Comparación de la imputación genética en cohortes individuales *versus* imputación conjunta: Factores implicados y propuesta de actuación

Pedro Sánchez García
Facultad de Informática
Universidade da Coruña
A Coruña, España
p.sanchezg@udc.es

Raquel Cruz Guerrero
CiMUS
Universidade de Santiago
Santiago de Compostela, España
raquel.cruz@usc.es

Silvia Diz de Almeida
CiMUS
Universidade de Santiago
Santiago de Compostela, España
silvia.diz.dealmeida@usc.es

Jorge González Domínguez
CITIC
Universidade da Coruña
A Coruña, España
jorge.gonzalezd@udc.es

***Resumen*—Los estudios de asociación del genoma completo (GWAS, por sus siglas en inglés) han demostrado ser una estrategia eficaz para identificar el enlace entre marcadores genéticos y fenotipos de interés, como enfermedades complejas, permitiendo comprender los mecanismos moleculares alterados y determinar nuevas dianas farmacológicas. En los GWAS, la fase de imputación es clave en su preprocesado y control de calidad, basándose en paneles de referencia disponibles para inferir marcadores ausentes en el *array* de genotipado e incrementar la resolución en las regiones asociadas. Si bien la imputación ha proporcionado un notable avance en los últimos años, esta puede implicar sesgos por diversos factores, entre los que destaca la ancestralidad de los individuos. En el caso de la población latina, su particular complejidad genética representa un desafío en proyectos de investigación internacionales, siendo necesaria la valoración de ajustes posteriores a la imputación. Este trabajo propone explorar, sobre un grupo de individuos latinos, la influencia por su imputación individual y en conjunto con otra cohorte geográficamente distinta. Los resultados muestran una notable diferencia en la estimación de las frecuencias alélicas para numerosos marcadores, siendo la calidad de imputación alcanzada, factor causal del impacto. De esta manera, según las características del proyecto, la aplicación de un filtro adecuado para la calidad de imputación puede definir un criterio eficiente en la robustez de los análisis.**

***Palabras clave*—GWAS, imputación, panel de referencia, control de calidad, latinos, *arrays* de genotipado**

I. Introducción

Los avances en las tecnologías de secuenciación y genotipado de alto rendimiento han permitido ampliar conocimientos en la base genética de numerosas enfermedades complejas, distinguiendo genes y regiones cromosómicas específicas que juegan un papel clave [1]. En los últimos años, mediante los estudios de asociación del genoma completo (GWAS, por sus siglas en inglés), se han publicado más de 6.000 proyectos donde se han identificado aproximadamente 128.550 asociaciones entre variantes genéticas y enfermedades, abriendo nuevas oportunidades para su prevención y tratamiento [2].

El fundamento de los GWAS se basa en la información sobre marcadores genéticos de dos grupos, que comprenden individuos portadores de una enfermedad o carácter en cuestión (casos) y otros que no lo son (controles). De esta forma, se estiman las correlaciones entre el estado de la enfermedad (fenotipo) y los marcadores genéticos recogidos. Estos últimos consisten en polimorfismos de un único nucleótido (SNPs, por sus siglas en inglés), es decir, variaciones en una posición específica del genoma (A, C, G o T). Cada SNP suele existir en dos formas diferentes (por ejemplo, A -alelo menor- frente a T -alelo alternativo-), que reciben el nombre de alelos y cuya frecuencia puede variar en función del fenotipo analizado [3]. Posteriormente, se lleva a cabo un test estadístico para indicar si cada una de las correlaciones estimadas resulta estadísticamente significativa, cuyo poder depende principalmente del tamaño muestral y de las frecuencias de los marcadores [4]. Los resultados se representan gráficamente mediante *Manhattan plots*, enfrentando los marcadores genéticos con la significación estadística, en forma del logaritmo negativo del *p*-valor obtenido ($-\log_{10}$(*p*-valor)) (Figura 1A). Cabe destacar que, debido a los costes asociados en esta técnica, se genotipa un conjunto representativo de cientos de miles a millones de marcadores, los cuales se disponen en *arrays* de genotipado desarrollados específicamente por las casas comerciales [5, 6].

Si bien los GWAS son cada vez más populares en la comunidad científica, estos poseen una metodología compleja, con intersección de conocimientos en genética, estadística y bioinformática. Su diseño experimental comprende, en primer lugar, un estricto control de calidad, donde se eliminan muestras duplicadas, inconsistencias entre el sexo asignado y el genético de los individuos y muestras con exceso o déficit de heterocigosidad (la portación de dos alelos diferentes en un SNP concreto) que indican baja calidad de muestras y posible endogamia, respectivamente. Para los marcadores, debido a un poder estadístico razonable en la detección de asociaciones, se excluyen aquellos casos con valores inferiores al 1% en la frecuencia del alelo menor (MAF, por sus siglas en inglés), que se trata de la frecuencia del alelo menos frecuente en una posición determinada. Además, se garantiza que los marcadores cumplan el equilibrio Hardy-Weinberg (HW). Bajo este supuesto, las frecuencias alélicas y genotípicas son constantes a lo largo de las generaciones, de tal forma que las observadas en estas últimas no deberían ser diferentes de las esperadas. En caso contrario, se atribuye a errores de genotipado que afectan a los GWAS [7].

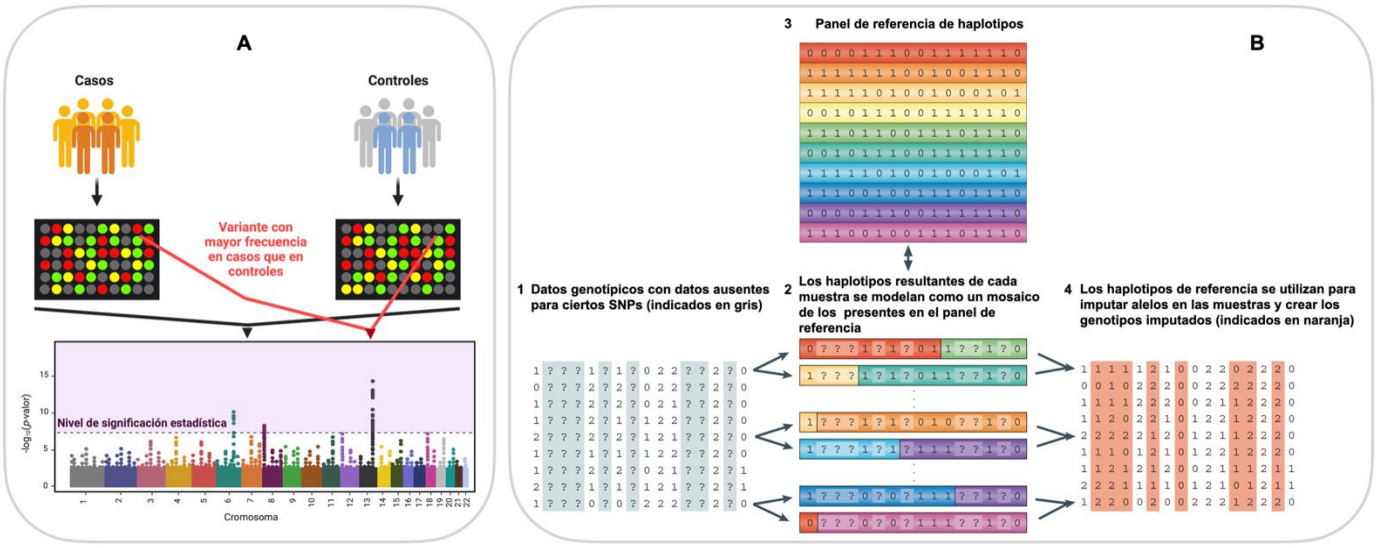


Figura 1. (A): Fases principales de un GWAS, que comprenden el diseño de grupos caso-control de pacientes, el genotipado de los marcadores presentes en los *arrays* y la comparación en las frecuencias alélicas para esos marcadores entre los grupos, donde el *Manhattan plot* ilustra las diferencias estadísticamente significativas, que indican una asociación entre el marcador y la enfermedad analizada. (B): Fases principales de la imputación genética para muestras de individuos no emparentados. Los datos crudos consisten en un conjunto de marcadores genotipados, que presenta un amplio número de marcadores sin genotipar (indicados en gris) (1). Se destacan los haplotipos de tres individuos concretos (2). Los haplotipos del panel de referencia son la base para efectuar la comparación con los haplotipos de los individuos (3). Los genotipos que faltan en la muestra de estudio se imputan utilizando los haplotipos coincidentes en el panel de referencia (4). Modificado de [10].

Como se ha mencionado anteriormente, los *arrays* de genotipado que ofrecen las casas comerciales poseen un conjunto de marcadores específicos. De esta forma, una vez realizado el control de calidad, la siguiente fase en los GWAS es la imputación, que, utilizando un panel de referencia, permite inferir el genotipo para marcadores ausentes en el *array*. En términos generales, esta fase mejora la precisión e interpretación de los análisis, pues permite la realización de los test estadísticos con un gran número de marcadores, la combinación de experimentos realizados en diferentes *arrays* y la replicación de resultados previos entre estos [8]. La imputación se basa en el modelo oculto de Markov (HMM, por sus siglas en inglés) [9], que usa la información del panel de referencia y el desequilibrio de ligamiento (LD, por sus siglas en inglés), indicando que dos marcadores próximos en el mismo cromosoma permanecen físicamente ligados a través de las generaciones. Las combinaciones de alelos que cumplen esta última condición se denominan haplotipos. Mediante esta clase de modelo estadístico ajustado, la distribución de las probabilidades calculadas para los haplotipos perdidos en cada iteración está condicionada por las observaciones del panel de referencia. El procedimiento general de la imputación se muestra en la Figura 1B: primero, se comparan los haplotipos para cada individuo en el estudio con los haplotipos del panel de referencia y posteriormente, se imputan los genotipos faltantes utilizando los haplotipos coincidentes estimados de la fase anterior [10].

