Trabajo Fin de Máster

# Comparación de técnicas estadísticas para evaluar el parentesco, estructura poblacional y ancestralidad genética.

Jorge Felpeto Evia

**Máster en Técnicas Estadísticas**

**Curso 2022-2023**

# Propuesta de Trabajo Fin de Máster

| |
|---|
| **Título en galego:** Comparación de técnicas estatísticas para avaliar o parentesco, estructura poboacional e ancestralidade xenética. |
| **Título en español:** Comparación de técnicas estadísticas para evaluar el parentesco, estructura poblacional y ancestralidad genética. |
| **English title:** Techniques statisticals comparison for evaluating relationship,population structure and genetic ancestry. |
| **Modalidad:** Modalidad B |
| **Autor:** Jorge Felpeto Evia, Universidad de Santiago de Compostela |
| **Director:** Manuel Febrero Bande, Universidad de Santiago de Compostela; |
| **Tutora:** Raquel Cruz Guerrero, CIBERER - U711. CIMUS; |
| **Breve resumen del trabajo:**<br>Se introducirán los análisis genéticos GWAS y la importancia del control de calidad en dichos estudios. Se definirán las técnicas estadísticas utilizadas y se aplicarán a un caso real. |
| **Recomendaciones:** |
| **Otras observaciones:** |

Don Manuel Febrero Bande, Catedrático del área de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela, doña Raquel Cruz Guerrero, Doctora del Grupo de Medicina Xenómica de CIBERER - U711. CIMUS, informan que el Trabajo Fin de Máster titulado

**Comparación de técnicas estadísticas para evaluar el parentesco, estructura poblacional y ancestralidad genética.**

fue realizado bajo su dirección por don Jorge Felpeto Evia para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 5 de junio de 2023.

FEBRERO BANDE MANUEL - 34961128R
Firmado digitalmente por FEBRERO BANDE MANUEL - 34961128R
Fecha: 2023.06.05 18:38:10 +02'00'

El director:
Don Manuel Febrero Bande

CRUZ GUERRERO RAQUEL - 12383132R
Firmado digitalmente por CRUZ GUERRERO RAQUEL - 12383132R
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-12383132R, givenName=RAQUEL, sn=CRUZ GUERRERO, cn=CRUZ GUERRERO RAQUEL - 12383132R
Fecha: 2023.06.05 14:16:16 +02'00'

La tutora:
Doña Raquel Cruz Guerrero

El autor:
Don Jorge Felpeto Evia

---

**Declaración responsable.** Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

# Agradecimientos

Esta memoria está dedicada a todas a aquellas personas que me apoyaron durante la evolución de este trabajo. Lo dedico a mi familia, a mis padres quienes me aconsejaron en los momentos difíciles y a mi hermano, quien fue la sorpresa de mi vida, dándome fuerzas en los momentos más tensos. A mis abuelos quienes se han preocupado por mi trayectoria durante estos meses y de quienes me aportaron útiles soluciones.

También quiero dedicar este trabajo a mis tutores. A Manuel Febrero Bande por su ayuda con Latex y la lectura contenido matemático de la memoria. También a mi tutora Raquel Cruz Guerrero y a Silvia Diz de Almeida , quienes me enseñaron el mundo de la genética partiendo de cero y me ayudaron a superar muchos obstáculos conceptuales de biología y matemáticas, son buenas investigadoras y tienen un potencial envidiable en la investigación y la enseñanza. Agradezco a mis compañeras del CIMUS, de quienes me siento orgulloso de haber estado en su departamento, además de Ángel Carracedo, un gran investigador que me enseñó la motivación de este campo y su amplio conocimiento.

Finalmente, quiero dedicar a mis amigos de la carrera de ADE, especialmente a Alejandro Fernández, Carlos Campos, Moisés Domínguez y a Liang Chen, quienes estuvieron siempre ahí apoyando desde mi comienzo con el máster. También a Berta, quién me ayudó a revisar las traducciones en inglés de la presente memoria y a quién agradezco su ayuda. Finalmente, dedico esta memoria a mis compañeras del piso de Santiago de este curso, a quienes siempre daba gusto escuchar después de un día extenuante de prácticas y clases y por su preocupación durante mi estancia.

# Índice general

# Resumen


## Resumen en español

Esta memoria estudiará la confusión del parentesco y la ancestralidad poblacional en individuos españoles y latinoamericanos adscritos al Proyecto SCOURGE. Se utilizará el GWAS, cuyo objetivo en este caso es evaluar el riesgo de padecer con mayor o menor intensidad los síntomas del virus respiratorio SARS-CoV-2, en base a las características genéticas de la población de estudio. El GWAS requiere una fase de control de calidad para excluir variantes con valores faltantes o individuos emparentados que distorsionen las asociaciones genéticas. Existen varios algoritmos o métodos que permiten inferir las relaciones de parentesco genético en base a la información genotípica de cada individuo. Además, se definirán los tres métodos de estimación del parentesco: PLINK, KING-robust y PC-Relate. A partir de su comparación, se podrá determinar cuál es el más preciso o adecuado para el control de calidad del GWAS partiendo de la población de estudio y utilizando las técnicas estadísticas convenientes.

## English abstract

This paper will examine the confusion of kinship and population ancestry in Spanish and Latin American individuals enrolled in the SCOURGE Project. The study will utilize GWAS, with the objective of assesing the risk of experiencing symptoms of the respiratory virus SARS-CoV-2 to a greater or lesser extent, based on the genetic characteristics of the study population. GWAS requires a quality control phase to exclude variants with missing values or related individuals that may distort genetic associations. There are several algorithms or methods that allow for the inference of genetic kinship relationships based on the genotypic information of each individual. In addition, three methods of estimating kinship will be defined: PLINK, KING-robust and PC-Relate. By comparing them, it will be possible to determine which one is the most accurate or suitable for GWAS quality control based on the study population and using appropriate statistical techniques.

# Introducción

Los procedimientos GWAS[1] son de uso frecuente en el campo de la Genética de Poblaciones y de la Bioestadística. En la actualidad, especialmente gracias a los avances en la investigación en Medicina Molecular, se ha incrementado el número de artículos de investigación referentes a los GWAS, cuyo objetivo es el análisis del efecto de los factores genéticos de un individuo con respecto a la incidencia de diversas enfermedades u otros trastornos.

### Conceptos básicos de genética

La información genética necesaria parte de la información del ADN[2], definiéndose como el elemento de almacenaje de información hereditaria que caracteriza a un individuo. Permite crear nuevos componentes de las células, como proteínas y moléculas de ARN[3] y tiene forma de doble hebra. Cada segmento o secuencia de ADN, definido como gen, está formado por conjuntos de pares de bases nitrogenadas, cuyas funciones son la uniformidad e integridad de los genes. Las bases serían: adenina(A), guanina(G), timina(T) y citosina(C).

El marcador genético o biomarcador forma parte de un segmento de ADN, cuya ubicación física se denomina *locus* o *loci* en plural. Los biomarcadores permiten situar la proximidad de varios segmentos de ADN, siendo más probable que se hayan heredado conjuntamente si sus *loci* son próximos. Los marcadores genéticos denominados SNPs[4], son variaciones genéticas de las secuencias de ADN en una sola base de un *locus* determinado. Con respecto a la representación de la información genética, las bases están representadas en genotipos o valores genotípicos [5] y los factores ambientales u externos que influyen genéticamente en el individuo se denominan fenotipos. Todos los conceptos genéticos aquí señalados fueron introducidos gracias a la literatura de la memoria (Solari 2004).

### Introducción al GWAS

Los GWAS permiten identificar relaciones o asociaciones entre genotipos y fenotipos a partir de las diferencias en las frecuencias alélicas de las denominadas variaciones genéticas o SNPs entre individuos. El protocolo de elaboración de un GWAS consta de ocho fases, como se observa en la literatura reciente (Uffelmann *et al.* 2021). En la Figura 1, se puede observar esquemáticamente el desarrollo del procedimiento de un GWAS donde, a partir de la tercera fase, el orden de los pasos puede divergir según los criterios de la investigación, así como el número de pasos a seguir. El procedimiento detallado es el siguiente:

[1] *Genome-Wide Association Studies*
[2] Ácido desoxirribonucleico
[3] Ácido ribonucleico traducido al español
[4] Polimorfismos de nucleótidos únicos. Para considerar una variación genética un SNP, es necesario que al menos el 1 % de la población lo comparta, en otro caso se considera una mutación aislada.
[5] Clasificación mediante símbolos de un tipo de variante genética en un *locus* determinado del genoma

1. **Recogida de los datos**

   Se selecciona la cohorte[6] para la extracción de datos a través de recogida directa y analizada en laboratorio o a través de repositorios o biobancos de datos.

2. **Genotipado y definición de los datos**

   Se codifican los datos recogidos pertenecientes a la cohorte y se identifican los SNPs para cada individuo. Los datos genotipados son categóricos, cada valor observado en un alelo es una de las cuatro categorías de bases nitrogenadas que puede contener un segmento de ADN en un *locus* dado[7]. En términos de la disposición de los datos, están representados en una tabla de doble entrada, donde cada fila corresponde a un individuo de la muestra y cada columna es un SNP, conteniendo a su vez dos subcolumnas que representan a cada alelo o gen con sus correspondientes bases.

3. **Control de calidad**

   A partir de los datos genotipados en la fase anterior, se realiza el control de calidad de los datos por SNPs y por individuos de la muestra. A continuación, se identifican a los individuos emparentados y se analizan sus ancestralidades a partir del Análisis de Componentes Principales y del análisis de la estratificación poblacional.

4. **Imputación**

   Los SNPs que no fueron codificados en las fases anteriores, se les imputan genotipos en base a la información de la estratificación poblacional de una población de referencia.

5. **Test de asociación genética**

   En este paso, se ajustan los modelos de asociación genética para cada SNP , tales como las regresiones aditivas o logísticas. En base a la estratificación poblacional detectada en el control de calidad, se corregirá la confusión de los SNPs observando valores o patrones inusuales y se realizarán contrastes y análisis estadísticos oportunos.

6. **Metaanálisis**

   Se combinan resultados de cohortes más pequeñas usando protocolos estadísticos estandarizados.

7. **Replicación**

   Se replican los resultados obtenidos en otras cohortes independientes. En el caso de replicación de una cohorte externa, no puede tener parientes o ancestralidad común con la cohorte replicadora.

8. **Análisis Post-GWAS**

   Se comparan los resultados con otras fuentes externas y se realiza la validación de los datos con otras técnicas experimentales.

### Contextualización de la memoria

En referencia al caso práctico estudiado en la memoria, se analizará el efecto de los factores genéticos en la severidad de la Covid-19 o SARS-CoV-2[8] a través de un GWAS. Desde que se detectó el primer caso de Covid-19, se registraron 676.609.955 casos diagnosticados en el mundo y 6.881.955 fallecimientos por este virus según los datos recogidos de la Universidad Johns Hopkins .En el caso de España, se recogieron 13.770.429 casos activos de Covid-19 y 119.479 fallecimientos a partir de los

[6] Población de interés en la investigación.
[7] Las bases son: Guanina(G),Citosina(C),Adenina(A) y Timina(T).
[8] Virus coronario respiratorio que se registró por primera vez en China en diciembre de 2019.La OMS decretó la emergencia sanitaria el 11 de marzo de 2020.

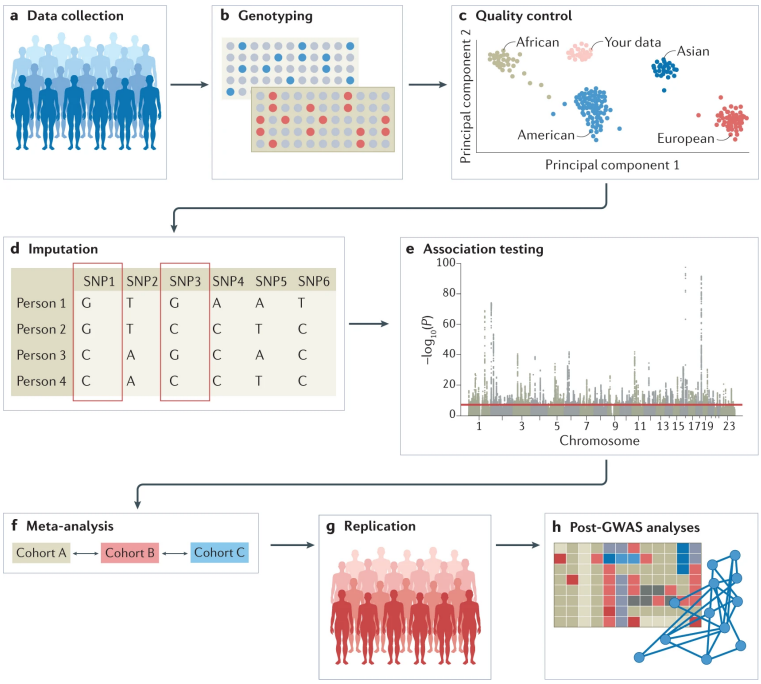


Figura 1: Fuente: Overview of steps for conducting GWAS. Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., *et al.* Genome-wide association studies. *Nature Reviews Methods Primers.*

datos de la misma fuente de información[9].

Existen numerosas investigaciones sobre el estudio de los efectos fenotípicos en la severidad del SARS-CoV-2. Algunos estudios sugieren que los efectos fenotípicos como el nivel de exposición o el tabaco están conectados con la gravedad de los síntomas (Niemi 2022), e incluso también se plantea que su severidad se ve influida por otros factores como el sexo y la edad del individuo (Cruz *et al.* 2022).

Desde el CIMUS[10], se trabaja con la metodología mencionada y se investiga en el campo de la Medicina Molecular. Todos los datos recogidos en el presente trabajo se recogieron bajo la dirección del Proyecto SCOURGE y analizados en el CIMUS. Este proyecto de colaboración investigadora internacional tiene como misión "encontrar biomarcadores de evolución y pronóstico que puedan tener un impacto inmediato en el manejo clínico y en las decisiones terapéuticas en pacientes infectados por SARS-CoV-2" (Spanish COalition to Unlock Research on host GEnetics on COVID-19 (SCOURGE) s.f.).

### Estructura de la memoria

La memoria constará de dos capítulos, siendo el primero la definición de los métodos de estimación y la presentación de las técnicas estadísticas utilizadas para el caso práctico. En el primer capítulo, se definirá la fase de control de calidad de un GWAS, además del funcionamiento teórico de cada uno de los métodos de estimación del parentesco genético: PLINK, KING y sus algoritmos KING-homo y KING-robust y PC-AiR junto con PC-Relate. Como extensión, se comentarán los detalles teóricos de

[9] Para más información, consultar su página web, así como la explicación de la herramienta de visualización de los datos utilizada (Dong *et al.* 2020) (Dong *et al.* 2022).

[10] *Center for Research in Molecular Medicine and Chronic Diseases*. Centro de investigación que colabora con la Universidad de Santiago de Compostela, para más información, visitar su página web.

los modelos PLS y PLSGLR[11] como técnicas sugeridas para evaluar las medidas de precisión de los métodos estimativos del parentesco mencionados.

El segundo capítulo se enfoca en el estudio de un caso práctico con la información genética real de una cohorte analizada en el CIMUS. Se presentarán los datos ya genotipados y el método de extracción de estos, para posteriormente realizar el control de calidad del GWAS. Como primer avance en el análisis estadístico, se representarán gráficamente los resultados que se obtienen de la inferencia de parentesco. Dicha representación se realizará para cada método de estimación bajo dos escenarios: considerando o no el *LD pruning*. Como comprobación del funcionamiento de los tres métodos, se identificarán los núcleos de parientes cercanos desde tercer grado hasta los de primer grado. La segunda parte del capítulo se centrará en el estudio de la estratificación poblacional a través del Análisis de Componentes Principales de cada método, así como la correlación de las componentes principales con las tres estructuras poblacionales de interés: europeas, africanas y nativo-americanas. La última parte del capítulo hará uso de los modelos PLS y PLSGLR para medir la precisión de los métodos en base a la estimación de las estructuras poblacionales. Primeramente, se centrará en considerar un modelo de regresión con variables respuesta numéricas en los modelos PLS. Finalmente, se realizará un modelo de clasificación a partir de cierto umbral fijado de la variable respuesta binaria para considerar a un individuo perteneciente a una estructura poblacional en los modelos PLSGLR. El último apartado de este capítulo constará de la discusión de los resultados y posibles conclusiones, así como limitaciones que pudieran surgir en la memoria.

[11] *Partial Least Squares* y *Partial Least Squares Generalized Linear Regression*, traducido al español: Mínimos Cuadrados Parciales y Regresión Lineal Generalizada por Mínimos Cuadrados Parciales respectivamente.

# Capítulo 1

# Materiales y métodos

A continuación, se presentan los métodos utilizados en la memoria con sus respectivas explicaciones teóricas, cuya notación teórica está descrita y recogida en la Tabla A.1 del Apéndice. Las siglas y abreviaturas de la memoria están recogidas en la Tabla A.2 del Apéndice.

## 1.1. Control de calidad en estudios GWAS

Los estudios de asociación genética caso-control o GWAS son de los más utilizados para extraer información sobre los datos genotípicos que pueda ofrecer una población de estudio. Entre todas sus fases, el control de calidad permite presentar resultados coherentes y correctos del GWAS a partir del genotipado de los datos, formulándose un protocolo que permite llevar a cabo dicho procedimiento, detallado en el artículo existente (Anderson *et al.* 2010). De forma adicional, se incluye la aplicación de un ejemplo práctico donde se explica profundamente toda la ejecución en PLINK del control de calidad además del resto de fases del GWAS en el material relacionado (Marees *et al.* 2018). El control de calidad se basa en dos vías de actuación: el control por individuos y el control por marcadores o SNPs. De este modo, se pretende filtrar aquellos SNPs inválidos para el estudio y excluir a aquellos individuos que puedan estar emparentados, provocando correlaciones genéticas que distorsionen los modelos de asociación genética del GWAS.

#### Control de calidad por individuos

A continuación, se detalla el procedimiento de control por individuos de la muestra de la cohorte como:

1. Identificación y posterior exclusión de individuos con información de sexo ambiguo.
2. Filtrado de individuos con baja tasa de genotipado o con déficit o exceso de heterocigosidad con respecto al promedio de la población de estudio.
3. Identificación de individuos emparentados o duplicados para su posterior exclusión.
4. Evaluación de la existencia de estratificación poblacional a partir de Análisis de Componentes Principales y análisis de la ancestralidad.

#### Control de calidad por marcadores o SNPs

Con el objetivo de obtener un conjunto de marcadores independientes con los que evaluar el parentesco y estratificación en las muestras, se realiza paralelamente al control por individuos con los siguientes pasos:

1. Se seleccionan los autosomas[1].
2. Se fija un umbral del MAF[2], para la selección de SNPs.
3. Se realiza un chequeo para filtrar los valores perdidos o no codificados(*missing values*).
4. Se considera el Equilibrio de Hardy-Weinberg, cuyo p-valor medio será el umbral de filtrado de SNPs.
5. Excluir a aquellos SNPs con alto Desequilibrio de Ligamiento[3], seleccionando un subconjunto de SNPs independientes.

## 1.2. Introducción al modelo de Hardy-Weinberg

El modelo de Hardy-Weinberg permite calcular las frecuencias genotípicas a partir de las frecuencias alélicas[4] (Kalmes y Huret 2001). Considerando, dentro de una población, una pareja alélica $A$ y $a$ como alelos heredados de cada uno de los progenitores del individuo en un *locus* dado, se presentan las frecuencias alélicas y sus restricciones:

- $p$ : *frecuencia del alelo* $A$ $\quad 0 \leq p \leq 1$.
- $q$ : *frecuencia del alelo* $a$ $\quad 0 \leq q \leq 1$.
- $p + q = 1$.

Las frecuencias genotípicas para un individuo en un *locus* dado se representan como:

- $p^2$ : *frecuencia conjunta de los alelos* $AA \rightarrow Homocigoto$.
- $2pq$ : *frecuencia conjunta de los alelos* $Aa \rightarrow Heterocigoto$.
- $q^2$ : *frecuencia conjunta de los alelos* $aa \rightarrow Homocigoto$.

Se consideran estas frecuencias constantes para los descendientes del individuo, sean homocigotos o heterocigotos.[5]

Se consideran las frecuencias alélicas para ambos sexos, y si se añade la generación siguiente, se obtienen las frecuencias genotípicas en un *locus* determinado, definiendo el modelo de Hardy-Weinberg en la Ecuación 1.1

$$(p + q)^2 = p^2 + 2pq + q^2 = 1. \tag{1.1}$$

[1] Cromosomas no sexuales, que no contengan ni el cromosoma X ni el cromosoma Y.

[2] *Minor Allelic Frequency* traducido del inglés, se define como la frecuencia del alelo menos común en un determinado *locus* de una población. Se suele utilizar para identificar variantes o SNPs raros frente a variantes más frecuentes.

[3] *Linkage Desequilibrium* o por sus siglas en inglés LD, es la asociación de dos genes de *loci* distintos, que se presentan juntos con mayor frecuencia que si fueran escogidos al azar en un cromosoma dado. Incumplen la condición de independencia estadística y deben ser excluidos para considerar solamente SNPs independientes. Véase para más información en (Nordborg y Tavaré 2002).

[4] Proporción de un alelo con respecto al número total de alelos de un determinado *locus* dentro de una población.

[5] Un individuo es homocigoto si sus dos alelos de un cromosoma son idénticos (o que posee dos copias idénticas del mismo gen),en caso contrario será heterocigoto.

El equilibrio de Hardy-Weinberg[6] es un modelo teórico basado en la Ecuación 1.1 que se basa en las hipótesis siguientes:

- La población es panmíctica: Todos los individuos tienen la misma probabilidad de aparearse , y el apareamiento es al azar.
- La población es suficientemente grande: Permite minimizar las diferencias entre los individuos.
- La población no está sometida a la migración, mutación o selección(no se pierden o se ganan alelos).
- Las frecuencias génicas y genotípicas se mantienen constantes para cada generación siguiente.

Este modelo teórico será necesario para la inferencia de parentesco en el control de calidad de los GWAS, ya que los softwares utilizados para este estudio asumen las condiciones del HWE.

