Imputación de genotipos faltantes mediante algoritmos de machine learning

M. Agustina Raschia[1,5[0000-0002-0662-4397]], Pablo J. Ríos[2,6[0000-0002-9768-7587]], Marcela E. Cordoba[1[0009-0000-9774-6295]], M. Eugenia Caffaro[1[0000-0002-5814-2293]], M. Valeria Donzelli[1,7[0009-0009-4243-4652]], Daniel O. Maizon[3,8[0000-0002-2701-4109]], Daniel Demitrio[4,6[0009-0008-3372-3340]], y Mario A. Poli[1,9[0000-0001-8775-2333]]

[1] Instituto Nacional de Tecnología Agropecuaria (INTA), CICVyA-CNIA, Instituto de Genética "Ewald A. Favret", Nicolás Repetto y de Los Reseros s/n, Hurlingham (B1686), Buenos Aires, Argentina raschia.maria@inta.gob.ar; cordoba.marcela@inta.gob.ar; caffaro.maria@inta.gob.ar; donzelli.maria@inta.gob.ar; poli.mario@inta.gob.ar
[2] Universidad de Buenos Aires, Buenos Aires, Argentina.pablo.javier.rios@gmail.com
[3] Instituto Nacional de Tecnología Agropecuaria (INTA), E.E.A. Anguil, Ruta 5 Km 580, Anguil (6326), La Pampa, Argentina maizon.daniel@inta.gob.ar
[4] Instituto Nacional de Tecnología Agropecuaria (INTA), Coordinación Nacional de Relaciones Institucionales y Vinculación Tecnológica, Chile 460 - Piso 1, Buenos Aires, Argentina demitrio.daniel@inta.gob.ar
[5] Facultad de Ciencias Médicas, Universidad Nacional de La Plata, Argentina.
[6] Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina.
[7] Facultad de Ciencias Agrarias, Universidad Nacional de Lomas de Zamora, Argentina.
[8] Facultad de Agronomía, Universidad Nacional de La Pampa, Argentina.
[9] Facultad de Ciencias Agrarias y Veterinarias, Universidad del Salvador, Argentina.

Resumen. La imputación o inferencia de genotipos faltantes utilizando correlaciones entre variantes obtenidas a partir de paneles de referencia puede ser llevada a cabo por programas específicos basados en la utilización de información genética familiar y/o poblacional o mediante la implementación de algoritmos de machine learning. El objetivo de este trabajo fue evaluar la precisión en la imputación lograda mediante distintas estrategias de machine learning, tras comparar genotipos imputados con los obtenidos por genotipificación con un microarreglo de mediana densidad de SNPs. Sobre una base de datos con genotipos de 966 ovinos en 57.876 SNPs, con 53,4% de genotipos faltantes, se exploraron tres estrategias de imputación basadas en el algoritmo random forest. Un subconjunto de los genotipos imputados, correspondientes a 232 animales en 30.924 SNPs, fue comparado con genotipos obtenidos por genotipificación. El porcentaje de concordancia obtenido para las tres estrategias fue de alrededor de 60%. Este bajo porcentaje puede atribuirse a la gran cantidad de genotipos no asignados del archivo de partida. Una estrategia para aumentar la precisión de la imputación podría ser aumentar el número de animales en la población de referencia y, de este modo, reducir la proporción de genotipos faltantes en el conjunto de datos.

Palabras clave: imputación, machine learning, random forest, polimorfismo de nucleótido único

Imputation of missing genotypes using machine learning algorithms

Abstract. The imputation or inference of missing genotypes using correlations between variants obtained from reference panels can be carried out by specific programs that utilize family and/or population genetic information or by implementing machine learning algorithms. The objective of this study was to evaluate the imputation accuracy achieved using different machine learning strategies by comparing imputed genotypes with those obtained by genotyping with a medium-density SNP microarray. To compare the performance of three imputation strategies using the random forest algorithm, we analyzed a database containing 966 sheep genotyped at 57,876 SNPs, where 53.4% of the data was missing. A subset of the imputed genotypes, corresponding to 232 animals at 30,924 SNPs, was compared with genotypes obtained by genotyping. The percentage of concordance obtained for the three strategies was approximately 60%. This low percentage can be attributed to the large number of missing genotypes in the source file. One strategy for increasing imputation accuracy would be to increase the number of animals in the reference population and thus reduce the proportion of missing genotypes in the data set.

