UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

FACULTAD DE ZOOTECNIA Y ECOLOGÍA

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

BIOINFORMÁTICA INTEGRATIVA PARA LA IDENTIFICACIÓN DE GENES DE IMPORTANCIA REPRODUCTIVA EN BOVINOS

POR:

M.V.Z. NYDIA JOANNA BADILLO SEGOVIA

TESIS PRESENTADA COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE MAESTRÍA EN CIENCIAS
ÁREA MAYOR: REPRODUCCIÓN Y GENÉTICA ANIMAL

CHIHUAHUA, CHIH., MÉXICO **NOVIEMBRE-2021**

Bioinformática integrativa para la identificación de genes de importancia reproductiva en bovinos. Tesis presentada por Nydia Joanna Badillo Segovia como requisito parcial para obtener el grado de Maestría en Ciencias, ha sido aprobada y aceptada por:

Ph.D. Carlos Ortega Ochoa
Director de la Facultad de Zootecnia y Ecología

D.Ph. Agustín Corral Luna
Secretario de Investigación y Posgrado

Ph.D. Iván Adrián García Galicia
Coordinador Académico

Dr. Francisco Joel Jahuey Martínez
Presidente

24 de noviembre del 2021
Fecha

Comité:
Dr. José Alfredo Martínez Quintana.
D.Ph. Joel Domínguez Viveros.
M.C. Beatriz Elena Castro Valenzuela.

AGRADECIMIENTOS

Se extiende un supremo agradecimiento a Dios, por darme fortaleza en momentos de debilidad y por darme la oportunidad de cursar la Maestría en Ciencias.

A mis padres, la inspiración misma de todos mis actos, el núcleo del cual surge toda la maravilla que permea mi vida. Se les agradece por su financiamiento económico, afectivo e intelectual.

Se agradece con cariño y afecto al Dr. Francisco Joel Jahuey Martínez por creer en mí, por su paciencia, por compartir sus conocimientos en todo momento, por siempre contar con su apoyo y por darme la oportunidad de demostrar el empeño y dedicación expuestos en este trabajo.

A Adrián Alvarado por ser parte de mi vida, por ser un gran amigo, fundamental apoyo y motivarme cada día.

A mis compañeros Álvaro Vargas, José Álvarez y Neón Larios por toda la ayuda brindada, y los buenos momentos que pasamos juntos.

Al CONACYT por el apoyo económico brindado a través de la beca CVU: 954774 de estudios para posgrado.

A la Universidad Autónoma de Chihuahua por haberme abierto sus puertas y permitirme cursar los estudios de maestría.

DEDICATORIA

A mis padres, hermanas y sobrino que fueron un factor importante para alcanzar este logro.

A mis profesores, por esforzarse en mí.

A mis compañeros que contribuyeron en alcanzar mi meta.

A la Facultad de Zootecnia y Ecología que me permitió ser parte de esta institución.

CURRICULUM VITAE

La autora nació el 15 de junio de 1995 en la Ciudad de Torreón, Coahuila, México.

2010-2013 Estudios medio superior en Informática en la preparatoria colegio nacional de educación profesional técnica en electrónica en la rama de robótica.

2013-2018 Estudios superiores en la Universidad Autónoma Agraria Antonio Narro, egresado con título en Medicina veterinaria y zootecnia. Titulado con tesis.

RESUMEN

BIOINFORMÁTICA INTEGRATIVA PARA LA IDENTIFICACIÓN DE GENES DE IMPORTANCIA REPRODUCTIVA EN BOVINOS

POR:

M. V. Z. NYDIA JOANNA BADILLO SEGOVIA

Maestría en Ciencias en Producción Animal

Secretaría de Investigación y Posgrado

Facultad de Zootecnia y Ecología

Universidad Autónoma de Chihuahua

Presidente: Dr. Francisco Joel Jahuey Martínez

Una característica de importancia económica en bovinos productores de leche es la tasa de preñez de las hijas (DPR). Sin embargo, seleccionar animales sobresalientes mediante métodos convencionales no ha generado los resultados esperados, debido a su compleja arquitectura genética, herencia poligénica y baja heredabilidad. Por lo que se requieren estrategias que identifiquen la variabilidad genética que subyace este rasgo. El objetivo de este trabajo fue emplear el método de redes neuronales para priorizar genes candidatos para DPR haciendo uso de bases de datos que contienen información biológica de bovinos. Se generaron 122 variables después de compilar información de Ensembl, AnimalQTLdb, STRING, OMIA y KEGG. Tras el filtrado de la información se construyó un modelo de red neuronal con 10 variables, 120 neuronas y un valor de decaimiento 0.1. El análisis se realizó en el software R y la librería caret. El entrenamiento de la red se realizó mediante validación cruzada obteniendo una exactitud de predicción del 83 %. La red neuronal se utilizó para

priorizar genes en 4 QTLs para DPR en los cromosomas 15, 18 y 25 reportados en AnimalQTLdb. Considerando un valor P de 0.95 se priorizaron 29 genes de los 387 localizados en los QTLs mencionados, seis de ellos (*KLC3, ERCC2, GIPR, SIX5, DMPK y DMWD*) se han relacionado a caracteres reproductivos en modelos de estudio como ratón y humano, y uno *FOSB* se ha asociado con DPR en bovinos. La priorización de genes mediante redes neuronales resultó eficaz para el estudio de QTLs en bovinos.

ABSTRACT

INTEGRATIVE BIOINFORMATICS FOR THE IDENTIFICATION OF GENES OF REPRODUCTIVE IMPORTANCE IN CATTLE

BY:

NYDIA JOANNA BADILLO SEGOVIA

An economically important characteristic in dairy cattle is the daughter pregnancy rate (DPR). However, selecting outstanding animals using conventional methods has not generated the expected results, due to their complex genetic architecture, polygenic inheritance, and low heritability. Therefore, it requires strategies that identify the genetic variability that underlies this trait. The objective of this work was to use the neural network method to prioritize candidate genes for DPR using databases that contain biological information of bovines. 122 variables were generated after compiling information from Ensembl, AnimalQTLdb, STRING, OMIA, and KEGG. After filtering the information, a neural network model was built with 10 variables, 120 neurons and a decay value of 0.1. The analysis was carried out in the R software and the caret library. The training of the network was carried out by cross validation obtaining a prediction accuracy of 83%. The neural network was used to prioritize genes in 4 QTLs for DPR on chromosomes 15, 18 and 25 reported in AnimalQTLdb. Considering a P value of 0.95, 29 genes of the 387 located in the QTLs were prioritized, six of them (*KLC3*, *ERCC2*, *GIPR*, *SIX5*, *DMPK* and *DMWD*) have been related to reproductive characters in study models such as mouse and human, and one (*FOSB)* has been associated with DPR in cattle. The prioritization of genes through neural networks was effective for the study of QTLs in bovines.

CONTENIDO

Página

LISTA DE CUADROS

LISTA DE CUADROS DEL APÉNDICE

LISTA DE GRÁFICAS

LISTA DE FIGURAS

INTRODUCCIÓN

Un buen desempeño reproductivo es un factor determinante para la rentabilidad del sistema de producción de bovinos leche (Lucy, 2001). No obstante, la selección intensiva hacia mayor producción de leche ha generado efectos desfavorables en la fertilidad, debido al antagonismo genético entre estos aspectos (Pritchard *et al.,* 2014; Norman *et al.,* 2009). Esto ha provocado una disminución en eficiencia reproductiva y pérdidas económicas considerables como consecuencia del incremento en el número de inseminaciones por concepción, tratamientos hormonales, servicios veterinarios, aumento en el número de días abiertos y animales de deshecho (Kargo *et al.,* 2014).