## 1.3. Inferencia de parentesco: PLINK

La inferencia de parentesco considera dos individuos con sus respectivos alelos o genes, seleccionándose un SNP único para ambos y en el mismo *locus*. El objetivo será estudiar las diferencias de las bases y así estimar la afinidad o parentesco entre ellos genéticamente a partir de los datos genotipados.

PLINK es un software especializado en análisis y procesamiento de datos genéticos, donde existen numerosos comandos de análisis estadístico y de control de calidad que ayudan a la interpretación de los resultados. Actualmente, es uno de los programas más utilizados para gran cantidad de investigaciones en genética, especialmente en el control de calidad y en la elaboración de tests de asociaciones genéticas. Dentro de la fase de control de calidad, es necesario identificar individuos que tengan algún vínculo genético que puedan distorsionar los resultados finales del GWAS. Cuando dos individuos tienen alelos IBS[7], implica que comparten bases idénticas de sus respectivos alelos en un SNP dado, independientemente de su descendencia. En el caso de los alelos IBD[8], se tiene en cuenta la ascendencia común de los individuos. Este razonamiento conlleva a que dos individuos que compartan alelos IBD sí son IBS, pero no tiene por qué darse a la inversa. Según el desarrollo explicado en (Purcell *et. al* 2007), una vez realizado el control de calidad previo de los datos genotipados, se procede con la inferencia de parentesco a través de la estimación de la probabilidad de que dos individuos compartan alguna copia genotípica IBD en sus respectivos alelos. La estimación de la probabilidad de que dos individuos contengan alelos IBD se obtiene de la estimación de la probabilidad de que dos individuos compartan alelos IBS.

Se denota como $I$ y $Z$ el número de alelos considerados IBS e IBD de un par de individuos respectivamente en un SNP $m \in \mathcal{M}$, siendo $\mathcal{M}$ un conjunto de SNPs seleccionados sin valores perdidos.

Asumiendo que tanto $I$ como $Z$ pueden tomar solamente tres valores en los que $l, z = \{0, 1, 2\}$, la probabilidad anterior de que dos individuos posean dos alelos IBS en un SNP dado queda reflejada en la Ecuación 1.2

$$P(I = l) = \sum_{z=0}^{z=l} P(I = l)|Z = z)\ P(Z = z), \tag{1.2}$$

calculada en términos de las frecuencias alélicas y aplicable para todos los SNPs seleccionados. Calculando previamente la esperanza matemática de $P(I = l|Z = z)$ para todos los SNPs, se determina la

[6] *Hardy-Weinberg Equilibrium* o por sus siglas en inglés HWE.
[7] *Identity By State*, o Identidad/Idénticos por Estado traducido del inglés.
[8] *Identity By Descendent*, traducido es: Identidad/Idénticos por Descendencia.

Ecuación 1.2 en base a las probabilidades $P(Z=0), P(Z=1)$ y $P(Z=2)$. De esta forma, se puede obtener la probabilidad de que dos individuos compartan alelos IBD en un SNP dado bajo la Ecuación 1.3

$$\hat{\pi} = \frac{P(Z=1)}{2} + P(Z=2). \tag{1.3}$$

La probabilidad de que los alelos de dos individuos tengan alelos IBD en un SNP, requiere de una corrección de sesgo del estimador $\pi$ basado en el recuento de alelos. Las estimaciones de $\pi$ se realizan en la práctica bajo condiciones de incertidumbre, y por tanto sobre una muestra generando cierto sesgo si no se corrige (Purcell *et. al* 2007, pág. 565).

A partir de la definición de la Ecuación 1.3, se calculan las probabilidades de que dos individuos compartan ningún, uno o dos copias genotípicas IBD en sus respectivos alelos en un SNP. Previamente, se considera la esperanza matemática del número de SNPs cuando los alelos son IBS $I=l$, expresada como $E[P(I=l|Z=z)] = \sum_{m=1}^{|\mathcal{M}|} P(I=l|Z=z)$, cumpliéndose si $E[P(I=0)] = \sum_{m=1}^{|\mathcal{M}|} P(I=l)$ y $E[P(Z=z)] = \sum_{m=1}^{|\mathcal{M}|} P(Z=z)$. Estas probabilidades globales calculadas por el método de los momentos se definen respectivamente en las Ecuaciones 1.4, 1.5 y 1.6 respectivamente.

$$P(Z=0) = \frac{E[P(I=0)]}{E[P(I=0|Z=0)]} \tag{1.4}$$

$$P(Z=1) = \frac{E[P(I=1)] - P(Z=0)E[P(I=1|Z=0)]}{E[P(I=1|Z=1)]} \tag{1.5}$$

$$P(Z=2) = \frac{E[P(I=2)] - P(Z=0)E[P(I=2|Z=0)] - P(Z=1)E[P(I=2|Z=1)]}{E[P(I=2|Z=2)]} \tag{1.6}$$

Las Ecuaciones 1.4, 1.5 y 1.6 y la Ecuación 1.3 serán utilizadas para la inferencia de parentesco, siendo estas probabilidades valores diferentes dependiendo del grado de parentesco del par de individuos. Se incluye en el Apéndice la tabla A.3 con los valores teóricos de cada probabilidad para cada grado de parentesco.

## 1.4. Inferencia de parentesco: KING

El software KING[9] permite realizar controles de calidad del GWAS basados en la estimación del coeficiente kinship, denotado como $\phi_{ij}$, en vez de la proporción de alelos considerados IBD $\pi_{ij}$ para cada par de individuos $i$ y $j$. Siguiendo la literatura relacionada (Manichaikul *et al.* 2010), asumiendo que la estructura poblacional no se conoce previamente, se define el coeficiente kinship como la probabilidad de que existan copias genotípicas IBD en los alelos de dos individuos muestreados aleatoriamente. Además, este estimador también es la mitad de la proporción de alelos IBD, demostrándose en la Ecuación 1.7

$$2\phi_{ij} = \frac{\pi_{1ij}}{2} + \pi_{2ij} = \pi_{ij}, \tag{1.7}$$

siendo $\pi_{1ij}$ y $\pi_{2ij}$ las probabilidades de que dos individuos tengan una o dos copias genotípicas IBD en sus alelos respectivamente en un SNP, denotadas alternativamente como $P(Z=1)$ y $P(Z=2)$.

[9] *Kinship-based Inference for Genome-wide association studies.*

### 1.4.1. KING-homo

El algoritmo KING-homo utiliza métodos de estimación del parentesco bajo la asunción de homogeneidad de la población[10]

Al igual que en PLINK, KING-homo también utiliza las frecuencias alélicas para la estimación de la probabilidad de tener alelos IBS e IBD. La proporción esperada de alelos IBS igual a cero se muestra en la Ecuación 1.8

$$P(I=0) = P(AA, aa|Z=0)\ P(Z=0) = 2p^2(1-p)^2\pi_{0ij}, \tag{1.8}$$

siendo $P(AA, aa)$ la probabilidad de que dos individuos sean homocigotos sin ninguna copia genotípica IBS y $\pi_{0ij}$. A partir del desarrollo de las probabilidades de los alelos IBS, se obtiene la definición del estimador $\hat{\pi}_{0ij}$ en la Ecuación 1.9

$$\hat{\pi}_{0ij} = \frac{\sum_m I_{ij}^m = 0}{\sum_m 2\hat{p}_m^2(1-\hat{p}_m)^2} = \frac{N_{AA,aa}}{\sum_m 2\hat{p}_m^2(1-\hat{p}_m)^2}, \tag{1.9}$$

siendo $I_{ij}^m = 0$ la variable indicadora del recuento de pares de individuos cuando no existen copias genotípicas IBS en sus alelos correspondientes para un $m$ SNP y $N_{AA,aa}$ su notación equivalente. $\hat{p}_m$ es la frecuencia alélica media estimada para cada SNP $m$.

Tal y como se observa en la Ecuación 1.9, se desconoce el valor real de la frecuencia alélica $p_m$ para cada $m$ SNP . En consecuencia, se estima a partir de la muestra SNPs sin genotipos con valores perdidos en la Ecuación 1.10

$$\hat{p}_m = \frac{\#AA + \#Aa/2}{\#AA + \#Aa + \#aa}, \tag{1.10}$$

siendo $\#AA, \#aa, \#Aa$ el recuento de individuos totales genotipados siendo homocigotos del alelo de referencia, homocigotos del segundo alelo u heterocigotos respectivamente.

El estimador del coeficiente kinship se calcula a partir de las distancias genéticas de los alelos de referencia de dos individuos. Se asumen las condiciones del Equilibrio de Hardy-Weinberg en la Ecuación 1.11

$$E(X^{(i)} - X^{(j)})^2 = 4p(1-p)(1-2\phi_{ij}), \tag{1.11}$$

siendo $X^{(i)}$ y $X^{(j)}$ variables categóricas que representan el número de copias genotípicas definidas respecto al alelo de referencia para el individuo $i$ y $j$ respectivamente. Estas variables toman los valores detallados en el siguiente esquema para el individuo $i$ y análogo para el individuo $j$:

- $X_m^{(i)} = 2$ si el individuo es homocigoto respecto al alelo de referencia. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor 2 si los dos alelos tienen genotipos idénticos $AA$ en un SNP $m$ dado.
- $X_m^{(i)} = 1$ si el individuo es heterocigoto respecto al alelo de referencia. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor 1 si los dos alelos tienen genotipos diferentes $Aa$ en un SNP $m$ dado.
- $X_m^{(i)} = 0$ si el individuo es homocigoto respecto al otro alelo. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor 0 si los dos alelos tienen genotipos idénticos respecto al otro alelo $aa$ en un SNP $m$ dado.

Las variables $X_m^{(i)}$ y $X_m^{(j)}$ poseen las siguientes características bajo las condiciones del Equilibrio de Hardy-Weinberg:

[10] La población de estudio presenta una única ascendencia o estructura poblacional común.

- $E[X_m^{(i)}] = 2p_m$, análogo para $X_m^{(j)}$
- $Var[X_m^{(i)}] = 2p_m(1-p_m)$, análogo para $X_m^{(j)}$

Ahora bien, si se consideran los alelos de cada par de individuos simultáneamente, será útil definir las distancias genéticas entre los alelos de cada individuo. Estas distancias pueden tomar tres valores$\{0, 1, 2\}$ en valor absoluto, en función de las diferencias entre los genotipos de los individuos , pudiéndose considerar heterocigotos u homocigotos. En la Ecuación 1.12 se reflejan los valores de las distancias genéticas en valor absoluto.

$$\left|X^{(i)} - X^{(j)}\right| = \begin{cases} 2 & si \quad \{AA, aa\} \\ 1 & si \quad \{AA, Aa\} \ o \ \{Aa, aa\} \\ 0 & en \ otro \ caso \end{cases} \tag{1.12}$$

Si los individuos del par son homocigotos diferentes, entonces la distancia genética absoluta entre los dos individuos será 2, si alguno de ellos es heterocigoto, será 1 y en otro caso será 0[11].

Para seguir construyendo el estimador del coeficiente kinship, se considerará el estimador $\frac{\hat{H}_{ij}}{|\mathcal{M}_{ij}|}$ , consistente a $\frac{\sum_m 2p_m(1-p_m)}{|\mathcal{M}|_{ij}}$, siendo $|\mathcal{M}_{ij}|$ el número total de SNPs sin valores perdidos para un par de individuos y $\hat{H}_{ij}$ la suma de frecuencias genotípicas cuando un individuo es heterocigoto para cada SNP. Se representa entonces este estimador en la Ecuación 1.13

$$\frac{\hat{H}_{ij}}{|\mathcal{M}_{ij}|} = \frac{\sum_m 2\hat{p}_m\left(1-\hat{p}_m\right)}{|\mathcal{M}_{ij}|}, \tag{1.13}$$

considerándose un momento respecto al origen de orden 1.[12]. A partir de este estimador, se procede a la estimación del coeficiente kinship para poblaciones homogéneas, representado en la Ecuación 1.14.

$$\hat{\phi}_{ij} = \frac{1}{2} - \frac{\sum_m \left(X_m^{(i)} - X_m^{(j)}\right)^2}{4\hat{H}_{ij}}. \tag{1.14}$$

Bajo las propiedades del estimador de la Ecuación 1.13, se puede afirmar que el estimador del coeficiente kinship es consistente a su valor teórico $\phi_{ij}$ . En el caso de que el número de copias IBD en los alelos de dos individuos sea distinto de 0, se realizaría la estimación del coeficiente kinship según las Ecuaciones 1.15 y 1.16.

$$\hat{\pi}_1 = 2 - 2\hat{\pi}_0 - 4\hat{\phi} \tag{1.15}$$

$$\hat{\pi}_2 = 4\hat{\phi} + \hat{\pi}_0 - 1 \tag{1.16}$$

### 1.4.2. KING-robust

El desarrollo anterior se basa siempre en el supuesto de que la población sea homogénea, es decir, que no presente estructuras de poblaciones diferentes. Sin embargo, en la práctica no suele ser frecuente trabajar con un único grupo ancestral de individuos. Por este motivo, se utiliza un método alternativo para estimar el coeficiente kinship cuando los individuos de la cohorte pertenecen a diferentes grupos

[11] Para más información acerca de las distancias genéticas del estimador kinship, ver Material Suplementario del artículo recogido en la bibliografía de esta memoria (Manichaikul *et al.* 2010).

[12] $\frac{\sum_m 2p_m(1-p_m)}{|\mathcal{M}_{ij}|}$ es equivalente a $\frac{\sum_m 2p_m q_m}{|\mathcal{M}_{ij}|}$.

ancestrales sin población mezclada.

Se asume que $Q$ es una variable aleatoria que representa a la frecuencia alélica en un SNP extraído aleatoriamente del conjunto de SNPs genotipados de un individuo. La distribución de probabilidad de $Q$ debe ser la misma para cada subpoblación o grupo ancestral y su valor puede ser diferente para cada individuo.

Denotada la variable aleatoria $Q$, se determina el valor esperado de las distancias genéticas que permiten calcular el coeficiente kinship para cualquier SNP en la Ecuación 1.17

$$E\left[X^{(i)} - X^{(j)}\right]^2 = 4E[Q(1-Q)](1-2\hat{\phi}_{ij}). \tag{1.17}$$

A continuación, asumiendo las condiciones del Equilibrio de Hardy-Weinberg, se define la heterocigosidad media a partir de la Ecuación 1.18

$$E\left[2Q(1-Q)\right] = E[P(Aa|Q)] = E\left[E\left[I_a|Q\right]\right] = E\left[I_a\right], \tag{1.18}$$

siendo $I_a$ una variable indicadora para cualquier individuo con frecuencia alélica $Q$ si y sólo si el individuo es heterocigoto, es decir, que tenga un genotipo $Aa$ en un SNP seleccionado aleatoriamente. La heterocigosidad media para un individuo puede ser estimada a partir del estimador consistente $\frac{N_{Aa}}{|\mathcal{M}_{ij}|}$. En el caso de que $N_{Aa}$, siendo el número de SNPs con individuos heterocigotos, sea diferente , entonces el estimador empírico de la heterocigosidad media $E(2Q(1-Q))$ se representa en la Ecuación 1.19

$$\frac{\hat{H}_{ij}}{|\mathcal{M}_{ij}|} = \frac{\left(N_{Aa}^{(i)} + N_{Aa}^{(j)}\right)}{2|\mathcal{M}_{ij}|}, \tag{1.19}$$

siendo $N_{Aa}^{(i)}, N_{Aa}^{(j)}$ el número total de SNPs en los que, para cada individuo i y j respectivamente, tienen genotipos heterocigotos. La Ecuación 1.20 presenta al estimador del coeficiente kinship de KING-robust, como un momento respecto al origen de orden 1

$$\hat{\phi}_{ij} = \frac{1}{2}\left[1 - \frac{\sum_m \left(X_m^{(i)} - X_m^{(j)}\right)^2}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}\right] = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)} + N_{Aa}^{(j)}}. \tag{1.20}$$

La estimación del coeficiente kinship genera diferentes valores en función de las probabilidades $Z$ cuando ninguna, una o dos bases son idénticas en los alelos de sus respectivos individuos. Calculando la media geométrica de los valores teóricos extremos del estimador kinship para clasificar según el grado de parentesco, se obtienen sus criterios de inferencia en función de la clasificación tanto para la probabilidad de que dos individuos no tengan dos copias genotípicas IBS como para el propio coeficiente kinship. La tabla A.4 recogida en el apéndice refleja dicha clasificación.

El estimador robusto del coeficiente kinship $\hat{\phi}_{ij}$ es consistente a su valor teórico si y sólo si el par de individuos observado se haya extraído del mismo grupo ancestral. En el caso de que esta condición no se cumpla, el estimador será consistente a un parámetro con signo negativo. Si dicho valor es lo suficientemente negativo, entonces los miembros de ese par provienen de subpoblaciones ancestrales distintas, es decir, pertenecerán a estructura poblacionales diferentes. Esta propiedad se refleja en la Ecuación 1.21

$$Sesgo\ Negativo\ Kinship = -\frac{E\left[Q_1 - Q_2\right]^2}{E[Q_1(1-Q_1)] + E[Q_2(1-Q_2)]}, \tag{1.21}$$

siendo $Q_1$ y $Q_2$ las variables aleatorias $Q$ para cada uno de los dos individuos.

A diferencia de KING-homo, KING-robust permite trabajar en poblaciones con estratificación poblacional, es decir, pueden existir varias subpoblaciones con distinta ancestralidad. Bajo este escenario, los individuos pueden pertenecer a una población mezclada cuando tienen la probabilidad de pertenecer a más de una subpoblación ancestral. Asumiendo el Equilibrio de Hardy-Weinberg, el estimador robusto puede dar lugar a una sobreestimación del coeficiente kinship. Una posible solución sería considerar las ratios de heterocigosidad mínima mín($\frac{N_{Aa}^{(i)}}{|\mathcal{M}_{ij}|}, \frac{N_{Aa}^{(j)}}{|\mathcal{M}_{ij}|}$) para determinar la heterocigosidad media $E(2Q(1-Q)$, dando lugar a seleccionar individuos con una menor ratio. Suponiendo que el individuo $i$ tiene una ratio de heterocigosidad menor que el individuo $j$, se determina el estimador robusto del coeficiente kinship para poblaciones heterogéneas en la Ecuación 1.22

$$\hat{\phi}_{ij} = \frac{1}{2} - \frac{1}{4}\frac{\sum_m \left(X^{(i)} - X^{(j)}\right)^2}{N_{Aa}^{(i)}} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{N_{Aa}^{(i)}} + \frac{1}{2} - \frac{1}{4}\frac{N_{Aa}^{(i)} + N_{Aa}^{(j)}}{N_{Aa}^{(i)}}. \tag{1.22}$$

A partir de la asunción de las condiciones del Equilibrio de Hardy-Weinberg, es posible estimar tanto la varianza como la media de la variable aleatoria $Q$ para cada individuo en un SNP seleccionado aleatoriamente en las Ecuaciones 1.23 y 1.24 respectivamente

$$Var[Q] = E[Q^2] - E[Q]^2 = P(AA) - \frac{1}{4}\left[E[X]\right]^2, \tag{1.23}$$

$$E[Q] = E\left[\frac{1}{2}E[X|Q]\right] = \frac{1}{2}E[X], \tag{1.24}$$

siendo $X$ la variable que representa a las variables $X_m^{(i)}$ y $X_m^{(j)}$ para todos los $|\mathcal{M}|$ SNPs.

### 1.4.3. Matriz GRM

La matriz GRM[13] se presenta como una representación del conjunto de valores o estimaciones del parentesco para cada par de individuos en formato matriz. A continuación, se presenta la idea básica de esta herramienta de inferencia del parentesco a partir del material referenciado (Yang *et al.* 2011).

Sean dos conjuntos de individuos de $i$ y $j$ denotados como $Y$ y $J$ respectivamente y $n \in \mathcal{N}$ cada individuo del conjunto total de la muestra se presenta la matriz GRM de las estimaciones de la proporción de alelos IBD para todos los SNPs denotada como $\hat{\mathbf{\Pi}}$, con dimensiones $Y \times J$. Cada valor de la matriz representa a la proporción de alelos IBD estimada $\hat{\pi}_{ij}$ para cada par de individuos, definiéndose en la Ecuación 1.25

$$\hat{\pi}_{ij} = \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\frac{\left(X_m^{(i)} - 2p_m\right)\left(X_m^{(j)} - 2p_m\right)}{2p_m\left(1-p_m\right)}, \tag{1.25}$$

siendo utilizada para la determinación de la proporción de alelos IBD en PLINK. Si se calcula la mitad de dicha matriz, se obtiene la matriz GRM $\hat{\mathbf{\Phi}}$ para la estimación de coeficientes kinship de KING en la Ecuación 1.26

$$\hat{\phi}_{ij} = \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\frac{\left(X_m^{(i)} - 2p_m\right)\left(X_m^{(j)} - 2p_m\right)}{4p_m\left(1-p_m\right)}. \tag{1.26}$$

El uso de la matriz GRM permite recoger toda la información sobre las estimaciones del parentesco para cada par de individuos.

[13] *Genetic Relationship Matrix* o por sus siglas en inglés GRM.

## 1.5. Inferencia de parentesco: PC-AiR y PC-Relate

En los siguientes dos apartados, se explicarán tanto el algoritmo PC-AiR como PC-Relate. Es importante señalar que PC-AiR se enfoca en la estimación de las estructuras poblacionales en base al Análisis de Componentes Principales[14] siendo necesario para la determinación del método de PC-Relate, destinado este último a la inferencia del parentesco. Cabe señalar que no pertenecen a ningún software especializado, sino que son paquetes estadísticos *GENESIS* y *SNPRelate* de Bioconductor para el lenguaje R (Gogarten *et al.* 2019) (Zheng *et al.* 2012) [15].

### 1.5.1. PC-AiR

El método PC-AiR[16] aplica un algoritmo para la identificación de conjuntos diversos de individuos no emparentados que son representativos para la muestra completa de todos los SNPs. Se basa en la predicción u estimación de la estructura poblacional en muestras de SNPs de individuos emparentados a partir del PCA de los individuos no emparentados. Como característica más relevante, permite trabajar con un subconjunto de individuos sin requerir paneles externos de referencia u otra información genealógica para la inferencia de la ancestralidad. Gran parte de la notación aquí presente, así como las ecuaciones, están basadas en la teoría de las referencias de esta memoria (Conomos *et al.* 2015).