Keywords: imputation, machine learning, random forest, single nucleotide polymorphism

1 Introducción

Los avances en las tecnologías del ADN, la disponibilidad y accesibilidad a servicios de genotipificación y el desarrollo de la bioinformática han contribuido al aumento en el uso de información genética en los programas de cría de ganado que ha tenido lugar en las últimas décadas. La selección de animales según evaluaciones genéticas tradicionales basadas en registros fenotípicos y genealógicos ha sido complementada con la selección asistida por marcadores (Fernando y Grossman, 1989) y, posteriormente, con la selección genómica (Meuwissen et al., 2001). Estas formas de selección se basan en la existencia de desequilibrio de ligamiento entre marcadores moleculares y loci de caracteres cuantitativos, por lo que hacen uso de información genotípica. El desarrollo de tecnologías de genotipificación de polimorfismos de nucleótido único (SNPs) eficientes y de alto rendimiento posibilitó reducir costos de genotipado y obtener información genotípica de muestras individuales en decenas de miles a millones de SNPs a costos relativamente bajos. En la actualidad existe variedad de microarreglos de distintas plataformas comerciales que evalúan diferente número de SNPs en numerosas especies (Ragoussis, 2009; Tosser-Klopp et al., 2016; Marina et al., 2021; Ghavi Hossein-Zadeh, 2024; Ogunbawo et al., 2024). Sin embargo, si el número de animales a seleccionar es grande, la implementación de la selección genómica sigue siendo costosa (Abolhassani Targhi et al., 2019). Una práctica común

es recurrir a la imputación, es decir, la inferencia de genotipos no asignados o en marcadores no evaluados utilizando correlaciones entre las variantes construidas a partir de paneles de referencia que facilitan una buena cobertura del genoma (Naito y Okada, 2024). Los procedimientos de imputación generalmente implican un conjunto de muestras a imputar, un panel de referencia de haplotipos en fase a partir del cual se inferirán los genotipos y un algoritmo para el procedimiento de imputación (Jewett et al., 2012). Entonces, suelen imputarse genotipos obtenidos con un microarreglo de baja densidad de marcadores utilizando como referencia un panel de genotipos obtenidos mediante genotipificación con microarreglos de mayor densidad. Se han desarrollado programas específicos que utilizan estrategias de imputación basadas en relaciones de parentesco y reglas de herencia mendeliana y/o métodos que infieren genotipos utilizando información poblacional basada en parámetros como el desequilibrio de ligamiento (Li et al., 2009). Entre ellos, pueden mencionarse los siguientes: fastPHASE (Scheet and Stephens, 2006), BEAGLE (Browning and Browning, 2007, 2009), IMPUTE2 (Howie et al., 2009), FImpute (Sargolzaei et al., 2014), PEDIMPUTE (Nicolazzi et al., 2013). La imputación de genotipos faltantes también puede llevarse a cabo mediante algoritmos de machine learning (Mikhchi et al., 2016; Naito y Okada, 2024; Mora-Márquez et al., 2025). En el aprendizaje supervisado, estos algoritmos (como random forest, Schwarz et al., 2009) requieren paneles haplotípicos o genotipos en fase. Como métodos no supervisados, otros algoritmos de machine learning (como k-nearest neighbors, Money et al., 2015), imputan genotipos faltantes sin requerir información adicional al conjunto de datos. La precisión de la imputación depende de numerosos factores, incluida la diversidad haplotípica de la población a imputar, el tamaño del panel de referencia, la similitud genética entre la población de referencia y la población a imputar, la frecuencia alélica mínima de los SNPs a imputar y el método usado (Jewett et al., 2012; Mikhchi et al., 2016).

