Concordance between genotype imputation methods

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]], Daniel O. Maizon[3,7[0000-0002-2701-4109]], Daniel Demitrio[4,6[0009-0008-3372-3340]], y Mario A. Poli[1,8[0000-0001-8775-2333]]

[1] Instituto Nacional de Tecnología Agropecuaria, 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; poli.mario@inta.gob.ar

[2] Universidad de Buenos Aires, Buenos Aires, Argentina.pablo.javier.rios@gmail.com

[3] Instituto Nacional de Tecnología Agropecuaria, E.E.A. Anguil, Ruta 5 Km 580, Anguil (6326), La Pampa, Argentina maizon.daniel@inta.gob.ar

[4] Instituto Nacional de Tecnología Agropecuaria, 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 Agronomía, Universidad Nacional de La Pampa, Argentina.

[8] Facultad de Ciencias Agrarias y Veterinaria, Universidad del Salvador, Argentina.

Abstract. Genotype imputation improves genomic coverage by combining data obtained from microarrays of different densities or platforms and inferring unassigned genotypes. This study compared three imputation strategies using a dataset of 1,153 Corriedale sheep genotyped for 54,718 SNPs, with 28.21% missing data: Random Forest (RF), Sparse Convolutional Denoising Autoencoders (SCDA), and FImpute. Random Forest was implemented using the *impute()* method, in 28 blocks corresponding to markers with undefined genomic positions, autosomes, and the X chromosome. The SCDA model was trained using a training (80% of the population) and a validation (20% of the population) partitions over 40 *epochs*, selecting the model with the lowest loss and reasonable accuracy in the validation partition. FImpute was run on all markers with defined positions, incorporating pedigree information. All three strategies achieved complete imputation of missing genotypes; FImpute additionally corrected 0.13% of assigned genotypes by detecting Mendelian errors. Comparison of the imputed genotypes showed higher concordance between RF and FImpute (63.54%) than between SCDA and FImpute (52.02%), with uniform values across autosomes and slightly higher concordance on the X chromosome. Assuming greater accuracy for FImpute, RF outperformed SCDA in this dataset. These results support further exploration of these strategies using datasets with varying proportions of missing data and masking some assigned genotypes to evaluate imputation accuracy.

Keywords: imputation, genotypes, random forest, autoencoders.

Concordancia entre métodos de imputación de genotipos

Resumen. La imputación de genotipos permite mejorar la cobertura genómica al combinar datos obtenidos con microarreglos de distinta densidad o plataforma e inferir genotipos no asignados. Este estudio comparó tres estrategias de imputación utilizando un dataset compuesto por genotipos de 1.153 ovinos Corriedale en 54.718 SNPs, con un 28,21% de datos faltantes: Random Forest (RF), autoencoders convolucionales dispersos para la eliminación de ruido (SCDA) y FImpute. Random forest se implementó a través del método *impute()*, ejecutándolo en 28 bloques correspondientes a marcadores con posición indefinida en el genoma, autosomas y cromosoma X. El modelo SCDA se entrenó usando una partición de training (80% de la población) y otra de validación (20% de la población) durante 40 *epochs*, seleccionándose el modelo que presentó la menor pérdida y una precisión razonable en la partición de validación. FImpute se ejecutó sobre todos los marcadores con posición definida, considerando la información de pedigree. Las tres estrategias condujeron a una imputación completa de los genotipos faltantes; FImpute, además, corrigió un 0,13% de genotipos asignados por hallar errores mendelianos. La comparación de los genotipos imputados mostró mayor concordancia entre RF y FImpute (63,54%) que entre SCDA y FImpute (52,02%), con valores uniformes en los autosomas y ligeramente superiores en el cromosoma X. Asumiendo una mayor precisión para FImpute, RF mostró mejor desempeño que SCDA en el dataset empleado. Estos resultados motivan continuar explorando estas estrategias, utilizando datasets con distintos porcentajes de datos faltantes y enmascarando algunos genotipos para evaluar la precisión en las imputaciones logradas.

Palabras clave: imputación, genotipos, random forest, autoencoders.