#### Metodología del algoritmo

Se denota $\mathcal{N}$ como la muestra de individuos genotipados cuya partición se divide en dos subconjuntos:

- $\mathcal{U}$: submuestra de pares de individuos no emparentados que son representativos para la ancestralidad de todo $\mathcal{N}$.
- $\mathcal{R}$: submuestra de pares emparentados que tienen al menos a un pariente dentro del subconjunto $\mathcal{U}$.
- $\mathcal{N} = \mathcal{U} \cup \mathcal{R}$.
- $\mathcal{U} \cap \mathcal{R} = \emptyset$.

Se asume que $\mathcal{N}$ proviene de una población heterogénea con subpoblaciones ancestrales $K$. $\mathcal{M}$ es el conjunto de SNPs de autosomas, y $\mathbf{p}_m = (p_m^1, ..., p_m^K)$ es el vector de frecuencias alélicas elevadas al número de subpoblaciones, siendo $m \in \mathcal{M}$ y $k \in K$.

Se asume que cada $p_m^k$ es una variable aleatoria e independiente respecto a los $m$ SNPs y dependiente con las $k$ subpoblaciones. El valor esperado de $p_m^k$ es $E[p_m] = p_m\mathbf{1}$ y su covarianza $Cov[p_m] = p_m(\mathbf{1} - p_m)\mathbf{\Theta}_K$ para cualquier $m$, siendo $\mathbf{\Theta}_K$ una matriz $K \times K$ y $\mathbf{1}$ un vector columna unitario de longitud $K$.

En modelos con presencia de estratificación poblacional, $p_m$ se interpreta como la media de las frecuencias alélicas de cada subpoblación. Se permite la presencia de ancestralidad o población mezclada en los $K$ subpoblaciones, denotando las proporciones de ancestralidad como $\mathbf{a}_i = (a_i^1 ..., a_i^K)^\top$ para cada individuo $i \in \mathcal{N}$ y subpoblación $k$, cumpliendo que $a_i^k > 0$ y $\sum_{k=1}^K a_i^k = 1$.

En el caso de que se desconozca la información familiar de los individuos emparentados, PC-AiR utiliza el estimador KING-robust para estimar el parentesco bajo estratificación poblacional.

[14] *Principal Component Analysis*, o por sus siglas en inglés PCA.
[15] Para más información, ver el contenido de los paquetes en el enlace de *GENESIS* y en el sitio web de *SNPRelate* de Bioconductor.
[16] *Principal Component Analysis in Related samples*, o Análisis de Componentes Principales en muestras Emparentadas traducido del inglés.

**Inferencia de parentesco y ancestralidad con Estructura Poblacional**

PC-AiR utiliza los coeficientes kinship para medir el parentesco genético entre cada par de individuos. En el caso de no conocer su información genealógica o tener poblaciones mezcladas, es conveniente utilizar el estimador empírico del coeficiente kinship KING-robust, presentando sesgo negativo si hay individuos no emparentados con diferente estructura poblacional. Al igual que en KING, no se recomienda utilizar el estimador KING-homo para aplicar PC-AiR cuando la población es heterogénea. Este motivo se debe a los posibles fallos de clasificación cuando varios individuos tienen ancestralidad similar o por inflación de la clasificación de parientes cercanos. PC-AiR utiliza previamente medidas de ancestralidad para mejorar la representatividad de los pares no emparentados que definen la diversidad ancestral de la población.

Se considera un par de individuos $(i, j) \in \mathcal{N}$ del subconjunto de SNPs autosomáticos sin valores perdidos de dicho par $\mathcal{M}_{ij} \in \mathcal{M}$. Se denotan las variables aleatorias $X_m^{(i)}, X_m^{(j)}$, como el número de copias idénticas de genotipos de los alelos que cada individuo $i$ y $j$ tienen respectivamente en un SNP $m \in \mathcal{M}_{ij}$. La medida utilizada para estudiar la divergencia ancestral de un par de individuos no emparentados es el estimador del coeficiente kinship KING-robust, mostrándose en la Ecuación 1.20 y las propiedades del estimador cuando los pares de individuos son no emparentados en la Ecuación 1.27

$$\hat{\phi}_{ij} \to 0, |\mathcal{M}_{ij}| \to \infty. \tag{1.27}$$

Cuando la población está mezclada y los individuos $i$ y $j$ son ancestralmente diferentes, se produce un sesgo negativo que permite medir la diversidad ancestral en $\hat{\phi}_{ij}$. Dicho sesgo se define en la Ecuación 1.28

$$\hat{\phi}_{ij} \to \frac{-\frac{1}{2}(\theta_k + \theta_{k'})}{1 - \frac{1}{2}(\theta_k + \theta_{k'})}. \tag{1.28}$$

siendo $\theta_k$ una medida estandarizada de Wright de la variabilidad $F_{ST}$ para un grupo ancestral $k$, asumiendo las condiciones del modelo de Balding-Nichols (Wright 1949). El sesgo negativo será mayor si se cumplen uno de estos tres casos:

- Si la medida $\theta_k$ tiene un valor elevado para una subpoblación $k$.
- Si $(i, j)$ tienen proporciones ancestrales muy diferentes.
- Si una proporción ancestral de algún individuo es cercana a 1 en alguna de las $K$ subpoblaciones.

**Matriz GSM del PC-AiR en la inferencia de individuos emparentados**

Una vez estimado el parentesco a partir del estimador robusto del coeficiente kinship, se procede con el PCA de los individuos no emparentados. Previamente, se deben estandarizar los valores genotípicos para un individuo $i$ en un SNP $m$ seleccionado según la Ecuación 1.29

$$z_{im} = \frac{X_m^{(i)} - 2\hat{p}_m}{\sqrt{2\hat{p}_m(1 - \hat{p}_m)}}. \tag{1.29}$$

Utilizando únicamente los pares no emparentados contenidos en $\mathcal{U}$, se cumple que $\hat{p}_m = \hat{p}_m^u$ cuando se define según la Ecuación 1.30

$$\hat{p}_m^u = \frac{1}{2|\mathcal{U}_m|} \sum_{i \in \mathcal{U}_m} X_m^{(i)}, \tag{1.30}$$

siendo $|\mathcal{U}_m|$ el número de individuos pertenecientes al subconjunto $\mathcal{U}_m$. Se demuestra que la frecuencia alélica para individuos no emparentados $p_m^u$ es más eficiente que utilizar la frecuencia alélica $p_m$ para

toda la muestra en PC-AiR, según (Conomos *et al.* 2015).

Se consideran $n, n_u$ y $n_r$ como el número de individuos de los conjuntos y subconjuntos $\mathcal{N}$,$\mathcal{U}$ y $\mathcal{R}$ respectivamente y $|\mathcal{M}^*|$ como el número de SNPs filtrados después de un proceso de control de calidad. A partir de la Ecuación 1.29, se construye una matriz $\mathbf{Z}$ de dimensión $n \times |\mathcal{M}^*|$, con valores $z_{ij}$, donde las primeras $n_u$ filas pertenecen al subconjunto $\mathcal{U}$ y el resto de $n_r$ corresponden al subconjunto $\mathcal{R}$.

Seleccionando únicamente la submuestra de pares no emparentados y calculando la matriz de genotipos estandarizados para los pares no emparentados $\mathbf{Z}_u$, se determina la Matriz de Similitudes Genéticas[17] en la Ecuación 1.31

$$\hat{\mathbf{\Psi}}_u = \frac{1}{|\mathcal{M}^*|}\mathbf{Z}_u\mathbf{Z}_u^\top, \tag{1.31}$$

siendo los primeros pares $(i, j)$ en $\hat{\mathbf{\Psi}}_\mathbf{u}$ una medida de similitudes genéticas medias entre individuos $(i, j) \in \mathcal{U}$.

Una vez definida la matriz GSM, se calculan las $D$ primeras componentes principales que sean representativas ancestralmente para el conjunto $\mathcal{N}$ utilizando el subconjunto de individuos no emparentados $\mathcal{U}$. Aplicando el PCA utilizando un procedimiento similar al aplicado en EIGENSOFT o EIGENSTRAT sobre la matriz GSM simétrica (Patterson *et al.* 2006) y (Price *et al.* 2006), se obtiene que $\hat{\mathbf{\Psi}}_u = \mathbf{V}_u\mathbf{L}_u\mathbf{V}_u^{-1}$ , resultando:

- $\mathbf{V}_u = [V_1^u, .., V_{n_u}^u]$ siendo una matriz $n_u \times n_u$ cuyo vector columna $d$-ésimo $\mathbf{V}_d^u$, es la $d$-ésima componente principal.
- $\mathbf{L}_u = diag(\lambda_1^u, ..., \lambda_{n_u}^u)$ siendo una matriz diagonal $n_u \times n_u$ de autovalores $\lambda_{n_u}^u$.
- $\mathbf{W}_u = \mathbf{Z}_u^\top\mathbf{V}_u$ siendo una matriz de pesos de SNPs $|\mathcal{M}^*| \times n_u$ que muestra la influencia ponderada de los SNPs en $n_u$.

Como resultado, se podrá determinar la matriz de componentes principales $\mathbf{V}_u$ a partir de su descomposición en la Ecuación 1.32

$$\mathbf{V}_u = \hat{\mathbf{\Psi}}_u\mathbf{V}_u\mathbf{L}_u^{-1} = \left(\frac{1}{|\mathcal{M}^*|}\mathbf{Z}_u\mathbf{Z}_u^\top\right)\mathbf{V}_u\mathbf{L}_u^{-1} = \frac{1}{|\mathcal{M}^*|}\mathbf{Z}_u\mathbf{W}_u\mathbf{L}_u^{-1}. \tag{1.32}$$

A partir de las componentes principales del subconjunto de pares no emparentados $\mathcal{U}$, se calculan las predicciones sobre los valores de las componentes principales del subconjunto de pares emparentados $\mathcal{R}$, reemplazando la matriz de genotipos estandarizados $\mathbf{Z_u}$ por $\mathbf{Z_r}$ en la Ecuación 1.32. Se denota $\mathbf{Q}_r$ como una matriz $n_r \times n_u$ de componentes principales predictivas para el subconjunto $\mathcal{R}$, reflejada en la Ecuación 1.33

$$\mathbf{Q}_r = \frac{1}{|\mathcal{M}^*|\mathbf{Z}_r\mathbf{W}_u\mathbf{L}_u^{-1}}, \tag{1.33}$$

siendo cada $d$-ésimo vector columna cada valor predictivo perteneciente a cada $d$-ésima componente principal en $\mathcal{R}$.

Si se consideran de forma conjunta los individuos de $\mathcal{R}$ y $\mathcal{U}$ simultáneamente, se define $\Gamma$ una matriz de $n \times n_u$ de componentes principales combinadas tanto en $\mathcal{R}$ como en $\mathcal{U}$ en la Ecuación 1.34

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{V}_u \\ \\ \mathbf{Q}_r \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1^u & \mathbf{V}_2^u & ... & \mathbf{V}_{n_u}^u \\ \\ \mathbf{Q}_1^r & \mathbf{Q}_2^r & ... & \mathbf{Q}_{n_u}^r \end{pmatrix}, \tag{1.34}$$

[17] *Genetic Similitary Matrix*, o por sus siglas en inglés GSM.

siendo cada vector columna de $\mathbf{\Gamma}$ la componente principal del conjunto $\mathcal{N}$ obtenida por el método PC-AiR. Las primeras componentes principales de $\mathbf{\Gamma}$ se considerarán representativas de la ancestralidad del conjunto completo de individuos $\mathcal{N}$.

#### FastPCA

El método FastPCA es un procedimiento alternativo de cálculo de componentes principales disponible en PC-AiR. El desarrollo de este algoritmo se basa en la teoría de matrices aleatorias[18], demostrando que es una técnica más eficiente que el cálculo tradicional de componentes principales en base a la literatura existente (Galinsky *et al.* 2016). De forma resumida, el procedimiento consiste en estimar los mayores autovalores y autovectores de una matriz $\mathbf{M}$, multiplicando un vector aleatorio y una matriz cuadrada. Esta matriz proyecta el vector anterior sobre los autovectores de $\mathbf{M}$ y lo reescala según los autovalores de la matriz $\mathbf{M}$. Este algoritmo será el utilizado para el cálculo de componentes principales en PC-AiR a través de la función *pcair* en la librería *GENESIS* del lenguaje de programación R.

### 1.5.2. PC-Relate

El método de PC-Relate se presenta como un método de estimación del parentesco que no requiere de la especificación previa de los grupos ancestrales ni de paneles externos de referencia de ancestralidad de individuos. Por este motivo, se considera un método de estimación libre para la inferencia de parentescos genéticos. En KING-robust, por ejemplo, el coeficiente kinship estimado podía tener sesgo negativo debido a que dos individuos emparentados podían pertenecer a dos estructuras poblacionales distintas, provocando una estimación errónea para poblaciones mezcladas. Por este motivo, se aplica PC-Relate como un método basado en el PCA previo de la ancestralidad de los individuos para poblaciones mezcladas y con estructuras poblacionales desconocidas (Conomos *et al.* 2016).

#### Parámetros poblacionales

Referente a la notación de PC-Relate, los parámetros de partida serán similares a los ya utilizados en PC-AiR. Se denotan específicamente las correlaciones intragrupo o intergrupal de las subpoblaciones, mostradas en la diagonal y en los elementos fuera de la diagonal de la matriz $\mathbf{\Theta}_K$ respectivamente.

El $k$-ésimo elemento de la diagonal de la matriz $\mathbf{\Theta}_K$ se denota como $\theta_k$, siendo la correlación de un par aleatorio de alelos de una subpoblación $k$ respecto al total de la población. Respecto a los elementos fuera de la diagonal $[k, k']$ denotados como $\theta_{kk'}$, se definen como la correlación de dos alelos escogidos aleatoriamente de dos grupos ancestrales distintos $k$ y $k'$ respectivamente.

Se denota como $\psi_{ij}$ el coeficiente kinship cuando la población de referencia es la población ancestral común de las que pertenecen las subpoblaciones $K$. $\phi_{ij}$ sería el coeficiente kinship cuando la población de referencia está compuesta de las $K$ subpoblaciones.

#### Matriz GRM de PC-Relate

La Matriz empírica Genética de Parentesco, tiene como función inferir sobre muestras sin parientes en presencia de estructura poblacional. La matriz GRM se construye bajo las condiciones del Equilibrio de Hardy-Weinberg, cuyos valores son las estimaciones del coeficiente kinship $\psi_{ij}$ en base a la Ecuación 1.35

[18] Para mayor detalle, consultar los artículos (Rokhlin *et al.* 2010), (Halko, Shkolnisky *et al.* 2011) y (Halko, Martinsson *et al.* 2011).

$$\hat{\psi}_{ij} = \frac{1}{|\mathcal{M}_{ij}|} \sum_{m \in \mathcal{M}_{ij}} \frac{\left(X_m^{(i)} - 2\hat{p}_m\right)\left(X_m^{(j)} - 2\hat{p}_m\right)}{4\hat{p}_m(1-\hat{p}_m)}. \tag{1.35}$$

A partir de la matriz GRM detallada en el artículo (Yang *et al.* 2011) y (Conomos *et al.* 2015), se muestra en la Ecuación 1.35, asumiendo que $p_m$ es conocida y que el número de SNPs independientes tiende a infinito se demuestra en la Ecuación 1.36 que

$$\hat{\psi}_{ij} \to \hat{\phi}_{ij} + \theta_{ij} - b_\psi(i,j), \tag{1.36}$$

siendo $\theta_{ij} = \mathbf{a}_i^\top \mathbf{\Theta}_K \mathbf{a}_j$ el coeficiente de coancestralidad afectado por las $K$ estructuras poblacionales para cada par de individuos y $b_\psi(i,j)$ la función de coancestralidad de los ancestros comunes de los individuos $i$ y $j$.

A partir de la Ecuación 1.36, utilizar el PCA tradicional con presencia de parientes en la población de estudio, puede provocar que la medida de coancestralidad $\theta_{ij}$ sea difícil de calcular con el coeficiente kinship $\hat{\phi}_{ij}$. Este escenario conlleva a una posible confusión de las componentes principales que permiten medir la ancestralidad. Por este motivo, es preferible utilizar PC-AiR para mejorar la representatividad de las componentes principales en la ancestralidad.

### Metodología de PC-Relate

PC-Relate presenta varias ventajas frente a otros algoritmos según queda demostrado en el artículo relacionado (Conomos *et al.* 2016), ya que no requiere de:

- La determinación previa de las poblaciones ancestrales
- Estimaciones previas de la ancestralidad de cada individuo
- Cálculo previo de frecuencias alélicas
- Panales de referencia externos

Considerando las $D$ primeras componentes principales en la matriz $\mathbf{V}$ de dimensiones $|\mathcal{N}| \times D$, se calcula $\mathbf{X}_m$ como el vector de valores genotípicos para todos los individuos en cada uno de los $m$ SNPs.

A continuación, se considera el siguiente modelo de regresión lineal en la Ecuación 1.37

$$E[\mathbf{X}_m|\mathbf{V}] = \mathbf{1}\beta_0 + \mathbf{V}\beta, \tag{1.37}$$

siendo $\beta = (\beta_1, ..., \beta_D)^\top$ un vector columna de regresores de tamaño $D$ para cada componente principal y $\mathbf{1}$ un vector de unos de tamaño $|\mathcal{N}|$. Por consiguiente, $E[\mathbf{X}_m|\mathbf{V}]$ es equivalente a la esperanza de $\mathbf{X}_m$ condicionada a los valores reales de ancestralidad de cada individuo de estudio. Ajustando la regresión de la Ecuación 1.37, se pueden utilizar los valores ajustados para estimar las frecuencias alélicas de las componentes principales en cada individuo. Como resultado, se aplica el estimador de $\mu_{im}$ para cada SNP $m \in \mathcal{M}$ en la Ecuación 1.38

$$\hat{\mu}_{im} = \frac{1}{2}E\left[X_m^{(i)}|V_i^1, ..., V_i^D\right] = \frac{1}{2}\left(\hat{\beta}_0 + \sum_{d=1}^{D} \hat{\beta}_d V_i^d\right), \tag{1.38}$$

siendo cada $V_i^d$ un valor de la $d$-ésima componente principal del vector columna $\mathbf{V}_d$ de un individuo $i$. Si se considera que cada componente principal tiene media cero, entonces el valor del intercepto de la regresión es $\frac{1}{2}\hat{\beta}_0 = \hat{p}_m$ para cada $m$ SNP y $\hat{\beta}_d$ sería la medida de desviación de la frecuencia alélica debido a la ancestralidad representada en$\mathbf{V_d}$.

La estimación del coeficiente kinship aplicando el método PC-Relate $\kappa_{ij}$ se define en la Ecuación 1.39

$$\hat{\kappa}_{ij} = \frac{\sum_{m\in\mathcal{M}_{ij}} \left(X_m^{(i)} - 2\hat{\mu}_{im}\right)\left(X_m^{(j)} - 2\hat{\mu}_{jm}\right)}{4\sum_{m\in\mathcal{M}_{ij}} \left[\hat{\mu}_{im}\left(1-\hat{\mu}_{im}\right)\hat{\mu}_{jm}\left(1-\hat{\mu}_{jm}\right)\right]^{\frac{1}{2}}}, \tag{1.39}$$

denotándose $\hat{\mu}_{im}, \hat{\mu}_{jm}$ como las frecuencias alélicas estimadas para un individuo $i$ y $j$ respectivamente en cada SNP $m$. El estimador del coeficiente kinship de PC-Relate tiene en cuenta el efecto de la ancestralidad a través de las frecuencias alélicas estimadas $\hat{\mu}_{im}, \hat{\mu}_{jm}$ para la inferencia del parentesco. Los valores genotípicos $X_m^{(i)}$ y $X_m^{(j)}$ están centrados y escalados con respecto las frecuencias alélicas $\hat{\mu}_{im}$ y $\hat{\mu}_{jm}$ respectivamente. El estimador $\hat{\kappa}_{ij}$ de PC-Relate se vería entonces como una ratio ponderada de las frecuencias alélicas medias en los *loci*, con resultados más estables si el MAF es bajo. Este estimador se construiría a partir de los residuos del modelo de regresión lineal formulado en la Ecuación 1.37 incluyendo a las componentes principales como predictores, siendo estos residuos ortogonales con respecto a las componentes principales.

A pesar de la existencia de un sesgo asintótico del estimador $\hat{\kappa}_{ij}$, si los pares de individuos están emparentados y presentan distintas estructuras poblacionales, dicho sesgo no es influyente en la estimación de su parentesco. Esta afirmación se demuestra en la Ecuación 1.40

$$\hat{\kappa}_{ij} \to \frac{E[X_{im}X_{jm}] - 4E[\mu_{im}\mu_{jm}]}{4E\left[\left[\mu_{im}(1-\mu_{im})\mu_{jm}(1-\mu_{jm})\right]^{\frac{1}{2}}\right]} = b_\kappa(i,j), \tag{1.40}$$

siendo $b_\kappa(i,j)$ el sesgo asintótico del estimador PC-Relate del coeficiente kinship para cada par de individuos. Se determina entonces el sesgo según la Ecuación 1.41

$$b_\kappa(i,j) = \sum_{m\in\mathcal{M}_{ij}} \kappa_{ij|m}\left(\frac{\theta_{mm} - d_\kappa(i,j)}{1 - d_\kappa(i,j)}\right), \tag{1.41}$$

siendo $d_\kappa(i,j) = \theta_k = \theta_{mm}$ para todo $m \in \mathcal{M}_{ij}$ tal que si $b_\phi(i,j) = 0$ entonces $\hat{\kappa}_{ij} \to \kappa_{ij}$.

**Estimación de las probabilidades de alelos IBD con presencia de estructura poblacional**

Al igual que en el resto de los métodos de estimación de parentesco, se determinan las probabilidades de que los alelos de dos individuos sean IBD. Primeramente, bajo las condiciones de una población homogénea y las asunciones del Equilibrio de Hardy-Weinberg , se denota una variable aleatoria $X_{im}^D$, como alternativa a la variable $X_m^{(i)}$, que representa el número de copias idénticas respecto al alelo de referencia del individuo $i$ en un SNP $m$ con $D$ componentes principales. Es ortogonal a la variable de valores genotípicos $X_m^{(i)}$ y análoga para la variable $X_{jm}^D$ siendo:

- $X_{im}^D = (1 - \hat{p}_m)$ si el individuo es homocigoto respecto al alelo de referencia. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor uno si los dos alelos tienen genotipos idénticos $AA$ en un SNP $m$ dado.
- $X_{im}^D = 0$ si el individuo es heterocigoto respecto al alelo de referencia. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor uno si los dos alelos tienen genotipos diferentes $Aa$ en un SNP $m$ dado.
- $X_{im}^D = \hat{p}_m$ si el individuo es homocigoto respecto al otro alelo. Si el genotipo del alelo de referencia se denota como $A$, entonces tomará valor cero si los dos alelos tienen genotipos idénticos respecto al otro alelo $aa$ en un SNP $m$ dado.

