array*, diseñado para el estudio genético de enfermedades comunes y raras, perfiles de riesgo genético, respuesta inmune, farmacogenómica y áreas relacionadas con la medicina de precisión; el *Axiom PharmacoScan array*, que contiene 4.627 marcadores ADME en 1.191 genes implicados en procesamiento de fármacos, incluyendo marcadores en genes críticos como *CYP2D6*, *CYP2C9* y *CYP4F2*, así como para el genotipado de CNVs.

Además, dispone de otras Plataformas de genotipado de media-alta capacidad, como OpenArray y Taqman (ThermoFisher), que proporcionan apoyo al diagnóstico genético y farmacogenético, bien con procedimientos de cribado o con herramientas para el propio diagnóstico, como el uso de paneles validados para esta finalidad, ofreciendo herramientas para el diagnóstico genético y farmacogenético e implementando procedimientos específicos para la prestación de estos servicios.

En relación al diagnóstico, la Unidad forma parte de la FPGMX, organización que apoya al Servizo Galego de Saúde (SERGAS), realizando los análisis de genética clínica (análisis molecular y citogenético) para todos los hospitales pertenecientes a la red SERGAS, abarcando más de 3,5 millones de habitantes. Esta es una de las principales organizaciones de apoyo de genética clínica en España, ofreciendo uno de los volúmenes de rendimiento más altos y uno de los compendios de pruebas más extensos de diagnóstico de enfermedades hereditarias. La FPGMX es también la responsable de gestionar la Plataforma de ADN del CIBERSAM, siendo CeGen quien se encarga de la extracción y control de calidad de las muestras incluidas en dicha Plataforma.

Equipos de dispensación automática:

El CeGen-FPGMX cuenta con Robots Tecan Freedom Evo con cabezales de 8 y 96 puntas para dispensación de muestras y reactivos y Robot Tecan Aquarius con cabezal de 96 puntas para dispensación de muestras y reactivos; también Robots Beckman NX con cabezal de 96 puntas para dispensación de muestras y reactivos.

Equipamiento auxiliar:

Fluorímetro Genios FL de Tecan para cuantificación de ácidos nucleicos, termocicladores Duales 384-Well deThermoFisher, termocicladores 96-Well GeneAmp PCR System 9700 de ThermoFisher, termociclador Dual 96-Well de ThermoFisher y equipamiento menor como centrífugas de placas, rotor de placas, centrífugas y mini-centrífugas de tubos, agitadores de tubos, agitador magnético, agitador de placas, incubadora de placas y dispensadores manuales: pipetas, pipetas automáticas monocanales, pipetas automáticas de 8 canales, pipetas electrónicas de 8 canales, pipetas automáticas de 12 canales, así como congeladores -80ºC, congeladores -20ºC y neveras 4ºC.

La Unidad CeGen-FPGMX mantiene una actualización constante de las tecnologías y metodologías existentes en el campo. Por este motivo, lleva a cabo proyectos propios para la mejora de sus protocolos y el desarrollo de nuevas metodologías, cuando las posibilidades técnicas así lo requieran. El responsable de la Unidad es, además, investigador principal de numerosos proyectos de investigación en los que se requiere del uso de tecnologías de genotipado y los miembros de la Unidad son igualmente requeridos para participar en proyectos de investigación liderados por otros investigadores colaboradores habituales. Con todos estos proyectos se busca incrementar el conocimiento de las bases genéticas de las enfermedades, de modo que se faciliten el establecimiento de medidas de prevención, un diagnóstico temprano y un pronóstico más preciso.

El principal objetivo del CeGen-FPGMX es proporcionar servicios de genotipado y análisis de metilación de muestras de ADN para proyectos de investigación en ciencias de la salud y de la vida. Los servicios prestados están relacionados, en su mayoría, con el diagnóstico genético y farmacogenético y la investigación biomédica, ofreciendo sus prestaciones a más de 100 instituciones vinculadas con el Sistema Nacional de Salud, como hospitales, fundaciones hospitalarias e institutos de investigación. Cabe destacar también, la prestación de servicios a estructuras del Instituto de Salud Carlos III, como el Instituto de Investigación de enfermedades

raras, el Centro Nacional de Microbiología, el Centro Nacional de Investigaciones Oncológicas, Centro de Investigación Biomédica En Red (Áreas de Enfermedades Raras, de Enfermedades Respiratorias, de Obesidad y Nutrición, de Salud Mental y de Enfermedades Neurodegenerativas) y Banco Nacional de ADN así como a centros del Consejo Superior de Investigaciones Científicas. Además, ha prestado servicios a centros y empresas privadas, nacionales e internacionales, dedicados al diagnóstico genético.

Entre los servicios realizados se han llevado a cabo estudios en oncología, salud mental, enfermedades metabólicas, neurogenética, dermatología, oftalmología, trasplantes, trastornos del neurodesarrollo, enfermedades neurodegenerativas, espondilitis anquilosante, esclerosis múltiple, periodontitis, angiopatía amieloide cerebral, asma, obesidad, fibrosis quística, fracturas vertebrales, sepsis, infección neumocócica, albinismo, alteraciones del sueño, rinitis alérgica, déficit de IgA, farmacogenética, terapia celular y estudios poblacionales entre otros.

En total se han analizado más de 900.000 muestras y se han obtenido más de 50.000 millones de genotipos.

La Unidad CeGen-FPGMX proporciona, también, servicios complementarios de extracción (a partir de diferentes tipos de muestras y volúmenes de partida con diferentes metodologías), cuantificación (con diferentes metodologías), normalización, amplificación, purificación y verificación de la integridad del ADN y control de calidad de los resultados, asesoramiento científico y técnico en las etapas de pre-genotipado, genotipado y post-genotipado y soporte en el análisis de los datos. La cartera de servicios ofertados por la Unidad es muy amplia y constantemente, se desarrollan y añaden nuevos productos a la misma.

El CeGen-FPGMX lleva a cabo constantemente numerosas colaboraciones desde sus inicios con el objetivo de facilitar el acceso transnacional de los usuarios externos a las tecnologías de análisis de ácidos nucleicos y apoyar la genética excepcional, la genómica funcional y la biología de sistemas en Europa. Como ejemplos cabe mencionar el proyecto europeo CHIBCHA con el *Wellcome Trust Sanger Institute* y colaboraciones con consorcios como Geuvadis (*Genetic European Variation in Health and Disease*, HELIX (*The Early Exposome*), en los que el CeGen-FPGMX ha participado realizando toda la parte de expresión, la iniciativa española dentro del Consorcio Inetrnacional del Genoma del Cáncer, en donde hemos realizado la parte de genotipado y los estudios de expresión, los proyectos europeos H2020, BCAST, PANCANRISK y VISAGE en los que realizamos y coordinamos todos los aspectos relacionados con genotipado, al igual que en los proyectos recientemente concedidos, como IMPaCT-GENóMICA, cuyo objetivo es dotar al Sistema Nacional de Salud de una estructura colaborativa para la implementación de la Medicina Genómica en el SNS, de forma que los pacientes puedan acceder con equidad y con tiempos adecuados de respuesta a todas las pruebas genómicas que sean precisas para mejorar su salud, y a la vez obtener datos genómicos que puedan ser utilizados en investigación, mejorando las capacidades de análisis de la infraestructura. Destacar, también, colaboraciones con el Institut Hospital del Mar d'Investigacions Mèdiques, la Universitat de Vic-Universitat Central de Catalunya y el IDIAP Jordi Gol i el IDIBGI en el proyecto GINA-COVID, que analiza las características genéticas que llevan a una COVID-19 grave, así como en el proyecto Cordelia, impulsado por Jaume Marrugat, coordinador del Programa de Epidemiología Cardiovascular del CIBERCV y del Grupo Registre Gironí del Cor (REGICOR) en el Instituto Hospital del Mar de Investigaciones Médicas, en el que el CeGen-FPGMX llevará a cabo el análisis del genoma completo de más de 101.000 personas para determinar el riesgo genético de sufrir determinadas enfermedades cardiovasculares.

Hay que destacar además que de la investigación conjunta y propia de la Unidad CeGen-FPGMX han surgido numerosas publicaciones en revistas de alto impacto y gran cantidad de proyectos obtenidos en concurrencia competitiva.

Otro de los objetivos de la Unidad CeGen-FPGMX es favorecer y contribuir a la formación de investigadores y tecnólogos así como contar con un plan de formación propio que contemple las acciones a implementar en materia formativa para los distintos perfiles de personal integrante de la Unidad. El CeGen-FPGMX contempla estancias formativas en sus instalaciones, organización de cursos formativos dirigidos a la comunidad científica/biosanitaria, dirección de trabajos académicos y eventos de divulgación científica dirigidos a la población general, en centros educativos, en asociaciones culturales, etc, así como visitas de centros educativos a las instalaciones de la Unidad.

La Unidad CeGen-FPGMX le da suma importancia a realizar dichas actividades de difusión que, junto con las de formación, contribuyan a visibilizar la actividad para, por un lado, fomentar el uso de sus plataformas en el entorno científico y empresarial, facilitando la absorción y la explotación de nuevas ideas y tecnologías por parte del tejido productivo y, por otro, para incrementar la cultura científica, tecnológica e innovadora entre la población general.

OBJETIVOS

2 OBJETIVOS

2.1 OBJETIVO GENERAL

El objetivo de este trabajo consiste en la creación de un *array* de SNPs de alta densidad – el Axiom Spain Biobank Array (SBA).

A continuación, generaremos un repositorio de datos de genotipado de muestras de población control española que se hará accesible a la comunidad científica para sus estudios de asociación del genoma completo (GWAS).

2.2 OBJETIVOS ESPECÍFICOS

- Diseño de un *array* de genotipado concreto para el análisis de variación funcional específica de población española.

- Obtención de genotipos de alta calidad mediante el genotipado llevado a cabo con la tecnología Axiom de ThermoFisher de acuerdo con las especificaciones del fabricante (https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0018132_702991_Axiom2_96F_Man_SPG.pdf).

- Generar un repositorio de datos de genotipado de muestras de población control española que se harán accesibles a la comunidad científica para sus estudios de asociación del genoma completo.

- Calcular las frecuencias alélicas y realizar una comparación con poblaciones de referencia en bases de datos públicas; contrastar los resultados en variantes comunes y raras.

- Análisis de la existencia de variabilidad local a escala microgeográfica. Caracterización de los diferentes patrones de estratificación poblacional en variantes comunes y raras para el estudio de trastornos genéticos complejos.

- Ofrecer su disponibilidad a la comunidad científica como patrones de comparación al ser representativos de la población española.

MATERIAL Y MÉTODOS

3 MATERIAL Y MÉTODOS

El protocolo para llevar a cabo este trabajo ha sido aprobado desde el punto de vista ético y científico por los organismos correspondientes a la Universidade de Santiago de Compostela en la categoría de investigación con riesgo mínimo.

3.1 ANÁLISIS GENÉTICO DE LAS MUESTRAS

3.1.1 El material biológico

3.1.1.1 Selección y origen de las muestras

Todas las muestras empleadas en este trabajo han sido solicitadas al Banco Nacional de ADN (BNADN) (Universidad de Salamanca; http://www.bancoadn.org/), por lo que ya han sido recogidas bajo los correspondientes consentimientos informados y dentro de la normativa aplicable a un Biobanco.

El ADN fue extraído a partir de sangre periférica de donantes voluntarios y habituales de más de 30 centros de transfusión y bancos de sangre distribuidos por todo el territorio nacional, lo que nos ha permitido conseguir una colección muy representativa de la población residente en España.

Se solicitó al BNADN una muestra total de **3.169 individuos**, seleccionados de forma aleatoria y pertenecientes a toda la población española con los siguientes datos asociados:

- sexo
- edad
- origen geográfico
- origen geográfico de los padres
- origen geográfico de los cuatro abuelos

3.1.1.2 Extracción del ADN

La extracción de ADN genómico se llevó a cabo de manera manual a partir de sangre total en EDTA (2,5-20 ml) como anticoagulante, obtenida mediante punción en el brazo del donante, y fue realizada por los técnicos de laboratorio del BNADN en el área de extracción de ADN, donde se trabaja cumpliendo las normas establecidas en la guía de trabajo "Trabajo en laboratorios bajo condiciones de seguridad" (NC-OD-000004).

3.1.2 Genotipado

3.1.2.1 Diseño del Axiom Spain BioBank Array Plate de ThermoFisher: selección de marcadores

Este estudio se realizó en la Fundación Pública Galega de Medicina Xenómica, Universidade de Santiago de Compostela y en el Centro Nacional de Genotipado (CeGen), nodo de Santiago de Compostela (CeGen-FPGMX).

Los *arrays* para análisis de genoma completo pueden ser una alternativa a la Secuenciación de Nueva Generación a la hora de estudiar variantes de baja frecuencia, siempre y cuando estén enriquecidos en variantes raras específicas de población.

Por este motivo hemos diseñado el *Axiom Spain BioBank Array Plate*, que abarca la variación común de una serie de enfermedades complejas y lo hemos enriquecido con variantes raras de población española. Nuestro principal objetivo es obtener una herramienta para explorar todas estas variaciones de una manera más fácil y factible, ya que parece una buena estrategia para estudiar enfermedades complejas, no solo analizando la variación común para

tales enfermedades, sino también una variación rara que puede ser específica para la población de estudio.

El panel ha sido desarrollado por la Unidad del Centro de Genotipado del nodo Santiago de Compostela en colaboración con ThermoFisher, con la aportación del grupo de Medicina Xenómica y bajo la dirección del Profesor Ángel Carracedo.

El *Axiom Spain BioBank Array Plate* de ThermoFisher ha sido desarrollado con la Plataforma *GeneTitan* de ThermoFisher para paneles de genotipado de SNPs, InDels y CNVs del genoma completo y pre-seleccionados (media-alta capacidad). Esta plataforma utiliza la tecnología *Axiom Genotyping*, un sistema de genotipado de ThermoFisher, empleado, fundamentalmente, para estudios de asociación de genoma completo (GWAS), replicación y estudios de asociación mediante estrategia de genes candidato, farmacogenética / farmacogenómica y medicina personalizada de precisión. Incluye paneles de SNPs, InDels y CNVs en formato de placa de 96 y 384 muestras, todos los reactivos y las herramientas de análisis de datos, así como el equipamiento para el procesado completamente automatizado con el equipo *GeneTitan-Multi-Channel Instrument*.

Este panel de genoma completo específico de población española se ha generado a partir del *UK Biobank Axiom® Array* también de ThermoFisher (Figura 3.1) Aa:

https://www.thermofisher.com/order/catalog/product/902502#:~:text=UK%20Biobank%20Axiom%E2%84%A2%20Array%20was%20designed%20by%20and%20for,single%20comprehensive%20low%2Dcost%20solution.

Se consideró la selección de este *array* porque contiene todas las ventajas de los paneles de SNPs de alta densidad junto con la posibilidad de personalizar módulos.

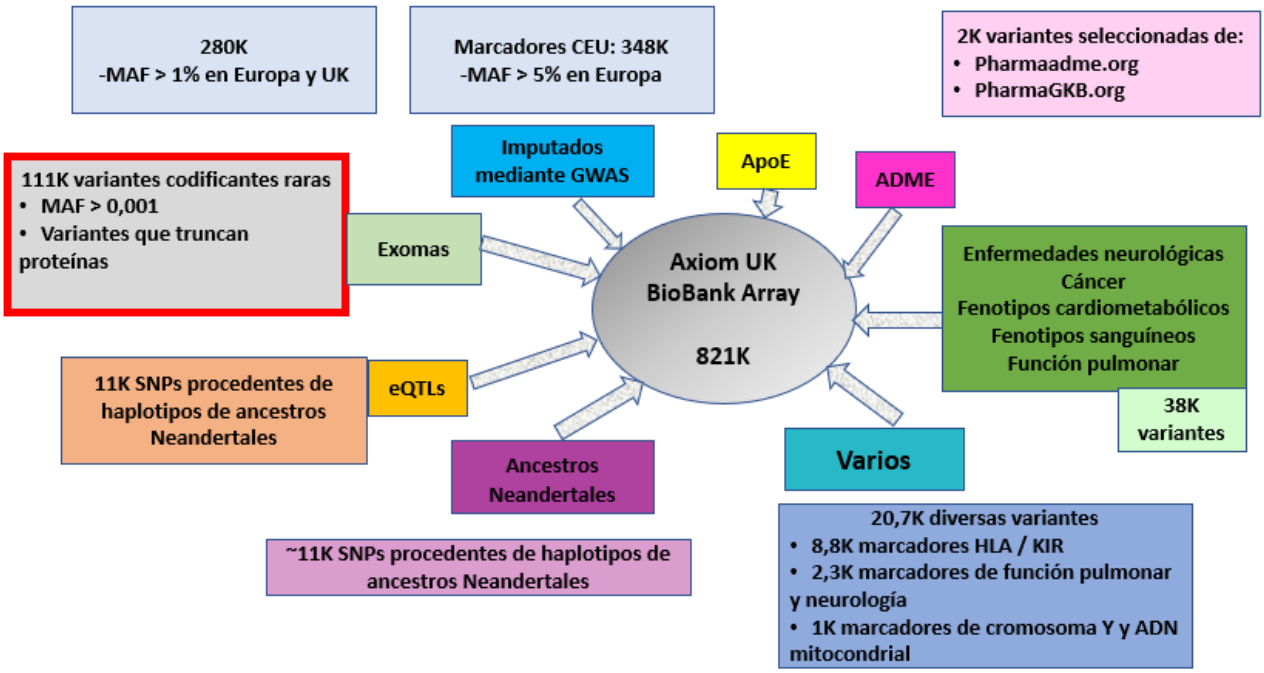


Figura 3.1. *UK Biobank Array Content Summary*. Adaptada de https://assets.thermofisher.com/TFS-Assets/LSG/brochures/uk_axiom_biobank_contentsummary_brochure.pdf.

El *UK Biobank Axiom® Array* (UKBB) estaba inicialmente compuesto por 820.967 marcadores, optimizado para estudios de asociación de variantes comunes y de baja frecuencia a nivel de todo el genoma involucradas en función biológica y enfermedad humana en poblaciones de Europa y ascendencia británica. Se trata de un *array* para GWAS que incorpora variación en regiones exónicas e intrónicas. A este *array* se le han añadido, además, otras regiones exónicas de interés, es decir, con variantes funcionales que específicamente han demostrado presentar asociación con ciertas enfermedades reportadas en artículos científicos.

Ya que la variación de baja frecuencia presenta gran variabilidad geográfica y habitualmente incluso hay variantes raras específicas de sitio, el módulo de 111.000 variantes raras del UKBB puede ser de limitada utilidad en el análisis de poblaciones no británicas. Para diseñar el SBA se ha eliminado dicho módulo de 111K y se ha sustituido por variación rara identificada en población española y otros marcadores seleccionados en diferentes ámbitos, tal y como se describe a continuación.

Incorporación de variación codificante rara en población española

Para capturar la diversidad única de la población española identificamos un conjunto de variantes que se han incluido en el panel, para cubrir variación importante de baja frecuencia y alelos funcionales asociados a diferentes patologías y trastornos dentro de la población española (Figura 3.2). La selección de estas variantes se llevó a cabo, previo consentimiento informado, a partir de la secuenciación de exomas (WES, del inglés *Whole Exome Sequencing*) de un total de 639 individuos no relacionados distribuidos en dos cohortes diferentes de origen español (muestras anonimizadas y fenotipadas como sanas). Una de las cohortes se corresponde con 115 muestras de exomas de población gallega (cohorte G) procedentes de la Fundación Pública Galega de Medicina Xenómica (exomas que estaban secuenciados en aquel momento en la

FPGMX) y las otras 524 muestras proceden de exomas de población valenciana (cohorte V) (Pena-Chilet et al., 2021).

La WES en el DNA germinal de la cohorte gallega fue realizada en el secuenciador 5500xl SOLiD™ (ThermoFisher Scientific, Waltham, Massachusetts, EEUU) usando el kit *SureSelect Human All Exon* V5 (Agilent Technologies, Santa Clara, CA, USA).

El proceso consiste en la preparación de librerías, que comienza con una fragmentación del ADN genómico en un sonicador Covaris™ seguida de una reparación de extremos (*ends repair*). A continuación, se lleva a cabo la purificación y la selección de tamaños de los fragmentos mediante el uso de *AMPure XP beads*. Una vez realizado el chequeo de calidad en el *2200 TapeStation System* (Agilent Technologies), se ligan los adaptadores *5500 SOLiD™ Fragmem Library Barcode*. Posteriormente se purifican los fragmentos con los adaptadores unidos a los extremos mediante el uso de *AMPure XP beads* y se realiza la amplificación de la librería con los adaptadores ligados. Una vez amplificados los fragmentos de ADN se lleva a cabo la hibridación de la librería mediante el uso de *SureSelect Capture Library*. A continuación, mediante *beads* magnéticas se procede a *Hybrid Capture Selection* y se realiza la amplificación por PCR de las regiones capturadas. Se purifica la muestra mediante el uso de *AMPure XP beads* y se realiza el chequeo de calidad en el *2200 TapeStation System*.

Una vez preparadas las librerías estas son amplificadas mediante una PCR en emulsión y las microesferas con los fragmentos de ADN generados son depositadas en un *SOLiD™ FlowChip* y en el *Ion Proton™ system chip* para su secuenciación en el 5500xl SOLiD™.

Las secuencias obtenidas fueron alineadas con el genoma humano de referencia hg19 (*Build* GRCh37). Para alinear las secuencias del 5500xl SOLiD™ se utilizó una *suite* completa de análisis de Lifescope. La WES se realizó con una cobertura media de, al menos, 50X.

Para el *calling* de las variantes se utilizaron los *softwares Lifescope* y *Genome Analysis Toolkit* 3.0 (GATK 3.0) (McKenna et al., 2010).

La anotación funcional de cada variante se realizó mediante ANNOVAR (https://annovar.openbioinformatics.org/en/latest/#reference; (K. Wang, Li, & Hakonarson, 2010).

Los resultados obtenidos directamente de la secuenciación se almacenaron en formatos FastQ o XSQ según la plataforma empleada.

El análisis de ambas cohortes se llevó a cabo con el s*oftware* R versión 3.2.4. El estudio se realizó por separado y en combinación, con el fin de explorar la posibilidad de encontrar variantes en común. Se aplicaron varios requisitos a ambas cohortes para filtrar las variantes que se incorporarían a la matriz. Se seleccionaron solo marcadores puntuales (no inserciones o deleciones) en regiones codificantes, áreas de *splicing* y UTR (*untraslated region*), con una frecuencia (MAF) < 0,01.

A continuación, se seleccionaron las variantes en función de SIFT y Polyphen-2:

Para el algoritmo SIFT, los valores por debajo de 0,05 se clasificaron como perjudiciales, lo que significa que se predice que la sustitución afectará la función de la proteína y por encima de 0,05 significa que se predice que la sustitución será funcionalmente neutra (Kumar, Henikoff, & Ng, 2009). En el caso de Polyphen-2 (Ramensky, Bork, & Sunyaev, 2002), este calcula la probabilidad, en función del Teorema de Bayes, de que una determinada mutación sea dañina, dando la posibilidad de que dicha variante se clasifique como dañina cuando en realidad no lo es. Esta se clasifica como dañina definiendo tres umbrales: umbral > 0,9 para las variantes dañinas que afectan la función de la proteína; umbral < 0,5 para variantes benignas que no conducen a cambios en el fenotipo y umbral = 0,447 > x > 0,908 para variantes posiblemente dañinas que tengan también en cuenta la estructura proteica. Bajo estas premisas se aplicaron otros dos filtros: se seleccionaron variantes "dañinas" con SIFT y finalmente variantes "dañinas" y "posiblemente dañinas" con Polyphen2 (HVID).

Tras aplicar este conjunto de filtros se obtuvo una lista de 52.464 variantes presentes en al menos una de las dos cohortes. La MAF fue recalculada en el subgrupo de 4.320 variantes comunes a ambas cohortes y aquellas cuya frecuencia resultante no pasó el corte de MAF 0,01 fueron excluidas. El hecho de que encontrásemos 4.320 variantes raras en común puede deberse a que el filtro aplicado para seleccionar este tipo de marcador fue aquellas con una frecuencia menor a 0,01, umbral no muy restringido, considerando que, la variación rara, se define como variantes que se presentan en una frecuencia menor a 0,001.

De esta forma, finalmente se obtuvo una lista de 50.536 variantes raras para la población española (Figura 3.2).

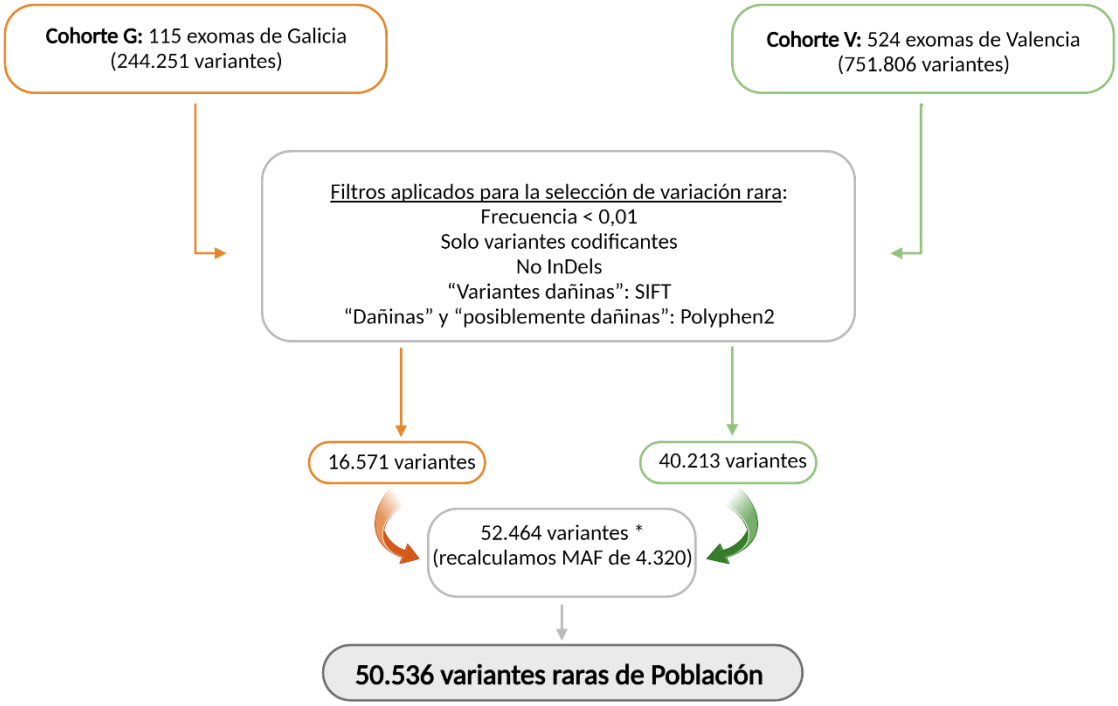


Figura 3.2. Filtros aplicados para la selección de variantes raras de población española. Fuente propia. Realizada con www.Biorender.com.

Dado que era posible incorporar más variantes en el módulo a diseñar, y para complementar y enriquecer el contenido del *array*, se decidió abrir el espectro de selección de variantes en dos direcciones: en primer lugar, agregar variantes codificantes en dos áreas de interés: psiquiatría e inmunología (Figura 3.3).

Incorporación de variantes de psiquiatría e inmunología

Como punto de partida, se tomaron dos listas de variantes que incluían combinaciones tanto raras como comunes, compuestas por 588.628 para psiquiatría (*InfiniumPsychArray-24 v1*; https://emea.illumina.com/products/by-type/microarray-kits/infinium-psycharray.html) y 253.702 para inmunología (*Infinium ImmunoArray-24 v2*; https://emea.illumina.com/products/by-type/microarray-kits/infinium-immunoarray.html). Para determinar las variantes que se agregarían a la matriz, se aplicaron varios filtros (Figura 3.3). En este caso fueron seleccionadas todas las variantes codificantes (incluyendo SNPs e

InDels) clasificadas como "dañinas" (SIFT) o "dañinas" o "posiblemente dañinas" (Polyphen2). Esto condujo a una lista final de 39.246 variantes.

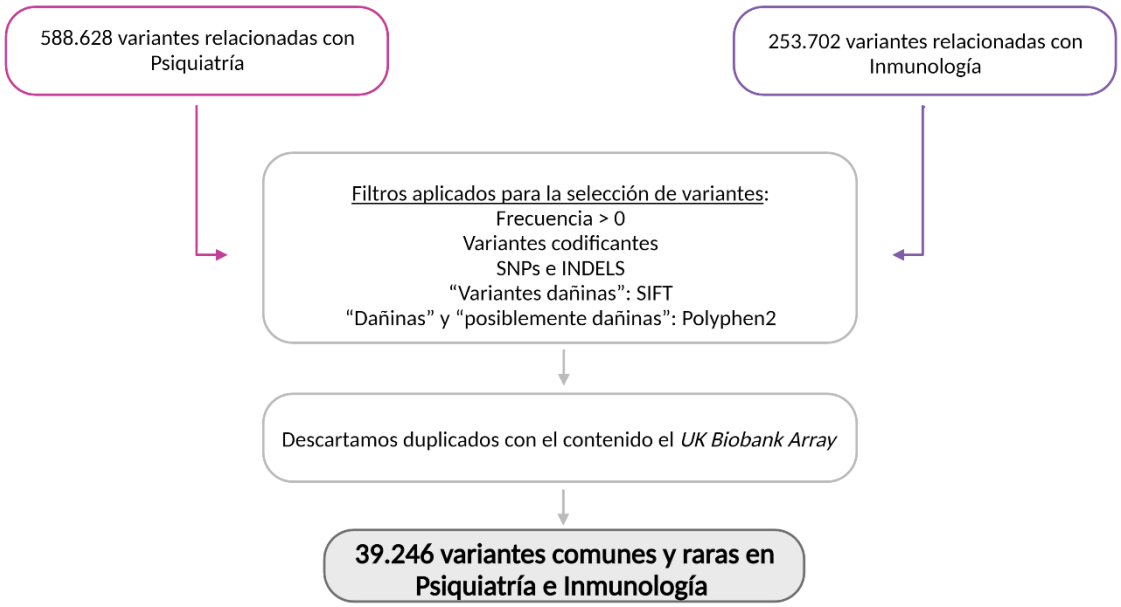


Figura 3.3. Filtros aplicados para la selección de variantes codificantes comunes y raras en psiquiatría e inmunología. Fuente propia. Realizada con www.Biorender.com.

Finalmente complementamos la matriz con variación común y de baja frecuencia, presente no solo en regiones codificantes, sino también en no codificantes y para otras enfermedades complejas.

Los filtros aplicados y las áreas que participaron son los que se detallan a continuación:

filtros establecidos: variación común (frecuencia > 0,05), SNPs e InDels presentes tanto en región codificante como no codificante. Las variantes fueron revisadas y excluidas y otras incluidas por un panel de expertos. Los especialistas que reclutaron las variantes son líderes nacionales en las distintas áreas y están participando en muchos proyectos nacionales e internacionales con acceso a diversas bases de datos e información privada de consorcios.

Así se incluyeron un total de 27.911 variantes distribuidas entre las siguientes áreas:

- Cáncer: incluyendo cáncer de mama, colon, próstata y ovario; teníamos un total de 257 variantes seleccionadas por las Dras. Ana Vega y Clara Ruiz, que se quedaron en 224 al descartar duplicados con el *UK Biobank Array*.
- Radiogenómica: la Dra. Ana Vega seleccionó 2.344 variantes implicadas en cáncer de mama, próstata y pulmón (radiogenómica), que se quedaron finalmente en 2.253.
- Esclerosis múltiple: se seleccionaron inicialmente 82 variantes que se quedaron en 75.
- Variantes en HLA: de las 440 variantes seleccionadas inicialmente, nos quedamos con 110 debido a su menor frecuencia.
- Cardiogenética: la Dra. María Brión aportó 7.676 variantes inicialmente, que finalmente se redujeron a 6.868.

- Nefrología: en este caso nos quedamos solo con regiones codificantes. El Dr. Miguel Ángel García aportó dos listados: por un lado, variantes codificantes y por otro no codificantes, pero por problemas de espacio en el *array* hubo que seleccionar el primer archivo. Tras descartar los marcadores comunes con el *UK Biobank Array*, definimos un total de 5.090 variantes a incluir.
- Farmacogenética: incluimos 2.062 variantes que no estaban presentes en el *UK Biobank Array*. En este figuraban otras 2.143.
- Variantes codificantes comunes de baja frecuencia: 11.229.

Una vez recopilada toda la información de las variantes implicadas en las diferentes enfermedades, se revisaron cuidadosamente tres aspectos: 1) secuencia de referencia utilizada para anotar variantes (debe ser utilizada la versión HG19 del genoma (GRCh37 / hg19)); 2) el cromosoma y la posición deben ser coincidentes con otras bases de datos, concretamente UCSC (https://genome.ucsc.edu/) y dbSNP-Q (https://bio.tools/dbsnp-q); 3) se confirmó el alelo alternativo y de referencia.

Tras esta revisión finalmente tenemos un total de 114.898 variantes comunes y raras seleccionadas de distintas áreas y que no estaban presentes en el *UK Biobank Array* ("*UKBB_content_reduced*").

Este listado final de 114.898 marcadores abarca variantes raras para la población española y variantes raras, de baja frecuencia y frecuentes para muchas enfermedades complejas. Debido a que el conjunto de variantes finalmente seleccionadas superaba el tamaño del módulo de variación rara a sustituir (111K) hubo que sustraer variantes de otro módulo del *UK Biobank Axiom® Array*, concretamente de la categoría "Cobertura genómica para variantes de baja frecuencia". A pesar de esta reducción, la cobertura genómica y la precisión de imputación del *Axiom Spain Biobank Array* tienen un buen rendimiento en comparación con los diseños de matrices globales para múltiples poblaciones (Tabla 3.1).

Se puede seleccionar un marcador para más de una categoría, pero solo aparece en los detalles de la matriz una vez. Cabe señalar que cuando nos referimos a la posible enfermedad causante o variantes asociadas a la enfermedad, no queremos dar a entender que se ha establecido relación, cualquier papel definitivo o asociación con la enfermedad.

Tabla 3.1. Cobertura genómica y la precisión de imputación del *Axiom Spain Biobank Array*. La cobertura genómica fue calculada frente a la población íbera de referencia de la base de datos 1000G (datos de la Fase II). La cobertura del SBA está infraestimada, ya que no se han incluido para el cálculo variantes nuevas no descritas en 1000G. Fuente: Centro Nacional de Genotipado, Nodo de Santiago de Compostela.

	Axiom Spain Biobank Array		Axiom Precision Medicine Research Array		Ilumina Global Screening Array	
MAF %	Cobertura Genómica	Precisión de imputación	Cobertura Genómica	Precisión de imputación	Cobertura Genómica	Precisión de imputación
1-5%	0,71	0,84	0,67	0,81	0,67	0,81
>5%	0,92	0,94	0,90	0,93	0,88	0,91
>1%	0,86	0,91	0,84	0,90	0,82	0,89

En definitiva, el *Axiom Spain BioBank Array Plate* contiene 758.740 marcadores distribuidos por todo el genoma. Incluye por tanto variantes comunes y variantes de baja frecuencia y ha sido específicamente diseñado para cubrir variación funcional específica de población española. 114.898 variantes fueron seleccionadas en población española, pero, debido a que muchos de estos marcadores son raros y no genotipados previamente en ThermoFisher, para 56.183 marcadores se incluyeron dos sondas. De este modo, el número total de posiciones anotadas (inicialmente genotipadas) en el *array* fue 814.923.

3.1.2.2 Requerimientos y preparación del ADN para el genotipado con el *Spanish Biobank Genotyping Array Plate de TermoFisher*

El genotipado se efectuó en el Centro Nacional de Genotipado - FPGMX (nodo de Santiago de Compostela) con la tecnología Axiom en un *GeneTitan® Multichannel* (MC) *Instrument*.

El protocolo de trabajo de genotipado precisa que las muestras se encuentren a determinadas condiciones, por lo que han sido sometidas a distintas técnicas para cumplir los requerimientos de la tecnología. El ADN debe ser de doble cadena y se necesita una cantidad total inicial de 250 - 500 ng, que serían 50 ng/µl (cuantificados por PicoGreen) en un volumen de 5 - 10 µl y debe estar libre de inhibidores de polimerasas o de reacciones enzimáticas (por ejemplo, altas concentraciones de agentes quelantes, sales, etc.…). Si el ADN genómico contuviese inhibidores (no es nuestro caso), se podría recurrir a protocolos de purificación. La muestra debe estar resuspendida en TE EDTA reducido (10mM Tris pH 8,0, 0,1mM EDTA pH 8,0) o agua. También se necesita ADN no degradado, para lo que se comprueba su integridad en gel de agarosa al 1% (El 90% del ADN debe tener un tamaño superior a 10 kb).

Una vez conocida la concentración de ADN para cada muestra, se ajustó la concentración y volumen requeridos, con agua ultrapura en placas de 96 pocillos de fondo redondeado. Se dispensó una muestra por pocillo limpio. Las placas se sellaron, se centrifugaron y se almacenaron a -20°C hasta su posterior genotipado. En los casos en los que se procedió a la amplificación del ADN inmediatamente, las placas se dejaron a temperatura ambiente (18 a 25 °C).

La cuantificación de las muestras de ADN se llevó a cabo con el QUANT-IT dsDNA BR Assay kit (Invitrogen_Q33130) mediante el *software* de laboratorio Freedom EVOware®, realizando el ensayo de cuantificación con el método Magellan™.

La normalización de las muestras a la concentración idónea se realizó de forma automatizada utilizando el robot Freedom EVOware®, que permite garantizar la integridad del proceso a la hora de manipular las muestras. Posteriormente y dado que la tecnología de genotipado requiere ADN no degradado como se ha mencionado, se evaluó la integridad mediante geles de agarosa 1% (Tecnología SYBR GREEN).

3.1.2.3 Principios metodológicos del sistema *Axiom* de ThermoFisher

Una vez preparadas las muestras de ADN a las condiciones requeridas, el proceso de genotipado mediante el sistema *Axiom* de ThermoFisher sigue un protocolo específico y semi-automatizado que consiste en una preparación manual seguido de procesamiento automatizado de las placas en el instrumento *GeneTitan Multi-Channel* (MC). El genotipado se llevó a cabo de acuerdo con las especificaciones del fabricante: (*Axiom® 2.0 Assay Manual Workflow*; https://www.thermofisher.com/es/es/home/life-science/microarray-analysis.html).

En el protocolo con esta plataforma (*AxiomTM Genome-Wide Human assay*), el ADN genómico total se amplifica y fragmenta hasta 25-125 pb. Estos fragmentos se purifican y resuspenden con la solución de hibridación que se transfiere al *GeneTitan Instrument* para seguir su procesamiento completamente automatizado (hibridación en las placas de 96, tinción, lavado y escaneado). Las imágenes se procesan automáticamente para obtener los genotipos con el algoritmo *Axiom* GT1, disponible a través del *software* de ThermoFisher *Genotyping Console* (GTC) y *Axiom Analysis Suite*: https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html.

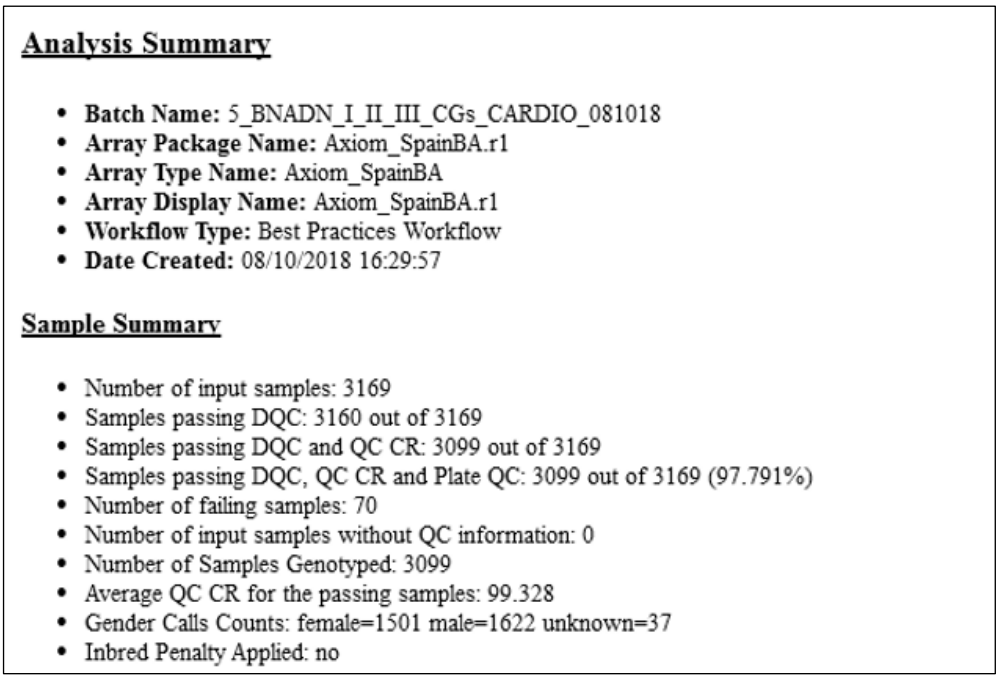
3.1.2.4 Obtención de los genotipos. Análisis bioestadístico y bioinformático. Controles de calidad

Las 3.169 muestras solicitadas al BNADN fueron genotipadas con el *Axiom Spain BioBank Array*, reduciéndose a 3.099 tras fallar 70 en esta fase del genotipado (Figura 3.5). Los umbrales establecidos para este filtrado inicial fueron los siguientes:

DQC (*Dish Quality Control*): ≥ 0,82: las muestras con un valor DQC inferior al predeterminado se vuelven a procesar o se excluyen del estudio.

QC (*Quality Control*) *call rate*: ≥ 97: este sería un segundo paso en el control de calidad, en el que se filtran las muestras (sometiéndolas a un proceso de genotipado con sondas determinadas) que han superado el paso anterior (valores DQC ≥ al establecido). Los genotipos producidos por este paso son solo para el propósito del control de calidad de la muestra y no están destinados al análisis posterior.

Tras el genotipado, lo habitual en el protocolo, es que los datos se filtren de acuerdo con las recomendaciones de ThermoFisher ("*Best Practice Supplement to Axiom® Genotyping Solution Data Analysis User Guide*"), utilizando el *software Axiom Analysis Suite* (Figura 3.6) y posteriormente un filtrado con *SNPolisher*. Esta herramienta categoriza los marcadores en función de la calidad de los clústeres y genera una lista de marcadores recomendados para poder usarla como filtro al exportar. Nosotros no hemos utilizado esta última herramienta a la hora de exportar los genotipos (en formato Plink), ya que en el listado de marcadores recomendados incluye la selección de un miembro de cada par de sondas duplicadas que no queremos realizar en este momento.



Analysis Summary

- **Batch Name:** 5_BNADN_I_II_III_CGs_CARDIO_081018
- **Array Package Name:** Axiom_SpainBA.r1
- **Array Type Name:** Axiom_SpainBA
- **Array Display Name:** Axiom_SpainBA.r1
- **Workflow Type:** Best Practices Workflow
- **Date Created:** 08/10/2018 16:29:57

Sample Summary

- Number of input samples: 3169
- Samples passing DQC: 3160 out of 3169
- Samples passing DQC and QC CR: 3099 out of 3169
- Samples passing DQC, QC CR and Plate QC: 3099 out of 3169 (97.791%)
- Number of failing samples: 70
- Number of input samples without QC information: 0
- Number of Samples Genotyped: 3099
- Average QC CR for the passing samples: 99.328
- Gender Calls Counts: female=1501 male=1622 unknown=37
- Inbred Penalty Applied: no

Figura 3.4. Filtrado inicial del *Axiom Analysis Suite* (https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html). (Sesión Axiom específica de este trabajo).

Este flujo de trabajo realiza un control de calidad para muestras y placas, caracteriza los genotipos de aquellas muestras que pasan los umbrales de control de calidad y luego, mediante el *SNPolisher*, categoriza los conjuntos de sondas que emiten intensidades bien agrupadas y cuyos genotipos son recomendados para el análisis posterior. Los detalles están disponibles en la Guía de análisis de datos de *Axiom Genotyping Solution* (N/P 702961).

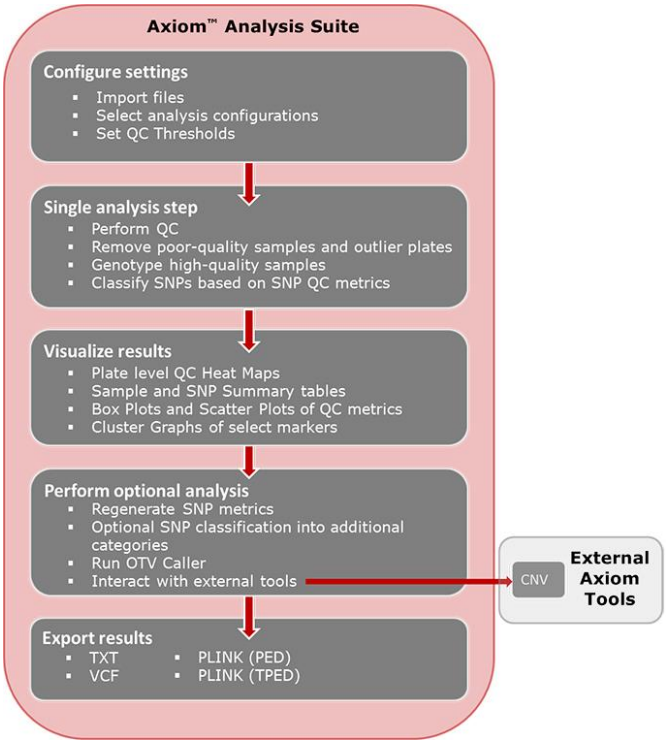


Figura 3.5. Flujo de trabajo completo del *Axiom Analysis Suite* (https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html).

Controles de calidad

El control de calidad de las muestras y genotipos se realizó con los *softwares* Plink y R.

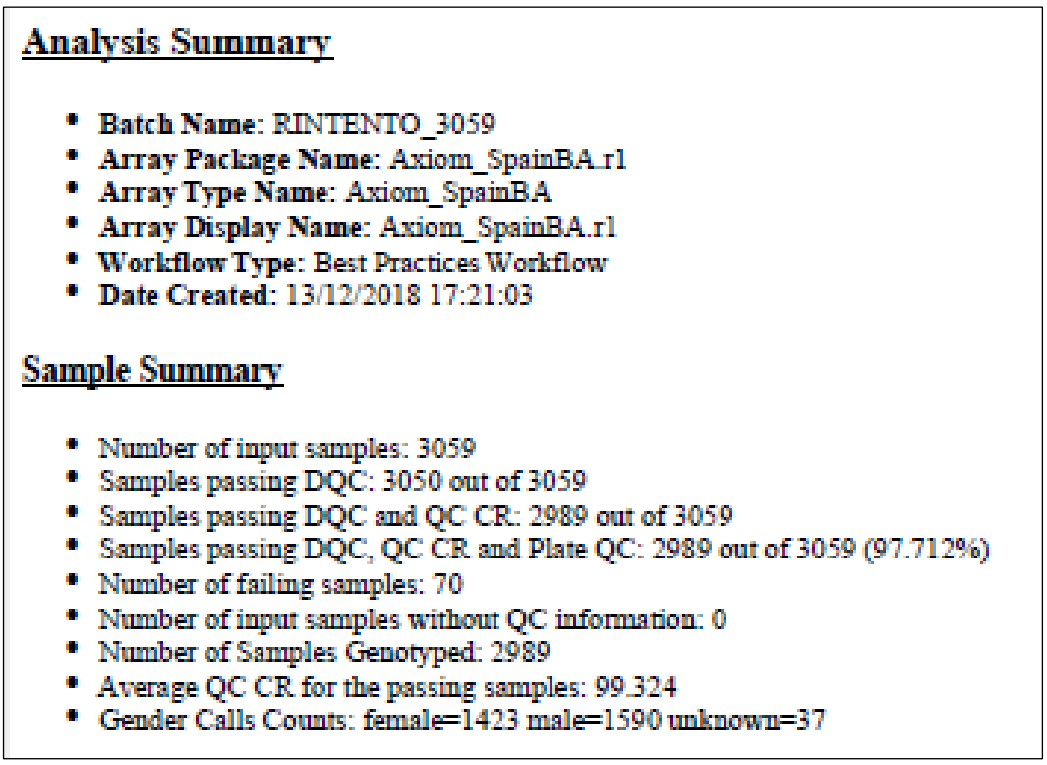
Como se ha descrito en la introducción en el apartado de genética computacional, Plink (Purcell et al., 2007) es un conjunto de herramientas con el que se pueden manipular y analizar rápidamente y en su totalidad grandes conjuntos de datos, que comprenden cientos de miles de marcadores genotipados en un elevado número de muestra. Además de proporcionar herramientas para hacer que los pasos analíticos básicos sean computacionalmente eficientes, Plink también admite algunos enfoques novedosos para el análisis de datos de todo el genoma. Los cinco dominios principales de la función son la gestión de datos, resumen de estadísticas, estratificación de la población, análisis de asociación y estimación de la IBD (del inglés *Identity By Descent*).

R (https://cran.r-project.org/) es un sistema para análisis estadísticos y gráficos creado por Ross Ihaka y Robert Gentleman (Ihaka & Gentleman, 1996). R tiene una naturaleza doble de programa y lenguaje de programación y es considerado como un dialecto del lenguaje S creado por los Laboratorios AT&T Bell. El desarrollo y distribución de R son llevados a cabo por varios estadísticos conocidos como el Grupo Nuclear de Desarrollo de R.