Hasta la actualidad, se han desarrollado diferentes algoritmos de imputación, siendo los más relevantes BIMBAM [11], MaCH [12], IMPUTE [13], BEAGLE [14] y Minimac [15], con diferentes particularidades, mejoras en eficiencia computacional y limitaciones. No obstante, la base para una precisión adecuada en la imputación es el panel de referencia, donde el servidor TOPMed [16], que utiliza Minimac, posee una reciente representación de más de 50.000 individuos de 26 cohortes y de diversos orígenes ancestrales, con un mayor número de marcadores frente a paneles como los del proyecto 1000 Genomas o el proyecto HapMap [17]. A pesar de las posibilidades y progresos existentes, la imputación se encuentra limitada por varios factores como el panel de referencia, densidad de marcadores en el *array* de genotipado, MAF del marcador imputado, la ancestralidad y el tamaño muestral, que pueden implicar falsos positivos en las posteriores asociaciones determinadas [18]. En el caso de la ancestralidad, es importante destacar que la mayoría de los estudios se han llevado a cabo en individuos de ancestralidad europea (91%), siendo muy limitados en poblaciones de ancestralidad latina y, por tanto, reduciendo su replicabilidad [19]. Este patrón se debe a una mayor complejidad en su estructura genética y desequilibrio de ligamiento. Por todo ello, es crucial analizar rigurosamente el diseño experimental en las colaboraciones internacionales con ancestralidad mixta y establecer posibles criterios para el control de calidad posterior a la imputación, que permitan ampliar sustancialmente los conocimientos biológicos y la asistencia sanitaria orientada a la medicina de precisión [19, 20].

II. Objetivos

El objetivo principal de este trabajo es caracterizar las diferencias en la imputación realizada en una subcohorte de pacientes latinos en dos situaciones: cuando se imputa de forma individual o en conjunto con el total de la cohorte. Para ello, se establecieron los siguientes objetivos concretos:

- Realizar un análisis exploratorio de factores que potencialmente influyen en la diferencia en estima de frecuencias alélicas: distribución y densidad de marcadores en el *array* de genotipado.
- Determinar un criterio de filtrado o control de calidad posterior a la imputación para obtener unos resultados fiables en un estudio típico.

III. Materiales y métodos

A. Cohortes de estudio

Para llevar a cabo los análisis de este trabajo, se utilizaron dos cohortes del consorcio SCOURGE *(Spanish Coalition to Unlock Research on Host Genetics)*, conformadas por 1.082 y 3.898 pacientes de ancestralidad mixta reclutados en hospitales de diferentes áreas geográficas de España (22 centros) y en conjunto con América Latina (5 países, 13 centros), respectivamente. La primera cohorte estaba compuesta por 533 hombres y 549 mujeres, con 293 controles y 789 casos y una edad media de 49,8 años. Esta cohorte comprende los individuos de ancestralidad mixta identificados en la primera fase del proyecto, con muestras reclutadas en España. En la segunda fase del proyecto SCOURGE, cuyo objetivo era analizar la base genética de la COVID-19 en población latina, la cohorte anterior fue combinada con el total de muestras reclutadas en América Latina. Esta segunda cohorte comprendió 1.806 hombres y 2.092 mujeres, distinguiendo 2.322 controles y 1.576 casos, con una edad media de 46,2 años.

La subcohorte sobre la que se centra el presente estudio implicó 1.071 individuos comunes entre las cohortes originales tras el control de calidad y la imputación, abarcando 527 hombres y 544 mujeres, 290 controles y 781 casos, con una edad media de 49,8 años.

B. Genotipado y control de calidad

Las muestras de pacientes se genotiparon mediante tecnología *Axiom Spain Biobank Array* de *Thermo Fisher Scientific*, en el nodo de Santiago de Compostela del CeGen (Centro Nacional de Genotipado). Este *array* contiene 757.836 marcadores, que incluyen variantes raras seleccionadas en la población de España.

Se llevó a cabo un riguroso protocolo para el control de calidad de los datos genéticos (tanto en muestras como en marcadores) en ambas cohortes utilizando PLINK 1.9 [21] y scripts propios de R (https://www.R-project.org/). Se descartaron muestras duplicadas, individuos con discrepancias en sexo o emparentados, muestras con déficit o exceso de heterocigosidad o aquellas con una tasa de genotipado inferior al 98%. Por otra parte, en lo que respecta a los marcadores, se excluyeron aquellas variantes con frecuencia del alelo menor (MAF) inferior al 1%, tasa de genotipado inferior al 98% o que se desviasen significativamente del equilibrio Hardy-Weinberg (HW). Las salidas para cada cohorte consistieron en archivos VCF *(Variant Call Format)* con todas las variantes genómicas filtradas y anotadas.

Detalles adicionales sobre el proceso de control de calidad se describen en el estudio llevado a cabo por Cruz et al. [22].

C. Imputación

Sobre los datos genéticos revisados y filtrados en cada cohorte, se realizó la imputación con el servidor TOPMed [16], empleando la versión r2 del panel de referencia con 308.107.085 variantes genéticas distribuidas y basándose en el ensamblaje del genoma humano GRCh38 (hg38). De esta forma, el archivo VCF imputado, contuvo 39.612.242 marcadores en la cohorte de individuos latinos en España, mientras que, para la cohorte en conjunto con individuos reclutados en América Latina, se alcanzaron 71.367.877 marcadores totales.

D. Comparativa de las MAFs entre tipos de imputación

En cada cromosoma, se evaluó la posible influencia de la imputación individual o en conjunto de una subcohorte para las MAFs de los marcadores. Para ello, se partió de los archivos VCF imputados con TOPMed.

Se generaron los siguientes archivos por PLINK 1.9, mediante los parámetros *--vcf* y *--make-bed*, seleccionando el archivo VCF de entrada y convirtiéndolo en formato BED *(Browser Extensible Data)*, respectivamente:

- *.bed: Fichero binario que contiene los identificadores de pacientes y sus genotipos.
- *.bim: Fichero de texto con amplia información sobre los marcadores, recogida en los siguientes campos: nombre o código del cromosoma, identificador del marcador, su posición en centimorgans, coordenadas en pares de bases, alelo menor y alelo alternativo.
- *.fam: Fichero de texto que presenta la siguiente información de pacientes: identificador familiar, identificador dentro de la familia, identificador del padre, identificador de la madre, sexo (‘1’=varón, ‘2’=mujer, ‘0’=desconocido) y fenotipo (‘1’=control, ‘2’=caso, ‘-9/0/no numérico’=desconocido).

A continuación, utilizando un script personalizado en R, en primer lugar, se realizó la combinación o *merge* entre los archivos *.fam de las cohortes, exportando un listado de la subcohorte. Posteriormente, especificando los ficheros *.bed, *.bim y *.fam como entrada en PLINK 1.9 y manteniendo los individuos de la lista generada, se calcularon las MAFs de los marcadores mediante el parámetro *--freq*. En este caso, los ficheros *.frq obtenidos para cada cohorte, poseen los siguientes campos: nombre o código del cromosoma, identificador del marcador, alelo menor, alelo alternativo, MAF y el número de pacientes con ese marcador.

En el mismo script de R, estos ficheros *.frq se combinaron, representándose gráficamente la correlación existente para las MAFs entre tipos de imputación. Por último, del resultado de la combinación, se filtró y exportó en un archivo *.csv, el subconjunto de marcadores donde la diferencia de MAFs entre los tipos de imputación era superior a 0,05, es decir, atípicos.

E. Análisis de la influencia en la estima de MAFs por densidad y distribución de marcadores

Se verificó, de forma independiente en ambas cohortes, si se producían variaciones destacadas en términos de densidad y distribución de los marcadores genotipados tras el proceso de control de calidad que pudiesen explicar las diferencias existentes para las MAFs entre los tipos de imputación.

Los archivos *.bim previos a la imputación (marcadores genotipados) fueron combinados en R a través de los marcadores comunes. A continuación, en esos archivos, se plantearon ventanas de 500.000 pares de bases, determinando para cada una, el número de marcadores y el porcentaje de estos sobre el total del cromosoma.

Por otro lado, se combinaron los archivos *.bim posteriores a la imputación, con los archivos *.frq generados anteriormente, dividiendo los cromosomas en ventanas del mismo tamaño, para las cuales se calculó la diferencia promedio en MAFs para marcadores entre las cohortes.

A partir de los datos de densidad y diferencia promedio en MAFs obtenidos en los análisis anteriores, se computó para cada ventana, una medida de la diferencia en número relativo de marcadores entre cohortes. Por último, estas medidas se representaron gráficamente, centrándonos en las ventanas que contenían los marcadores con diferencias en MAFs atípicas, recogidos en el archivo *.csv logrado anteriormente. De esta forma, se evaluó la posible correspondencia entre la diferencia atípica en MAF y dicha medida calculada para esos marcadores concretos.