1 Introducción

La disponibilidad y accesibilidad a servicios de genotipificación de ADN de ganado ha ido en aumento en las últimas décadas, en paralelo a las aplicaciones de la información genotípica en la producción y el mejoramiento animal. El desarrollo de tecnologías de genotipificación de polimorfismos de nucleótido único (SNPs) de alto rendimiento ha permitido reducir significativamente los costos de obtención de información genotípica individual para decenas de miles hasta millones de marcadores. Una herramienta muy utilizada en el manejo de información genotípica es la imputación, es decir, la inferencia de genotipos no asignados o en marcadores no evaluados. Constituye una herramienta poderosa para ampliar la cobertura genómica cuando se pretende combinar genotipos obtenidos con distintas plataformas o paneles, siempre que exista un solapamiento de marcadores suficiente entre ellos, así como para inferir genotipos no asignados con el fin de mejorar la tasa de genotipificación. Una estrategia de imputación que permite reducir los costos de genotipificación consiste en genotipificar una población de animales jóvenes candidatos a selección con un panel de baja densidad de marcadores, e imputar los genotipos de los loci no evaluados utilizando

información de una población de referencia genotipificada con un panel de mayor densidad (Berry y Kearney, 2011). Existen programas específicos de imputación que tienen en cuenta relaciones de parentesco y principios de herencia mendeliana y métodos que emplean información de desequilibrio de ligamiento a nivel poblacional para la inferencia de genotipos (Li et al., 2009). Entre los más utilizados en ganado se encuentran BEAGLE (Browning y Browning, 2007, 2009), IMPUTE2 (Howie et al., 2009) y FImpute (Sargolzaei et al., 2014). Estos programas requieren utilizar un panel de referencia, el que contiene información de haplotipos de múltiples muestras, generalmente provenientes de la misma población que se desea imputar o de una similar. Otras estrategias reportadas para la imputación de genotipos faltantes emplean algoritmos de machine learning, entre ellos, random forest (Schwarz et al., 2009; Mikhchi et al., 2016; Faux et al., 2019) y autoencoders convolucionales dispersos para la eliminación de ruido (Chen y Shi, 2019; Song et al., 2022). En particular, los algoritmos de *deep learning* han tenido un gran impacto en varias áreas, como bioinformática, por su capacidad para manejar grandes conjuntos de datos y modelar relaciones altamente no lineales (Naito y Okada, 2024).

El objetivo de este trabajo fue comparar los resultados de la imputación de genotipos realizada empleando los algoritmos de machine learning random forest y autoencoders convolucionales dispersos para la eliminación de ruido (Sparse Convolutional Denoising Autoencoders, SCDA), con los obtenidos de la imputación realizada con el programa FImpute.

2 Materiales y métodos

2.1 Base de datos genotípicos

Se trabajó con un único conjunto de datos de entrada, consistente en genotipos de 1.153 ovinos de la raza Corriedale en 54.718 polimorfismos de nucleótido simple (SNPs), con una tasa de genotipificación del 71,79%, es decir, un porcentaje de datos faltantes del 28,21%. Este archivo resultó de la combinación de cinco bases de datos genotípicos obtenidas tras genotipificar ovinos de raza Corriedale con microarreglos de distinta densidad de las plataformas Illumina y Affymetrix. Las bases de datos obtenidas con la plataforma Illumina fueron tres y consistieron en genotipos de 677 animales en 15.000 SNPs, de 24 animales en 54.241 SNPs, y de 285 animales en 53.516 SNPs, evaluados por los microarreglos 15K, 50Kv1 y 50Kv2, respectivamente. Mientras que las bases de datos obtenidas con la plataforma Affymetrix fueron dos y consistieron en genotipos de 768 animales en 39.747 SNPs y de 291 animales en 42.769 SNPs, evaluados por los microarreglos 50K y 60K, respectivamente. A cada una de estas bases de datos se le realizó un control de calidad que consistió en descartar los individuos con tasa de genotipificación < 95%, con una heterocigosidad observada mayor a 3 DS de la media, y los SNPs duplicados, con tasa de genotipificación <95%, con frecuencia alélica mínima <0,03 y con un valor p para la prueba del equilibrio de Hardy-Weinberg $<1.10^{-6}$. Luego del control de calidad de genotipos y previamente a realizar la combinación de las cinco bases de datos, se analizó la equivalencia entre los SNPs evaluados por las plataformas Illumina y Affymetrix a través de la determinación de su posición referida al mismo ensamblado y estudio de su secuencia flanqueante. Además, mediante la

función --flip del programa PLINKv1.9 (Chang et al., 2015) se intercambiaron los alelos A↔T y C↔G en aquellos SNPs que habían sido evaluados por los microarreglos de ambas plataformas en hebras complementarias del ADN. Finalmente, la combinación de las bases de datos se realizó utilizando las funciones --merge, --bmerge y --merge-list del programa PLINKv1.9. Para la combinación se utilizó el modo consenso de estas funciones 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. En todos los casos se utilizó la notación nucleotídica para designar los alelos. En la Tabla 1 puede observarse el porcentaje de animales genotipificados con los distintos microarreglos o combinaciones de microarreglos, mientras que en la Tabla 2 se indica el porcentaje de SNPs compartidos o evaluados por los distintos microarreglos. Del total de animales evaluados, un 60,6% fue genotipificado con un único microarreglo, un 38,5% con dos microarreglos y un 0,9% con tres microarreglos. Del total de SNPs evaluados, más del 65% son compartidos por cuatro o cinco microarreglos, mientras que sólo un 6% fue evaluado por un único microarreglo.