Genéticamente, estos rasgos son considerados complejos, es decir, que son controlados por múltiples *loci* (que contribuyen con una pequeña proporción de la variación fenotípica) y que se ven afectados por factores ambientales (Möser *et al*., 2015); y generalmente tienen baja heredabilidad desde de 0.01 a 0.06 (Pryca y Veer Kamp, 2001), limitando el mejoramiento genético de las características reproductivas. Aunado a esto se tiene la limitante de que a muchos ganaderos les preocupa que la producción de leche disminuya (Liu *et al*., 2009) al priorizar sobre estos caracteres.

Debido a la dificultad para mejorar la reproducción se han implementado nuevas herramientas que permiten seleccionar animales con alto merito genético. Tradicionalmente, la selección animal mejorada por los índices de selección ha ayudado en el mejoramiento genético de los caracteres productivos (Snelling *et al.,* 2013), pero actualmente estrategias como la selección genómica y el mapeo

genético en ganado lechero son técnicas de rutina en países desarrollados (Michelozzi *et al.,* 2011). Además, las tecnologías de genotipificación basadas en polimorfismos de un solo nucleótido (SNPs), junto con tecnologías de secuenciación de ADN y ARN (Nguyen *et al.,* 2018), han permitido explorar la estructura del genoma y su función sobre esta clase de fenotipos complejos.

En cuanto al mapeo genético, hasta ahora se han identificado más de 220 mil regiones genómicas asociadas a fenotipos (QTLs, por sus siglas en inglés Quantitative Trait *Locus*) de importancia económica en bovinos (Hu *et al.,* 2019) incluyendo caracteres reproductivos. No obstante, la ubicación de los genes causales en estos QTLs todavía se desconoce y su identificación es difícil ya se requieren estudios adicionales como de mapeo fino, de expresión diferencial, de validación y de la disponibilidad de poblaciones. Además, los estudios de mapeo genético suelen exigir un gran número de muestras, y requieren una alta precisión del fenotipo evaluado para obtener un resultado consistente y significativo (Hoglund *et al.,* 2014).

Lo anterior ilustra los desafíos de estudiar la arquitectura genética de las características y de ubicar los genes cuya variación se correlaciona con el fenotipo, principalmente en caracteres poligénicos y de baja heredabilidad como los de reproducción. Afortunadamente, el rápido desarrollo de tecnologías de alto rendimiento para el estudio de la genómica ha acelerado la generación de datos en múltiples niveles de información, por lo que cada vez se dispone de más conocimiento sobre la función y propiedades de los genes (Ensembl, 2020). Bases de datos tanto de secuencias genómicas, anotaciones génicas, rutas biológicas, interacción entre proteínas y expresión de genes aportan información

esencial para el conocimiento de las especies y sus mecanismos vitales (Klasberg *et al.,* 2016). Debido a la disponibilidad de datos biológicos ha surgido la necesidad de analizar toda esta información, con el fin de extraer el máximo conocimiento en el menor tiempo posible y con alto grado de fiabilidad, (Hassani-Pak y Rawlings, 2017) y ayudar a revelar la importancia funcional de los genes desde un enfoque bioinformático.

Hipotéticamente, la información en bases de datos pudiera contener patrones importantes indicadores de la relevancia funcional de los genes. Por lo tanto, el objetivo de esta investigación fue integrar información de diferentes repositorios de datos biológicos aplicando técnicas bioinformáticas y de aprendizaje automático (redes neuronales) con el fin de localizar genes de importancia relacionados con uno de los caracteres de reproducción más importantes en bovinos de leche como es la tasa de preñez de la hija.

REVISIÓN DE LITERATURA

Genómica de Bovinos y Diversas Fuentes de Información Biológica

La genómica es la subdisciplina de la genética enfocada al estudio de los genes y genomas, así como también de la relación que existe entre ellos y con distintos factores externos. Incluye los mapas genómicos, las secuencias genómicas y las funciones génicas (López-López *et al*., 2005).

Desde el lanzamiento del primer borrador del genoma bovino en 2009 (The Bovine Genome Sequencing and Analysis Consortium), su estructura y anotación se actualiza y mejora constantemente (Tellam *et al.,* 2009). En la actualidad se cuenta con varias versiones del genoma bovino (Zimin *et al*., 2009) que se pueden consultar en las bases de datos de Ensembl, de la Universidad de California en Santa Cruz (UCSC) y del Centro Nacional para la Información Biotecnológica (NCBI), principalmente. Las versiones más utilizadas para fines comparativos han sido el Btau 4.6 y UMD 3.1, pero recientemente, se lanzó un nuevo genoma de referencia denominado ARS-UCD 1.2 (Ensembl, 2020).

El tamaño, organización y variabilidad son aspectos importantes del genoma bovino (Kiser *et al.,* 2019). De acuerdo con la base de datos de Ensembl, el genoma bovino tiene un tamaño de 2.7 Giga pares de bases (Gpb) y contiene un total de 24,616 genes (19,994 codificantes, 3,895 no codificantes y 797 pseudogenes). Estructuralmente, está organizado en 29 autosomas, 2 cromosomas sexuales y 1 cromosoma mitocondrial. En cuanto a variaciones genéticas, se reportan casi 100 millones de variaciones cortas (principalmente SNPs) y cerca de 20,000 variaciones estructurales.

Derivado de la secuenciación del genoma, se ha generado información sobre la función de los genes, su participación en rutas metabólicas y de su interacción, dando como resultado el surgimiento de muchas bases de datos de información biológica (Hagen *et al*., 2018, Shamimuzzaman *et al*., 2019).

En cuanto a anotaciones funcionales de los genes, la base de datos Pantherdb (versión 16), la cual se deriva de la base de datos Gene Ontology, reporta que, de los 24,616 genes identificados en el genoma bovino, aproximadamente el 50 % cuenta con anotaciones del tipo procesos biológicos (PB), componentes celulares (CC) y funciones moleculares (FM). Sin embargo, a pesar de que 13,832 productos génicos han sido asignados a familias de proteínas, la participación en rutas metabólicas para el 90 % de los genes todavía se desconoce (Pantherdb, 2019).

Respecto a la expresión de los genes, el sitio web denominado “All Of Gene Expression” que vincula la información de tres bases de datos (Gene Expression Omnibus, GEO; ArrayExpress, AE; y Sequence Read Archive, SRA) reporta hasta la fecha una total de 950 estudios enfocados al análisis de expresión génica en bovinos (Bono, 2020). No obstante, los estudios en bovinos, así como en otros animales de granja todavía siguen siendo pocos comparado con los casi 60,000 estudios reportados en humanos (Bono, 2020). Esto se debe principalmente a la complejidad para experimentar con este tipo de especies y a la dificultad para medir las características de interés, aunado al alto costo de las tecnologías de microarreglos y de secuenciación del ARN (Hayes *et al.,* 2009).

Por otro lado, la red de interacción de proteínas de bovino, según la base de datos de STRING (versión 11.5), tiene un total de 13,795,040 interacciones.

Esto demuestra que los sistemas biológicos son complejos y que constan de diferentes componentes que interactúan entre sí de forma simultánea generando el fenotipo observable e influyendo en la variación fenotípica (Orgogozo *et al.,* 2015).

Mapeo Genético de Características de Importancia en Bovinos

Los esquemas de selección genética en bovinos se han desarrollado con base en la información genealógica y productiva, generando predicciones, las cuales permiten caracterizar y categorizar a los animales respecto a una característica de interés (Martínez *et al*., 2012). Sin embargo, para variables reproductivas en ganado lechero la técnica convencional para la selección animal no ha tenido un avance sustancial y no se toman en cuenta todas las fuentes de variabilidad genética (Gutiérrez-Reinoso *et al*., 2021). Por lo tanto, al analizar e incluir las diferencias a nivel de ADN mediante información genómica es posible enfrentar los problemas asociados con la selección tradicional y, por lo tanto, seleccionar animales genéticamente superiores a edades tempranas con alta confiabilidad (Goddard *et al.,* 2010).