El objetivo de este trabajo fue evaluar la precisión en la imputación lograda mediante distintas estrategias de machine learning. Para ello se compararon genotipos imputados de un panel de 15K SNPs a uno de 50K, con los obtenidos por genotipificación con un microarreglo de mediana densidad.

2 Materiales y métodos

2.1 Bases de datos genotípicos

Como input se utilizaron tres bases de datos genotípicos obtenidas tras genotipificar ovinos de raza Corriedale con microarreglos de distinta densidad de la plataforma Illumina. Una de ellas contenía los genotipos de 677 animales en 14.943 SNPs evaluados por el microarreglo 15K, mientras que las dos restantes contenían los genotipos de 24 y 285 animales en 54.241 y 53.511 SNPs evaluados por los microarreglos 50K v1 y v2, respectivamente. En todos los casos, la notación alélica utilizada fue la codificación A/B de Illumina. Del total de SNPs evaluados, 11.197 son compartidos por los tres microarreglos, mientras que otros son compartidos por dos microarreglos o evaluados únicamente por un microarreglo (Figura 1). Del total de

animales evaluados, 17 fueron genotipificados con los microarreglos 15K y 50Kv1 y tres con los microarreglos 15K y 50Kv2.

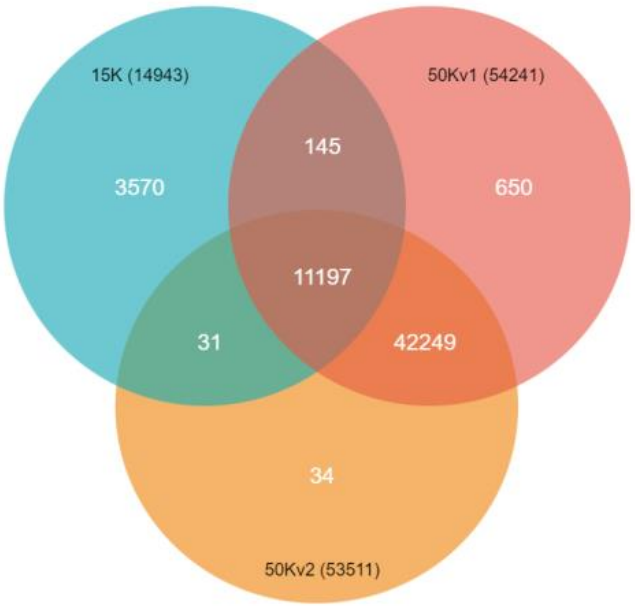

Fig. 1. Número de SNPs compartidos por los distintos microarreglos.

Las tres bases de datos mencionadas se combinaron mediante la función merge del programa PLINKv1.9 (Chang et al., 2015; PLINK 1.9 Homepage, 2025), resultando en una base combinada con genotipos de 966 animales en 57.876 SNPs únicos. Para la combinación se utilizó el modo consenso de merge que, en caso de encontrar para un mismo animal genotipos diferentes asignados para un mismo SNP evaluado por distintos microarreglos, califica como no asignado el genotipo de ese animal en ese SNP. Los genotipos del dataset combinado, expresados con notación alélica A/B, fueron recodificados mediante la función recodeA de PLINK para ser expresados con la codificación aditiva 0/1/2 que requieren los algoritmos de machine learning. Dado que por default la función recodeA computa el alelo que posee la menor frecuencia, se utilizó además la función recode-allele para computar, para todos los SNPs, los alelos "A". El archivo combinado, con genotipos de 966 animales en 57.876 SNPs, expresados en codificación 0/1/2, se utilizó como input para el algoritmo de machine learning utilizado para realizar la imputación. Este archivo tenía 53,38% de genotipos no asignados, correspondientes a SNPs que no son evaluados por algún microarreglo o que fueron evaluados pero resultaron no asignados por algún problema en el proceso de genotipificación. De los 57.876 SNPs evaluados, 65 poseían una tasa de genotipificación=0, es decir, los 966 animales analizados poseían genotipo no asignado para este locus, por lo que no fue posible imputar valores a estos SNPs.