Si $p_m$ es conocido, entonces $X_{im}^D$ posee las siguientes características:

- $E[X_{im}^D] = p_m(1 - p_m)$

- $Var[X_{im}^D] = [p_m(1-p_m)]^2$

Bajo estas propiedades, si no existe estratificación poblacional, se define $\delta_{ij}$ como la correlación no condicionada entre $X_{im}^D$ y $X_{jm}^D$, reflejada en la Ecuación 1.42

$$\hat{\delta}_{ij} = \frac{1}{|\mathcal{M}_{ij}|} \sum_{m \in \mathcal{M}_{ij}} \frac{\left[X_{im}^D - \hat{p}_m(1-\hat{p}_m)\right]\left[X_{jm}^D - \hat{p}_m(1-\hat{p}_m)\right]}{[\hat{p}_m(1-\hat{p}_m)]^2}. \tag{1.42}$$

El estimador expresado en la Ecuación 1.42 se podría expresar como la probabilidad de que los alelos sean IBD de dos individuos en un SNP , $\hat{\delta}_{ij} \to \hat{\varpi}_{ij}^{(2)}$ si $M_{ij} \to \infty$ , siendo $\hat{\varpi}_{ij}^{(2)}$ un estimador consistente de la probabilidad $\varpi_{2ij}$. Entonces, si se considera la estratificación poblacional, el estimador $\hat{\delta}_{ij}$ no es consistente a $\varpi_{2ij}$. Por este motivo, se reemplaza $\hat{p}_m$ por $\hat{\mu}_{im}$ en la Ecuación 1.43

$$\hat{\varpi}_{2ij} = \frac{\sum_{m \in \mathcal{M}_{ij}} [X_{im}^D - \hat{\mu}_{im}(1-\hat{\mu}_{im})(1+\hat{f}_i)][X_{jm}^D - \hat{\mu}_{jm}(1-\hat{\mu}_{jm})(1+\hat{f}_j)}{\sum_{m \in \mathcal{M}_{ij}} \hat{\mu}_{is}(1-\hat{\mu}_{is})\hat{\mu}_{js}(1-\hat{\mu}_{js})}, \tag{1.43}$$

denotando $\hat{f}_i$ y $\hat{f}_j$ el recuento de frecuencias alélicas bajo la asunción del Equilibrio de Hardy-Weinberg en presencia de estratificación poblacional de los individuos $i$ y $j$ respectivamente. El estimador es consistente para pares no emparentados y asintóticamente sesgado para los pares de individuos emparentados con población mezclada por ancestralidad común.

PC-Relate también permite calcular las probabilidades cuando un genotipo tiene una copia idéntica en los alelos de los dos individuos y cuando no coinciden en ningún individuo[19],identificándose estimadores distintos en función del grado de parentesco que se presente. El punto de corte para utilizar uno de los dos estimadores será el umbral del coeficiente kinship a partir del cual se consideran parentescos de primer grado. Por consecuencia, se toma el resultado de la Ecuación 1.44

$$\hat{\varpi}_{0ij} = \begin{cases} \frac{\sum_{m \in \mathcal{M}_{ij}} \mathbf{1}_{\left[\left|X_m^{(i)} - X_m^{(j)}\right| = 2\right]}}{\sum_{m \in \mathcal{M}_{ij}} [\hat{\mu}_{im}^2](1-\hat{\mu}_{jm})^2 + (1-\hat{\mu}_{im})^2\hat{\mu}_{jm}^2} & si \quad \hat{\kappa}_{ij} > 2^{-5/2} \approx 0{,}177 \\ 1 - 4\hat{\kappa}_{ij} + \hat{\varpi}_{2ij} & si \quad \hat{\kappa}_{ij} \leq 2^{-5/2} \approx 0{,}177 \end{cases} \tag{1.44}$$

siendo $\mathbf{1}_{\left[\left|X_m^{(i)} - X_m^{(j)}\right| = 2\right]}$ una variable indicadora que toma valor 1 cuando las distancias genéticas en valor absoluto en $\left|X_m^{(i)} - X_m^{(j)}\right|$ sean iguales a dos. Finalmente, a partir de la Ecuación 1.44, se puede obtener la estimación de la probabilidad de que los dos individuos compartan una copia idéntica de un genotipo en los alelos de $i$ y $j$ a través de la Ecuación 1.45.

$$\hat{\varpi}_{1ij} = 1 - \hat{\varpi}_{0ij} - \hat{\varpi}_{2ij} \tag{1.45}$$

## 1.6. PCA y Modelos de Regresión por Mínimos Cuadrados Parciales

### 1.6.1. PCA y matrices GSM

El Análisis de Componentes Principales o PCA es una técnica estadística multivariante no supervisada que tiene como objetivo reducir la dimensión de los datos, simplificando su representación y manteniendo toda la información de las variables originales. Todos los métodos de estimación de esta

[19]Serían similares a las probabilidades $P(Z=1)$ y $P(Z=0)$ respectivamente en el formato mencionado en PLINK.

memoria utilizan aplicaciones del PCA tales como EIGENSTRAT O EIGENSOFT (Patterson *et al.* 2006) y (Price *et al.* 2006)[20]. A continuación, se detallará el funcionamiento de estos métodos para calcular las componentes principales con datos genotípicos.

El primer paso de la aplicación del PCA , será utilizar una matriz de valores genotípicos estandarizados, denotada como $\mathbf{Z}$ de dimensiones $|\mathcal{N}| \times |\mathcal{M}^*|$, cuyos valores $z_{im}$ fueron presentados en la Ecuación 1.29. Se consideran adicionalmente las frecuencias alélicas estimadas $\hat{p}_m$ para todos los individuos $n \in \mathcal{N}$.

Todos los métodos de estimación permiten determinar el PCA a partir de la matriz de genotipos estandarizados $\mathbf{Z}$, utilizando el software EIGENSOFT o EIGENSTRAT . Manteniendo un desarrollo similar a la teoría de PC-AiR, se considera una matriz GSM $\mathbf{\Psi}$ de dimensiones $n \times n$, definida como la covarianza muestral de los vectores columna de la matriz $\mathbf{Z}$ en la Ecuación 1.46

$$\hat{\mathbf{\Psi}} = \frac{1}{|\mathcal{M}^*|}\mathbf{Z}\mathbf{Z}^\top. \tag{1.46}$$

Dicha matriz está determinada a partir de la Descomposición Singular de Valores o SVD de la matriz $\mathbf{Z}$ y ejecutada computacionalmente con el paquete LAPACK[21], cuyo número total de individuos de la muestra debe ser menor que el número total de SNPs. Como resultado, se obtiene la descomposición de la matriz GSM donde $\hat{\mathbf{\Psi}} = \mathbf{V}\mathbf{L}\mathbf{V}^{-1}$, obteniéndose qué:

- $\mathbf{V} = [V_1, ..., V_n]$ siendo una matriz $n \times n$ cuyo vector columna $d$-ésimo $\mathbf{V}_d$ es la $d$-ésima componente principal.
- $\mathbf{L} = diag(\lambda_1, ..., \lambda_n)$ siendo una matriz diagonal $n \times n$ de autovalores $\lambda_n$
- $\mathbf{W} = \mathbf{Z}^\top\mathbf{V}$ siendo una matriz de pesos de dimensiones $|\mathcal{M}^*| \times n$ que muestra la influencia ponderada de los SNPs en los $n$ individuos.

El resultado final será la matriz de componentes principales $\mathbf{V}$, pudiéndose expresar en la Ecuación 1.47

$$\mathbf{V} = \hat{\mathbf{\Psi}}\mathbf{V}\mathbf{L}^{-1} = \left(\frac{1}{|\mathcal{M}^*|}\mathbf{Z}\mathbf{Z}^\top\right)\mathbf{V}\mathbf{L}^{-1} = \frac{1}{|\mathcal{M}^*|}\mathbf{Z}\mathbf{W}\mathbf{L}^{-1}, \tag{1.47}$$

donde cada $V_i$ de la matriz $\mathbf{V}$ es un autovector correspondiente a cada individuo $i$. Este procedimiento es similar al aplicado en el PCA calculado en PC-AiR, siendo aplicable para los métodos de estimación de PLINK y KING. A partir de los comandos –pca de ambos softwares especificando el número de componentes principales, permiten la aplicación del PCA a partir de la matriz GSM[22].

### 1.6.2. Modelos de regresión PLS y PLSGLR

#### Modelos PLS

El modelo de regresión PLS[23] es un método estadístico no lineal de aprendizaje supervisado que permite simplificar el análisis de datos de alta dimensionalidad. La idea general de esta técnica proviene de la consideración de las componentes principales del PCA como variables predictoras o explicativas de un modelo de regresión lineal. La propiedad de la ortogonalidad de estas componentes permite la

[20] Cabe destacar que PC-AiR utiliza un caso particular de PCA, donde considera únicamente al subconjunto de pares no emparentados para luego predecir las componentes principales del subconjunto de pares emparentados.

[21] SVD es *Singular Value Decomposition*, o Valor Singular de Descomposición traducido del inglés. Para más información acerca del método y la aplicación de LAPACK, consultar (Golub y Van Loan 2013).

[22] Para más información sobre el PCA de cada software, visitar el sitio web de PLINK y el sitio web de KING.

[23] *Partial Least Squares* o Mínimos Cuadrados Parciales traducido al español.

independencia de las variables explicativas, evitando el problema de la colinealidad. El objetivo de esta sección será determinar los modelos PLS a partir de la submuestra de pares de individuos no parientes $u \in \mathcal{U}$. Por este motivo, se adaptará la notación de la teoría para incluir únicamente a este subconjunto.

La teoría de los modelos PLS y PLSGLR se estudian con mayor detalle en la literatura existente (Vinzi *et al.* 2010), (Bertrand *et al.* 2014), (Wold *et al.* 1983) y (Meyer *et al.* 2010).

Sea $\mathbf{Z}_u$ la matriz de variables explicativas para el subconjunto de individuos $\mathcal{U}$, definida como la matriz de genotipos estandarizados para todos los individuos $i$ en la Ecuación 1.29. Se considera a cada variable $z_m$ como la variable predictora original, definida en un vector columna de la matriz $\mathbf{Z}_u$. Esta variable representa a los valores genotípicos $X_m^{(i)}$ estandarizados para cada SNP y cada individuo $i$. Se considera entonces a la matriz $\mathbf{V}_u$ de componentes principales obtenida de la matriz GSM $\hat{\mathbf{\Psi}}_u$, siendo cada vector columna $\mathbf{V}_d^u$ la $d$-ésima componente principal ortogonal maximizadora de la covarianza de la variable respuesta denotada como $\mathbf{y}$ . Se presenta entonces el modelo de regresión PLS en la Ecuación 1.48

$$\mathbf{y} = \mathbf{V}_u^\top \mathbf{c} + \epsilon, \tag{1.48}$$

donde $\mathbf{c}$ representa el vector compuesto de coeficientes de regresión de las componentes $\mathbf{V}_d^u$, $\epsilon$ es el vector de residuos del modelo y $\top$ la traspuesta.

Si se considera entonces que $\mathbf{V}_u = \mathbf{Z}_u\mathbf{R}^*$, siendo $\mathbf{R}^*$ una matriz de coeficientes de los predictores de cada componente $\mathbf{V}_d^u$, bajo la condición de que $1 \leq d \leq D$, el modelo de la Ecuación 1.48 se podrá reescribir según la Ecuación 1.49 como

$$\mathbf{y} = \mathbf{Z}_u{\mathbf{R}^*}^\top c + \epsilon. \tag{1.49}$$

A partir de la descripción anterior del modelo PLS, se puede desglosar el modelo para cada individuo $i$ en la Ecuación 1.50 como

$$y_i = \sum_{d=1}^{D} (c_d r_{1d}^* z_{im} + ... + c_d r_{md}^* z_{im}) + \epsilon_i, \tag{1.50}$$

siendo $D$ el número de componentes principales y $m$ cada vector columna de SNPs de la matriz $\mathbf{Z}_u$, con $D \leq rango(\hat{\mathbf{\Psi}}_u)$. Denotándose $s$ como un SNP escogido arbitrariamente, los coeficientes $c_d r_{sd}^*$ son las relaciones entre la variable respuesta y las variables predictoras originales $z_{im}$ de la matriz $\mathbf{Z}_u$ a través de las componentes principales $\mathbf{V}_d^u$, siendo $1 \leq s \leq m$ y $1 \leq d \leq D$.

### PLSGLR: Conceptos y utilidades del modelo

Los modelos PLS pueden generalizarse al modelo PLSGLR[24], con la ventaja de permitir la presencia de valores faltantes en los datos y trabajar con datos que provengan de distribuciones no continuas. Considerando $\mathbf{V}_d^u = r_{1d}^* z_1 + ... + r_{md}^* z_m$, se define el modelo en la Ecuación 1.51

$$g(\theta)_i = \sum_{d=1}^{D} \left( c_d \sum_{s=1}^{m} r_{sd}^* z_{is} \right), \tag{1.51}$$

siendo $\theta$:

- La esperanza condicional de la variable respuesta $\mathbf{y}$ si su distribución es continua.
- El vector de probabilidades si la distribución de la respuesta $\mathbf{y}$ es discreta con dominio finito.

Las componentes $\mathbf{V}_u^d$ se consideran ortogonales y la función link $g$ se selecciona en función de la distribución más adecuada para el ajuste del modelo.

[24] *Partial Least Squares Generalized Linear Regression Model* o por sus siglas en inglés PLSGLR.

**Métodos analíticos de modelos de clasificación**

La evaluación de la precisión de modelos predictivos resulta interesante cuando se requiere seleccionar algún modelo en términos de capacidad predictiva. El ajuste de dichos modelos, así como su selección, se realiza a partir de una muestra de entrenamiento, cuyos valores ajustados se utilizan posteriormente para predecir sobre los datos de una muestra de test. Dichas muestras representan un porcentaje de la muestra total considerada y son elegidas según los criterios de la persona investigadora. En función del objetivo del análisis estadístico existen dos tipos de modelos: modelos de regresión y modelos de clasificación. Estos últimos presentan una variable respuesta categórica, que puede tomar desde valores binarios hasta valores en diferentes categorías. En referencia a los modelos PLSGLR, se considerarían modelos de clasificación, ya que su respuesta es binaria. Después de utilizar la muestra de entrenamiento para ajustar y seleccionar el modelo adecuado, se procede con su evaluación, cuyos análisis se centrarían en el estudio de la precisión de las predicciones sobre la muestra de test de observaciones reales. Uno de estos análisis, utilizados en el caso práctico de la memoria, es la determinación de las curvas ROC[25], frecuentemente usados en aprendizaje estadístico (Gonçalves *et.al* 2014).

Se denotan dos variables aleatorias $X$ y $W$ independientes, representando la medición del test de diagnóstico de una población sana o enferma respectivamente. Se considera una población sana cuando $D = 0$ y enferma cuando $D = 1$ aplicando un punto de corte $c$, considerándose individuos positivos si superan dicho umbral y en otro caso negativos.

$F$ y $G$ son las funciones de distribución de las variables $X$ y $W$ respectivamente, definiéndose la sensibilidad como $Sen(c) = 1 - G(c)$ y la especificidad como $Esp(c) = F(c)$. Como resultado, se define a la curva ROC como un gráfico de la relación entre la sensibilidad y la inversa de la especificidad $1 - Esp(c)$ siendo el punto de corte $-\infty \leq c \leq \infty$. Expresada de forma analítica, la curva ROC se define en la Ecuación 1.52

$$ROC(t) = 1 - G\left(F^{-1}(1-t)\right), \tag{1.52}$$

siendo $t \in [0,1]$ y $F^{-1} = \text{ínf}\{x \in \mathbb{R} : F(x) \geq 1 - t\}$. Las curvas ROC son crecientes e invariantes en su monotonía bajo transformaciones de las variables $X$ y $W$. Una de las medidas más utilizadas para determinar la precisión de las variables a partir de las curvas ROC es el AUC[26], definida en la Ecuación 1.53

$$AUC = \int_0^1 ROC\,(u)\,du. \tag{1.53}$$

La AUC puede ser definida como la probabilidad de que el valor del test de diagnóstico para casos positivos o enfermos sea mayor que para los casos negativos o sanos, seleccionando aleatoriamente un par de individuos negativos y positivos $AUC = P(W > Y)$. Un valor cercano a uno de la AUC, indica una mayor precisión del test utilizado para clasificar a los individuos. Será especialmente útil en el análisis estadístico de la memoria para la comparación de la precisión de los diferentes métodos de estimación del parentesco con relación a la predicción de la ancestralidad.

[25] *Receiver Operating Characteristic*
[26] *Area under the Curve* o por sus siglas en inglés AUC

# Capítulo 2

# Aplicación práctica

Los GWAS tienen como principal objetivo estudiar la base genética del riesgo de padecer una enfermedad u otro trastorno o característica fenotípica (gravedad, capacidad de respuesta, etc.). En tal análisis, se debe controlar o ajustar el posible efecto de las diferencias ancestrales y el parentesco genético entre individuos. Por consiguiente, la caracterización de dichas relaciones familiares y ancestralidad es un paso fundamental en el control de calidad previo. La memoria se enfocará en estudiar diferentes métodos de control de calidad del GWAS, orientado a la incidencia de la Covid-19 sobre individuos latinoamericanos y españoles adscritos al Proyecto SCOURGE (Spanish COalition to Unlock Research on host GEnetics on COVID-19 (SCOURGE) s.f.). La población de estudio o cohorte no necesariamente incluye individuos ingresados en centros sanitarios, ya que se muestran aquellos que fuesen asintomáticos y sintomáticos leves hasta individuos con síntomas graves.

## 2.1. Presentación de los datos

El caso práctico de esta memoria se presenta dentro de un proyecto de investigación sobre la severidad del SARS-CoV-2, analizando datos genotipados desde centros adscritos al Proyecto SCOURGE en Latinoamérica y en España. El objetivo de dicho proyecto es elaborar un GWAS completo para determinar el factor genético que pueda afectar a la gravedad de los síntomas del virus a partir de la muestra de individuos seleccionados de la cohorte. La presente memoria se centrará únicamente sobre la fase de control de calidad, aunque tendrá en cuenta las directrices del GWAS para el resto de las fases.

### 2.1.1. Presentación de los datos genotipados del GWAS aplicado

#### Población y muestra de estudio

La muestra seleccionada de la cohorte está constituida por 3952 personas, tanto hombres como mujeres, de origen latinoamericano y español adscritos al Proyecto SCOURGE. Se identificaron 757836 marcadores genotipados de la muestra. [1]. Los individuos presentan diferentes estados de gravedad, desde asintomáticos o sintomáticos leves hasta aquellos que presentaron cuadros de síntomas graves y requirieron el ingreso hospitalario.

[1] Los marcadores son equivalente a los SNPs

**Recogida de datos y genotipado**

En base al protocolo del GWAS, la recogida de los datos genéticos y su genotipado de la cohorte las realizó el CEGEN y fueron analizados posteriormente en el CIMUS[2]. Los datos presentados en la memoria fueron extraídos de la información genotípica analizada desde el CIMUS y gracias a las colaboraciones del Proyecto SCOURGE desde Brasil, Ecuador, México, Paraguay y España.

**Objetivos**

Comprobar la eficacia de diversos métodos para la fase de control de calidad y recomendar el que mejor se adecúe a los objetivos del GWAS. Entre estos métodos, estarían los métodos de estimación del parentesco y de la ancestralidad poblacional.

### 2.1.2. Aplicación de la fase de control de calidad del GWAS

Los datos genotipados utilizados en el control de calidad del GWAS son categóricos. Cada categoría corresponde a una base nitrogenada de un gen o alelo en cada SNP, denotadas como: guanina(G), citosina(C), adenina(A) y timina(T). Cada SNP está formado por una combinación de estas bases de los dos genes de un segmento de ADN en un *locus* dado, siendo cada base única para cada gen o alelo. La disposición de los datos genotipados está representada en una tabla de doble entrada, donde cada fila corresponde a un individuo y cada columna a un SNP, donde a su vez se subdivide en dos columnas. Estas subcolumnas son los dos alelos que posee un individuo recogido en el SNP y las celdas de la tabla son los denominados valores genotípicos o genotipos representados por una de las categorías de bases nitrogenadas.

La extracción de los datos genotipados para su posterior control de calidad fue realizada a través del software PLINK [3]. Los ficheros de datos utilizados son archivos binarios que representan a los valores genotípicos en la tabla de doble entrada, siendo cada fila un individuo y cada columna un SNPs, con su nombre, su *locus* y su subdivisión en base a los dos alelos de cada individuo. La lectura con PLINK de estos ficheros será el punto de partida para el control de calidad de los datos según el GWAS. El procedimiento simultáneo de control por SNPs e individuos es el siguiente:

**Control de calidad por SNPs**

1. Exclusión de los cromosomas X e Y para todos los SNPs[4]
2. Selección de SNPs con un MAF menor al 5 %.
3. Filtrado de datos según el p-valor medio de Hardy-Weinberg fijado en $10^{-6}$.
4. Exclusión de SNPs con ratios de valores perdidos superiores al 20 %.
5. Extracción de los SNPs con alto Desequilibrio de Ligamiento, seleccionando a aquellos con un $R^2$ superior a la ventana seleccionada[5].

**Control de calidad por individuos**

1. Exclusión de individuos sin codificación del sexo.
2. Exclusión de individuos con baja tasa de genotipado o de valores perdidos mayores al 20 %.

[2] Centro Nacional de Genotipado(CEGEN) y *Center for Research in Molecular Medicine and Chronic Diseases* respectivamente

[3] Ver detalles en (Purcell *et. al* 2007), (Chang *et al.* 2015) y en la guía web de PLINK.