F. Relevancia de la calidad de imputación en la estima de MAFs

Para este análisis se partió de los ficheros *.info obtenidos en la imputación. Los ficheros *.info contienen las frecuencias alélicas y calidad de imputación de los marcadores (r^2) como información adicional, distinguiéndose los siguientes campos: nombre o código del cromosoma, identificador del marcador, alelo menor, alelo alternativo, MAF, frecuencia del alelo alternativo y calidad asociada al marcador.

Se combinaron los archivos *.info de las cohortes, calculando la diferencia existente en la calidad de imputación para los marcadores entre estas. Este *data frame* resultante se unió con el fichero *.csv obtenido anteriormente, que contiene el subconjunto de marcadores atípicos en términos de MAFs. De esta forma, para estos marcadores concretos, se representó gráficamente la diferencia en la calidad de imputación entre los tipos de imputación, así como una comparación de la tendencia en calidad de imputación para cada tipo de imputación.

Además, esta exploración se amplió al cromosoma completo. Por tanto, estableciendo ventanas de 500.000 pares de bases sobre la combinación de los *.info, se halló la calidad de imputación y el promedio de la diferencia de esta entre tipos, para determinar la tendencia en cada una de las ventanas.

G. Filtrado de la calidad de imputación

Para comprobar las implicaciones de la selección de aquellos marcadores que cumpliesen una calidad de imputación superior a un umbral, se procedió a su aplicación para los cromosomas que mostraron marcadores atípicos.

Usando los ficheros *.info de cada imputación por separado, se estableció un filtro de 0,8 en la calidad de imputación en base a las observaciones de la sección anterior. Se exportó una lista con los identificadores de aquellos marcadores que superasen este umbral. Los archivos *.frq fueron combinados con las correspondientes listas generadas, conformando la entrada para los análisis posteriores, siguiendo las fases descritas anteriormente en la sección III-D.

H. Entorno de ejecución

Los análisis desarrollados en el estudio comprendieron ficheros de gran tamaño, scripts de R y ordenes en PLINK 1.9, necesitando un alto poder computacional. Las ejecuciones se llevaron a cabo en el superordenador Finis Terrae III del Centro de Supercomputación de Galicia (CESGA), que consiste en un clúster heterogéneo con 357 nodos de computación, basados en procesadores Intel Xeon Ice Lake 8352Y (32 núcleos y 256 GB de memoria RAM) e interconectados por una red Infiniband HDR 100.

IV. RESULTADOS Y DISCUSIÓN

A. Comparativa de las MAFs entre tipos de imputación

Comenzando con la exploración visual del gráfico para la tendencia en MAFs de los marcadores entre tipos de imputación, se puede apreciar, para cada cromosoma, una fuerte dependencia lineal, reflejada por los coeficientes de correlación de Pearson asociados (r) y comprendidos entre 0,9999601 y 0,9992513 (Anexo 1). A modo de ejemplo, la Figura 2 muestra la relación entre la MAF generada para los dos tipos de imputación en los cromosomas 11 y 15. A pesar de que el ajuste es muy bueno, en todos los casos se distinguen abundantes puntos que se desvían y que corresponden con marcadores donde se está produciendo una variación relevante en MAFs, debido al tipo de imputación realizada.

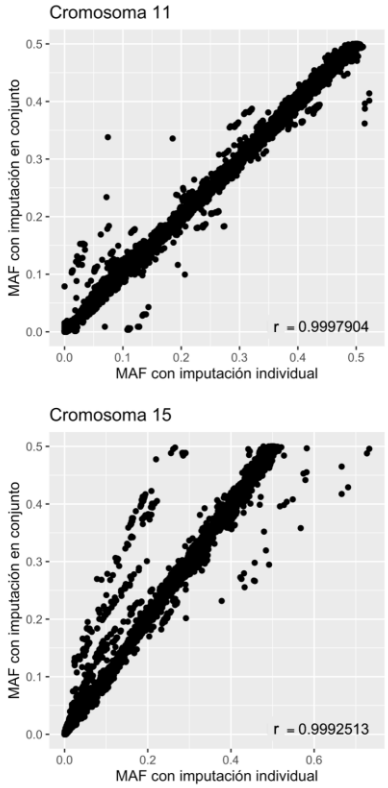


Figura 2. Gráfico de correlación y r asociado para las MAFs entre la subcohorte imputada de forma individual y en conjunto (cromosomas 11 y 15). Se aprecian numerosos marcadores que se desvían de la diagonal.

La Tabla I muestra la cantidad de marcadores filtrados (aquellos con una diferencia de MAFs entre tipos de imputación superior a 0,05) para cada cromosoma. Vemos que el caso extremo son los cromosomas 11 y 15, con 357 y 302 marcadores atípicos, respectivamente. En cambio, se produce una ausencia de marcadores atípicos en los cromosomas 5 y 12, lo que pone de manifiesto que no influye el tipo de imputación en los mismos para la subcohorte evaluada.

TABLA I. RESULTADOS DEL NÚMERO DE MARCADORES ATÍPICOS EN TÉRMINOS DE DIFERENCIA EN MAFS PARA LOS DOS TIPOS DE IMPUTACIÓN.

Cromosoma	*Marcadores totales*	*Marcadores atípicos*	*r*
1	8.553.267	219	0,9997864
2	9.223.282	140	0,9999255
3	7.655.401	2	0,9999565
4	7.561.380	7	0,9999548
5	7.005.543	-	0,9999601
6	6.668.396	4	0,9999552
7	6.275.350	16	0,9999506
8	6.000.476	1	0,9999560
9	4.704.138	162	0,9997886
10	5.157.212	6	0,9999370
11	5.270.952	357	0,9997904
12	5.133.523	-	0,9999519
13	3.862.684	6	0,9999565
14	3.431.884	1	0,9999380
15	3.147.065	302	0,9992513
16	3.480.859	11	0,9999121
17	3.073.840	10	0,9999306
18	3.066.725	11	0,9999518
19	2.407.721	1	0,9999429
20	2.419.620	1	0,9999550
21	1.430.516	82	0,9998406
22	1.468.096	69	0,9997846
X	3.982.189	93	0,9998425

Si bien los análisis posteriores se han realizado en todos los cromosomas que presentaron marcadores atípicos, esta memoria se centra en los cromosomas 11 y 15, donde el recuento ha sido más elevado.

B. Análisis de la influencia en la estima de MAFs por densidad y distribución de marcadores

Debido a que el proceso de control de calidad se realizó por separado en cada cohorte, el número final de marcadores genotipados a partir de los cuales se realizó la imputación puede variar a lo largo de cada cromosoma, teniendo consecuencias en la estima de las MAFs para los marcadores imputados. Esto puede implicar la introducción de artefactos y los correspondientes sesgos en los estudios de asociación [23-25].

En la Figura 3 se muestran para los cromosomas 11 y 15 las diferencias en número relativo de marcadores entre cohortes. Tal y como se puede observar, en términos generales, para ambos casos se producen notables variaciones de valores inferiores a 0,2, con dos picos próximos a 0,25 y 0,3. Atendiendo a las ventanas del cromosoma que comprenden los marcadores atípicos en términos de MAFs (indicados con el rectángulo rojo), se aprecia que, en estas, no tiene lugar un valor elevado para la medida analizada.

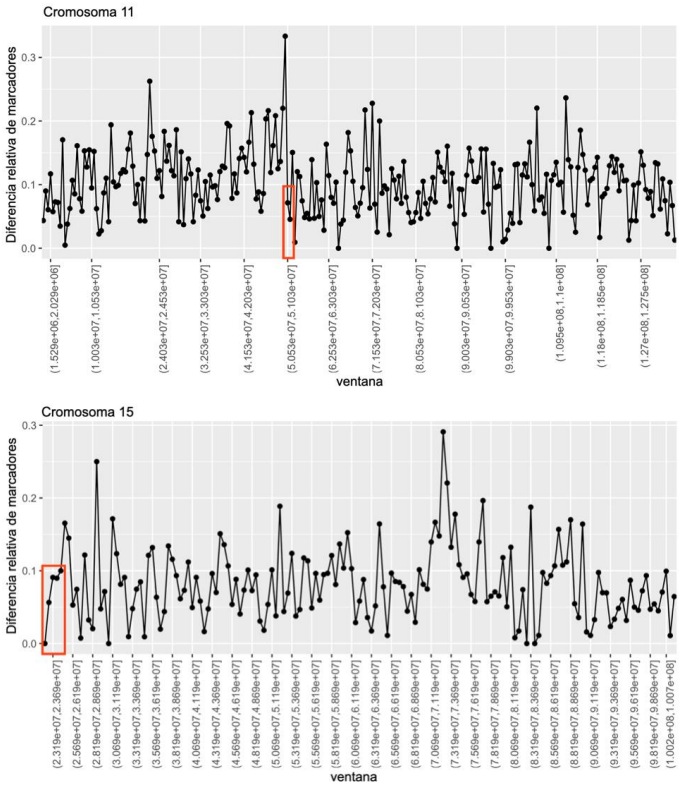


Figura 3. Representación gráfica de las diferencias en número relativo de marcadores entre las cohortes (cromosomas 11 y 15). En el eje *x* se representan las diferentes ventanas de 500.000 pares de bases generadas para el cromosoma y en el eje *y* las correspondientes diferencias determinadas. Los rectángulos rojos indican las ventanas donde se localizan los marcadores atípicos extraídos.