2.2 Algoritmo de imputación

Se utilizó el algoritmo de random forest (Ishwaran et al., 2021) usando la implementación en R randomForestSRC versión 3.3.1 para imputar datos faltantes. Se exploraron múltiples estrategias de imputación variando los siguientes parámetros del método impute (Fast Unified Random Forests with randomForestSRC 3.3.1, Impute Only Mode, 2025): max.iter (máximo número de iteraciones usadas), ntree (número de árboles creados), nodesize (tamaño promedio del nodo terminal del bosque), mf.q (especifica la fracción de variables utilizadas como respuestas en la imputación multivariada de missForest) y fast (utiliza bosques aleatorios rápidos). Los valores utilizados de los parámetros del algoritmo de imputación se asignaron mediante una exploración manual, extrapolando los tiempos de ejecución de la imputación de todas las variables. Finalmente, se seleccionaron tres estrategias de imputación. El criterio que se utilizó para la selección se basó en considerar el menor error de imputación y tiempos de procesamiento que no excedieran las 8 horas en el hardware utilizado para las corridas (CPU Intel® Core™ i7-7820X CPU @ 3.60GHz; 32 GB de memoria RAM; disco NVMe SSD de 500 GB). Cada imputación consistió en ejecutar el método impute() de a bloques de SNPs contiguos dado que la complejidad computacional de este algoritmo es cuadrática en función del número de SNPs (O(N2)). Las tres estrategias de imputación se resumen en la Tabla 1.

Tabla 1. Parámetros del método impute() para cada estrategia de imputación seguida.

# estrategia de imputación	Parámetros impute()	Tamaño de bloque de imputación
1	*max.iter=2, ntree=100, nodesize=5, mf.q=0,2, fast=FALSE*	1.000 SNPs contiguos (se realizaron 58 ejecuciones de imputación para imputar todos los datos faltantes del dataset).
2	*max.iter=5, ntree=100, nodesize=5, mf.q=0,2, fast=FALSE*	1.000 SNPs contiguos
3	*max.iter=5, ntree=100, nodesize=5, mf.q=0,2, fast=FALSE*	Se definieron 28 bloques, constituidos por los SNPs de un mismo cromosoma (cromosoma 0, desconocido; 1 a 26, autosomas; 27, cromosoma X). El cromosoma 28 contaba con un sólo SNP, que se incluyó en el cromosoma 27 para la imputación.

Se generaron tres archivos de salida (imputados), uno para cada una de las estrategias de imputación descriptas anteriormente. Los datasets imputados constan de 966 filas (animales) y 57.811 columnas (57.876 - 65 SNPs, dado que los 65 SNPs con tasa de genotipificación=0 se excluyeron del dataset de salida). Se imputaron todos los genotipos faltantes del dataset combinado, correspondientes a SNPs evaluados por los microarreglos 15K, 50Kv1 y 50Kv2. Los datasets de salida se recodificaron a notación

alélica A/B. El flujo de trabajo que describe el procesamiento de archivos para su imputación puede visualizarse en la Figura 2.