[4] Cromosomas X e Y son los cromosomas sexuales, se seleccionan únicamente autosomas.

[5] A través del comando –indep-pairwise en PLINK, se puede seleccionar el tamaño de la ventana y el número de SNPs necesarios para cambiar de ventana en cada fase, además del umbral $R^2$.

3. Inferencia del parentesco y Análisis de Componentes Principales.

4. Análisis de la ancestralidad poblacional.

Cabe comentar que los umbrales fijados no son únicos, sino seleccionados bajo el criterio del juicio de la persona investigadora. Esta aclaración se comenta adicionalmente para todos los umbrales aquí descritos en la memoria[6].

### 2.1.3. Tratamiento de los datos y estructura del análisis estadístico

#### Lectura y tratamiento de los datos

Una vez realizado el control de calidad del GWAS por individuos y por SNPs, se procede a la inferencia del parentesco como siguiente medida de control. Los datos de las estimaciones del parentesco se presentan en un *data frame*, siendo cada fila un par de individuos y cada columna un indicador o estimador del parentesco, presentados en la teoría de la memoria. Para los métodos de estimación de parentesco: PLINK,KING-robust y PC-Relate junto con PC-AiR, se creó una variable categórica adicional que permite determinar el grado de parentesco de cada par de individuos[7] según los umbrales fijados. Dichos umbrales están recogidos en las tablas de clasificación según los valores teóricos de la proporción de alelos IBD y del coeficiente kinship en las Tablas A.3 y A.4 respectivamente del Apéndice. A continuación, se calcularon las diez primeras componentes principales sobre la matriz GSM de cada método de estimación para cada individuo en función de los SNPs. Cada individuo observado se codifica con cuatro letras, seguidas de una barra baja y cuatro números. Las cuatro letras son las iniciales del centro de extracción de la información genética y los números son el código de identificación del individuo.[8]

Si se da alguna posible confusión del parentesco, es decir, pares de individuos clasificados como parientes cuando realmente no lo son, se aplicarán arbitrariamente umbrales que permitan reconocer a los parientes confusos o no creíbles. Los criterios de selección de parientes confusos o no creíbles son:

- El umbral mínimo de los estimadores $\hat{\pi}_{ij}$ , $\hat{\phi}_{ij}$ o $\hat{\kappa}_{ij}$ a partir del cual dos individuos se consideran parientes cercanos, siendo un 12.5 % para la proporción de dos alelos IBD $\hat{\pi}_{ij}$ y del 4.42 % para los coeficientes kinship $\hat{\phi}_{ij}$ y $\hat{\kappa}_{ij}$.
- El umbral máximo para excluir a los parientes considerados monocigotos o duplicados. Se fijó en un 98 % para la proporción de alelos IBD $\hat{\pi}_{ij}$ y en un 35.4 % para los coeficientes kinship $\hat{\phi}_{ij}$ y $\hat{\kappa}_{ij}$.
- El umbral máximo del 5 % de la probabilidad de que los pares de individuos tengan una copia genotípica IBD respecto al alelo de referencia, expresada como $P(Z = 1)$, $\hat{\pi}_{1ij}$ o $\hat{\varpi}_{1ij}$.

#### Estructura del análisis estadístico

El análisis estadístico constará de tres análisis diferentes: el análisis exploratorio e identificación del parentesco, el análisis de la ancestralidad poblacional y la evaluación de la precisión de los métodos de inferencia de parentesco una vez finalizado el control de calidad. La primera parte comenzará con una representación de gráficos de dispersión clasificados con diferentes colores según el grado de parentesco.

En el eje de abscisas se representa la probabilidad de que dos individuos tengan alelos IBD en PLINK y PC-Relate y la probabilidad de que dos individuos tengan alelos IBS para KING-robust. Por

[6] Excepto los criterios de inferencia en la clasificación de parientes en KING y los valores teóricos de la proporción de alelos IBD, mostrados en las Tablas A.3 y A.4 del Apéndice.

[7] La notación de los grados de parentesco está recogida en el Glosario A.2.

[8] Como ejemplo, los datos extraídos de un individuo ficticio en el Centro Hospitalario Universitario de Santiago de Compostela(CHUS) sería *CHUS_0000*.

otro lado, el eje de ordenadas representa el valor del coeficiente kinship para los tres métodos para cada individuo $i$ y $j$ respectivamente. Todos los gráficos distinguieron los escenarios posibles aplicando o no el *LD pruning*. La clasificación de los individuos se realizó utilizando la media aritmética entre los valores teóricos en PLINK correspondientes a cada intervalo del grado de parentesco y utilizando la media geométrica en KING y PC-Relate para los mismos grados de parentesco. El gráfico correspondiente a KING no utiliza la probabilidad IBD, debido a la falta de un *pedigree* necesario para estimar Z0 en KING-robust, en su lugar se utilizó la probabilidad IBS como alternativa similar.

Dentro de la primera parte, se identificaron aquellos núcleos o conjuntos de individuos emparentados en una misma familia genética para el primer, segundo y tercer grado de parentesco, comentándose posibles incongruencias de los métodos en la identificación de los núcleos parentales.

La segunda parte se enfocará en la representación gráfica del PCA para cada método estimativo considerando únicamente el escenario bajo la aplicación del *LD pruning*[9]. De forma arbitraria, se fijó una proporción mínima del 80 % para considerar a un individuo perteneciente a una estructura poblacional concreta. En el caso de que no se cumpla dicho umbral ,se dirá que tiene ancestralidad mezclada o perteneciente a una población *ADMIXED*. Las denominadas estructuras poblacionales o grupos ancestrales son las siguientes: europea, africana, nativo-americana y población mezclada o *ADMIXED*, cuyas proporciones fueron estimadas a través de STRUCTURE (Raj *et al.* 2013)[10]. Las notaciones de los grupos ancestrales están recogidas en el Glosario A.2 del Apéndice. Los gráficos de los PCA, siendo el eje de abscisas la primera componente y el eje de ordenadas la segunda, están recogidos sobre el total de individuos. Adicionalmente, se considerará simultáneamente la representación de las componentes principales del subconjunto de parientes, permitiendo la relación entre la ancestralidad poblacional y las estimaciones del parentesco.

Dentro de la segunda parte, se representarán gráficamente las correlaciones a partir de la función *corrPlot* de la librería *pysch* en el lenguaje de R (Revelle y Condon 2018) de las dos primeras componentes principales con respecto a las estructuras poblacionales. Sean dos variables $X$ e $Y$, se puede determinar su correlación lineal de Pearson (Dagnino 2014) a partir de la Ecuación 2.1

$$r = \frac{Cov(XY)}{\sqrt{Var(X)Var(Y)}}. \tag{2.1}$$

A partir de la definición de la Ecuación 2.1, se mostrarán simultáneamente los correlogramas del conjunto total de individuos indicando en una escala de color la intensidad de dicha correlación y el valor del coeficiente de correlación de Pearson. Además, también se representarán los correlogramas para el subconjunto de parientes de cada método de estimación. Cabe advertir que el subconjunto de parientes no es un PCA, sino que filtra las componentes principales en función de la presencia de parentesco de los individuos. A pesar de que las componentes principales son ortogonales, es posible que para el subconjunto de parientes pueda darse correlaciones entre componentes, sin poder ser interpretables al no tratarse de un PCA.

La tercera y última parte, evaluará la precisión de los métodos de estimación del parentesco sobre las proporciones ancestrales a través de los modelos PLS y una aproximación al PLSGLR. La muestra de los modelos será el subconjunto de individuos no emparentados con sus correspondientes valores de sus diez componentes principales y sus proporciones ancestrales. Tal y como se demuestra en la literatura referenciada en esta memoria (Zhu *et al.* 2008), es posible representar el efecto del tipo de método de inferencia del parentesco en poblaciones mezcladas o ADMIXED utilizando modelos de regresión lineales. Se construirán dos modelos para cada grupo ancestral y cada método estimativo. Uno

[9]Para los Análisis de Componentes Principales y siguientes apartados, se utilizarán los datos aplicando la exclusión de marcadores de SNPs de alto LD, para evitar dependencia estadística de los SNPs.

[10]Estas proporciones también pueden ser calculadas a través de otros métodos de estimación de la ancestralidad como ADMIXTURE (Alexander, *et al.* 2009) bajo el escenario de población mezclada.

será un modelo PLS de regresión , considerando a la variable respuesta como la proporción ancestral de los individuos no emparentados y como variables explicativas los valores de las diez componentes principales. El segundo modelo PLSGLR será de clasificación, donde la variable respuesta se representa como una variable binaria. Si la proporción ancestral supera el 70 %, se considerará una probabilidad alta de pertenencia a dicho grupo ancestral, en caso contrario se clasificará con probabilidad baja. Las variables explicativas del modelo serán los valores de las diez componentes principales. En ambos modelos se dispondrá de una muestra de entrenamiento que represente al 80 % de la muestra y un 20 % como la muestra de test para las predicciones de los modelos. Además, se representarán gráficamente los resultados de las predicciones, así como otros indicadores que sean pertinentes. Los modelos PLS fueron ejecutados en la función *plsr* del paquete *pls* y los modelos PLSGLR con la función *glm* del paquete *stats* del lenguaje R. Esta parte fue realizada a partir de la literatura de mencionada (Mevik y Wehrens 2007), (Wold *et al.* 1983) y (James *et al.* 2021) para los modelos PLS y para los modelos PLSGLR adicionalmente (Bertrand *et al.* 2014).

Por último, toda la notación y siglas utilizadas en el análisis estadístico están descritas en el Glosario A.2 del Apéndice, para una mejor lectura de este.

## 2.2. Inferencia del parentesco

El análisis estadístico de los datos parte del procedimiento del control de calidad del GWAS previo a la inferencia del parentesco utilizando el software PLINK. Sin embargo, esta fase se puede replicar con cualquier otro software o método que permita la elaboración de GWAS. A continuación, se mostrarán con mayor detalle los resultados obtenidos de la inferencia del parentesco en los métodos utilizados.

### 2.2.1. Análisis gráfico de las relaciones de parentesco

En los gráficos de dispersión de la Figura 2.1, quedan reflejados los grados de parentesco por colores a partir de los umbrales explicados en las Tablas A.3 y en A.4 del Apéndice. Aquellos pares de individuos con mayor valor estimado del coeficiente kinship son aquellos que tienen menor probabilidad de que sus alelos no contengan copias IBS e IBD. En el caso extremo, los que tienen mayor valor estimado del coeficiente kinship son los monocigotos considerados también como gemelos, o los duplicados cuando un individuo fue codificado más de una vez por error. En cambio, los que presentan menor coeficiente kinship estimado son los parientes de tercer grado, con una probabilidad Z0 o I0 más elevada.

En un escenario ideal, todos los métodos de estimación deberían identificar los mismos parientes independientemente del método. De hecho, todos los softwares identificarían de la misma forma a los parientes con una relación más estrecha: 23 pares monocigotos o duplicados y sobre 70 pares de primer grado(padres-hijos o hermanos completos)[11]. Sin embargo, tal y como se demuestra en la Figura 2.1, el número de parientes identificados es distinto según el procedimiento, especialmente en los parientes de tercer grado. Bajo esta casuística, será necesario comprobar el funcionamiento de los tres métodos de estimación del parentesco de los pares de individuos. Solamente se podría observar que PLINK presenta una mayor cantidad de parientes de tercer grado con respecto al resto de métodos.

Con respecto a la comparativa del *LD pruning*, no se aprecian grandes diferencias en los algoritmos, siendo más notable dicha diferencia en los parientes de tercer grado en PLINK. Por último, se podría deducir que KING y PC-Relate presentan una menor confusión de los parientes según su grado de parentesco. Dicho de otro modo, cada grupo de parientes clasificados por un grado de parentesco está claramente definido y separado con respecto a los otros grupos de diferente grado.

[11] Se observaron pequeñas diferencias en los pares de primer grado entre métodos, posiblemente debido a pares con valores de las estimaciones muy próximos a los umbrales o debido a errores de la codificación de los datos.

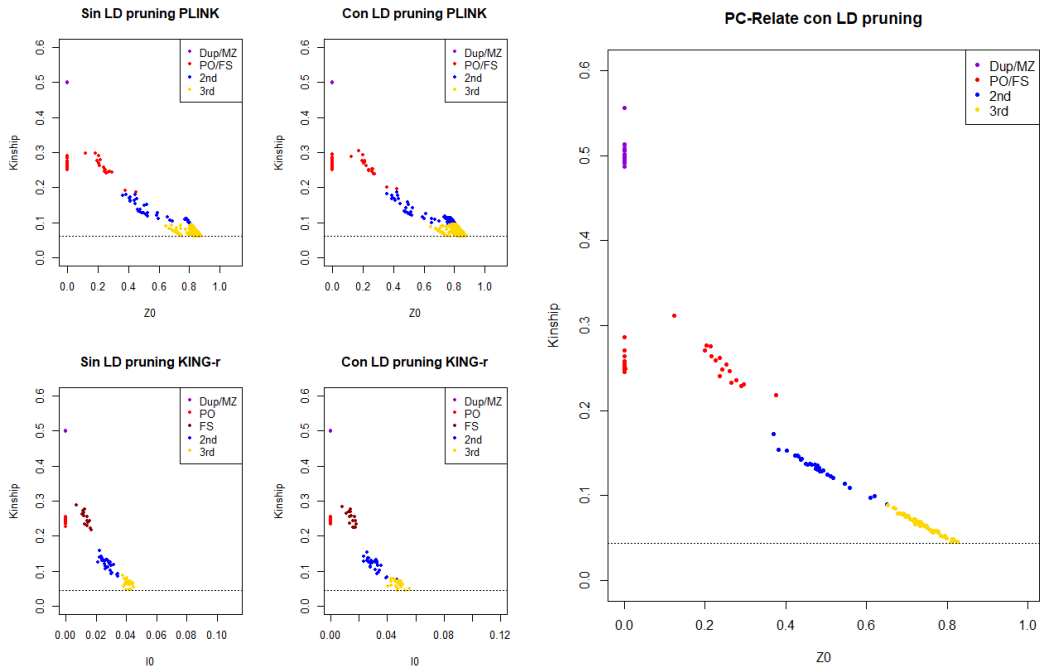


Figura 2.1: Gráficos de dispersión entre Z0,I0 y Kinship. Fuente: Elaboración propia

### 2.2.2. Identificación de núcleos de parientes cercanos

A continuación, a partir de la información de los métodos de estimación del parentesco, se comparará el número de parientes cercanos distinguiendo el *LD pruning*. Para detectar que los algoritmos estiman de forma correcta el parentesco, la situación ideal sería que todos detectasen los mismos parientes y/o el mismo número de núcleos de parentesco[12]. A pesar de distinguir diferencias entre los parientes detectados con o sin *LD pruning*, en la práctica se utilizarán los datos una vez aplicados en el procedimiento de exclusión de alto LD.

#### Núcleos parentales de primer y segundo grado en PLINK

El recuento de núcleos parentales se realizó identificando grupos de individuos que estuvieran emparentados con unos pocos o entre sí. Según se aprecia en la información recogida en las Tablas A.5 y A.6 del Apéndice, para parientes de primer grado se identificaron cinco núcleos sin aplicar *LD pruning* a los SNPs y siete núcleos una vez aplicado. Cada uno de ellos presentan entre tres y cuatro individuos por núcleo, y cada miembro de cada núcleo pertenece a un centro de extracción común, a pesar de que este caso no tiene por qué darse siempre.

En parientes de segundo grado, se identifican tres núcleos parentales creíbles considerando o no el *LD pruning* indistintamente. El núcleo de mayor tamaño son aquellos individuos del núcleo TECM, observando seis individuos miembros.

[12]Grupo o familias de individuos que están emparentados entre sí, formando un conjunto o núcleo parental.

### Núcleos parentales de primer y segundo grado en KING

La información sobre los núcleos parentales, recogida en las Tablas A.7 y A.8 del Apéndice, recogen a los núcleos parentales detectados con KING-robust. Sin necesidad de distinguir los datos con o sin *LD pruning*, se contabilizan ocho núcleos de parientes de primer grado y tres de segundo grado. Ambos son idénticos a los ya detectados con PLINK en parientes de segundo grado, salvo que si no se considera el*LD pruning*, el núcleo TECM consta de seis individuos en vez de los cinco detectados aplicando a exclusión de alto LD, concretamente se añade el individuo TECM_0745, a quién no se considera en el último escenario en KING.

### Núcleos parentales de primer y segundo grado en PC-Relate

A partir de las Tablas A.9 y A.10 del Apéndice, se recogen los núcleos parentales identificados con PC-Relate. Para parientes de primer grado, se cuantifican ocho núcleos parentales, idénticos a los ya obtenidos con PLINK. Sin embargo, en parientes de segundo grado se considera un nuevo núcleo parental con respecto a PLINK y a KING, que incluye tres individuos pertenecientes a los grupos UFPA,UFRN y UFPE. El motivo de esta diferencia se debería estudiar más en detalle en la fase de recogida de datos y genotipado del GWAS, por si hubiera algún error en la codificación.

### Identificación de núcleos parentales de tercer grado

Hasta el momento, se consideraron aquellos parientes que fuesen creíbles, es decir, aquellos en los que un método de estimación clasifica correctamente su grado de parentesco. Sin embargo, como se demuestra en el artículo (Ramstetter *et al.* 2017), es probable que a medida que aumente el grado de parentesco, los algoritmos de los métodos de estimación puedan presentar incongruencias en la clasificación. Debido a la gran cantidad de parientes de tercer grado, solamente se representó el recuento de individuos emparentados con unos pocos, de forma similar a la identificación de los núcleos parentales de primer y segundo grado. Se recogen los recuentos de los cinco mayores núcleos en función a su tamaño en la Tabla 2.1. Además, se distinguirán los parientes según el criterio del *LD pruning*.

A partir de la información sobre los tamaños de los núcleos en la Tabla 2.1, PLINK detecta los cinco núcleos parentales más grandes con un tamaño muy superior a los cinco más grandes detectados con KING o PC-Relate, siendo ligeramente superiores si se aplica el *LD pruning*. Tal y como se observa en los núcleos de tercer grado de PLINK, puede afirmarse que gran parte de los parientes de tercer grado detectados no son creíbles. Además, se presenta una diferencia sustancial con respecto al resto de métodos de estimación, donde apenas alcanzan los cuatro individuos por núcleo. El motivo de esta diferencia se puede deber a que PLINK esté confundiendo a los pares que tienen una ancestralidad común como parientes reales, siendo necesario un análisis de las estructuras poblacionales para conocer el origen de esa confusión.

| **NÚCLEOS DE PARIENTES DE TERCER GRADO CON *LD pruning*** | | | | | |
|---|---|---|---|---|---|
| **PLINK** | | **KING-robust** | | **PC-Relate** | |
| **IID** | **Nº parientes** | **IID** | **Nº parientes** | **IID** | **Nº parientes** |
| CNMI_0001 | 692 | CNMI_0389 | 4 | UFPA_0296 | 4 |
| CNMI_0469 | 687 | H12O_004_R | 4 | UFPE_0193 | 4 |
| CNMI_0608 | 683 | CNMI_0514 | 3 | PARG_0166 | 3 |
| CNMI_0642 | 683 | COVI_0569 | 3 | PARG_0197 | 3 |
| COVI_0152 | 659 | FJD_2183 | 3 | PARG_0205 | 3 |
| **NÚCLEOS DE PARIENTES DE TERCER GRADO SIN *LD pruning*** | | | | | |
| CNMI_0001 | 636 | CNMI_0389 | 4 | | |
| CNMI_0004 | 619 | H12O_004_R | 4 | | |
| CNMI_0389 | 616 | CNMI_0514 | 3 | | |
| CNMI_0469 | 612 | COVI_0569 | 3 | | |
| CNMI_0684 | 611 | FJD_2183 | 3 | | |

Cuadro 2.1: Núcleos de parientes de tercer grado. Fuente: Elaboración propia

## 2.3. Análisis de la estratificación poblacional

En este apartado, el análisis estadístico se centrará en la estratificación poblacional que presenta la cohorte haciendo uso de técnicas estadísticas exploratorias y de reducción de la dimensión, concretamente del Análisis de Componentes Principales y del coeficiente de correlación de Pearson. Se utilizarán las tres estructuras poblacionales expresadas en proporciones con rango entre cero y uno, denotadas en el glosario A.2 del Apéndice: europea, africana y nativo-americana. Sin embargo, bajo presencia de población mezclada, es necesario incluir la estructura poblacional mezclada, representada cuando un individuo tiene una ancestralidad no definida o ambigua[13]. Las estimaciones de las proporciones de ancestralidad se determinaron a partir del software STRUCTURE (Raj *et al.* 2013).

### 2.3.1. Análisis de Componentes Principales y parentesco

El Análisis de Componentes Principales[14] resume la información genotípica en un número determinado de componentes con máxima varianza (Zhu *et al.* 2008) y (Chang *et al.* 2015). El PCA está estrechamente relacionado con el estudio de la ancestralidad genética , ya que cada componente capta parte de la variabilidad de la estructura poblacional o ancestralidad de cada individuo.

[13] En inglés *ADMIXED*, definida cuando un individuo no tiene una proporción predominante de ancestralidad.
[14] *Principal Component Analysis* o por sus siglas en inglés PCA.

Por regla general, el PCA tiende a presentar grupos definidos y delimitados, formados por los valores de las componentes principales. Sin embargo, con los datos actuales, se puede observar que hay dos grupos que están entrelazados por valores intermedios de las componentes principales, formando una .ªla delta.en los gráficos de la izquierda de las Figuras 2.2, 2.3 y 2.4. Esta representación inusual se debería a la presencia de población mezclada, ya que permite la existencia de individuos que no tengan una estructura poblacional dominante. Estos gráficos representan al PCA del conjunto de individuos de la muestra clasificando sus valores por sus estructuras poblacionales. Las representaciones del lado derecho representan al subconjunto de componentes principales correspondientes a los parientes de primer, segundo y tercer grado. Estos últimos no son PCA, sino una submuestra de las componentes principales del PCA considerando únicamente a los parientes cercanos detectados en la inferencia del parentesco. Además, se incluyen en puntos de color rojo a los parientes que no se consideran creíbles bajo el criterio de selección de parientes confusos.