Tabla 1. Porcentaje de animales genotipificados con los distintos microarreglos o combinaciones de microarreglos.

microarreglos	% de animales
50Kv2	23,59
15K, 60K	19,08
50K	19,08
15K, 50K	17,95
15K	13,53
60K	4,08
15K, 50Kv1, 60K	0,87
15K, 50Kv1	0,43
50Kv1, 60K	0,43
50Kv1	0,35
50Kv2, 60K	0,35
15K, 50Kv2	0,26

Los microarreglos 15K, 50Kv1 y 50Kv2 corresponden a la plataforma Illumina, mientras que los microarreglos 50K y 60K corresponden a la Affymetrix.

Los alelos del dataset combinado, expresados con notación nucleotídica, fueron recodificados mediante la función --recode12 de PLINKv1.9 para ser expresados en notación 1/2/0, donde 1 y 2 representan los distintos alelos posibles para cada SNP y 0 indica que el alelo no fue asignado. Luego, se utilizaron las funciones --recodeA junto con --recode-allele para obtener la codificación genotípica aditiva 0/1/2 requerida por los algoritmos de machine learning computando, para todos los SNPs, el alelo codificado como "1". El archivo combinado, con genotipos de 1.153 animales en 54.718 SNPs, expresados en codificación 0/1/2, se utilizó como input para los algoritmos de machine learning utilizados para realizar la imputación. Este archivo presentaba 28,21% de genotipos no asignados, correspondientes a SNPs que no fueron evaluados por algún microarreglo o que fueron evaluados pero resultaron no asignados por algún problema en el proceso de genotipificación.

Tabla 2. Porcentaje de SNPs compartidos o evaluados por los distintos microarreglos.

microarreglos	% de SNPs
50Kv1-50Kv2-50K-60K	48,89
15K-50Kv1-50Kv2-50K-60K	14,27
50Kv1-50Kv2	13,58
50Kv1-50Kv2-60K	8,41
15K	5,70
15K-50Kv1-50Kv2	3,53
15K-50Kv1-50Kv2-60K	2,11
50Kv1-50Kv2-50K	1,64
50Kv1-50K-60K	0,74
15K-50Kv1-50Kv2-50K	0,44
50Kv1	0,21
15K-50Kv1-50K-60K	0,18
50Kv1-60K	0,10
50K	0,05
15K-50Kv1	0,05
50Kv1-50K	0,03
15K-50Kv1-60K	0,03
50Kv2	0,02
60K	0,01
15K-50Kv1-50K	0,01
50K-60K	0,00

Los microarreglos 15K, 50Kv1 y 50Kv2 corresponden a la plataforma Illumina, mientras que los microarreglos 50K y 60K corresponden a la Affymetrix.

2.2 Estrategias de imputación

Imputación mediante el algoritmo de random forest. Se utilizó el algoritmo de random forest (Ishwaran et al., 2021) usando la implementación en R versión 4.5.2 del paquete randomForestSRC versión 3.4.5 para imputar datos faltantes. Se realizaron múltiples imputaciones de exploración variando los siguientes parámetros del método *impute()* (Impute Only Mode, Fast Unified Random Forests with randomForestSRC, 2026): *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). Considerando 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), se decidió realizar una imputación por cromosoma (28 bloques, del 0 al 27, constituidos por los SNPs de un mismo cromosoma, o SNPs con posición indefinida en el genoma), utilizando los siguientes parámetros: *max.iter*=5, *ntree*=100, *nodesize*=5, *mf.q*=0.2, *fast*=FALSE. Cada imputación consistió en ejecutar el método *impute()* de a bloques contiguos de SNPs dado que la complejidad computacional de este algoritmo es cuadrática en función del número de SNPs (O(N2)). En el bloque 0 se incluyeron los SNPs sin posición definida, en los bloques 1 al 26 los SNPs de los autosomas ovinos, y en el 27 los SNPs ubicados en el cromosoma X. En la Figura 1 se indica la cantidad de SNPs por bloque. El output

o resultado de la imputación, inicialmente con codificación genotípica aditiva 0/1/2 fue recodificado a notación alélica A/B. Se imputaron todos los genotipos faltantes del dataset combinado.