Para apoyar la selección animal y el mejoramiento genético se han desarrollado metodologías que implican la identificación de *loci* responsables de caracteres cuantitativos, a esta metodología se le denomina mapeo genético. Los estudios de asociación genotipo-fenotipo mediante asociación del genoma completo (GWAS, por sus siglas en inglés Genome Wide Association Studies) han sido una herramienta poderosa para explicar la base genética de caracteres poligénicos o también denominados rasgos complejos (Sharma *et al.,* 2015). Este

tipo de estudios se realizan en cuatro pasos: 1) se genotipifican miles de SNPs en cientos de individuos, 2) se realiza la asociación estadística, 3) se exploran las regiones genómicas estadísticamente más importantes, así como los genes allí localizados, 4) se validan los SNPs que demuestran ser significativos mediante técnicas de mapeo fino analizando además los SNPs adyacentes (Sevilla, 2007). La estrategia de GWAS ha demostrado ser un método ideal para identificar genes asociados con varios fenotipos y para dilucidar los mecanismos genéticos que subyacen los rasgos complejos, lo que ha incrementado los estudios en diferentes especies como bovinos, porcinos, equinos, ovinos, caprinos y aves (Sharma *et al.,* 2015).

Como resultado de la identificación de QTLs han surgido dos fuentes de información importantes para especies animales. Estas son la base de datos AnimalQTLdb (Hu *et al.,* 2019) y OMIA (Nicholas *et al.,* 2003).

La base de datos de QTLs, denominada AnimalQTLdb, es un recurso valioso para conocer los mecanismos genéticos de características económicamente relevantes en especies destinadas a la producción (Zhang *et al*., 2016). Desde el 2005, esta base de datos ha presentado un crecimiento importante, debido a que se deposita una gran cantidad de QTLs de diversas investigaciones (Hu *et al*., 2019). Actualmente, AnimalQTLdb (versión 44) contiene 220,401 QTLs asociados a más de 2000 fenotipos de importancia en diferentes especies, principalmente el bovino; en la Figura 1 se muestra la ubicación de los QTLs reportados para características reproductivas y producción de leche en bovinos.

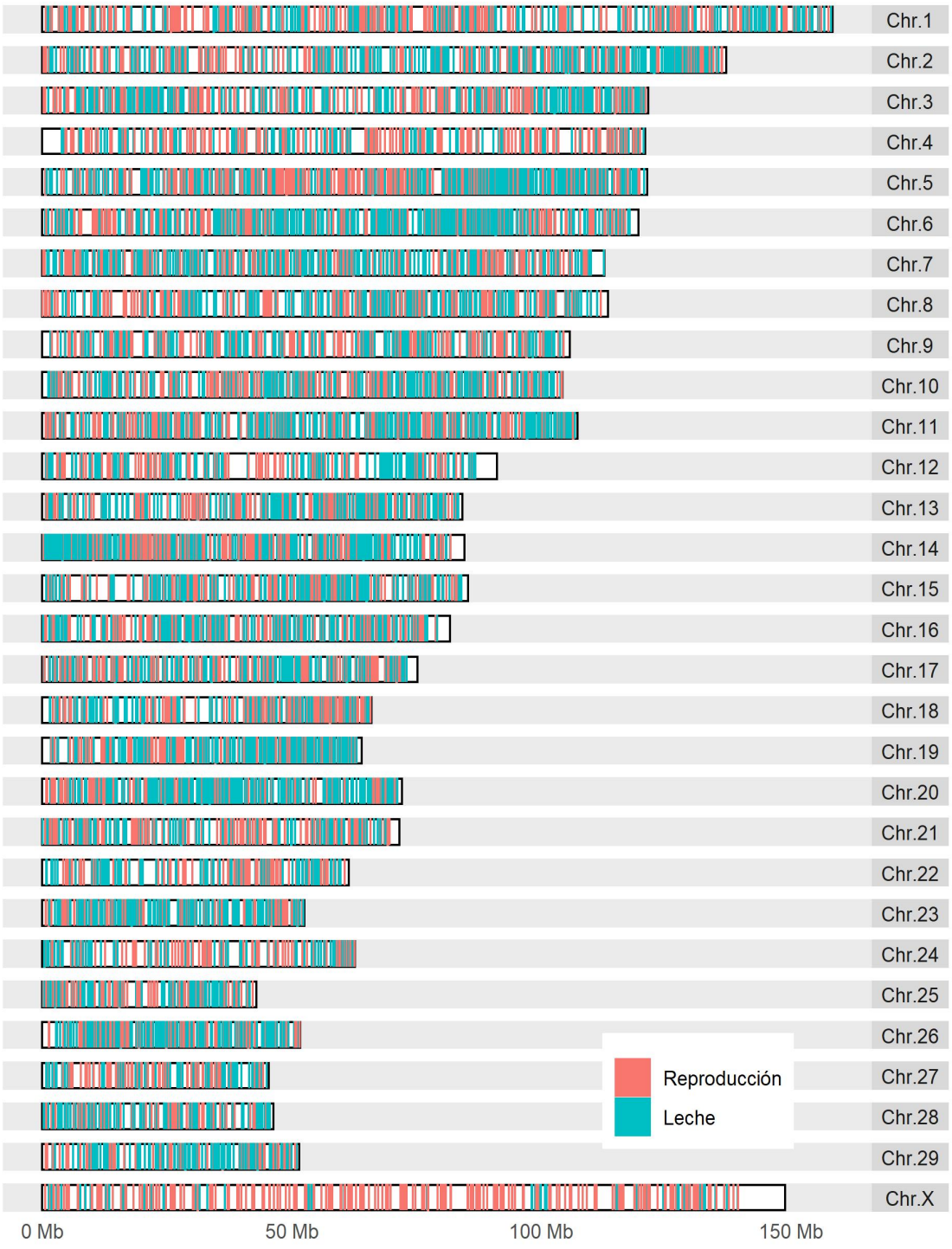


Figura 1. Cariograma de la distribución de los QTLs asociados a variables reproductivas y producción de leche en el genoma bovino. En cada cromosoma, las barras en rojo representan los QTLs de caracteres de reproducción y en azul los de producción de leche. La posición de los QTLs fue obtenida de la base de datos AnimalQTLdb (versión 44).

Con base en la información que brinda la AnimalQTLdb se han realizado diversos estudios de minería de datos. Salih y Adelson (2009) analizaron los QTLs de bovino y asociaron su ubicación con la distribución de genes pertenecientes a familias génicas; Wu *et al*. (2011) realizaron un meta-análisis para la ganancia diaria de peso en porcinos con base en los QTLs reportados; Zhang *et al.* (2016) identificaron los QTLs que influyen tanto características de crecimiento como de salud en porcinos. La AnimalQTLdb ofrece la posibilidad de realizar nuevas investigaciones, por ejemplo, estudios comparativos y bioinformáticos que faciliten la identificación de los genes causales de la variación fenotípica (Hassani-Pak *et al*., 2017).

Por otro lado, la base de datos OMIA contiene las mutaciones causales de genes mendelianos, o también conocidos como genes de efecto mayor (Nicholas, 2003). Esta base de datos tiene anotaciones génicas para diversas especies como *Bos taurus*, *Sus scrofa*, *Gallus gallus*, *Homo sapiens*, entre otras.

Los efectos de los genes depositados en OMIA han sido validados en diversas poblaciones y especies, prueba de ello son los genes para el color de capa (*MC1R*, *ASIP*, *TYR*, *KIT*) y doble músculo (*MSTN)*, por mencionar algunos. Otros genes de la base de datos OMIA han sido asociados a características de la salud animal (OMIA, 2019). Específicamente en bovinos se han identificado genes de efecto mayor relacionados con enfermedades y otros fenotipos (presencia de cuernos, color del pelaje, abortos), estos genes se encuentran distribuidos en los 30 cromosomas del genoma (Figura 2).