En primer lugar, se eliminaron los marcadores de mala calidad en función de la categorización realizada con la herramienta *SNPolisher* del *software Axiom Analysis Suite.* La tabla que se genera tras la ejecución del algoritmo y que contiene la información de la calidad de los marcadores fue exportada y manipulada en R; todos los marcadores definidos como *other*, OTV (del inglés *off-target variants*) y *CallRateBelowThreshold* fueron seleccionados para generar un listado de marcadores a eliminar en Plink.

En cuanto a los controles de calidad de las muestras se evaluó en primer lugar, la posible existencia de muestras duplicadas o emparentadas mediante el cálculo de la matriz de IBD a partir del total de SNPs autosómicos independientes (opción *indep-pairwise* en Plink, analizando la existencia de r2>0,2 en ventanas deslizantes de 50 marcadores). De esta forma, pudimos identificar 110 pares emparentados o muestras duplicadas (PI_HAT> 0,1875). Estas 3.059 muestras no emparentadas fueron nuevamente genotipadas para evitar cualquier posible sesgo debido a la inclusión de individuos emparentados y sobre todo muestras duplicadas, especialmente en el genotipado de variantes raras.

Tras este nuevo genotipado volvieron a fallar 70, por lo que finalmente obtuvimos 2.989 muestras válidas para incluir en el estudio (Figura 3.7).



Analysis Summary

- **Batch Name**: RINTENTO_3059
- **Array Package Name**: Axiom_SpainBA.r1
- **Array Type Name**: Axiom_SpainBA
- **Array Display Name**: Axiom_SpainBA.r1
- **Workflow Type**: Best Practices Workflow
- **Date Created**: 13/12/2018 17:21:03

Sample Summary

- Number of input samples: 3059
- Samples passing DQC: 3050 out of 3059
- Samples passing DQC and QC CR: 2989 out of 3059
- Samples passing DQC, QC CR and Plate QC: 2989 out of 3059 (97.712%)
- Number of failing samples: 70
- Number of input samples without QC information: 0
- Number of Samples Genotyped: 2989
- Average QC CR for the passing samples: 99.324
- Gender Calls Counts: female=1423 male=1590 unknown=37

Figura 3.6. Filtrado del *Axiom Analysis Suite* tras eliminar 110 muestras emparentadas excluidas por IBD. (Sesión Axiom específica de este trabajo).

En el archivo resultante, nuevamente exportado en formato plink y tras eliminar los marcadores de mala calidad, se confirmó la ausencia de parentesco (figura 3.7) o estratificación, mediante PCA con el mismo *subset* de marcadores independientes usados en la evaluación de parentesco. Posteriormente se aplicaron los filtros de control de calidad estándar establecidos: las muestras y los marcadores con una tasa de genotipado inferior al 98% o los marcadores que se desviaron significativamente del equilibrio HW ($p < 0,0001$) fueron eliminados.

Control de calidad adicional: evaluación de sondas duplicadas

Tras el control de calidad se calculó la frecuencia alélica de cada marcador y se fusionó con la información de frecuencia alélica obtenida de 1000G en población europea. Para facilitar la comparación de frecuencias observadas en nuestro estudio y lo registrado en población europea, todos los marcadores fueron categorizados en 4 grupos:

-Monomórficos (*Monomorphic*): alelo menor no encontrado en la población analizada.
-Raros (*Rare*): MAF<1%.
-Baja frecuencia (*Low-frequency*): 1%<MAF<5%.
-SNPs: MAF>5%.

Esta clasificación ya había sido utilizada en estudios previos, en los que, los objetivos, en su mayoría, iban enfocados al análisis del impacto de las variantes raras y de baja frecuencia en las enfermedades comunes (Bomba, Walter, & Soranzo, 2017), (Babron et al., 2012), (Ingles et al., 2018).

Esto se hizo en función tanto de la frecuencia alélica observada en las muestras del estudio como en función de la frecuencia registrada en 1000G-*europe* (población europea e íbera), y se llevó a cabo con el *software* R (https://cran.r-project.org/).

En el caso de los marcadores asociados a dos sondas se analizó en primer lugar la concordancia de categoría en ambos casos y paralelamente su concordancia con 1000G. Todos los casos en los que la categoría del marcador discrepaba en las dos sondas y aquellos marcadores (con dos o una única sonda) que discrepaban de la categoría de 1000G (*rare* / SNP o al revés) fueron seleccionados, y sus *plots* fueron revisados exhaustivamente en el *Axiom analysis Suite* para confirmar la calidad del genotipado (80 en concreto). En general, en los casos en los que la categoría del marcador discrepaba en las dos sondas y una de ellas concordaba con 1000G se seleccionó esta última (salvo que el *plot* mostrara un claro fallo de genotipado). En las Figuras 3.7, 3.8 y 3.9 se pueden ver ejemplos de clúster *plot* específicos de este genotipado.

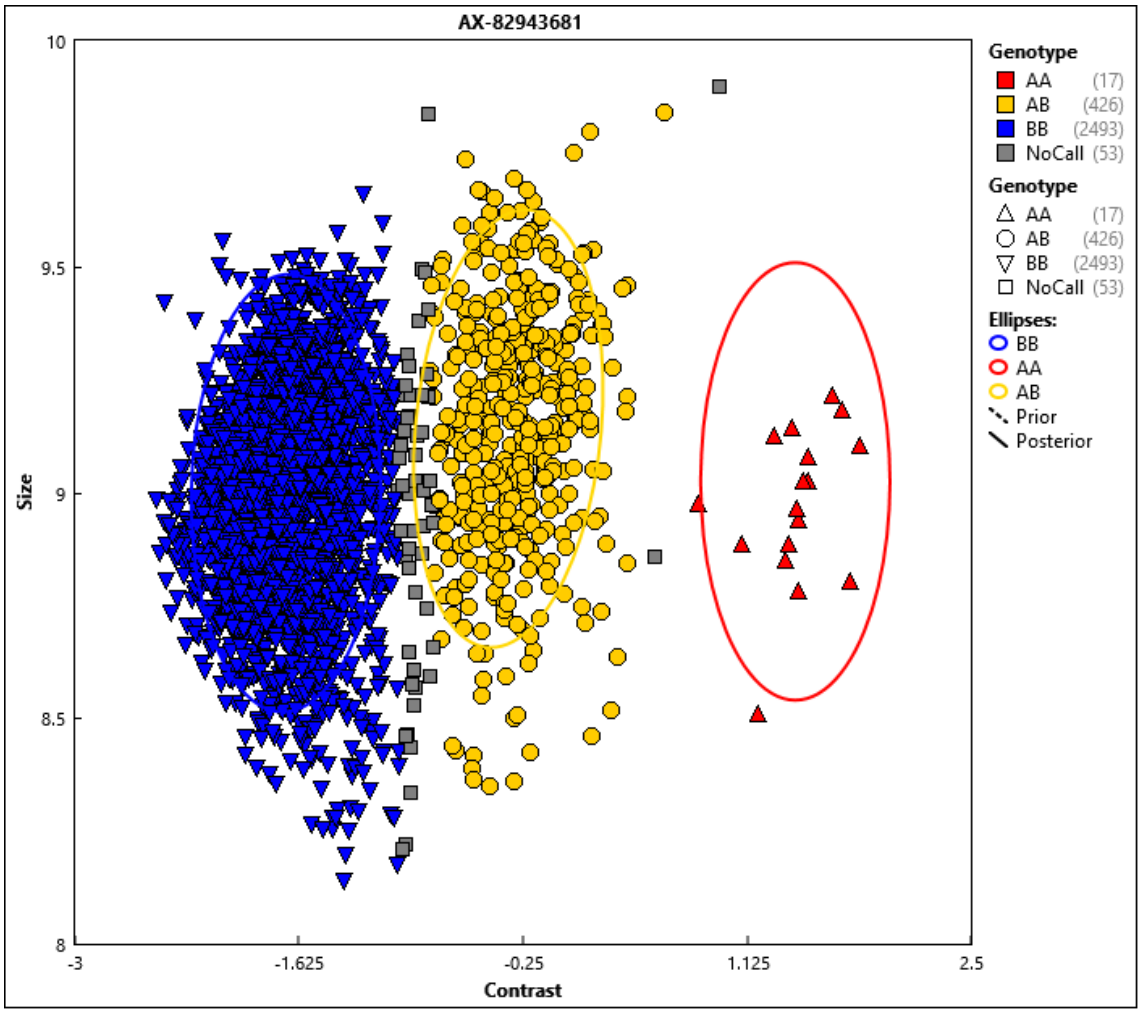


Figura 3.7. Clúster *plot* de SNPs producido por el *Axiom Analysis Suite* (Sesión de Axiom específica de este trabajo).

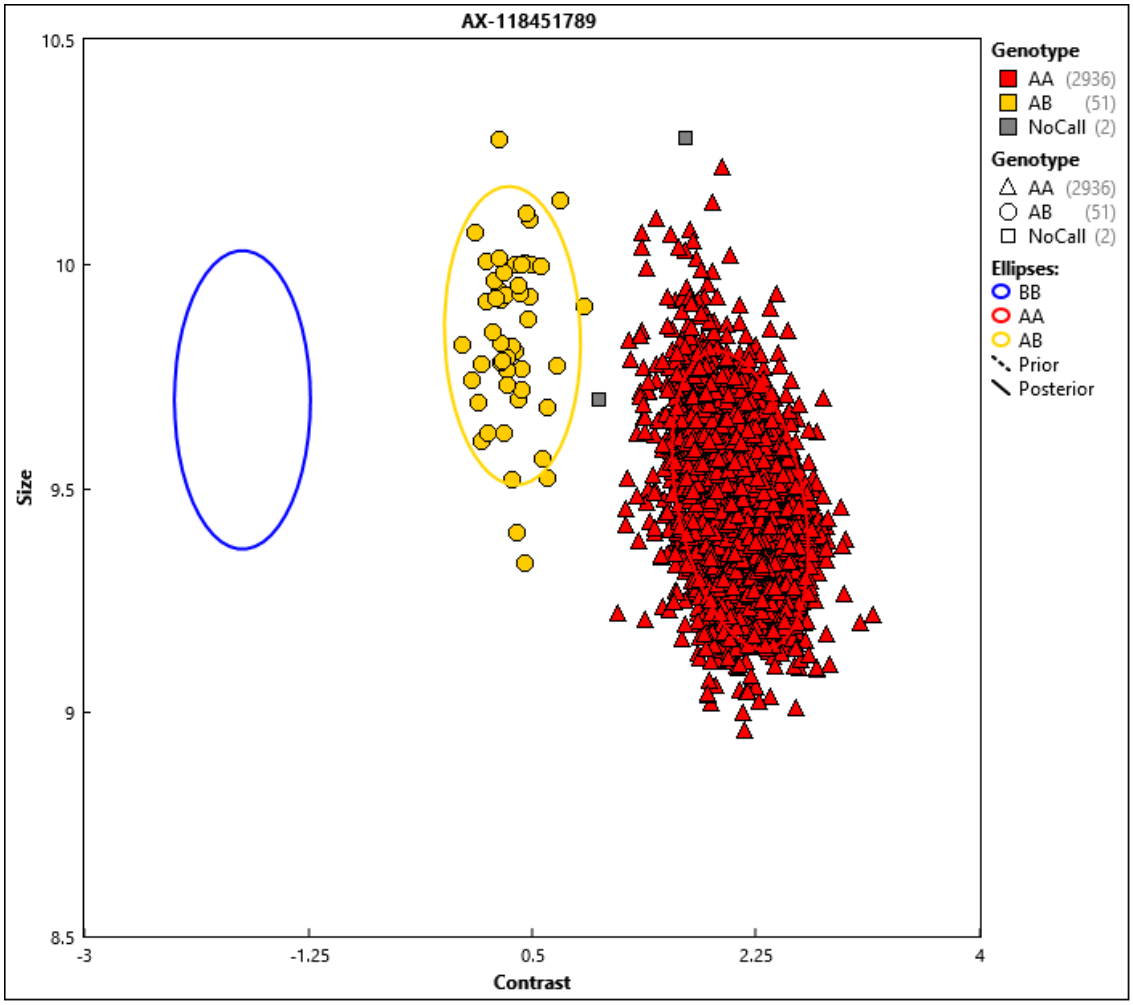


Figura 3.8. Ejemplo de un clúster de SNPs bien definido (Sesión de Axiom específica de este trabajo).

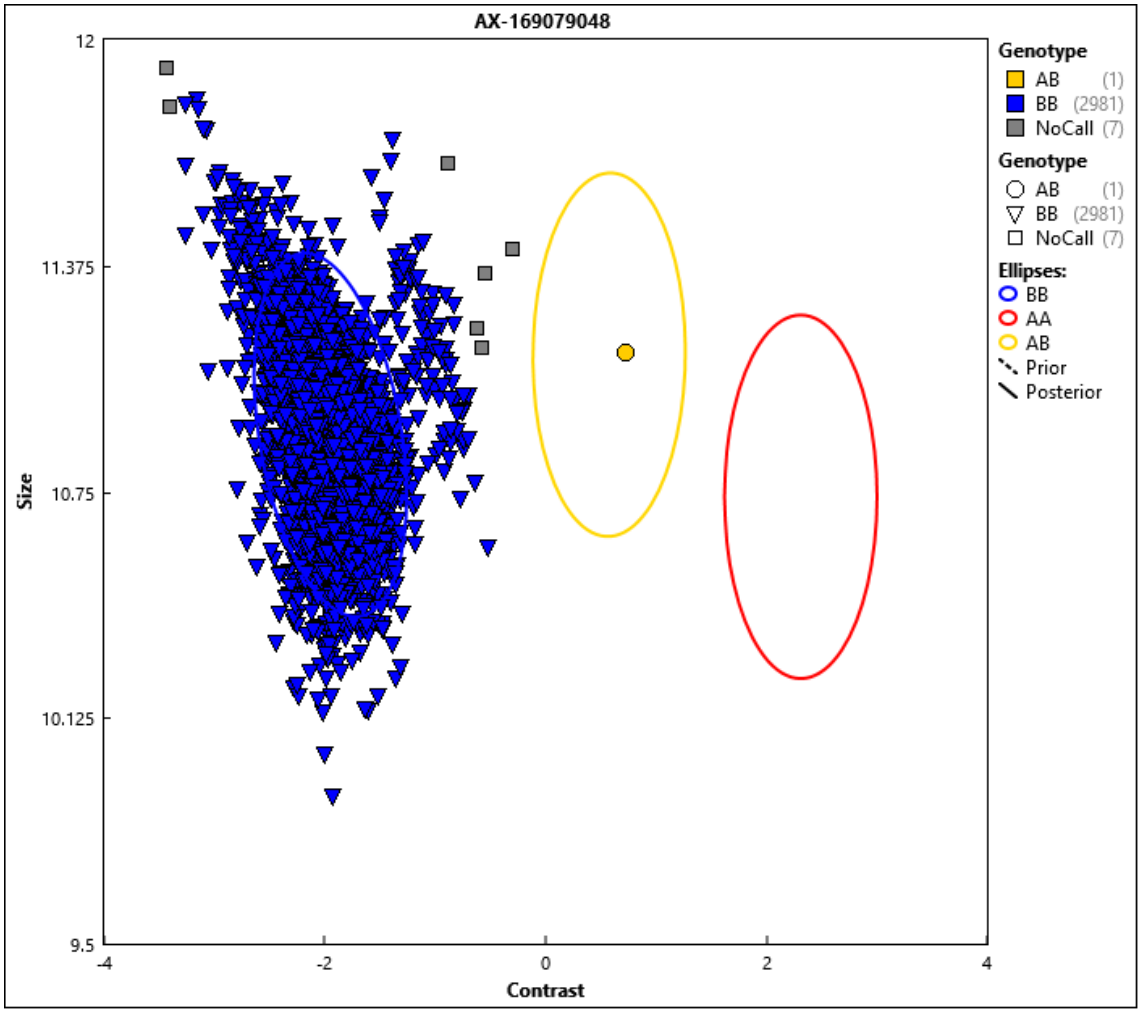


Figura 3.9. Ejemplo de un clúster de SNPs menos claro por no haber separación entre los grupos (Sesión de Axiom específica de este trabajo).

Tras la selección "manual" de estos casos revisados concretos se realizó la eliminación aleatoria de los duplicados idénticos (marcadores con el genotipado de ambas sondas de calidad con idéntica categorización en función de su frecuencia alélica) con el *software* SPSS (IBM, 2011).

Comparación de categorías de marcadores

Tras el control de calidad y eliminación de duplicados, se caracterizó la frecuencia relativa de cada categoría de marcador en el conjunto de las muestras del BNADN y su comparación con la categorización en población europea mediante tablas de frecuencias con el *software* SPSS (IBM, 2011)[4]. Los análisis se llevaron a cabo en autosomas y también incluyendo el cromosoma X. De este modo comparamos muestra población control española con las referencias establecidas en 1000G y podemos explorar si hay diferencias en cuanto a la variación rara y a los otros grupos de variantes categorizadas.

[4] Tablas de frecuencias detalladas en resultados.

Caracterización de la estructura poblacional

Normalmente "estratificación" alude a diferenciación genética de base entre casos y controles, pero en este trabajo utilizamos el PCA con el *dataset* completo para identificar *outliers* y caracterizar la posible estructura poblacional a escala geográfica local. Para ello se solicitó al BNADN la información acerca del origen geográfico de cada muestra, sus padres y sus abuelos, para poder seleccionar, posteriormente, el grupo de muestras con origen geográfico homogéneo con el que analizar detalladamente la estructura poblacional a escala local, uno de los objetivos principales del presente estudio.

La caracterización de la estructura poblacional a escala local se realizó usando un subgrupo de individuos con origen geográfico definido (el individuo y sus cuatro abuelos de la misma zona), asignándolos a diferentes regiones, intentando maximizar el número de zonas geográficas, estando estas siempre representadas por el máximo número de individuos.

Así y basándonos en el origen geográfico solicitado de las muestras incluidas en el estudio y de los parientes, hemos creado la variable "ORIGEN" teniendo en cuenta una serie de criterios. Nuestro objetivo era seleccionar muestras procedentes de una zona geográfica concreta, y por ese motivo se integró la información de la localidad de origen de padres y abuelos de cada muestra. En los casos en los que los 4 abuelos procedían del mismo lugar, la asignación era clara, pero, para intentar incluir más casos aprovechables se identificaron también aquellos casos en los que cada par de abuelos era de una misma zona (para valorar la posibilidad de reunirlas en una unidad geográfica más amplia)2 o casos incluso en los que, faltando información de los abuelos, los padres eran del mismo lugar. A partir de ahí, examinando las frecuencias de la variable "ORIGEN" (definida como la provincia/localidad común a los 4 o la combinación de dos pares de abuelos) se redefinieron zonas geográficas más amplias agrupando provincias cercanas para optimizar el número de observaciones en las diferentes regiones. Tras este análisis de frecuencias tomamos la decisión de trabajar con 12 puntos geográficos por proximidad, asignándoles un código numérico (Tabla 3.2).

Tabla 3.2. Origen geográfico de las muestras analizadas y su agrupación en puntos geográficos por proximidad.

PUNTOS GEOGRÁFICOS	N GLOBAL	"LOCALIDAD / PROVINCIA" N
1_ANDALUCIA	398	Almería = 12
		Cádiz = 33
		Ceuta = 7
		Córdoba = 109
		Granada = 39
		Huelva = 9
		Jaén = 52
		Málaga = 69
		Melilla = 7
		Sevilla = 61
2_ARAGÓN	64	Huesca = 19
		Teruel = 16
		Zaragoza = 29
3_CANTABRIA	47	Cantabria = 47
4_CASTILLA	169	Ávila = 22
		Burgos = 20
		Palencia = 14
		Segovia = 25
		Valladolid = 88
5_CATALUÑA	630	Barcelona = 465
		Cataluña = 2
		Gerona = 46
		Lleida = 67
		Tarragona = 50
6_LEVANTE	106	Comunidad Valenciana = 48
		Alicante = 36
		Castellón = 22
7_EXTREMADURA	102	Badajoz = 41
		Cáceres = 39
		Salamanca = 22
8_GALICIA	290	A Coruña = 25
		Galicia = 207
		Lugo = 18
		Ourense = 11
		Pontevedra = 29
9_LEÓN	430	Asturias = 97
		León = 86
		Salamanca = 192
		Zamora = 55
10_MANCHA_MADRID	452	Ciudad Real = 22
		Cuenca = 32
		Guadalajara = 11
		Madrid = 368
		Toledo = 19
11_PAÍSVASCO_NAVARRA	117	Álava = 16
		Gipuzkoa = 21
		Navarra = 27
		País Vasco = 2
		Pamplona = 1
		Vitoria = 25
		Vizcaya = 25
12_SORIA_LOGROÑO	43	La Rioja = 32
		Soria = 11

Análisis estructura poblacional -PCA

El análisis de componentes principales (PCA) se utilizó para llevar a cabo una primera caracterización de la variabilidad existente. Como se ha apuntado PCA es un tipo de análisis multivariante que permite que la información multidimensional se represente gráficamente con la pérdida mínima de información (Jobling MA, 2004). PCA transforma una serie de frecuencias alélicas correlacionadas en un menor número de variables no correlacionadas o componentes principales (PCs). Es útil cuando la mayor parte de la información provista por los datos puede resumirse en los primeros pocos PCs que representan la mayor fracción de la variación global.

Análisis discriminante de componentes principales -DAPC

Posteriormente llevamos a cabo DAPC *(*T. Jombart et al., 2010*)*, que es una técnica estadística multivariante que combina un análisis de componentes principales (PCA) y análisis discriminante (DA), maximizando la diferenciación entre los grupos, ya sean predefinidos o identificados por el programa en un análisis *cluster* (*K-means*) inicial (opción *find.clusters*). DAPC se lleva a cabo con el paquete "adegenet" para el *software* R (Thibaut Jombart, 2008). Este paquete fue diseñado como solución al problema que se presentaba al realizar el análisis multivariante directamente en los genotipos obtenidos de marcadores genéticos, ya que suponían una enorme cantidad de datos. Adegenet permite recodificar numéricamente en una matriz de frecuencias alélicas, que es almacenada en objetos de la clase *genind* o *genpop* dependiendo si se refiere a genotipos individuales o frecuencia de alelos por población respectivamente. (Thibaut Jombart, 2008).

DAPC tiene como finalidad identificar y describir agrupaciones (clústeres) de individuos genéticamente relacionados. Es una aproximación metodológica que integra las ventajas de los análisis de PCA y DA y que optimiza la varianza entre grupos y minimiza la varianza dentro de grupos con el fin de buscar variables sintéticas, las funciones discriminantes (DFs), que maximizan las diferencias entre grupos y minimizan las distancias dentro de los grupos (T. Jombart et al., 2010).

DAPC funciona en dos pasos: primero transforma los datos genéticos mediante PCA y después identifica los grupos o clústeres usando DA. En el primer paso, los datos son convertidos en variables no correlacionadas (componentes principales, PCs), que representan la mayor parte de la variación genética, y las ordena por importancia. Con esto se reduce la dimensionalidad del conjunto de datos e intuitivamente sirve para hallar las causas de la variabilidad de este. Estos componentes no relacionados son evaluados con DA, encontrando una combinación lineal de alelos (funciones discriminantes, DFs) que separe de la mejor manera los clústeres.

El método gráfico para DAPC es documentado en *scatter.plot*, que produce un diagrama de dispersión de los componentes principales (o funciones discriminantes), con una ventana de valores propios en su interior (Montano & Jombart, 2017) en donde se puede ver la distribución de los clústeres. La probabilidad de pertenencia de los individuos a cada clúster se puede estimar con *assign.prop*, lo que nos indica qué tan definidos o dispersos están los clústeres genéticos. Un clúster disperso puede estar apuntando hacia una posible mezcla. Esta información se puede ver gráficamente con *assign.plot*, que muestra cuánto coinciden los grupos originales (*a priori*) con los grupos inferidos por la función DAPC (T. Jombart, 2014). En definitiva el DAPC representa una valiosa herramienta para investigar patrones genéticos espaciales (Montano & Jombart, 2017).

RESULTADOS

4 RESULTADOS

4.1 VALIDACIÓN DEL *AXIOM SPAIN BIOBANK ARRAY PLATE*

4.1.1 Categorización de las variantes

La validación del *array* de genotipado para el análisis de variación funcional específica de población española que hemos diseñado, se realizó, como se ha descrito en el apartado 3, analizando un total de 3.059 muestras del Banco Nacional de ADN (BNADN), de las cuales 70 de ellas no cumplieron los criterios de calidad de ThermoFisher y fueron descartadas automáticamente. De las 814.923 variantes incluidas en nuestro *array* Axiom, 528.029 variantes se clasificaron como "marcadores *PolyHighResolution*" mostrando tres grupos con buena resolución y al menos dos ejemplos del alelo menor; 114.421 variantes como "marcadores *NoMinorHomozygous*" que exhiben dos grupos sin ejemplos del alelo menor, y 78.389 variantes se clasificaron como "marcadores *MonoHighResolution*", mostrando un solo grupo. Por último, se descartaron 29.729 + 3.906 + 3.266 variantes: 3.906 de ellas eran "*OffTargetVariant*", que denota marcadores reproducibles pero que aún no están caracterizados, causados por doble deleción o no homología de secuencia; 3.266 variantes "*CallRateBelowThreshold*" por no alcanzar el umbral de calidad y finalmente 29.729 de ellos por "otros" motivos, eliminadas por el *Axiom Analysis Suite* por ser sondas homocigotas no recomendadas. Tras estos descartes y eliminar los marcadores duplicados, finamente fueron evaluados 741.373.

Se exploró la distribución de las frecuencias alélicas del conjunto final de variantes seleccionadas en nuestra cohorte española, y los resultados se clasificaron en 4 categorías diferentes dependiendo de los valores de frecuencia de alelos menores, de la siguiente manera: a) si MAF=0 la variante se clasificó como "monomórfica" en nuestra población; b) las variantes con 0<MAF<0,01 se clasificaron como "raras (*rare*)" (K. L. Williams et al.); c) las variantes con 0,01<MAF<0,05 se identificaron como variantes de "baja frecuencia (*low frequency*)" y finalmente d) si MAF>0,05: entonces se utilizó el término "SNP", clasificación ya empleada en estudios previos como se apuntó en la metodología (Bomba et al., 2017), (Babron et al., 2012), (Ingles et al., 2018).

Siguiendo esta clasificación, en nuestro conjunto de 741.373 variantes, 63.464 resultaron monomórficas, 66.253 variantes raras, 208.082 variantes de baja frecuencia y 403.574 fueron clasificadas como SNP.

Se realizó la misma clasificación a partir de los datos de frecuencias descargados de 1000G, tanto en la súper-población europea como en la población IBS (*Iberian Population in Spain*). Los datos de frecuencia y categoría de marcador fueron entonces fusionados con los resultados en nuestra cohorte de forma que se pudo realizar la comparación de la categorización de marcadores para el total del *array*.

Tras explorar con SPSS (Shek & Ma, 2011) (IBM, 2011) las variantes incluidas en el *array* en cuanto a su cambio de categoría con respecto a sus frecuencias en la base de datos 1000G, se pudo comprobar que existen más monomórficas y raras en España y prácticamente el mismo número de SNPs y variantes de baja frecuencia en ambas poblaciones (íbera y europea).

Se aprecia también que hay 57.433 variantes específicas de población española que no se encuentran en Europa en la base de datos de referencia 1000G (Tablas 4.1 y 4.2) y que corresponden con un 8% no incluidas en 1000G.

Tabla 4.1. Comparación de la clasificación en categorías (cat) de las variantes en población española (CATEGORÍA_BNADN) respecto a población europea de 1000G (CATEGORÍA_1000G EUR).

		CATEGORÍA_1000G EUR				TOTAL
CATEGORÍA_BNADN		Monomórfica	Raras	Baja frecuencia	SNP	
Monomórfica	N	6.085	6.791	628	512	14.016
	% dentro cat BNADN	43,40%	48,50%	4,50%	3,70%	100%
	% dentro cat 1000G-EUR	49,90%	16%	0,30%	0,10%	2,10%
Raras	N	6.095	29.674	22.172	288	58.199
	% dentro cat BNADN	10,40%	51%	38,10%	0,50%	100%
	% dentro cat 1000G-EUR	49,70%	70,10%	10,10%	0,10%	8,90%
Baja frecuencia	N	19	5.835	179.111	14.057	199.022
	% dentro cat BNADN	0,00%	2,90%	90%	7,10%	100%
	% dentro cat 1000G-EUR	0,20%	13,80%	85%	3,70%	30,50%
SNP	N	24	22	16.599	364.212	380.857
	% dentro cat BNADN	0,00%	0,00%	4,40%	95,60%	100%
	% dentro cat 1000G-EUR	0,20%	0,10%	7,60%	96,10%	58,40%
Total	N	12.193	42.322	218.510	379.069	652.094
	% dentro cat BNADN	1,90%	6,50%	33,50%	58,10%	100%
	% dentro cat 1000G-EUR	100%	100%	100%	100%	

Tabla 4.2. Cambio de categoría de las variantes en población española respecto a población europea según 1000G.

CATEGORÍA – BNADN	presente en 1000G	No presente en1000G
Monomórfica	14.016 (2,1%)	42.025 (73,2%)
Raras	58.199 (8,9%)	12.597 (21,9%)
Baja frecuencia	199.022 (30,5%)	761 (1,3%)
SNP	380.857 (58,4%)	2.050 (3,6%)

En la Tabla 4.1 llaman especialmente la atención los cambios de categorías más extremos (de SNP a variante rara o monomórfica). Destaca el hecho del relativamente alto número de marcadores tipo SNP en 1000G que presentan frecuencias mucho más bajas en la cohorte española. El detalle de la lista de tales marcadores se puede ver en la tabla de resultados en el Anexo I, en el que se puede observar como muchos marcadores que resultan ser comunes en Europa cambian a categorías raro y monomórfico en España: 30 variantes ADME, 42 relacionadas con nefrología, 152 vinculadas a psiquiatría e inmunología, 36 relacionadas con radiogenómica y 286 del UK Biobank *array* previo.

Tal y como se puede ver en la siguiente tabla (Tabla 4.3), el *Spain Biobank Array Plate* muestra diferencias en los marcadores añadidos específicos de población española y los contenidos previamente en el *array* (comunes al UKBB). En esta tabla, que se muestra a continuación, se resume el recuento de marcadores genotipados en nuestra población control española con una muestra total final de 2.989 individuos no emparentados.

Tabla 4.3. Distribución en categorías de las variantes genotipadas específicas de población española respecto al contenido previo común en el *UK Biobank Array*.

	monomórficas	raras	Baja frecuencia	SNPs
Marcadores seleccionados población española	40.547 (63%)	26.908 (41%)	7.489 (4%)	31.828 (8%)
Resto *array*	23.841 (37%)	39.263 (59%)	197.540 (96%)	364.336 (92%)
TOTAL	64.388	66.171	205.029	396.164

Hay que aclarar que los resultados de la comparación cuando se hizo con 1000G-IBS son muy semejantes a los mostrados con población europea, por lo que no se especifica nada más al respecto considerando que las deducciones son las mismas.

De los marcadores seleccionados en población española, unos 48.000 correspondían a variantes raras identificadas en exomas de población control. Estas variantes raras han sido encontradas como tales en unos 19.000 casos (serían la mayoría de las 26.908 indicadas en la Tabla 4.3), mientras que unas 28.000 entran en la categoría de los monomórficos (de ahí su alta frecuencia relativa mostrada en la Tabla 4.1). Hay que tener en cuenta que, debido a su baja frecuencia, la capacidad de detección de estas variantes es muy dependiente de la N, por lo que en diferentes estudios pueden encontrarse variantes raras ahora incluidas en monomórficas.

Por otra parte, dentro del grupo de los marcadores previos del *array* (comunes al UK Biobank), las variantes categorizadas como raras en esta muestra de población control española corresponden, en su mayoría (unas 31.000 de 39.000), a marcadores GWAS de baja frecuencia, apuntando la posibilidad de diferencias locales en frecuencia alélica.

Tras cuantificar el porcentaje de marcadores que cambiaron de categoría en población española respecto a las poblaciones tenidas en cuenta como referencia (ibérica y europea de la base de datos 1000G), en el apartado siguiente se muestra el análisis de las variantes que contribuyen a la discriminación poblacional en nuestra muestra. Tras esta evaluación se podría depurar este *array* generando una versión 2 del mismo, trabajo que ya estamos llevando a cabo, eliminando sondas redundantes e incluyendo otras de interés.

4.1.2 Caracterización de la estructura poblacional a escala local

Investigamos por tanto la estructura de la población a nivel local en nuestra población ibérica, finalmente en una N de 1.859 individuos que catalogamos con origen geográfico definido (el individuo y sus cuatro abuelos de la misma zona, ya que buscábamos un grupo de muestras con origen geográfico homogéneo), a través del análisis con nuestro *array* específico de población española. El objetivo principal es ver si existen diferentes patrones de estratificación que dependan de la frecuencia del alelo menor, siendo particularmente interesantes las variantes raras. Para lograr eso realizamos PCA y luego ejecutamos DAPC.

El análisis PCA, realizado con un *subset* de SNPs independientes, es la aproximación estándar para evaluar la existencia de subestructura. En nuestro caso se evaluó con un grupo de 72.323 SNPs independientes; al realizar el *plot* PC1-PC2 (Figura 4.1) diferenciando el origen poblacional de los individuos, se observa una nube de puntos en la que se aprecia un gradiente este-oeste, de Galicia a País Vasco.

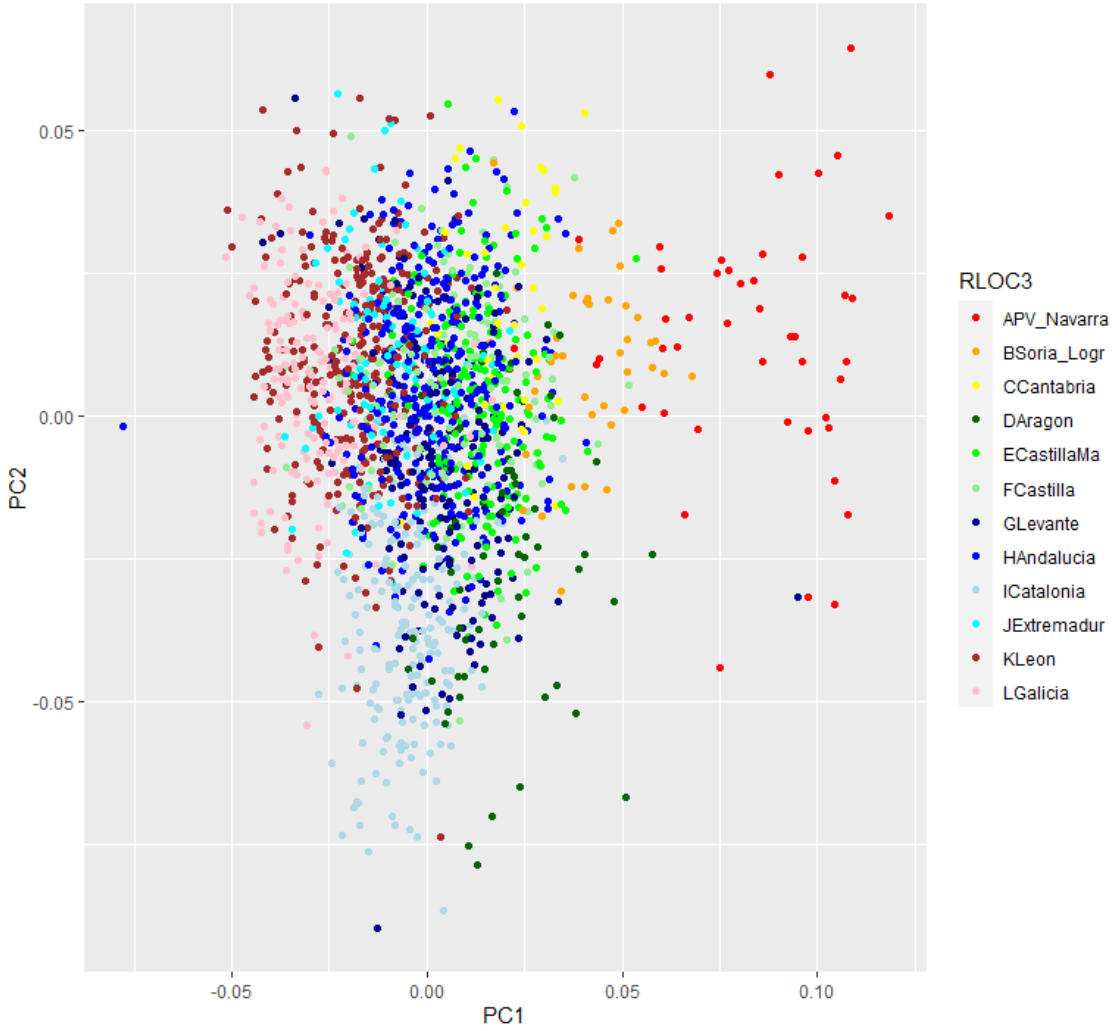


Figura 4.1. Análisis PCA inicial para evaluar la subestructura poblacional (PC1 y PC2).

Para analizar en detalle la diferenciación entre las subpoblaciones definidas, solo intuida en el PCA, se realizó el análisis DAPC que, tal y como se explicó en la metodología, busca maximizar la discriminación entre los grupos definidos.

Como también se ha descrito en la metodología de este trabajo, el análisis se llevó a cabo con la librería adegenet para realizar DAPC a partir de los genotipos para un subgrupo de 308.547 marcadores independientes.

DAPC se realizó con el total de marcadores independientes y diferenciando 3 categorías de marcadores, para comparar la capacidad discriminatoria en función de la MAF: utilizando solo SNPs, marcadores de baja frecuencia y variantes raras, utilizando en todos los casos la definición de clústeres *a priori* en función de la variable población de origen, tal y como se definió en Material y Métodos.

Mediante el Análisis Discriminante de los Componentes Principales hemos podido ver cómo discriminan los diferentes tipos de marcadores (SNPs, variantes de baja frecuencia y raras) a nuestra población de estudio, de forma que hemos podido comparar el nivel de discriminación alcanzado. Este análisis se realizó adicionalmente sin los cromosomas 6 (en el que se encuentra el Complejo Mayor de Histocompatibilidad, HLA) y 8 con una inversión común en población española descrita por Baran et al. (Baran et al., 2013)) para comprobar su posible influencia en la discriminación, pero los resultados obtenidos fueron similares cuando

estos se incluyeron en el análisis. Las Figuras 4.2, 4.3 y 4.4 muestran el *scatter plot* para las dos primeras funciones discriminantes realizadas con SNPs, *low* y *rare* respectivamente; estas dos primeras funciones son las más discriminantes, especialmente LD1, como se aprecia en el gráfico de barras embebido en la esquina inferior derecha de cada *scatter plot*. En el Anexo II pueden verse los *plots* correspondientes a otras combinaciones de las 4 funciones discriminantes testadas. Tras los filtrados realizados el análisis basado en SNPs se llevó a cabo con un total de 135.577 variantes; en el caso de los marcadores de baja frecuencia se tuvieron en cuenta 114.581 y 58.492 para el análisis correspondiente a las variantes raras.

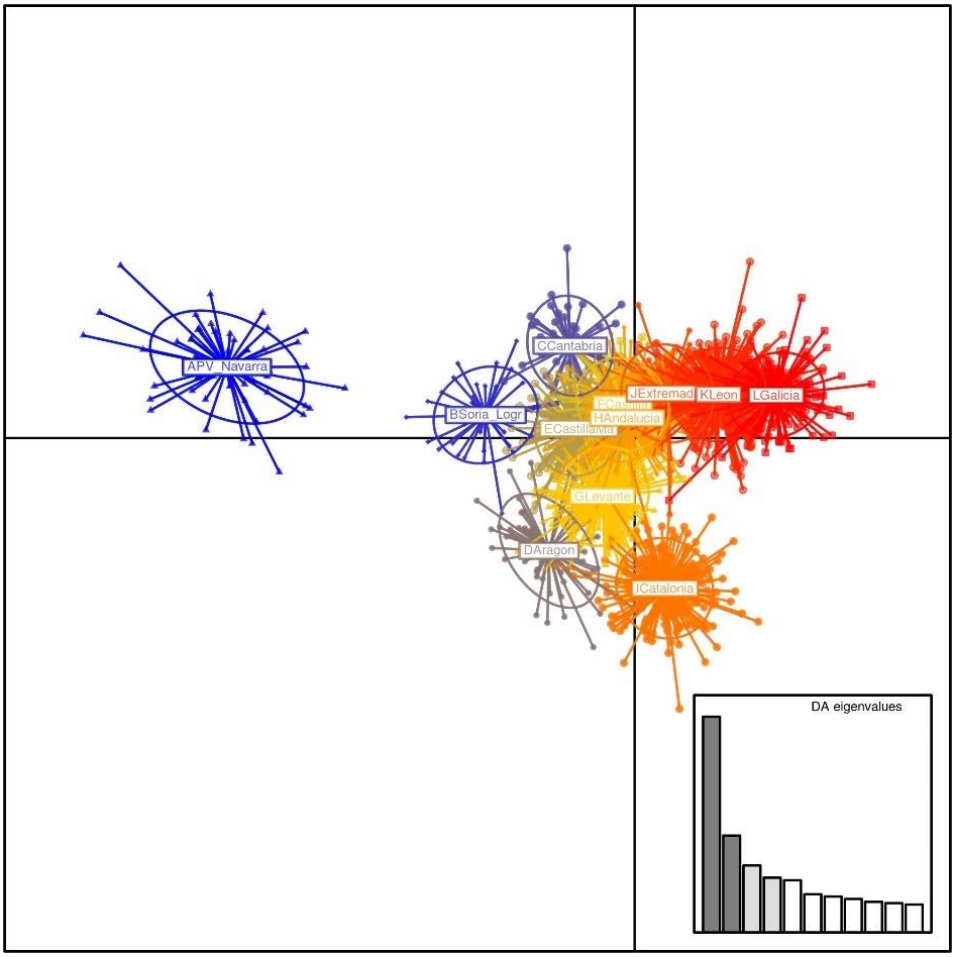


Figura 4.2. Caracterización de la estructura de nuestra población de estudio en función de SNPs (135.577) (LD1 y LD2).

La Figura 4.2 muestra la distribución de los individuos en 12 clústeres en relación con las dos funciones discriminantes utilizadas, donde se observa el clúster "País Vasco_Navarra" claramente diferenciado de los otros 11, los que quedan más alineados a lo largo de LD1, quedando Galicia en el extremo opuesto, apuntando así al gradiente este-oeste ya documentado (Figura 4.1).

En la Figura A5 (Anexo II), también basada en SNPs, se puede apreciar, además, una diferenciación de las poblaciones cántabra y gallega.

En función de las variantes de baja frecuencia - *low* (Figura 4.3), podemos ver diferenciadas estas tres poblaciones (País Vasco_Navarra, Cantabria y Galicia) a lo largo de LD1, pero también se aprecia una diferenciación de "Cataluña" en LD2, que es observada además en otros

plots que implican la segunda función discriminante: LD2 / LD3 (Figura A6, Anexo II) y LD2 / LD4 (Figura A7, Anexo II).

En el Anexo II además se muestran los *plots* correspondientes a otras combinaciones de las 4 funciones discriminantes testadas.

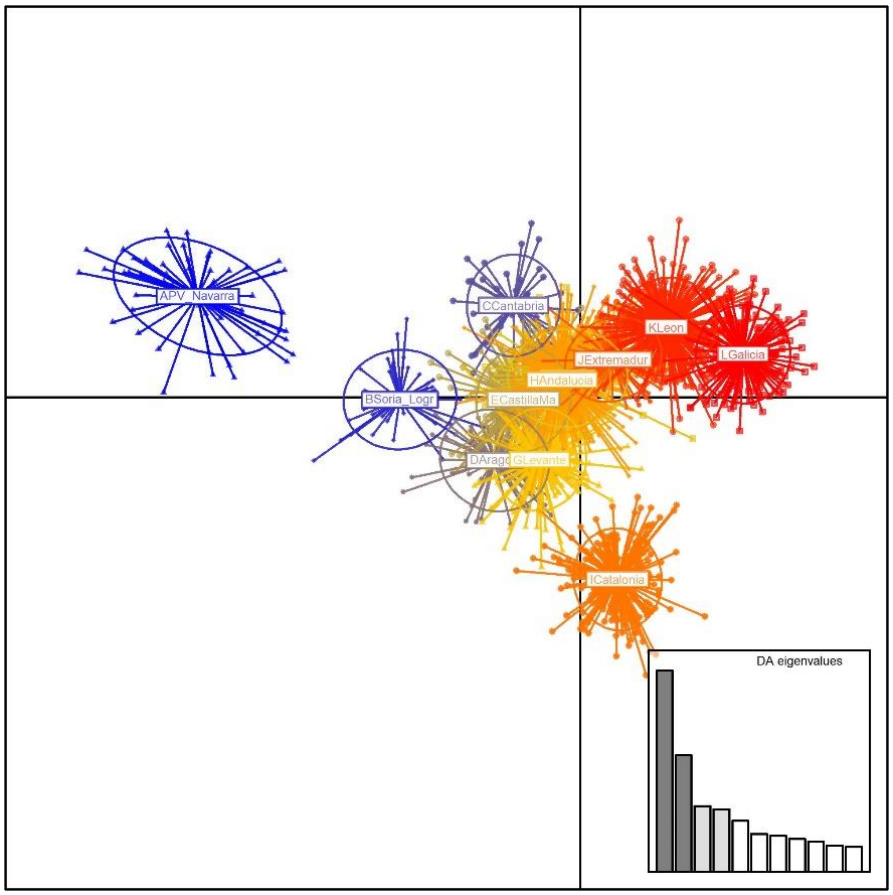


Figura 4.3. Caracterización de la estructura de nuestra población de estudio en función de variantes de baja frecuencia (*low*) (114.581); (LD1 y LD2).

Cuando el análisis se lleva a cabo con las variantes raras, continúan diferenciándose aún más "Galicia", "Cantabria", "País Vasco_Navarra" y "Cataluña", sobre todo al tener en cuenta LD1 y LD2 (Figura 4.4), LD2 y LD3 (Figura 4.5) y LD2 y LD4 (Figura 4.6).

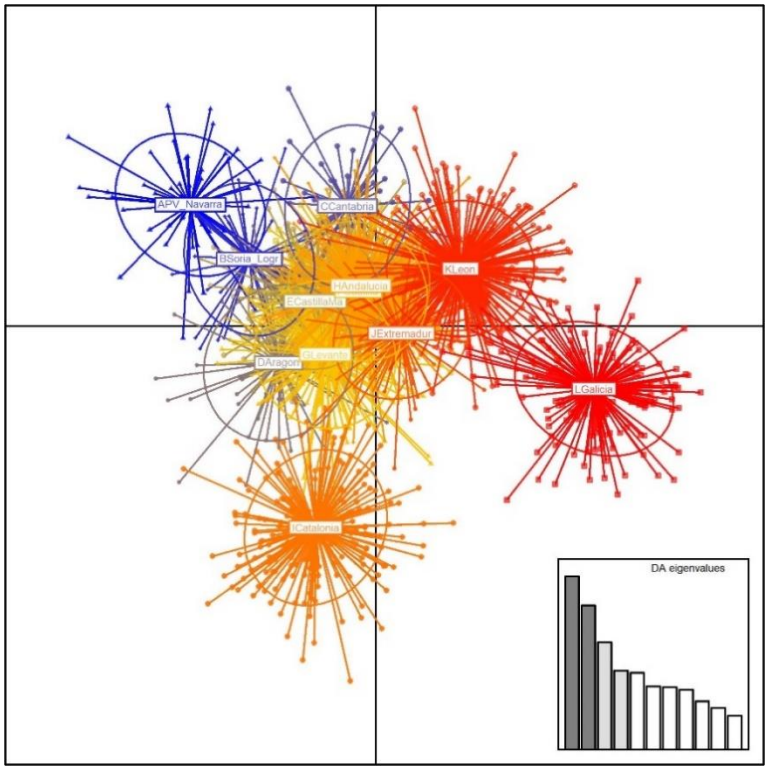


Figura 4.4. Caracterización de la estructura de nuestra población de estudio en función de variantes raras (*rare;* 58.492); (LD1 y LD2).

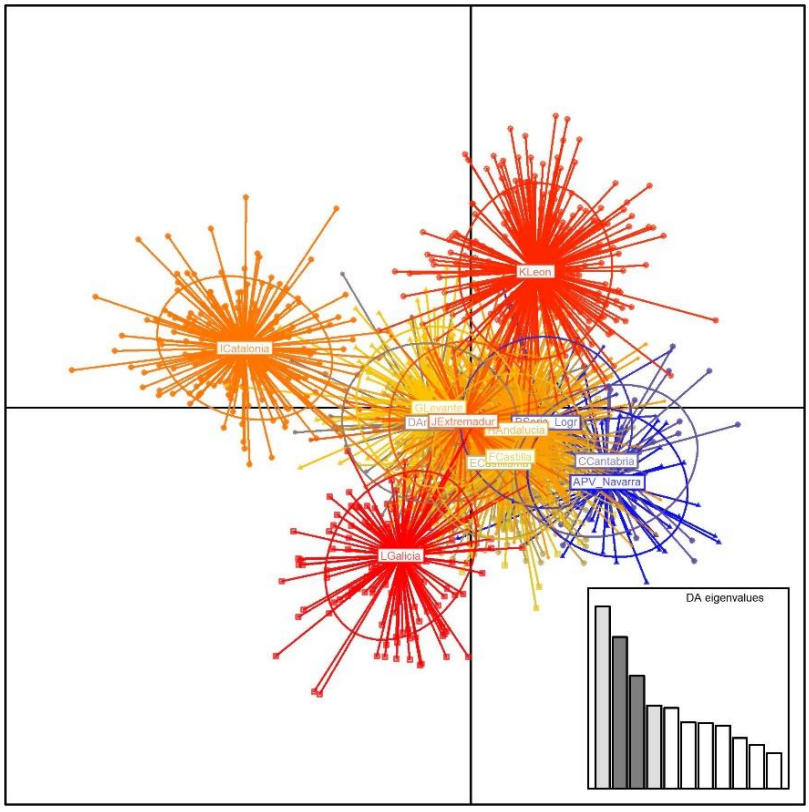


Figura 4.5. Caracterización de la estructura de nuestra población de estudio en función de variantes raras (*rare*; 58.492); (LD2 y LD3).