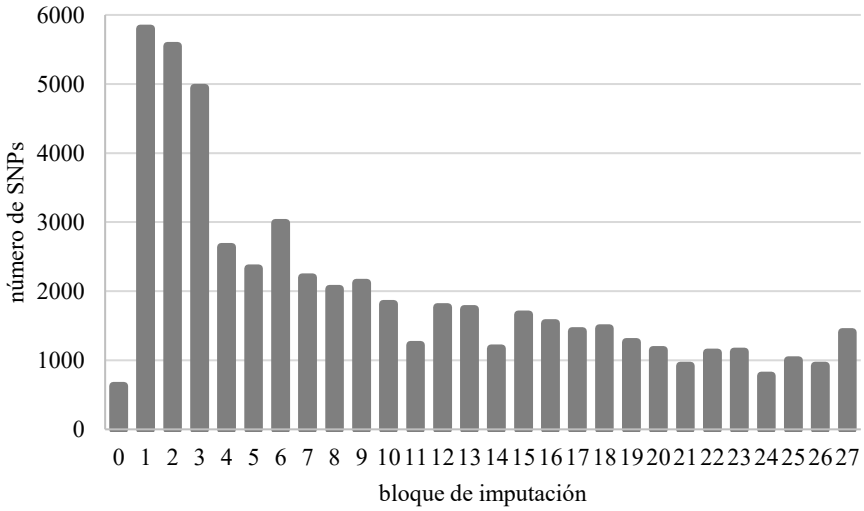

Fig. 1. Cantidad de SNPs por bloque de imputación. El bloque 0 contiene los SNPs sin posición definida en el genoma, los bloques 1 a 26 los SNPs de los autosomas ovinos, y el 27 los SNPs ubicados en el cromosoma X.

Imputación mediante autoencoders convolucionales dispersos para la eliminación de ruido. En base al trabajo de Chen y Shi (Chen y Shi, 2019), se implementó un modelo de *deep learning* llamado autoencoder de convolución disperso para la eliminación de ruido (SCDA) para imputar genotipos faltantes. El modelo SCDA utiliza capas convolucionales capaces de detectar patrones locales y relaciones espaciales (como el desequilibrio de ligamiento) entre marcadores genéticos y aplica una matriz de pesos dispersos derivada de la regularización L1 para manejar datos de alta dimensión donde muchos valores son cero, lo que optimiza el manejo de grandes volúmenes de datos genómicos. La eliminación de ruido a la que hace referencia la técnica implica que el modelo aprende a reconstruir datos originales a partir de versiones incompletas, lo cual es esencial para la imputación.

Para la imputación de genotipos utilizando el modelo SCDA, en primer lugar, se entrenó el modelo usando una partición de training (80% de la población, o 922 animales) y otra de validación (20% de la población, o 231 animales). Al 10% de los SNPs de la partición de training que tenían genotipo asignado (no faltante) se los enmascaró como valores faltantes (ruido), y es lo que se intentó reconstruir minimizando una función de pérdida como parte del entrenamiento. La función de pérdida utilizada fue la *categorical crossentropy*. En cada *epoch* del entrenamiento se computó la pérdida (*loss*) y la precisión (*accuracy*) tanto en el 10% de los SNPs de la partición de training, como en todos los SNPs con valores no faltantes de la partición de validación. El entrenamiento se realizó durante 40 *epochs*, seleccionándose el

modelo que presentó la menor pérdida y una precisión razonable en la partición de validación, para predecir los valores faltantes del conjunto de datos utilizados como input (imputación del 28,21% de los genotipos del dataset). El entrenamiento del modelo y posterior imputación se realizaron utilizando un procesador CPU Intel® CoreTM i7-7820X CPU @ 3.60GHz; 32 GB de memoria RAM; disco NVMe SSD de 500 GB y una GPU NVIDIA GeForce GTX 1080 Ti (11 GB). El output, inicialmente con codificación genotípica aditiva 0/1/2 fue recodificado a notación alélica A/B. Se imputaron todos los genotipos faltantes del dataset combinado.