En el resto de los cromosomas con marcadores atípicos, se ha observado el mismo resultado. Por tanto, se descarta la posible relevancia de la densidad y distribución de marcadores en las diferencias de MAFs observadas por cada imputación realizada.

C. Relevancia de la calidad de imputación en la estima de MAFs

La potencia en la detección de asociaciones significativas en este tipo de estudios se encuentra determinada por el número de individuos en la cohorte empleada y la cobertura de los marcadores. Esta última depende principalmente del número y calidad de los genotipos imputados [26, 27]. Tal y como se ha mencionado anteriormente, TOPMed proporciona la calidad de imputación para cada marcador mediante la métrica r^2, que se ubica entre 0 y 1, reflejando la correlación entre los genotipos que se han imputado y los genotipos esperados [28]. En base a sus valores se pueden determinar aquellos marcadores cuya imputación no es fiable para los posteriores análisis.

La Figura 4 muestra, para la totalidad de marcadores en los cromosomas 11 y 15, las correspondientes diferencias en calidad de imputación al realizarla con la subcohorte individual y en conjunto. Se aprecian unos picos, mostrados con el rectángulo rojo, que corresponden a las regiones donde se ubican los marcadores atípicos en MAFs de cada cromosoma. Por lo tanto, en estas tiene lugar una importante diferencia en calidad de imputación si se compara con el resto del cromosoma.

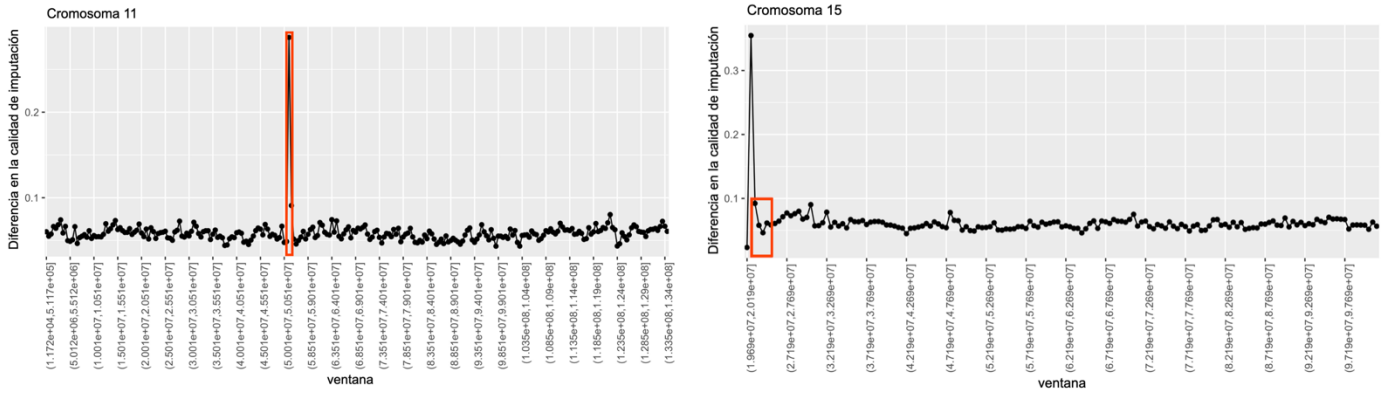


Figura 4. Representación gráfica de la diferencia en calidad de imputación para los cromosomas 11 y 15. El eje *x* muestra cada una de las ventanas y el eje *y* representa el valor en la diferencia de calidad entre cada tipo de imputación. Los rectángulos rojos comprenden las ventanas con los marcadores atípicos extraídos.

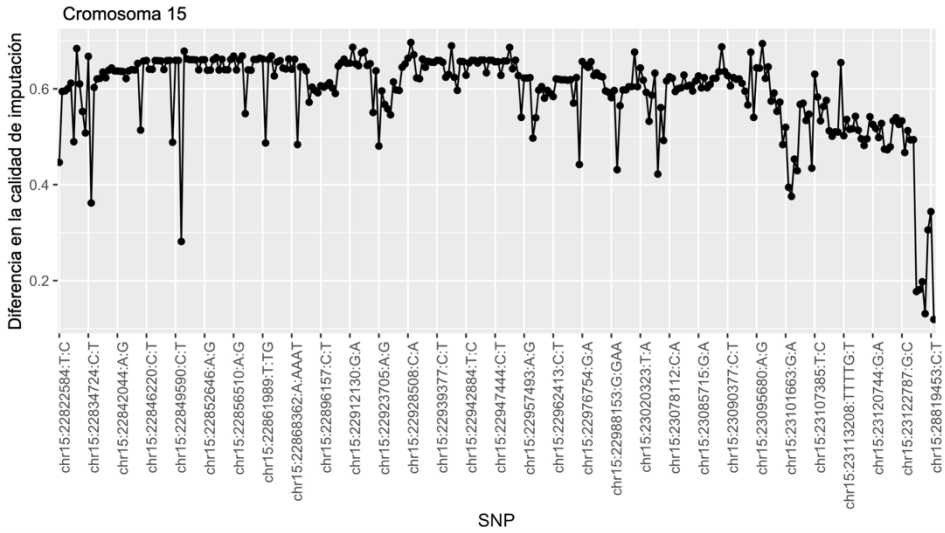


Figura 5. Representación de la diferencia en la calidad de imputación para los marcadores atípicos en el cromosoma 15. El eje *x* representa cada uno de los marcadores y el eje *y* muestra los valores en la diferencia de calidad entre los tipos de imputación.

Figura 6. Representación gráfica de la calidad de imputación por separado para los marcadores atípicos en el cromosoma 15. El eje *x* es cada uno de los marcadores y el eje *y* es el valor en la calidad para cada tipo de imputación.

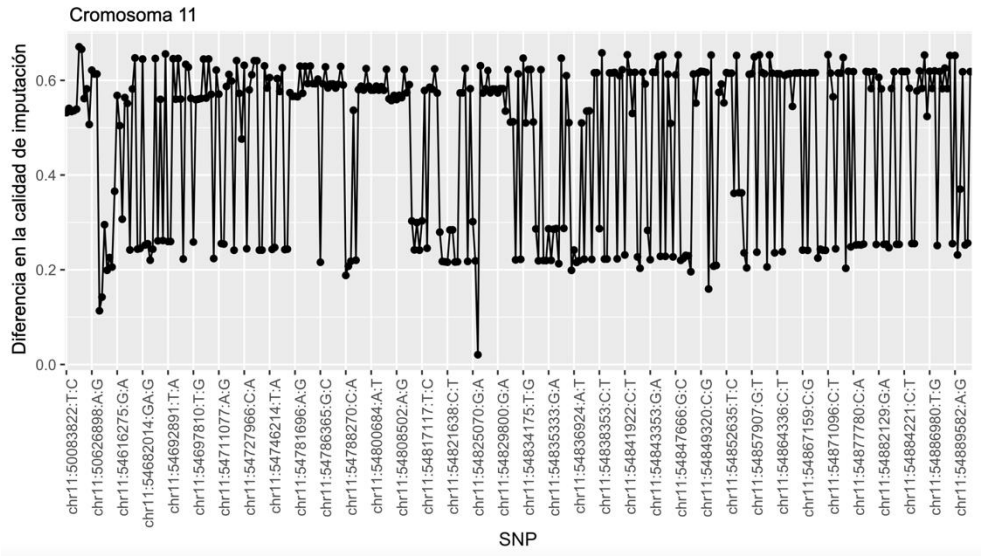


Figura 7. Representación de la diferencia en la calidad de imputación para los marcadores atípicos en el cromosoma 11. El eje *x* representa cada uno de los marcadores y el eje *y* muestra los valores en la diferencia de calidad entre los tipos de imputación.