A partir de la Figura 2.2 de PLINK, se aprecia una concentración de puntos en valores reducidos del PC1 y relativamente altos para el PC2. Estos individuos a su vez coinciden con el subconjunto de individuos que pertenecen a los parientes no creíbles, también muy concentrados en la misma área. Comparando el gráfico del PCA completo con el de parientes cercanos, se deduce que los parientes no creíbles en PLINK, clasificados como de tercer grado detectados en la inferencia del parentesco, tienen una estructura poblacional nativo-americana. Esta deducción conlleva a que esos parientes no creíbles realmente están relacionados por su ancestralidad común nativo-americana y no por una relación de parentesco genético.

Siguiendo la misma línea con las Figuras 2.3 y 2.4, KING-robust y PC-AiR presentan el mismo PCA completo, debido a la unicidad de las componentes principales, solamente diferentes en signo. Con respecto a las representaciones del subconjunto de los parientes, muestran una menor concentración de valores de componentes principales con respecto a PLINK , además de un menor número de parientes no creíbles. En alusión a la teoría de KING (Manichaikul *et al.* 2010), también se pueden incluir aquellas componentes principales de pares de individuos que tengan coeficientes kinship negativos, siendo en este caso aquellos con un coeficiente kinship estimado menor a $-0{,}25$. Se podría deducir una mayor concentración de individuos con el coeficiente kinship negativo con una probabilidad alta de pertenecer a la estructura poblacional nativo-americana, coincidente con los parientes no creíbles de PLINK , como se muestra en la Figura B.1 del Apéndice.

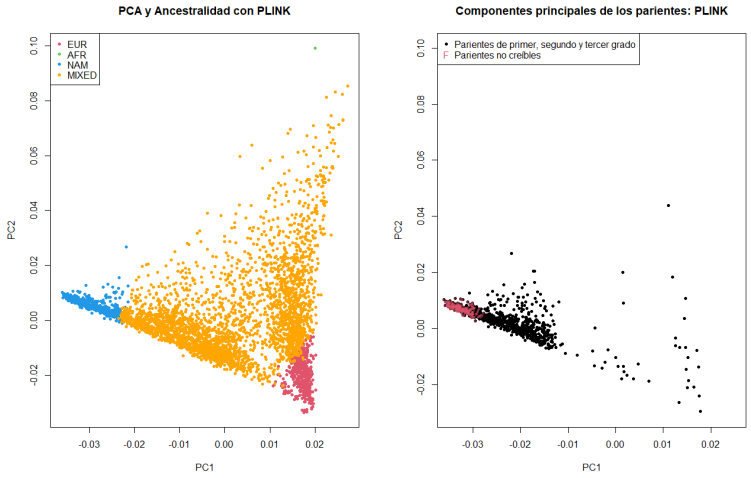


Figura 2.2: PCA ,estructura poblacional e identificación de parientes: PLINK. Fuente: Elaboración propia

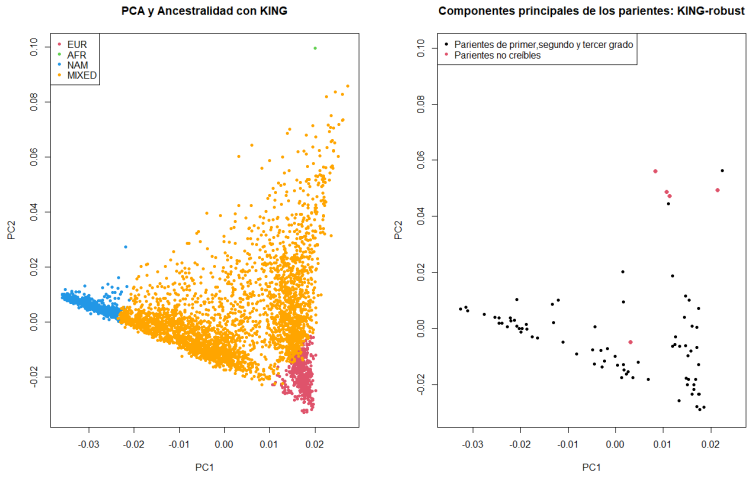


Figura 2.3: PCA, estructura poblacional e identificación de parientes: KING-robust. Fuente: Elaboración propia

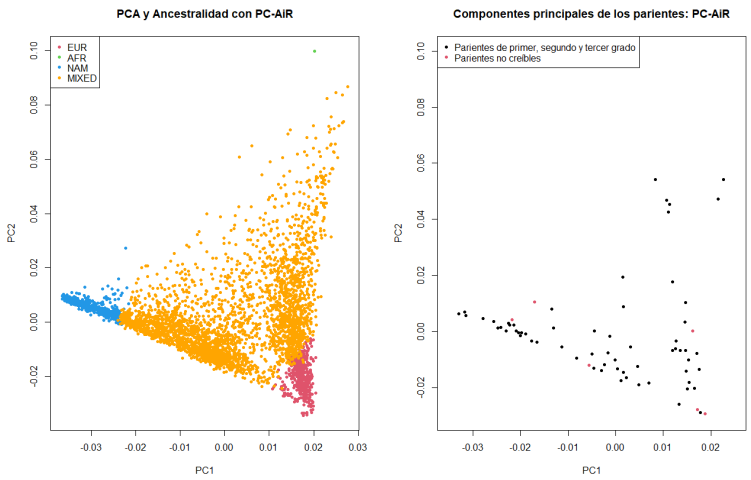


Figura 2.4: PCA, estructura poblacional e identificación de parientes: PC-AiR. Fuente: Elaboración propia

### 2.3.2. Estudio de la correlación con estratificación poblacional

En base al estudio del PCA y la ancestralidad, es posible relacionar los valores de las componentes principales sobre el efecto en la ancestralidad genética de la población. Como paso previo a una futura modelización del efecto de las componentes principales sobre la ancestralidad, se recomienda un análisis exploratorio de la correlación entre estas variables. La estructura del apartado seguirá la misma línea que la del anterior, distinguiendo entre el PCA completo y el subconjunto de parientes cercanos. Cabe señalar que las componentes principales mantienen su ortogonalidad y unicidad en el PCA original, por lo tanto, aunque se presente cierta correlación entre componentes en el análisis de parientes cercanos, no es interpretable al tratarse de un subconjunto de las componentes principales de la muestra y no de un PCA.

A continuación, se representan los correlogramas en las Figuras 2.5, 2.6 y 2.7. Estos gráficos muestran correlogramas representados en mapas de calor e integrados con el valor del coeficiente de correlación lineal de Pearson. Los gráficos del extremo izquierdo de las figuras representan a las correlaciones considerando el PCA completo de todos los individuos. Los gráficos del extremo derecho consideran el subconjunto de componentes principales de los parientes de primer, segundo y tercer grado con sus correspondientes proporciones de ancestralidad.

Si se considera el PCA original completo, incluyendo a todos los individuos, los resultados son similares para todos los métodos de estimación, tanto en PLINK, como KING como en PC-AiR. Se presenta una correlación negativa fuerte entre la primera componente y la estructura poblacional nativo-americana a diferencia del resto de estructuras poblacionales, donde se recogen correlaciones positivas más discretas. Con respecto a la segunda componente, los algoritmos presentan únicamente una correlación notable con el grupo ancestral africano. Con respecto a la correlación entre subpoblaciones ancestrales, se observa que entre la estructura poblacional europea y nativo-americana, existe una correlación negativa fuerte, y menor entre el grupo ancestral africano y nativo-americano.

Si se consideran solamente parientes reales en PLINK , la primera componente está más correlacionada con la estructura poblacional europea y la segunda más con la nativo-americana. Si se observa la correlación entre subpoblaciones ancestrales, se observa que el único cambio destacable es el incremento de la correlación del grupo ancestral africano con respecto al europeo del PCA completo. Por otro lado, KING-robust presenta una mayor correlación positiva con los grupos ancestrales nativo-americano y africano e incorrelación entre la población europea y africana a diferencia de PLINK. Según PC-Relate, se observa que la segunda componente presenta incorrelación con respecto al grupo ancestral nativo-americano, ya observado en las diferencias entre PLINK y KING-robust.

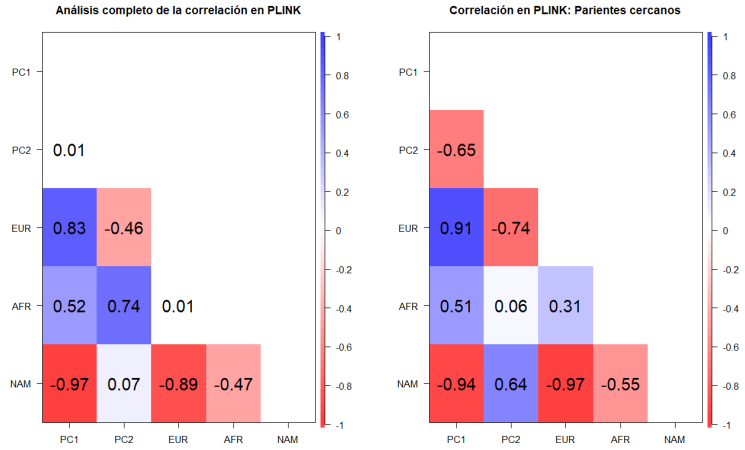


Figura 2.5: Correlogramas PCA-Estructura poblacional:PLINK. Fuente: Elaboración propia

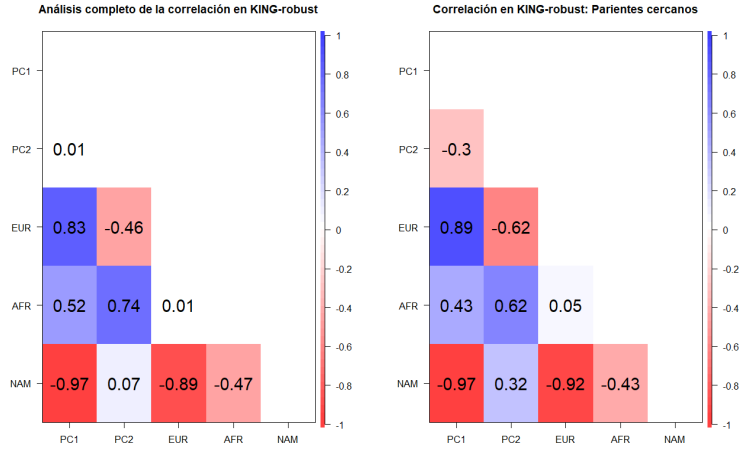


Figura 2.6: Correlogramas PCA-Estructura poblacional: KING-robust. Fuente: Elaboración propia

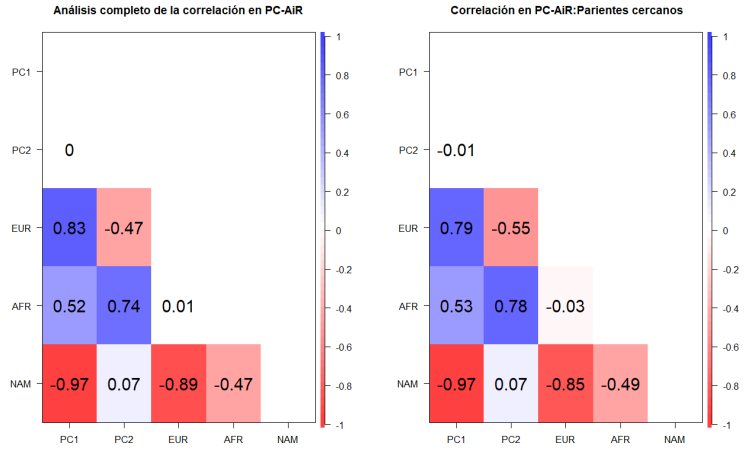


Figura 2.7: Correlogramas PCA-Estructura poblacional: PCAIR. Fuente: Elaboración propia

## 2.4. Evaluación de la precisión de los métodos de estimación del parentesco

### 2.4.1. Modelos de regresión PLS

Una vez identificados los parientes y estudiadas sus estructuras poblacionales, se excluyen a los parientes del conjunto de datos para continuar con el GWAS, seleccionando solamente aquellos individuos no emparentados. En base a este procedimiento, se analizará cuál de los métodos de estimación es más preciso, estudiando la relación de la estimación de cada estructura poblacional con las componentes principales de los pares de individuos no emparentados.

Tal y como se demuestra en la teoría relacionada (Zhu *et al.* 2008), es posible vincular las componentes principales con una variable cuantitativa que represente a una proporción de una estructura poblacional. Este vínculo se podría estudiar para este caso a partir de un modelo de regresión lineal, donde la variable respuesta será una proporción con rango entre cero y uno bajo las condiciones de una población mezclada, y las variables explicativas serían los valores que toma cada componente principal con sus correspondientes cargas o *loadings*. El resultado será un modelo que represente la proporción de variabilidad explicada por las componentes principales de la proporción de ancestralidad de un individuo en cada estructura poblacional.

Debido a la dificultad de realizar un modelo que determine la precisión de un método estimativo, por las grandes dimensiones de los datos, se recurrirá a modelos de regresión PLS, utilizados y explicados en numerosas investigaciones (Mevik y Wehrens 2007), (Wold *et al.* 1983) y (James *et al.* 2021).

Se dispondrá de una muestra de entrenamiento del 80 % del total de individuos no emparentados y de un 20 % como muestra de test.[15].

Se presentan nueve modelos PLS a partir de las muestras de entrenamiento, cuyas variables explicativas son las diez primeras componentes principales y sus variables respuesta una proporción de pertenencia a una estructura poblacional. Primeramente, se ajustan los modelos estandarizando las variables predictivas obteniéndose los valores de los coeficientes de regresión en la Figura 2.8. Si se considera el grupo ancestral europeo EUR como respuesta, se obtiene que las dos primeras componentes principales presentan valores muy alejados del valor nulo, siendo similar para todos los métodos estimativos. Si la respuesta es el grupo ancestral nativo-americano NAM, el valor de la primera componente es mucho menor al valor nulo y la segunda componente presenta un valor positivo más reducido. Finalmente, si se considera a la respuesta como el grupo ancestral africano AFR, se observa también que las dos primeras componentes son las que presentan mayores diferencias respecto al valor nulo. Si se observan las diferencias entre métodos, prácticamente todos los modelos presentan valores similares de las componentes principales para cada estructura poblacional. Sin embargo, para el grupo ancestral africano, se observan sutiles diferencias a partir de la tercera componente entre PLINK, KING-robust y PC-AiR.

Adicionalmente, se analizan las proporciones de variabilidad explicada de cada componente en el modelo en la Figura 2.9. Considerando a PLINK, se observa que la componente que mejor explica el modelo es la cuarta en el grupo ancestral nativo-americano NAM, la segunda y la quinta para el grupo europeo y la segunda para el grupo africano. Sin embargo, estas proporciones máximas difieren con el resto de los métodos estimativos. KING presenta la proporción de variabilidad máxima en la octava componente con el grupo ancestral nativo-americano y en la segunda con el grupo europeo y africano. Finalmente, en PC-AiR no hay una componente con una proporción de variabilidad máxima clara, ya que todas ellas presentan proporciones similares y muy repartidas. Esto puede deberse a

[15]Es necesario señalar que los algoritmos pueden presentar diferencias en el número de individuos.

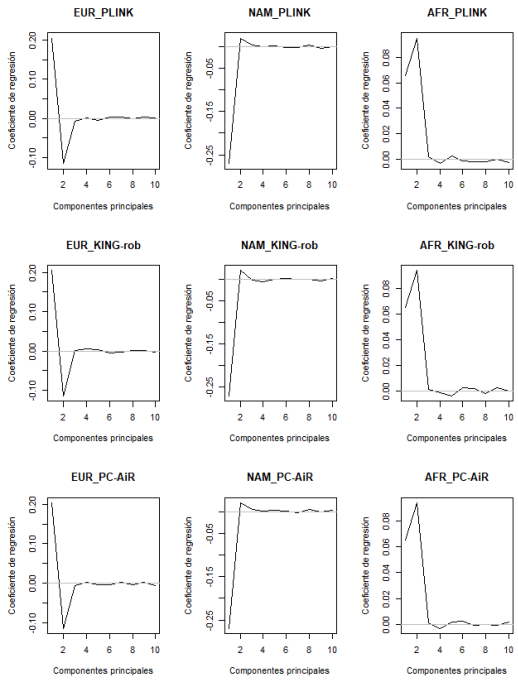


Figura 2.8: Interpretación de los coeficientes de regresión PLS. Fuente: Elaboración propia

la consideración previa del PCA sobre los pares no emparentados, propio de la metodología de PC-AiR.

Una vez analizado el modelo de entrenamiento, será necesario seleccionar el número de componentes óptimo para la predicción de los modelos. Según la documentación del método PLS (Mevik y Wehrens 2007), esta selección se realiza en base a la Validación Cruzada utilizando la medida de error predictivo RMSEP[16] para cada componente y seleccionando aquel que tenga un RMSEP menor.

Cabe advertir que, si se considerasen todos los individuos, todos los métodos tendrían el mismo número de componentes óptimos para cada grupo ancestral. Sin embargo, se consideraron a los individuos no emparentados seleccionados de la inferencia de parentesco, cuyo recuento es diferente según el método utilizado. Las diferencias del número óptimo de componentes se deben a los resultados implícitos de la inferencia de los individuos no emparentados, determinándose diferentes números de componentes óptimos. Los resultados mostrados en la tabla 2.2 muestran el número óptimo de componentes para cada modelo ajustado, mostrándose en PLINK un mayor número de componentes necesarios para la predicción del modelo. PC-AiR sería el algoritmo que presenta menor número de componentes óptimo, siendo dos componentes para todas las estimaciones de los grupos ancestrales. Este último caso, se debería a la propia metodología de PC-AiR, ya que utiliza los mismos valores de las componentes principales de los individuos no emparentados para luego predecir dichos valores en los parientes. Este procedimiento se realiza considerando previamente a los grupos ancestrales, obteniéndose un único modelo de regresión para cada uno de los grupos ancestrales descrito en la Ecuación 1.37 y en la teoría de PC-AiR.

A partir de los modelos ajustados con el número de componentes óptimo, se realiza el ajuste predictivo sobre la muestra de test para cada uno de los modelos, tal y como se representa en la Figura B.2 del Apéndice, sobre los valores observados de la muestra de test. Para esta fase, se recurrió a la

[16] *Root Medium Square Error of Prediction*, o Raíz del Error Cuadrático Medio de Predicción traducido del inglés.

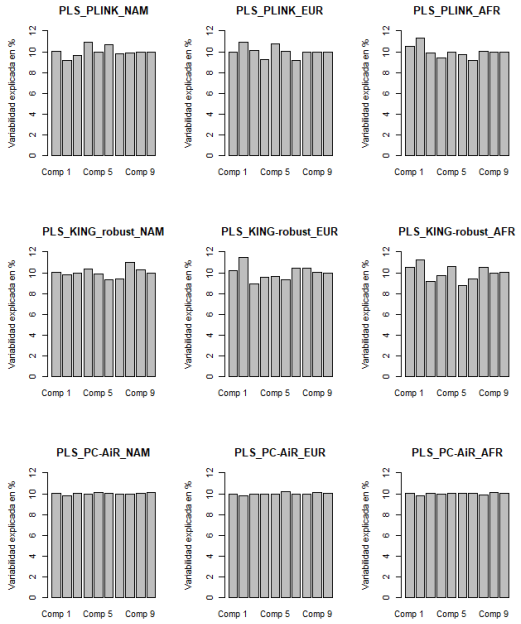


Figura 2.9: Proporciones de variabilidad explicada. Fuente: Elaboración propia

| **Selección del número de componentes** | | | |
|---|---|---|---|
| | **NAM** | **EUR** | **AFR** |
| **PLINK** | 5 | 3 | 5 |
| **KING-robust** | 4 | 3 | 2 |
| **PC-AiR** | 2 | 2 | 2 |

Cuadro 2.2: Selección del número de componentes óptimos por CV. Fuente: Elaboración propia

comparación del $R^2$ ajustado en la tabla 2.3 siendo el coeficiente de correlación de Pearson elevado al cuadrado (Dagnino 2014). La motivación de su uso es debido a que los modelos presentan diferente número de variables explicativas y representa la proporción de variabilidad explicada por las predicciones sobre las observaciones reales.

En general, todos los modelos con el grupo ancestral nativo-americano como respuesta, explican más del 95 % de su variabilidad, cerca del 92 % del grupo ancestral europeo y entre el 79.82 % y el 83.06 % del grupo africano. En este último se presentan las mayores diferencias entre métodos, siendo KING-robust el método que mejor explica la variabilidad de todas las estructuras poblacionales.

| $R^2$ ajustado de los modelos predictivos | | | |
|---|---|---|---|
| | Variable respuesta | | |
| | **NAM** | **EUR** | **AFR** |
| **PLINK** | 0.9536 | 0.9192 | 0.8028 |
| **KING-rob** | 0.9551 | 0.9228 | 0.8306 |
| **PC-AIR** | 0.9540 | 0.9160 | 0.7982 |

Cuadro 2.3: Variabilidad explicada de los modelos predictivos. Fuente: Elaboración propia

### 2.4.2. Comparativa de la precisión en modelos PLSGLR de clasificación

Como extensión al análisis del modelo PLS y al de la precisión de los métodos estimativos, se llevó a cabo un estudio sobre la capacidad predictiva de cada método a la hora de clasificar correctamente a un individuo en cada grupo ancestral. La clasificación se realizó ajustando un umbral arbitrario del 70 %, a partir del cual, existe una probabilidad alta de que un individuo pertenezca a la estructura poblacional.[17]

Para este análisis, se realizó una aproximación de los modelos generalizados de los PLS denominados PLSGLR, ajustando modelos de regresión logística para cada uno de los métodos. Las variables explicativas son las componentes principales y la variable respuesta un grupo ancestral, tomando únicamente dos categorías: probabilidad alta o baja de pertenecer a dicho grupo ancestral.

Disponiendo de las mismas proporciones para definir las muestras de entrenamiento y de test que en el estudio de los modelos PLS, se dispone a estudiar la capacidad predictiva de los métodos.