Imputación mediante FImpute. La tercera estrategia de imputación empleada consistió en el uso del programa FImpute versión 2.2, que utiliza tanto información familiar como poblacional. Inicialmente, emplea la información de pedigree para realizar la determinación de fases e imputación de genotipos mediante un enfoque iterativo. Posteriormente, los genotipos faltantes remanentes se imputan utilizando una estrategia de ventanas deslizantes superpuestas. Este enfoque en primer lugar captura la información de parientes cercanos y, en sucesivos barridos cromosómicos, a medida que se reduce progresivamente el tamaño de las ventanas, va incorporando información de parientes más distantes para la construcción de una librería de haplotipos para cada ventana (Sargolzaei et al., 2014). El método se basa en el concepto de que los parientes cercanos comparten haplotipos largos, mientras que los parientes más distantes comparten haplotipos más cortos.

En este caso se excluyeron del archivo de entrada los SNPs sin posición definida, por lo que sólo se imputaron los genotipos faltantes de SNPs localizados en autosomas y en el cromosoma X. El input consistió, entonces, en genotipos de 1.153 animales en 54.087 SNPs, con un 28,05% de genotipos faltantes. Por otro lado, se proporcionó al programa un archivo con información de pedigree. La población a imputar está estructurada en 714 tríos padre-madre-hijo con padre y madre genotipificados, 342 dúos padre-hijo con padre genotipificado, 44 dúos madre-hijo con madre genotipificada y 53 animales sin padre ni madre genotipificados. Los animales con padre conocido (1.064) son hijos de 52 carneros, con entre 66 y 1 cría cada uno. El proceso de imputación requirió un tiempo de ejecución de 21 min 25 s en una computadora equipada con un procesador Intel® Pentium(R) 4 de 3,20 GHz, 3,2 GiB de memoria RAM y un disco rígido de 245,1 GB. El sistema operativo utilizado fue Ubuntu 18.04.1 LTS, con entorno de escritorio GNOME 3.28.2. El output o resultado de la imputación, inicialmente con codificación genotípica aditiva 0/1/2, fue recodificado a notación alélica A/B y luego adaptado al formato PLINK.

2.3 Comparación de la imputación lograda mediante random forest, SCDA y FImpute

Los outputs (archivos con genotipos imputados) obtenidos a partir de las tres estrategias empleadas, inicialmente en notación alélica A/B, se recodificaron, a través de la función --update-alleles de PLINKv1.9, a notación nucleotídica. Los archivos imputados se compararon, de a pares, mediante las funciones --bmerge y --merge-mode 6 de PLINKv1.9, calculándose la concordancia entre los genotipos imputados (no asignados en el archivo original) por cada par de métodos. Al comparar los genotipos imputados por FImpute vs los imputados por los algoritmos de machine learning empleados, se

analizaron únicamente los SNPs con ubicación conocida. En todos los casos, la concordancia de genotipos imputados se calculó para todos los SNPs del dataset en su conjunto, así como por cromosoma (los SNPs de cada cromosoma se extrajeron con la función --chr de PLINKv1.9).

3 Resultados y discusión

En este trabajo se utilizaron tres estrategias de imputación de genotipos faltantes, dos de las cuales, random forest (RF) y un algoritmo de deep learning consistente en un autoencoder, no requieren un panel de referencia, mientras que la tercera, llevada a cabo a través del empleo del programa FImpute, sí lo requiere. El material de partida consistió en cinco bases de datos genotípicos obtenidas a partir de la genotipificación con microarreglos de distinta densidad de las plataformas Illumina y Affymetrix. Luego de realizar el control de calidad y descartar los animales y SNPs que no alcanzaron los umbrales preestablecidos, en las bases de datos correspondientes a los microarreglos 15K, 50Kv1, 50Kv2, 50K y 60K permanecieron genotipos de 601 animales en 14.141 SNPs, 24 animales en 49.273 SNPs, 279 animales en 49.230 SNPs, 427 animales en 35.696 SNPs y 286 animales en 40.210 SNPs, respectivamente. La combinación de estos archivos condujo a la obtención del archivo de genotipos utilizado como input para la imputación con algoritmos de machine learning, consistente en genotipos de 1.153 animales en 54.718 SNPs, con un 28,21% de datos faltantes. De este archivo se excluyeron los 631 SNPs con posición indefinida, obteniendo el archivo de genotipos de 1.153 animales en 54.087 SNPs que, previa conversión de formato, se utilizó como input del programa FImpute.