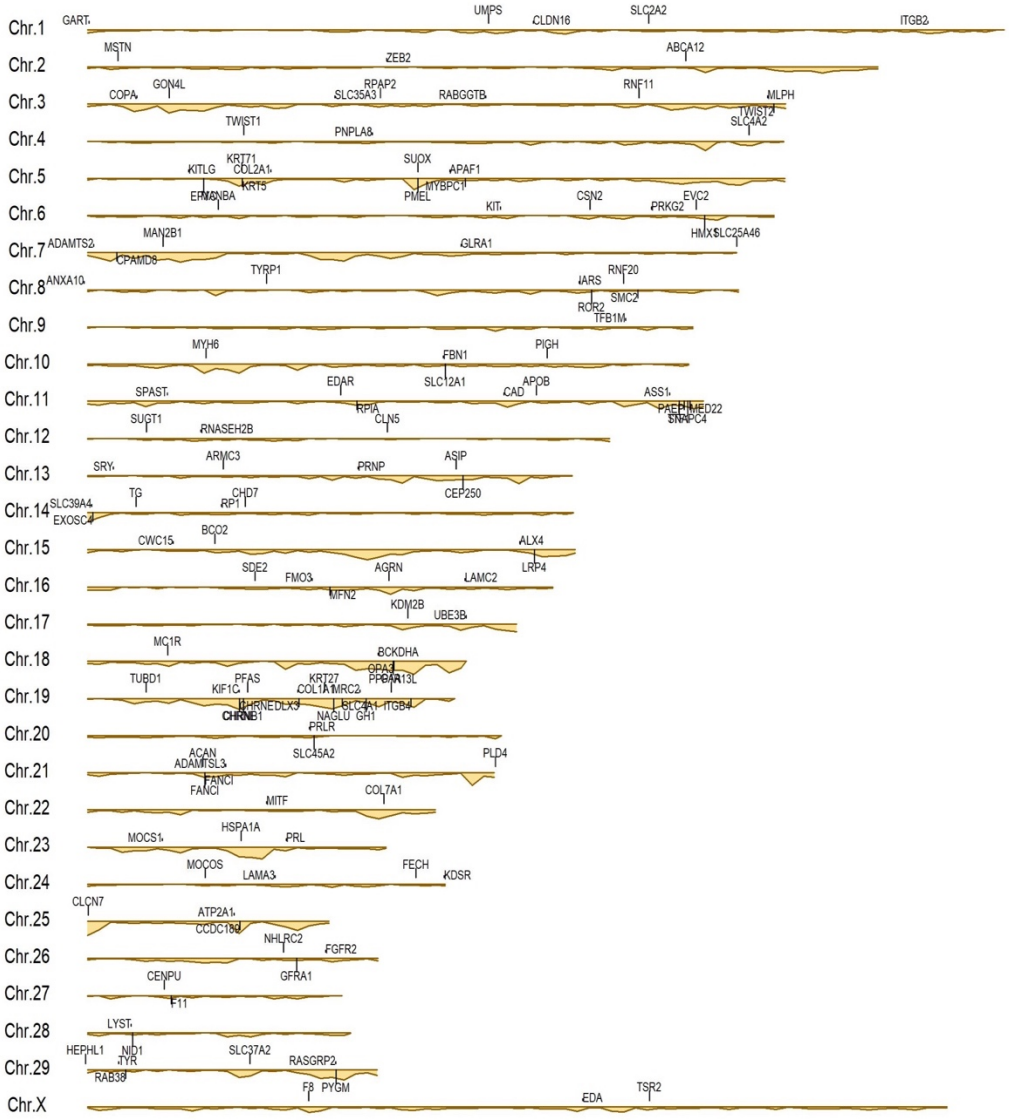


Figura 2. Cariograma de la distribución de los genes de efecto mayor en el genoma bovino. La información de los genes mayores fue obtenida de la base de datos OMIA y la ubicación de los genes se obtuvo de la base de datos Ensembl.

Tasa de Preñez de la Hija (DPR)

En los últimos años la selección de animales sobresalientes con rasgos deseados, incluido el rendimiento reproductivo ha recibido énfasis en los índices de selección de ganado lechero, con la finalidad de lograr objetivos de reproducción más equilibrados y sostenibles (Egger-Danner *et al*., 2015). Según Bo (2009) una meta de mejoramiento debe incluir los siguientes aspectos: mayores ingresos, facilidad de manejo y costos reducidos que incluyen mejor fertilidad, menos enfermedades y tasas de sacrificio reducidas. Muchos países se basan en un grupo de diversos rasgos económicamente importantes, de acuerdo a los productores, para clasificar el ganado sobresaliente mediante selección genética; dentro de estos rasgos la fertilidad es de los más importantes (Egger-Danner *et al*., 2015).

En bovinos de leche, la tasa de preñez de hijas (DPR, por sus siglas en inglés Daughter Pregnancy Rate), es una de las características reproductivas y de fertilidad más importantes, y es considerada económicamente relevante. DPR es un rasgo que evalúa a los toros por la capacidad que tienen sus hijas de quedar preñadas, por ejemplo, cada punto en la tasa de preñez equivale a un cambio del 1 % en la tasa de preñez del hato, es decir a cuatro días abiertos. Un punto positivo en la tasa de preñez de las hijas equivale a cuatro días abiertos menos y un punto negativo a cuatro días abiertos más, es decir, que se espera que las hijas de un toro con +1.0 queden preñadas cuatro días antes que las hijas de un toro con +0.0 DPR (VanRaden, 2003).

A pesar de ser un carácter poligénico y de baja heredabilidad, se ha logrado acelerar el mejoramiento genético y aumentar la confiabilidad de

selección para DPR, aún y considerando los continuos aumentos de producción de leche por vaca (Chebel y Ribeiro, 2016). Detrás de estos cambios están los programas de mejoramiento genético y genómico que enfatizan rasgos de salud y aptitud, como la fertilidad en lugar de sólo producción de leche (Miglior *et al.,* 2005). Por ejemplo, Wiggans *et al. (*2011) realizaron un estudio mediante selección genómica con un microarreglo de 50 mil SNPs y notaron un aumento del 17 % en la exactitud de predicción de los valores genéticos en Holstein de Estados Unidos de América (E.U.A.). Sin embargo, este aumento fue uno de los más bajos en comparación con otros 31 rasgos de conformación corporal, salud y producción de leche que fue del 30 %.

En cuanto a parámetros genéticos, se ha estimado que la heredabilidad de DPR es aproximadamente de 0.04 (Gaddis *et al.,* 2016 citado por Shao, *et al.,* 2021). Por otro lado, Cochran *et al*. (2013) reportaron que DPR tiene una correlación genética positiva con otros rasgos reproductivos como la tasa de concepción de vaquillas (0.61), tasas de concepción de vacas (0.91), vida productiva (0.81), merito neto (0.49), porcentaje de grasa (0.16). Carmo *et al.* (2019) realizaron un estudio con el objetivo de verificar si la selección de toros basada en DPR implicaba la reducción de los días abiertos en las hijas. Para esto, recolectaron datos de 325 hijas de 42 toros Holstein a partir de la evaluación genética de ganado Holstein de E.U.A. Demostraron que la selección basada en las DPR estimadas para tres grupos de toros (<-2, DPR> -2; <0, DPR> 0; y <2, DPR> 2) impactó positivamente reduciendo los días abiertos por vaca (153 $\pm$ 5.1, 130 $\pm$ 4.3, 128 $\pm$ 7.4, 115 $\pm$ 5.2, respectivamente).