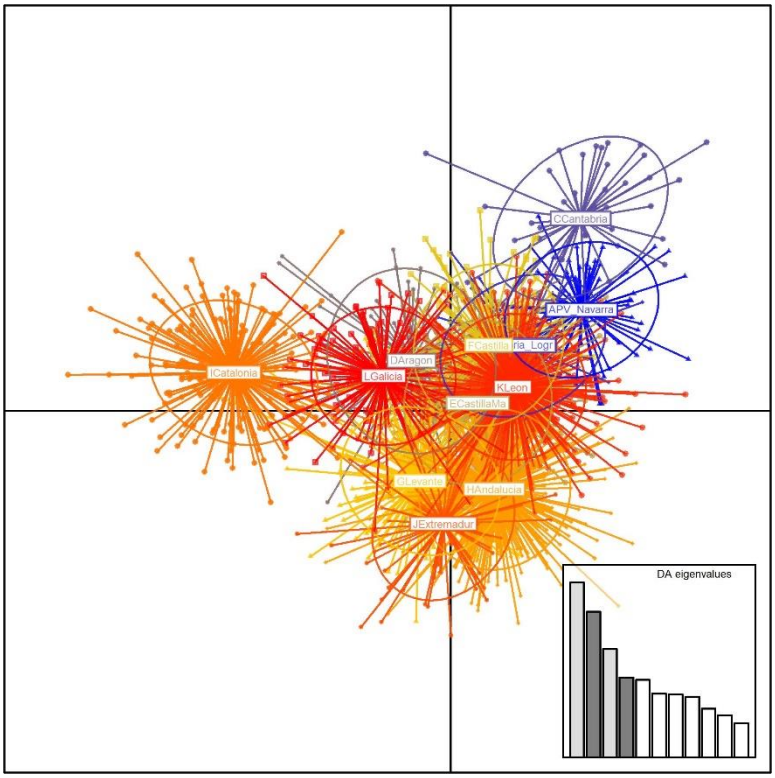


Figura 4.6. Caracterización de la estructura de nuestra población de estudio en función de variantes raras (*rare*; 58.492); (LD2 y LD4).

Estos resultados sugieren que un aspecto interesante de este análisis consiste en poder confirmar que algunas de las subpoblaciones se pueden agrupar en unidades mayores, teniendo en cuenta su pertenencia al mismo clúster, pero considerando también su localización geográfica.

4.1.3 Marcadores que más contribuyen a la discriminación de los diferentes patrones de estratificación poblacional a nivel del conjunto de población española y a escala microgeográfica

El análisis PCA muestra un gradiente este-oeste (de País Vasco a Galicia) a lo largo de la cornisa cantábrica, ya documentado.

DAPC confirma la clara diferenciación de País Vasco_Navarra, quedando Galicia en el otro extremo y diferencia también Cantabria y Cataluña. El patrón de discriminación mostrado es muy coherente tanto usando SNPs como variantes de menor frecuencia, si bien, en ese caso, la diferenciación de alguna subpoblación se acentúa.

Para evaluar la existencia de alguna región cromosómica que sea particularmente responsable de la diferenciación observada, se extrajo en cada DAPC la contribución de cada marcador a la discriminación en LD1 y LD2 (opción *var.contr*). En cada análisis (DAPC con SNPs, *low* o *rare*) se clasificaron los marcadores en función de su carga o *loading*, determinando como más discriminantes aquellos cuya carga superaba el tercer cuartil de la distribución. A continuación, se definieron ventanas deslizantes de 500.000 pares de bases (pb) antes y después de cada marcador y en cada ventana se analizó, mediante test de Fisher, si la frecuencia relativa de marcadores discriminantes era significativamente mayor a la global.

Los resultados, a nivel de toda la población española analizada, se pueden ver en las Figuras mostradas a continuación, donde, de modo similar a un *Manhattan Plot* (MP) se muestran las

probabilidades, en este caso del test de Fisher, de las ventanas deslizantes ordenadas por su posición cromosómica:

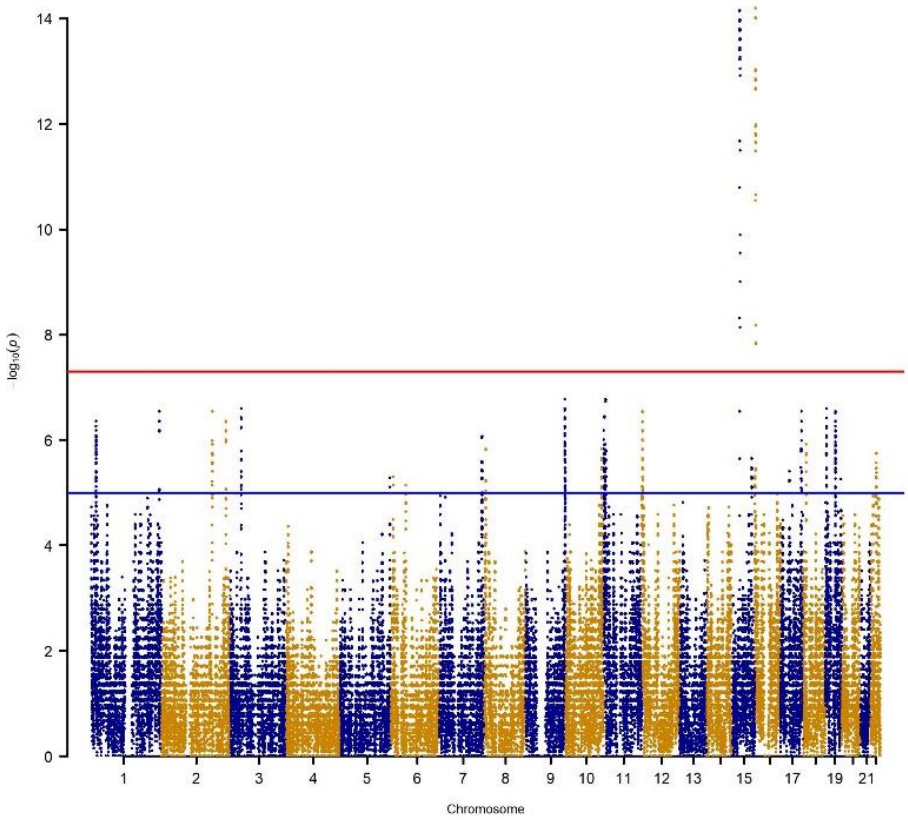


Figura 4.7.a. Test de Fisher: ventanas deslizantes_LD1_SNPs. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de SNPs; (LD1).

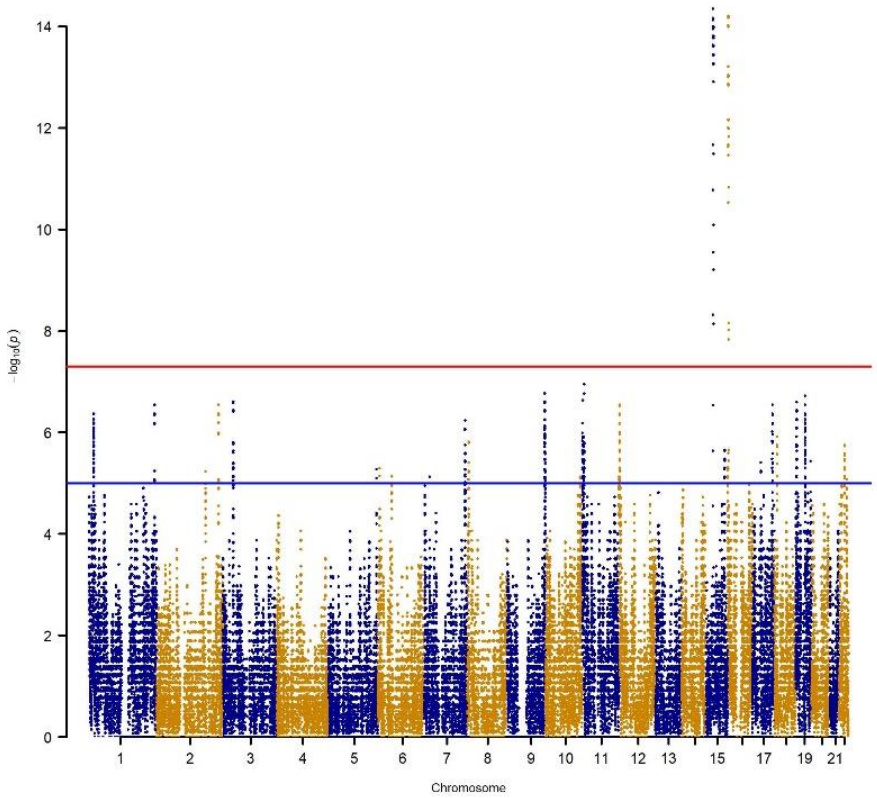


Figura 4.7.b. Test de Fisher: ventanas deslizantes_LD2_SNPs. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de SNPs; (LD2).

Para analizar en detalle qué variantes discriminan específicamente las subpoblaciones definidas en la población española, se estableció el corte en $p<5*10^{-8}$(umbral de significación GWAS).

Bajo esta premisa, en la Figura 4.7.a se puede apreciar una región bien diferenciada tanto en el cromosoma 15 (que se corresponde con 302 marcadores) como en el 16 (342 marcadores). En la Figura 4.7.b, en la que se tiene en cuenta LD2, se distinguen las mismas regiones correspondientes prácticamente a los mismos marcadores.

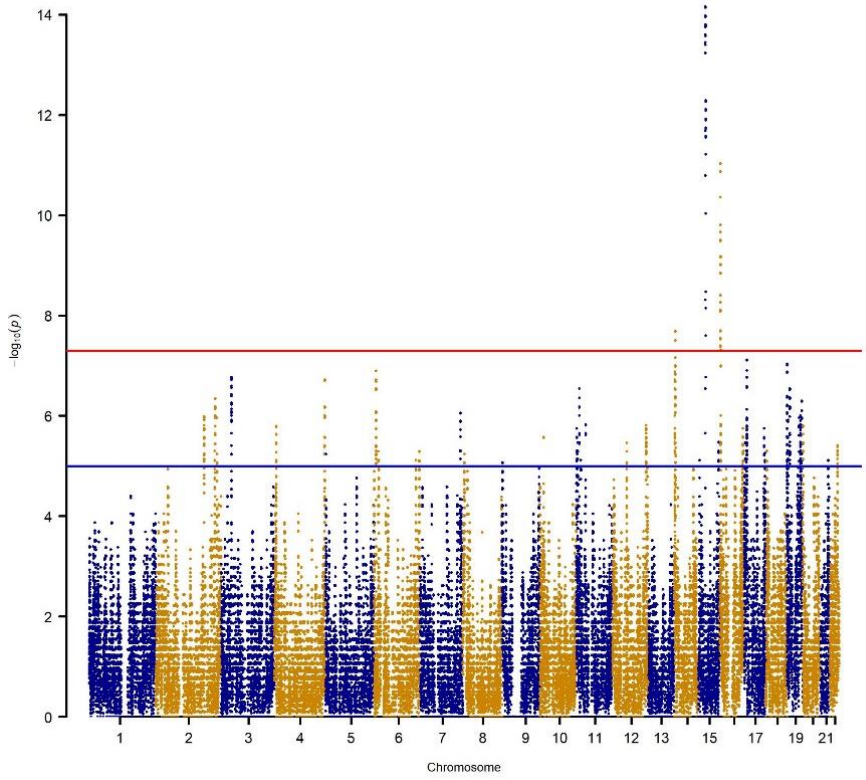


Figura 4.8.a. Test de Fisher: ventanas deslizantes_LD1_variantes de baja frecuencia. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de variantes de baja frecuencia (*low*); (LD1).

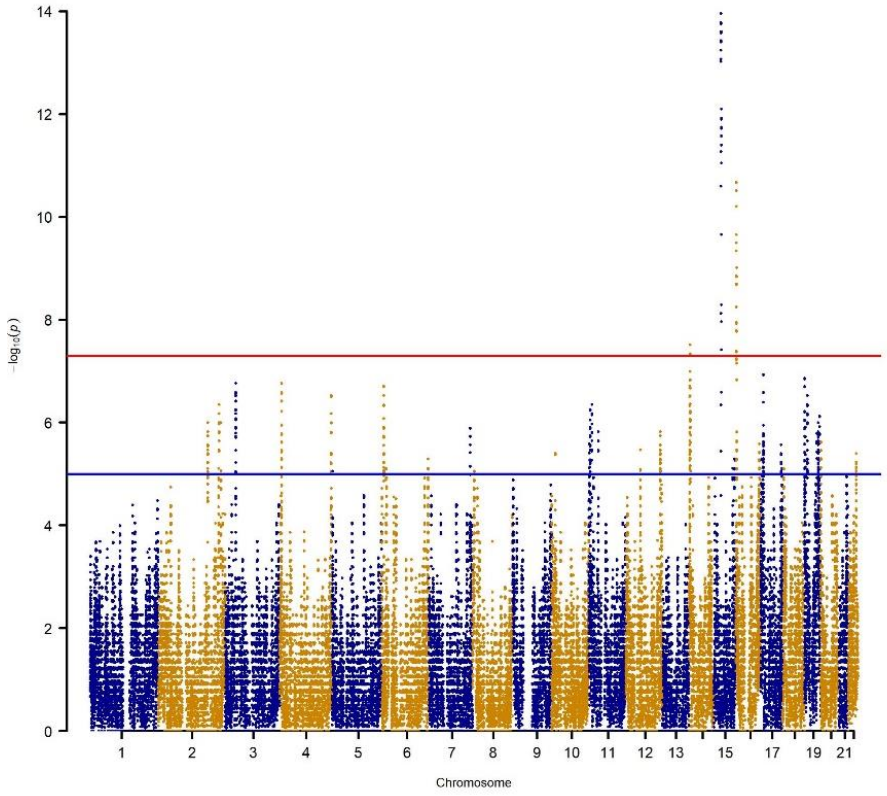


Figura 4.8.b. Test de Fisher: ventanas deslizantes_LD2_variantes de baja frecuencia. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de variantes de baja frecuencia (*low*); (LD2).

Haciendo el análisis con las variantes de baja frecuencia se aprecia el mismo patrón de discriminación que con SNPs, pero se aprecia una señal menor en el caso de marcadores *low* en el cromosoma 16.

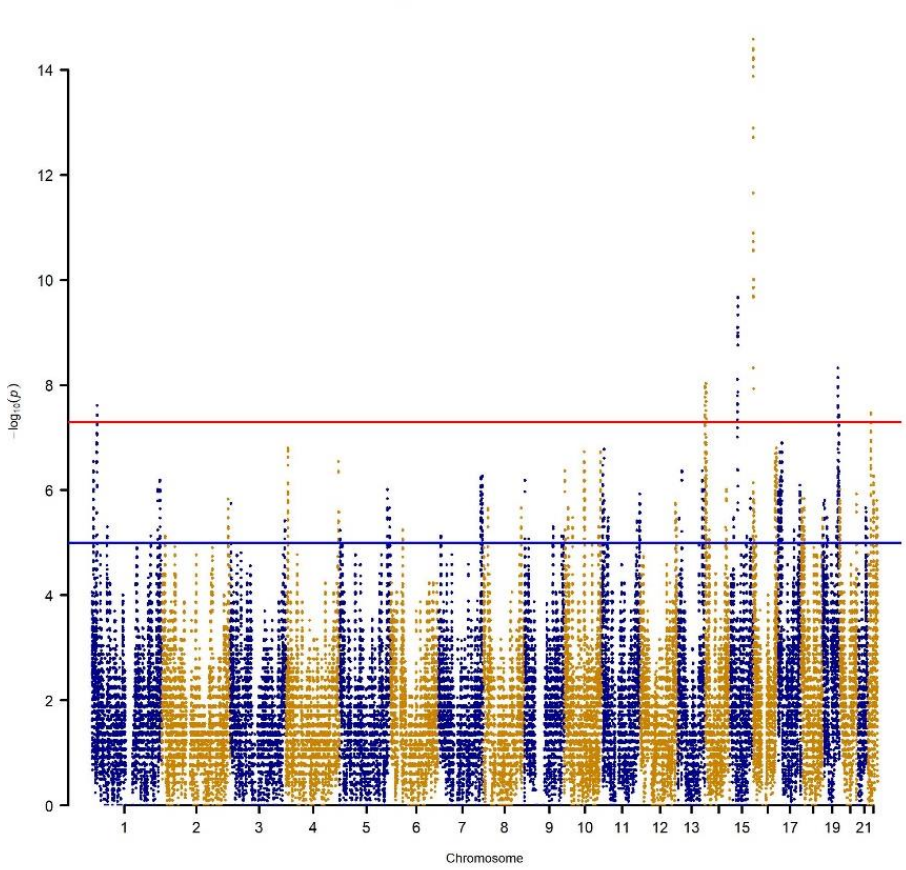


Figura 4.9.a. Test de Fisher: ventanas deslizantes_LD1_variantes raras. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de variantes raras (*rare*); (LD1).

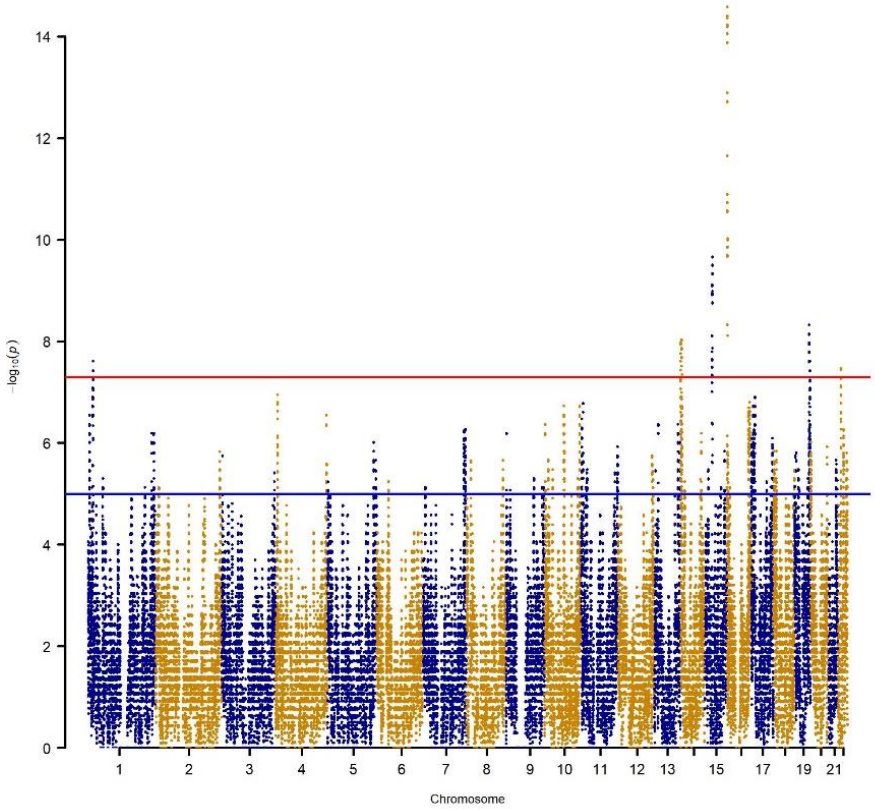


Figura 4.9.b. Test de Fisher: ventanas deslizantes_LD2_variantes raras. Distribución de marcadores responsables de la discriminación en nuestra población total de estudio en función de variantes raras (*rare*); (LD2).

Cuando tenemos en cuenta las variantes raras destaca sobre todo una región de 39 variantes en el cromosoma 16. En el cromosoma 15 volvemos a encontrarnos señal, pero además se aprecia también en el cromosoma 14, lo que no se había visto en los análisis llevados a cabo anteriormente con los otros tipos de marcadores.

El análisis de ventanas deslizantes nos permitió encontrar zonas cromosómicas en las que se concentraron marcadores que contribuyeron a la discriminación a lo largo de ambas LDs, pero la frecuencia de dichos marcadores puede mostrar diferencias graduales entre las poblaciones que se distribuyen a lo largo de estos ejes (Tabla 4.4).

Tabla 4.4. Distribución de marcadores destacados responsables de la discriminación en nuestra población total de estudio en función de SNPs, variantes de baja frecuencia (*low*) y variantes raras (*rare*) y patologías asociadas.

CHR	GEN	PATOLOGÍA ASOCIADA
15	*FBN1*	SÍNDROME DE MARFAN
15	*SLC12A1*	ENFERMEDAD RENAL
15	*SLC24A5*	ENFERMEDAD RENAL
16	*ABCA17P*	DISCAPACIDAD INTELECTUAL
16	*ABCA3*	DISCAPACIDAD INTELECTUAL
16	*AMDHD2*	DISCAPACIDAD INTELECTUAL
16	*DNASE1L2*	DISCAPACIDAD INTELECTUAL
16	*E4F1*	DISCAPACIDAD INTELECTUAL
16	*MLST8*	DISCAPACIDAD INTELECTUAL
16	*NTN3*	DISCAPACIDAD INTELECTUAL
16	*PGP*	DISCAPACIDAD INTELECTUAL
16	*PKD1*	DISCAPACIDAD INTELECTUAL
16	*RAB26*	DISCAPACIDAD INTELECTUAL
16	*RNPS1*	DISCAPACIDAD INTELECTUAL
16	*SLC9A3R2*	DISCAPACIDAD INTELECTUAL
16	*SYNGR3*	DISCAPACIDAD INTELECTUAL
16	*TBC1D24*	DISCAPACIDAD INTELECTUAL
16	*TRAF7*	DISCAPACIDAD INTELECTUAL
16	*TSC2*	DISCAPACIDAD INTELECTUAL
16	*ZNF598*	DISCAPACIDAD INTELECTUAL
16	*NOXO1*	DISCAPACIDAD INTELECTUAL Y CÁNCER
16	*NTHL1*	DISCAPACIDAD INTELECTUAL Y CÁNCER
16	*BRICD5*	DISCAPACIDAD INTELECTUAL Y SORDERA
16	*CASKIN1*	SÍNDROME X FRÁGIL
16	*NHERF2*	FIBROSIS QUÍSTICA
16	*TEDC2*	CÁNCER
16	*TBL3*	AZOOSPERMIA Y LINFOMA CEREBRAL
16	*FAHD1*	AZOOSPERMIA Y LINFOMA CEREBRAL
16	*MAPK8IP3*	ENFERMEDADES NEURODEGENERATIVAS
16	*HAGH*	ENFERMEDADES NEURODEGENERATIVAS
16	*EME2*	SÍNDROME DE LEIGH
16	*CCNF*	ESCLEROSIS LATERAL AMIOTRÓFICA

Las señales observadas son las mismas tanto en LD1 como en LD2; sin embargo, no ocurre lo mismo con las poblaciones diferenciadas, ya que, como se ha visto, las poblaciones se diferencian de distinta manera en función de la LD testada.
Por este motivo, para analizar si existen genes o zonas cromosómicas en las que una de las poblaciones definidas se diferencia significativamente del resto de poblaciones españolas, se realizó un análisis de asociación tipo GWAS para cada una de las comparaciones (población "x" vs resto población española). Los resultados se muestran en forma de MP como en cualquier GWAS, indicando las señales altas, en este caso una zona o gen en el que se aprecian diferencias significativas en frecuencia entre la población comparada y el resto de poblaciones españolas.
Para determinar las regiones de interés, como se ha mencionado, se estableció el corte en $p<5*10^{-8}$. A continuación, se muestran estas comparaciones entre poblaciones:

Andalucía *versus* Resto población española:

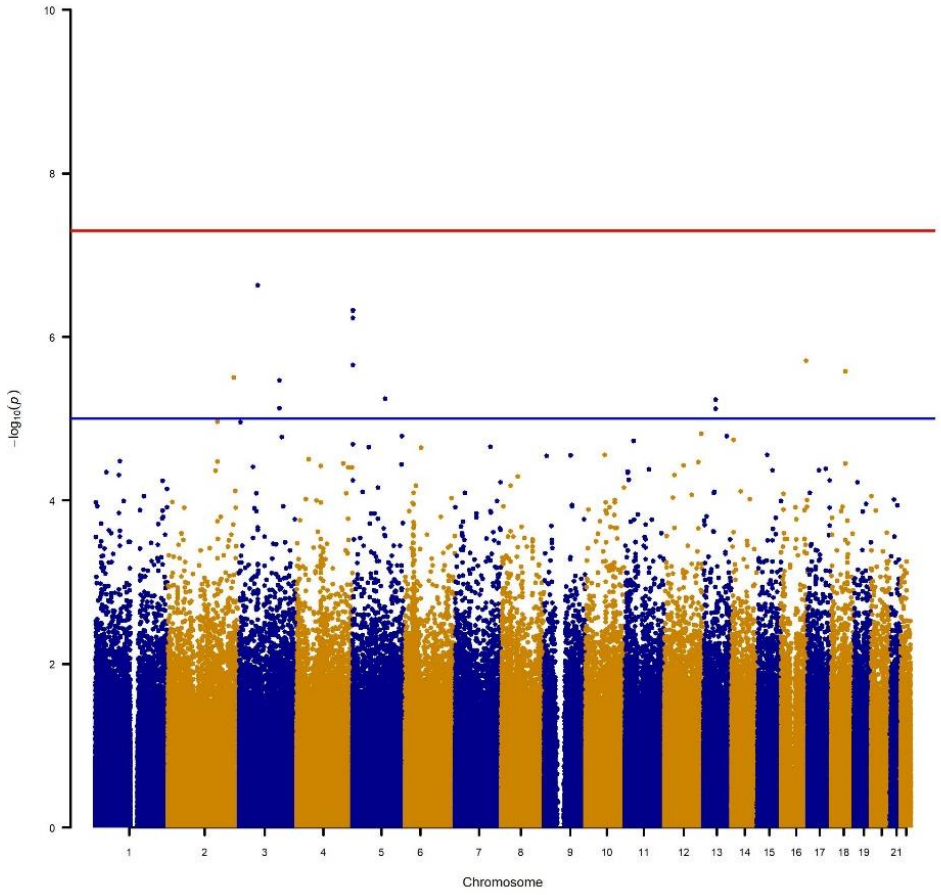


Figura 4.10. Distribución de marcadores en función de su probabilidad de asociación con la población andaluza comparada con el resto de población española.

Aragón *versus* Resto población española:

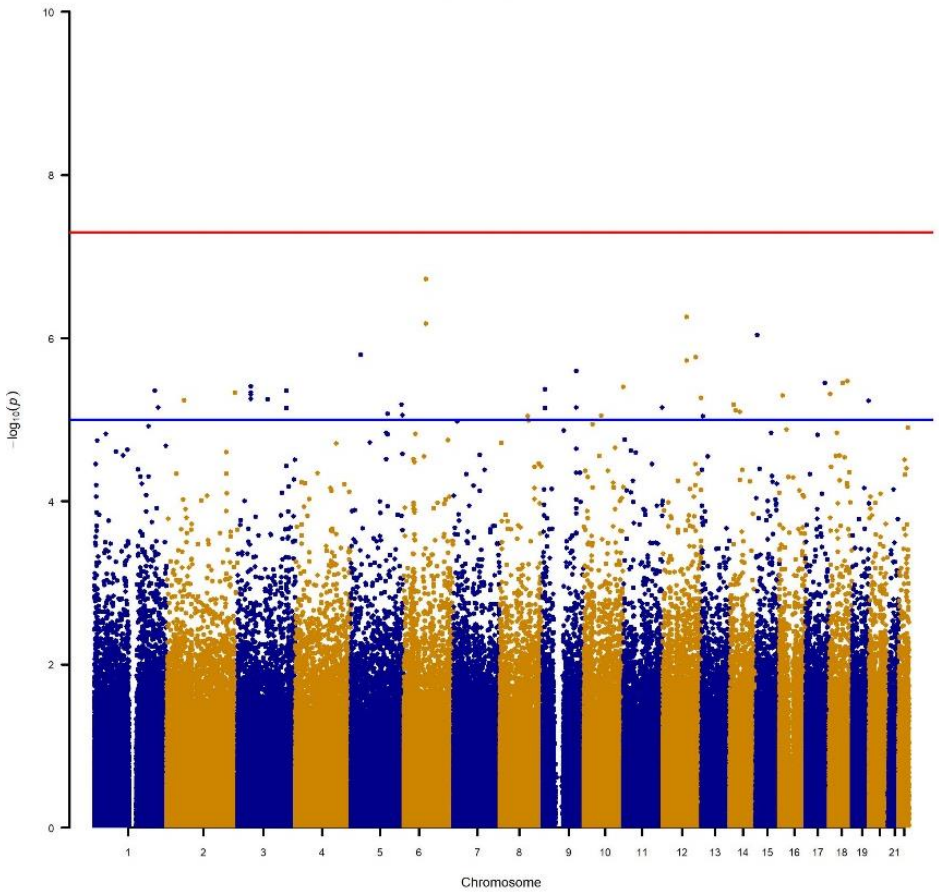


Figura 4.11. Distribución de marcadores en función de su probabilidad de asociación con la población de Aragón comparada con el resto de población española.

Cantabria *versus* Resto población española:

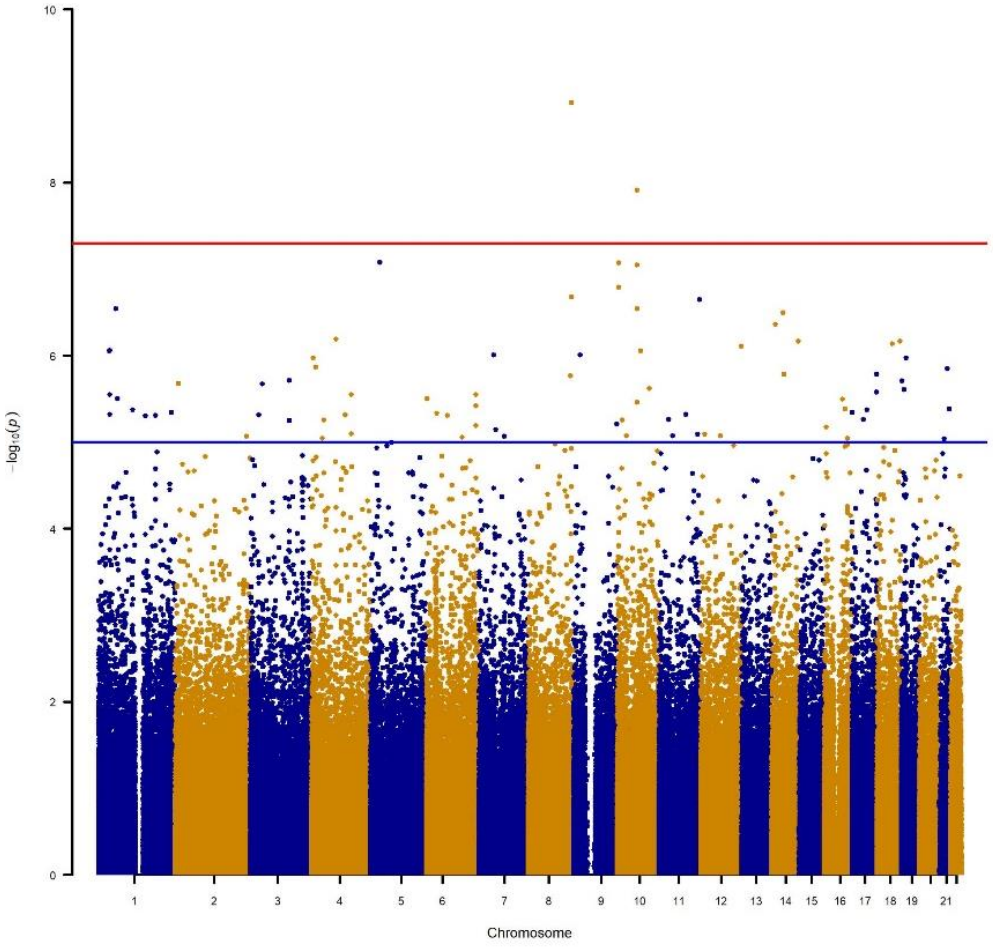


Figura 4.12 Distribución de marcadores en función de su probabilidad de asociación con la población de Cantabria comparada con el resto de población española.

En las Figuras 4.10, en la que se compararon 360 individuos pertenecientes a nuestra población de Andalucía frente a 1.496 del resto de la población española y en la 4.11 (64 individuos de Aragón frente a 1.796 del resto de la población ibérica), no se aprecia ninguna señal cromosómica significativa que diferencie las poblaciones analizadas frente al resto. No así en la Figura 4.12, en la que se distingue un marcador (poco significativo al tratarse de una señal puntual) en el cromosoma 5 ubicado en el gen *CDH6* y otro en el cromosoma 10. En este caso se comparaban 44 individuos de origen cántabro frente a 1.814, razón que puede explicar este posible artefacto.

Castilla *versus* Resto población española:

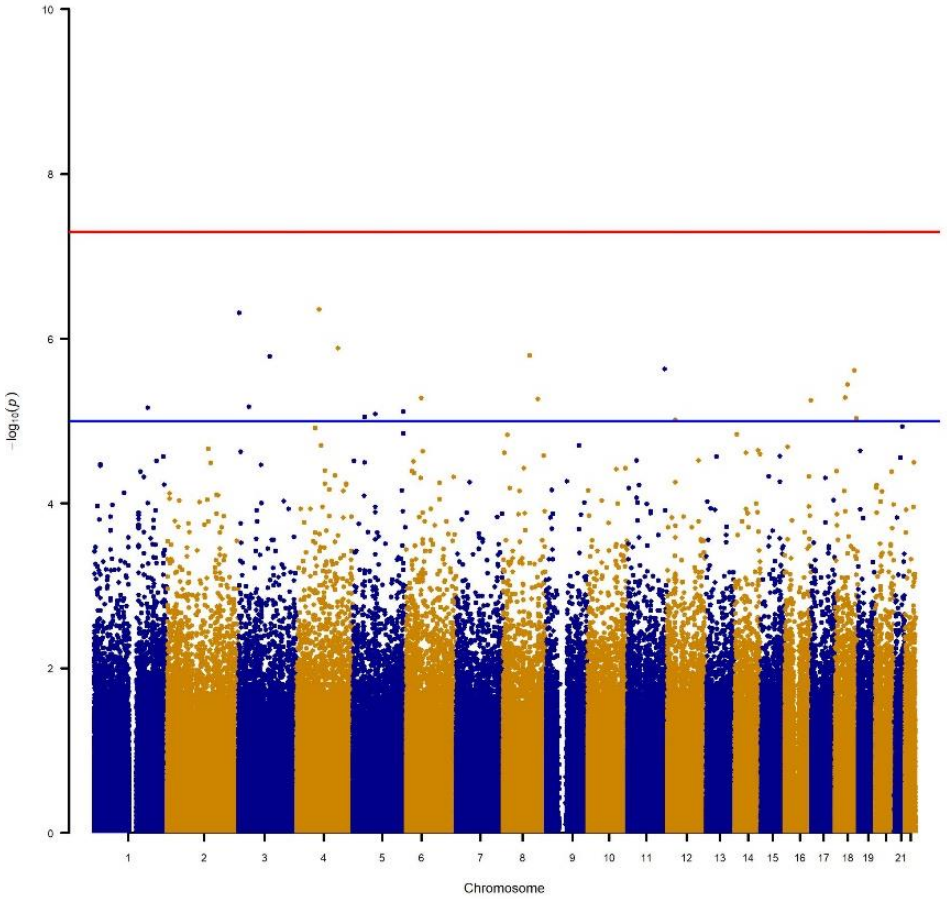


Figura 4.13. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Castilla comparada con el resto de población española.

Castilla-Mancha-Madrid *versus* Resto población española:

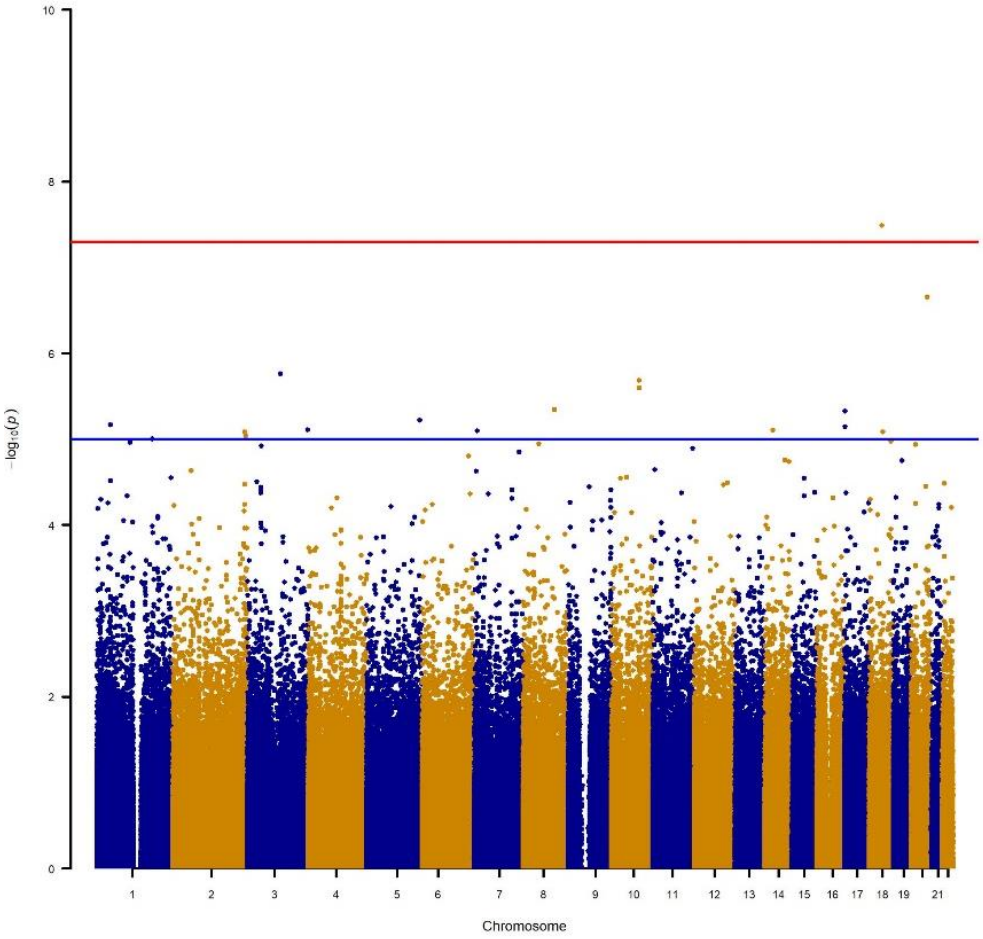


Figura 4.14. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Castilla-Mancha-Madrid comparada con el resto de población española.

En el MP de la Figura 4.13 no se aprecia ningún marcador significativo que diferencie la población analizada frente al resto. Se compararon 142 individuos de la población Castilla frente a 1.714. En la Figura 4.14, sin embargo y aunque la señal puede ser poco fiable al tratarse de un solo marcador aislado, se distingue un marcador en el cromosoma 18. En este caso se compararon 187 individuos de la población Castilla-Mancha-Madrid frente a 1.669. Esta diferencia tan elevada en la N puede dar lugar a ese posible artefacto.

Soria-Logroño *versus* Resto población española:

Al comparar nuestra población "Soria-Logroño" (41 individuos) frente al resto (1.815), también se distinguen dos marcadores que podrían diferenciar a nuestra población de estudio frente al resto de población española, aunque al tratarse solamente de dos tampoco podemos definirlo como significativo (Figura 4.15).

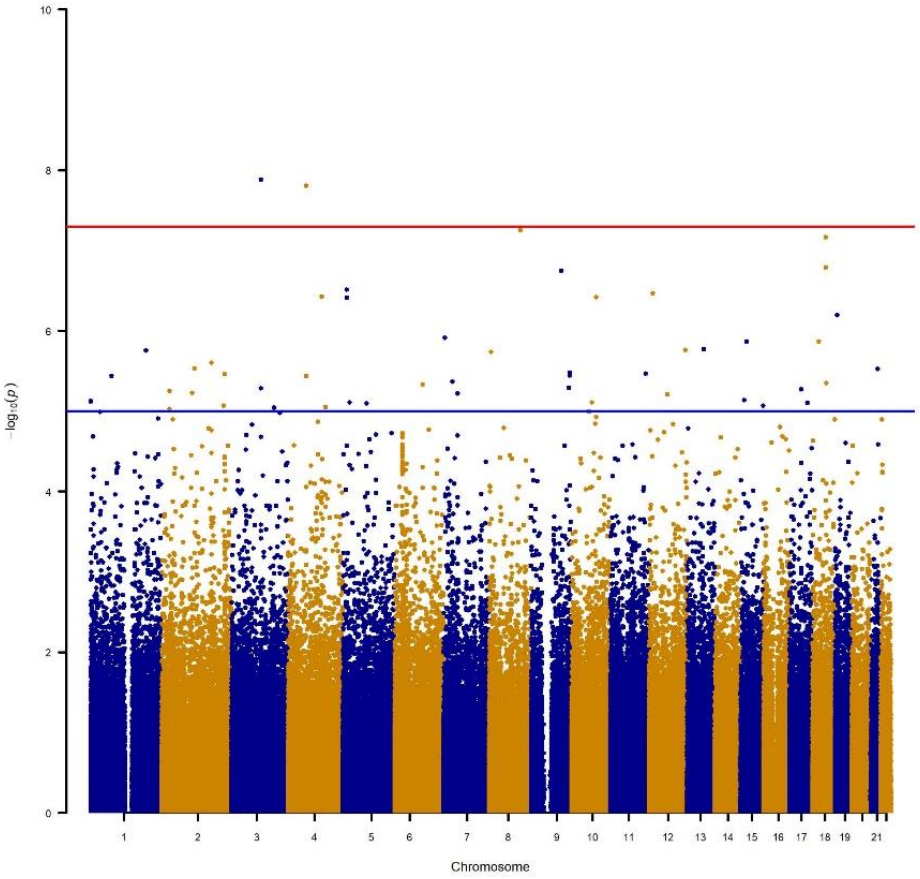


Figura 4.15. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como "Soria-Logroño" comparada con el resto de población española.

Continuando con el análisis de nuestras poblaciones y como se puede apreciar en los siguientes MP, no se vio ninguna señal que superase nuestro umbral establecido en las poblaciones de León (333 individuos comparados con 1.523) y Levante, donde se estableció una comparación de 187 individuos frente a 1.669 (Figuras 4.16 y 4.17), pero donde se si apreciaron muchas variantes significativas fue en nuestras poblaciones de Galicia (155 vs 1.701) y País Vasco - Navarra (59 vs 1.797), como se muestra en la Figura 4.18, Tabla 4.4, Figura 4.19 y Tabla 4.5 respectivamente:

León *versus* Resto población española:

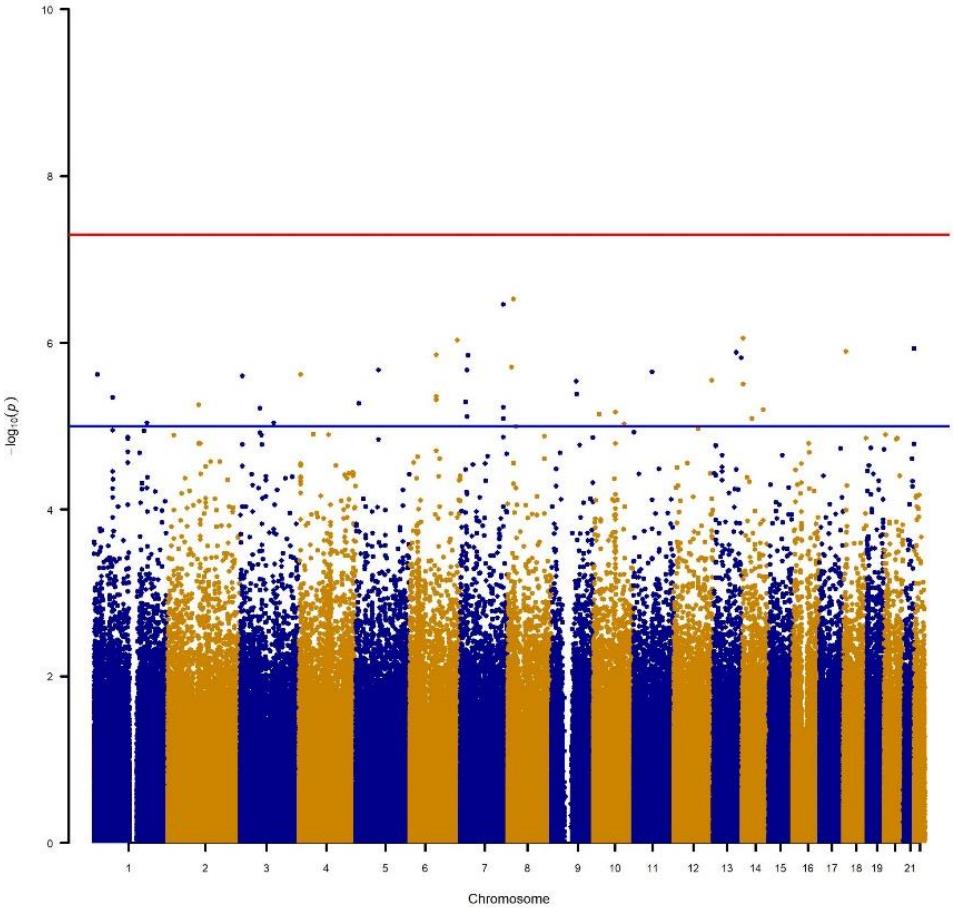


Figura 4.16. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como León comparada con el resto de población española.

Levante *versus* Resto población española:

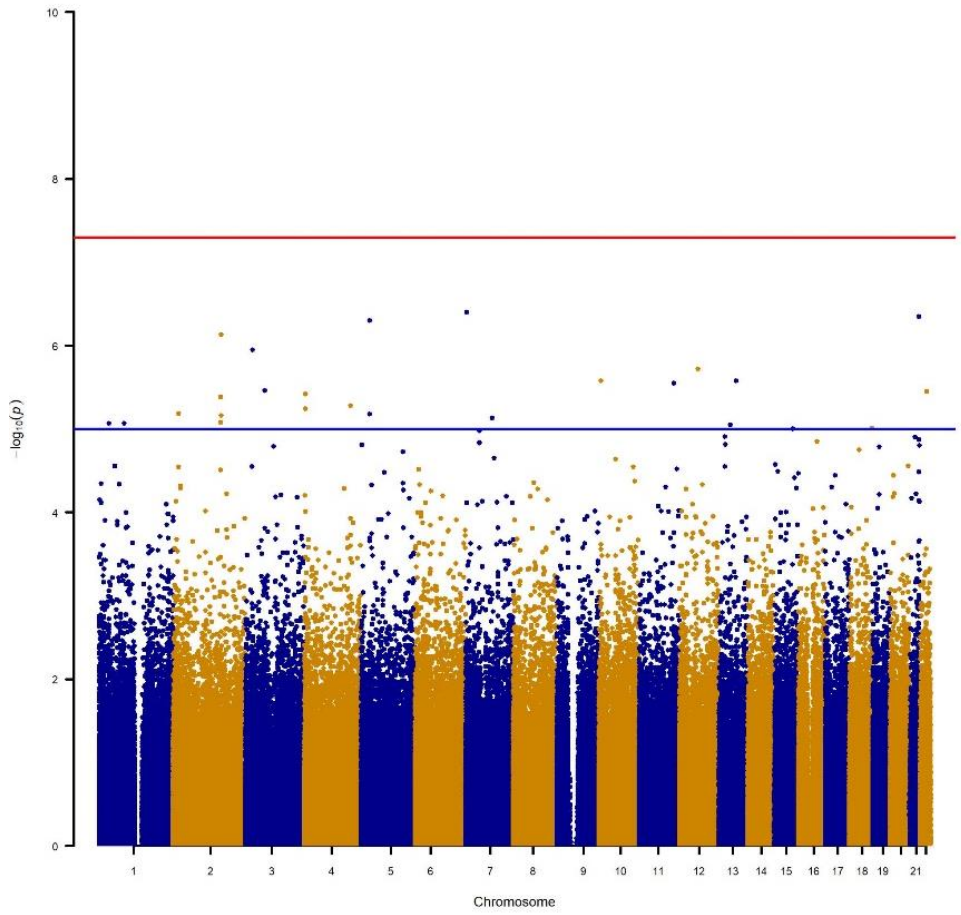


Figura 4.17. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Levante comparada con el resto de población española.