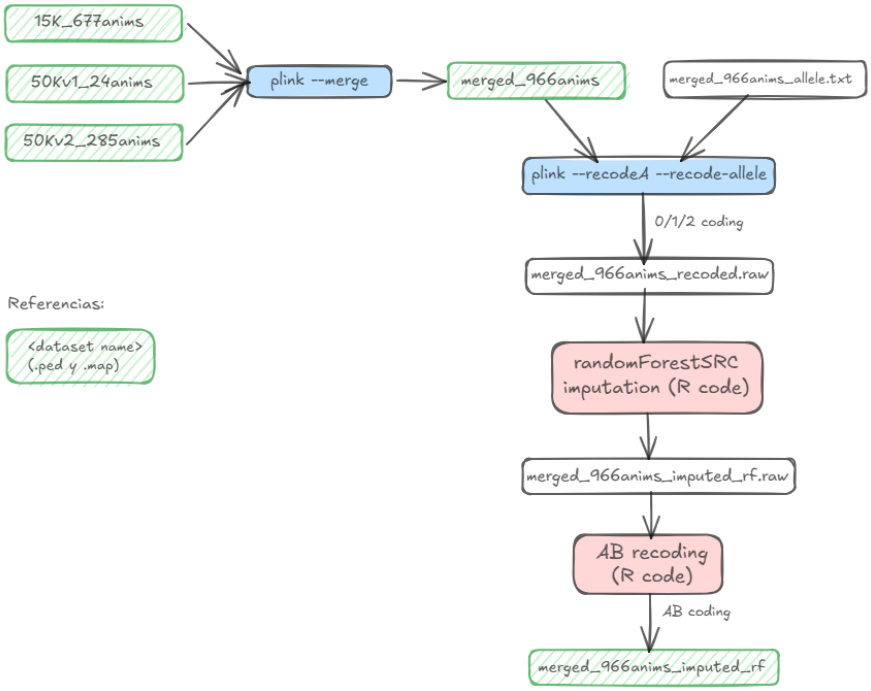

Fig. 2. Flujo de trabajo que describe el procesamiento de archivos para su imputación.

Las estrategias de imputación empleadas se seleccionaron en función del error de imputación que presentaron durante el procedimiento y los tiempos de procesamiento en el hardware utilizado. No obstante, se probaron otras implementaciones en R de algoritmos de machine learning de imputación de datos faltantes. Entre ellas, cabe mencionar el método Multivariate Imputation by Chained Equations (MICE; van Buuren y Groothuis-Oudshoorn, 2011) y una implementación basada en modelos k-nearest neighbour (Hastie et al., 2001). Sin embargo, para nuestro dataset combinado el tiempo de procesamiento de estas dos implementaciones resultó prohibitivo debido a la alta cantidad de datos faltantes.

2.3 Comprobación de la precisión de la imputación

La comprobación de la precisión en la imputación se realizó a través de la comparación de un subconjunto de los genotipos imputados con genotipos obtenidos con el microarreglo triespecie de Affymetrix (Axiom™ Bovine-Ovine-Caprine Genotyping Array). Los genotipos comparados correspondían a 232 animales en 30.924 SNPs. Esos

232 animales habían sido genotipificados con el microarreglo 15K, pero no con los microarreglos 50Kv1 ni 50Kv2.

A partir de los tres datasets de salida (imputados), mediante el programa PLINK (funciones keep y extract), se extrajo el subconjunto de genotipos a comparar, uno por cada estrategia de imputación. La notación alélica A/B se convirtió a notación nucleotídica mediante la función update-alleles de PLINK, para permitir la comparación con el mismo subconjunto de genotipos extraído a partir de los archivos resultantes de la genotipificación con el microarreglo de Affymetrix, con alelos expresados en notación nucleotídica. El archivo de genotipos resultantes de la genotipificación con el microarreglo de Affymetrix poseía una tasa de genotipificación del 99,56%. La comparación se realizó utilizando el modo merge-mode 7 de la función merge de PLINK, la cual detecta genotipos asignados no coincidentes entre los archivos comparados (imputado vs. genotipificado), pudiéndose determinar a continuación el porcentaje de concordancia.

La precisión de la imputación se calculó, para cada animal, como el número de genotipos imputados correctamente dividido por el número total de genotipos imputados. Mientras que la precisión de la imputación para cada SNP se calculó como el número de animales con genotipos imputados correctamente dividido por el número total de animales imputados.