En cuanto a la arquitectura genética, DPR es una característica que no sigue un patrón de herencia simple, por lo cual, se pueden encontrar genes que se expresen en diferentes tejidos, células y procesos biológicos (Koyama y Takahashi, 2020). Cole *et al. (*2011) identificaron genes y procesos biológicos relacionados con caracteres reproductivos incluyendo DPR. En su estudio, genotipificaron toros Holstein de alto y bajo valor genético para DPR con un panel de 434 SNPs. Evaluaron tres tipos de marcadores: SNPs asociados con rasgos reproductivos o físicamente cercanos a marcadores genéticos para la reproducción, SNPs en genes involucrados en procesos reproductivos y SNPs en genes que se expresan de manera diferencial entre condiciones fisiológicas. Encontraron 40 genes importantes para DPR, principalmente implicados en el sistema endócrino, la señalización celular, la función inmunológica y la inhibición de la apoptosis. Otro estudio de GWAS para DPR fue reportado por Gaddis *et al.* (2016), donde lograron identificar algunos genes (*ESPNL, FAM132B, ILKAP, LINGO2*) y SNPs importantes para este caracter.

Se han propuesto otros enfoques de investigación que abarquen los diferentes sistemas de la biología de la DPR para explorar más a fondo su arquitectura genética, ya sea mediante redes de correlación de genes, análisis de expresión diferencial, redes de proteínas, entre otras técnicas (Fair *et al.,* 2007; Kommadath *et al.,* 2011; Cerri *et al.,* 2012), incluyendo metodologías de minería de datos y técnicas bioinformáticas (Karami *et al.,* 2019).

Estudios de Priorización de Genes Utilizando Métodos Basados en Bioinformática Integrativa

Uno de los primeros estudios bioinformáticos para la priorización de genes en características reproductivas fue el realizado por Hulsegge *et al.* (2013). Estos investigadores utilizaron información de expresión génica, interacciones proteína-proteína y literatura extraída por minería de texto para predecir y clasificar genes relacionados al estro en ganado lechero, e identificar procesos biológicos importantes en partes del cerebro relacionadas con el control de la reproducción. Mediante un análisis basado en redes, analizaron información de STRING, ENSEMBL, Bioconductor y datos de microarreglos de expresión. Primero identificaron grupos de genes interactuantes con base en anotaciones biológicas similares (proceso conocido como asociación por culpabilidad) para luego combinar sus valores de relevancia en cuanto a expresión y citación. De 6,027 genes analizados encontraron 295 diferencialmente expresados en la pituitaria, amígdala, hipocampo e hipotálamo dorsal y ventral, los cuales fueron propuestos como candidatos para el comportamiento durante el estro. Entre las anotaciones biológicas más importantes de estos genes se encuentran: neurotransmisores/canal iónico/sinapsis, receptor de hormonas esteroideas, unión iónica, región extracelular y regulación positiva de procesos metabólicos y de la transcripción.

Siguiendo un análisis bioinformático también basado en redes, Jiang *et al.* (2012) reportaron un estudio de priorización de genes para la susceptibilidad a la mastitis en bovinos. Implementaron un enfoque que utiliza el mapeo de genes ortólogos (de humano y ratón), información sobre enfermedades y otros

fenotipos, QTLs, interacciones proteína-proteína, así como perfiles de expresión de la respuesta a la infección por *E.coli* en la glándula mamaria. Utilizaron diferentes bases de datos como: OMIM, PubMed, GeneRif, Bioconductor, STRING, ENSEMBL. Calcularon diversos valores de relevancia dependiendo de la información de origen, así como un valor global para cada uno de los genes del bovino. Los genes más relevantes estuvieron relacionados con la respuesta inmune y la inflamación; además, utilizaron sus predicciones para identificar genes candidatos dentro de un QTL en el cromosoma 9. De los 188 genes localizados en la región genómica de estudio, pudieron priorizar 88 de ellos. Los autores comprobaron que la priorización de genes basado en la integración de datos, proporciona información relevante para identificar y clasificar genes, y puede aplicarse a diversas enfermedades complejas incluso en diferentes organismos.

Fonseca *et al*. (2018) realizaron una revisión sistemática de resultados de GWAS sobre rasgos testiculares y espermáticos en bovinos, donde utilizaron un análisis de sobreposición para identificar QTLs compartidos entre diferentes rasgos y razas; identificaron 8,201 genes, de los cuales 2,054 estaban asociados al menos a dos características fenotípicas. Estos genes se evaluaron mediante un análisis de asociación por culpabilidad (software GUILDify y ToppGene) con el fin de identificar los mejores candidatos funcionales. El análisis funcional se realizó utilizando información de ontología genética, fenotipos de humano y ratón, rutas metabólicas, minería de textos, patrón de expresión y enfermedades. Finalmente, el análisis de priorización identificó 222 genes entre los cuales se encontraron genes que desempeñan un papel crucial en procesos biológicos

como progresión de la espermatogénesis, control de la actividad ciliar, desarrollo de células de Sertoli, integridad del ADN en espermatozoides y la homeostasis de células testiculares. En algunos casos, estos genes se agruparon en regiones cromosómicas específicas, lo que indica cierto nivel de especialización para rasgos de fertilidad en dichas regiones del ADN. Los genes más importantes se ubicaron en los cromosomas X, 14, y 17.

Bajo una estrategia similar a la antes mencionada, pero añadiendo información de datos de expresión, Fonseca *et al.* (2020) reportaron un estudio para priorizar genes de la fertilidad en bovinos para carne; con base en un análisis de redes de coexpresión, identificaron módulos de genes que se activaban o reprimían bajo condiciones de fertilidad o subfertilidad. La priorización de genes se realizó también mediante un análisis de asociación por culpabilidad. En este estudio lograron priorizar 32 genes que participan en procesos como el desarrollo embrionario, proliferación de células germinales y la regulación de las hormonas ováricas. Entre los genes más importantes de este estudio se encuentran: ENSBTAG00000046047, *PWP2, DACT2, MIA3, COLGALT2, SKA2, MAPKAP1, PPP1R12B, SLC25A15, EEF1AKMT1, PARP4, FMO2, MDM4, RABIF, CCDC181, F2RL2, IQCG, HACL1, PIGR, ARHGEF16, NRDE2, IFT80*. Los genes reguladores de la fertilidad fueron: *EGFR, ETV5, KLF4, TCHP, COX7A2, PIK3C2A, ARID4A, CUX1, PGR, IPO9, DLG1, AGER, SORT1, HNRNPAB, TBX6, HSF1, NUB1, DPH5, LEPR, DYSF, API5, BCKDK, DUSP16, CHFR, PROM1, ERN2, UTP3, RDH10*.

Redes Neuronales y otras Técnicas de Aprendizaje Automático para la Priorización de Genes

El campo de la priorización de genes utilizando herramientas bioinformáticas está incrementado gracias al análisis masivo de bases de datos y al uso de técnicas estadísticas como las basadas en aprendizaje automático (ML, por sus siglas en inglés Machine Learning). Algunas aplicaciones importantes basadas en ML que se han desarrollado recientemente son las herramientas Iswine (Fu *et al.,* 2020) y Endeavour (Tranchevent *et al*., 2016) que han servido de apoyo en la interpretación de resultados de mapeo genético en porcinos y en humanos, respectivamente. Estas herramientas bioinformáticas integran la información de diferentes fuentes y permiten reducir una lista larga de genes, priorizando aquellos biológicamente relevantes y en asociación con el fenotipo de estudio (Bargsten *et al.,* 2014).

Las técnicas de ML proporcionan a los sistemas de análisis la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programadas específicamente (Von-Lilienfeld *et al.,* 2020). Para construir un modelo basado en ML es clave la disponibilidad de datos, los cuales se pueden encontrar de diversas formas: estructurados, semiestructurados o no estructurados. Para analizar dichos datos y extraer la información relevante se pueden utilizar diferentes técnicas de ML (Sarker, 2021). Las técnicas más populares en todas las divisiones son las redes neuronales, seguida de algoritmos genéticos, K- vecino más cercano, regresión logística, árboles de clasificación simple, entre otras (Keith *et al.,* 2021).