Galicia *versus* Resto población española:

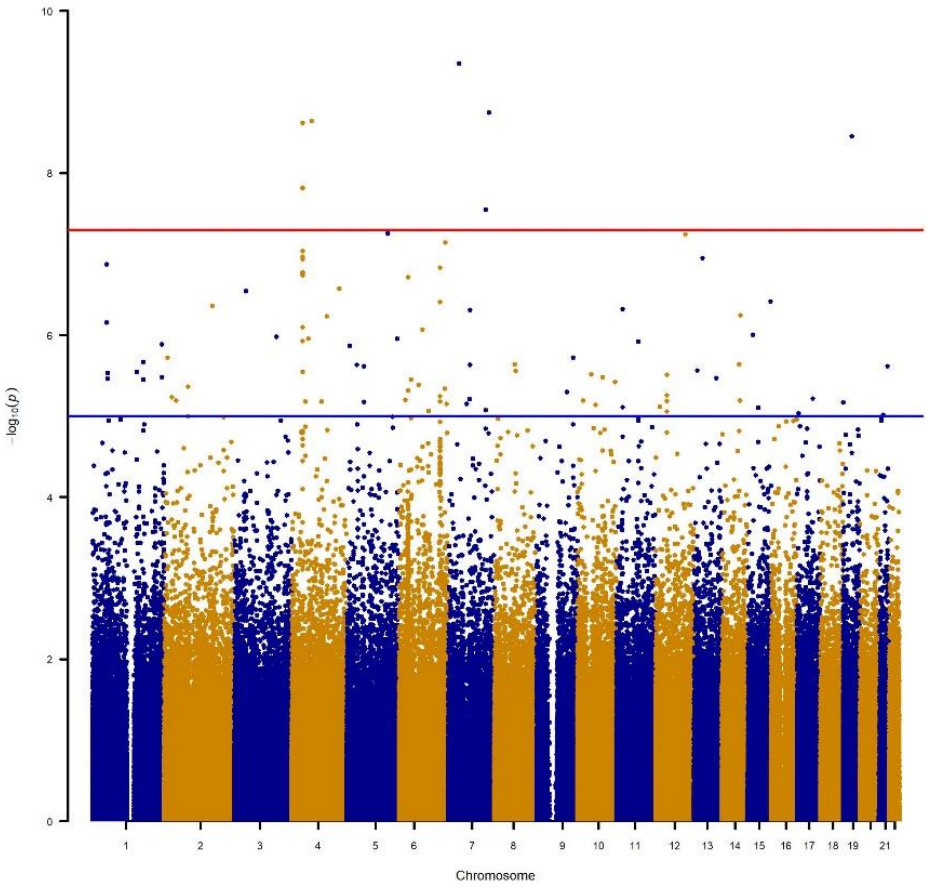


Figura 4.18. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Galicia comparada con el resto de población española.

Tabla 4.5. Variantes identificadas como significativas en nuestra población de estudio correspondiente a Galicia.

CHR	POSICIÓN (GRCh38)	A1	N_Total	low_GAL	P_GAL	dbSNP	tipo1	GEN
4	38756620	T	1848	0,2911	2,44E-06	rs6853255	*downstream*	*TLR10*
4	38804321	C	1855	0,3	9,26E-08	rs5743566	UTR-5	*TLR1*
4	38837698	C	1852	0,2857	1,52E-05	rs73236633	intrón	*TLR6*
4	38843988	A	1849	0,3348	1,55E-05	rs56245262	intrón	*TLR6*
4	69495194	C	1827	6,687	2,28E-06	rs185495830	*missense*	*UGT2B4*
7	36474489	T	1821	1,696	4,47E-07	rs76827509	*downstream*	*ANLN*
7	129817601	T	1836	2,093	2,85E-05	rs2631610	*downstream*	*UBE2H*
7	141988783	T	1850	1,682	1,81E-6	rs56982032	*upnstream*	*MGAM*

En nuestra población gallega de estudio encontramos señales especialmente en el cromosoma 4 y en el cromosoma 7 (Figura 4.18).

Las señales del cromosoma 4 parecen corresponder, fundamentalmente a los genes *TLR* (*TLR1, TLR6* y *TLR10* (Tabla 4.5).

País Vasco-Navarra *versus* Resto población española:

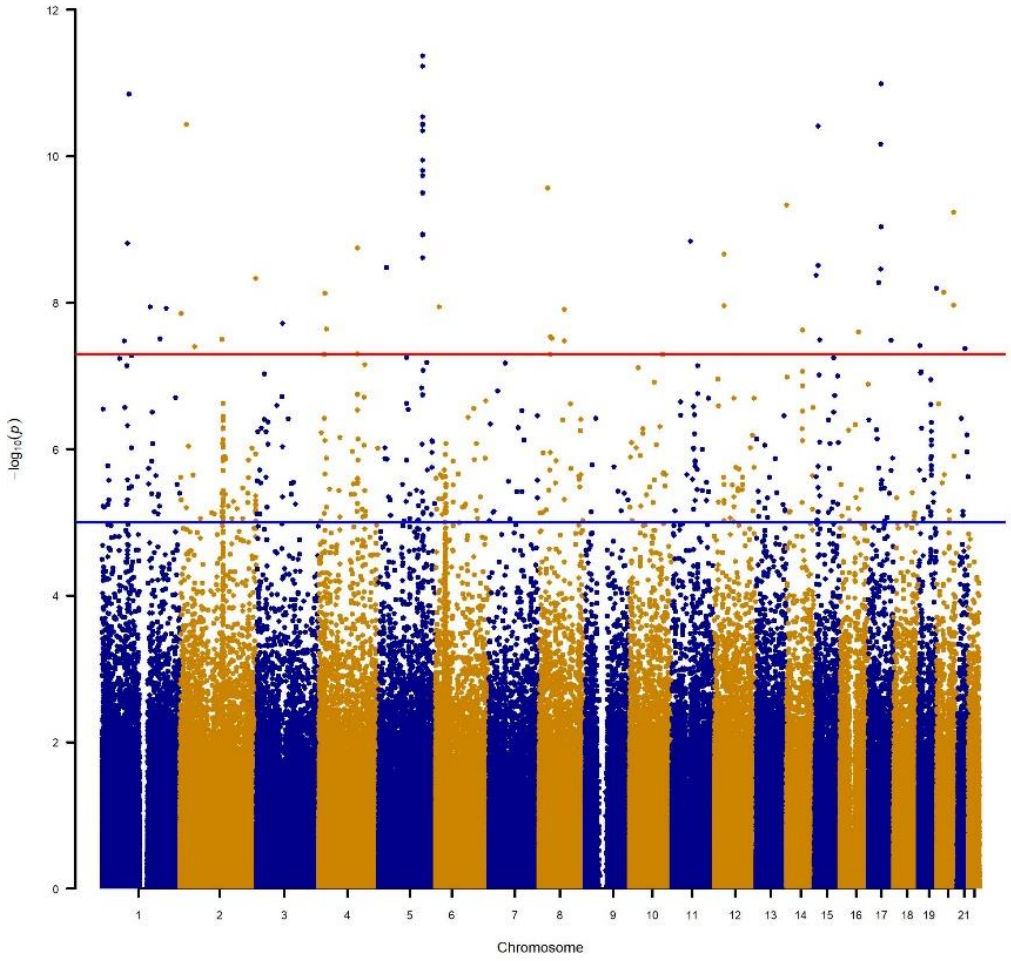


Figura 4.19. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como País Vasco - Navarra comparada con el resto de población española.

Tabla 4.6. Variantes identificadas como significativas en nuestra población de estudio correspondiente a País Vasco – Navarra.

CHR	POSICIÓN (GRCh38)	A1	N_Total	low_PV	P_PV	dbSNP	tipo1	GEN
2	19937945	A	1847	9.277	3,70E-08	rs140196566	*splice-sit*	*WDR35*
10	28912347	A	1852	3,62	7,84E-8	rs78854530	intrón	*RYR3*
15	33629208	A	1847	3.36	3,88E-08	rs77587145	intrón	RYR3
8	36385464	A	1855	2,822	5,12E-08	rs72612123	*upstream*	*KCNU1*
17	42969185	C	1855	8.443	1,04E-08	rs564812850	intrón	*PTGES3L*
17	42242349	T	1853	4.847	6,85E-08	rs117770911	intrón	*STAT5B*
17	42573724	G	1850	3.958	9,29E-10	rs75716493	intrón	*PSMC3IP*
1	86399845	A	1855	4.441	1,42E-08	rs112520043	*upstream*	*ODF2L*
5	141338999	C	1850	4.452	5,95E-09	rs191299749	*missense*	*PCDHGA2*
5	141365803	A	1856	4.219	4,51E-08	rs138451800	intrón	*PCDHGA3*
5	141394152	A	1855	4.268	3,64E-08	rs141035845	intrón	*PCDHGA3*
5	141399193	A	1854	4.318	2,94E-08	rs138699584	intrón	*PCDHGA3*
5	141409459	C	1852	4.261	3,77E-08	rs139792503	intrón	*PCDHGA3*
5	141418260	A	1854	4.266	3,69E-08	rs144490159	intrón	*PCDHGA3*
5	141628992	A	1854	4.792	4,29E-09	rs41290605	intrón	*HDAC3*

En nuestra población de estudio agrupada como País Vasco - Navarra nos encontramos 7 señales significativas correspondientes con variantes de baja frecuencia que caracterizan a esta población frente al resto, principalmente en los cromosomas 5 y 17.

Cataluña *versus* Resto población española:

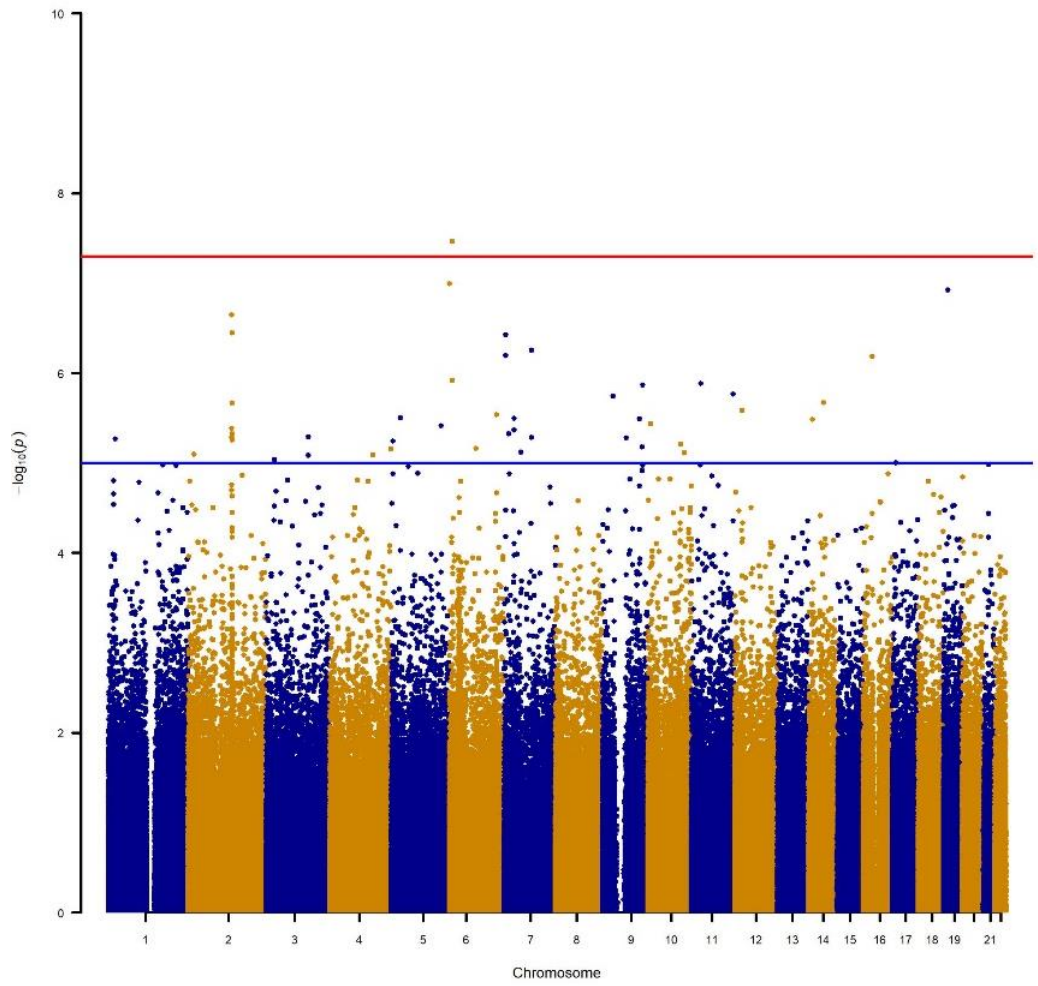


Figura 4.20. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Cataluña (201 individuos) comparada con el resto de población española (1.655).

En este caso se aprecia, en el cromosoma 6, un único marcador que alcanza el umbral establecido. En estos casos, como ya se ha dicho previamente, no podemos considerar que se trate de una señal significativa.

Extremadura *versus* Resto población española:

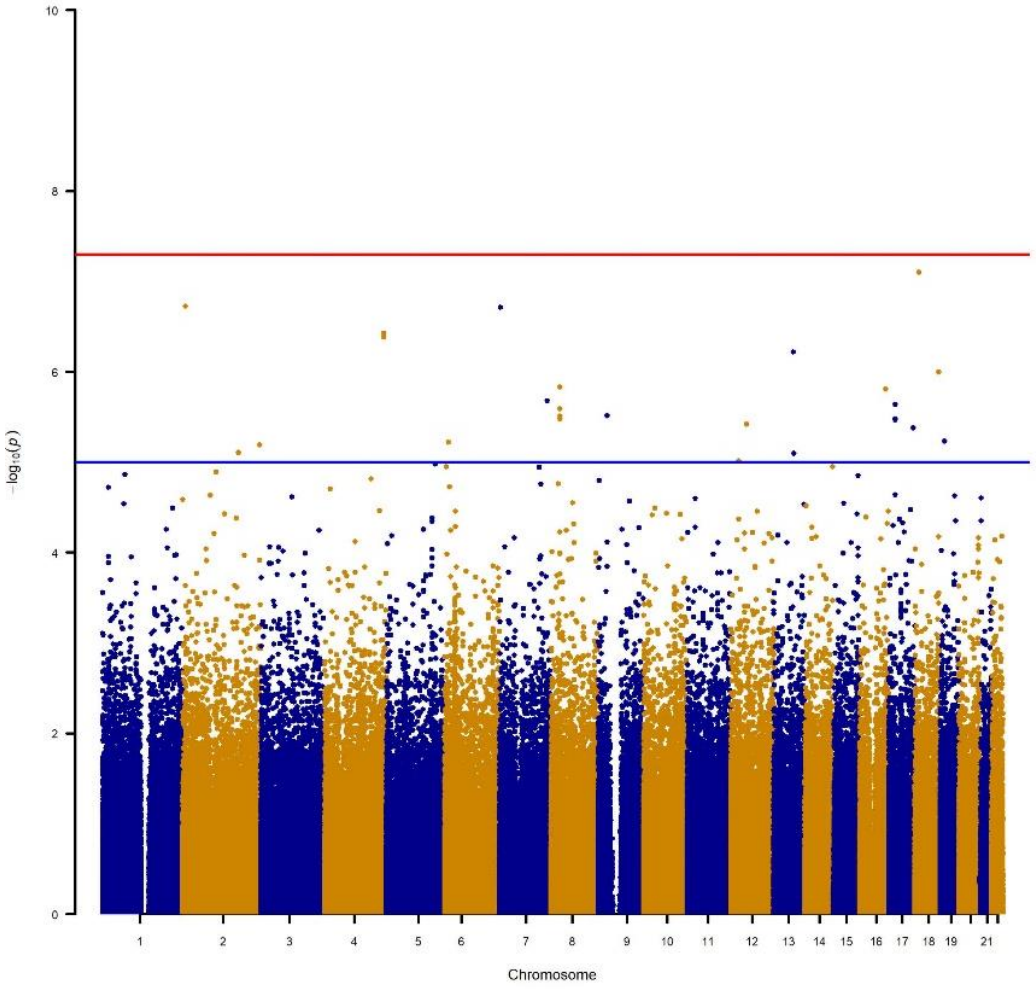


Figura 4.21. Distribución de marcadores responsables de la discriminación de nuestra población agrupada como Extremadura (89 individuos) comparada con el resto de población española (1.767).

En el MP de la Figura 4.21 no se aprecia ninguna señal cromosómica significativa que diferencie la población analizada frente al resto.

En todos los casos expuestos anteriormente, excepto Galicia y País Vasco_Navarra, no se aprecian realmente señales significativas que diferencien las subpoblaciones; es probable que no se observen debido a la baja N, que siempre supone una limitación.

En el Anexo III se muestran MP análogos a los mostrados en el texto, pero realizados con datos imputados en el servidor TopMed (Taliun et al.).

Los recursos de TopMed incluyen un buscador de variantes e imputación de datos genómicos y fenotípicos que están disponibles a través de dbGaP. En las primeras 53.831 muestras de TOPMed, se detectaron más de 400 millones de variantes de un solo nucleótido y de inserción o deleción no descritas anteriormente después de su alineación con el genoma de referencia. Entre los más de 400 millones de variantes detectadas, el 97% tiene frecuencias > 1% y el 46% son singletons. Estas raras variantes proporcionan información sobre los procesos mutacionales y la historia evolutiva humana reciente.

TOPMed contribuye mediante la identificación y caracterización de variantes raras, que comprenden la mayoría de la variación genómica humana. Las variantes raras representan cambios recientes y potencialmente perjudiciales que pueden afectar la función de las proteínas, la expresión génica u otros eventos biológicamente importantes.

Como es lógico, a través de los datos imputados, se apuntan las mismas regiones cromosómicas, pero con una mayor densidad de las señales e incluso pueden apreciarse nuevas asociaciones.

Además, se puede ver que los SNPs están casi completamente presentes en todas y cada una de las poblaciones. Ese no es el caso, como se esperaba, con respecto a variantes más raras (Taliun et al., 2021).

Finalmente, la demanda del *Axiom Spain Biobank Array Plate* así como del uso de los datos de genotipado de nuestras muestras de población control española para estudios de GWAS, quedan resumidos en el Anexo IV.

DISCUSIÓN

5 DISCUSIÓN

Tras describir y analizar los diferentes resultados obtenidos, procede ahora realizar una discusión para analizar con más detalle alguno de los resultados a la vez que aportar ideas para futuras investigaciones.

En el objetivo general que planteábamos en nuestra investigación, incluíamos la valoración del *array* de alta densidad enriquecido en variantes raras específicas de población española que hemos diseñado, aplicado sobre el grupo de muestras escogidas para el estudio, en función, entre otras variables, del origen geográfico. Por tanto, vamos a centrar la discusión en aquellos aspectos más relevantes que se han extraído de los resultados obtenidos, estableciendo una comparación con elementos específicos de trabajos previos con los que contrastaremos nuestros resultados y aportaciones.

5.1 IMPORTANCIA DE LA SELECCIÓN DE BIOMARCADORES: ENFERMEDAD COMÚN / VARIANTE RARA

Como ya se ha recalcado es muy importante prescindir del desequilibrio de ligamiento en estos estudios de asociación de genoma completo, ya que los métodos basados en el DL funcionan mejor cuando hay un solo alelo de susceptibilidad implicado en cualquier enfermedad dada, y por lo general funcionan mal si hay una heterogeneidad alélica sustancial. Las frecuencias alélicas de los *loci* de susceptibilidad en enfermedades complejas son muy variables. El número de alelos distintos que afectan a la susceptibilidad de enfermedad en un locus dado y sus respectivas frecuencias, incluiría también las penetrancias de todas las combinaciones del genotipo (Pritchard & Cox, 2002).

La comunidad genética ha logrado un gran éxito a la hora de encontrar genes responsables de una amplia gama de enfermedades mendelianas (Cooper, Ball, & Krawczak, 1998). Por el contrario, la búsqueda de variantes relacionadas con enfermedades complejas, como se ha visto a lo largo de este trabajo, había sido relativamente frustrante (N. J. Risch, 2000) a pesar del intenso esfuerzo de investigación tanto en el sector académico como en el comercial hasta que se aumentó la N lo suficiente para detectar muchas más variantes y explicar mucho más del componente genético de la heredabilidad. Así el nivel de predicción de algunas enfermedades comunes es mayor que el de enfermedades mendelianas (The Esophageal Adenocarcinoma Genetics et al.).

Como se señaló en un principio, la arquitectura alélica de enfermedades complejas representa un reto a la hora de diseñar estudios, en concreto los GWAS. La hipótesis CD/CV representa la mejor de las opciones para el mapeo de asociaciones a gran escala, por lo que es importante preguntarse si esta hipótesis podría ser, de alguna manera, definitiva. Parece importante plantearse y llevar a cabo más estudios basándose en la hipótesis CD/RV.

Según estudios científicos previos, las mutaciones implicadas en la enfermedad mendeliana son muy penetrantes y generalmente se encuentran bajo una selección muy fuerte. En cambio, las variantes de susceptibilidad involucradas en enfermedades complejas, parecen tener una penetrancia baja o media y probablemente no estén sujetas a una selección tan fuerte, lo que sugiere que las frecuencias se mantienen en niveles altos (D. E. Reich & Lander, 2001), (Pritchard, 2001).

Tener en cuenta el crecimiento de la población parece conducir a una descripción considerablemente mejor de alelos de muy baja frecuencia (D. E. Reich & Lander, 2001).

En el caso de que deseemos explorar alelos de baja frecuencia al mismo tiempo que la variación común en un GWAS, será necesario crear *arrays* de SNPs enriquecidos en variantes raras y de baja frecuencia que, además, son mucho más específicos de población.

Otro aspecto a tener en cuenta es la comprensión del desequilibrio de ligamiento en las distintas poblaciones humanas tal como se ha comentado en la introducción, ya que sin duda facilita el descubrimiento de genes que influyen en enfermedades humanas complejas. Con el proyecto HapMap se pusieron de manifiesto problemas prácticos a la hora de escoger marcadores, la influencia del modelo estadístico escogido para analizar estas estructuras "en bloque" y los niveles de genotipado necesarios para los estudios de enfermedades comunes. El conocimiento de los patrones de desequilibrio local puede ayudar a identificar polimorfismos comunes involucrados en enfermedades complejas, pero se requerirán métodos analíticos y diseños experimentales completamente nuevos para identificar importantes variantes raras (Cardon & Abecasis, 2003).

Del mismo modo cabe destacar que los bloques haplotípicos están altamente correlacionados entre poblaciones, lo que proporciona un poder estadístico sustancial en estudios de asociación de variación genética común en cada región (Gabriel et al., 2002). En cualquier caso, la importancia de estos estudios radica en el análisis de la distribución de la variabilidad genética de los individuos, caracterizando diferencias en la composición genética de las poblaciones y permitiendo evaluar la posible existencia de diferentes patrones de estratificación poblacional para las variantes genéticas comunes y raras. Así los GWAS y el estudio genético de las poblaciones están íntimamente ligados, ya bien sea para analizar la estructura genética de las poblaciones humanas modernas como para evaluar el efecto que las variantes de baja frecuencia pueden tener sobre la subestructura genética total. Mediante este tipo de estudios y cuanto más exhaustivos sean estos, será posible llevar a cabo la caracterización de la estructura poblacional y el efecto de variantes comunes y raras en la patogénesis de la enfermedad. Son estas variantes las que explican gran parte de la diversidad genética en nuestra especie, una consecuencia del corto período evolutivo y la ascendencia compartida de la población humana.

5.2 IMPORTANCIA DE LA SELECCIÓN DE CASOS Y CONTROLES

Se debe incidir en la importancia de disponer de una muestra homogénea. Los casos y los controles en los estudios de GWA deben pertenecer a la misma población para evitar la estratificación poblacional, aunque existan herramientas para corregirla. El fenotipo también debe ser homogéneo, evitando patologías relacionadas, fenocopias, etc…El tamaño muestral es muy importante, ya que cuanto mayor sea la N también se incrementará la potencia del estudio.

La consecuencia de no tener en cuenta estas recomendaciones puede llevar a conseguir resultados contradictorios: en este sentido se han publicado numerosos estudios basados en metaanálisis, que proporcionan enfoques para combinar resultados de varios estudios sobre el mismo tema y poder estimar y explicar su diversidad.

Una fuente de confusión obvia es que las poblaciones de estudio sean diferentes y/o la selección de casos y controles no sea homogénea. En cualquier caso, la heterogeneidad ofrece la posibilidad de observar una relación entre las poblaciones de estudio y sus circunstancias. En la heterogeneidad también existen beneficios que fundamentalmente residen en la necesidad de explorar y describir las fuentes de esta, ya que la exclusión de estudios con resultados extremos puede sesgar una síntesis de investigación y subestimar la verdadera varianza de los resultados, lo que contribuye a una inferencia engañosa. La búsqueda minuciosa es la mejor protección contra el sesgo (Mosteller & Colditz, 1996).

En el trabajo "*Replication validity of genetic association studies*" (Ioannidis, Contopoulos-Ioannidis, & Lau, 1999), evaluaron 370 estudios basados en metaanálisis. Pudieron mostrar que la heterogeneidad es muy frecuente y significativa entre estudios del mismo tipo, y que los resultados de un primer estudio se correlacionan sólo modestamente con investigaciones posteriores sobre la misma asociación. El primer estudio a menudo sugiere un efecto genético más fuerte que el encontrado en estudios posteriores. Tanto el sesgo como la diversidad de la población podrían explicar por qué los primeros estudios de asociación tienden a sobrestimar la protección o predisposición a la enfermedad conferida por un polimorfismo. En este trabajo pudieron concluir que las probabilidades de obtener una heterogeneidad estadísticamente significativa entre los estudios de un mismo tema son mayores cuantos más estudios se realicen. En los diez metaanálisis que analizaron con menos de seis estudios cada uno, no detectaron heterogeneidad significativa estadísticamente hablando entre los resultados de los estudios incluidos. Por el contrario, observaron una heterogeneidad estadísticamente significativa entre los estudios combinados en 7 de 9 metaanálisis con al menos 15 estudios. El poder de un metaanálisis para detectar heterogeneidad aumenta con estudios adicionales: la situación típica es que se detecta una asociación muy fuerte en un primer estudio y esta se vuelve paulatinamente menos intensa o incluso desaparece a medida que se acumulan más datos (Ioannidis et al., 1999).

Tal comportamiento puede sugerir un hallazgo falso que no es validado por investigaciones posteriores, un hallazgo exagerado o un efecto genético que es más fuerte en algunas subpoblaciones que en otras. De ahí la importancia de tener en cuenta la estructura poblacional. Cuando este factor no se valora o la elección de los controles no es acorde con los casos analizados, estudios posteriores no logran validar la propuesta original. De hecho, hay estudios en los que no se encuentran diferencias estadísticamente significativas entre los casos de enfermedad y controles, y esto puede ser debido a la incorrecta elección de estos últimos.

En el mismo trabajo (Ioannidis et al., 1999) también se pone de manifiesto que las probabilidades de encontrar una discrepancia estadísticamente significativa entre la investigación inicial y posteriores es mayor cuantos más estudios se lleven a cabo sobre el mismo tema, y obviamente un pequeño tamaño muestral también influye a la hora de arrojar resultados contundentes.

En resumen, los estudios de asociación genética requieren una replicación cautelosa. La heterogeneidad en la fuerza de una asociación es común incluso entre estudios de, aparentemente, poblaciones similares, que pueden diferir en parámetros que no son aún conocidos o en parámetros que los estudios originales no han capturado. El metaanálisis puede detectar diversidad no revelada previamente, y esto debería ser un objetivo en futuros estudios. La evaluación de los efectos de subgrupos (como diferencias geográficas o interacciones gen-ambiente) es difícil (Oxman & Guyatt, 1992) y requiere un gran número de sujetos (Ioannidis & Lau, 1998).

Muchos estudios de asociación basados en enfermedades se centran, en realidad, simplemente en la detección de genes en lugar de estimar el tamaño del efecto asociado con un gen en particular. La significación estadística aislada no garantiza una asociación genética, y la falta de significancia estadística formal no excluye la posibilidad de una asociación. Se reconoce la variabilidad genética de la población local como uno de los factores más relevantes en el descubrimiento de nuevas variantes de la enfermedad. Sin embargo, los datos genómicos de individuos sanos pertenecientes a la población local de interés son a menudo escasos o no están disponibles. De aquí la necesidad de la creación de catálogos de variación específica de población, que se discutirá en el apartado 5.5 de este capítulo.

5.3 INTERACCIONES GEN-GEN O GEN-AMBIENTE. ADAPTACIÓN Y SELECCIÓN

Está claro que genotipos diferentes responden a la variación ambiental de diversas maneras, así como que las interacciones entre diferentes formas alélicas de los genes dan lugar a distintos fenotipos.

Los genes no suelen actuar de forma individual. Forman parte de un sistema complejo, el genoma, en el que se producen interacciones genéticas que llevan a que los efectos de un gen o de una variante genética sean modificados por la acción de otro elemento genético o se vean influenciados por un tercero.

Identificar las interacciones entre los genes es un paso esencial para poder entender el funcionamiento de las células y tejidos, así como para conocer cómo se producen muchas de las enfermedades humanas. Este conocimiento podría también contribuir a determinar por qué la presencia de mutaciones que deberían desencadenar una enfermedad hereditaria no siempre deriva en la patología o por qué una misma mutación puede manifestarse de forma diferente en dos personas distintas. Además, puesto que el análisis del genoma está siendo utilizado cada vez más en el ámbito de la medicina, tanto en diagnóstico como en la toma de decisiones sobre el tratamiento, la identificación de las interacciones génicas resulta especialmente importante para proporcionar diagnósticos de más calidad (D. E. Reich et al., 2002).

Parece que el origen geográfico subyace a diversos aspectos de la biología, incluida la evolución del sexo, la especiación y las enfermedades complejas (P. C. Phillips, 2008). A pesar de los avances tecnológicos en secuenciación y genotipado, el mapeo de las indicaciones geográficas asociadas con la variación natural dentro del genoma de un individuo sigue siendo difícil (Zuk et al., 2014). Sin embargo, diversos modelos como levaduras y gusanos, así como cultivos celulares derivados de Drosophila y mamíferos, brindan un formato experimental para explorar estas indicaciones geográficas que se podría aplicar a otras especies.

En principio, se podría usar cualquier fenotipo para detectar interacciones génicas, pero un fenotipo en crecimiento celular, que se puede medir y cuantificar fácilmente, integra la fisiología celular y ha demostrado ser especialmente informativo para explorar interacciones génicas, tanto en levaduras como en células animales.

Sin embargo, la influencia de los factores ambientales en la estructura y topología general de la red génica y cómo esta cambia en respuesta a diferentes entornos, sigue sin estar clara. Según estudios publicados es posible que la estructura general y la topología de la red genética global siga siendo relativamente sólida a las influencias ambientales, aunque esto debe ser explorado sistemáticamente (Bandyopadhyay et al., 2010) (Guenole et al., 2013).

También cabe tener en cuenta las interacciones genéticas-químicas, ya que existen combinaciones alélicas hipersensibles o resistentes a un compuesto dado. Del mismo modo la supresión genética y efectos modificadores, procesos en los que un alelo mutante se ve compensado parcial o totalmente por una mutación supresora. Las mutaciones supresoras pueden ser intragénicas, ocurriendo en el mismo gen que la mutación original, o extragénicas, involucrando mutaciones en dos genes diferentes. Para poder detectar estos efectos se trabaja con organismos modelo como los detallados anteriormente (Costanzo et al., 2019). Son innumerables los factores que habría que tener en cuenta a la hora de describir interacciones génicas. Por supuesto no podemos olvidar las múltiples, en las que están involucrados más de 3 genes. Un objetivo importante del análisis de redes genéticas es descubrir las indicaciones geográficas impulsadas por la variación natural, que, según sugiere una creciente evidencia, contribuye a la relación genotipo-fenotipo de los individuos (Costanzo et al., 2019).

Costanzo et al. (2019), en su trabajo, repasan los diferentes tipos de interacciones génicas que pueden identificarse, así como los modelos de estudio disponibles en la actualidad y los

tipos de análisis que pueden realizarse. Además, exponen cómo la información obtenida en sistemas sencillos puede trasladarse a organismos más complejos, o cómo puede resultar especialmente relevante en enfermedades humanas como el cáncer.

Desde luego resulta de absoluto interés explorar las redes de interacción genética humana y las enfermedades. Los avances en las tecnologías de edición de genes durante la última década han transformado los modelos funcionales y proporcionando nuevas herramientas para explorar esta interacción. El desarrollo de CRISPR (del inglés *Clustered Regularly Interspaced Short Palindromic Repeats* - repeticiones palindrómicas cortas agrupadas regularmente interespaciadas) promete una caracterización funcional sin precedentes del genoma humano. Estos estudios han sentado las bases para una nueva ola de genómica funcional para caracterizar el papel de los genes esenciales, cómo este depende de los contextos genéticos y tisulares y cómo evolucionan e interactúan con la enfermedad (Bartha, di Iulio, Venter, & Telenti, 2018) (Rancati, Moffat, Typas, & Pavelka, 2018). Por supuesto también han revolucionado por completo el análisis de mutaciones de significado incierto, ya que se pueden hacer *knockout* y *knockin* con rapidez. En laboratorios como el nuestro podemos desarrollar organoides (trabajamos mucho en cerebroides), sistemoides o humanoides y utilizamos como modelo animal el pez cebra (Pensado-Lopez, Veiga-Rua, Carracedo, Allegue, & Sanchez, 2020), (The Esophageal Adenocarcinoma Genetics et al.).

El concepto de interacción genética considera que dos genes interaccionan cuando la combinación de un alelo en uno de ellos con un alelo del segundo genera un doble mutante con fenotipo inesperado (es decir, cuando la combinación de ciertas formas de ambos genes desencadena un efecto no esperado). Un ejemplo de interacción sería la letalidad sintética, que es aquella interacción extrema en la que la presencia de mutaciones en uno cualquiera de dos genes resulta viable para la célula, pero en la que la presencia de mutaciones en ambos genes de forma simultánea induce letalidad.

Identificar una interacción entre genes no es fácil, especialmente en los organismos complejos. Por esta razón, la mayoría de los estudios destinados a identificar interacciones genéticas se llevan a cabo en modelos animales simples o en levaduras como se ha mencionado anteriormente. Estos sistemas más sencillos permiten disponer de paneles de células en las que generar las dobles mutaciones u organismos con mutaciones que pueden ser cruzados entre sí para ver el resultado.

Además, muchas de las interacciones génicas están organizadas, bien a través de las rutas bioquímicas o bien por los compartimentos celulares en las que actúan las proteínas resultantes de la actividad génica. Estas interacciones génicas pueden formar redes génicas dentro de la célula que pueden ser detectadas y analizadas a partir de diferentes aproximaciones.

En definitiva, las variantes de múltiples genes y sus combinaciones van a ser diferentes para las distintas personas, y estas combinaciones específicas podrían no solo afectar profundamente la susceptibilidad a la enfermedad, sino que probablemente también prescribirán nuevas terapias personalizadas.

5.4 DISEÑO DE UN CATÁLOGO DE VARIACIÓN ESPECÍFICA RARA

Las poblaciones presentan diferencias en frecuencias alélicas, especialmente en variantes raras, que es necesario caracterizar para distinguir asociaciones reales de polimorfismos específicos de población y controlar correctamente la estratificación en los estudios de asociación.

En los últimos años, un gran número de proyectos genómicos de población han explorado la existencia de una enorme cantidad de variaciones raras en los genomas de poblaciones humanas que se ha observado que son específicas de la población (Fu et al., 2013; Keinan & Clark,

2012; Tennessen et al., 2012). En consecuencia, entendemos que se deben realizar más esfuerzos para generar bases de datos locales de variantes raras, objetivo que nos planteamos en el presente trabajo.

En este sentido, la obtención de una población control española, genotipada con un *array* de alta densidad enriquecido en variantes raras específicas de población, permitirá confirmar la existencia de diferencias en frecuencias alélicas respecto a bases de datos de referencia. Además, el uso de un *array* específicamente diseñado para cubrir variación funcional específica de población española permitirá caracterizar la variación rara local de una forma más detallada y comparar la estructura poblacional a escala microgeográfica en variación común y rara.

La idea de diseñar catálogos de variación específica de población está la orden del día, tal como se ha visto en las revisiones detalladas en este trabajo y en las distintas y numerosas publicaciones que se están llevando a cabo actualmente.

En cuanto a población española, ya ha sido creada la primera base de datos de la variabilidad genética, como se ha visto, y que además ha sido utilizada para la selección de exomas con los que hemos trabajado (Pena-Chilet et al., 2021). La plataforma, denominada Servidor Colaborativo de Variabilidad Española (CSVS, por sus siglas en inglés *Collaborative Spanish Variability Server*), recoge un total de 2.027 genomas y exomas de individuos españoles no emparentados.

Este repositorio ayudará al desarrollo y aplicación de la medicina personalizada, ya que el conocimiento de la variabilidad genética de la población local es de suma importancia en este ámbito y se ha revelado, además, como un factor determinante para el descubrimiento de nuevas variantes de la enfermedad.

La principal aportación de CSVS es la disponibilidad de las frecuencias en las que se observan las variantes en los genomas de población española, lo que es muy importante para el descubrimiento de nuevas variantes de patologías hereditarias (normalmente raras). Su importancia radica en que, cuando se secuencian genomas de pacientes sin diagnóstico, aparecen del orden de un millón de variantes. Para encontrar la nueva variante causal hay que descartar aquellas que con gran probabilidad no están asociadas a afección, como son las variantes conocidas, o variantes que están en una alta proporción en la población y no son compatibles con una enfermedad rara, por lo que se aplica un filtro que permite descartar una enorme cantidad de variantes y centrar los estudios en las más probablemente patogénicas, de manera similar a lo hecho aquí para el diseño del SBA.

El CSVS proporciona un catálogo exacto de la variación en población española, algo que no está recogido en las bases de datos internacionales. Ahora se sabe que hay muchas ramificaciones polimórficas típicamente españolas que no aparecen en las bases de datos internacionales y que, por tanto, sin CSVS podrían ser tomadas por variantes raras y retrasar mucho el descubrimiento de la verdadera variante causal. Por lo tanto, la plataforma contribuye al descubrimiento más eficiente de nuevas vías de enfermedad, lo que permitirá aumentar el conocimiento sobre las patologías y ser capaces de diagnosticar y estratificar pacientes previamente. En muchas enfermedades, como los cánceres hereditarios, la nueva herramienta permitirá tomar medidas preventivas antes, con lo que ciertamente contribuye, a la larga, a reducir la mortalidad en este tipo de afecciones.

En la actualidad, más de 4.500 patologías monogénicas (causadas por la mutación de un solo gen) pueden ser diagnosticadas directamente mediante genómica personalizada. Esta capacidad diagnóstica podría aplicarse en un futuro cercano a todo el espectro de enfermedades raras de origen genético.

En nuestros datos mostrados en el Anexo I es destacable que las variantes que han cambiado de categoría de comunes en Europa a raras o monomórficas en España, coinciden en

la categoría de farmacogenómica, lo que resulta muy interesante para la utilización del *array* en estudios de GWA con objetivos de imputación. Sería importante confirmar en detalle si esto ocurre exactamente del mismo modo comparando nuestra población aquí analizada con la 1000G-IBS, ya que es probable que las diferencias locales derivasen en resultados diferentes.

Incidiendo en la importancia de disponer de datos genómicos de población control, por norma general, se tiende a secuenciar principalmente los genomas de personas con alguna dolencia y no los de personas sanas. Sin embargo, es necesario contar con la información genómica de los sujetos sanos para discriminar e identificar nuevas variantes genómicas de las enfermedades. La iniciativa europea de 1M *Genomes* pretende, precisamente, compilar un millón de genomas de poblaciones europeas y minorías, al menos 500.000 de población control y el resto hasta el millón de genomas de casos de enfermedades genéticas y cáncer.

En esta línea volvemos a destacar la gran importancia de las variantes de interés farmacogenómico, que son claves para pautar las dosis para administrar los medicamentos a los pacientes y que tienen un interés particular.

En el Servidor Colaborativo de Variabilidad Española las secuencias se han agrupado por categorías superiores ICD10 (por sus siglas en inglés *The International Classification of Diseases*), por lo que se puede consultar la base de datos eliminando una o más categorías de dicha clasificación. De este modo, se pueden obtener recuentos de variantes en genomas de la población española sana que servirían como pseudocontroles para estudios de búsqueda de nuevos genes e incluso algunos estudios poblacionales como, por ejemplo, la prevalencia de variantes farmacogenómicas.

Es importante que todos los investigadores continuemos recogiendo nuevos genomas y exomas de individuos españoles para ampliar datos sobre la variabilidad genética española y las iniciativas del CIBERER en enfermedades raras y la infraestructura IMPaCT, con sus ejes de epidemiología, datos y medicina genómica, son claves en la generación y organización de datos.

5.5 DISCUSIÓN SOBRE LA SUBESTRUCTURA POBLACIONAL RELACIONADA CON VARIANTES RARAS Y DE BAJA FRECUENCIA; IMPLEMENTANDO NUEVOS ENFOQUES

Para llevar a cabo un estudio basado en la genética de una enfermedad y por supuesto, en el contexto de una población, también deben ser tenidos en cuenta los procesos evolutivos. Por tanto, no podemos olvidar en este apartado los pilares de la Biología evolutiva, entre ellos la selección natural, aunque también hay que mencionar la deriva genética, los cuellos de botella, mutación, endogamia, consanguinidad, expansión, migración, mezcla de poblaciones y la tasa de recombinación entre otros. Destacar que los "puntos calientes" de recombinación son una característica general del genoma humano y tienen un papel principal en la configuración de la variación genética en poblaciones humanas (D. E. Reich et al., 2002).

La selección natural (https://es.wikipedia.org/wiki/Selecci%C3%B3n_natural) es un proceso evolutivo que fue descrito por Charles Darwin en su libro El origen de las especies e inspirado en las ideas del Ensayo sobre el principio de la población de Thomas Malthus, que establece la supervivencia del más apto o la preponderancia de la ley del más fuerte en un medio natural sin intervención externa, por lo que los individuos menos aptos o más débiles perecen y sus rasgos no se transmiten a las generaciones siguientes al no reproducirse, en contraposición al concepto de selección artificial, donde sí existe una intervención directa, por el humano, con el propósito de mejorar los rasgos de los individuos manipulándolos a voluntad. Estrictamente hablando, se define como la supervivencia y reproducción diferencial de los fenotipos de una población biológica. La formulación clásica de la selección natural

establece que las condiciones de un medio ambiente favorecen o dificultan, es decir, seleccionan la reproducción de los organismos vivos según sean sus peculiaridades. La selección natural fue propuesta por Darwin como medio para explicar la evolución biológica. Esta explicación parte de tres premisas; la primera de ellas es que el rasgo sujeto a selección debe ser heredable. La segunda sostiene que debe existir variabilidad del rasgo entre los individuos de una población. La tercera premisa plantea que la variabilidad del rasgo debe dar lugar a diferencias en la supervivencia o éxito reproductor, haciendo que algunas características de nueva aparición se puedan extender en la población. La acumulación de estos cambios a lo largo de las generaciones produciría todos los fenómenos evolutivos.

En su formulación inicial, la teoría de la evolución por selección natural constituye el gran aporte de Charles Darwin (e, independientemente, por Alfred Russel Wallace), y es un pilar fundamental del darwinismo, posteriormente reformulado en la actual teoría de la evolución conocida como neodarwinismo o síntesis evolutiva moderna. En biología evolutiva, el proceso de selección natural se considera la principal causa del origen de las especies y de su adaptación al medio.

La selección natural puede ser expresada con la siguiente ley general, tomada de la conclusión de El origen de las especies:

"Existen organismos que se reproducen y la progenie hereda características de sus progenitores, existen variaciones de características si el medio ambiente no admite a todos los miembros de una población en crecimiento. Entonces aquellos miembros de la población con características menos adaptadas (según lo determine su medio ambiente) morirán con mayor probabilidad. Entonces aquellos miembros con características mejor adaptadas sobrevivirán más probablemente".

"*Darwin, El origen de las especies*"

Para establecer un modelo genético de poblaciones, también es necesario hacer algunas suposiciones sobre la historia de la población. El modelo más simple asume una población de apareamiento aleatorio con un crecimiento constante del tamaño de esta, pero esto ignora varias complejidades, incluyendo, en particular, el dramático aumento de la población humana hasta la actualidad, lo que se puede traducir en la frecuencia de distribución de los alelos; parece que el impacto del crecimiento puede ser lo suficientemente modesto como para que, aparte de afectar a variantes de baja frecuencia (discutidas a continuación), también lo haga sobre el espectro de frecuencias alélicas general (Moorad & Wade, 2005).

Como hemos visto en capítulos anteriores, hemos evaluado aquí la aplicabilidad de varias técnicas multivariantes para discriminar poblaciones separadas geográficamente que fueron tipificadas inicialmente con un *array* específicamente diseñado para analizar variación rara en población española realizando el análisis a una escala más local.

La comparación de los resultados cuando se aplican estas aproximaciones a las diferentes clases de marcadores nos permite discutir sobre los patrones fuertes de estratificación esperados en cuanto a variación rara (Babron et al., 2012; Mathieson & McVean, 2012).

Primero, abordamos los datos por PCA, uno de los métodos más populares para corregir estratificación en estudios de asociación y que también podría utilizarse para discriminar poblaciones geográficamente separadas (Price et al., 2006).

Dado que nuestro estudio se llevó a cabo de manera local en cuanto a grupos geográficamente definidos, la acción de la selección natural puede no ser muy fuerte, e incluso si en algún momento lo fue podría haber sido mitigada por el flujo constante de genes.

Con respecto a las variantes raras, esperaríamos ver más diferencias entre las poblaciones, ya que se pueden conservar a frecuencias muy bajas (escondidas en los heterocigotos), e incluso

ser perjudiciales. Sin embargo, en nuestro caso, el estudio tiene poco poder para sacar conclusiones, ya que una variante rara puede estar presente en una población, pero no se genotipó porque la N no fue suficiente para detectarla.

Primero evaluamos la distribución de los marcadores *rare* y todos los demás (SNPs y *low*) en nuestras 12 poblaciones de estudio agrupadas en función de origen geográfico definido y por proximidad geográfica. Las tablas resultantes (Tablas 4.1, 4.2 y 4.3) destacan algo obvio: a medida que descendemos en el nivel de frecuencia y consideramos los marcadores más raros, el número de casos únicos (presentes en una sola población) aumenta.

Nuestros paneles de variantes raras y de baja frecuencia dieron como resultado una buena diferenciación entre las 12 poblaciones cuando se compararon las primeras PCs. Esto puede compararse con el trabajo de Babron, ya que, a pesar de que nuestro conjunto de datos es muy pequeño, ambos estudios se realizan a escala local. En el estudio del Reino Unido, los valores propios son mucho más altos, especialmente en las clases de frecuencias más bajas, y el primer PC puede ser usado como representante de la variabilidad total. En nuestro estudio la variabilidad se evaluó a lo largo de varios PCs de cada panel, quedándonos finalmente con los dos primeros. De esta manera, podríamos conseguir una idea general de cómo los diferentes tipos de marcadores capturan información distinta y por lo tanto discriminan las poblaciones.

En la Tabla 4.4 se muestran las señales encontradas evaluando las tres categorías de marcadores, se aprecian en el cromosoma 16 en los tres casos (tanto para SNPs, *low* y *rare*) e incluyen genes que pueden estar relacionados con patologías prevalentes en nuestra población. Este podría ser el caso del gen *TEDC2*; este gen se expresa en tejidos en desarrollo, como células madre y tejido fetal diferenciado, por lo que es probable que desempeñe un papel en la replicación del ADN o la división celular. Es posible que *TEDC2* desempeñe una función en el desarrollo del tumor cuando muta (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TEDC2). También relacionado con la regulación génica nos encontramos en este cromosoma con los genes *TBL3* y *FAHD1*, cuyas enfermedades asociadas incluyen azoospermia y linfoma cerebral (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TBL3&keywords=TBL3, https://www.genecards.org/cgi-bin/carddisp.pl?gene=FAHD1&keywords=FAHD1).

En cuanto a enfermedades neurodegenerativas encontramos señales correspondientes a genes cuyas mutaciones originan patologías de este tipo, como es el caso de *MAPK8IP3*, que codifica una proteína de andamiaje en las células neuronales (https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAPK8IP3&keywords=MAPK8IP3) y *EME2*, implicado en la Deficiencia Combinada de Fosforilación Oxidativa debido a un defecto en la síntesis de proteínas mitocondriales que origina Síndrome de Leigh, una condición neurológica degenerativa hereditaria rara (https://www.genecards.org/cgi-bin/carddisp.pl?gene=EME2&keywords=EME2).

También podemos enmarcar en este tipo de patologías las relacionadas con mutaciones en el gen *HAGH*, que al igual que la anterior se manifiesta a edades tempranas y da lugar a neurodegeneración entre otros problemas (https://www.genecards.org/cgi-bin/carddisp.pl?gene=HAGH&keywords=HAGH).

No existen evidencias científicas de que estas patologías se presenten con una mayor incidencia en la población española.

Además, nos encontramos en esta señal cromosómica genes relacionados con la codificación de proteínas ribosómicas: *MRPS34*, *RPL3L* y *RPS2*.

Por último, se puede destacar el gen *CCNF*. Mutaciones en este gen originan esclerosis lateral amiotrófica (ELA), una enfermedad progresiva del sistema nervioso que afecta las neuronas en el cerebro y en la médula espinal causando pérdida del control muscular

(https://www.genecards.org/cgi-bin/carddisp.pl?gene=CCNF&keywords=CCNF). En España hay 4.000 pacientes (44% en Andalucía y Madrid). La incidencia habitual en otros países europeos es 1 caso por cada 100.000 habitantes; en España se presentan 3 por cada 100.000 (https://www.sen.es/saladeprensa/pdf/Link191.pdf).