3 Resultados y discusión

Todos los SNPs con genotipo no asignado del dataset combinado fueron imputados, es decir, los evaluados por el microarreglo de 15K y/o los microarreglos de 50K (v1 y v2). Luego de la imputación no quedaron genotipos no asignados en el dataset. Se concluye entonces que las tres estrategias de imputación seguidas condujeron a una imputación completa. Por otro lado, de los 7.174.368 (=232*30.924) genotipos del subconjunto de genotipos obtenidos con el microarreglo de Affymetrix que se analizó, 7.142.988 eran genotipos asignados, no faltantes. Estos genotipos se utilizaron para la comparación.

El resultado de la comparación de los genotipos obtenidos por genotipificación con los imputados a partir de la estrategia #1, que llevó a cabo un máximo de 2 iteraciones de imputación en bloques de 1.000 SNPs, fue una concordancia del 60,4%. Por otro lado, al comparar los genotipos imputados a partir de la estrategia #2, en la que se aumentó a 5 el máximo número de iteraciones permitidas, manteniendo constantes los demás parámetros y el tamaño del bloque de imputación, la coincidencia con los genotipos obtenidos por genotipificación fue de un 60,9%. Aunque se esperaba que la estrategia #2 condujera a menos errores de imputación que la #1 por ejecutar un máximo de 5 iteraciones en lugar de 2 hasta que el error de imputación fuera menor que un valor por defecto, se observa que las concordancias obtenidas tras la comparación fueron muy similares para ambas estrategias. Finalmente, se esperaba que la estrategia #3 disminuyera el error de la estrategia #2 ya que se ejecutó utilizando los mismos parámetros que en la estrategia #2 pero agrupando los SNPs por cromosoma, con lo que se esperaba aportar más información. Sin embargo, el resultado de concordancia

tras la comparación con los resultados de la genotipificación por microarreglo fue de un 59,8%, aunque similar, la menor de las tres estrategias seguidas (Tabla 2).

Tabla 2. Rangos de precisión en la imputación y porcentajes de concordancia entre los genotipos imputados por tres estrategias de machine learning y los obtenidos por genotipificación con microarreglo.

# estrategia de imputación	Concordancia (%)	Rango de precisión en la imputación por animal	Rango de precisión en la imputación por SNP
1	60,4	0,55 - 0,70	0,12 - 0,97
2	60,9	0,55 - 0,71	0,12 - 0,98
3	59,8	0,55 - 0,71	0,12 - 0,97

El porcentaje de concordancia obtenido para las tres estrategias, aproximadamente de un 60%, puede atribuirse a la gran cantidad de genotipos no asignados del archivo de partida. En general, a mayor cantidad de datos faltantes en el dataset, menor es la calidad de la imputación. Un 53,38% de datos faltantes es una cantidad muy alta si se espera un buen desempeño en la imputación (tasa de error de imputación menor al 20%). Los algoritmos de imputación de datos de machine learning tienen, en general, un buen desempeño con una proporción de datos faltantes menor al 5% (Abolhassani Targhi et al., 2019), aunque esto puede variar en función a la naturaleza u origen de los datos faltantes.

Los rangos de precisión en la imputación, evaluada tanto en animales como en SNPs, fueron muy similares para las tres estrategias empleadas. El número de animales con una precisión en la imputación superior a la media + 2 DS fue 5, 5 y 7 para las estrategias #1, #2 y #3, respectivamente. Por otro lado, el número de SNPs con una precisión en la imputación superior al 80% fue similar para las tres estrategias: 3740, 3836 y 3708 para las estrategias #1, #2 y #3, respectivamente. En todos los casos, los cromosomas que presentaron mayor número de SNPs con alta precisión en la imputación se correspondieron con aquellos para los cuales se había imputado un mayor número de SNPs.