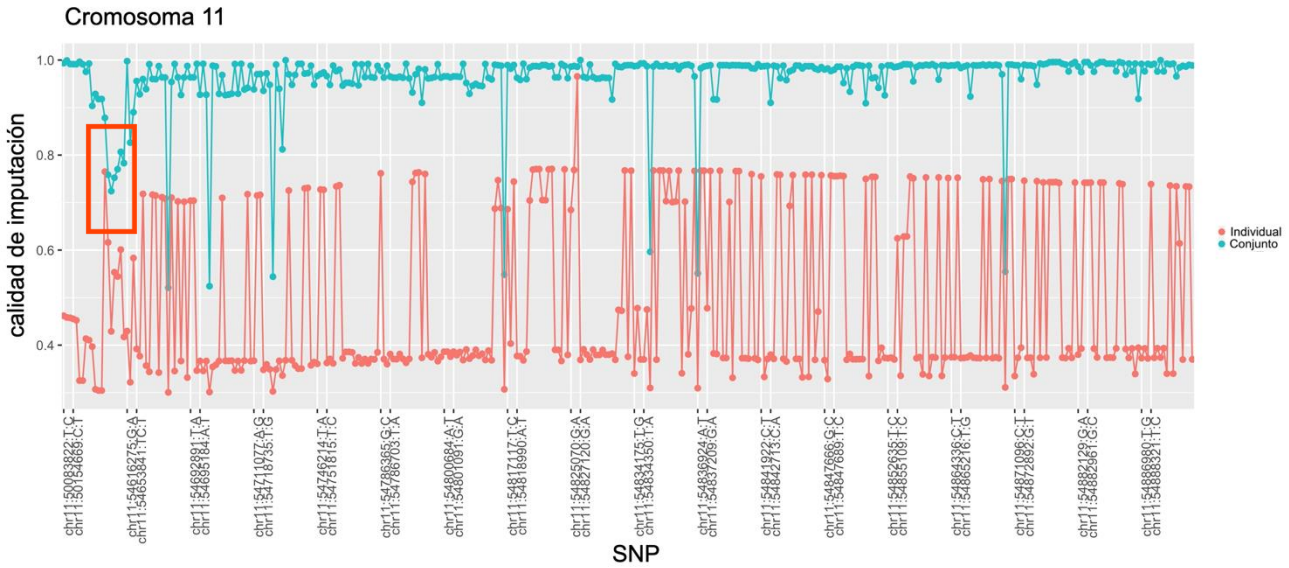


Figura 8. Representación gráfica de la calidad de imputación por separado para los marcadores atípicos en el cromosoma 11. El eje *x* es cada uno de los marcadores y el eje *y* es el valor en la calidad para cada tipo de imputación. El rectángulo rojo muestra una región donde la calidad de imputación es similar para ambos tipos.

Centrando los análisis para los marcadores atípicos en MAFs de los cromosomas 15 y 11, en las Figuras 5 y 7, se muestran las diferencias en calidad de imputación al realizarla con la subcohorte individual y en conjunto. Para el cromosoma 11 se pueden observar unos valores elevados en la diferencia, comprendidos aproximadamente entre 0,2 y 0,6. En cambio, aunque se produce una tendencia similar en el cromosoma 15, los valores mínimos ascienden aproximadamente a 0,4, con unas variaciones más suaves. En las Figuras 6 y 8 de los cromosomas 15 y 11, se ve una calidad de imputación notablemente superior al realizarse en conjunto, destacando una mayor magnitud de diferencia en el cromosoma 15. No obstante, para el cromosoma 11, en el caso de la imputación individual, hay abundantes marcadores que presentan una calidad de imputación notablemente elevada. Esto puede deberse al total de marcadores de entrada, el panel de referencia empleado y la calidad de los haplotipos iniciales para la imputación realizada [6]. En términos generales, tal y como se esperaba, con la imputación en conjunto el mayor tamaño de cohorte contribuye a una mejora en los resultados de calidad [29].

D. Filtrado de la calidad de imputación

Encontrar un balance entre mantener marcadores con alta calidad de imputación asociada y eliminar los que poseen baja calidad es un paso fundamental para el control de falsos positivos en el posterior análisis de asociación [30]. Mediante la fijación de un umbral óptimo para la calidad de imputación, es posible modular dicho balance, que varía en robustez de acuerdo con el proyecto y, además, contribuye a la reducción de los costes computacionales [31].

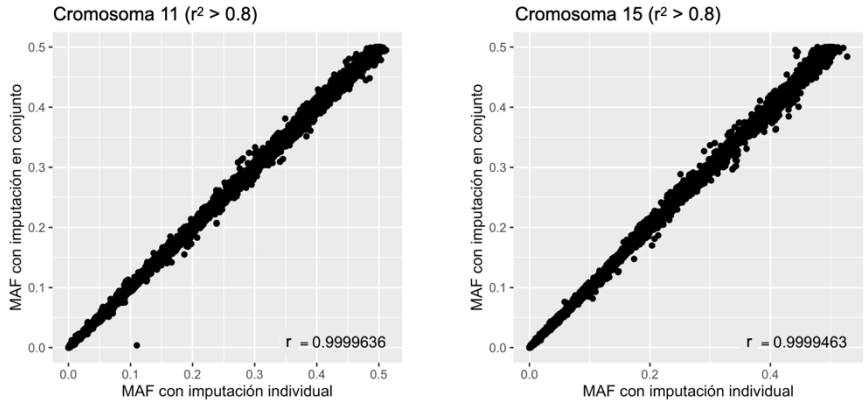


Figura 9. Gráfico de correlación y r asociado para las MAFs entre la subcohorte imputada de forma individual y en conjunto tras el filtrado en la calidad de imputación (cromosomas 11 y 15). Se aprecia la notable reducción en los marcadores que se desvían de la diagonal.

De acuerdo con aquellos marcadores atípicos en los que sus valores de la calidad de imputación son cercanos a 0,8 tanto al llevarse a cabo individualmente como en conjunto, tal y como sucede en una región relevante del cromosoma 11 (Figura 8), se fijó un valor de corte estricto de 0,8. Como resultado, se logra una drástica reducción en el número de marcadores atípicos en términos de MAFs por el tipo de imputación realizada. Además, tal y como se muestra en la Figura 9, mejora la correlación entre las MAFs de los marcadores según la imputación de la subcohorte.

TABLA II. RESULTADOS DEL NÚMERO DE MARCADORES ATÍPICOS TRAS APLICAR EL FILTRADO EN LA CALIDAD DE IMPUTACIÓN.

Cromosoma	***Marcadores totales***	***Marcadores filtrados***	***Marcadores atípicos***	***r***
1	8.553.267	6.009.109	-	0,9999602
2	9.223.282	6.529.383	-	0,9999525
3	7.655.401	5.466.344	-	0,9999662
4	7.561.380	5.413.192	-	0,9999651
6	6.668.396	4.840.996	-	0,9999636
7	6.275.350	4.457.000	-	0,9999677
8	6.000.476	4.284.320	-	0,9999639
9	4.704.138	3.316.225	-	0,9999550
10	5.157.212	3.755.212	-	0,9999611
11	5.270.952	3.743.241	1	0,9999636
13	3.862.684	2.745.075	-	0,9999646
14	3.431.884	2.428.213	-	0,9999576
15	3.147.065	2.204.444	-	0,9999463
16	3.480.859	2.388.588	3	0,9999437
17	3.073.840	2.096.336	-	0,9999641
18	3.066.725	2.156.883	-	0,9999632
19	2.407.721	1.651.122	-	0,9999558
20	2.419.620	1.667.101	-	0,9999642
21	1.430.516	978.949	-	0,9999618
22	1.468.096	987.398	1	0,9999474
X	3.982.189	2.439.908	6	0,9999444

Atendiendo a los resultados en detalle, el descenso en el número de marcadores atípicos posterior al filtrado es generalizado (Tabla II). Si bien en los cromosomas 11, 16, 22 y X quedan algunos casos, posiblemente debido a marcadores intermedios de regiones con alta densidad, se mantiene una tendencia similar, por lo que un ligero incremento en el umbral definido sería lo recomendable en términos de robustez. Por otra parte, en la mayoría de los cromosomas, el número de marcadores que superan el filtrado no representan una pérdida del 50% sobre el total, lo que tiene especial interés para garantizar un poder estadístico en los posteriores análisis que se realicen [32].

V. CONCLUSIONES

Los GWAS son estudios observacionales muy efectivos en la identificación de variantes genéticas asociadas con enfermedades o caracteres. Junto con el control de calidad inicial, la imputación conforma una fase clave, permitiendo la inferencia de marcadores sin determinar o perdidos en las muestras de estudio, así como de variantes raras o específicas de ancestralidad. No obstante, la imputación se puede ver alterada por multitud de factores, desencadenando falsos positivos en los posteriores análisis. Con respecto a la ancestralidad, las poblaciones no europeas suponen un desafío, debido a su arquitectura genética más compleja. En consecuencia, es fundamental abordar criterios para controlar el efecto de los diferentes factores, adaptando los rangos de parámetros en cada fase de los GWAS para mantener la precisión y el rendimiento de acuerdo con las características del proyecto.

En este trabajo se ha llevado a cabo una exploración de la posible influencia al imputar una subcohorte de individuos latinos del consorcio SCOURGE, individualmente o en conjunto, con el total de otra cohorte geográficamente distinta. Se analizaron las frecuencias del alelo menor (MAFs) para los marcadores genotipados, determinando aquellos marcadores donde tenían lugar diferencias atípicas en su estima. En base a estos, posteriormente, se verificó la relevancia de la densidad de los marcadores y la calidad lograda en la imputación.

Los resultados confirmaron la existencia de notables diferencias en las MAFs entre los tipos de imputación, con la existencia de un mayor número de marcadores atípicos en términos de estas para los cromosomas 11 y 15. Se descartó la

influencia de la densidad de marcadores genotipados en esas regiones, comparando con el resto de cada cromosoma. En cambio, presentaron unas notables diferencias en la calidad de imputación. El posterior filtrado para restringir aquellos marcadores que superasen el umbral de 0,8, puso de manifiesto una efectiva reducción en el número de marcadores atípicos. De esta forma, para estudios internacionales que impliquen la imputación de poblaciones de latinos combinadas con otras de origen europeo y no europeo, se propone en la fase posterior a la imputación, fijar un filtro estricto en la calidad de imputación, para el control del ruido asociado a este tipo de marcadores.