El gen *CCNF* codifica para la proteína ciclina F, que cataliza la transferencia de ubiquitina a las proteínas que van a ser degradadas por el proteosoma. Así, *CCNF* forma parte de un sistema destinado a mantener la homeostasis de proteínas evitando la formación de agregados o precipitados proteicos en las células. Williams y colaboradores (2016), (K. L. Williams et al., 2016) observaron que la introducción de las variantes genéticas de *CCNF* asociadas a ELA y demencia frontotemporal en células nerviosas, lleva a la acumulación de proteínas ubiquitiniladas como TDP-43, cuya acumulación es un rasgo patológico característico de la ELA que se presenta también en la mayoría de los casos de demencia frontotemporal.

Es, cuando menos llamativo, que esta enfermedad se presente en una incidencia mayor en España que en otros países y que podamos apreciar una señal significativa en nuestra población española a nivel de este gen.

Por otro lado, teniendo en cuenta solamente las señales encontradas para SNPs y *low*, vemos que, además de genes comunes en el cromosoma 16 hay también una región destacada en el cromosoma 15, donde resultan significativas muchas variantes localizadas en el gen *FBN1*, gen que produce fibrilina 1, proteína que forma fibras elásticas en el tejido conectivo para brindar soporte a los huesos, los músculos y los órganos. Mutaciones en este gen ocasionan síndrome de Marfan, viéndose afectado el tejido conectivo(https://www.nhlbi.nih.gov/es/salud/sindrome-demarfan/causas#:~:text=El%20gen%20FBN1%20produce%20fibrilina,a%20controlar%20c%C3%B3mo%20se%20crece). Actualmente en España, la prevalencia de Síndrome de Marfan es de 3 personas por cada 10.000 habitantes. Este dato lo convierte en una de las más frecuentes dentro de las llamadas Enfermedades Raras, lo que también resulta interesante analizar en profundidad. (https://ec.europa.eu/health/archive/ph_threats/non_com/docs/r299_es.pdf).

Otros genes localizados en este cromosoma que parecen diferenciar a la población española son *CEP152*, *DUT*, *LOC100506059*, *MYEF2*, *SHC4*, *SLC12A1* y *SLC24A5*. Mientras que los primeros 5 genes no parecen tener relevancia clínica en nuestra población, las mutaciones en los genes *SLC12A1* y *SLC24A5*, relacionadas con tubulopatías (grupo heterogéneo de entidades definidas por anomalías de la función tubular renal), pueden llegar a originar enfermedad renal crónica, presente en población española en uno de cada 7 adultos (15%), una prevalencia más elevada que la estimada en estudios previos en nuestro país y similar a la observada en Estados Unidos (Gorostidi et al., 2018).

En el cromosoma 16 nos encontramos además con regiones correspondientes a los genes *ABCA17P*, *ABCA3*, *AMDHD2*, *BRICD5*, *CASKIN1*, *DNASE1L2*, *E4F1*, *MLST8*, *NOXO1*, *NTHL1*, *NTN3*, *PGP*, *PKD1*, *RAB26*, *RNPS1*, *SLC9A3R2*, *SYNGR3*, *TBC1D24*, *TRAF7*, *TSC2* y *ZNF598*.

Se puede destacar que la mayoría de las mutaciones en estos genes dan lugar a discapacidad intelectual y sordera ((*BRICD5*, *CASKIN1*, este último implicado en el síndrome X frágil, que es la forma más común de discapacidad intelectual hereditaria, y se estima que la frecuencia en España es de 1 por cada 4.000 varones en la población general, 1 portadora por cada 800 y 1 portador por cada 5.000 nacidos vivos (https://www.ahedysia.org/patologias/127-sindrome-x-fragil), *E4F1*, *NTN3*, *SYNGR3*, *TBC1D24* y *TRAF7*)).

Otros genes destacados en la señal detectada en este cromosoma están relacionados con cáncer, como *NOXO1* y *NTHL1*. Además, también se distingue el gen *NHERF2* (denominado antes *SLC9A3R2*), implicado en el desarrollo de fibrosis quística, enfermedad genética

multisistémica más grave y frecuente de la raza caucásica. En España afecta aproximadamente a 1 de cada 5.000 recién nacidos vivos, mientras que 1 de cada 35 personas son portadoras sanas de la enfermedad (https://fibrosisquistica.org/).

Los resultados muestran las peculiaridades de Galicia y País Vasco – Navarra cuando se utilizan marcadores de baja frecuencia, tal como se muestra en las Tablas 4.5 y 4.6. Esta diferenciación en la clase rara es similar a la caracterizada por DAPC.

En el caso de población gallega, las proteínas codificadas por la familia de genes *TLR* juegan un papel fundamental en el reconocimiento de patógenos y la activación de la inmunidad innata. Los genes *TLR* están muy conservados desde *Drosophila* hasta humanos y comparten similitudes estructurales y funcionales. Reconocen patrones moleculares asociados a patógenos (Iloro & Pampliega) que se expresan en agentes infecciosos y median en la producción de citocinas necesarias para el desarrollo de una inmunidad eficaz. Los diversos *TLR* exhiben diferentes patrones de expresión. Este gen se expresa más en tejidos linfoides como el bazo, los ganglios linfáticos, el timo y las amígdalas.
(https://www.genecards.org/cgi-bin/carddisp.pl?gene=TLR10&keywords=TLR10
https://www.genecards.org/cgi-bin/carddisp.pl?gene=TLR6&keywords=tlr
https://www.genecards.org/cgi-bin/carddisp.pl?gene=TLR1&keywords=tlr1).

Por tanto, según los datos expuestos podemos decir que existe una región clara en el cromosoma 4 (genes *TLR1*, *TLR6* y *TLR10*) que diferencia a nuestra población de estudio "Galicia". En el Anexo III se muestra el mismo análisis con datos imputados en los que se aprecia que esta misma región está presente en población gallega, lo que la diferencia del resto de poblaciones analizadas. Esta diferencia localizada en el cromosoma 4 podría indicar una inmunidad específica de la población gallega frente a determinadas patologías (aunque no hay evidencias científicas previas), ya que *TLR* juega un papel crucial en el inicio y la coordinación de la inmunidad innata antimicobacteriana. Lo que sí demostraron J. G. Ocejo-Vinyals y colaboradores (2013), (Ocejo-Vinyals et al., 2013) en población del norte de España y tras muchos estudios contradictorios (probablemente con resultados inconsistentes debido a la omisión de la subestructura poblacional), es que el marcador rs5743618, ubicado en el cromosoma 4 y en *TLR1*, es protector frente a la tuberculosis en la población mencionada, mientras que ofrece susceptibilidad a la enfermedad en otras poblaciones. Como ya se comentó en la introducción de este trabajo, *The Wellcome Trust Case Control Consortium* demostró en 2007 diferencias de frecuencia de esta variante de norte a sur en el Reino Unido.

En la población País Vasco – Navarra analizada, en el cromosoma 5 se aprecian 7 marcadores, el rs41290605 (G>A), sin relevancia clínica descrita en ClinVar, es una variante intrónica ubicada en el gen *HDAC3*. Las funciones de este gen son similares a las de *HDAC4*, con actividad histona desacetilasa, que reprime la transcripción cuando se une a un promotor. Las enfermedades asociadas con *HDAC3* incluyen el síndrome de Rett y retinoblastoma. Entre sus vías relacionadas se encuentran el reloj circadiano y la activación transcripcional de la biogénesis mitocondrial (https://www.genecards.org/cgi-bin/carddisp.pl?gene=HDAC3&keywords=HDAC3).

Los otros 6 marcadores se encuentran en genes de la familia *PCDHGA*, codificadores de las cadherinas. Todas ellas sin relevancia clínica descrita en ClinVar, este grupo tiene una organización similar a la de las inmunglobulinas, lo que parece sugerir que está involucrado un mecanismo novedoso en su regulación y expresión. Estas proteínas de adhesión celular, parecidas a las cadherinas neurales, probablemente desempeñen un papel crítico en el establecimiento y la función de conexiones específicas entre células en el cerebro. Las

enfermedades asociadas con *PCDHGA* incluyen trastorno del desarrollo neurológico con crecimiento deficiente y anomalías esqueléticas (https://www.genecards.org/Search/Keyword?queryString=PCDHGA).

En el cromosoma 17 también se distingue una señal, que significativamente comprende 3 marcadores situados muy cerca, y en los siguientes genes:

PTGES3L. Se prevé que esté activo en el citosol y el núcleo. Parece que participa en el ensamblaje del complejo de proteínas mediado por chaperonas y el plegamiento de proteínas. Las enfermedades asociadas con *PTGES3L* incluyen el síndrome de sinostosis espondilocarpotarsal y la deficiencia del complejo mitocondrial IV (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTGES3L&keywords=PTGES3L).
Variante rs117770911 (C>T). Sin relevancia clínica descrita en ClinVar, se trata de un biomarcador intrónico del gen *STAT5B*. La proteína codificada por este gen es miembro de la familia *STAT* de factores de transcripción. En respuesta a las citoquinas y los factores de crecimiento, los miembros de la familia *STAT* son fosforilados por las quinasas asociadas al receptor y luego forman homo- o heterodímeros que se translocan al núcleo celular donde actúan como activadores de la transcripción. Esta proteína media la transducción de señales desencadenada por varios ligandos celulares, como IL2 (*interleukin 2*), IL4 (*interleukin 4*), CSF1 (*Colony stimulating factor* 1) y diferentes hormonas de crecimiento. Se ha demostrado que este gen está involucrado en diversos procesos biológicos, como la apoptosis, el desarrollo de la glándula mamaria adulta y el dimorfismo sexual de la expresión génica del hígado. Se descubrió que este gen se fusiona con el gen del receptor alfa del ácido retinoico en un pequeño subconjunto de leucemias promielocíticas agudas. Las enfermedades asociadas con *STAT5B* incluyen el síndrome de insensibilidad a la hormona del crecimiento con desregulación inmunitaria (https://www.genecards.org/cgi-bin/carddisp.pl?gene=STAT5B&keywords=STAT5B).

PSMC3IP: este gen codifica una proteína que funciona en la recombinación meiótica. Es una subunidad del complejo *PSMC3IP/MND1*, que interactúa con *PSMC3/TBP1* para estimular el intercambio de hebras mediado por *DMC1* y *RAD51* durante la meiosis. La proteína codificada por este gen también puede coactivar la transcripción impulsada por ligandos mediada por receptores nucleares de estrógenos, andrógenos, glucocorticoides, progesterona y tiroides. Las mutaciones en este gen causan disgenesia gonadal femenina XX. El *splicing* alternativo de este gen da como resultado múltiples variantes de transcripción. (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PSMC3IP).

Como podemos ver en los datos expuestos se distinguen varias regiones en distintos cromosomas que diferencian a nuestra población de estudio “País Vasco_Navarra”, principalmente en el cromosoma 5 y concretamente la familia de genes *PCDHGA*.

No parece haber evidencias científicas previas acerca de que las frecuencias de las variantes ubicadas en *PCDHGA* sean diferentes en País Vasco_Navarra al resto de la población española y que, por tanto, discriminen a la población. Tampoco las patologías asociadas a este gen parecen presentar mayor prevalencia y/o incidencia en esta región.

En cualquier caso, la literatura científica nos ofrece trabajos referentes a la singularidad genética de la población vasca. En uno de los más recientes, Flores-Bello y colaboradores (Flores-Bello et al., 2021), llevaron a cabo el mayor estudio realizado con muestras de ADN, que comprendió 1.970 individuos actuales y antiguos, el muestreo geográfico más exhaustivo hasta el momento de la población vasca, con más de 600.000 marcadores genéticos a lo largo de todo el genoma para cada individuo, pudiendo confirmar la "singularidad genética vasca". Esta la explican por sus largos "periodos de aislamiento" y su "falta de flujo genético". "La barrera

cultural del lenguaje pudo promover el aislamiento de la población vasca ante los contactos poblacionales posteriores".

Este estudio también permite a los investigadores postular que los dialectos del euskera "pueden haber surgido mucho antes de la Edad de Hierro, y por eso se relacionan con la estructura genética".".

La diferenciación de País Vasco, que con Galicia se muestran también en los extremos de la península (Bycroft et al., 2019), parece que no se puede atribuir a la prehistoria, sino a eventos recientes de aislamiento genético.

En este estudio, primer mapa genético de España publicado (Bycroft et al., 2019), los investigadores Ángel Carracedo, Simon Myers, Peter Donnelly y Clare Bycroft pusieron de manifiesto que los movimientos poblacionales que se produjeron entre los siglos IX y XIII se tradujeron en un mapa genético que se parece a cómo era la división política y lingüística del siglo XIV.

Respecto a Galicia destacan que existe mucha continuidad genética con Portugal, seguramente fruto de muchos movimientos sur-norte a lo largo de la historia, y muchísima subestructura genética que no ocurre en el resto de España. Estas diferencias genéticas observadas a pequeña escala en Galicia podrían explicar por qué algunas enfermedades con componente genético parecen concentrarse en áreas concretas. En este trabajo destacan que el componente genético norteafricano más alto se observa en Portugal y en Galicia. El equipo de investigadores estima que este ADN dataría de alrededor de los años 860-1120, un período de migración de población procedente de la actual Marruecos durante el dominio musulmán.

Nuestros resultados podrían apoyar la idea general de un patrón de estratificación diferente en los marcadores más raros (Babron et al., 2012; Mathieson & McVean, 2012), pero también podría estar relacionado con nuestro pequeño tamaño de muestra, que es un problema más fuerte en las clases de frecuencias más bajas donde la diferenciación podría ser causada por error de muestreo. En cualquier caso, la magnitud de nuestra correlación entre las clases de marcadores es más baja que la encontrada en el Reino Unido (Babron et al., 2012) y Europa (Heath et al., 2008). Se necesitan estudios adicionales (la base de datos de genotipado de población control española continúa en aumento), con un tamaño de muestra más alto, para confirmar si estas correlaciones bajas (pero significativas) se deben al uso de una escasa N.

Probamos el enfoque multivariante aplicando análisis discriminante de componentes principales (DAPC). Es importante recordar que DAPC crea variables sintéticas como combinaciones lineales de las variables originales (alelos), que representan la mayor variación entre nuestras poblaciones predefinidas y las representa gráficamente. Hay que tener en cuenta que la capacidad de discriminación de este enfoque está "inflada" porque hemos incluido una predefinición de los clústeres (nuestras poblaciones -POP-) que deben ser discriminados. Cuando el *software* utiliza una agrupación no predefinida (mediante un procedimiento de *K-means* para encontrar agrupaciones que se definen en el paquete), la asociación entre los grupos encontrados y las poblaciones reales no es muy correcta.

Por lo tanto, cuando usamos DAPC, pretendíamos comparar el patrón de discriminación entre las clases de marcadores e inferir los más informativos en cada análisis. Si bien es cierto que la discriminación entre las poblaciones prefijadas es más fuerte cuando se usan marcadores de menor frecuencia (Figuras 4.2, 4.3, 4.4 y 4.5), parte de esta discriminación puede haber sido provocada por variantes que están presentes en una sola población, especialmente variantes raras.

Debemos tener en cuenta que el número de PCs retenidas viene recomendado por el propio *software* y es específico para cada análisis, aumentando gradualmente en número a medida que

consideramos marcadores con menor MAF. Por lo tanto, no se debe olvidar que puede haber algún sesgo en estas representaciones gráficas.
De forma similar a lo encontrado en PCA, todos los LD muestran una alta y significativa correlación entre ellos. Una vez más, la población gallega y la de País Vasco - Navarra se aprecian más discriminadas en el análisis de variantes raras.

En general, los marcadores más importantes en la discriminación de una población también se encontraron como significativos cuando esta población se comparó con el resto mediante un procedimiento más directo (de manera similar a un estudio de asociación de casos y controles). De hecho, una selección muy simple de marcadores en función de la probabilidad en cada uno de los análisis individuales fue suficiente para reproducir el patrón de discriminación (sutil) del conjunto.

En definitiva, deben llevarse a cabo más análisis sobre la subestructura diferencial de marcadores de alta y baja frecuencia, aunque en cierto modo hemos demostrado que las variantes raras se comportan de manera diferente, describiendo patrones específicos de estructura de la población para estos marcadores.

Los estudios adicionales con un tamaño de muestra mayor para poblaciones españolas y europeas nos permitirán definir el uso y las limitaciones de los marcadores codificantes y no codificantes en la diferenciación geográfica, un tema muy interesante en estudios de epidemiología genética entre otros. La importancia sugerida de variantes raras en la discriminación de algunas de nuestras poblaciones señala la necesidad de estudios más exhaustivos sobre la presencia y distribución de variantes raras en población española.

En este sentido, se ha considerado aplicar diferentes métodos para probar la subestructura provocada por distintas variantes genéticas (incluyendo variantes raras y de baja frecuencia), considerando no solo las posiciones, sino regiones genómicas completas que podrían ir acumulando pequeños efectos de muchos marcadores únicos en la estructura de las poblaciones.

Podrían considerarse, para esta tarea, enfoques como ChromoPainter y FineSTRUCTURE, pero deberían hacerse modificaciones de algunos parámetros para analizar un número más pequeño de marcadores, ya que *softwers* como estos son altamente dependientes del desequilibrio de ligamiento. Simplemente podríamos eliminar aquellos marcadores que se encuentran en esta situación tal como hemos hecho en nuestros análisis. A pesar del modesto tamaño de nuestra muestra, pensamos que algunos de los resultados de los enfoques multivariantes que utilizamos, como DAPC, también podrían utilizarse para analizar y comparar genes o secciones cromosómicas definidas por diferentes clases de marcadores.

A la hora de llevar a cabo este tipo de análisis multivariantes se debe tener en cuenta que la representación de LD1 y LD2 (o como máximo también LD3 y LD4) puede no ser suficiente, ya que pueden no ser mucho más representativos de la variabilidad total que LD9 o LD10, por ejemplo. Por lo tanto, la representación gráfica de individuos basada en LD1 y LD2 puede no mostrar una clara diferenciación geográfica de los clústeres.

En este trabajo hemos confirmado la existencia de diferencias (alguna importante) en la frecuencia de marcadores respecto a bases de datos de referencia. La evaluación de las diferencias excesivamente grandes (cambios drásticos de categoría SNP-rara) puede contribuir a la depuración y rediseño del *array* (marcadores con excesiva diferencia debido a mal genotipado) en posteriores versiones.

La existencia de diferencias creíbles en MAF confirma la importancia de diseñar este tipo de *arrays* específicos de población, para una mejor caracterización poblacional e incorporación de dicha variabilidad en los estudios GWA. Tras los resultados obtenidos en los estudios de GWA parece interesante analizar la lista de variantes más importantes para detectar posibles superposiciones.

Actualmente se están analizando en detalle los marcadores que han cambiado de categoría con respecto al UKBB array para poder identificar qué marcadores son los que más contribuyen a la discriminación en cada caso, y estamos evaluando su distribución a lo largo del genoma. De este modo este *array* ha sido depurado y hemos generado una versión 2 del mismo.

5.6 CAMINO HACIA LA MEDICINA PERSONALIZADA

Uno de los desafíos centrales en la genética humana moderna es desentrañar la base genética de las enfermedades humanas y otros fenotipos de interés. Además del interés inherente en comprensión de los determinantes biológicos de los distintos fenotipos, se espera que este trabajo conduzca a importantes avances médicos. En particular, la determinación de las variantes genéticas involucradas en una enfermedad particular debe proporcionar nuevo conocimiento sobre la etiología de la enfermedad, puede sugerir nuevos objetivos farmacéuticos, y podría conducir a la detección genética para identificar a las personas con mayor riesgo.

El tipo de análisis llevado a cabo en este trabajo podría contribuir a concretar la etiología de las enfermedades en poblaciones específicas. Es muy importante tener en cuenta que es necesaria la combinación de los datos genéticos con bases de datos clínicos como parte de un sistema integrado de salud, y bases de datos que reflejen el efecto fenotípico de la variación genética. Objetivo: medicina personalizada.

Los *arrays* específicos de población y dirigidos al estudio de marcadores farmacogenéticos, son clave para el correcto funcionamiento de esta disciplina. A pesar de que la farmacogenética para nada es una idea nueva (en 1959 Friederich Vogel ya la definió como la "variación hereditaria de importancia clínica en la respuesta a los fármacos" y el mismo Vogel, junto con Motulsky, sentaron las bases teóricas de esta nueva disciplina y poco después, en 1962, Werner Kalow sentó las bases de la farmacogenética como ciencia con su monografía "*Pharmacogenetics: Heredity and Response to Drugs*"), fue con el desarrollo de las técnicas de genotipado, técnicas de secuenciación de nueva generación y los análisis de expresión, los que posibilitaron que se encontrasen numerosos ejemplos de genes relacionados con la respuesta a fármacos.

El uso de biomarcadores para medicina personalizada tiene tres fases bien diferenciadas: descubrimiento, validación regulatoria y traslación.

La mayoría de los biomarcadores que conocemos en la actualidad provienen de la genómica, y comprenden tanto variantes en el ADN como marcadores a nivel de ARN (expresión) y más recientemente marcadores epigenéticos; puede tratarse de marcadores tanto en línea germinal como somática, y afectar tanto a la farmacocinética como a la farmacodinámica de los medicamentos.

La integración de los datos (genómica, imagen, microbioma, otras ciencias ómicas y datos clínicos y epidemiológicos) y la bioinformática y la biología de sistemas, están siendo claves en esta fase de identificación de grupos de biomarcadores y su interacción.

La bioinformática se ha convertido rápidamente en un componente importante de la farmacogenómica, proporcionando la capacidad de diseñar y analizar experimentos complejos y también de gestionar la gran cantidad de información recopilada en estos estudios. La farmacogenómica interacciona sobre dos áreas principales de la bioinformática: los recursos disponibles en internet y el uso de la bioinformática aplicada. La naturaleza misma de la arquitectura abierta de la Web y su flexibilidad casi ilimitada la convierten en una plataforma oportuna para acceder e intercambiar información genómica y genética. Las bases de datos masivas mantenidas por instituciones públicas y privadas brindan una amplia gama de

capacidades de extracción de datos (Hansi & Heimpel, 1979). Tras el descubrimiento de un biomarcador este tiene que ser validado. Si se trata de un biomarcador ligado a un fármaco esta labor le corresponde a la Agencia Europea del Medicamento a través de su grupo de trabajo de Farmacogenómica.

La traslación de los biomarcadores descubiertos y validados es el gran reto. Los estudios de coste-eficacia son importantes para los sistemas de salud, hospitales o clínicas, pero son independientes respecto a la validación y su aprobación.

Solo una aplicación correcta en la clínica de lo que ya está validado y aprobado supondría una disminución importante del gasto sanitario. En este reto hay dos puntos claves: la educación de los profesionales en este campo, lo que incluye la formación continuada, y la organización del análisis de biomarcadores a nivel hospitalario, autonómico y nacional, con criterios de equidad y eficacia, y en laboratorios que dispongan de métodos con los que la validez en la aplicación esté garantida.

La Medicina personalizada o de precisión es un camino ya trazado y sin retorno de la medicina moderna. Implica el uso de biomarcadores de estratificación de la enfermedad y de respuesta a fármacos.

Todo lo anterior está en línea con la visión actual de que la recopilación de *Big Data* puede conducir a la medicina de precisión, orientada a satisfacer las necesidades de los pacientes y fortalecer la cooperación paciente / médico en la elección conjunta de la mejor opción terapéutica. Las tres grandes necesidades a cubrir en la medicina personalizada son la identificación del mecanismo molecular de la enfermedad, la disponibilidad de herramientas de diagnóstico y la disponibilidad de un tratamiento capaz de bloquear el mecanismo. En cuanto al diagnóstico, hay que tener en cuenta que la historia clínica recogida por un médico experto es la base para decidir las investigaciones posteriores de acuerdo con las características individuales. Las herramientas actuales y futuras para alcanzar una prescripción personalizada son variables, pero es obvio que la atención personalizada es valiosa y además puede ser preventiva (en cuanto a calidad de vida), predictiva (al permitir ajustar el tratamiento en función de la respuesta del individuo) y participativa (colaboración del paciente). Este enfoque personalizado para el diagnóstico, la toma de decisiones, la elección de productos y programas de tratamiento, pueden mejorar la eficacia, minimizar los eventos adversos, mejorar la calidad de vida del paciente y reducir el impacto socioeconómico. Dependiendo del perfil de sensibilización y de los signos clínicos, el especialista puede desarrollar un abordaje específico. Además, permite adaptar la dosis a las necesidades para lograr una eficacia óptima (Ángel Carracedo, 2022; https://isanidad.com/228154/angel-carracedo-papel-medico-atencion-primaria-va-a-ser-vital-medicina-personalizada/).

En un informe realizado para la Fundación Instituto Roche por Ángel Carracedo y Marina Pollán junto con otros colaboradores expertos, se pone de manifiesto que la incorporación de la información genómica a los modelos de predicción de riesgo, a través de las Estimas de Riesgo Poligénico (PRS, por sus siglas en inglés "*Polygenic Risk Score*), abre nuevos horizontes para el avance de la Medicina Preventiva y la Salud Pública de Precisión.

Los PRS, que son una medida global del riesgo genético de desarrollar una enfermedad por parte de una persona respecto de la población general, permiten incorporar información sobre variantes en las secuencias de los genes asociadas al riesgo de desarrollar enfermedades. Esto podría contribuir a mejorar de manera sustancial la capacidad predictiva de los modelos de predicción de riesgo desarrollados hasta el momento. Así, se espera que estos PRS constituyan una herramienta clave para la predicción de riesgo tanto a nivel poblacional como de manera personalizada, guiando la toma de decisiones clínicas o la selección o priorización de tratamientos en individuos o grupos de pacientes.

Por todo ello, aunque es necesario avanzar en el estudio y validación de los PRS, se prevé que estos tengan un papel relevante en la Medicina del Futuro: (https://www.institutoroche.es/static/archivos/Informes_anticipando_2022_Prediccion_riesgo_DEF.pdf).

Queda un camino largo y desafiante para llegar a tener una comprensión completa de la amplia variación en el genoma humano, pero sabemos que parte de esa comprensión proviene de la genética de poblaciones, que se debe tener en cuenta para mejorar la salud humana.

CONCLUSIONES

6 CONCLUSIONES

En cuanto a los resultados obtenidos tras los objetivos planteados en este trabajo se puede concluir que:

1- Hemos diseñado un *array*: el *Axiom Spain Biobank Array Plate* con la compañía ThermoFisher que permite el análisis de 643.842 variantes comunes y 114.898 de variación rara funcional, recogidos de análisis de exomas de población española, y enriquecido con variación compilada por expertos en las principales áreas, que ha sido depurado y validado con criterios de calidad. El *Axiom Spain Biobank Array Plate* ha resultado óptimo para llevar a cabo estudios de GWA analizando y captando variación común, así como variación rara funcional específica de población española, y está siendo utilizado para estudios de asociación de genoma completo por diversos grupos de investigación a través del Centro Nacional de Genotipado-FPGMX.

2- La obtención de genotipos de alta calidad y las herramientas y filtros aplicados para llevar a cabo los controles de calidad tanto de los marcadores como de las muestras, han permitido comprobar que las variantes infrecuentes y especialmente las variantes funcionales, son muy dependientes de la población de origen.

3- En cuanto al análisis de la existencia de variabilidad local a escala microgeográfica y a la caracterización de los diferentes patrones de estratificación poblacional en variantes comunes y raras para el estudio de trastornos genéticos complejos, llevado a cabo con el Análisis Discriminante de Componentes Principales (DAPC), hemos podido ver que los marcadores que más discriminan a escala local son las variantes de baja frecuencia, sobre todo teniendo en cuenta LD1 y LD2. La prueba de asociación entre las poblaciones consideradas produjo una serie de marcadores que destacan en diferenciar cada una de las poblaciones del resto. Galicia (variación rara) y País Vasco-Navarra (SNPs y marcadores de baja frecuencia), superaron claramente al resto, aunque Cantabria y Cataluña también están entre las poblaciones que obtuvieron valores más altos. Esta variabilidad parece no estar sujeta a efecto de muestreo.

4- En población gallega se diferencia especialmente una región correspondiente a los genes *TLR* ubicados en el cromosoma 4 y en la población País Vasco – Navarra destaca el gen *PCDHGA*, en el cromosoma 5. Esto parece sugerir que pueden existir diferencias en estas dos subpoblaciones con respecto al resto de población española en cuanto a determinados fenotipos que, en el caso de la población gallega podría indicar una inmunidad específica frente a determinadas patologías de carácter infeccioso.

5- La existencia de patrones de estratificación diferentes apunta la necesidad de aplicar los métodos de corrección diseñados y generalmente aplicados a estudios de GWAS. Por su parte, las variantes raras responsables de la diferenciación poblacional, aparecieron diseminadas por todo el genoma, lo que estaría acorde con su aparición más reciente, de forma que incluso aquellas variantes potencialmente deletéreas pueden no haber sido eliminadas aún por selección natural.

6- La heterogeneidad geográfica en la arquitectura genética de la enfermedad, es decir, el hecho de que diferentes variantes, a menudo en diferentes genes, estén implicadas en la predisposición a la enfermedad en diferentes poblaciones (Fernandez-Rozadilla et al., 2013) puede ser más frecuente de lo esperado, destacando nuevamente la necesidad de la creación de "catálogos" de variaciones locales (Bustamante, Burchard, & De la Vega, 2011), (Dopazo et al., 2016).

7- Con los datos de genotipado obtenidos hemos generado un repositorio de datos de genotipado de muestras de población control española que ya es accesible a la comunidad científica.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

Abdellaoui, A., Hottenga, J. J., de Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., . . . Boomsma, D. I. (2013). Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet, 21*(11), 1277-1285. doi: 10.1038/ejhg.2013.48

Adeyemo, A., & Rotimi, C. (2010). Genetic variants associated with complex human diseases show wide variation across multiple populations. *Public Health Genomics, 13*(2), 72-79. doi: 10.1159/000218711

Ahmed, M., Dorling, L., Kerns, S., Fachal, L., Elliott, R., Partliament, M., . . . West, C. M. (2016). Common genetic variation associated with increased susceptibility to prostate cancer does not increase risk of radiotherapy toxicity. *Br J Cancer, 114*(10), 1165-1174. doi: 10.1038/bjc.2016.94

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat Rev Genet, 16*(4), 197-212. doi: 10.1038/nrg3891

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664. doi: 10.1101/gr.094052.109

Associating., F. (1999). Freely associating. *Nat Genet, 22*(1), 1-2. doi: 10.1038/8702

Attwood, T. P.-S., David. (1999). *Introduction to Bioinformatics*

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296. doi: 10.1093/bioinformatics/btm108

Babron, M. C., de Tayrac, M., Rutledge, D. N., Zeggini, E., & Genin, E. (2012). Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One, 7*(10), e46519. doi: 10.1371/journal.pone.0046519

Bacolod, M. D., Schemmann, G. S., Giardina, S. F., Paty, P., Notterman, D. A., & Barany, F. (2009). Emerging paradigms in cancer genetics: some important findings from high-density single nucleotide polymorphism array studies. *Cancer Res, 69*(3), 723-727. doi: 10.1158/0008-5472.CAN-08-3543

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M. K., Chuang, R., Jaehnig, E. J., . . . Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Science, 330*(6009), 1385-1389. doi: 10.1126/science.1195618

Baran, Y., Quintela, I., Carracedo, A., Pasaniuc, B., & Halperin, E. (2013). Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am J Hum Genet, 92*(6), 882-894. doi: 10.1016/j.ajhg.2013.04.023

Barnett, G. C., Thompson, D., Fachal, L., Kerns, S., Talbot, C., Elliott, R. M., . . . West, C. M. (2014). A genome wide association study (GWAS) providing evidence of an association between common genetic variants and late radiotherapy toxicity. *Radiother Oncol, 111*(2), 178-185. doi: 10.1016/j.radonc.2014.02.012

Bartha, I., di Iulio, J., Venter, J. C., & Telenti, A. (2018). Human gene essentiality. *Nat Rev Genet, 19*(1), 51-62. doi: 10.1038/nrg.2017.75

Bauer, A. E., Avery, C. L., Shi, M., Weinberg, C. R., Olshan, A. F., Harmon, Q. E., . . . Engel, S. M. (2018). A Family Based Study of Carbon Monoxide and Nitric Oxide Signalling Genes and Preeclampsia. *Paediatric and perinatal epidemiology, 32*(1), 1-12. doi: 10.1111/ppe.12400

Beck, T., Rowlands, T., Shorter, T., & Brookes, A. J. (2022). GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res, 51*(D1), D986-D993. doi: 10.1093/nar/gkac1017 %J Nucleic Acids Research

Beck, T., Rowlands, T., Shorter, T., & Brookes, A. J. (2023). GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res, 51*(D1), D986-D993. doi: 10.1093/nar/gkac1017

Bergen, S. E., & Petryshen, T. L. (2012). Genome-wide association studies of schizophrenia: does bigger lead to better results? *Curr Opin Psychiatry, 25*(2), 76-82. doi: 10.1097/YCO.0b013e32835035dd

Bingham, S., & Riboli, E. (2004). Diet and cancer--the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer, 4*(3), 206-215. doi: 10.1038/nrc1298

Bishop, D. T. (1994). Fundamentals of Genetic Epidemiology. *J Med Genet, 31*(11), 900-900.

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet, 40*(6), 695-701. doi: 10.1038/ng.f.136

Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol, 18*(1), 77. doi: 10.1186/s13059-017-1212-4

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell, 169*(7), 1177-1186. doi: 10.1016/j.cell.2017.05.038

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet, 81*(5), 1084-1097. doi: 10.1086/521987

Browning, S. R., & Weir, B. S. (2010). Population structure with localized haplotype clusters. *Genetics, 185*(4), 1337-1344. doi: 10.1534/genetics.110.116681

Buchanan, C. C., Torstenson, E. S., Bush, W. S., & Ritchie, M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J Am Med Inform Assoc, 19*(2), 289-294. doi: 10.1136/amiajnl-2011-000652

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., . . . Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nat Genet, 47*(11), 1236-1241. doi: 10.1038/ng.3406

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol, 8*(12), e1002822. doi: 10.1371/journal.pcbi.1002822

Bustamante, C. D., Burchard, E. G., & De la Vega, F. M. (2011). Genomics for the world. *Nature, 475*(7355), 163-165. doi: 10.1038/475163a

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, A., Donnelly, P., & Myers, S. (2019). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun, 10*(1), 551. doi: 10.1038/s41467-018-08272-w

Cade, B. E., Chen, H., Stilp, A. M., Gleason, K. J., Sofer, T., Ancoli-Israel, S., . . . Redline, S. (2016). Genetic Associations with Obstructive Sleep Apnea Traits in Hispanic/Latino Americans. *Am J Respir Crit Care Med, 194*(7), 886-897. doi: 10.1164/rccm.201512-2431OC

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., . . . Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science, 296*(5566), 261-262.

Cardon, L. R., & Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. *Trends Genet, 19*(3), 135-140. doi: 10.1016/S0168-9525(03)00022-2

Carlborg, Ö., & Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics, 5*, 618. doi: 10.1038/nrg1407

Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature, 429*(6990), 446-452. doi: 10.1038/nature02623

Carvalho, B., Bengtsson, H., Speed, T. P., & Irizarry, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics, 8*(2), 485-499. doi: 10.1093/biostatistics/kxl042

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics, 6*, 333. doi: 10.1038/nrg1596

Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., & Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis, 20*(8), 1682-1696. doi: 10.1002/(SICI)1522-2683(19990101)20:8<1682::AID-ELPS1682>3.0.CO;2-Z

Chan, S. L., Jin, S., Loh, M., & Brunham, L. R. (2015). Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. *Pharmacogenomics, 16*(10), 1161-1178. doi: 10.2217/pgs.15.54

Chandak, G. R., Janipalli, C. S., Bhaskar, S., Kulkarni, S. R., Mohankrishna, P., Hattersley, A. T., . . . Yajnik, C. S. (2007). Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population. *Diabetologia, 50*(1), 63-67. doi: 10.1007/s00125-006-0502-2

Chang, D., Nalls, M. A., Hallgrimsdottir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., . . . Graham, R. R. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet, 49*(10), 1511-1516. doi: 10.1038/ng.3955

Clark, A. G., Boerwinkle, E., Hixson, J., & Sing, C. F. (2005). Determinants of the success of whole-genome association testing. *Genome Res, 15*(11), 1463-1467. doi: 10.1101/gr.4244005

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., & Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science, 305*(5685), 869-872. doi: 10.1126/science.1099870

Colhoun, H. M., McKeigue, P. M., & Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet, 361*(9360), 865-872.

Companioni, O., Rodriguez Esparragon, F., Fernandez-Aceituno, A. M., & Rodriguez Perez, J. C. (2011). [Genetic variants, cardiovascular risk and genome-wide association studies]. *Rev Esp Cardiol, 64*(6), 509-514. doi: 10.1016/j.recesp.2011.01.010

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet, 38*(11), 1251-1260. doi: 10.1038/ng1911

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi: 10.1038/nature11247

Consortium, U. K. P. s. D., Wellcome Trust Case Control, C., Spencer, C. C., Plagnol, V., Strange, A., Gardner, M., . . . Wood, N. W. (2011). Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum Mol Genet, 20*(2), 345-353. doi: 10.1093/hmg/ddq469

Cooper, D. N., Ball, E. V., & Krawczak, M. (1998). The human gene mutation database. *Nucleic Acids Res, 26*(1), 285-287. doi: 10.1093/nar/26.1.285

Costanzo, M., Kuzmin, E., van Leeuwen, J., Mair, B., Moffat, J., Boone, C., & Andrews, B. (2019). Global Genetic Networks and the Genotype-to-Phenotype Relationship. *Cell, 177*(1), 85-100. doi: 10.1016/j.cell.2019.01.033

Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., . . . Consortia, F. O. N. o. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet, 7*(8), e1002254. doi: 10.1371/journal.pgen.1002254

Cotton, R. G., Auerbach, A. D., Beckmann, J. S., Blumenfeld, O. O., Brookes, A. J., Brown, A. F., . . . den Dunnen, J. T. (2008). Recommendations for locus-specific databases and their curation. *Hum Mutat, 29*(1), 2-5. doi: 10.1002/humu.20650

Cotton, R. G., Auerbach, A. D., Brown, A. F., Carrera, P., Christodoulou, J., Claustres, M., . . . Human Variome Project Diagnostic Laboratory Working, G. (2007). A structured simple form for ordering genetic tests is needed to ensure coupling of clinical detail (phenotype) with DNA variants (genotype) to ensure utility in publication and databases. *Hum Mutat, 28*(10), 931-932. doi: 10.1002/humu.20631

Cotton, R. G., Human Variome, P., Appelbe, W., Auerbach, A. D., Becker, K., Bodmer, W., . . . Watson, M. (2007). Recommendations of the 2006 Human Variome Project meeting. *Nat Genet, 39*(4), 433-436. doi: 10.1038/ng2024

Cotton, R. G. H. K. J., Haig H. . (2005). Toward a Human Variome Project. *Hum Mutat*.

Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., . . . Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun, 1*, 131. doi: 10.1038/ncomms1130

Cristianini, N. H., M. (2006). *Introduction to Computational Genomics: A Case Studies Approach.* . Cambridge: Cambridge University Press. .

Cross-Disorder Group of the Psychiatric Genomics, C., Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., . . . International Inflammatory Bowel Disease Genetics, C. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet, 45*(9), 984-994. doi: 10.1038/ng.2711

Cruz, R., Diz-de Almeida, S., Lopez de Heredia, M., Quintela, I., Ceballos, F. C., Pita, G., . . . Carracedo, A. (2022). Novel genes and sex differences in COVID-19 severity. *Hum Mol Genet, 31*(22), 3789-3806. doi: 10.1093/hmg/ddac132

Daly, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics, 11*, 241. doi: 10.1038/nrg2751

Davis, O. S. P., Band, G., Pirinen, M., Haworth, C. M. A., Meaburn, E. L., Kovas, Y., . . . Spencer, C. C. A. (2014). The correlation between reading and mathematics ability at age twelve has a substantial genetic component. *Nat Commun, 5*, 4204. doi: 10.1038/ncomms5204

de With, M., Sadlon, A., Cecchin, E., Haufroid, V., Thomas, F., Joerger, M., . . . The Working Group on the Implementation of, D. P. D. d. T. i. E. (2023). Implementation of dihydropyrimidine dehydrogenase deficiency testing in Europe. *ESMO Open, 8*(2), 101197. doi: 10.1016/j.esmoop.2023.101197

Deak, J. D., Zhou, H., Galimberti, M., Levey, D. F., Wendt, F. R., Sanchez-Roige, S., . . . Gelernter, J. (2022). Genome-wide association study in individuals of European and African ancestry and multi-trait analysis of opioid use disorder identifies 19 independent genome-wide significant risk loci. *Mol Psychiatry, 27*(10), 3970-3979. doi: 10.1038/s41380-022-01709-1

Delaneau, O., Marchini, J., & Zagury, J. F. (2011). A linear complexity phasing method for thousands of genomes. *Nat Methods, 9*(2), 179-181. doi: 10.1038/nmeth.1785

Devlin, B., Roeder, K., & Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol, 60*(3), 155-166. doi: 10.1006/tpbi.2001.1542

Diabetes Genetics Initiative of Broad Institute of, H., Mit, L. U., Novartis Institutes of BioMedical, R., Saxena, R., Voight, B. F., Lyssenko, V., . . . Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science, 316*(5829), 1331-1336. doi: 10.1126/science.1142358

Distefano, J. K., & Taverna, D. M. (2011). Technological issues and experimental design of gene association studies. *Methods Mol Biol, 700*, 3-16. doi: 10.1007/978-1-61737-954-3_1

Dong, S., Wang, E., Hsie, L., Cao, Y., Chen, X., & Gingeras, T. R. (2001). Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res, 11*(8), 1418-1424. doi: 10.1101/gr.171101

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature, 456*(7223), 728-731. doi: 10.1038/nature07631

Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Aleman, A., Garcia-Garcia, F., . . . Antinolo, G. (2016). 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Mol Biol Evol, 33*(5), 1205-1218. doi: 10.1093/molbev/msw005

Dumancas, G. G., Rachal, M., Zamora, P. R. F. C., & de Castro, R. (2022). Examining Barriers and Opportunities of Conducting Genome-Wide Association Studies in Developing Countries. *Current Epidemiology Reports, 9*(4), 376-386. doi: 10.1007/s40471-022-00303-x

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., . . . Ponder, B. A. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature, 447*(7148), 1087-1093. doi: 10.1038/nature05887

Ellinghaus, D., Jostins, L., Spain, S. L., Cortes, A., Bethune, J., Han, B., . . . Franke, A. (2016). Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet, 48*(5), 510-518. doi: 10.1038/ng.3528

Esparza-Castro, D., Andrade-Ancira, F. J., Merelo-Arias, C. A., Cruz, M., & Valladares-Salgado, A. (2015). [Genome-wide association in type 2 diabetes and its clinical application]. *Rev Med Inst Mex Seguro Soc, 53*(5), 592-599.

Evans, D. S., Avery, C. L., Nalls, M. A., Li, G., Barnard, J., Smith, E. N., . . . Sotoodehnia, N. (2016). Fine-mapping, novel loci identification, and SNP association transferability in a genome-wide association study of QRS duration in African Americans. *Hum Mol Genet, 25*(19), 4350-4368. doi: 10.1093/hmg/ddw284

Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour, 10*(3), 564-567. doi: 10.1111/j.1755-0998.2010.02847.x

Fachal, L., Gomez-Caamano, A., Barnett, G. C., Peleteiro, P., Carballo, A. M., Calvo-Crespo, P., . . . Vega, A. (2014). A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1. *Nat Genet, 46*(8), 891-894. doi: 10.1038/ng.3020

Fan, H. C., Wang, J., Potanina, A., & Quake, S. R. (2011). Whole-genome molecular haplotyping of single cells. *Nat Biotechnol, 29*(1), 51-57. doi: 10.1038/nbt.1739

Fan, J. B., Gunderson, K. L., Bibikova, M., Yeakley, J. M., Chen, J., Wickham Garcia, E., . . . Barker, D. (2006). Illumina universal bead arrays. *Methods Enzymol, 410*, 57-73. doi: 10.1016/S0076-6879(06)10003-8

Fearnhead, N. S., Winney, B., & Bodmer, W. F. (2005). Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. *Cell Cycle, 4*(4), 521-525. doi: 10.4161/cc.4.4.1591

Fernandez-Rozadilla, C., Cazier, J. B., Tomlinson, I. P., Carvajal-Carmona, L. G., Palles, C., Lamas, M. J., . . . Ruiz-Ponte, C. (2013). A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics, 14*, 55. doi: 10.1186/1471-2164-14-55

Feulner, T. M., Laws, S. M., Friedrich, P., Wagenpfeil, S., Wurst, S. H., Riehle, C., . . . Riemenschneider, M. (2010). Examination of the current top candidate genes for AD in a genome-wide association study. *Mol Psychiatry, 15*(7), 756-766. doi: 10.1038/mp.2008.141

Flores-Bello, A., Bauduer, F., Salaberria, J., Oyharcabal, B., Calafell, F., Bertranpetit, J., . . . Comas, D. (2021). Genetic origins, singularity, and heterogeneity of Basques. *Curr Biol, 31*(10), 2167-2177 e2164. doi: 10.1016/j.cub.2021.03.010

Foo, J. N., Tan, L. C., Irwan, I. D., Au, W. L., Low, H. Q., Prakash, K. M., . . . Tan, E. K. (2017). Genome-wide association study of Parkinson's disease in East Asians. *Hum Mol Genet, 26*(1), 226-232. doi: 10.1093/hmg/ddw379

Fradin, D. D., & Fallin, M. D. (2009). Influence of control selection in genome-wide association studies: the example of diabetes in the Framingham Heart Study. *BMC Proc, 3 Suppl 7*, S113.

Frank, L. (2000). Give and Take—Estonia's New Model for a National Gene Bank

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., . . . Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet, 42*(12), 1118-1125. doi: 10.1038/ng.717

Freimer, N. B., & Sabatti, C. (2007). Human genetics: variants in common diseases. *Nature, 445*(7130), 828-830. doi: 10.1038/nature05568

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K., & Sander, J. D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol, 31*(9), 822-826. doi: 10.1038/nbt.2623

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., . . . Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science, 296*(5576), 2225-2229. doi: 10.1126/science.1069424

Garcia-Etxebarria, K., Bracho, M. A., Galan, J. C., Pumarola, T., Castilla, J., Ortiz de Lejarazu, R., . . . Controls in Pandemic Influenza Working, G. (2015). No Major Host Genetic Risk Factor Contributed to A(H1N1)2009 Influenza Severity. *PLoS One, 10*(9), e0135983. doi: 10.1371/journal.pone.0135983

Gattepaille, L. M., & Jakobsson, M. (2012). Combining markers into haplotypes can improve population structure inference. *Genetics, 190*(1), 159-174. doi: 10.1534/genetics.111.131136

Genetic Analysis of Psoriasis, C., & the Wellcome Trust Case Control, C. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat Genet, 42*, 985. doi: 10.1038/ng.694

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi: 10.1038/nature09534

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., . . . Easton, D. F. (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet, 44*, 312. doi: 10.1038/ng.1049

Glessner, J. T., Li, J., Liu, Y., Khan, M., Chang, X., Sleiman, P. M. A., & Hakonarson, H. (2023). ParseCNV2: efficient sequencing tool for copy number variation genome-wide association studies. *European Journal of Human Genetics, 31*(3), 304-312. doi: 10.1038/s41431-022-01222-7

Goldstein, D. B. (2009). Common genetic variation and human traits. *N Engl J Med, 360*(17), 1696-1698. doi: 10.1056/NEJMp0806284

Gonzalez, J. R., Estevez, M. N., Giralt, P. S., Caceres, A., Perez, L. M., Gonzalez-Carpio, M., . . . Rodriguez-Lopez, R. (2014). Genetic risk profiles for a childhood with severe overweight. *Pediatr Obes, 9*(4), 272-280. doi: 10.1111/j.2047-6310.2013.00166.x

Gorlov, I. P., Meyer, P., Liloglou, T., Myles, J., Boettger, M. B., Cassidy, A., . . . Amos, C. I. (2007). Seizure 6-like (SEZ6L) gene and risk for lung cancer. *Cancer Res, 67*(17), 8406-8411. doi: 10.1158/0008-5472.CAN-06-4784

Gorostidi, M., Sanchez-Martinez, M., Ruilope, L. M., Graciani, A., de la Cruz, J. J., Santamaria, R., . . . Banegas, J. R. (2018). Chronic kidney disease in Spain: Prevalence and impact of accumulation of cardiovascular risk factors. *Nefrologia (Engl Ed), 38*(6), 606-615. doi: 10.1016/j.nefro.2018.04.004

Grant, S. F., & Hakonarson, H. (2009). Genome-wide association studies in type 1 diabetes. *Curr Diab Rep, 9*(2), 157-163.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., . . . Bustamante, C. D. (2011). Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A, 108*(29), 11983-11988. doi: 10.1073/pnas.1019276108

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., . . . Open Regulatory Annotation, C. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res, 36*(Database issue), D107-113. doi: 10.1093/nar/gkm967

Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., Helgason, A., . . . Stefansson, K. (2007). Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet, 39*, 631. doi: 10.1038/ng1999

Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J. T., Thorleifsson, G., Manolescu, A., . . . Stefansson, K. (2007). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet, 39*(8), 977-983. doi: 10.1038/ng2062

Guenole, A., Srivas, R., Vreeken, K., Wang, Z. Z., Wang, S., Krogan, N. J., . . . van Attikum, H. (2013). Dissection of DNA damage responses using multiconditional genetic interaction maps. *Mol Cell, 49*(2), 346-358. doi: 10.1016/j.molcel.2012.11.023

Gunderson, K. L. (2009). Whole-genome genotyping on bead arrays. *Methods Mol Biol, 529*, 197-213. doi: 10.1007/978-1-59745-538-1_13

Gunderson, K. L., Steemers, F. J., Ren, H., Ng, P., Zhou, L., Tsan, C., . . . Shen, R. (2006). Whole-genome genotyping. *Methods Enzymol, 410*, 359-376. doi: 10.1016/S0076-6879(06)10017-8

Hageman, G. S., Anderson, D. H., Johnson, L. V., Hancox, L. S., Taiber, A. J., Hardisty, L. I., . . . Allikmets, R. (2005). A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc Natl Acad Sci U S A, 102*(20), 7227-7232. doi: 10.1073/pnas.0501536102

Hakonarson, H., & Grant, S. F. (2009). Genome-wide association studies in type 1 diabetes, inflammatory bowel disease and other immune-mediated disorders. *Semin Immunol, 21*(6), 355-362. doi: 10.1016/j.smim.2009.06.001

Hakonarson, H., & Grant, S. F. A. (2011). Planning a genome-wide association study: Points to consider. *Annals of Medicine, 43*(6), 451-460. doi: 10.3109/07853890.2011.573803

Han, J. K., M. (2001). *Data mining: concepts and techniques* (M. Kaufmann Ed.).