En un trabajo publicado hace unos años, Hayes et al. estudiaron la precisión de la imputación de genotipos desde paneles de SNPs de baja densidad a 50K, en distintas razas ovinas, utilizando métodos de imputación basados en el desequilibrio de ligamiento de la población (Hayes et al., 2012). Las precisiones obtenidas variaron entre 61 (para la raza Merino) y 81% (para la raza Border Leicester). Asimismo, evaluaron la influencia de la densidad de SNPs de los paneles de partida y de la diversidad genética intrarracial sobre las precisiones obtenidas. Hallaron una baja precisión de imputación usando paneles de menos de 5.000 SNPs en la población target. En nuestro dataset, el panel de la población target compartía más de 11.200 SNPs con los microarreglos con los que se genotipificó la población de referencia. Por otro lado, Hayes et al. hallaron mejores precisiones de imputación cuanto menor era la distancia genética en la población bajo estudio y cuanto más emparentados estaban los animales de la población target con la de referencia. En nuestro caso, los animales bajo estudio

son de la misma raza, pertenecen a la misma población y están emparentados. Sin embargo, en el presente trabajo, la población genotipificada con los microarreglos de 50K representa menos de la mitad de la genotipificada con el microarreglo de 15K, mientras que en el trabajo mencionado previamente la población de referencia representaba el 75% de la población bajo estudio. Esta gran diferencia podría explicar la obtención de precisiones comparables sólo a los valores inferiores del rango de precisiones de imputación obtenido por Hayes et al.

4 Conclusión

En este trabajo se investigó la precisión en la imputación de genotipos mediante algoritmos de machine learning que utilizaron como target una población de ovinos genotipificada con un microarreglo que evalúa 14.943 SNPs y, como referencia, animales emparentados con los anteriores (y en algunos casos los mismos animales) genotipificados con microarreglos que evalúan alrededor de 53.000 SNPs. Las tres estrategias de imputación seguidas condujeron a una precisión del 60% aproximadamente, la cual es relativamente baja para utilizar los genotipos obtenidos en futuros análisis como, por ejemplo, para selección genómica. Una estrategia para aumentar la precisión de la imputación podría ser aumentar el número de animales en la población de referencia y, de este modo, reducir la proporción de genotipos faltantes en el conjunto de datos completo.

Referencias

Abolhassani Targhi, M. V., Asgari Jafarabadi, G., Aminafshar, M. y Emam Jomeh Kashan, N. (2019). The effect of genotype imputation and some important factors on the accuracy of genomic prediction and its persistency over time. Gene Reports, 16, 100425.

Browning, S. R. y Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics, 81, 1084-1097.

Browning, B. L. y Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American Journal of Human Genetics, 84, 210-223.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M. y Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Fast Unified Random Forests with randomForestSRC 3.3.1, Impute Only Mode Homepage, https://www.randomforestsrc.org/reference/impute.rfsrc.html, last accessed 2025/03/31.

Fernando, R. y Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution, 21, 467.

Ghavi Hossein-Zadeh, N. (2024). An overview of recent technological developments in bovine genomics. Veterinary and animal science, 25, 100382.

Hastie, T., Tibshirani, R., Eisen, M., Brown, P. y Botstein, D. (2001). Imputing Missing Data for Gene Expression Arrays. Technical report, Stanford Statistics Department. 1.

Hayes, B. J., Bowman, P. J., Daetwyler, H. D., Kijas, J. W. y van der Werf J. H. (2012). Accuracy of genotype imputation in sheep breeds. Animal Genetics, 43(1), 72-80.

Howie, B. N., Donnelly, P. y Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics, 5, e1000529.

Ishwaran, H., Lu, M. y Kogalur, U. B. (2021). "randomForestSRC: getting started with randomForestSRC vignette." http://randomforestsrc.org/articles/getstarted.html.

Jewett, E. M., Zawistowski, M., Rosenberg, N. A. y Zöllner, S. (2012). A coalescent model for genotype imputation. Genetics, 191(4), 1239-1255.