Como trabajos futuros, para comprobar la tendencia observada, se plantea la ampliación por estudios adicionales con diferentes tamaños de la subcohorte y variabilidad en torno a las áreas geográficas, balance casos-controles y porcentajes de ancestralidad nativoamericana, africana y europea a nivel individual. Además, posteriormente se podría incluir el umbral para la calidad de imputación que se considere adecuado en un *pipeline* bioinformático, basado en el software empleado para este tipo de proyectos. Esto permitiría un eficiente control de calidad centrado en uno de los factores con mayor impacto en los análisis.

REFERENCIAS

[1] E. Cano-Gamez and G. Trynka, "From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases," Front. Genet., vol. 11, no. 424, 2020.

[2] H. Fitipaldi and P. W. Franks, "Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005-2022," Hum. Mol. Genet., vol. 32, no. 3, pp. 520-532, 2023.

[3] A. T. Marees et al., "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," Int. J. Methods Psychiatr. Res., vol. 27, no. 2, p. e1608, 2018.

[4] P. M. Visscher et al., "10 years of GWAS discovery: Biology, function, and translation," Am. J. Hum. Genet., vol. 101, no. 1, pp. 5-22, 2017.

[5] Y. Wu, F. Hormozdiari, J. W. J. Joo, and E. Eskin, "Improving imputation accuracy by inferring causal variants in genetic studies," J. Comput. Biol., vol. 26, no. 11, pp. 1203-1213, 2019.

[6] E. Porcu, S. Sanna, C. Fuchsberger, and L. G. Fritsche, "Genotype imputation in genome-wide association studies," Curr. Protoc. Hum. Genet., vol. 78, no. 1, pp. 1-14, 2013.

[7] V. Q. Truong et al., "Quality control procedures for genome-wide association studies," Curr. Protoc., vol. 2, no. 11, p. e603, 2022.

[8] R. De, W. S. Bush, and J. H. Moore, "Bioinformatics challenges in genome-wide association studies (GWAS)," Methods Mol. Biol., vol. 1168, pp. 63-81, 2014.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[10] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," Nat. Rev. Genet., vol. 11, no. 7, pp. 499-511, 2010.

[11] Y. Guan and M. Stephens, "Practical issues in imputation-based association mapping," PLoS Genet., vol. 4, no. 12, 2008.

[12] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes," Genet. Epidemiol., vol. 34, no. 8, pp. 816-834, 2010.

[13] B. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," PLoS Genet., vol. 5, no. 6, 2009.

[14] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering," Am. J. Hum. Genet., vol. 81, no. 5, pp. 1084-1097, 2007.

[15] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," Nat. Genet., vol. 44, no. 8, pp. 955-959, 2012.

[16] S. Das et al., "Next-generation genotype imputation service and methods," Nat. Genet., vol. 48, no. 10, pp. 1284-1287, 2016.

[17] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini, "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip," PLoS Genet., vol. 5, no. 5, 2009.

[18] K. Hao, E. Chudin, J. McElwee, and E. E. Schadt, "Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies," BMC Genetics, vol. 10, no. 27, 2009.

[19] K. Bryc et al., "Genome-wide patterns of population structure and admixture among Hispanic/Latino populations," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 2, pp. 8954-8961, 2010.

[20] H. R. Dueñas, C. Seah, J. S. Johnson, and L. M. Huckins, "Implicit bias of encoded variables: frameworks for addressing structured bias in EHR-GWAS data," Hum. Mol. Genet., vol. 29, no. 1, pp. 33-41, 2020.

[21] S. Purcell et al., "PLINK: A tool set for whole-genome association and population-based linkage analyses," Am. J. Hum. Genet., vol. 81, no. 3, pp. 559-575, 2007.

[22] R. Cruz et al., "Novel genes and sex differences in COVID-19 severity," Hum. Mol. Genet., vol. 31, no. 22, pp. 3789-3806, 2022.

[23] E. O. Johnson et al., "Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy," Hum. Genet., vol. 132, no. 5, pp. 509-522, 2013.

[24] Z. Zhang, X. Xiao, W. Zhou, D. Zhu, and C. I. Amos, "False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy," Hum. Mol. Genet., vol. 31, no. 1, pp. 146-155, 2022.

[25] P. Deelen et al., "Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'," Eur. J. Hum. Genet., vol. 22, no. 11, pp. 1321-1326, 2014.

[26] J. Zheng, Y. Li, G. R. Abecasis, and P. Scheet, "A comparison of approaches to account for uncertainty in analysis of imputed genotypes," Genet. Epidemiol., vol. 35, no. 2, pp. 102-110, 2011.

[27] M. H. Kowalski et al., "Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations," PLoS Genet., vol. 15, no. 12, 2019.

[28] Y. Li, C. J. Willer, S. Sanna, and G. R. Abecasis, "Genotype imputation," Annu. Rev. Genomics Hum. Genet., vol. 10, pp. 387-406, 2009.

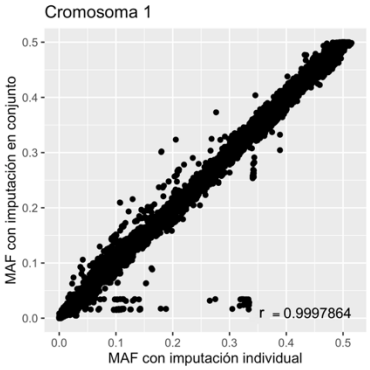
[29] T. Pook et al., "Improving imputation quality in BEAGLE for crop and livestock data," G3, vol. 10, no. 1, pp. 177-188, 2020.

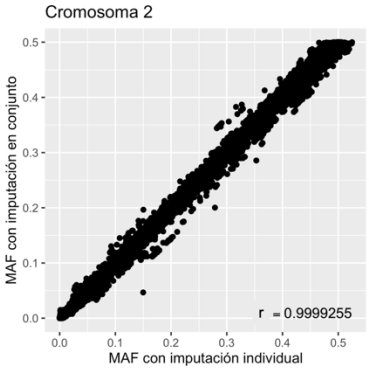
[30] Y. Xie, D. B. Hancock, E. O. Johnson, and J. P. Rice, "Two adjustment strategies for imputation across genotyping arrays," Hum. Hered., vol. 78, no. 2, pp. 73-80, 2014.

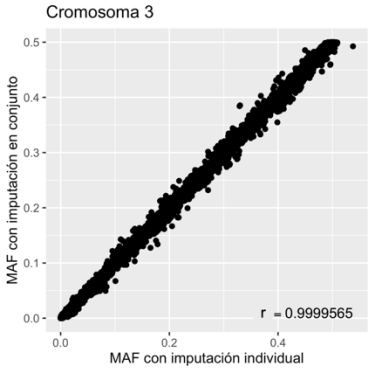
[31] E. Y. Liu et al., "Genotype imputation of Metabochip SNPs using a study-specific reference panel of 4,000 haplotypes in African Americans from the Women´s Health Initiative," Genet. Epidemiol., vol. 36, no. 2, pp. 107-117, 2012.

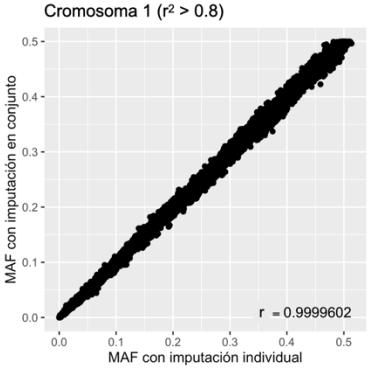
[32] C. A. Anderson et al., "Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms," Am. J. Hum. Genet., vol. 83, no. 1, pp. 112-119, 2008.