Han, J. W., Zheng, H. F., Cui, Y., Sun, L. D., Ye, D. Q., Hu, Z., . . . Zhang, X. J. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet, 41*(11), 1234-1237. doi: 10.1038/ng.472

Hansi, W., & Heimpel, H. (1979). [Treatment of panmyelopathies]. *Dtsch Med Wochenschr, 104*(27), 964-966. doi: 10.1055/s-0028-1129018

Heath, S. C., Gut, I. G., Brennan, P., McKay, J. D., Bencko, V., Fabianova, E., . . . Lathrop, M. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet, 16*(12), 1413-1429. doi: 10.1038/ejhg.2008.210

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., . . . Stefansson, K. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science, 316*(5830), 1491-1493. doi: 10.1126/science.1142842

Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S., . . . Stefansson, K. (2007). Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet, 39*(2), 218-225. doi: 10.1038/ng1960

Hellenthal, G., Auton, A., & Falush, D. (2008). Inferring human colonization history using a copying model. *PLoS Genet, 4*(5), e1000078. doi: 10.1371/journal.pgen.1000078

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science, 343*(6172), 747-751. doi: 10.1126/science.1243518

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol, 36*(1), 39-56.

Hill, A. V., & Jeffreys, A. J. (1985). Use of minisatellite DNA probes for determination of twin zygosity at birth. *Lancet, 2*(8469-70), 1394-1395.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367. doi: 10.1073/pnas.0903103106

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108. doi: 10.1038/nrg1521

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet Med, 4*(2), 45-61. doi: 10.1097/00125817-200203000-00002

Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., . . . Risch, N. (2011). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics, 98*(2), 79-89. doi: 10.1016/j.ygeno.2011.04.005

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., . . . Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using

high-density SNP genotyping microarrays. *PLoS Genet, 4*(8), e1000167. doi: 10.1371/journal.pgen.1000167

Horaitis, O., & Cotton, R. G. (2004). The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum Mutat, 23*(5), 447-452. doi: 10.1002/humu.20038

Horikoshi, M., Hara, K., Ito, C., Nagai, R., Froguel, P., & Kadowaki, T. (2007). A genetic variation of the transcription factor 7-like 2 gene is associated with risk of type 2 diabetes in the Japanese population. *Diabetologia, 50*(4), 747-751. doi: 10.1007/s00125-006-0588-6

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441. doi: 10.1037/h0071325

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 27: 321–377.

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi: 10.1371/journal.pgen.1000529

Huang, M. C., Chuang, T. P., Chen, C. H., Wu, J. Y., Chen, Y. T., Li, L. H., & Yang, H. C. (2016). An integrated analysis tool for analyzing hybridization intensities and genotypes using new-generation population-optimized human arrays. *BMC Genomics, 17*, 266. doi: 10.1186/s12864-016-2478-8

Hunt, L. T. (1983). Margaret O. Dayhoff 1925-1983. *DNA, 2*(2), 97-98. doi: 10.1089/dna.1983.2.97

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., . . . Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet, 39*(7), 870-874. doi: 10.1038/ng2075

IBM, C. (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY.

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314. doi: 10.1080/10618600.1996.10474713

Ikegawa, S. (2012). A short history of the genome-wide association study: where we were and where we are going. *Genomics Inform, 10*(4), 220-225. doi: 10.5808/GI.2012.10.4.220

Iloro, I., & Pampliega, O.

Ingles, M., Gimeno-Mallench, L., Mas-Bargues, C., Dromant, M., Cruz-Guerrero, R., Garcia-Garcia, F. J., . . . Vina, J. (2018). [Identification of single nucleotide polymorphisms related to frailty]. *Rev Esp Geriatr Gerontol, 53*(4), 202-207. doi: 10.1016/j.regg.2017.11.003

International HapMap, C. (2003). The International HapMap Project. *Nature, 426*(6968), 789-796. doi: 10.1038/nature02168

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi: 10.1038/nature04226

International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., . . . McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi: 10.1038/nature09298

Ioannidis, J. P., Contopoulos-Ioannidis, D. G., & Lau, J. (1999). Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J Clin Epidemiol, 52*(4), 281-291. doi: 10.1016/s0895-4356(98)00159-0

Ioannidis, J. P., & Lau, J. (1998). Uncontrolled pearls, controlled evidence, meta-analysis and the individual patient. *J Clin Epidemiol, 51*(8), 709-711. doi: 10.1016/s0895-4356(98)00042-0

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nat Genet, 29*(3), 306-309. doi: 10.1038/ng749

Ioannidis, J. P., Trikalinos, T. A., & Khoury, M. J. (2006). Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol, 164*(7), 609-614. doi: 10.1093/aje/kwj259

Irish Schizophrenia Genomics, C., & the Wellcome Trust Case Control, C. (2012). Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility

complex locus in schizophrenia. *Biol Psychiatry, 72*(8), 620-628. doi: 10.1016/j.biopsych.2012.05.035
Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., . . . Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature, 451*(7181), 998-1003. doi: 10.1038/nature06742
Jobling MA, H. M., Tyler-Smith C. (2004). Human Evolutionary Genetics: origins, peoples and disease. *London/New York: Garland Science Publishing*, 523.
Jobling, M. A., & Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. *Trends Genet, 11*(11), 449-456.
Jolliffe, I. (2002). *Principal Component Analysis*.
Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405.
Jombart, T. (2014). adegenet: an R package for the exploratory analysis of genetic and genomic data.
Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet, 11*, 94. doi: 10.1186/1471-2156-11-94
Jonathan L. Haines, M. P. V. (2005). *Genetic Analysis of Complex Diseases*: A JOHN WILEY & SONS, INC., PUBLICATION.
Jordahl, K. M., Shcherbina, A., Kim, A. E., Su, Y. R., Lin, Y., Wang, J., . . . Peters, U. (2022). Beyond GWAS of Colorectal Cancer: Evidence of Interaction with Alcohol Consumption and Putative Causal Variant for the 10q24.2 Region. *Cancer Epidemiol Biomarkers Prev, 31*(5), 1077-1089. doi: 10.1158/1055-9965.EPI-21-1003
Kawai, Y., Mimori, T., Kojima, K., Nariai, N., Danjoh, I., Saito, R., . . . Nagasaki, M. (2015). Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *Journal Of Human Genetics, 60*, 581. doi: 10.1038/jhg.2015.68
Keen, J. C., & Moore, H. M. (2015). The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. *J Pers Med, 5*(1), 22-29. doi: 10.3390/jpm5010022
Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science, 336*(6082), 740-743. doi: 10.1126/science.1217283
Kitzman, J. O., Mackenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., . . . Shendure, J. (2011). Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol, 29*(1), 59-63. doi: 10.1038/nbt.1740
Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389. doi: 10.1126/science.1109557
Klempner, S. J., Janjigian, Y. Y., & Wainberg, Z. A. (2023). Claudin18.who? Examining biomarker overlap and outcomes in claudin18.2-positive gastroesophageal adenocarcinomas. *ESMO Open, 8*(2), 100778. doi: 10.1016/j.esmoop.2022.100778
Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet, 43*(4), 520-526.
Koonin, E. V. (2001). Computational genomics. *Curr Biol, 11*(5), R155-158.
Kramer, I., Hooning, M. J., Mavaddat, N., Hauptmann, M., Keeman, R., Steyerberg, E. W., . . . Schmidt, M. K. (2020). Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. *Am J Hum Genet, 107*(5), 837-848. doi: 10.1016/j.ajhg.2020.09.001
Ku, C. S., Loy, E. Y., Pawitan, Y., & Chia, K. S. (2010). The pursuit of genome-wide association studies: where are we now? *J Hum Genet, 55*(4), 195-206. doi: 10.1038/jhg.2010.19

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi: 10.1038/nprot.2009.86

Kunkle, B., Grenier-Boley, B., Sims, R., Bis, J., Naj, A., Boland, A., . . . Pericak-Vance, M. (2018). Meta-analysis of genetic association with diagnosed Alzheimer's disease identifies novel risk loci and implicates Abeta, Tau, immunity and lipid processing. *bioRxiv*, 294629. doi: 10.1101/294629

Kvale, M. N., Hesselson, S., Hoffmann, T. J., Cao, Y., Chan, D., Connell, S., . . . Risch, N. (2015). Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics, 200*(4), 1051-1060. doi: 10.1534/genetics.115.178905

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi: 10.1038/35057062

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., . . . Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Curr Biol, 18*(16), 1241-1248. doi: 10.1016/j.cub.2008.07.049

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet, 8*(1), e1002453. doi: 10.1371/journal.pgen.1002453

Lee, J. C., Biasci, D., Roberts, R., Gearry, R. B., Mansfield, J. C., Ahmad, T., . . . Smith, K. G. (2017). Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet, 49*(2), 262-268. doi: 10.1038/ng.3755

Lee, J. C., & Parkes, M. (2011). Genome-wide association studies and Crohn's disease. *Brief Funct Genomics, 10*(2), 71-76. doi: 10.1093/bfgp/elr009

Leish, G. E. N. C., Wellcome Trust Case Control, C., Fakiola, M., Strange, A., Cordell, H. J., Miller, E. N., . . . Donnelly, P. (2013). Common variants in the HLA-DRB1–HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. *Nat Genet, 45*, 208. doi: 10.1038/ng.2518

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M. L., Alavere, H., Snieder, H., . . . Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol, 44*(4), 1137-1147. doi: 10.1093/ije/dyt268

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., . . . Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104. doi: 10.1126/science.1153717

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics, 165*(4), 2213-2233.

Li, Q., & Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol, 32*(3), 215-226. doi: 10.1002/gepi.20296

Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., . . . Wang, J. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet, 42*(11), 969-972. doi: 10.1038/ng.680

Li, Y. R., Zhao, S. D., Li, J., Bradfield, J. P., Mohebnasab, M., Steel, L., . . . Hakonarson, H. (2015). Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nat Commun, 6*, 8442. doi: 10.1038/ncomms9442

Lillioja, S., & Wilton, A. (2009). Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies. *Diabetologia, 52*(6), 1061-1074. doi: 10.1007/s00125-009-1324-9

Liu, J., Cheng, Y., Li, M., Zhang, Z., Li, T., & Luo, X. J. (2023). Genome-wide Mendelian randomization identifies actionable novel drug targets for psychiatric disorders. *Neuropsychopharmacology, 48*(2), 270-280. doi: 10.1038/s41386-022-01456-5

Liu, Y., Wang, Y., Qin, S., Jin, X., Jin, L., Gu, W., & Mu, Y. (2022). Insights Into Genome-Wide Association Study for Diabetes: A Bibliometric and Visual Analysis From 2001 to 2021. *Front Endocrinol (Lausanne), 13*, 817620. doi: 10.3389/fendo.2022.817620

Lluis-Ganella, C., Lucas, G., Subirana, I., Senti, M., Jimenez-Conde, J., Marrugat, J., . . . Elosua, R. (2010). Additive effect of multiple genetic variants on the risk of coronary artery disease. *Rev Esp Cardiol, 63*(8), 925-933.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet, 33*(2), 177-182. doi: 10.1038/ng1071

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res, 45*(D1), D896-D901. doi: 10.1093/nar/gkw1133

Magi, R., & Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics, 11*, 288. doi: 10.1186/1471-2105-11-288

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., . . . Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet, 39*(10), 1181-1186. doi: 10.1038/ng1007-1181

Malhotra, A. K. (2010). The pharmacogenetics of depression: enter the GWAS. *Am J Psychiatry, 167*(5), 493-495. doi: 10.1176/appi.ajp.2010.10020244

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *N Engl J Med, 363*(2), 166-176. doi: 10.1056/NEJMra0905980

Manolio, T. A. (2017). In Retrospect: A decade of shared genomic associations. *Nature, 546*(7658), 360-361. doi: 10.1038/546360a

Manolio, T. A., Bailey-Wilson, J. E., & Collins, F. S. (2006). Genes, environment and the value of prospective cohort studies. *Nat Rev Genet, 7*(10), 812-820. doi: 10.1038/nrg1919

Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *J Clin Invest, 118*(5), 1590-1605. doi: 10.1172/JCI34772

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi: 10.1038/nature08494

Maronas, O., Phillips, C., Sochtig, J., Gomez-Tato, A., Cruz, R., Alvarez-Dios, J., . . . Lareu, M. V. (2014). Development of a forensic skin colour predictive test. *Forensic Sci Int Genet, 13*, 34-44. doi: 10.1016/j.fsigen.2014.06.017

Marth, G. T., Yu, F., Indap, A. R., Garimella, K., Gravel, S., Leong, W. F., . . . Genomes, P. (2011). The functional spectrum of low-frequency coding variation. *Genome Biol, 12*(9), R84. doi: 10.1186/gb-2011-12-9-r84

Masotti, M., Guo, B., & Wu, B. (2019). Pleiotropy informed adaptive association test of multiple traits using genome-wide association study summary data. *Biometrics, 75*(4), 1076-1085. doi: 10.1111/biom.13076

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nat Genet, 44*(3), 243-246. doi: 10.1038/ng.1074

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet, 9*(5), 356-369. doi: 10.1038/nrg2344

McCarthy, M. I., & Zeggini, E. (2009). Genome-wide association studies in type 2 diabetes. *Curr Diab Rep, 9*(2), 164-171.

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell, 141*(2), 210-217. doi: 10.1016/j.cell.2010.03.032

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi: 10.1101/gr.107524.110

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., . . . Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science, 316*(5830), 1488-1491. doi: 10.1126/science.1142447

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet, 5*(10), e1000686. doi: 10.1371/journal.pgen.1000686

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science, 201*(4358), 786-792.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., . . . Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature, 551*, 92. doi: 10.1038/nature24284

Montano, V., & Jombart, T. (2017). An Eigenvalue test for spatial principal component analysis. *BMC Bioinformatics, 18*(1), 562. doi: 10.1186/s12859-017-1988-y

Moorad, J. A., & Wade, M. J. (2005). A genetic interpretation of the variation in inbreeding depression. *Genetics, 170*(3), 1373-1384. doi: 10.1534/genetics.104.033373

Morgan, L., McGinnis, R., Steinthorsdottir, V., Svyatova, G., Zakhidova, N., Lee, W. K., . . . Laivuori, H. (2014). InterPregGen: genetic studies of pre-eclampsia in three continents. *Nor Epidemiol, 24*(1-2), 141-146.

Mosteller, F., & Colditz, G. A. (1996). Understanding research synthesis (meta-analysis). *Annu Rev Public Health, 17*, 1-23. doi: 10.1146/annurev.pu.17.050196.000245

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol, 155*, 335-350.

Naj, A. C. (2019). Genotype Imputation in Genome-Wide Association Studies. *Curr Protoc Hum Genet, 102*(1), e84. doi: 10.1002/cphg.84

Nakamura, M., Nishida, N., Kawashima, M., Aiba, Y., Tanaka, A., Yasunami, M., . . . Ishibashi, H. (2012). Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population. *Am J Hum Genet, 91*(4), 721-728. doi: 10.1016/j.ajhg.2012.08.010

Nakamura, Y., Carlson, M., Krapcho, K., Kanamori, M., & White, R. (1988). New approach for isolation of VNTR markers. *Am J Hum Genet, 43*(6), 854-859.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., . . . Singleton, A. B. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet, 46*(9), 989-993. doi: 10.1038/ng.3043

Neale, B. M., & Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *Am J Hum Genet, 75*(3), 353-362. doi: 10.1086/423901

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., . . . Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science, 337*(6090), 100-104. doi: 10.1126/science.1217876

Niitsu, T., Fabbri, C., Bentini, F., & Serretti, A. (2013). Pharmacogenetics in major depression: a comprehensive meta-analysis. *Prog Neuropsychopharmacol Biol Psychiatry, 45*, 183-194. doi: 10.1016/j.pnpbp.2013.05.011

Niu, T. (2004). Algorithms for inferring haplotypes. *Genet Epidemiol, 27*(4), 334-347. doi: 10.1002/gepi.20024

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., . . . Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature, 456*(7218), 98-101. doi: 10.1038/nature07331

Ocejo-Vinyals, J. G., Puente de Mateo, E., Ausin, F., Aguero, R., Arroyo, J. L., Gutierrez-Cuadra, M., & Farinas, M. C. (2013). Human toll-like receptor 1 T1805G polymorphism and susceptibility to pulmonary tuberculosis in northern Spain. *Int J Tuberc Lung Dis, 17*(5), 652-654. doi: 10.5588/ijtld.12.0767

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., & Chee, M. S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques, Suppl*, 56-58, 60-51.

Ottman, R. (1990). An epidemiologic approach to gene-environment interaction. *Genet Epidemiol, 7*(3), 177-185. doi: 10.1002/gepi.1370070302

Oxman, A. D., & Guyatt, G. H. (1992). A consumer's guide to subgroup analyses. *Ann Intern Med, 116*(1), 78-84. doi: 10.7326/0003-4819-116-1-78

Pal Choudhury, P., Brook, M. N., Hurson, A. N., Lee, A., Mulder, C. V., Coulson, P., . . . Garcia-Closas, M. (2021). Comparative validation of the BOADICEA and Tyrer-Cuzick breast cancer risk models incorporating classical risk factors and polygenic risk in a population-based prospective cohort of women of European ancestry. *Breast Cancer Res, 23*(1), 22. doi: 10.1186/s13058-021-01399-7

Parkes, M., Cortes, A., van Heel, D. A., & Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet, 14*(9), 661-673. doi: 10.1038/nrg3502

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572. doi: 10.1080/14786440109462720

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA, 299*(11), 1335-1344. doi: 10.1001/jama.299.11.1335

Peiffer, D. A., & Gunderson, K. L. (2009). Design of tag SNP whole genome genotyping arrays. *Methods Mol Biol, 529*, 51-61. doi: 10.1007/978-1-59745-538-1_4

Pena-Chilet, M., Roldan, G., Perez-Florido, J., Ortuno, F. M., Carmona, R., Aquino, V., . . . Dopazo, J. (2021). CSVS, a crowdsourcing database of the Spanish population genetic variability. *Nucleic Acids Res, 49*(D1), D1130-D1137. doi: 10.1093/nar/gkaa794

Pennisi, E. (2007). Breakthrough of the year. Human genetic variation. *Science, 318*(5858), 1842-1843. doi: 10.1126/science.318.5858.1842

Pensado-Lopez, A., Veiga-Rua, S., Carracedo, A., Allegue, C., & Sanchez, L. (2020). Experimental Models to Study Autism Spectrum Disorders: hiPSCs, Rodents and Zebrafish. *Genes (Basel), 11*(11). doi: 10.3390/genes11111376

Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., . . . Consortium, S. N. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet, 1*(3-4), 273-280. doi: 10.1016/j.fsigen.2007.06.008

Phillips, P. C. (2008). Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet, 9*(11), 855-867. doi: 10.1038/nrg2452

Pickrell, J. K., Berisa, T., Liu, J. Z., Segurel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet, 48*(7), 709-717. doi: 10.1038/ng.3570

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature, 538*(7624), 161-164. doi: 10.1038/538161a

Preuss, M., Konig, I. R., Thompson, J. R., Erdmann, J., Absher, D., Assimes, T. L., . . . Consortium, C. A. (2010). Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ Cardiovasc Genet, 3*(5), 475-483. doi: 10.1161/CIRCGENETICS.109.899443

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet, 38*(8), 904-909. doi: 10.1038/ng1847

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet, 69*(1), 124-137. doi: 10.1086/321272

Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet, 11*(20), 2417-2423. doi: 10.1093/hmg/11.20.2417

Pritchard, J. K., & Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theor Popul Biol, 60*(3), 227-237. doi: 10.1006/tpbi.2001.1543

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Psychosis Endophenotypes International, C., Wellcome Trust Case-Control, C., Bramon, E., Pirinen, M., Strange, A., Lin, K., . . . Spencer, C. C. (2014). A genome-wide association analysis of a broad psychosis phenotype identifies three loci for further investigation. *Biol Psychiatry, 75*(5), 386-397. doi: 10.1016/j.biopsych.2013.03.033

Pulker, H., Lareu, M. V., Phillips, C., & Carracedo, A. (2007). Finding genes that underlie physical traits of forensic interest using genetic tools. *Forensic Sci Int Genet, 1*(2), 100-104. doi: 10.1016/j.fsigen.2007.02.009

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Qin, Z., Wang, Y., Cao, S., He, Y., Ma, H., Jin, G., . . . Shen, H. (2013). Genetic variants at 12p11 and 12q24 are associated with breast cancer risk in a Chinese population. *PLoS One, 8*(6), e66519. doi: 10.1371/journal.pone.0066519

Rahmani, M., Earp, M. A., Ramezani Tehrani, F., Ataee, M., Wu, J., Treml, M., . . . Brooks-Wilson, A. (2013). Shared genetic factors for age at natural menopause in Iranian and European women. *Hum Reprod, 28*(7), 1987-1994. doi: 10.1093/humrep/det106

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res, 30*(17), 3894-3900. doi: 10.1093/nar/gkf493

Rancati, G., Moffat, J., Typas, A., & Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. *Nat Rev Genet, 19*(1), 34-49. doi: 10.1038/nrg.2017.74

Rao, S., Yao, Y., & Bauer, D. E. (2021). Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Med, 13*(1), 41. doi: 10.1186/s13073-021-00857-3

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature, 461*(7263), 489-494. doi: 10.1038/nature08365

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends Genet, 17*(9), 502-510. doi: 10.1016/s0168-9525(01)02410-6

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., . . . Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet, 32*(1), 135-142. doi: 10.1038/ng947

Reme, T., Hose, D., De Vos, J., Vassal, A., Poulain, P. O., Pantesco, V., . . . Klein, B. (2008). A new method for class prediction based on signed-rank algorithms applied to Affymetrix microarray experiments. *BMC Bioinformatics, 9*, 16. doi: 10.1186/1471-2105-9-16

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273*(5281), 1516-1517.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature, 405*(6788), 847-856. doi: 10.1038/35015718

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi: 10.1038/nature14248

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet, 70*(Pt 6), 841-847. doi: 10.1111/j.1469-1809.2006.00285.x

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., . . . Lareu, M. V. (2013). Further development of forensic eye color predictive tests. *Forensic Sci Int Genet, 7*(1), 28-40. doi: 10.1016/j.fsigen.2012.05.009

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., . . . International, S. N. P. M. W. G. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature, 409*(6822), 928-933. doi: 10.1038/35057149

Sale, M. M., Mychaleckyj, J. C., & Chen, W. M. (2009). Planning and executing a genome wide association study (GWAS). *Methods Mol Biol, 590*, 403-418. doi: 10.1007/978-1-60327-378-7_25

Salmela, E., Lappalainen, T., Fransson, I., Andersen, P. M., Dahlman-Wright, K., Fiebig, A., . . . Lahermo, P. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS One, 3*(10), e3519. doi: 10.1371/journal.pone.0003519

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., . . . the Cardiogenics, C. (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med, 357*(5), 443-453. doi: 10.1056/NEJMoa072366

Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J., & Roe, B. A. (1980). Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J Mol Biol, 143*(2), 161-178.

Saunders, C. L., Chiodini, B. D., Sham, P., Lewis, C. M., Abkevich, V., Adeyemo, A. A., . . . Collier, D. A. (2007). Meta-analysis of genome-wide linkage studies in BMI and obesity. *Obesity (Silver Spring), 15*(9), 2263-2275. doi: 10.1038/oby.2007.269

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet, 78*(4), 629-644. doi: 10.1086/502802

Schillert, A., & Ziegler, A. (2012). Genotype calling for the Affymetrix platform. *Methods Mol Biol, 850*, 513-523. doi: 10.1007/978-1-61779-555-8_28

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., . . . Boehnke, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science, 316*(5829), 1341-1345. doi: 10.1126/science.1142382

Shek, D. T., & Ma, C. M. (2011). Longitudinal data analyses using linear mixed models in SPSS: concepts, procedures and illustrations. *ScientificWorldJournal, 11*, 42-76. doi: 10.1100/tsw.2011.2

Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., . . . Oliphant, A. (2005). High-throughput SNP genotyping on universal bead arrays. *Mutat Res, 573*(1-2), 70-82. doi: 10.1016/j.mrfmmm.2004.07.022

Sherry, S. T., Ward, M., & Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res, 9*(8), 677-679.

Siva, N. (2008). 1000 Genomes project. *Nat Biotechnol, 26*(3), 256. doi: 10.1038/nbt0308-256b

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., . . . Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet, 89*(5), 607-618. doi: 10.1016/j.ajhg.2011.10.004

Song, M., Hao, W., & Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nat Genet, 47*(5), 550-554. doi: 10.1038/ng.3244

Steemers, F. J., & Gunderson, K. L. (2007). Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J, 2*(1), 41-49. doi: 10.1002/biot.200600213

Studies, N.-N. W. G. o. R. i. A., Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., . . . Collins, F. S. (2007). Replicating genotype-phenotype associations. *Nature, 447*(7145), 655-660. doi: 10.1038/447655a

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81. doi: 10.1038/nature15394

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., . . . Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature, 590*(7845), 290-299. doi: 10.1038/s41586-021-03205-y

Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat Rev Genet, 20*(8), 467-484. doi: 10.1038/s41576-019-0127-1

Tansey, K. E., Guipponi, M., Hu, X., Domenici, E., Lewis, G., Malafosse, A., . . . Uher, R. (2013). Contribution of common genetic variants to antidepressant response. *Biol Psychiatry, 73*(7), 679-682. doi: 10.1016/j.biopsych.2012.10.030

Taylor, L. C., Law, K., Hutchinson, A., Dennison, R. A., & Usher-Smith, J. A. (2023). Acceptability of risk stratification within population-based cancer screening from the perspective of healthcare professionals: A mixed methods systematic review and recommendations to support implementation. *PLoS One, 18*(2), e0279201. doi: 10.1371/journal.pone.0279201

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., . . . Project, N. E. S. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science, 337*(6090), 64-69. doi: 10.1126/science.1219240

Teo, Y. Y. (2008). Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol, 19*(2), 133-143. doi: 10.1097/MOL.0b013e3282f5dd77

Teo, Y. Y. (2012). Genotype calling for the Illumina platform. *Methods Mol Biol, 850*, 525-538. doi: 10.1007/978-1-61779-555-8_29

The Australo-Anglo-American Spondyloarthritis, C., the Wellcome Trust Case Control, C., Evans, D. M., Spencer, C. C. A., Pointon, J. J., Su, Z., . . . Donnelly, P. (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet, 43*, 761. doi: 10.1038/ng.873

The Blue Mountains Eye, S., & The Wellcome Trust Case Control, C. (2013). Genome-wide association study of intraocular pressure identifies the GLCCI1/ICA1 region as a glaucoma susceptibility locus. *Hum Mol Genet, 22*(22), 4653-4660. doi: 10.1093/hmg/ddt293

The Esophageal Adenocarcinoma Genetics, C., The Wellcome Trust Case Control, C., Su, Z., Gay, L. J., Strange, A., Palles, C., . . . Jankowski, J. A. Z. (2012). Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus. *Nature Genetics, 44*, 1131. doi: 10.1038/ng.2408

The Go, D., Group, U. D. P. S., & The Wellcome Trust Case Control, C. (2010). Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nat Genet, 43*, 117. doi: 10.1038/ng.735

The International HapMap, C. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*, 851. doi: 10.1038/nature06258

The International Multiple Sclerosis Genetics, C., & The Wellcome Trust Case Control, C. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature, 476*, 214. doi: 10.1038/nature10251

The International Stroke Genetics, C., the Wellcome Trust Case Control, C., Bellenguez, C., Bevan, S., Gschwendtner, A., Spencer, C. C. A., . . . Markus, H. S. (2012). Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet, 44*, 328. doi: 10.1038/ng.1081

The Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*, 661. doi: 10.1038/nature05911

Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., . . . Chanock, S. J. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet, 40*(3), 310-315. doi: 10.1038/ng.91

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., & Feldman, M. W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A, 97*(13), 7360-7365.

Todd, J. A. (2006). Statistical false positive or true disease pathway? *Nat Genet, 38*(7), 731-733. doi: 10.1038/ng0706-731

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., . . . Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet, 39*, 984. doi: 10.1038/ng2085

Torgerson, D. G., Ampleford, E. J., Chiu, G. Y., Gauderman, W. J., Gignoux, C. R., Graves, P. E., . . . Nicolae, D. L. (2011). Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet, 43*(9), 887-892. doi: 10.1038/ng.888

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., . . . Maeda, S. (2008). SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet, 40*, 1098. doi: 10.1038/ng.208

Vaquero, A. R., Ferreira, N. E., Omae, S. V., Rodrigues, M. V., Teixeira, S. K., Krieger, J. E., & Pereira, A. C. (2012). Using gene-network landscape to dissect genotype effects of TCF7L2 genetic variant on diabetes and cardiovascular risk. *Physiol Genomics, 44*(19), 903-914. doi: 10.1152/physiolgenomics.00030.2012

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi: 10.1126/science.1058040

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet, 90*(1), 7-24. doi: 10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet, 101*(1), 5-22. doi: 10.1016/j.ajhg.2017.06.005

Visscher, P. M., & Yang, J. (2016). A plethora of pleiotropy across complex traits. *Nat Genet, 48*(7), 707-708. doi: 10.1038/ng.3604

Waage, J., Standl, M., Curtin, J. A., Jessen, L. E., Thorsen, J., Tian, C., . . . Bonnelykke, K. (2018). Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet, 50*(8), 1072-1080. doi: 10.1038/s41588-018-0157-1

Wacholder, S., Rothman, N., & Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst, 92*(14), 1151-1158.

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

Wang, X., Strizich, G., Hu, Y., Wang, T., Kaplan, R. C., & Qi, Q. (2016). Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *J Diabetes, 8*(1), 24-35. doi: 10.1111/1753-0407.12323

Warren, H. R., Evangelou, E., Cabrera, C. P., Gao, H., Ren, M., Mifsud, B., . . . group, U. K. B. C. C. B. w. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet, 49*(3), 403-415. doi: 10.1038/ng.3768

Watson, J. D., & Crick, F. H. (1995). Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. 1953. *Ann N Y Acad Sci, 758*, 13-14.

Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678. doi: 10.1038/nature05911

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi: 10.1093/nar/gkt1229

Williams, H. J., Owen, M. J., & O'Donovan, M. C. (2009). Schizophrenia genetics: new insights from new approaches. *Br Med Bull, 91*, 61-74. doi: 10.1093/bmb/ldp017

Williams, K. L., Topp, S., Yang, S., Smith, B., Fifita, J. A., Warraich, S. T., . . . Blair, I. P. (2016). CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nat Commun, 7*, 11253. doi: 10.1038/ncomms11253

Xiao, Y., Segal, M. R., Yang, Y. H., & Yeh, R. F. (2007). A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics, 23*(12), 1459-1467. doi: 10.1093/bioinformatics/btm131

Yang, H.-C., Huang, M.-C., Li, L.-H., Lin, C.-H., Yu, A. L., Diccianni, M. B., . . . Fann, C. S. (2008). MPDA: Microarray pooled DNA analyzer. *BMC Bioinformatics, 9*(1), 196. doi: 10.1186/1471-2105-9-196

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet, 88*(1), 76-82. doi: 10.1016/j.ajhg.2010.11.011

Yang, W. Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nat Genet, 44*(6), 725-731. doi: 10.1038/ng.2285

Yang, X., Eriksson, M., Czene, K., Lee, A., Leslie, G., Lush, M., . . . Antoniou, A. C. (2022). Prospective validation of the BOADICEA multifactorial breast cancer risk prediction model in a large prospective cohort study. *J Med Genet, 59*(12), 1196-1205. doi: 10.1136/jmg-2022-108806

Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., . . . Kasuga, M. (2008). Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet, 40*, 1092. doi: 10.1038/ng.207

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., . . . Thomas, G. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet, 39*, 645. doi: 10.1038/ng2022

Zanke, B. W., Greenwood, C. M. T., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., . . . Dunlop, M. G. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet, 39*, 989. doi: 10.1038/ng2089

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., . . . Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet, 40*(5), 638-645. doi: 10.1038/ng.120

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., . . . Hattersley, A. T. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science, 316*(5829), 1336-1341. doi: 10.1126/science.1142364

Zhang, X., Mu, W., Liu, C., & Zhang, W. (2014). Ancestry-informative markers for African Americans based on the Affymetrix Pan-African genotyping array. *PeerJ, 2*, e660. doi: 10.7717/peerj.660

Zhang, Y. M., Jia, Z., & Dunwell, J. M. (2019). Editorial: The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits. *Front Plant Sci, 10*, 100. doi: 10.3389/fpls.2019.00100

Zheng, W., Long, J., Gao, Y. T., Li, C., Zheng, Y., Xiang, Y. B., . . . Shu, X. O. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet, 41*(3), 324-328. doi: 10.1038/ng.318

Zhu, X., Bai, W., & Zheng, H. (2021). Twelve years of GWAS discoveries for osteoporosis and related traits: advances, challenges and applications. *Bone Res, 9*(1), 23. doi: 10.1038/s41413-021-00143-3

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A, 109*(4), 1193-1198. doi: 10.1073/pnas.1119675109

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., . . . Lander, E. S. (2014). Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A, 111*(4), E455-464. doi: 10.1073/pnas.1322563111

RECURSOS WEB

https://glosarios.servidor-alicante.com/genetica/ 27/02/2023

https://ccge.medschl.cam.ac.uk/boadicea/ 15/03/2023

http://www.ebi.ac.uk/gwas/ 27/02/2023

http://www.cardiogramplusc4d.org/data-downloads/ 27/02/2023

https://pgc.unc.edu/ 27/02/2023

https://github.com/CAG-CNV/ParseCNV2 09/07/2023

http://www.jurgott.org/linkage/liped.html 27/02/2023

https://www.ncbi.nlm.nih.gov/gap 27/02/2023

https://www.framinghamheartstudy.org/ 27/02/2023

https://www.ebi.ac.uk/gwas/diagram 27/02/2023

http://www.well.ox.ac.uk/home 28/02/2023

https://forensemolecular.es.tl/STRs.htm 28/02/2023

https://www.illumina.com/ 28/02/2023

https://www.thermofisher.com/es/es/home/life-science/microarray-analysis.html 28/02/2023

https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi 28/02/2023

https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/ 28/02/2023

https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/ 28/02/2023

https://www.genome.gov/19518624/2006-release-nih-announces-two-integral-components-of-the-cancer-genome 28/02/2023

https://www.genome.gov/26524200/2007-release-nih-launches-human-microbiome-project 28/02/2023

https://www.nature.com/ 28/02/2023

http://www.sciencemag.org/ 28/02/2023

https://www.genome.gov/about-nhgri/Brief-History-Timeline 28/02/2023

https://conogasi.org/diccionario/genoma-de-referencia/	28/02/2023
http://www.cephb.fr/en/hgdp_panel.php	01/03/2023
https://www.hagsc.org/hgdp/	01/03/2023
https://www.ncbi.nlm.nih.gov/nlmcatalog/101200656	01/03/2023
https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/	01/03/2023
http://www.humanvariomeproject.org/	01/03/2023
https://www.omim.org	01/03/2023
https://www.ncbi.nlm.nih.gov/genbank/	01/03/2023
https://www.ncbi.nlm.nih.gov/snp/	01/03/2023
https://www.ncbi.nlm.nih.gov/gap_	01/03/2023
https://www.genome.gov/es/genetics-glossary/HapMap-mapa-de-haplotipos	01/03/2023
https://www.ncbi.nlm.nih.gov/	01/03/2023
http://varnomen.hgvs.org/	01/03/2023
https://grenada.lumc.nl/LSDB_list/lsdbs/DMD	01/03/2023
https://www.ebi.ac.uk/eva/	01/03/2023
https://www.pharmgkb.org/	01/03/2023
https://www.ncbi.nlm.nih.gov/gtr/	01/03/2023
https://www.ensembl.org/index.html	01/03/2023
https://genome.ucsc.edu/	01/03/2023
http://www.1000genomes.org	01/03/2023
http://www.ukbiobank.ac.uk/	01/03/2023
https://www.ukbiobank.ac.uk/resources	01/03/2023
http://www.ckbiobank.org/site/	01/03/2023
http://www.genomenewsnetwork.org/	01/03/202

https://www.nlm.nih.gov/about/2017CJ.html 01/03/2023

https://www.sanger.ac.uk/ 01/03/2023

http://www.ebi.ac.uk/ 01/03/2023

http://www.embl.org/ 01/03/2023

https://inb-elixir.es/ 01/03/2023

https://www.isciii.es/QueHacemos/Financiacion/Documents/IMPaCT%20Web/PLAN_ESTRATEGICO_IMPaCT.pdf 12/07/2023

https://cran.r-project.org/ 01/03/2023

http://bioperl.org/ 01/03/2023

http://biopython.org/ 01/03/2023

https://www.hsph.harvard.edu/alkes-price/software/ 01/03/2023

http://cmpg.unibe.ch/software/arlequin35/ 01/03/2023

http://www.fluxus-engineering.com/sharenet.htm 01/03/2023

http://bioperl.org 01/03/2023

http://biopython.org/ 01/03/2023

http://cnsgenomics.com/software/gcta/#Overview 01/03/2023

http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html 01/03/2023

https://bioinformaticshome.com/tools/imputation/descriptions/fastPHASE.html#gsc.tab=0 01/03/2023

https://es.mathworks.com/ 01/03/2023

https://www.cog-genomics.org/plink2 01/03/2023

https://www.rdocumentation.org/packages/skatMeta/versions/1.4.3/topics/skatMeta 01/03/2023

https://cran.r-project.org/web/packages/SNPassoc/index.html 01/03/2023

http://mathgen.stats.ox.ac.uk/impute/impute_v2.html 01/03/2023

https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html 01/03/2023

https://genome.sph.umich.edu/wiki/METAL_Documentation#Brief_Description 01/03/2023

http://www.well.ox.ac.uk/GWAMA 01/03/2023

https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0018132_702991_Axiom2_96F_Man_SPG.pdf 01/03/2023

http://www.bancoadn.org/ 01/03/2023

https://www.thermofisher.com/order/catalog/product/902502#:~:text=UK%20Biobank%20Axiom%E2%84%A2%20Array%20was%20designed%20by%20and%20for,single%20comprehensive%20low%2Dcost%20solution. 01/03/2023

https://assets.thermofisher.com/TFS-Assets/LSG/brochures/uk_axiom_biobank_contentsummary_brochure.pdf 18/10/2023

https://annovar.openbioinformatics.org/en/latest/#reference 01/03/2023

https://emea.illumina.com/products/by-type/microarray-kits/infinium-psycharray.html 01/03/2023

https://emea.illumina.com/products/by-type/microarray-kits/infinium-immunoarray.html 01/03/2023

https://genome.ucsc.edu/ 02/03/2023

https://bio.tools/dbsnp-q 02/03/2023

https://www.thermofisher.com/es/es/home/life-science/microarray-analysis.html 01/03/2023 02/03/2023

https://www.thermofisher.com/es/es/home/life-science/microarray-analysis/microarray-analysis-instruments-software-services/microarray-analysis-software/axiom-analysis-suite.html 01/03/2023

https://www.genecards.org/ 09/02/2023

https://www.sen.es/saladeprensa/pdf/Link191.pdf 09/02/2023

https://www.nhlbi.nih.gov/es/salud/sindrome-de-marfan/causas#:~:text=El%20gen%20FBN1%20produce%20fibrilina,a%20controlar%20c%C3%B3mo%20se%20crece. 31/01/2023

https://ec.europa.eu/health/archive/ph_threats/non_com/docs/r299_es.pdf 31/01/2023

https://www.ahedysia.org/patologias/127-sindrome-x-fragil 08/02/2023

https://fibrosisquistica.org/ 08/02/2023

https://www.ncbi.nlm.nih.gov/gene/101927412 02/03/2023

http://www.thedonnellycentre.utoronto.ca/news/understanding-gene-interactions-holds-key-personalized-medicine-donnelly-centre-scientists-say 26/02/2023

https://es.wikipedia.org/wiki/Selecci%C3%B3n_natural 26/02/2023

https://isanidad.com/228154/angel-carracedo-papel-medico-atencion-primaria-va-a-ser-vital-medicina-personalizada/ 26/02/2023

https://www.biorender.com/ 07/07/2023

https://www.institutoroche.es/static/archivos/Informes_anticipando_2022_Prediccion_riesgo_DEF.pdf 12/07/2023

ANEXO I

Anexo I

COMPROBACIÓN DE LAS CATEGORÍAS Y FRECUENCIAS EN POBLACIÓN ESPAÑOLA DE LAS VARIANTES INCLUIDAS EN EL THERMOFISHER *AXIOM SPAIN BIOBANK ARRAY PLATE* VS SUS FRECUENCIAS EN POBLACIÓN EUROPEA (SEGÚN 1000G)

Tabla A1. "Categorías vs frecuencias de las variantes incluidas en el array"

			Frecuencias			Total
			1*	2**	3***	
Categoría	ADME	Recuento	30	8	595	633
		% dentro de categoría	4,70%	1,30%	94,00%	100,00%
		% dentro de comprobación	3,70%	0,10%	0,10%	0,10%
	CARDIOGENÉTICA	Recuento	0	2	221	223
		% dentro de categoría	0,00%	0,90%	99,10%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CARDIOGENÉTICA//ADME	Recuento	0	0	2	2
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CARDIOGENÉTICA//DR. CARLOS FLORES	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CARDIOGENÉTICA//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	0	10	10
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CCR	Recuento	2	0	41	43
		% dentro de categoría	4,70%	0,00%	95,30%	100,00%
		% dentro de comprobación	0,20%	0,00%	0,00%	0,00%
	CCR//NEFROLOGÍA	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CCR//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	0	3	3
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
		% dentro de categoría	2,00%	1,00%	97,00%	100,00%

Anexo I

COMPROBACIÓN DE LAS CATEGORÍAS Y FRECUENCIAS EN POBLACIÓN ESPAÑOLA DE LAS VARIANTES INCLUIDAS EN EL THERMOFISHER *AXIOM SPAIN BIOBANK ARRAY PLATE* VS SUS FRECUENCIAS EN POBLACIÓN EUROPEA (SEGÚN 1000G)

Tabla A1. "Categorías vs frecuencias de las variantes incluidas en el array"

			Frecuencias			Total
			1*	2**	3***	
Categoría		% dentro de comprobación	21,80%	1,50%	1,50%	1,50%
	CM	Recuento	4	0	70	74
		% dentro de categoría	5,40%	0,00%	94,60%	100,00%
		% dentro de comprobación	0,50%	0,00%	0,00%	0,00%
	CM//CO	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CM//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	0	2	2
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	HLA	Recuento	0	3	51	54
		% dentro de categoría	0,00%	5,60%	94,40%	100,00%
		% dentro de comprobación	0,00%	0,10%	0,00%	0,00%
	HLA//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	1	13	14
		% dentro de categoría	0,00%	7,10%	92,90%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	MS	Recuento	0	1	62	63
		% dentro de categoría	0,00%	1,60%	98,40%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	MS//HLA	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	NEFROLOGÍA	Recuento	42	27	844	913
		% dentro de categoría	4,60%	3,00%	92,40%	100,00%
		% dentro de comprobación	5,20%	0,50%	0,10%	0,20%

Anexo I

COMPROBACIÓN DE LAS CATEGORÍAS Y FRECUENCIAS EN POBLACIÓN ESPAÑOLA DE LAS VARIANTES INCLUIDAS EN EL THERMOFISHER *AXIOM SPAIN BIOBANK ARRAY PLATE* VS SUS FRECUENCIAS EN POBLACIÓN EUROPEA (SEGÚN 1000G)

Tabla A1. "Categorías vs frecuencias de las variantes incluidas en el array"

			Frecuencias			Total
			1*	2**	3***	
Categoría	NEFROLOGÍA//ADME	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	NEFROLOGÍA//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	8	157	165
		% dentro de categoría	0,00%	4,80%	95,20%	100,00%
		% dentro de comprobación	0,00%	0,10%	0,00%	0,00%
	CO	Recuento	0	0	7	7
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	CP	Recuento	0	0	63	63
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	PSIQUIATRÍA_INMUNOLOGÍA	Recuento	152	670	20229	21051
		% dentro de categoría	0,70%	3,20%	96,10%	100,00%
		% dentro de comprobación	19,00%	11,30%	3,50%	3,60%
	PSIQUIATRÍA_INMUNOLOGÍA//ADME	Recuento	1	1	50	52
		% dentro de categoría	1,90%	1,90%	96,20%	100,00%
		% dentro de comprobación	0,10%	0,00%	0,00%	0,00%
	RADIOGENÓMICA	Recuento	36	1	1784	1821
		% dentro de categoría	2,00%	0,10%	98,00%	100,00%
		% dentro de comprobación	4,50%	0,00%	0,30%	0,30%
	RADIOGENÓMICA//ADME	Recuento	0	0	1	1
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%

Anexo I

COMPROBACIÓN DE LAS CATEGORÍAS Y FRECUENCIAS EN POBLACIÓN ESPAÑOLA DE LAS VARIANTES INCLUIDAS EN EL THERMOFISHER *AXIOM SPAIN BIOBANK ARRAY PLATE* VS SUS FRECUENCIAS EN POBLACIÓN EUROPEA (SEGÚN 1000G)

Tabla A1. "Categorías vs frecuencias de las variantes incluidas en el array"

			Frecuencias			Total
			1*	2**	3***	
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	RADIOGENÓMICA//PSIQUIATRÍA_INMUNOLOGÍA	Recuento	0	0	10	10
		% dentro de categoría	0,00%	0,00%	100,00%	100,00%
		% dentro de comprobación	0,00%	0,00%	0,00%	0,00%
	SELECCIÓN RARAS	Recuento	42	279	13579	13900
		% dentro de categoría	0,30%	2,00%	97,70%	100,00%
		% dentro de comprobación	5,20%	4,70%	2,30%	2,40%
	UK_BIOBANK_ARRAY	Recuento	286	4800	525108	530194
		% dentro de categoría	0,10%	0,90%	99,00%	100,00%
		% dentro de comprobación	35,70%	81,20%	90,60%	90,50%
TOTAL		Recuento	802	5.909	579.316	586.027
		% dentro de categoría	0,10%	1,00%	98,90%	100,00%
		% dentro de comprobación	100,00%	100,00%	100,00%	100,00%

1*Categoría_España = rara o monomórfica y Categoría_Europa_1000G = SNP

2** Categoría_España = SNP o baja frecuencia y Categoría_Europa_1000G = rara o monomórfica

3*** Categoría_España = Categoría_Europa_1000G

ANEXO II

ANEXO II

CARACTERIZACIÓN DE LA ESTRUCTURA DE LA POBLACIÓN ESPAÑOLA EN FUNCIÓN DE LOS DIFERENTES BIOMARCADORES Y LDs TESTADAS

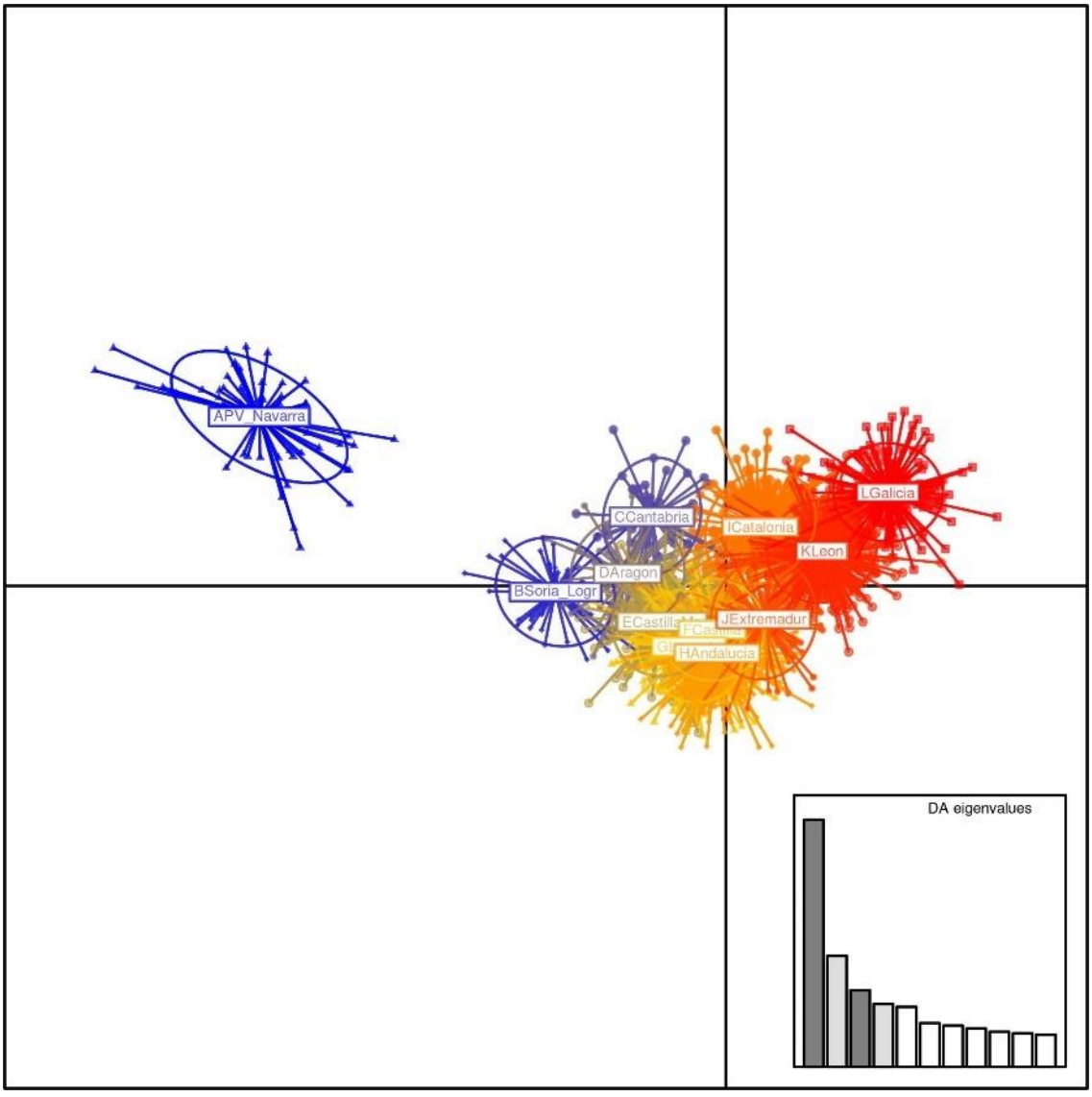


Figura A1. Caracterización de la estructura de nuestra población española de estudio en función de SNPs (LD1 y LD3).