En relación al algoritmo random forest, el criterio que se utilizó para la selección de parámetros del método *impute()* se basó en considerar el menor error de imputación y tiempos de procesamiento que no excedieran las 8 horas. En el caso del SCDA, el modelo seleccionado para la imputación fue el de la *epoch* 30 por presentar la menor pérdida y una precisión mayor al 70% en la partición de validación (Figura 2). La pérdida (*loss* o *categorical crossentropy*) mide qué tan confiables son las predicciones del modelo, penalizando fuertemente las predicciones incorrectas que se hicieron con un alto nivel de confianza, mientras que la precisión (*accuracy*) sólo mide si la categoría final predicha es correcta o no, sin considerar el nivel de confianza de la predicción. En la Figura 2 se observa que en la partición de validación la pérdida es más baja y la precisión es más alta que en la partición de entrenamiento. Esto se explica por las técnicas de regularización que se utilizaron, en particular, las capas de *Dropout* de la red. El *Dropout* sólo se activa durante el entrenamiento. La capa *Dropout* desactiva aleatoriamente una fracción de las neuronas (25% en la red propuesta) durante cada paso del entrenamiento. Esto hace que la tarea de entrenamiento sea intencionalmente más difícil para el modelo, evitando que el modelo desarrolle una dependencia excesiva de ciertos “caminos neuronales” o patrones de activación muy específicos. Al forzar a la red a utilizar un subconjunto diferente de neuronas en cada paso del entrenamiento, se asegura que aprenda características más robustas y generalizables. En consecuencia, el ajuste del modelo a la partición de entrenamiento se ve afectado negativamente. Por otro lado, la validación utiliza la red completa. Cuando el modelo se evalúa en la partición de validación al término de cada *epoch*, la capa *Dropout* se desactiva. El

modelo puede usar su red completa y ya entrenada, con todas las neuronas activas. Debido a que ha aprendido características robustas, tiene un muy buen rendimiento en los nuevos datos de validación, a menudo incluso mejor del que tuvo en la partición de entrenamiento. Finalmente, la regularización L1 agrega una penalización a la función de pérdida en la partición de entrenamiento, la cual no se incluye al calcular la pérdida en la partición de validación.

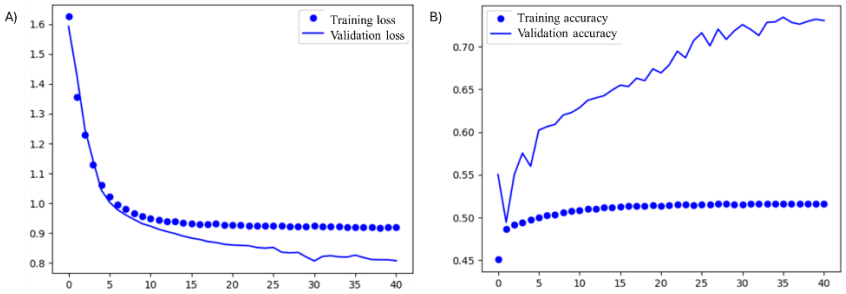

Fig. 2. A) Función de pérdida a lo largo de las 40 *epochs* evaluadas, en las particiones de entrenamiento y validación. B) Precisión a lo largo de las 40 *epochs* evaluadas, en las particiones de entrenamiento y validación.

Las tres estrategias de imputación seguidas condujeron a una imputación completa, dado que todos los SNPs con genotipo no asignado del dataset combinado utilizado como input fueron imputados. Luego de la imputación no quedaron genotipos no asignados en el dataset. Los algoritmos random forest y SCDA no modificaron los genotipos asignados del archivo de entrada, mientras que el programa FImpute modificó el 0,13% de los genotipos asignados del archivo a imputar (57.493 de 44.869.852 genotipos). Este comportamiento de FImpute se debe a que encontró, y corrigió, errores mendelianos en la población. Del análisis de los errores mendelianos hallados se desprende que el animal con mayor cantidad de errores presentó 866, 101 animales presentaron entre 100 y 434 errores, y 986 animales entre 99 y 1 error. Aunque una gran cantidad de errores mendelianos pueden sugerir conflictos de paternidad, un bajo número de errores es aceptable y atribuible a errores en la asignación de genotipos durante el proceso de genotipificación. FImpute corrige estos errores asignando un genotipo compatible con los de los miembros de la familia con los que se presentó el error. En este trabajo estas reasignaciones de genotipos no se consideraron para la comparación dado que los algoritmos de machine learning empleados, al no trabajar con información de parentesco, no detectan conflictos de paternidad. Por lo tanto, en las comparaciones realizadas sólo se estudiaron los genotipos no asignados en el archivo original (input), siendo éstos 17.795.622 para el caso de las estrategias que emplearon algoritmos de machine learning, y 17.492.459 en la que empleó el programa FImpute.