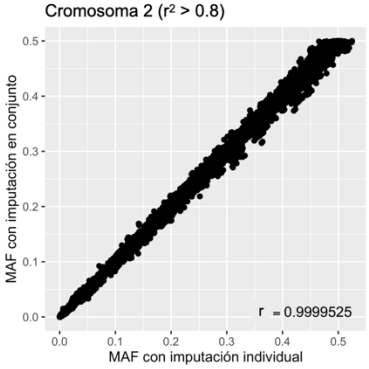
ANEXO

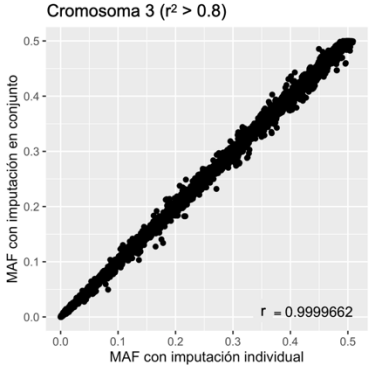

Cromosoma 1
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9997864

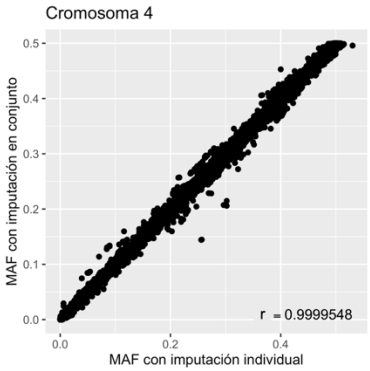

Cromosoma 2
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999255

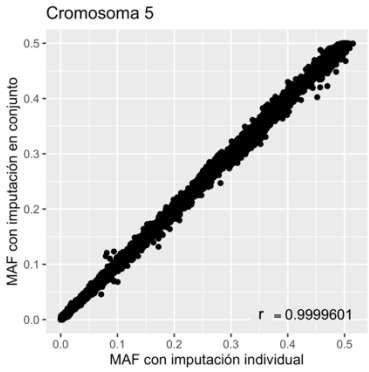

Cromosoma 3
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999565

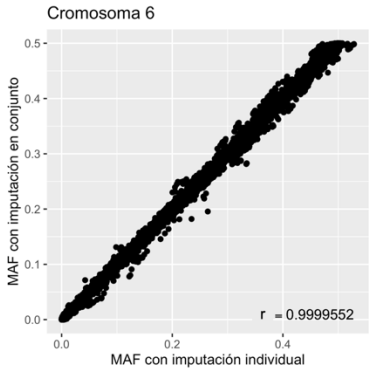

Cromosoma 1 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999602

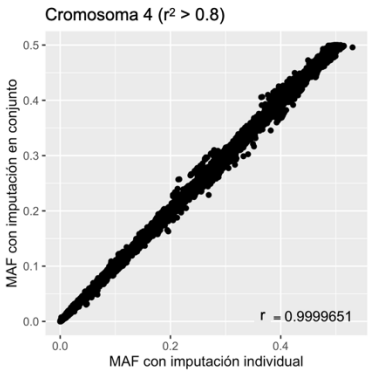

Cromosoma 2 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999525

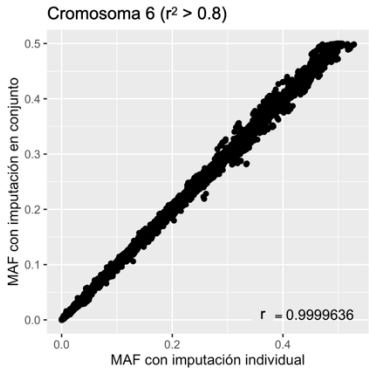

Cromosoma 3 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999662

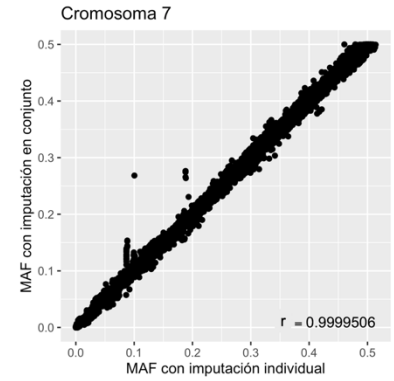

Cromosoma 4
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999548

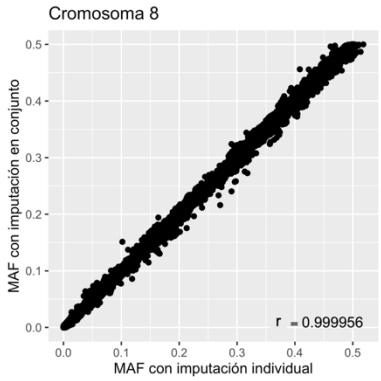

Cromosoma 5
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999601

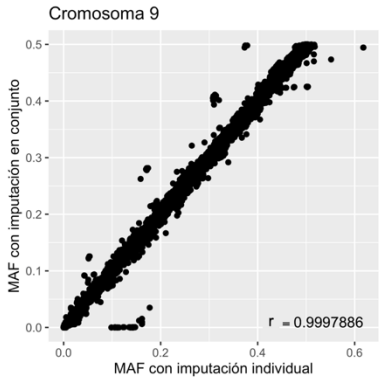

Cromosoma 6
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999552

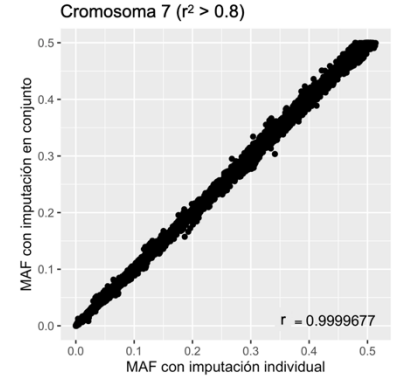

Cromosoma 4 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999651

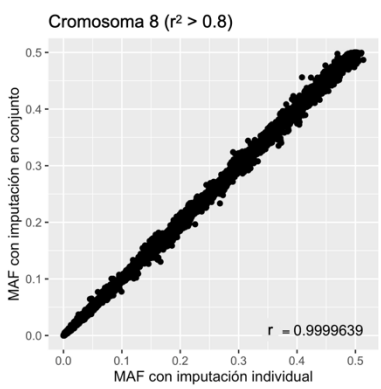

Cromosoma 6 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999636

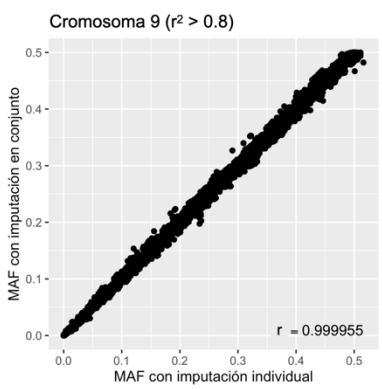

Cromosoma 7
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999506

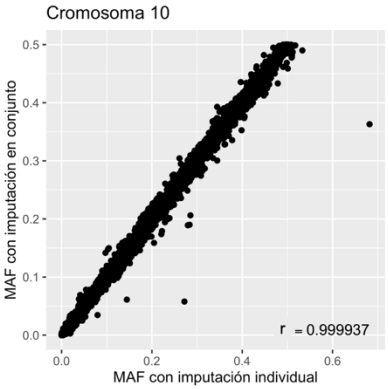

Cromosoma 8
MAF con imputación en conjunto
MAF con imputación individual
r = 0.999956

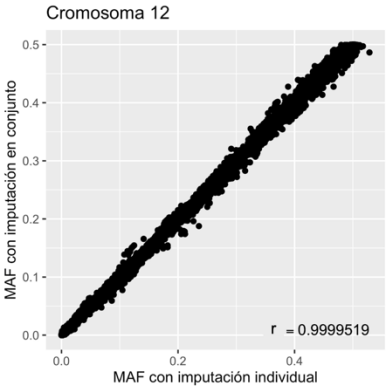

Cromosoma 9
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9997886

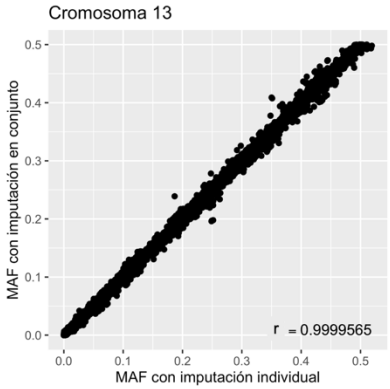

Cromosoma 7 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999677

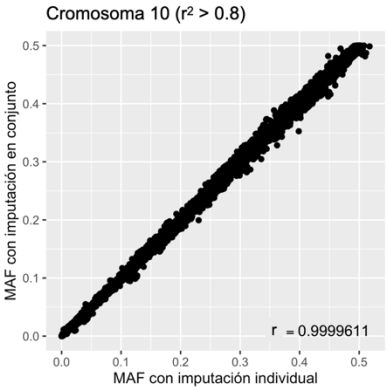

Cromosoma 8 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999639

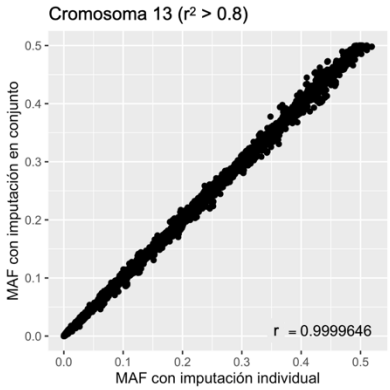

Cromosoma 9 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.999955

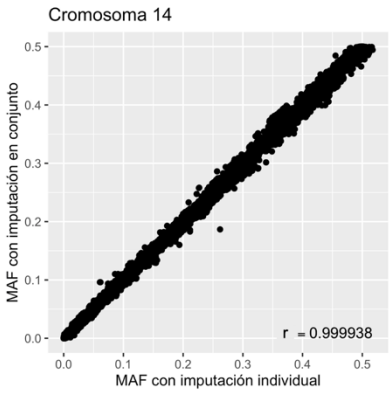

Cromosoma 10
MAF con imputación en conjunto
MAF con imputación individual
r = 0.999937