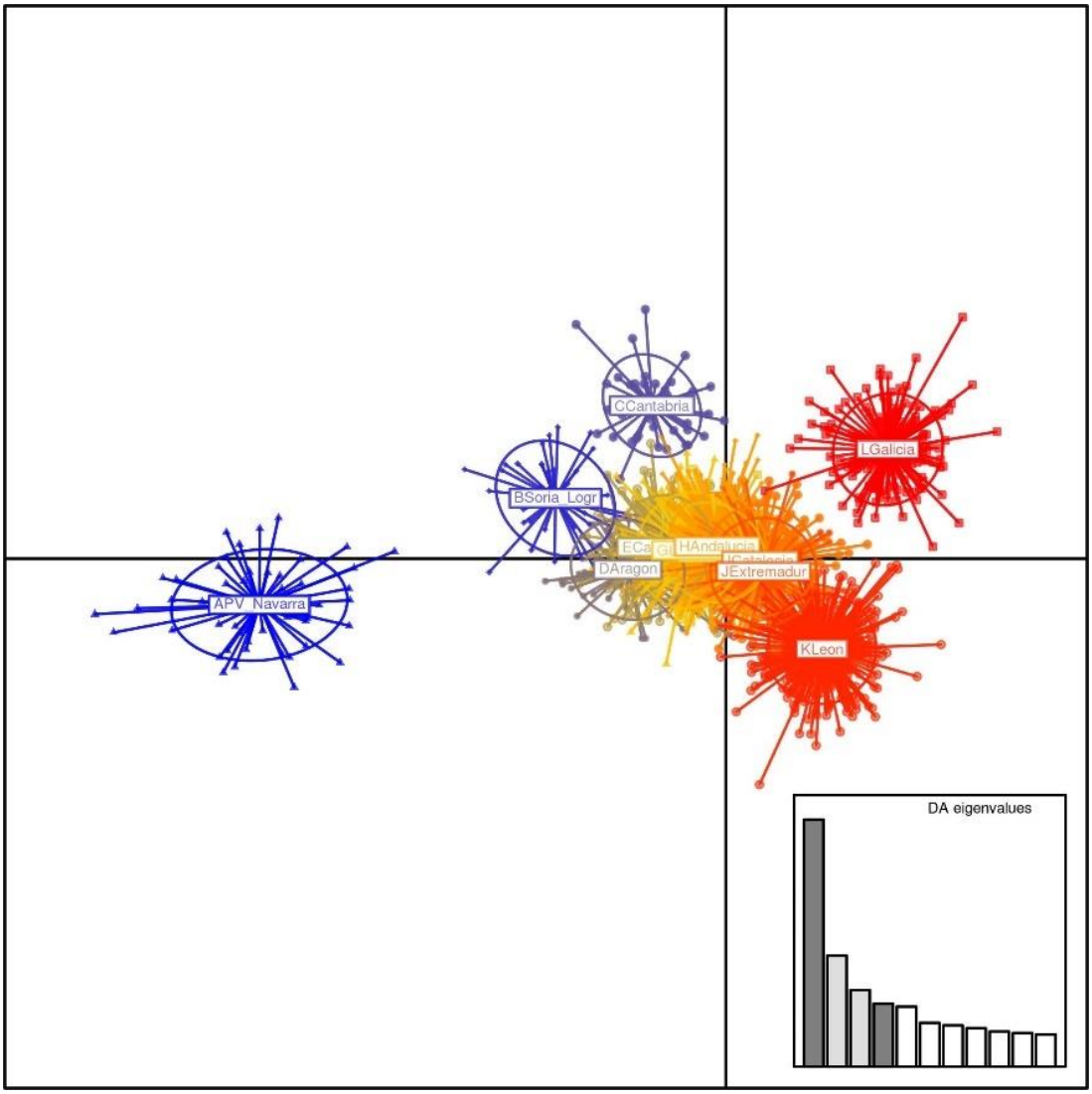


Figura A2. Caracterización de la estructura de nuestra población española de estudio en función de SNPs (LD1 y LD4).

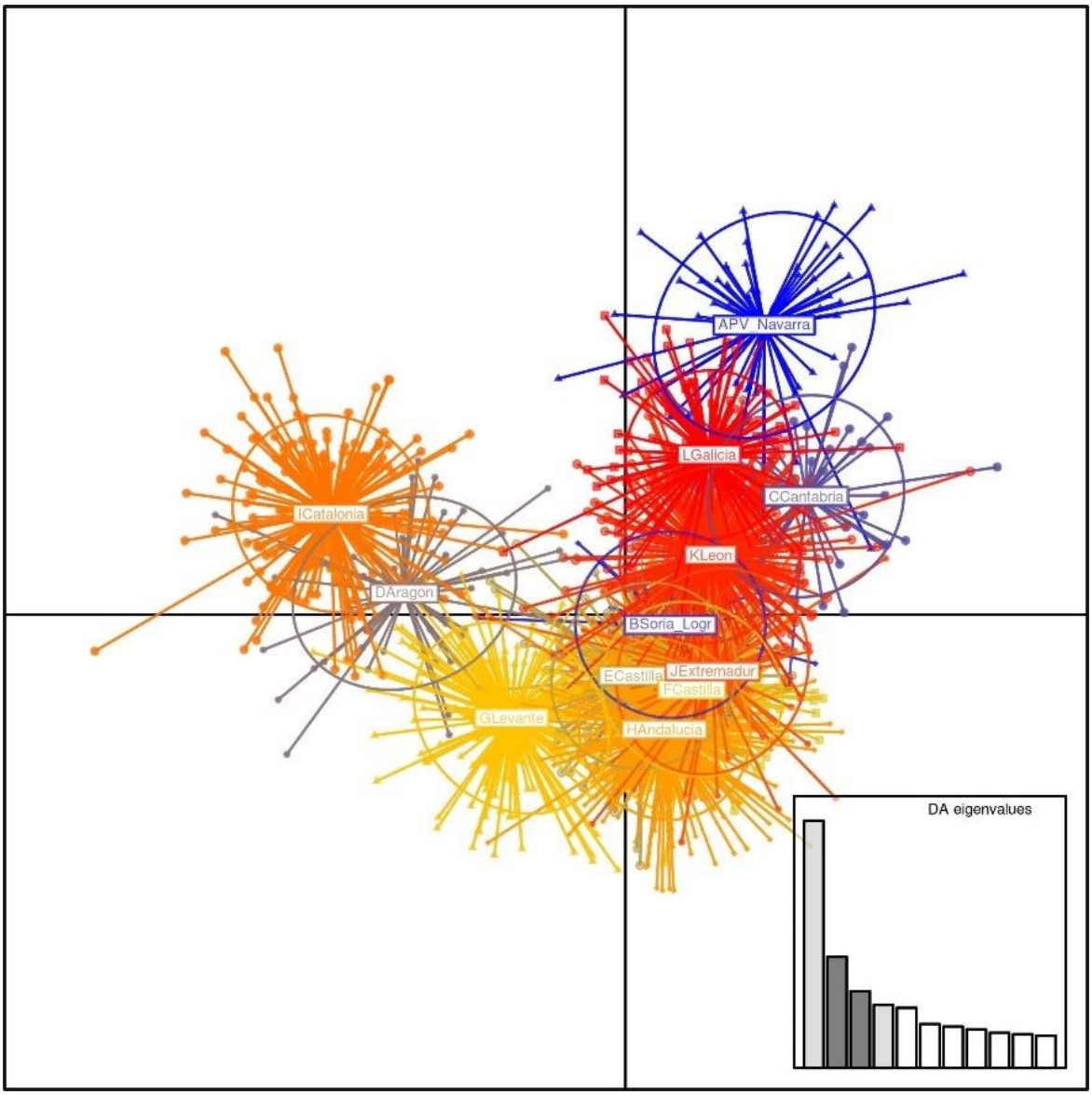


Figura A3. Agrupación de nuestra población española de estudio en función de SNPs (LD2 y LD3).

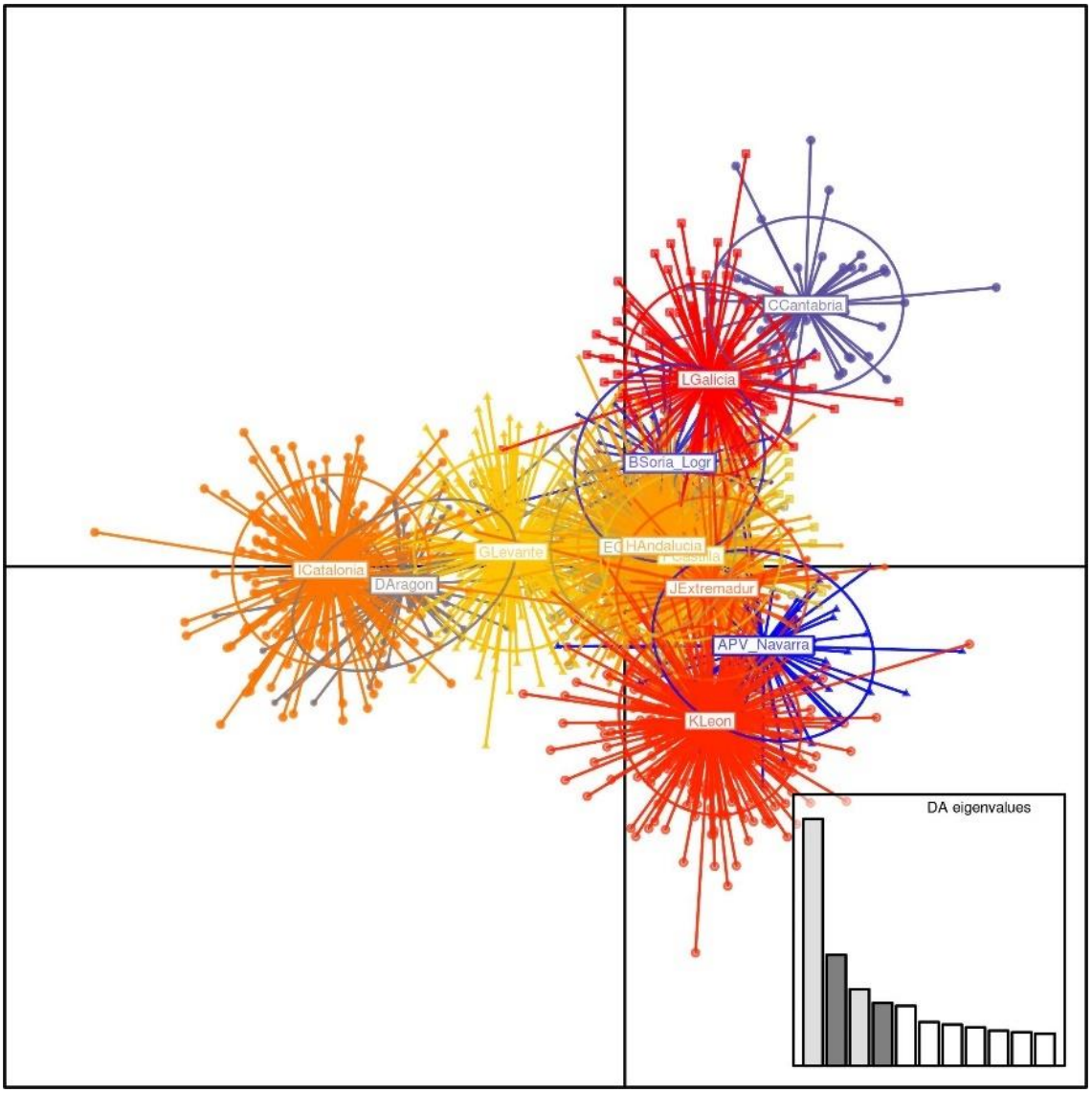


Figura A4. Caracterización de la estructura de nuestra población española de estudio en función de SNPs (LD2 y LD4).

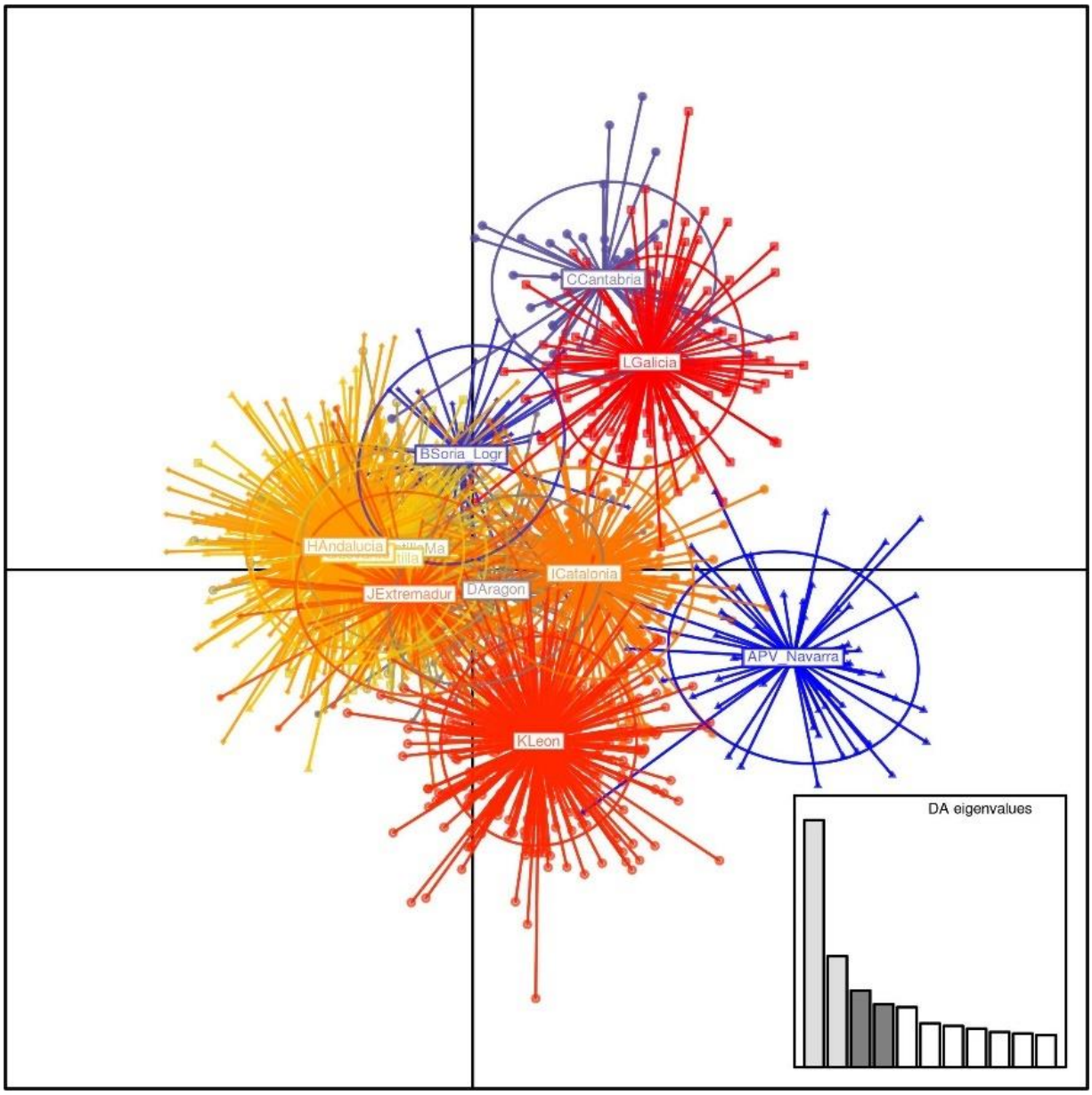


Figura A5. Caracterización de la estructura de nuestra población española de estudio en función de SNPs (LD3 y LD4).

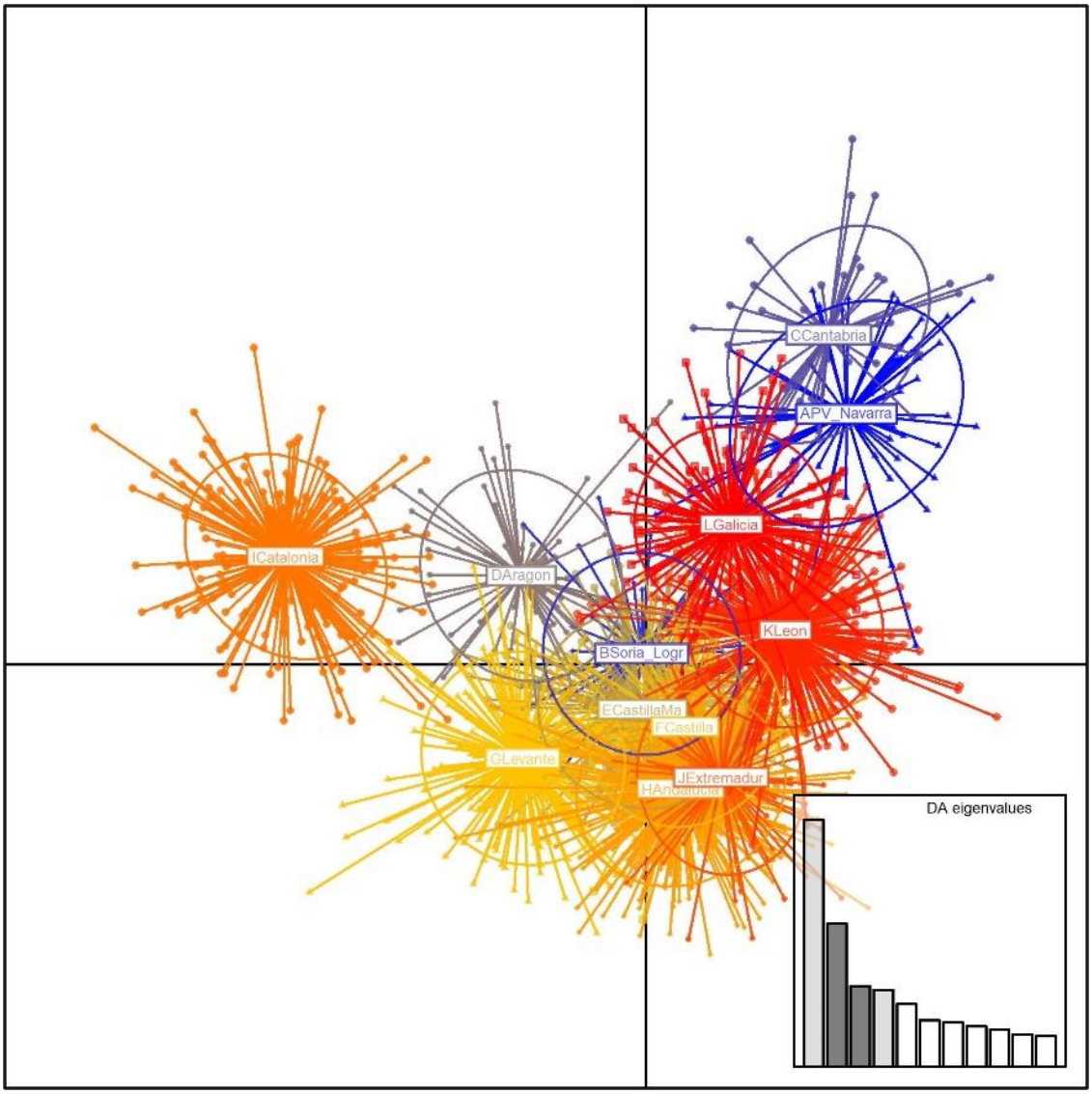


Figura A6. Caracterización de la estructura de nuestra población española de estudio en función de variantes de baja frecuencia (*low*); (LD2 y LD3).

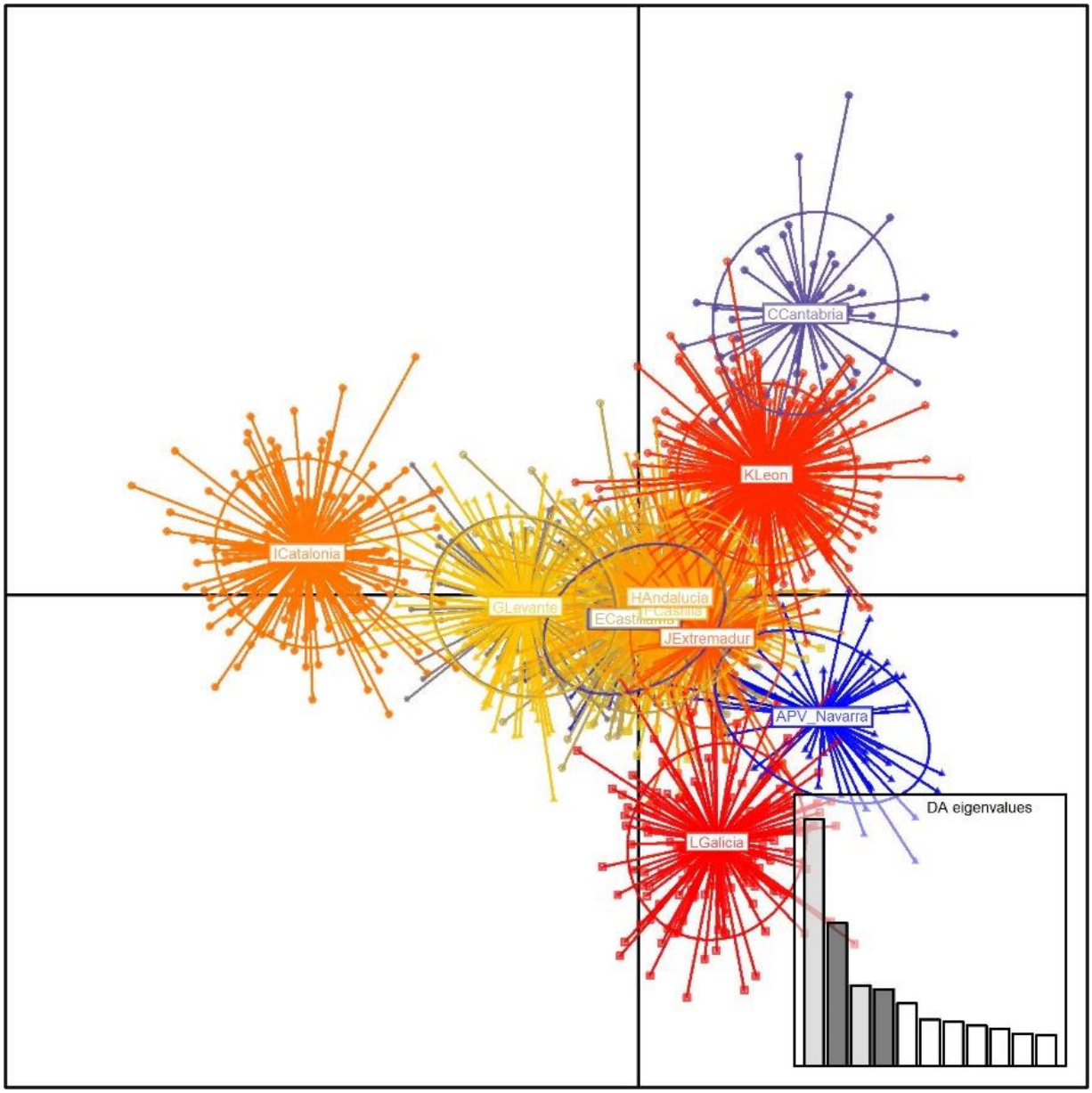


Figura A7. Caracterización de la estructura de nuestra población española de estudio en función de variantes de baja frecuencia (*low*); (LD2 y LD4).

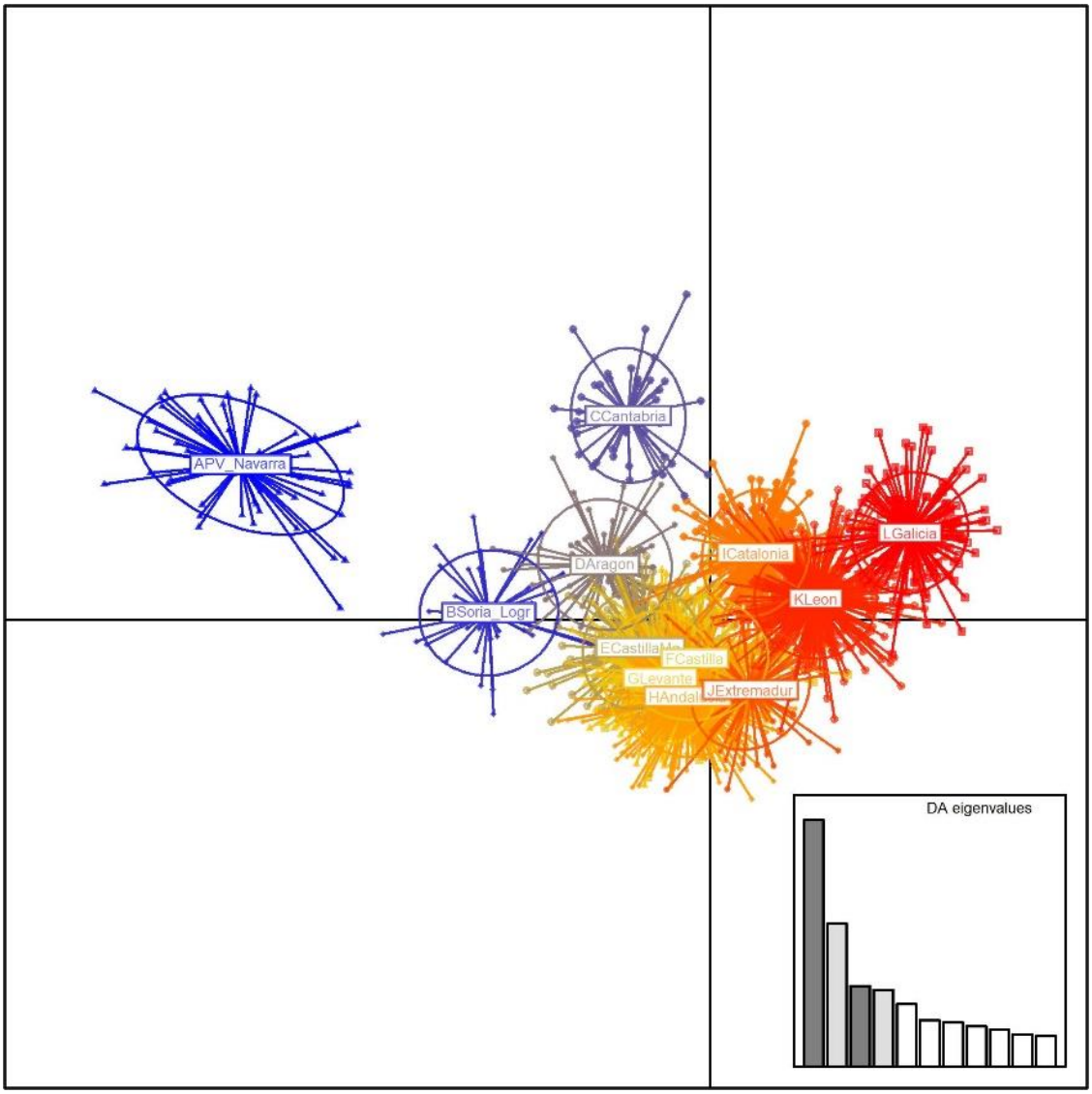


Figura A8. Caracterización de la estructura de nuestra población española de estudio en función de variantes de baja frecuencia (*low*); (LD1 y LD3).

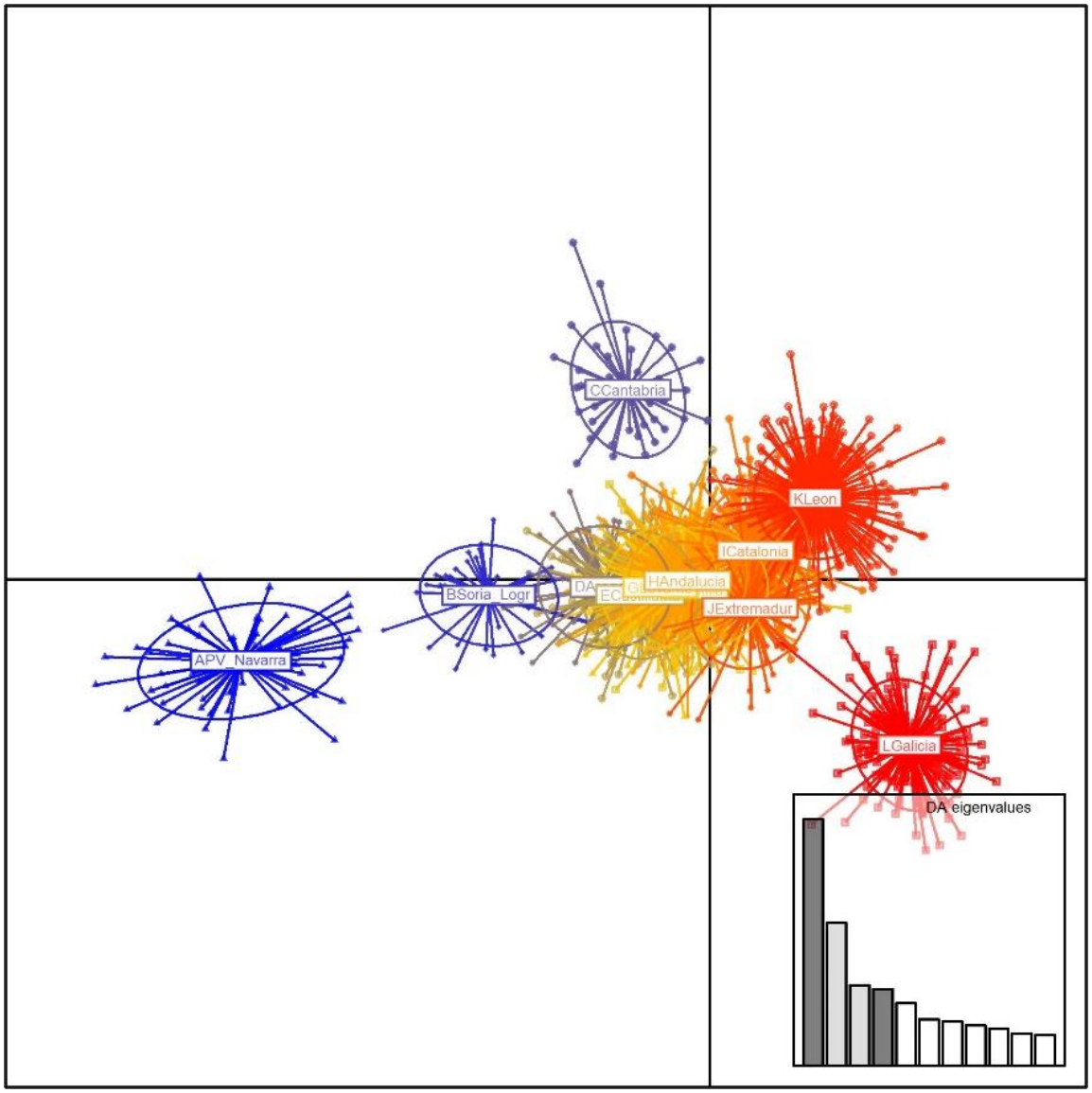


Figura A9. Caracterización de la estructura de nuestra población española de estudio en función de variantes de baja frecuencia (*low*); (LD1 y LD4).

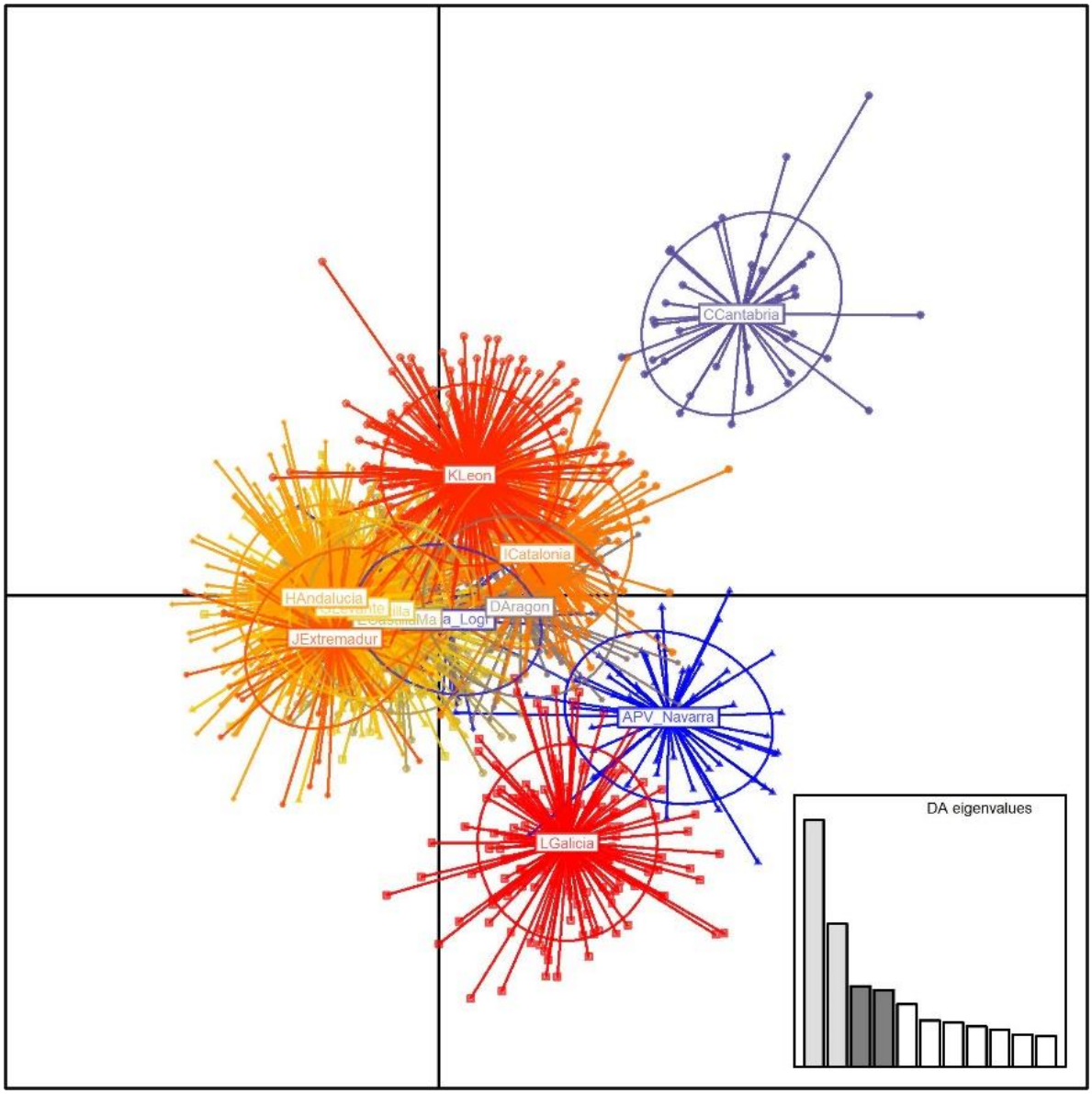


Figura A10. Caracterización de la estructura de nuestra población española de estudio en función de variantes de baja frecuencia (*low*); (LD3 y LD4).

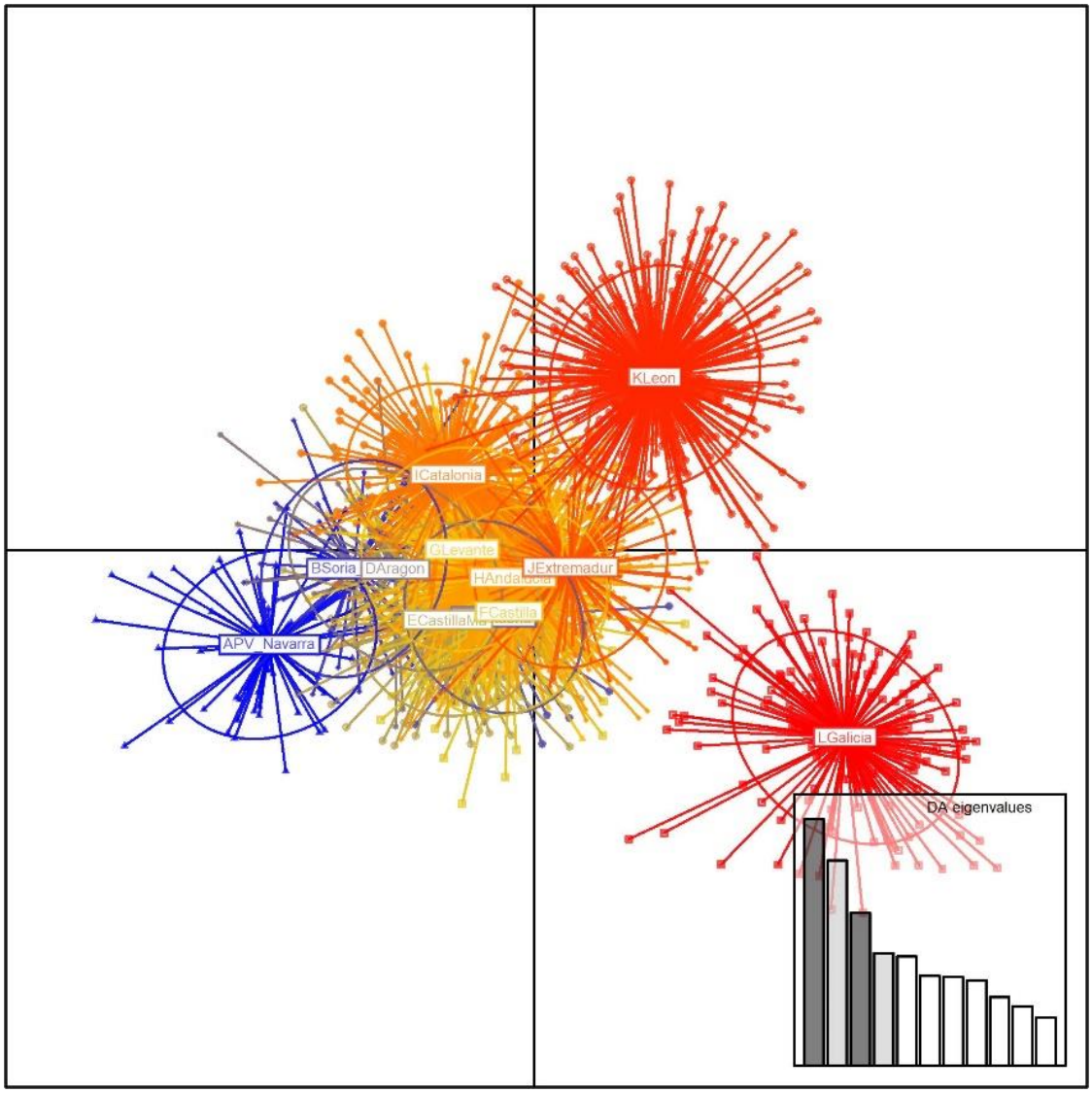


Figura A11. Caracterización de la estructura de nuestra población española de estudio en función de variantes raras (*rare*); (LD1 y LD3).

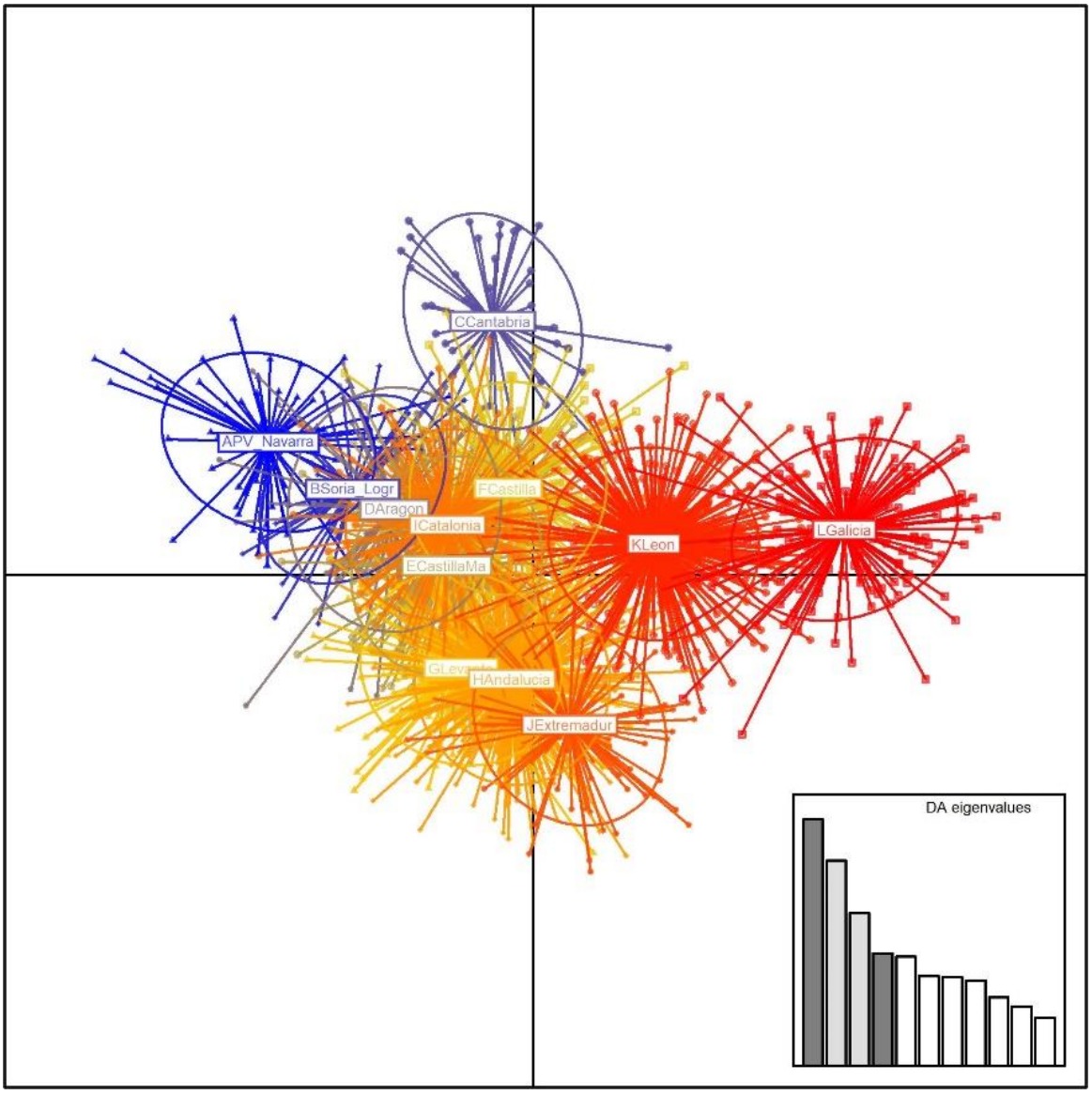


Figura A12. Caracterización de la estructura de nuestra población española de estudio en función de variantes raras (*rare*); (LD1 y LD4).

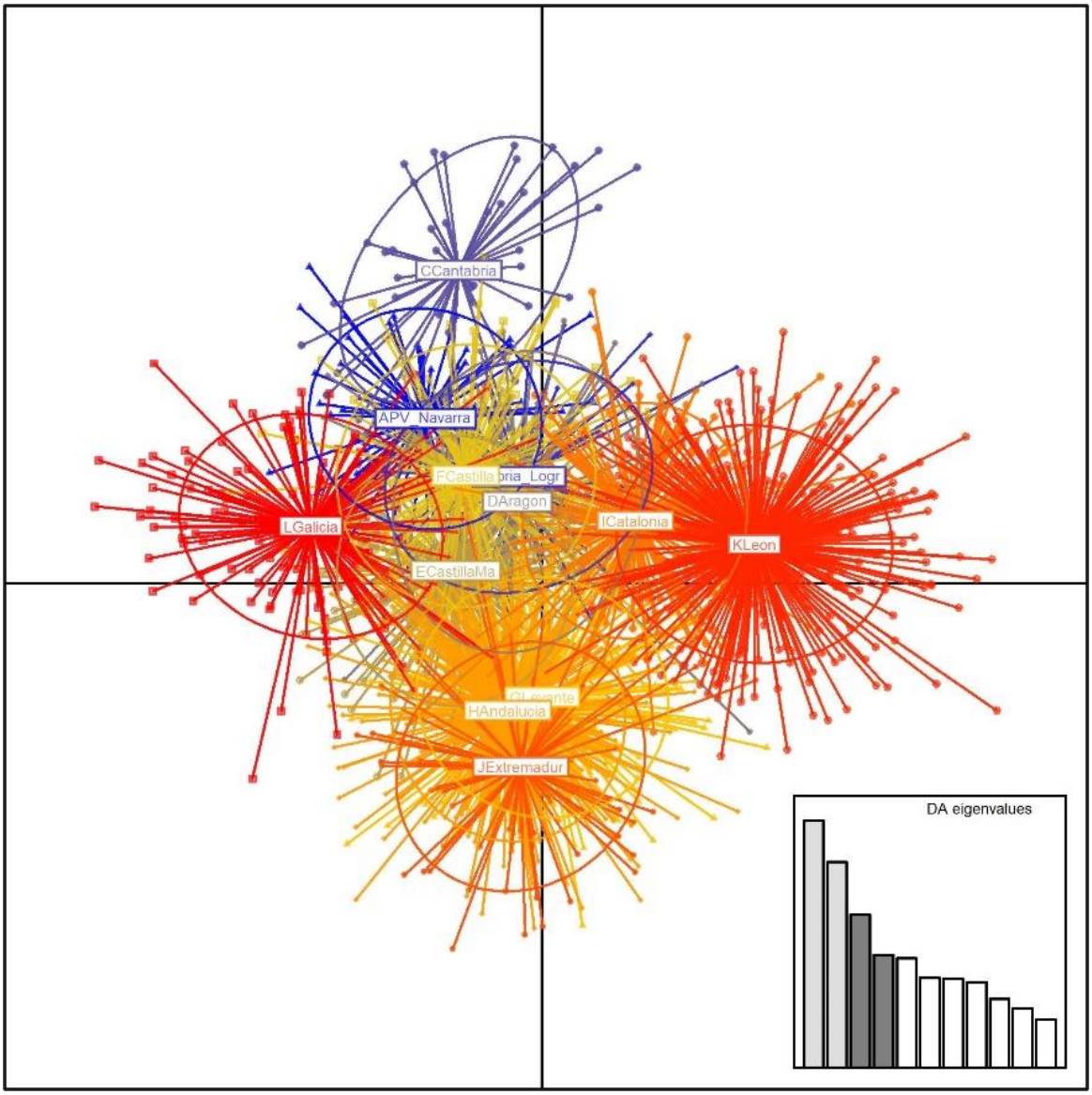


Figura A13. Caracterización de la estructura de nuestra población española de estudio en función de variantes raras (*rare*); (LD3 y LD4).