De los 17.795.622 genotipos imputados por los algoritmos RF y SCDA, 9.561.192 son concordantes, lo que representa un 53,73% de concordancia en la imputación por ambas estrategias. Por otro lado, al comparar los resultados obtenidos con FImpute vs. RF, se encontraron 11.114.469 genotipos concordantes de 17.492.459 imputados, es decir, un 63,54% de concordancia. Finalmente, la comparación de resultados de

FImpute vs. SCDA mostró un 52,02% de concordancia (9.100.208 genotipos concordantes de 17.492.459 imputados). Al evaluar la concordancia por cromosoma, se encontraron valores en los rangos 52,37 a 66,84, 61,26 a 69,63 y 50,20 a 59,80, para las comparaciones entre RF y SCDA, FImpute y RF, y FImpute y SCDA, respectivamente (Tabla 3). En las tres comparaciones realizadas, la mayor concordancia de imputación se obtuvo para los genotipos faltantes en SNPs del cromosoma X.

Tabla 3. Concordancia entre estrategias de imputación evaluada por cromosoma.

Cromosoma	Número de SNPs	% de genotipos imputados	% de concordancia entre estrategias de imputación		
			RF y SCDA	FImpute y RF	FImpute y SCDA
"0"	631	41,67	53,10		
1	5.801	27,95	53,05	61,45	51,90
2	5.549	27,93	53,04	61,93	51,87
3	4.946	27,65	53,08	61,26	51,64
4	2.640	28,33	53,46	62,66	52,08
5	2.332	26,39	52,69	62,58	51,37
6	2.990	32,08	52,37	64,69	51,72
7	2.199	27,15	53,27	63,71	52,72
8	2.033	28,55	52,70	62,84	50,89
9	2.119	28,44	54,54	63,13	52,40
10	1.814	29,35	53,32	65,59	52,72
11	1.223	26,19	55,26	67,14	53,05
12	1.769	27,59	53,04	63,32	51,31
13	1.741	26,57	53,88	63,76	52,28
14	1.171	26,20	52,59	64,53	50,81
15	1.662	27,23	54,09	64,95	52,30
16	1.539	27,82	53,71	63,83	51,47
17	1.421	28,10	53,74	63,76	51,40
18	1.464	27,98	53,71	64,86	51,66
19	1.266	26,36	54,07	65,24	52,39
20	1.150	27,30	53,18	64,24	50,78
21	920	25,67	54,38	65,13	52,22
22	1.110	26,88	52,88	63,84	50,79
23	1.123	27,09	52,38	64,42	50,30
24	778	26,12	54,34	65,67	51,30
25	999	28,45	53,65	66,34	51,30
26	920	26,88	52,54	63,20	50,20
27	1.408	36,17	66,84	69,63	59,80
Todos	54.718	28,21	53,73		
	54.087	28,05		63,54	52,02

Cromosoma "0" hace referencia al bloque 0 de imputación, constituido por los genotipos de los SNPs sin posición definida en el genoma. RF: random forest. SCDA: autoencoders convolucionales dispersos para la eliminación de ruido.

En el presente trabajo, la concordancia entre los genotipos inferidos por las estrategias que emplearon FImpute y RF fue mayor que la que presentaron los inferidos mediante FImpute y SCDA. En un trabajo en el que se comparó el desempeño de ocho

métodos de machine learning en la imputación de genotipos de tríos padre-madre-hijo respecto de la imputación realizada por el programa Beagle, se encontró una mayor precisión de imputación en los genotipos inferidos por Beagle que en los inferidos por cualquiera de los otros métodos (Mikhchi et al., 2016). Otros estudios compararon la precisión de imputación obtenida por FImpute y Beagle, encontrando que FImpute tiene un desempeño similar o superior al de Beagle (Piccoli et al., 2014; He et al., 2015). Entonces, asumiendo que la imputación por FImpute sería más precisa que la que se logra empleando algoritmos de machine learning, los resultados obtenidos en el presente trabajo sugieren que RF se desempeñó mejor que SCDA para inferir los genotipos no asignados del dataset empleado. No obstante, la baja concordancia observada con los genotipos imputados por FImpute, aún para el modelo que empleó el algoritmo de random forest, puede atribuirse a la cantidad de genotipos no asignados del archivo de partida (28,21%). En general, a mayor cantidad de datos faltantes en el dataset, menor es la calidad de la imputación. 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.