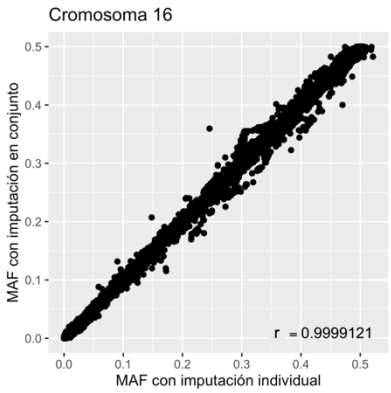

Cromosoma 12
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999519

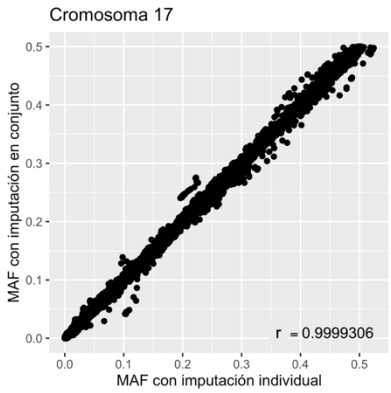

Cromosoma 13
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999565

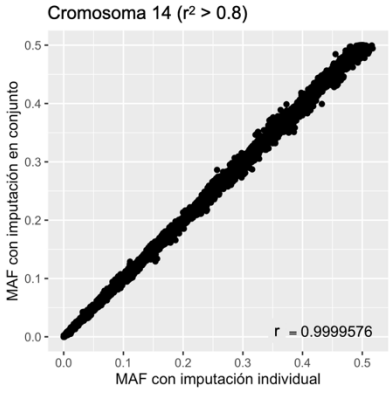

Cromosoma 10 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999611

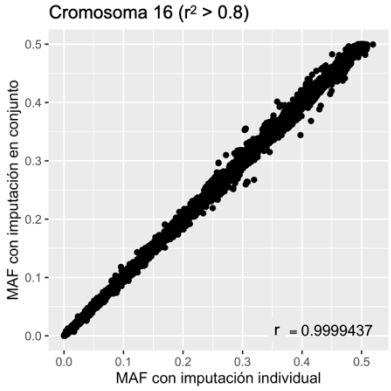

Cromosoma 13 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999646

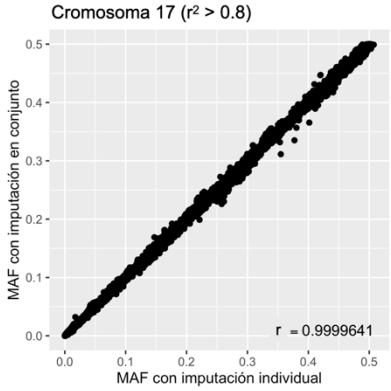

Cromosoma 14
MAF con imputación en conjunto
MAF con imputación individual
r = 0.999938

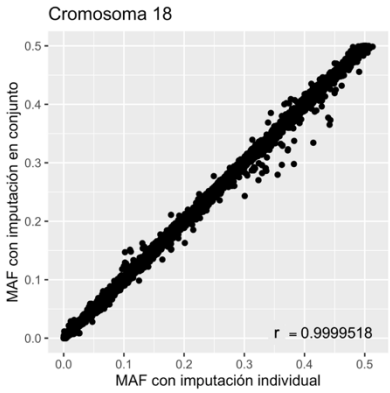

Cromosoma 16
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999121

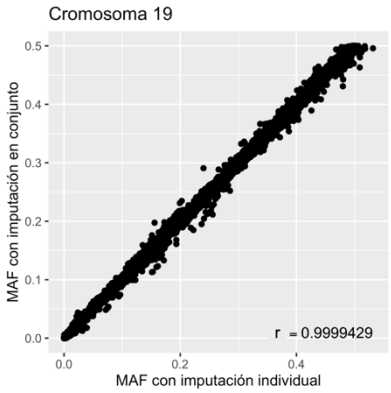

Cromosoma 17
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999306

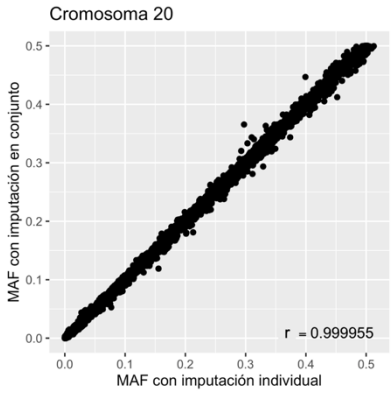

Cromosoma 14 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999576

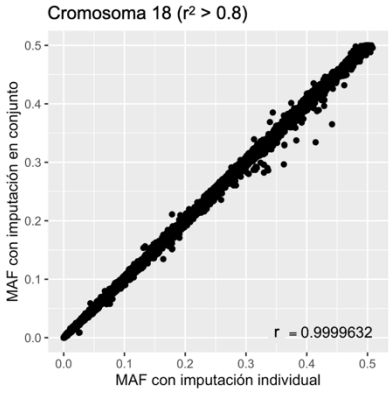

Cromosoma 16 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999437

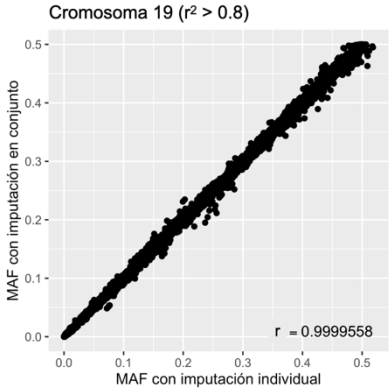

Cromosoma 17 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999641

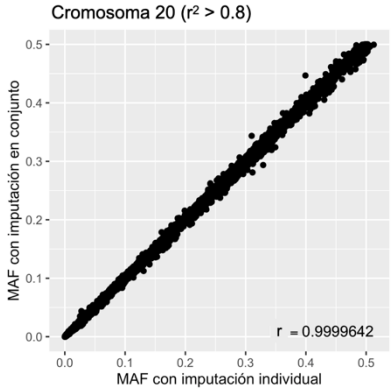

Cromosoma 18
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999518

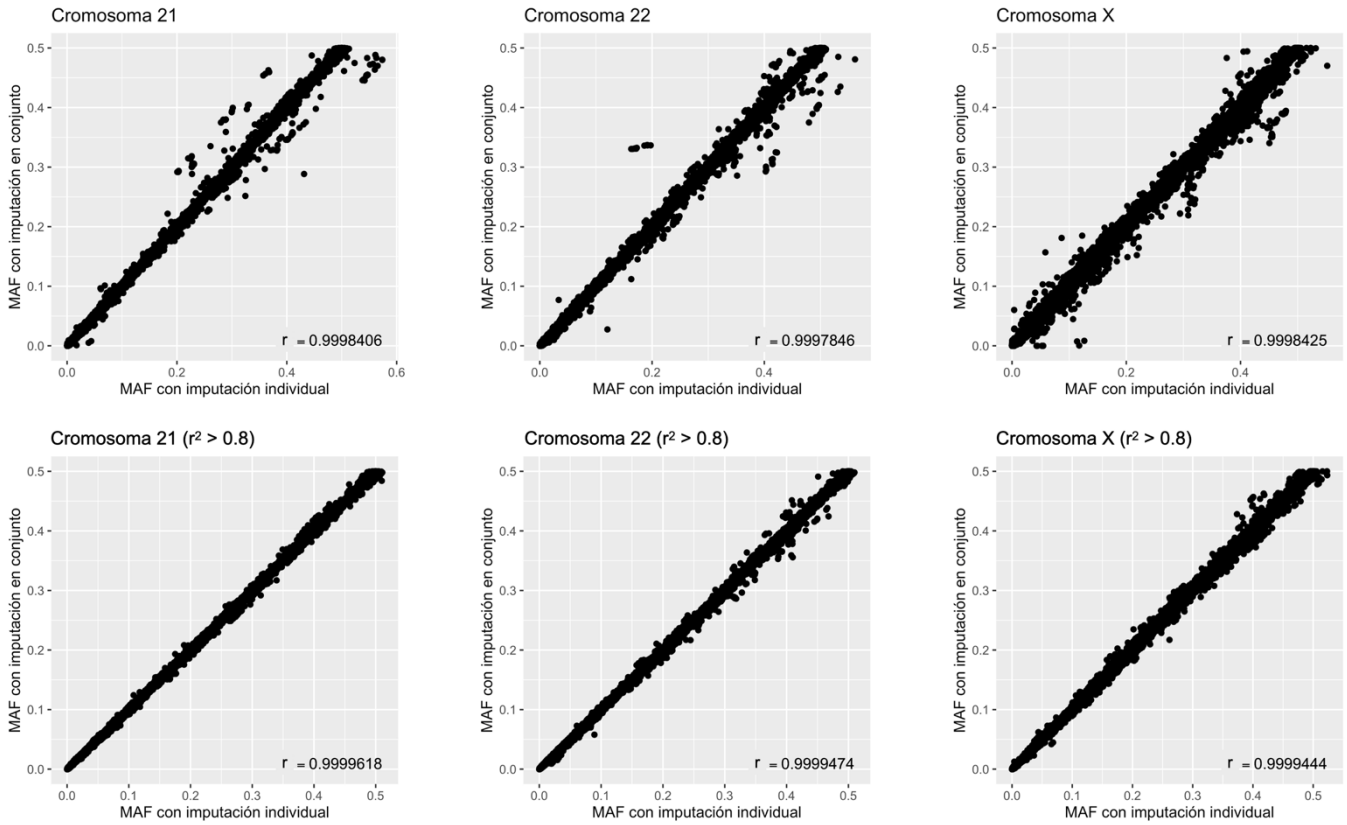

Cromosoma 19
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999429


Cromosoma 20
MAF con imputación en conjunto
MAF con imputación individual
r = 0.999955


Cromosoma 18 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999632


Cromosoma 19 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999558


Cromosoma 20 (r² > 0.8)
MAF con imputación en conjunto
MAF con imputación individual
r = 0.9999642



Anexo 1. Gráfico de correlación y r asociado para las MAFs entre la subcohorte imputada de forma individual y en conjunto, previo y posterior al filtrado en la calidad de imputación (superior e inferior, respectivamente).