Particularmente, las redes neuronales artificiales (ANN, por sus siglas en inglés Artificial Neural Network) se han convertido en una técnica de análisis de ML útil para predecir, clasificar, agrupar y reconocer patrones en bases de datos (Abiodun *et al.,* 2018). Estos sistemas aprenden y se forman a sí mismos, en lugar de ser programados de forma explícita, y sobresalen en áreas donde la detección de soluciones o características es difícil de expresar con la programación convencional. En la genética y la genómica las ANN se emplean para resolver los numerosos problemas que existen dentro de estas disciplinas, por ejemplo, para predecir la función de genes (Yandell y Ence, 2012).

Estructural y funcionalmente, las ANN están formadas por un conjunto de unidades llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. Como se muestra en la Figura 3, las neuronas artificiales forman una capa de nodos de entrada y una capa de nodos de salida, conectados por una o más capas de nodos ocultos. Los nodos de la capa de entrada pasan información a los nodos de la capa oculta mediante funciones de activación, y los nodos de la capa oculta se activan o permanecen inactivos. Las capas ocultas aplican funciones de ponderación a la evidencia (a los datos), y cuando el valor de un nodo particular o conjunto de nodos en la capa oculta alcanza algún umbral se pasa un valor a uno o más nodos en la capa de salida. (Hernández-Ramos *et al.,* 2020). Las ANN aprenden examinando los registros individuales generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realizan una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada (Angermueller *et al.,* 2016). En resumen, en una ANN la información

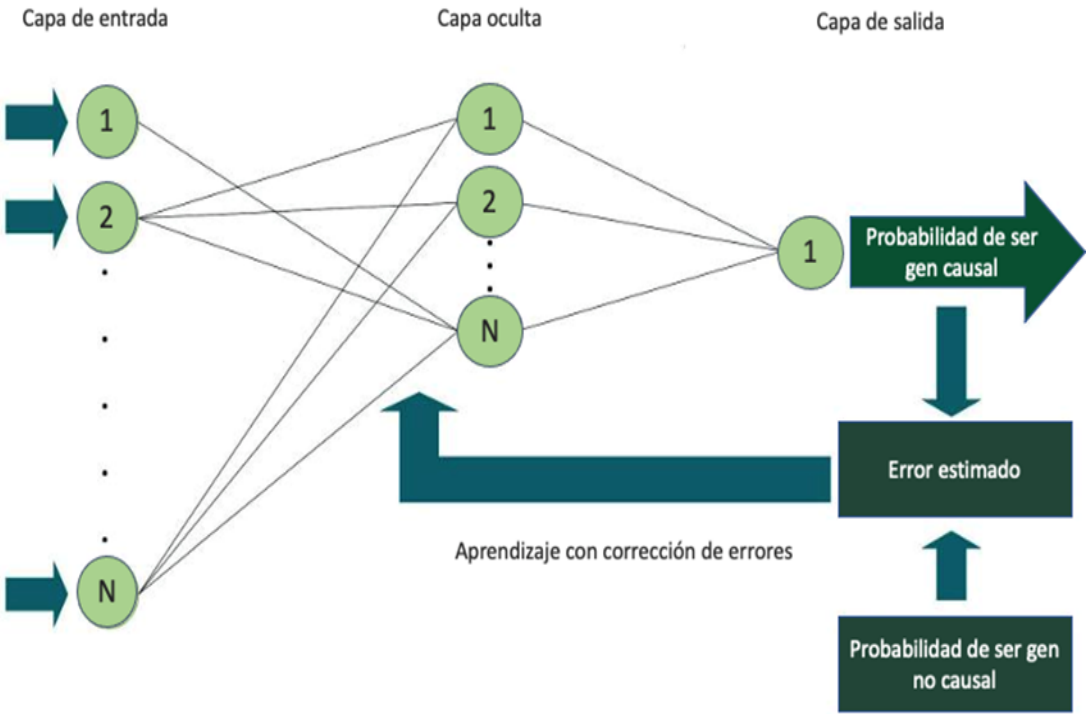


Figura 3. Estructura general de una red neuronal para priorización de genes. Adaptada de Hernández-Ramos *et al.,* 2020.

de entrada, atraviesa la red neuronal donde se somete a diversas operaciones produciendo unos valores de salida o predicciones.

En comparación con las técnicas convencionales de modelado predictivo, las ANN poseen las siguientes ventajas: tienen la capacidad de aprender de los datos durante y después del entrenamiento; no necesitan comprender los mecanismos internos o externos de los datos antes de su implementación (Oludare *et al.,* 2018); pueden analizar una gran cantidad de datos en paralelo; realizan predicciones rápidas y consistentes, tolerando fallas en los datos, e incluso asumiendo no linealidad entre las variables de estudio (Dave y Dutta, 2014). Además, existen aplicaciones que permiten realizar el proceso de creación del modelo (red neuronal) facilitando su ejecución y éxito en la aplicación de este método (Junior *et al.,* 2011). Por lo anterior, las ANN se consideran un método poderoso y flexible que simplifican, facilitan y aceleran el análisis de datos (Patel y Goyal, 2007). A pesar de esto las ANN, siguen siendo problemáticas en el sentido de que ofrecen poca o nula información sobre el proceso por el cual aprenden la totalidad del conocimiento incrustado en ellas (Tickle *et al.,*1998). Además, el desarrollo del modelo es empírico y requiere varios intentos para desarrollar un modelo aceptable (Tu, 1996; Joaquin, 2019).

La mayoría de los estudios de priorización de genes basados en ANN se han realizado utilizando información de especies modelo como el humano (Deo *et al.*, 2014), ratón (Tian *et al.,* 2018) y algunas especies de plantas (Lin *et al*., 2020). Por mencionar un ejemplo, Deo *et al.* (2014) desarrollaron un enfoque de aprendizaje automático llamado OPEN (por sus siglas en inglés Objective Prioritization for Enhanced Novelty) que prioriza genes para enfermedades. Para

crear el modelo de predicción, se utilizaron más de 40,000 variables genómicas extraídas de datos de expresión, información de unión de factores de transcripción, perfiles filogenéticos, dominios de proteínas, objetivos de microARN y análisis de mircroarreglos. Utilizando OPEN, Deo *et al.* (2014) priorizaron el gen FLNC y validaron experimentalmente su descubrimiento en un modelo de pez cebra, además, encontraron una nueva mutación en el sitio de corte y empalme de FLNC en un paciente con una enfermedad cardiovascular grave. OPEN también se probó con 26 rasgos no cardiacos adicionales, incluidos fenotipos autoinmunes, hematológicos, metabólicos y cancerosos, siendo exitoso en la priorización de genes causales.

MATERIALES Y MÉTODOS

Obtención de Datos de Estudio

Para este estudio bioinformático de priorización de genes se analizó e integró información de diferentes bases de datos de bovino como se describe en la Figura 4. Se descargó la información presente en las bases de datos de Ensembl, AnimalQTLdb, OMIA, STRING y KEGG para posteriormente convertirlas en elementos que permitan a un algoritmo de red neuronal extraer información y aprender de ellos. La información obtenida de cada base de datos fue:

Ensembl. Se utilizó la base de datos Ensembl (versión 99) que contiene la información del genoma bovino UMD3.1 y fue analizada con la librería biomartr (Durinck *et al*., 2009). De esta base de datos se obtuvo la información de 24,559 genes (descartando los mitocondriales) incluyendo su ubicación, identificación y conservación; particularmente se obtuvo el nombre, identificador, cromosoma, posición inicial y final, número de transcritos, longitud de transcritos, porcentaje de guanina y citosina. A partir de la información obtenida se calcularon otras variables como el tamaño del gen, la posición del gen respecto al extremo más cercano del cromosoma, número de genes vecinos en ventanas de 1 a 5 MB, grado (promedio) de conservación de genes ortólogos en diferentes especies de mamíferos. Estas variables fueron elegidas considerando que tanto el tamaño, la ubicación y la conservación de los genes son aspectos relevantes que podrían determinar su influencia en el fenotipo.