En cuanto a la concordancia de imputación obtenida entre pares de estrategias para los distintos autosomas, no se evidenció una relación con el porcentaje de genotipos imputados. Por otro lado, en las tres comparaciones realizadas, se observó una mayor concordancia de imputación para el cromosoma X, lo que indica un desempeño diferente de los modelos empleados para la inferencia de genotipos de SNPs localizados en este cromosoma, en comparación con los de autosomas. Trabajos previos reportaron que, en general, las variantes autosómicas son imputadas con mayor precisión que las variantes del cromosoma X (Su et al., 2014; de Souza et al., 2025). Distintos factores como la composición de machos y hembras en las poblaciones de referencia y validación, el software de imputación utilizado y la posibilidad de imputar en el cromosoma completo o en las regiones pseudoautosómica y no-pseudoautosómica del mismo, determinan la variabilidad en la precisión de la imputación de genotipos en el cromosoma X (de Souza et al., 2025). En nuestro caso, la población a imputar estaba constituida por machos y hembras y no se discriminó entre las regiones pseudoautosómica y no-pseudoautosómica del cromosoma X durante el proceso de imputación.

4 Conclusión

En este trabajo se investigó la concordancia en la imputación de genotipos sobre un dataset con 28,21% de datos faltantes, realizada mediante dos algoritmos de machine learning, random forest y SCDA, en comparación con la realizada utilizando el programa FImpute que considera información familiar y poblacional para la inferencia de genotipos. Se observó una mayor concordancia entre RF y FImpute que entre SCDA y FImpute, uniforme en marcadores autosómicos y ligeramente superior en los del cromosoma X. Para confirmar estos resultados preliminares, se propone continuar con la exploración de algoritmos que usen redes neuronales de convolución y autoencoders para la imputación de genotipos, utilizando datasets con distintos porcentajes de datos faltantes y enmascarando genotipos asignados por genotipificación, de modo de poder evaluar la precisión en las imputaciones logradas.

Referencias

Abolhassani Targhi, M. V., Asgari Jafarabadi, G., Aminafshar, M., 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.

Berry, D. P., Kearney, J. F. (2011). Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5(8), 1162-1169. doi:10.1017/S1751731111000309

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

Browning, B. L., 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., Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4.

Chen, J., Shi, X. (2019). Sparse Convolutional Denoising Autoencoders for Genotype Imputation. Genes (Basel) 10(9), 652. doi: 10.3390/genes10090652.

de Souza, T. C., Pinto, L. F. B., da Cruz, V. A. R., Chud, T. S., Pedrosa, V. B., Junior, G. A. O., Rojas de Oliveira, H., Mulim, H. A., Miglior, F., Schenkel, F. S., Brito, L. F. (2025). Genotype imputation accuracy of X chromosome variants in Holstein cattle based on different software and imputation strategies. Journal of Dairy Science 108(12), 13487-13498. doi: 10.3168/jds.2025-26715.

Faux, P., Geurts, P., Druet, T. (2019) A Random Forests Framework for Modeling Haplotypes as Mosaics of Reference Haplotypes. Frontiers in Genetics 10:562. doi: 10.3389/fgene.2019.00562.

He, S., Wang, S., Fu, W., Ding, X., Zhang, Q. (2015). Imputation of missing genotypes from low- to high-density SNP panel in different population designs. Animal Genetics 46(1), 1-7. doi: 10.1111/age.12236.

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

Impute Only Mode, Fast Unified Random Forests with randomForestSRC, https://www.randomforestsrc.org/reference/impute.rfsrc.html, último acceso 10/02/2026.

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

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

Mikhchi, A., Honarvar, M., Kashan, N. E., 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.

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

Piccoli, M. L., Braccini, J., Cardoso, F.F., Sargolzaei, M., Larmer, S. G., Schenkel, F. S. (2014). Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genetics 15, 157. doi: 10.1186/s12863-014-0157-9.

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

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

Song, M., Greenbaum, J., Luttrell, J., Zhou, W., Wu, C., Luo, Z., Qiu, C., Zhao, L. J., Su, K-J., Tian, Q., Shen, H., Hong, H., Gong, P., Shi, X., Deng, H-W., Zhang, C. (2022). An autoencoder-based deep learning method for genotype imputation. Frontiers in Artificial Intelligence 5:1028978. doi: 10.3389/frai.2022.1028978.

Su, G., Guldbrandtsen, B., Aamand, G. P., Strandén, I., Lund, M. S. (2014). Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genetics Selection Evolution 46(1), 47. doi: 10.1186/1297-9686-46-47.

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 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.