A partir de las curvas ROC definidas en la Figura 2.10, se presenta la comparación de la precisión de la clasificación de los distintos métodos para cada grupo ancestral en base a la especificidad y la sensibilidad de los modelos. Como se demuestra en las curvas, los métodos presentan una menor especificidad para clasificar a un individuo del conjunto de no emparentados con estructura poblacional europea que con el resto de los grupos. Los modelos presentan los mayores valores del AUC si la respuesta es la estructura poblacional africana. Considerando el AUC de las curvas ROC de la Tabla 2.4, PC-AiR clasifica mejor a aquellos que sean de estructura poblacional nativo-americana o africana que el resto de los algoritmos. Sin embargo, para el caso de la estructura poblacional europea, se considera mejor el método de KING-robust para la clasificación. Finalmente, PLINK presenta el menor AUC en los grupos ancestrales nativo-americanos y europeos y un AUC mayor en la estructura poblacional africana.

[17] Se redujo a este umbral con respecto al utilizado en el apartado de PCA debido a la nula representatividad de los individuos con estructura poblacional africana con un umbral del 80 %.

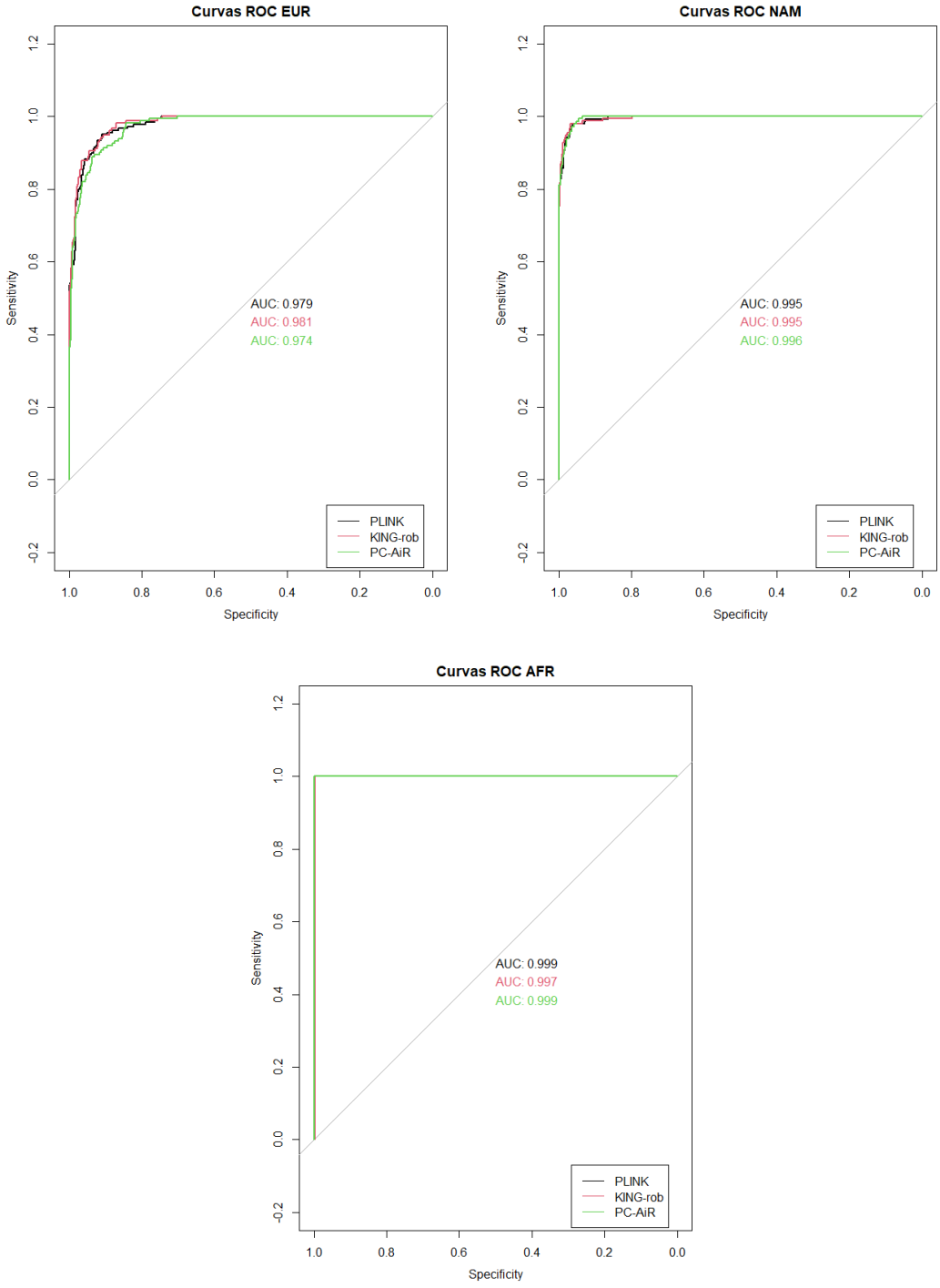


Figura 2.10: Correlogramas PCA-Estructura poblacional:PLINK. Fuente: Elaboración propia

| AUC de las curvas ROC | | | |
|---|---|---|---|
| | **EUR** | **NAM** | **AFR** |
| **PLINK** | 0.9786 | 0.9949 | 0.9987 |
| **KING-robust** | 0.9813 | 0.9949 | 0.9974 |
| **PC-AiR** | 0.9740 | 0.9959 | 0.9993 |

Cuadro 2.4: Área bajo la curva(AUC) de cada modelo

## 2.5. Discusión y conclusiones

El objetivo de la memoria ha sido comparar los tres métodos de estimación del parentesco y evaluar su precisión en la estimación de la ancestralidad, utilizando datos genotípicos de la cohorte latinoamericana y española del Proyecto SCOURGE. Dentro del procedimiento del GWAS, es necesario trabajar con individuos que no presenten relaciones genéticas más allá de las provocadas por la ancestralidad común para construir los test de asociaciones genéticas. Dicho de otro modo, se excluyen del GWAS a los individuos parientes genéticos que puedan distorsionar los resultados por su alta correlación genética en la fase de control de calidad. PLINK, KING-robust y PC-Relate utilizando PC-AiR, son métodos igualmente válidos para la fase de control de calidad de un GWAS, aunque ha sido de interés investigar posibles diferencias en los procedimientos.

Con respecto a PLINK, presenta una mayor proporción de parientes no creíbles que el resto de los métodos, considerados de tercer grado. Esto se demuestra con el recuento de núcleos parentales en la Tabla 2.1 donde se recoge una diferencia drástica de parientes en cada núcleo en PLINK con respecto al resto de métodos de estimación del parentesco, sin que sea posible en la práctica que un individuo esté emparentado en tercer grado con más de cien individuos. Se puede comprobar que estos parientes no creíbles pueden estar realmente relacionados por tener una ancestralidad común, siendo en este caso el grupo ancestral nativo-americano tal y como se observa en la Figura 2.2. Se demuestra también que la correlación entre la subpoblación ancestral nativo-americana y la europea es elevada y negativa, interpretándose que un individuo de este conjunto de parientes tiene mayor probabilidad de presentar una estructura poblacional nativo-americana cuanto menor sea la probabilidad de pertenencia al grupo europeo, como se retrata en los correlogramas de la Figura 2.5.

KING-robust y PC-Relate tienen la ventaja de considerar el efecto de la estratificación poblacional en la inferencia de parentesco con respecto a PLINK bajo el escenario de población mezclada. En tal situación, cuando dos individuos de un par presentan estructuras poblacionales distintas, KING-robust los representa con la estimación del coeficiente kinship con un sesgo negativo, indicando tal condición. Casualmente, se ha observado que dichos pares de individuos con este sesgo tienen mayor proporción de estructura poblacional nativo-americana, coincidiendo con las estimaciones de PLINK de los parientes no creíbles. Se podría presuponer que KING presenta a los parientes que son confusos en PLINK como individuos no emparentados asociados a dicha ancestralidad en KING-robust. Tales hipótesis se pueden mostrar en las Figuras B.1 del Apéndice y en 2.2.

En cuanto a los PCAs, son idénticos independientemente del método estimativo utilizado para todos los individuos, corroborando la unicidad de las componentes principales. La forma de 'ala delta' de las representaciones de las dos componentes principales se puede explicar a través de los elevados coeficientes de correlación observados en los tres métodos entre componentes y estructuras poblacio-

nales. En las Figuras 2.5,2.6 y 2.7 se aprecia que en vez de formar grupos de dispersión homogénea cada una de las componentes, están linealmente correlacionadas con las proporciones de los grupos ancestrales debido a la presencia de población mezclada. Bajo este escenario, no existe una clasificación definida y perfectamente delimitada de la pertenencia a los grupos ancestrales para cada individuo.

En base a los resultados de los modelos PLS y PLSGLR, se demuestra empíricamente que los modelos PLS, con el método estimativo de KING-robust, son los que mejor explican la variabilidad de las proporciones de ancestralidad. Además, a través de los modelos PLSGLR, KING-robust también muestra un AUC igual o superior que el resto de los métodos para los grupos ancestrales nativo-americano y europeo, fijando la probabilidad alta de pertenencia en el umbral mínimo del 70 %. La determinación de la complejidad de los modelos PLS demuestran las diferencias entre métodos, siendo PC-AiR el que menor componentes principales necesita para explicar la variabilidad de las proporciones de grupos ancestrales. Sin embargo, tanto en los modelos PLS como en los modelos PLSGLR, es el método con menor precisión para estimar o clasificar a un individuo dentro del grupo ancestral europeo, siendo preferible el método de KING-robust.

Se considera, en términos metodológicos, que el método o algoritmo más robusto para la inferencia del parentesco, bajo condiciones de estratificación poblacional y con población mezclada, sería PC-Relate utilizando PC-AiR. Esta recomendación está fundamentada en la consideración del PCA previa para realizar la inferencia de parentesco, incluyendo los análisis de las estructuras poblacionales. Por otro lado, aunque KING-robust presenta en general estimaciones más precisas que PC-Relate y PC-AiR, dichas diferencias no son demasiado amplias, siendo ambos métodos adecuados en términos de precisión, tal y como se demuestra en las Tablas 2.4 y 2.3. Finalmente, en PC-AiR, considerando la inferencia con PC-Relate, presenta modelos PLS con menor complejidad que el resto de los métodos, según la Tabla 2.2. Esto se debe a la consideración previa de la estratificación poblacional para los pares no emparentados en PC-AiR. Al aplicar el mismo PCA para todos los grupos ancestrales en dicho método, no se representan las diferencias entre grupos ancestrales en términos del número de componentes.

Una de las limitaciones de la memoria fue no detectar funciones similares a las de PLINK, que permitiesen realizar la fase completa de control de calidad ni en KING ni en PC-Relate. Sin embargo, no ha sido inconveniente debido a que el procedimiento sería idéntico o muy similar con cualquier otro software siguiendo el protocolo descrito (Anderson *et al.* 2010).

No ha sido posible determinar el PC-Relate sin la exclusión de SNPs de Alto Desequilibrio de Ligamiento, aunque en la práctica el objetivo de la fase de control de calidad es mostrar los datos después de aplicar tal procedimiento.

Todos los umbrales mencionados en la memoria, salvo los criterios de clasificación de la inferencia de parentesco de KING y PC-Relate, fueron seleccionados arbitrariamente y por criterio del investigador siguiendo las directrices de la memoria.

# Apéndice A

# Tablas complementarias

| GLOSARIO DE ABREVIATURAS Y NOTACIÓN TEÓRICA | |
|---|---|
| **Abreviatura o siglas** | **Descripción** |
| $p$,$p_m$ | Frecuencia alélica del alelo de referencia A en un locus dado. Frecuencia alélica para cada $m$ SNP. |
| $q$ | Frecuencia alélica del alelo que no es el de referencia $a$ en un *locus* dado. |
| Z | Variable que representa el número de copias idénticas por descendencia común IBD en los alelos de dos individuos (ej. un par de individuos con alelos AA-Aa), tomará tres valores: 0, 1 o 2 denotados como $z$. |
| I | Variable que representa el número de copias genotípicas idénticas por estado o IBS en los alelos de dos indivviduos (ej. un par de individuos con alelos AA-Aa), tomará valor 0, 1 o 2 denotados como $l$. |
| $i$ | Individuo $i$. |
| $j$ | Individuo $j$. |
| $\mathcal{M}$,$\lvert\mathcal{M}\rvert$, $\lvert\mathcal{M}^*\rvert$ | Conjunto de SNPs. Número de SNPs sin valores perdidos. Número de SNPs resultantes del control de calidad del GWAS. |
| $\hat{\pi}_{ij}$ | Proporción estimada de alelos IBD para un par de individuos $i$ y $j$. |
| $\phi_{ij}$ | Coeficiente kinship para cada par de individuos $i$ y $j$. Es la mitad de $\pi_{ij}$. |
| $\pi_{0ij}$,$\pi_{1ij}$,$\pi_{2ij}$ | Probabilidades de que los alelos de dos individuos contengan 0, 1 o 2 copias genotípicas IBD(ej. AA-Aa=$\pi_{1ij}$) |
| $N_{AA,aa}$ | Número total de SNPs cuando el número de copias genotípicas IBS es igual a cero. |

| | |
|---|---|
| $X_m^{(i)}$,$X_m^j$ | Variables categóricas que representan el número de copias respecto al alelo de referencia para el individuo $i$ en un $m$ SNP. Ídem para el individuo $j$ |
| $\lvert X_m^{(i)} - X_m^{(j)} \rvert$ | Distancias genéticas en base al número de copias genéticas entre dos individuos $i$ y $j$. |
| $H_{ij}$ | Suma de frecuencias genotípicas si el individuo es heterocigoto para un SNP. |
| $Q$ | Variable aleatoria que representa la frecuencia alélica en un SNP extraído al azar del conjunto de SNPs de un individuo. |
| $I_a$ | Variable indicadora para un individuo con frecuencia alélica $Q$ si es heterocigoto de un SNP escogido aleatoriamente. |
| $N_{Aa}$,$N_{Aa}^{(i)}$,$N_{Aa}^{(j)}$ | Número total de SNPs con alelos heterocigotos para todos los individuos. Ídem para individuos $i$ y $j$ respectivamente. |
| $Q_1$,$Q_2$ | Variables aleatorias de la frecuencia alélica en un SNP extraído al azar del conjunto de SNPs para un par de individuos $i$ y $j$. |
| $\hat{\Pi}$,$\hat{\Phi}$ | Matrices GRM para la estimación de la proporción de alelos IBD $\hat{\pi}_{ij}$ y del coeficiente kinship $\hat{\phi}_{ij}$ respectivamente. |
| $\mathcal{U}, \mathcal{R}, \mathcal{N}$ | Conjunto de individuos no emparentados, de parientes cercanos y del total de individuos de la población de estudio respectivamente. |
| $K$ | Conjunto de subpoblaciones ancestrales o estructuras poblacionales. |
| $\mathbf{\Theta}_K$ | Matriz de covarianzas para cada una de las subpoblaciones $K$ de dimensiones $K \times K$. |
| $\mathbf{a}_i$ | Vector columna de proporciones de ancestralidad de cada subpoblación $k \in K$. |
| $\mathbf{1}_{[X_m^{(i)}=1]}$ | Variable indicadora que toma valor 1 cuando $X_m^{(i)} = 1$. Ídem para el individuo $j$. |
| $\mathbf{Z}$ | Matriz de genotipos estandarizados de dimensiones $n \times \lvert\mathcal{M}^*\rvert$, cuyos valores son $z_{ij}$. Si se considera el subconjunto de individuos no emparentados, será $\mathbf{Z}_u$ de dimensiones $n_u \times \lvert\mathcal{M}^*\rvert$. |
| $n_u, n_r, n$ | Número de individuos no emparentados, de parientes cercanos y del conjunto total de individuos respectivamente. |
| $\mathbf{V}_u, \mathbf{L}_u, \mathbf{W}_u$ | Matriz de componentes principales de tamaño $n_u \times n_u$. Matriz diagonal de autovalores $n_u \times n_u$. Matriz de ponderación de los SNPs $\lvert\mathcal{M}^*\rvert \times n_u$ en $n_u$. Ídem para el conjunto total de individuos $n \in \mathcal{N}$. |

| | |
|---|---|
| $\mathbf{Q}_r$ | Matriz de componentes principales predictivas de tamaño $n_r \times n_u$ sobre los parientes cercanos $\mathcal{R}$. Permite predecir las componentes del subconjunto $\mathcal{R}$. |
| $\mathbf{\Gamma}$ | Matriz de componentes principales de tamaño $n \times n_u$ combinadas tanto en $\mathcal{R}$ como en $\mathcal{U}$. |
| $\hat{\psi}_{ij}$ | Coeficiente kinship estimado considerando una población de referencia común para todos los individuos en PC-Relate. Se utiliza como valor de la matriz GRM en PC-Relate. |
| $\theta_{ij}$ | Coeficiente de coancestralidad para cada par de individuos y para cada subpoblación siendo $\theta_{ij} = \mathbf{a}_i \Theta_K \mathbf{a}_j$. |
| $b_\psi(i,j)$ | Función de coancestralidad que determina los ancestros comunes de $i$ y $j$. |
| $\theta_k, \theta_{kk'}$ | Elementos de la diagonal y fuera de la diagonal de la matriz $\mathbf{\Theta}_K$ respectivamente. Se establece $\theta_k$ como una medida de la variabilidad $F_S T$ para un grupo ancestral $k$, bajo las condiciones del modelo de Balding-Nichols. |
| $\beta$ | Vector columna de regresores de tamaño $D$. |
| $\mu_{im}$ | Frecuencias alélicas estimadas de cada individuo y en cada SNP en base a los valores de la matriz de componentes principales $\mathbf{V}$. |
| $\kappa_{ij}$ | Coeficiente kinship utilizado para la estimación con PC-Relate. |
| $\varpi_{0ij}$, $\varpi_{1ij}$, $\varpi_{2ij}$ | Probabilidad de que dos individuos no contengan ninguna, una o dos copias genotípicas IBD en sus alelos respectivamente en PC-Relate. |
| $b_\kappa(i,j)$ | Sesgo asintótico del estimador del coeficiente kinship para PC-Relate para cada par de individuos. |
| $d_\kappa(i,j)$ | Función expresada de $\mathbf{a}_i$,$\mathbf{a}_j$ y $\mathbf{\Theta}_K$ en PC-Relate. |
| $\hat{\delta}_{ij}$ | Estimador de la probabilidad de que dos individuos contengan alelos IBD específico de PC-Relate. |
| $\hat{f}_i$ | Recuento de frecuencias alélicas bajo la asunción del Equilibrio de Hardy-Weinberg. |
| $\hat{\delta}_{ij}$ | Estimador de la probabilidad de que dos individuos contengan alelos IBD específico de PC-Relate. |
| $\mathbf{c}$ | Vector fila compuesto de coeficientes de regresión de las componentes principales $\mathbf{V}_d$ para el modelo PLS. |

| | |
|---|---|
| $\mathbf{R}^*$ | Matriz de coeficientes de predictores de cada componente principal $\mathbf{V}_d$ del modelo PLS. |
| $\mathbf{y}$ | Vector de la variable respuesta del modelo PLS. |
| $\alpha$ | Valor de la función link en los modelos PLSGLR tomando dos valores: la esperanza condicional de la respuesta $\mathbf{y}$ si su distribución es continua. -Si es discreta con dominio finito, será un vector de probabilidades. |
| $g$ | Función link del modelo PLSGLR. |
| $X, W$ | Variables aleatorias independientes para denotar el subconjunto de individuos sanos(negativos) y enfermos(positivos) de una población respectivamente. |
| $c$ | Punto de corte para la clasificar entre individuos positivos o negativos. |
| $s$ | Un SNP seleccionado arbitrariamente perteneciente al conjunto total de SNPs siendo $s \in \mathcal{M}$. |
| $F, G$ | Funciones de distribución de $X$ y $W$ |
| $t$ | Valor con rango entre cero y uno para las funciones de ditribución de $X$ y $W$ en las curvas ROC. |

Cuadro A.1: Glosario de notaciones teóricas. Fuente: Elaboración propia

| **GLOSARIO DE ABREVIATURAS Y SIGLAS** | |
|---|---|
| **Notación** | **Descripción** |
| SNP | Polimorfismo de nucleótidos únicos. Definido como variante o marcadores genéticos |
| GWAS | *Genome-Wide Association Studies*, traducido es sería: Estudios de Asociación del Genoma Completo |
| MAF | *Minor Allelic Frequency*, traducido es: Menor Frecuencia Alélica. |
| HWE | *Hardy-Weinberg Equilibrium*, traducido es: Equilibrio de Hardy-Weinberg |
| PCA | *Principal Component Analysis*, traducido es: Análisis de Componentes Principales |

| | |
|---|---|
| GRM | *Genetic Relationship Matrix*, traducido al español es: Matriz de Parentesco Genético |
| GSM | *Genetic Similarity Matrix*, traducido al español es: Matriz de Similitudes Genéticas |
| SVD | *Singular Value Decomposition*, traducido al español es: Descomposición Singular de Valores. |
| IID1/ID1 | Individuo $i$. |
| IID2/ID2 | Individuo $j$. |
| *LD pruning* | Exclusión de alto Desequilibrio de Ligamiento. |
| ADMIXED | Población mezclada. Pares de individuos pertenecientes a distintas estructuras poblacionales. |
| Z0 | Probabilidad de que dos individuos no tengan ninguna copia genotípica IBD en sus alelos, siendo $\hat{\pi}_{0ij}$. |
| Z1 | Probabilidad de que dos individuos tengan una copia genotípica IBD en sus alelos, siendo $\hat{\pi}_{1ij}$. |
| Z2 | Probabilidad de que dos individuos tengan dos copias genotípicas IBD en sus alelos, siendo $\hat{\pi}_{2ij}$. |
| PI_HAT | Proporción de valores genotípicos IBD en los alelos de dos individuos, siendo $\hat{\pi}_{ij}$. |
| Kinship | Coeficiente kinship para dos individuos, siendo $\hat{\phi}_{ij}$. |
| I0 | Probabilidad de que dos individuos no tengan ninguna copia IBS en sus alelos, siendo $P(I = 0)$. |
| MZ/Dup | Individuos gemelos o duplicados por errores de codificación. |
| PO | Individuos con relación de padres con hijos. |
| FS | Individuos hermanos completos. |
| 2nd | Individuos parientes de segundo grado. |
| 3rd | Individuos parientes de tercer grado. |
| UN | Individuos no emparentados. |
| EUR | Proporción de pertenencia a la estructura poblacional o grupo ancestral europeo. |

| | |
|---|---|
| AFR | Proporción de pertenencia a la estructura poblacional o grupo ancestral africano. |
| NAM | Proporción de pertenencia a la estructura poblacional o grupo ancestral nativo-americano. |
| PC1,PC2,... | Primeras componentes principales. |
| CV | Validación Cruzada. |
| PLS | *Partial Least Squares*, traducido es Mínimos Cuadrados Parciales. |
| PLSGLR | *Partial Least Squares Generalized Linear Regression*, traducido es Regresión Lineal por Mínimos Cuadrados Parciales Generalizados. |
| ROC | *Receiver Operating Characteristic.* |
| AUC | *Area Under the Curve*, traducido al español Área bajo la Curva. |