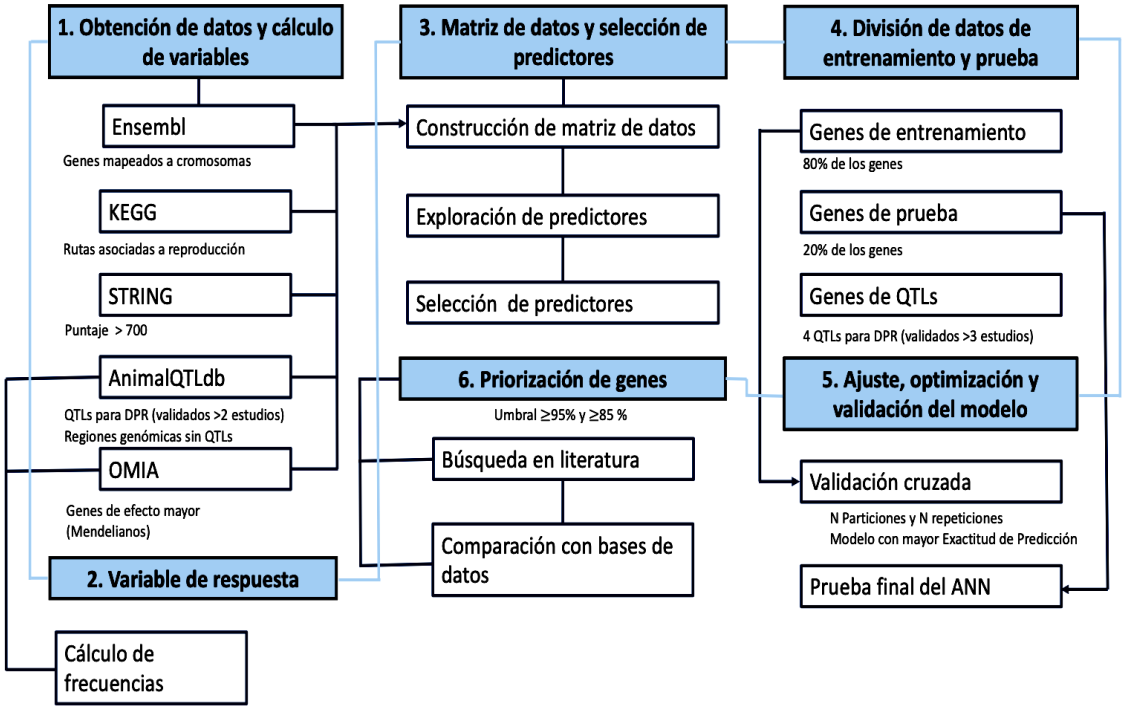


Figura 4. Diagrama general del procedimiento bioinformático para la creación del modelo predictivo de red neuronal.

AnimalQTLdb. Se utilizó la base de datos de QTLs (versión 38) como fuente principal de datos para este estudio y fue analizada con la librería GenomicRanges (Lawrence *et al.,* 2013). Esta base de datos contiene la ubicación de los QTLs asociados a más de 2000 fenotipos en bovinos, así como información adicional (metadatos) como el estudio de origen, la raza o población de descubrimiento, la característica asociada, entre otras variables. Esta base de datos se utilizó junto con la ubicación de los genes para calcular variables como el número de QTLs localizados alrededor de cada gen, el número de estudios y el número de características reportadas, número de tipos y clases fenotípicas de los QTLs. Para estas variables se consideró una ventana de 1 a 5 Mb alrededor de cada gen. El tamaño de las ventanas se definió considerando que los QTLs no se encuentran ubicados exactamente sobre la posición de los genes.

STRING. Se utilizó la versión 11.0 de STRING y fue analizada con las librerías CINNA (Ashtiani *et al.,* 2018) e igraph versión 1.2.6. Sólo se consideraron las interacciones con un puntaje ≥ 700. A partir de estas interacciones, se calcularon 30 variables de centralidad para cada uno los genes anotados en el interactoma bovino. Las variables de centralidad describen la relevancia de los genes considerando el número de conexiones con otros genes en el interactoma. Las variables que aportaron más información a la variabilidad en la centralidad fueron: intermediación, centralidad de apalancamiento, coeficiente de agrupamiento, centralidad de puente local, grado de centralidad, componente máximo de vecindad, puntuaciones de centralidad de autoridad de kleinbergs, puntuaciones de centralidad del centro kleinbergs, centralidades de vectores propios, excentricidad centralidad, distancia promedio y centralidad del

índice winer. Estas variables fueron posteriormente consideradas en la creación de la matriz de datos y el resto de las variables de centralidad fueron descartadas.

OMIA. Se utilizó la base de datos de OMIA (versión 2019) que contiene los genes de efecto mayor y trastornos mendelianos descubiertos en 251 especies animales. Se obtuvieron los genes de efecto mayor reportados como causales para enfermedades en bovinos, así como los genes ortólogos reportados en otras especies animales (n = 216). Esta información sirvió para realizar una clasificación de los genes considerando principalmente su efecto fenotípico. Con base en esto, se dividió el conjunto de genes de bovino en 2 grupos: causales (con efecto en caracteres mendelianos) y no causales (que no se encuentran dentro de QTLs para caracteres poligénicos).

KEGG. Se hizo una búsqueda en la literatura para determinar qué rutas metabólicas participaban en DPR; las rutas identificadas fueron: glucólisis/gluconeogénesis (00010), cascadas de complemento y coagulación (04610), vía de señalización de GnRH (04912), maduración de ovocitos mediada por progesterona (04914), biosíntesis de esteroides (00100), biosíntesis de hormonas esteroides (00140), metabolismo de taurina e hipotaurina (00430), las cuales fueron reportadas por Cochran *et al.* (2013) y Weller *et al.* (2018). Estas rutas fueron exploradas en la base de datos KEGG (versión 2019). Posteriormente fueron utilizadas como variables explicativas categóricas y numeradas del 1 al 7 respectivamente. Las clases utilizadas fueron: 1 si el gen participaba en la ruta y 2 si no lo hacía. Estas variables predictoras fueron utilizadas para incluir en el modelo un “componente reproductivo”.

Con la información obtenida de las diferentes bases de datos se creó una matriz de datos con n filas (total de genes) y n columnas (total de variables). De forma preliminar, se analizaron las variables obtenidas de cada base de datos y se realizaron procedimientos de exploración y visualización. Se analizó, por ejemplo, el tipo de información que contiene, los valores faltantes y la distribución de variables cuantitativas, entre otros aspectos descriptivos. Todas las librerías y técnicas de análisis fueron utilizadas en el software R versión 3.6.1 ya que, esta herramienta estadística permite aplicar metodologías tanto de ANN como de ciencia de datos.

Exploración de la base de datos de QTLs e identificación de QTLs Importantes para DPR

Se realizó una descripción cualitativa y cuantitativa de la base de datos de QTLs, con el propósito de calcular las variables antes mencionadas y también para seleccionar los QTLs de interés para DPR.

El análisis de la base de datos de QTLs consistió en identificar y eliminar QTLs que carecían de ubicación cromosómica, así como en descartar QTLs o anotaciones repetidas (asociados al mismo carácter, ubicados en la misma posición y provenientes del mismo estudio). Posteriormente, se eliminaron también los QTLs con un tamaño mayor a 10 Mb o que se localizaban fuera de los límites cromosómicos. Finalmente, se agruparon los QTLs localizados dentro de segmentos de 3.5 Mb que provenían de un mismo estudio. Esto sirvió para reducir el número de anotaciones de QTLs y para realizar un análisis de sobreposición entre diferentes estudios (similar a lo realizado por Fonseca *et al.*, 2018). Con esto, se identificaron los QTLs para DPR reportados por tres o más

estudios, los cuales fueron considerados como validados. Estos QTLs fueron los sitios blanco del presente trabajo para llevar a cabo los procedimientos de priorización y lograr la identificación de genes candidatos para DPR. En cada QTL validado se investigó el número y la distribución de genes presentes.