ANEXO III

ANEXO III

DISTRIBUCIÓN DE LOS MARCADORES RESPONSABLES DE LA DISCRIMINACIÓN POBLACIONAL DETECTADOS MEDIANTE IMPUTACIÓN

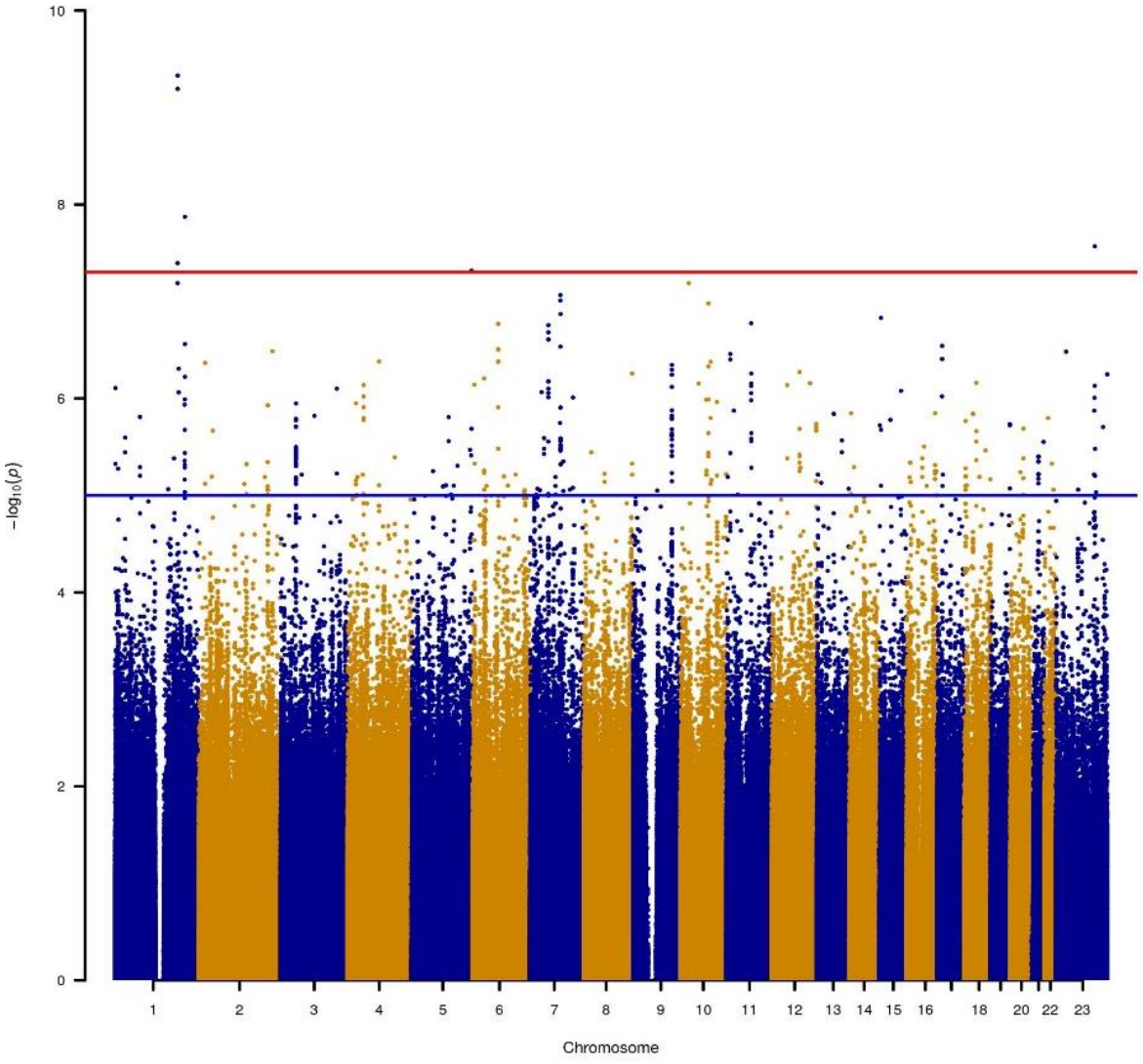


Figura A14. Distribución de marcadores responsables de la discriminación detectados en población "Aragón" mediante imputación.

En la imagen se pueden ver señales que diferencian a la población "Aragón" en los cromosomas 1, 5 y en el X, lo que no se apreciaba en nuestro análisis con datos de genotipos.

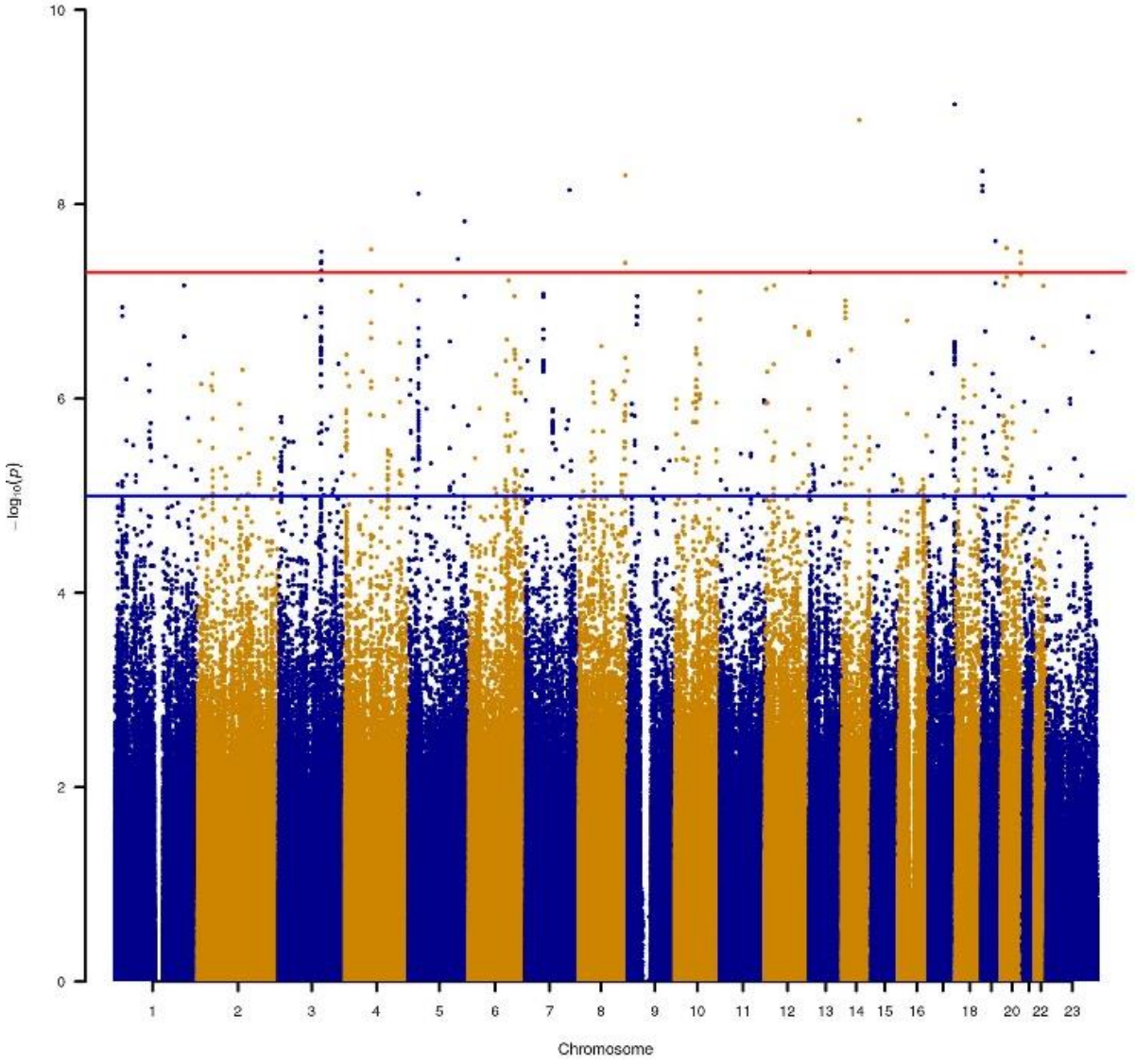


Figura A15. Distribución de marcadores responsables de la discriminación detectados población "Cantabria" mediante imputación.

En cuanto a la población "Cantabria" son detectables muchas más señales en la imagen que representa el GWAS con datos imputados, y esto lo veremos en todos los casos. Aquí, además de la señal en nuestra N genotipada en el cromosoma 5, se pueden ver otras en los cromosomas 7, 8, 14, 17, 19 y 20.

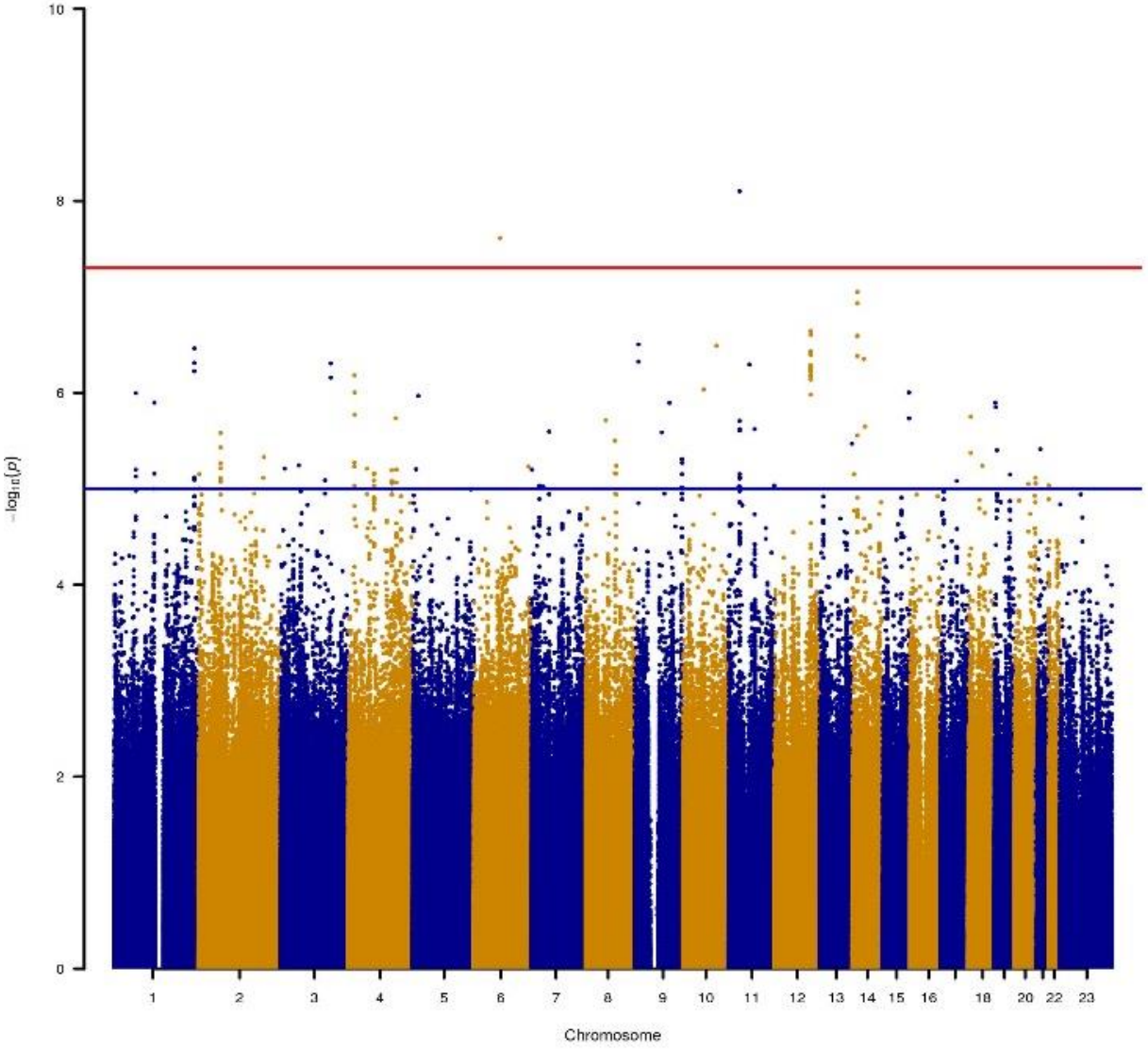


Figura A16. Distribución de marcadores responsables de la discriminación detectados en población "Castilla" mediante imputación.

En nuestra población "Castilla" no habíamos detectado ninguna señal significativa, mientras que en el GWAS mostrado podemos ver señales mínimas en los cromosomas 6 y 11.

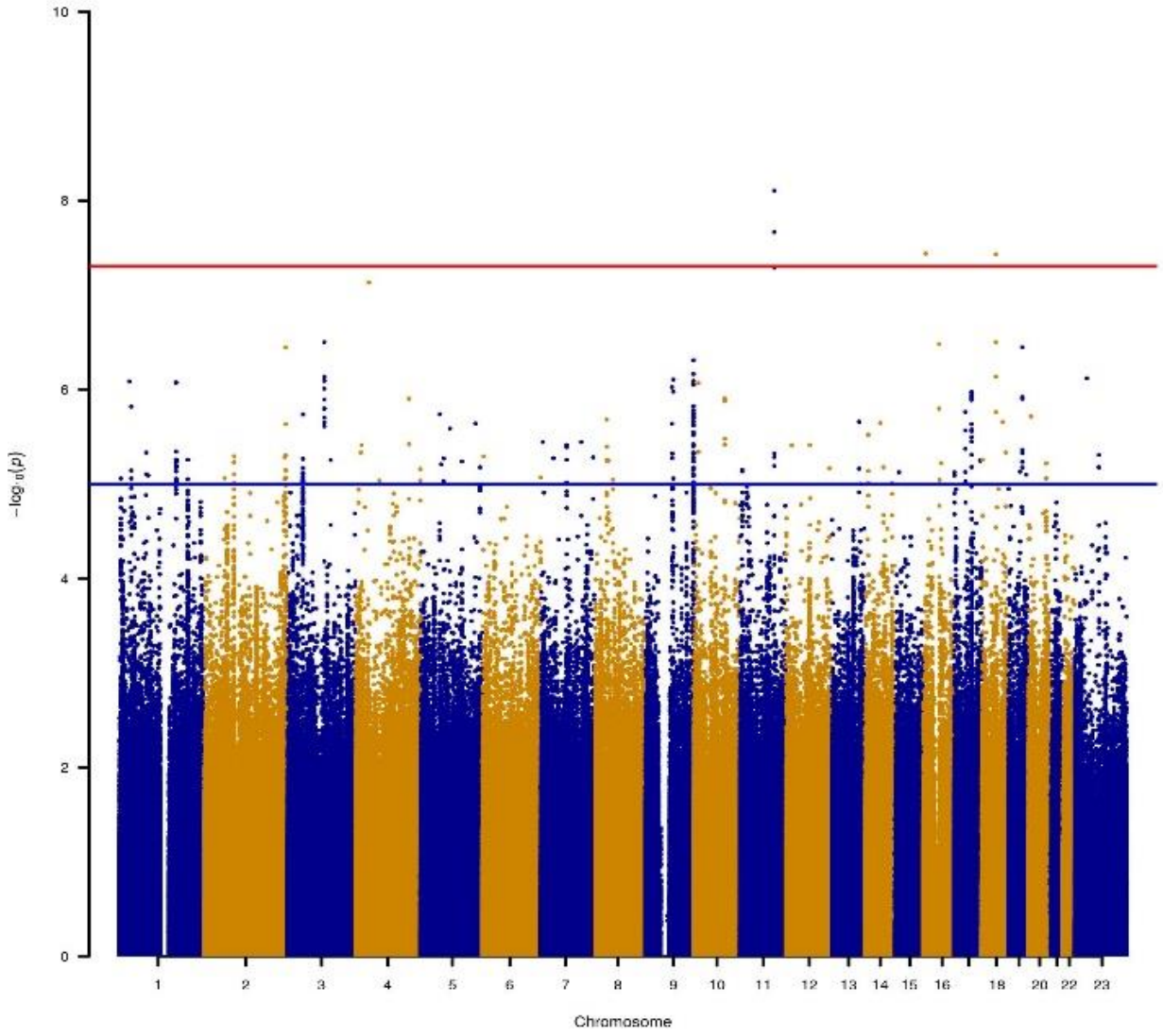


Figura A17. Distribución de marcadores responsables de la discriminación detectados en población "Castilla-Mancha" mediante imputación.

En nuestro análisis de la población "Castilla-Mancha" habíamos detectado una señal en el cromosoma 18, también encontrada en el GWAS con datos imputados. Además, en este último, se aprecian también marcadores en los cromosomas 11 y 16.

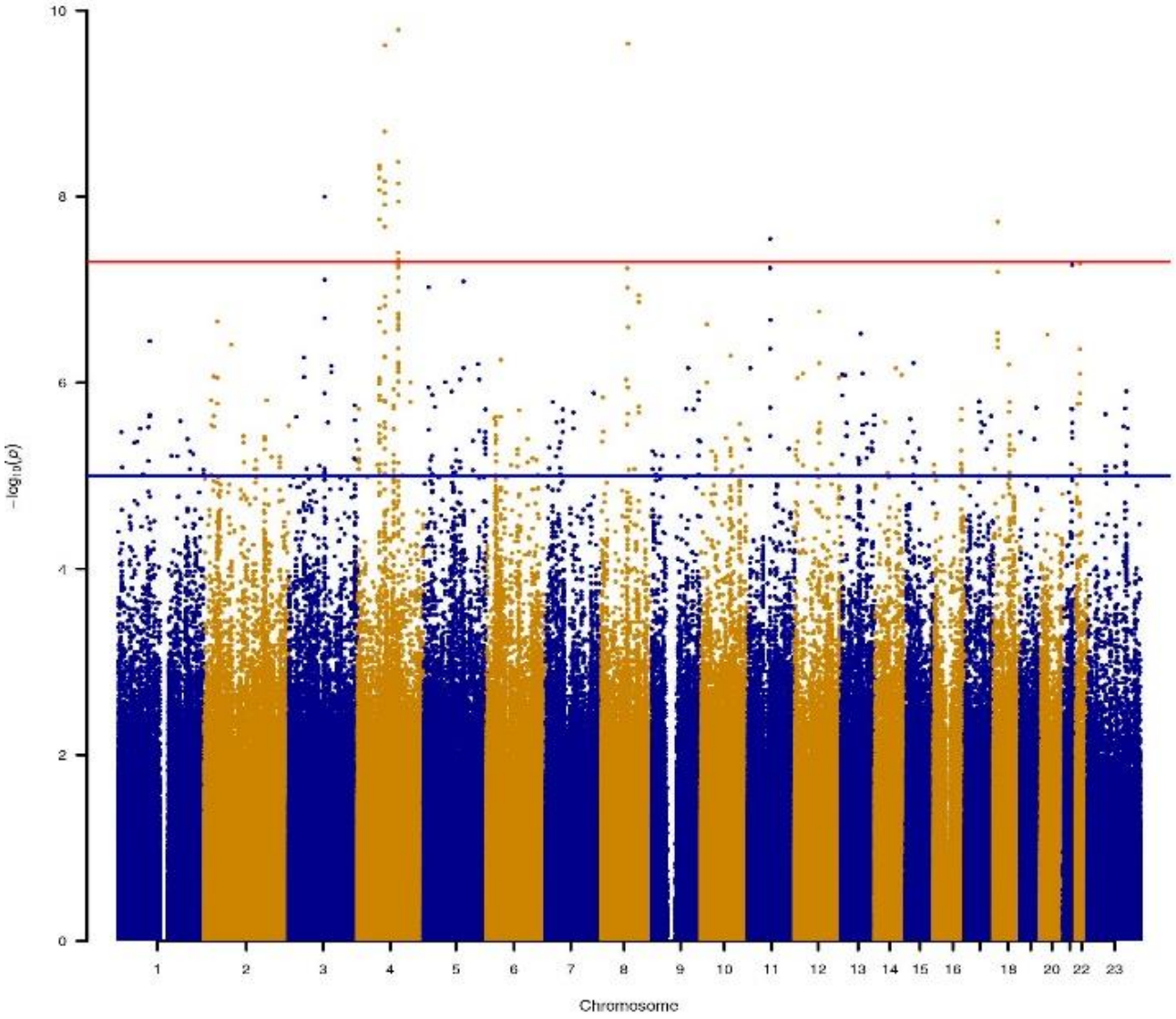


Figura A18. Distribución de marcadores responsables de la discriminación detectados en población "Soria-Logroño" mediante imputación.

En nuestra población "Soria-Logroño" pudimos ver señales en los cromosomas 3 y 4. En el MP con datos imputados se aprecia la misma señal en el cromosoma 3 y mucho más intensa en el cromosoma 4; además otras que no se habían detectado en los cromosomas 8, 11 y 18.

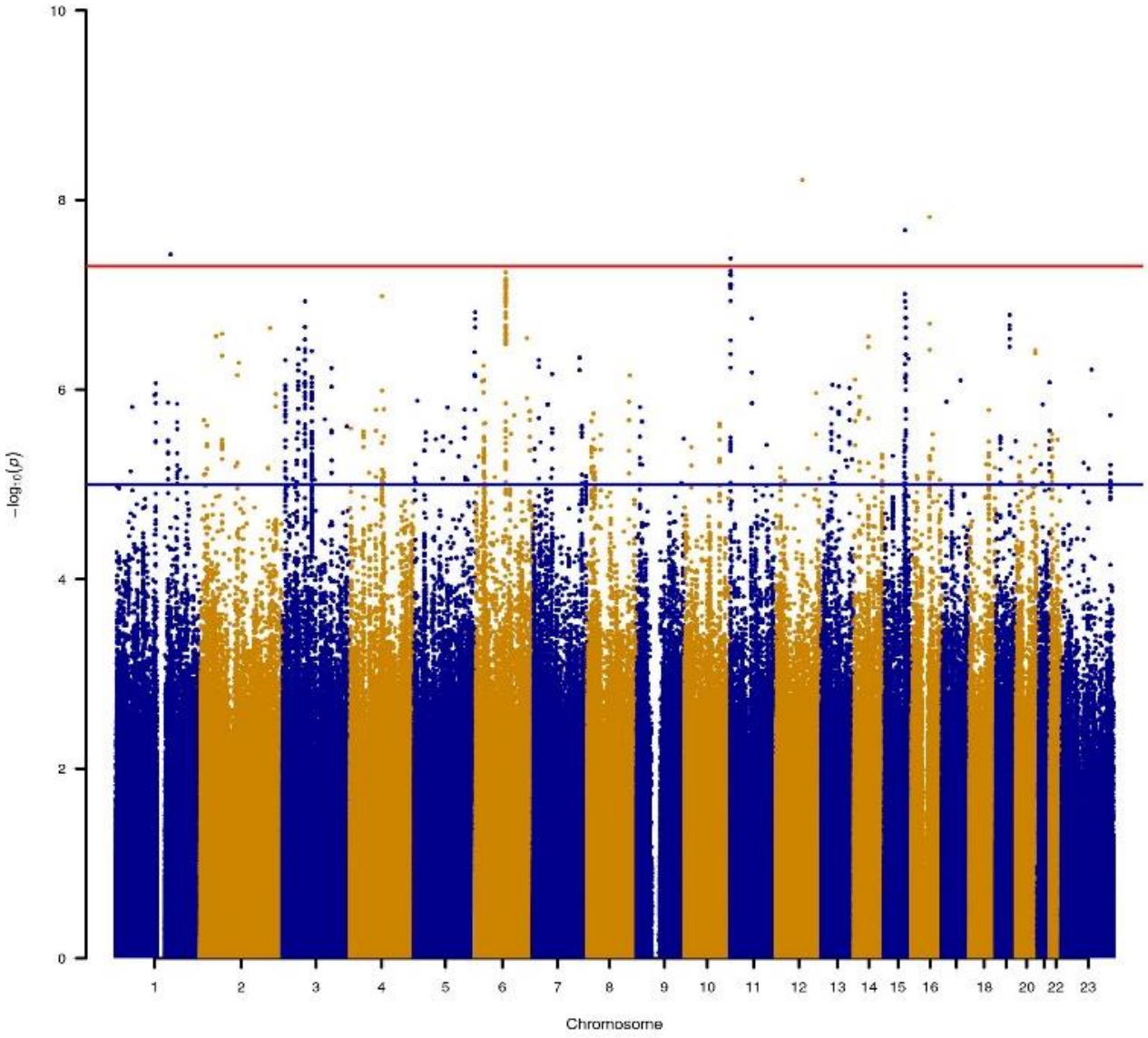


Figura A19. Distribución de marcadores responsables de la discriminación detectados en población "León" mediante imputación.

En nuestra población "León" no habíamos detectado ninguna señal. No así con datos imputados, donde se ve algún marcador significativo en los cromosomas 1, 12, 15 y 16.

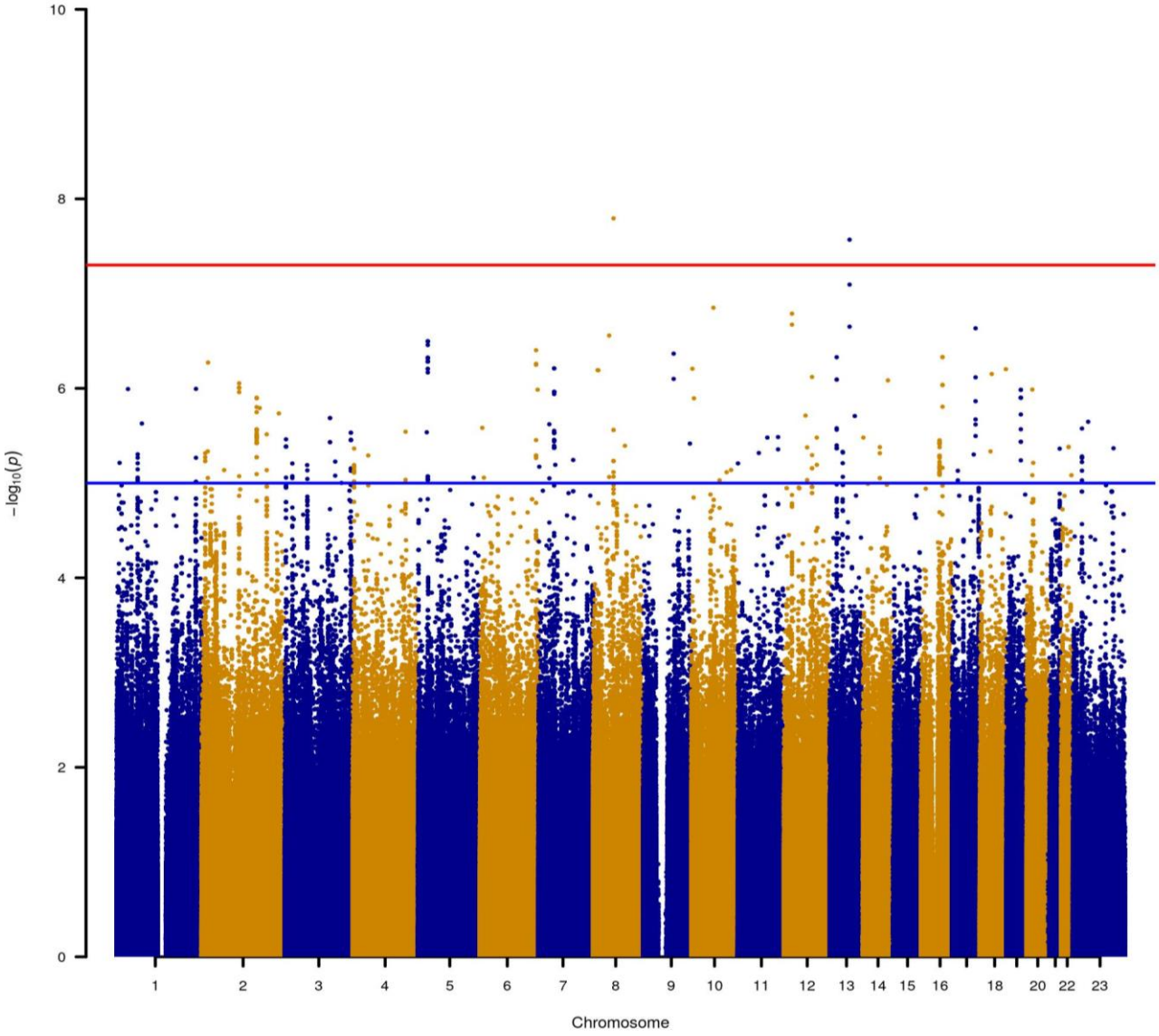


Figura A20. Distribución de marcadores responsables de la discriminación detectados en población "Levante" mediante imputación.

Al igual que en nuestra población "León", con "Levante" nos encontramos ante una situación similar, en la que a través de datos imputados se detecta algún marcador en los cromosomas 8 y 13.

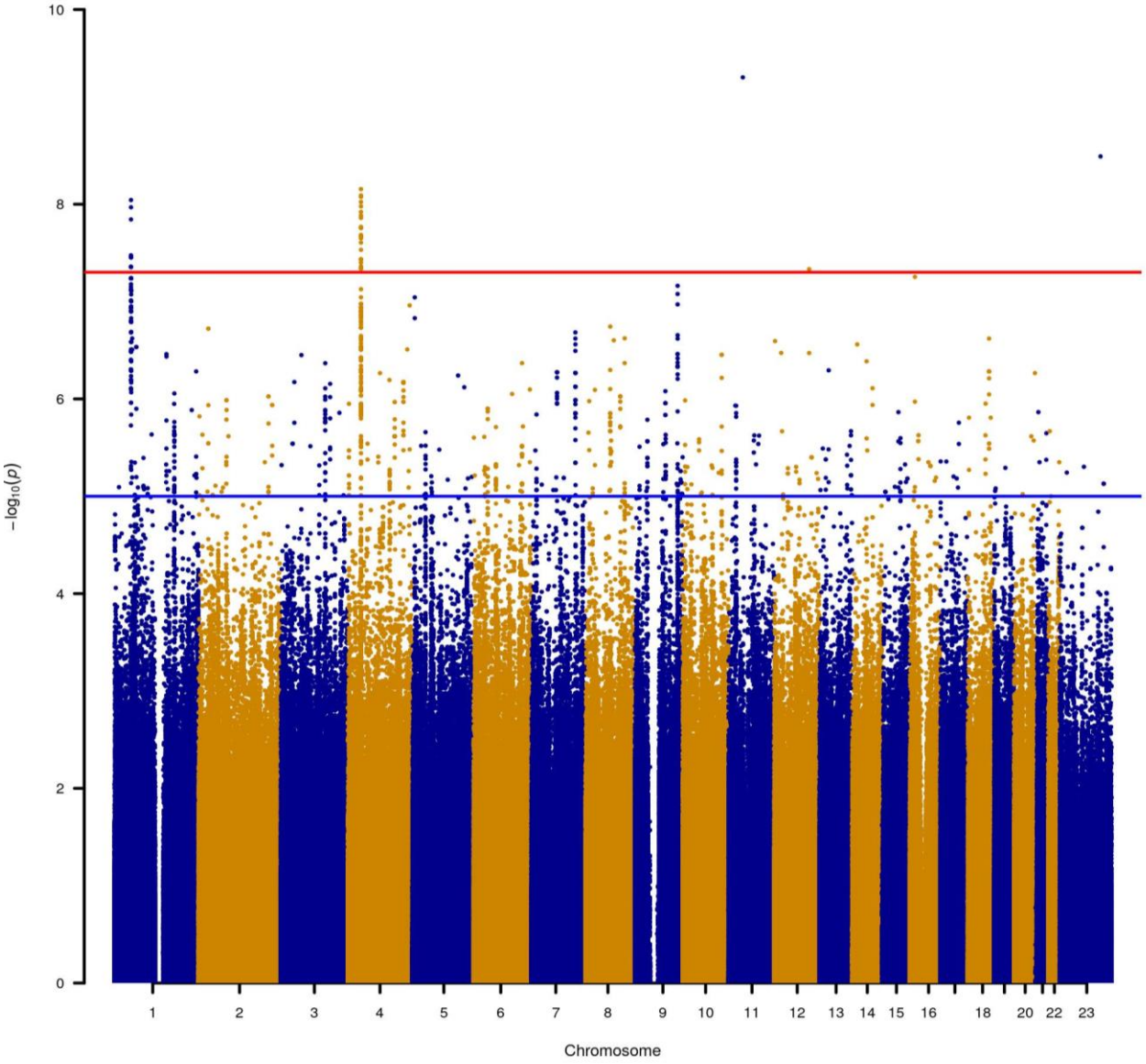


Figura A21. Distribución de marcadores responsables de la discriminación detectados en población "Galicia" mediante imputación.

En nuestra población gallega habíamos podido detectar señales significativas en los cromosomas 4 y 7 fundamentalmente, viendo un marcador aislado en el cromosoma 19. Mediante el GWAS de imputación se pusieron de manifiesto muchas más en el cromosoma 4 además de una señal importante en el cromosoma 1. No se aprecia la señal en el cromosoma 7 y aparecen marcadores aislados en los cromosomas 11 y X.

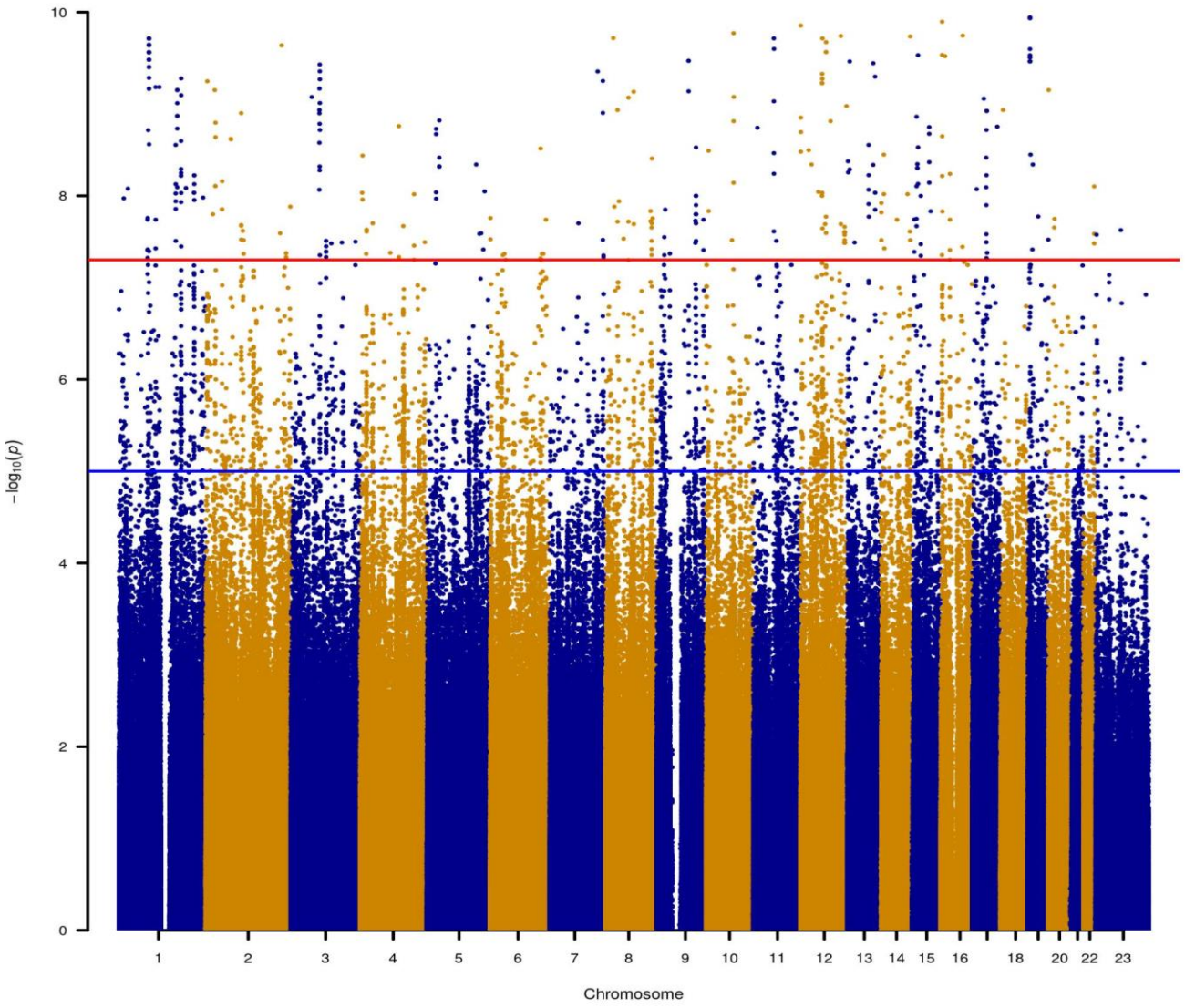


Figura A22. Distribución de marcadores responsables de la discriminación detectados en población "País Vasco-Navarra" mediante imputación.

Al igual que los resultados obtenidos en el análisis de nuestra población "País Vasco-Navarra", en el GWAS para imputaciones aparecieron muchas señales significativas que comprenden regiones de todos los cromosomas, lo que confirma nuestros resultados en cuanto a la subestructura poblacional de este *subset* frente al resto de la población española.

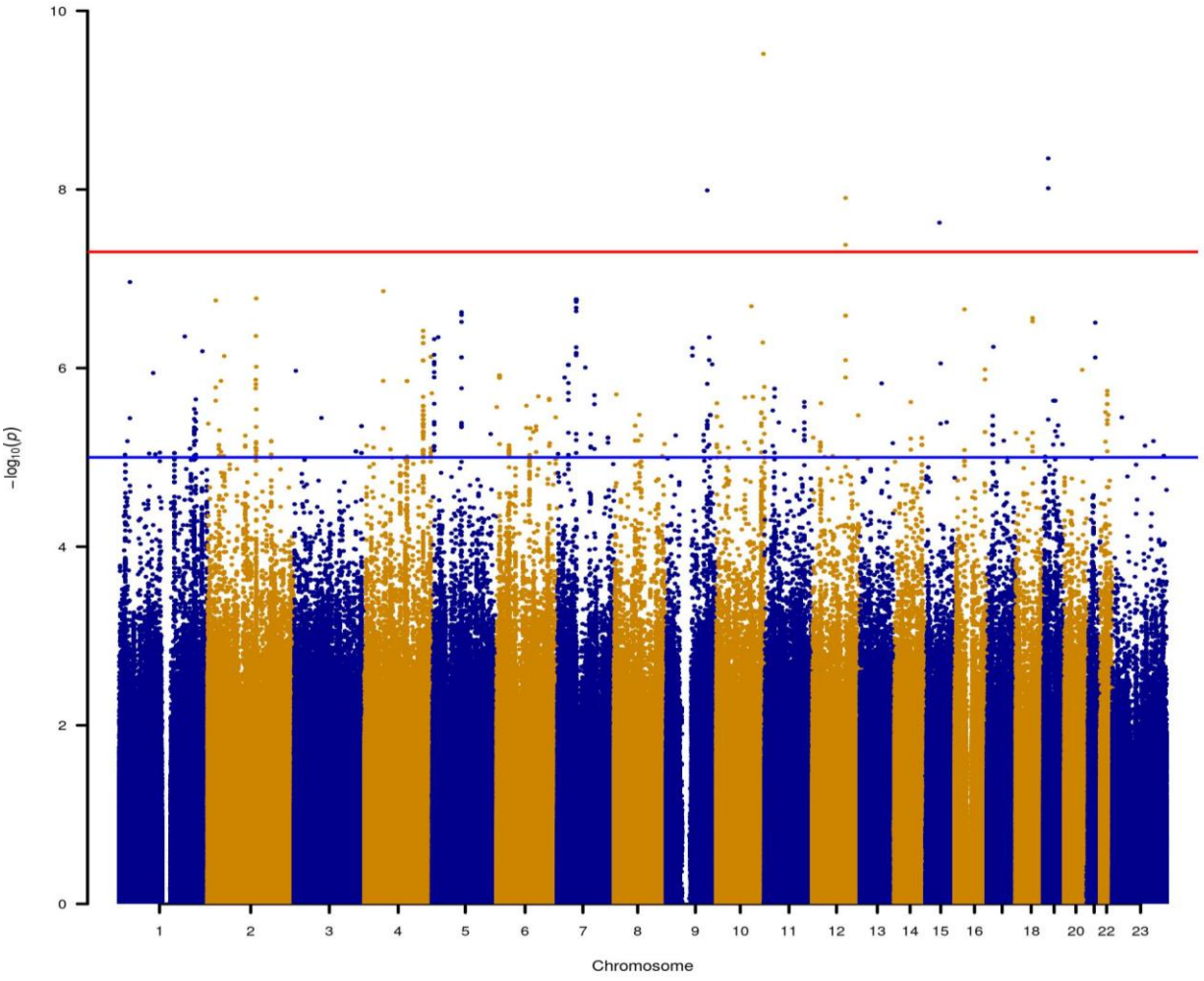


Figura A23. Distribución de marcadores responsables de la discriminación detectados en población "Cataluña" mediante imputación.

En la población "Cataluña" se pueden apreciar señales significativas en los cromosomas 9, 10, 12, 15 y 19, que no fueron detectados en nuestro análisis previo con datos de genotipos.

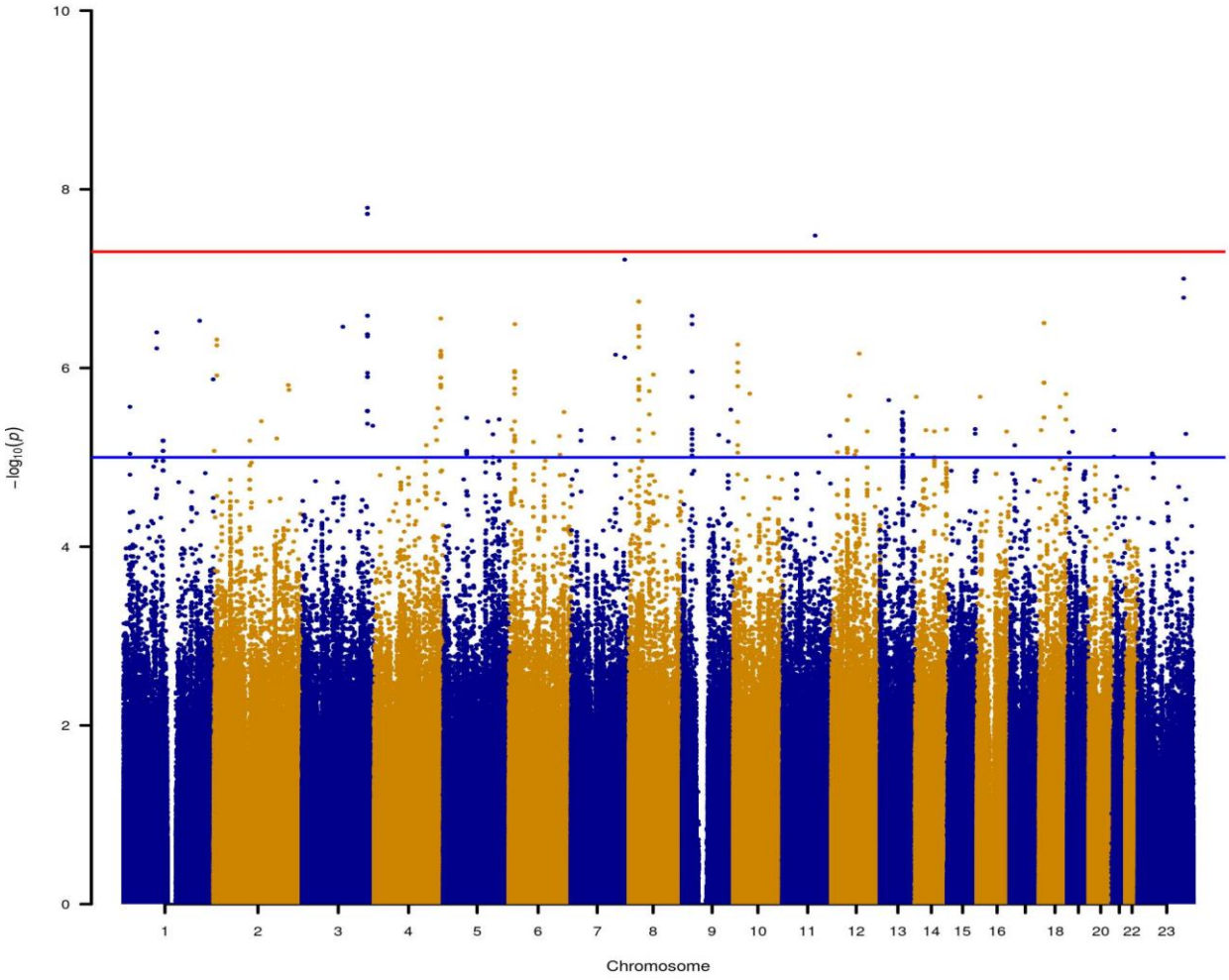


Figura A24. Distribución de marcadores responsables de la discriminación detectados en población "Extremadura" mediante imputación.

En el análisis de imputación con población "Extremadura" se pueden ver dos marcadores aislados en los cromosomas 3 y 11 que no habían sido captados con nuestros datos genotipados.

ANEXO IV

ANEXO IV

DEMANDA DEL *AXIOM SPAIN BIOBANK ARRAY PLATE* POR PARTE DE LA COMUNIDAD CIENTÍFICA, ASÍ COMO DE LOS DATOS DE GENOTIPADO DE NUESTRAS MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA

Los datos de genotipado de las muestras control obtenidos en este trabajo están ya disponibles para los investigadores que deseen genotipar su cohorte de casos. Estos datos han sido incorporados a un repositorio accesible a la comunidad científica para sus estudios de GWA.

En la Figura A25 se puede ver el número de proyectos solicitados con este *array* (2016 – julio 2023):

Figura A25. Número de proyectos solicitados al CeGen-FPGMX desde la optimización del *array*.

El chip ha sido utilizado por distintas instituciones (Tabla A2), tanto nacionales como internacionales.

Tabla A2. Instituciones solicitantes de proyectos de genotipado con el SBA y de las muestras control.

Institución solicitante
Consorcio Centro de Investigación Biomédica en Red (CIBERER)
Consorcio Centro de Investigación Biomédica en Red (CIBERSAM)
Fundació ACE: Institut Català de Neurociències
Fundació Clínic per la Recerca Biomédica
Fundació Docencia i Recerca Mutua Terrassa
Fundació Institut Mar d'Investigacions Mèdiques
Fundació Privada Institut d'Investigació Biomèdica de Bellvitge
Fundación Biomédica Galicia Sur
Fundación Canaria Instituto de Investigación Sanitaria de Canarias
Fundación Instituto de Investigación Marqués de Valdecilla
Fundación Instituto de Investigación Sanitaria de Santiago de Compostela
Fundación para la Investigación Biomédica del Hospital Gregorio Marañón
Hospital Universitario de Bellvitge
Institut d'investigacions Biomèdiques August Pi i Sunyer
Institut d'Investigació Sanitària Pere Virgili
Institut de Recerca Biomèdica de Lleida
Institut de Recerca l'Hospital de la Santa Creu i Sant Pau
Institut de Recerca Hospital vall d´Hebron
Institute of Molecular Pathology and Immunology of the University of Porto
Instituto de Investigaciones Biomédicas de Barcelona
Instituto de Investigación Biosanitaria de Granada
Instituto de Investigación Sanitaria Fundación Jiménez Díaz
Instituto de Investigación Sanitaria La Fe
Instituto Tecnológico y de Energías Renovables
Universidad Complutense de Madrid
Universidad Nacional de Educación a Distancia
Universidad de Valladolid
Universidade da Coruña
Universidade de Santiago de Compostela
Universitat Autónoma de Barcelona
Uniwersytet Jagielloński Collegium Medicum

Santiago de Compostela, 6 de febrero de 2024

Fe de errores anexada a la Tesis Doctoral "GENERACIÓN DE UN REPOSITORIO DE DATOS DE GENOTIPADO DE MUESTRAS DE POBLACIÓN CONTROL ESPAÑOLA Y DISEÑO DE UN NUEVO *ARRAY* PARA ESTUDIOS DE *GWAS*"

Autoría: María Soledad Otero Piñeiro
Dirección: Carracedo Álvarez, Angel Maria
Programa: Programa de Doutoramento en Medicina Molecular

Página 104: se especifica que se secuenciaron exomas de dos cohortes: la llamada cohorte G, compuesta por 115 individuos con origen gallego, y la llamada cohorte V, formada por 524 individuos de **población valenciana**. **Las 524 muestras mencionadas son originarias de varios puntos de España**, cuyos datos de exomas fueron recopilados en Valencia por el CSVS, (*Collaborative Spanish Variability Server*), y proceden del *Medical Genome Project*, del NAGEN 1000 (1000 genomas de Navarra), del proyecto *RareGenomics* de Madrid y otros grupos de investigación a lo largo de España.

Página 175: CONCLUSIÓN Nº5:
"La existencia de patrones de estratificación diferentes apunta la necesidad de aplicar los métodos de corrección diseñados y generalmente aplicados a estudios de GWAS. Por su parte las variantes raras responsables de la diferenciación poblacional aparecieron diseminadas por todo el genoma, lo que estaría acorde con su aparición más reciente, de forma que incluso aquellas variantes potencialmente deletéreas pueden no haber sido eliminadas aún por selección natural".

Decidimos eliminar la segunda parte por no ser una conclusión directa del trabajo y la palabra "diferente" en la primera frase por lo que quedaría como:
"La existencia de patrones de estratificación apunta la necesidad de aplicar los métodos de corrección diseñados y generalmente aplicados a estudios de GWAS".

Disculpas por los inconvenientes que ambos errores hayan podido causar.

Fdo.: Prof. Dr. Ángel Carracedo Álvarez
Tutor y Director

Fdo.: María Soledad Otero Piñeiro
Alumna de doctorado

La finalización del Proyecto Genoma Humano en 2003 y el Proyecto Internacional HapMap en 2005, ha supuesto la generación de nuevas herramientas de investigación que permiten analizar las contribuciones genéticas a enfermedades complejas. Una de las posibles complicaciones de los estudios de asociación es la existencia de estratificación, diferencias genético-poblacionales entre casos y controles que pueden sesgar la interpretación de los resultados. En esta tesis se ha diseñado un *array* específico de población española y se ha generado un repositorio de datos de genotipado de muestras de población control española que se harán accesibles a la comunidad científica para sus estudios de asociación del genoma completo (GWAS).