Li, Y., Willer, C., Sanna, S. y Abecasis, G. (2009). Genotype imputation. Annual Review of Genomics and Human Genetics, 10, 387-406.

Marina, H. Chitneedi, P., Pelayo, R., Suárez-Vega, A., Esteban-Blanco, C., Gutiérrez-Gil, B. y Arranz J. J. (2021). Study on the concordance between different SNP-genotyping platforms in sheep. Animal Genetics, 52(6), 868-880.

Meuwissen, T. H., Hayes, B. J. y Goddard, M. E. (2001). Prediction of total genetic value using genome wide dense marker maps. Genetics, 157, 1819-1829.

Mikhchi, A., Honarvar, M., Kashan, N.E. y Aminafshar, M. (2016). Assessing and comparison of different machine learning methods in parent-offspring trios for genotype imputation. Journal of Theoretical Biology, 399, 148-158.

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G. Y. y Myles, S. (2015). LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genetics, 5(11), 2383-2390.

Mora-Márquez, F., Nuño, J. C., Soto, Á. y López de Heredia, U. (2025). Missing genotype imputation in non-model species using self-organizing maps. Molecular Ecology Resources, 25(3), e13992.

Naito, T. y Okada, Y. (2024). Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology. Journal of Human Genetics, 69, 481-486.

Nicolazzi, E. L., Biffani, S. y Jansen, G. (2013). Short communication: imputing genotypes using PedImpute fast algorithm combining pedigree and population information. Journal of Dairy Science, 96, 2649-2653.

Ogunbawo, A. R., Mulim, H. A., Campos, G. S., Schinckel, A. P. y Oliveira, H. R. (2024). Tailoring Genomic Selection for Bos taurus indicus: A Comprehensive Review of SNP Arrays and Reference Genomes. Genes, 15, 1495.

PLINK 1.9 Homepage, https://www.cog-genomics.org/plink/1.9/, last accessed 2025/03/31.

Ragoussis J. (2009). Genotyping technologies for genetic research. Annual Review on Genomics and Human Genetics, 10, 117-133.

Sargolzaei, M., Chesnais, J. P. y Schenkel, F. S. (2014). A new approach for efficient genotype imputation using information from relatives. BMC Genomics, 15(1), 478.

Scheet, P. y Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics, 78, 629-644.

Schwarz, D. F., Szymczak, S., Ziegler, A. y König, I. R. (2009). Evaluation of single-nucleotide polymorphism imputation using random forests. BMC Proceedings, 3(7), S65.

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., Donnadieu-Tonon, C., Eggen, A., Heuven, H. C., Jamli, S., Jiken, A. J., Klopp, C., Lawley, C. T., McEwan, J., Martin, P., Moreno, C. R., Mulsant, P., Nabihoudine, I., Pailhoux, E., Palhière, I., Rupp, R., Sarry, J., Sayre, B. L., Tircazes, A., Wang, J., Wang, W., Zhang, W. e International Goat Genome Consortium. (2014). Design and characterization of a 52K SNP chip for goats. PLoS One, 9(1):e86227. Erratum in: PLoS One 11(3), e0152632 (2016)

van Buuren, S. y Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

Agradecimientos. Agradecemos la participación de personal de las EEAs Mercedes y Concepción del Uruguay de INTA por criar los animales y brindarnos acceso a bases de datos de registros fenotípicos y genealógicos y a muestras biológicas para realizar los análisis. Este trabajo fue financiado por proyectos del Instituto Nacional de Tecnología Agropecuaria (PD I108, PD I115 y PT I180) y de la división conjunta FAO-IAEA (Organización de las Naciones Unidas para la Alimentación y la Agricultura - Agencia Internacional de Energía Atómica; CRP D3.10.30).

Declaración de conflicto de intereses. Los autores no tienen conflictos de intereses que declarar que sean relevantes para el contenido de este artículo.