Cuadro A.2: Glosario de abreviaturas y siglas. Fuente: Elaboración propia

| | **Z0** | **Z1** | **Z2** | **IBD proportion ($\hat{\pi}$)** |
|---|---|---|---|---|
| **Monozygote twins/duplicate** | 0 | 0 | 1 | 1 |
| **Parent-offspring** | 0 | 1 | 0 | 1/2 |
| **Full sibling** | 1/4 | 1/2 | 1/4 | 1/2 |
| **Avuncular pair** | 1/2 | 1/2 | 0 | 1/4 |
| **Half sibling** | 1/2 | 1/2 | 0 | 1/4 |
| **Double first cousins** | 9/36 | 3/36 | 1/16 | 1/4 |
| **First cousins** | 3/4 | 1/4 | 0 | 1/8 |
| **Grandparent-grandchild** | 3/4 | 1/4 | 0 | 1/8 |
| **Unrelated** | 1 | 0 | 0 | 0 |

Cuadro A.3: Tabla de clasificación y valores teóricos de las probabilidades $P(Z = 0), P(Z = 1), P(Z = 2)$ y $\hat{\pi}_{ij}$. Fuente: Elaboración propia

| **Parentesco** | **Kinship teórico** | **Criterios de inferencia** | **I0 teórico** | **Criterios inferencia para I0** |
|---|---|---|---|---|
| **MZ/Dup** | $\frac{1}{2}$ | $\leq \frac{1}{2}$ | 0 | $< 0{,}1$ |
| **PO** | $\frac{1}{4}$ | $(\frac{1}{2^{\frac{5}{2}}}, \frac{1}{2^{\frac{5}{2}}})$ | 0 | $< 0{,}1$ |
| **FS** | $\frac{1}{4}$ | $(\frac{1}{2^{\frac{5}{2}}}, \frac{1}{2^{\frac{5}{2}}})$ | $\frac{1}{4}$ | $(0{,}1, 0{,}365)$ |
| **2nd** | $\frac{1}{8}$ | $(\frac{1}{2^{\frac{7}{2}}}, \frac{1}{2^{\frac{5}{2}}})$ | $\frac{1}{2}$ | $(0{,}365, 1 - \frac{1}{2^{3/2}})$ |
| **3rd** | $\frac{1}{16}$ | $(\frac{1}{2^{\frac{9}{2}}}, \frac{1}{2^{\frac{7}{2}}})$ | $\frac{3}{4}$ | $(1 - \frac{1}{2^{3/2}}, 1 - \frac{1}{2^{5/2}})$ |

Cuadro A.4: Tabla de clasificación del parentesco en KING. Fuente: Elaboración propia

| **NÚCLEOS DE PARIENTES DE PRIMER GRADO EN PLINK** | | |
|---|---|---|
| | **Sin LD pruning** | **Con LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** | **IID** |
| N1_FUCS | FUCS_0412<br>FUCS_0416<br>FUCS_0182 | FUCS_0412<br>FUCS_0416<br>FUCS_0182 |
| N2_PARG1 | | PARG_0169<br>PARG_0176<br>PARG_0182 |
| N3_PARG2 | PARG_0190<br>PARG_0203<br>PARG_0192<br>PARG_0206 | PARG_0190<br>PARG_0203<br>PARG_0192<br>PARG_0206 |
| N4_TECM1 | | TECM_0561<br>TECM_0589<br>TECM_0614 |
| N5_TECM2 | TECM_0578<br>TECM_0587<br>TECM_0596 | TECM_0578<br>TECM_0587<br>TECM_0596 |
| N6_TECM3 | TECM_0734<br>TECM_0732<br>TECM_0735 | |
| N7_UFPA | UFPA_0281<br>UFPA_0236<br>UFPA_0206 | UFPA_0281<br>UFPA_0236<br>UFPA_0206 |
| N8_UFPE | | UFPE_0248<br>UFPE_0259<br>UFPE_0260 |

Cuadro A.5: Tabla de núcleos de parentesco de primer grado en PLINK. Fuente: Elaboración propia

| **NÚCLEOS DE PARIENTES DE SEGUNDO GRADO EN PLINK** | |
|---|---|
| | **Con/Sin LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** |
| N1_HUGM/HULP | HUGM_0769 |
| | HUGM_1120 |
| | HULP_1395 |
| N2_PARG | PARG_0182 |
| | PARG_0197 |
| | PARG_0205 |
| N3_TECM | TECM_0516 |
| | TECM_0517 |
| | TECM_0744 |
| | TECM_0712 |
| | TECM_0731 |
| | TECM_0745 |

Cuadro A.6: Tabla de bloques de parientes de segundo grado con PLINK. Fuente: Elaboración propia

| **NÚCLEOS DE PARIENTES DE PRIMER GRADO EN KING** | |
|---|---|
| | **Sin/Con LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** |
| N1_FUCS | FUCS_0412 |
| | FUCS_0416 |
| | FUCS_0182 |
| N2_PARG1 | PARG_0169 |
| | PARG_0176 |
| | PARG_0182 |
| N3_PARG2 | PARG_0190 |
| | PARG_0192 |
| | PARG_0203 |
| | PARG_0206 |
| N4_TECM1 | TECM_0561 |
| | TECM_0589 |
| | TECM_0614 |
| N5_TECM2 | TECM_0578 |
| | TECM_0587 |
| | TECM_0596 |
| N6_TECM3 | TECM_0734 |
| | TECM_0735 |
| | TECM_0732 |
| N7_UFPA | UFPA_0236 |
| | UFPA_0206 |
| | UFPA_0281 |
| N7_UFPE | UFPE_0248 |
| | UFPE_0259 |
| | UFPE_0260 |

| **NÚCLEOS DE PARIENTES DE SEGUNDO GRADO EN KING** | | |
|---|---|---|
| | **Sin LD pruning** | **Con LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** | **IID** |
| N1_HUGM/HULP | HUGM_0769 | HUGM_0769 |
| | HUGM_1120 | HUGM_1120 |
| | HULP_1395 | HULP_1395 |
| N2_PARG | PARG_0182 | PARG_0182 |
| | PARG_0197 | PARG_0197 |
| | PARG_0205 | PARG_0205 |
| | PARG_0166 | PARG_0166 |
| N3_TECM | TECM_0517 | TECM_0517 |
| | TECM_0516 | TECM_0516 |
| | TECM_0712 | TECM_0712 |
| | TECM_0731 | TECM_0731 |
| | TECM_0744 | TECM_0744 |
| | TECM_0745 | |

Cuadro A.8: Bloques de parientes de segundo grado con KING-robust. Fuente: Elaboración propia

| **NÚCLEOS DE PARIENTES DE PRIMER GRADO EN PC-RELATE** | |
|---|---|
| | **Con LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** |
| N1_FUCS | FUCS_0412 |
| | FUCS_0416 |
| | FUCS_0182 |
| N2_PARG1 | PARG_0169 |
| | PARG_0176 |
| | PARG_0182 |
| N3_PARG2 | PARG_0190 |
| | PARG_0203 |
| | PARG_0192 |
| | PARG_0206 |
| N4_TECM1 | TECM_0561 |
| | TECM_0589 |
| | TECM_0614 |
| N5_TECM2 | TECM_0578 |
| | TECM_0587 |
| | TECM_0596 |
| N6_TECM3 | TECM_0732 |
| | TECM_0734 |
| | TECM_0735 |
| N7_UFPA | UFPA_0281 |
| | UFPA_0236 |
| | UFPA_0206 |
| N8_UFPE | UFPE_0248 |
| | UFPE_0259 |
| | UFPE_0260 |

Cuadro A.9: Bloques de parientes de primer grado con PC-Relate.Fuente: Elaboración propia

| **NÚCLEOS DE PARIENTES DE SEGUNDO GRADO EN PC-RELATE** | |
|---|---|
| | **Con LD pruning** |
| **NÚCLEOS DE PARIENTES** | **IID** |
| N1_HUGM/HULP | HUGM_0769 |
| | HUGM_1120 |
| | HULP_1395 |
| N2_PARG | PARG_0182 |
| | PARG_0197 |
| | PARG_0205 |
| N3_TECM | TECM_0517 |
| | TECM_0516 |
| | TECM_0712 |
| | TECM_0731 |
| | TECM_0744 |
| | TECM_0745 |
| N4_UFPA/UFRN/UFPE | UFPA_0296 |
| | UFPE_0193 |
| | UFRN_0362 |

Cuadro A.10: Fuente: Bloques de parientes de segundo grado con PC-Relate. Fuente: Elaboración propia

# Apéndice B

# Figuras complementarias

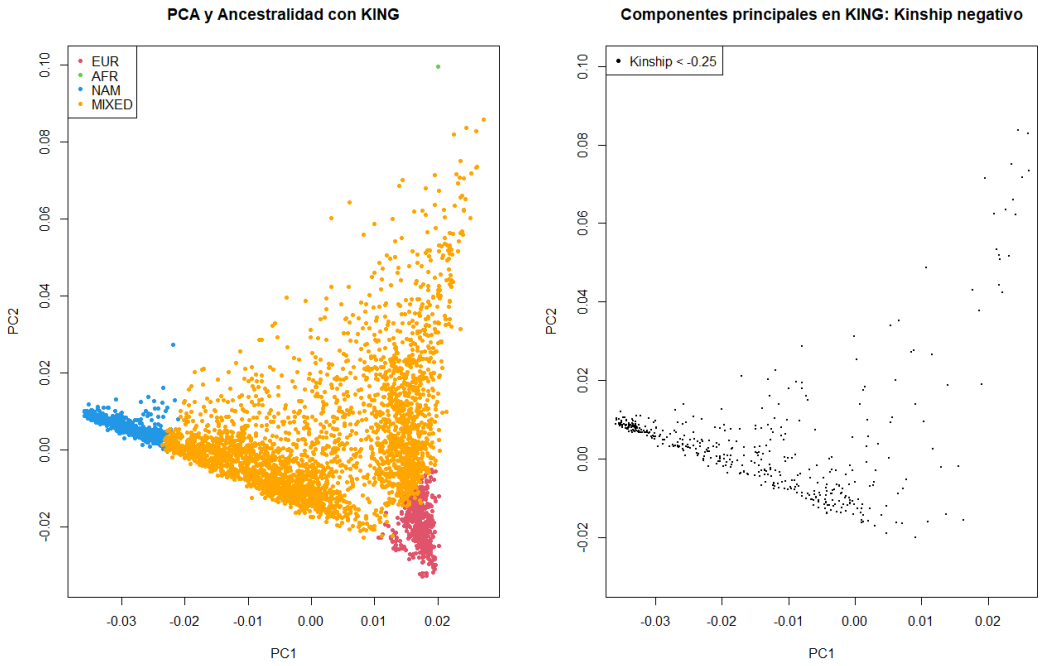


Figura B.1: PCA y estimaciones del coeficiente kinship negativo con KING-robust. Fuente: Elaboración propia

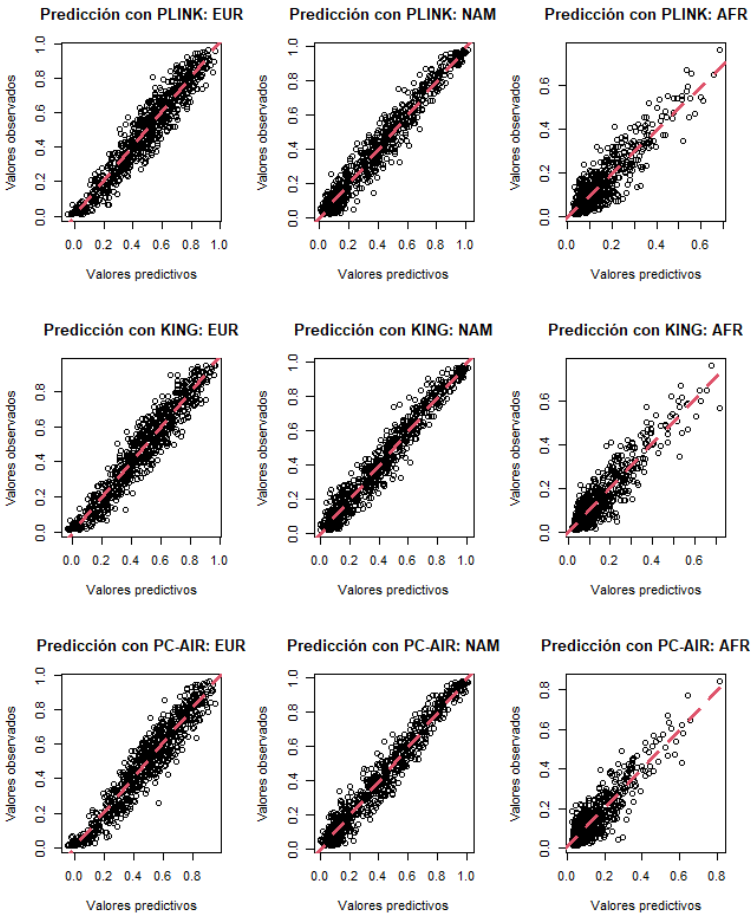


Figura B.2: Gráficos de predicción PLS: Comparación de valores observados vs. valores predictivos. Fuente: Elaboración propia

# Bibliografía


[Alexander, *et al.* 2009] Alexander, D. H., Novembre, J., y Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*(9), 1655-1664. https://doi.org/10.1101/gr.094052.109

[Anderson *et al.* 2010] Anderson, C. S., Pettersson, F., Clarke, G.M., Cardon, L.R., Morris, A.P., y Zondervan, K.T.(2010) Data quality control in genetic case-control association studies *Nature Protocols* 5, 1564-1573. https://doi.org/10.1038/nprot.2010.116. Accedido 21 de diciembre de 2022.

[Raj *et al.* 2013] Raj, A., Stephens, M., y Pritchard, J. K. (2013). Variational Inference of Population Structure in Large SNP Datasets. *bioRxiv (Cold Spring Harbor Laboratory).* https://doi.org/10.1101/001073

[Bertrand *et al.* 2014] Bertrand, F., Magnanensi, J., Meyer, N. y Maumy-Bertrand, M. (2014). plsRglm: Algorithmic Insights and Applications. *R package version 1.5.1.* http://127.0.0.1:28859/library/plsRglm/doc/plsRglm.pdf. Accedido el 3 de junio de 2023.

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

[Conomos *et al.* 2015] Conomos, M. P., Miller, M. B., y Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology, 39*(4), 276-293. https://doi.org/10.1002/gepi.21896

[Conomos *et al.* 2016] Conomos, M. P., Reiner, A. P., Weir, B. S., y Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics, 98*(1), 127-148. https://doi.org/10.1016/j.ajhg.2015.11.022

[Cruz *et al.* 2022] Cruz, R., Almeida, S. M. Z., De Heredia, M. L., Quintela, I., Ceballos, F., Pita, G., Lorenzo-Salazar, J. M., González-Montelongo, R., Gago-Domínguez, M., Porras, M. R., Castaño, J. A. T., Nevado, J., Aguado, J. M., Aguilar, C. A., Aguilera-Albesa, S., Almadana, V., Almoguera, B., Alvarez, N. B., Andreu-Bernabeu, A., . . . y Carracedo, A. (2022). Novel genes and sex differences in COVID-19 severity. *Human Molecular Genetics, 31*(22), 3789-3806. https://doi.org/10.1093/hmg/ddac132

[Dong *et al.* 2020] Dong, E., Du, H., y Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5), 533-534. *The Lancet. Infectious Diseases, 20*(5), 533-534. https://doi.org/10.1016/S1473-3099(20)30120-1

[Dong *et al.* 2022] Dong, E., Ratcliff, J., Goyea, T. D., Katz, A., Lau, R., Ng, T. K., ... y Gardner, L. M. (2022). The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned. *The Lancet.Infectious Diseases, 22*(12), 370-376. https://doi.org/10.1016/S1473-3099(22)00434-0

[Dagnino 2014] Dagnino, J. (2014). Coeficiente de correlación lineal de pearson. *Revista chilena de anestesia, 43*(2), 150-153. https://doi.org/10.25237/revchilanestv43n02.15

[Galinsky *et al.* 2016] Galinsky, K., Bhatia, G., Loh, P., Georgiev, S., Mukherjee, S., Price, A. L., y Price, A. L. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics, 98*(3), 456-472. https://doi.org/10.1016/j.ajhg.2015.12.022

[Gogarten *et al.* 2019] Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K., y Conomos, M. P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics, 35*(24), 5346-5348. https://doi.org/10.1093/bioinformatics/btz567

[Golub y Van Loan 2013] Golub, G. H., y Van Loan, C. F. (2013). *Matrix computations*. Johns Hopkins University Press. Recuperado de https://books.google.es/books?id=5U-l8U3P-VUC

[Gonçalves *et.al* 2014] Gonçalves, L., Subtil, A., Oliveira, M. R., y de Zea Bermudez, P. (2014). ROC curve estimation: An overview. *REVSTAT-Statistical journal, 12*(1), 1-20. https://doi.org/10.57805/revstat.v12i1.141

[Halko, Shkolnisky *et al.* 2011] Halko, N., Martinsson, P., Shkolnisky, Y., y Tygert, M. (2011). An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing, 33*(5), 2580-2594. https://doi.org/10.1137/100804139

[Halko, Martinsson *et al.* 2011] Halko, N., Martinsson, P. G., y Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review, 53*(2), 217-288. https://doi.org/10.1137/090771806

[James *et al.* 2021] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *Statistical Learning. In: An Introduction to Statistical Learning.*(pp. 15-57). Springer Texts in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-0716-1418-1_2.

[Kalmes y Huret 2001] Kalmes, R., y Huret, J. L. (2001). Hardy-weinberg model. *Atlas of Genetics and Cytogenetics in Oncology and Haematology, 5*(2), 156-163. http://atlasgeneticsoncology.org/teaching/30100/modelo-de-hardy-weinberg. Accedido el 24 de marzo de 2023.

[Manichaikul *et al.* 2010] Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., y Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics, 26*(22), 2867-2873. https://doi.org/10.1093/bioinformatics/btq559.

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

[Meyer *et al.* 2010] Meyer, N., Maumy-Bertrand, M., y Bertrand, F. (2010). Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives: application aux données d'allélotypage. *Journal de la Société Française de Statistique, 151*(2), 1-18. http://www.numdam.org/item/JSFS_2010__151_2_1_0/

[Mevik y Wehrens 2007] Mevik, B., y Wehrens, R. (2007). ThepIsPackage: Principal Component and Partial Least Squares Regression inR. *Journal of Statistical Software, 18*(2). https://doi.org/10.18637/jss.v018.i02

[Niemi 2022] Niemi, M., Daly, M. J., y Ingelsson, E. (2022). The human genetic epidemiology of COVID-19. *Nature Reviews Genetics, 23*(9), 533-546. https://doi.org/10.1038/s41576-022-00478-5

[Nordborg y Tavaré 2002] Nordborg, M., y Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics, 18*(2), 83-90. https://doi.org/10.1016/s0168-9525(02)02557-x

[Patterson *et al.* 2006] Price, A. L., Price, A. L., y Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics, 2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

[Price *et al.* 2006] Price, A. L., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., y Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909. https://doi.org/10.1038/ng1847

[Purcell *et. al* 2007] Purcell, S., Neale, B. M., Todd-Brown, K., Thomas, L. L., Ferreira, M., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., y Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, 81(3), 559-575. https://doi.org/10.1086/519795

[Ramstetter *et al.* 2017] Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Mezey, J. G., y Williams, A. W. (2017). Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives. *Genetics, 207*(1), 75-82. https://doi.org/10.1534/genetics.117.1122

[Revelle y Condon 2018] Revelle, W., y Condon, D. M. (2019). Reliability from $\alpha$ to $\omega$: A tutorial. *Psychological assessment, 31*(12), 1395. https://psycnet.apa.org/doi/10.1037/pas0000754

[Rokhlin *et al.* 2010] Rokhlin, V., Szlam, A., y Tygert, M. (2010). A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications, 31*(3), 1100-1124. https://doi.org/10.1137/080736417

[Solari 2004] Solari, A. J. (2004). *Genética humana: Fundamentos y aplicaciones en medicina.* Ed. Médica Panamericana. Recuperado de https://books.google.es/books?id=e-slX7S1KdsC

[Spanish COalition to Unlock Research on host GEnetics on COVID-19 (SCOURGE) s.f.] Spanish COalition to Unlock Research on host GEnetics on COVID-19 (s.f.). CIBER. Recuperado el 01/05/2023, de https://www.scourge-covid.org

[Uffelmann *et al.* 2021] Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., ... y Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers, 1*(1), 59.

[Vinzi *et al.* 2010] Vinzi, V. E., Chin, W. W., Henseler, J., y Wang, H. (2010). *Handbook of partial least squares* (Vol. 201, No. 0). Berlin: Springer.

[Wright 1949] Wright, S. (1949). The genetical structure of populations. *Annals of Eugenics, 15*(1), 323-354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

[Wold *et al.* 1983] Wold, S., Martens, H., Wold, H. (1983). Wold, S., Martens, H., y Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. En *Lecture Notes in Mathematics* (pp. 286-293). Springer Nature. https://doi.org/10.1007/bfb0062108

[Yang *et al.* 2011] Yang, J., Lee, S. H., Goddard, M. E. y Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics, 88*(1), 76-82. https://doi.org/10.1016/j.ajhg.2010.11.011

[Zheng *et al.* 2012] Zheng, X., Levine, D. K., Shen, J., Gogarten, S. M., Laurie, C. C., y Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*(24), 3326-3328. https://doi.org/10.1093/bioinformatics/bts606

[Zhu *et al.* 2008] Zhu, X., Li, S., Cooper, R. S., y Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *The American Journal of Human Genetics, 82*(2), 352-365. https://doi.org/10.1016/j.ajhg.2007.10.009