Además de los QTLs validados, se obtuvo también una lista de QTLs para DPR que estuvieran reportados por al menos dos estudios (descartando los QTLs de interés) y sirvieron para obtener una lista de genes asociados a DPR. Estos genes fueron añadidos a la matriz de datos y fueron utilizados posteriormente durante el proceso de entrenamiento de la ANN.

Variable de Respuesta

Un requisito de las técnicas de aprendizaje supervisado es que se necesita contar con conocimiento previo sobre la variable de respuesta para el entrenamiento del modelo (Agatonovic-Kustrin y Beresford, 2000). Para este análisis se necesitaba contar con una lista de genes asociados al carácter de interés, así como con una lista de genes no asociados con el fenotipo. Sin embargo, a pesar de que existe una cantidad considerable de QTLs reportados para DPR aún se desconocen los genes causales para esta característica.

Para resolver el problema anterior y poder definir la variable de respuesta se siguió el siguiente procedimiento: 1) debido a que se busca identificar genes con un efecto considerable sobre el fenotipo, se decidió utilizar la lista completa de genes mendelianos (de efecto mayor) reportados en OMIA como genes causales, a pesar de que estos no se relacionan con el fenotipo de estudio; 2) además de los genes mendelianos, se utilizó también una lista de genes

localizados en QTLs para DPR. El criterio de selección de estos genes se basó en que la asociación estuviera reportada por al menos 2 estudios. Estos genes junto con los mendelianos fueron considerados como causales para DPR; 3) para definir el grupo de genes no asociados con DPR (no causales) se utilizaron todos los genes localizados en regiones genómicas que no contienen QTLs para ninguna característica fenotípica de acuerdo con la base de datos AnimalQTLdb previamente filtrada.

Exploración de Variables y Selección de predictores

Los siguientes procedimientos, desde la exploración de variables, la selección de predictores, la construcción y validación de la red, así como el análisis de predicción (priorización de genes) fueron realizados siguiendo el tutorial de “Machine learning con R y caret” (Joaquín, 2020).

Después del análisis exploratorio de las variables se realizó un análisis para estudiar su importancia e identificar las mejores predictoras para la red neuronal. Para esto, se utilizó la técnica de Random Forest la cual pertenece a las técnicas basadas en árboles de decisión. La selección de predictores mediante Random Forest genera una jerarquización de las variables de acuerdo con dos métricas denominadas: pureza de nodos y reducción de precisión (Joaquin, 2021), las cuales fueron utilizadas para seleccionar las variables más importantes. Entre más alto sea el valor de estas métricas, indica una mayor relevancia de las variables. Para este procedimiento, se utilizó la librería y función denominada randomForest (Liaw y Wiener, 2007) con los argumentos mtry = 2 (número de variables candidatas para iniciar el modelo) y ntree = 1000 (número

de árboles). Cabe mencionar que este análisis se realizó considerando únicamente las variables con información completa en la matriz de datos.

Construcción y Validación de la Red Neuronal

Una vez que se estructuró la matriz de datos y que se eligieron las variables predictoras, se construyó la red neuronal (ajuste del modelo) y se realizó el procedimiento de entrenamiento y prueba de la misma. Para esto, la matriz de datos (sin incluir los genes en los QTLs validados) fue dividida aleatoriamente en 2 partes: 80% de los genes (matriz de entrenamiento) fueron utilizados para el entrenamiento de la red neuronal y la otra parte (20 %, matriz de prueba) para probar el desempeño de la red neuronal entrenada. La construcción del modelo se realizó mediante la función “train” de la librería caret (Joaquín, 2020) del software R.

Para el entrenamiento de la red neuronal se utilizó la función “trainControl” y se empleó la estrategia de validación cruzada, con 10 particiones y 5 repeticiones creados a partir de los datos de entrenamiento. Una validación cruzada con 10 particiones y 5 repeticiones implica ajustar y evaluar el modelo 10 x 5 = 50 veces, cada vez con una partición distinta, más un último ajuste con todos los datos de entrenamiento para crear el modelo final. Durante el entrenamiento de la red neuronal se obtuvo una estimación de la exactitud (precisión) de predicción para cada repetición. Al final se obtiene el promedio de todas las exactitudes y esta métrica es una estimación de la capacidad que tiene la red neuronal para predecir nuevas observaciones.

El procedimiento de validación cruzada también se utilizó para investigar el comportamiento de los hiperparámetros del modelo y para definir los valores que proporcionan la mayor exactitud de predicción (optimización). Estos hiperparámetros fueron: el número de neuronas en la capa oculta (size) y el valor que controla la regularización durante el entrenamiento de la red (decay). Esto se realizó debido a que no existe una forma de conocer de antemano cuál es el valor exacto de estos hiperparámetros que da lugar al mejor modelo. Se evaluaron los siguientes valores de size = 10, 20, 50, 80, 100, 120 y los siguientes valores del hiperparámetro decay = 0.0001, 0.1, 0.5. Además, se definió otro hiperparámetro que determina el número máximo de pesos (MAXWits) el cual quedó con un valor fijo de 2000.

Análisis de predicción y priorización de genes para DPR

Finalmente, la librería caret a través de las funciones "train" y “trainControl” arrojan el modelo (red neuronal) que proporciona la mayor predicción de acuerdo con la validación cruzada y con el proceso de optimización de hiperparámetros. Este modelo final es utilizado junto con la función “predict” para predecir nuevos datos, es decir, para predecir (clasificar) los genes que el modelo “no ha visto” (genes de la matriz de prueba) y para predecir (clasificar) los genes en los QTLs validados (priorización).

Los genes en la matriz de prueba pueden ser utilizados para realizar una segunda evaluación de la red neuronal entrenada. Para comprobar el desempeño del modelo y evaluar la exactitud de predicción en datos nuevos, se utilizaron los genes de la matriz de prueba (datos conocidos por el analista, pero “no

observados por el modelo”) y se obtuvo la exactitud alcanzada con este conjunto de datos mediante las funciones “predict” y “confusionMatrix”.

Como resultado del proceso de predicción, para cada gen se obtiene un valor de probabilidad (P) de ser causal o de no ser causal. Para este estudio se utilizó un valor de $P \geq 0.95$ como punto de referencia para identificar genes con muy alta probabilidad de ser causales para DPR. Los genes causales que cumplían con la condición anterior fueron investigados en la literatura, así como en bases de datos para comparar si existe información que los relacione con la característica de interés.

RESULTADOS Y DISCUSIÓN

Identificación de QTLs validados para DPR

La base de datos de QTLs (versión 38) contiene un total 45,406 QTLs asociados con caracteres de reproducción en bovinos, de los cuales 920 han sido reportados para DPR. Tras el filtrado de la información se obtuvo un total de 595 QTLs únicos y con posiciones bien definidas en el genoma bovino. En la Gráfica 1 se muestra la ubicación de los QTLs para DPR y se puede observar que estos se encuentran distribuidos a lo largo de los 30 cromosomas del genoma bovino. Existen algunas regiones genómicas importantes con una mayor densidad de QTLs en los cromosomas 1, 3, 5, 6, 17, 18 y 22 donde es posible que algunos de estos QTLs estén validados por diferentes estudios. Cabe mencionar que la base de datos de QTLs contiene anotaciones provenientes de estudios basados en genes candidatos, GWAS y de secuenciación, los cuales difieren en tamaño y estructura de las poblaciones de estudio, densidad de marcadores, metodología estadística, entre otros aspectos, por lo que no todos los QTLs reportados tienen el mismo nivel de significancia. Por ello, la importancia de investigar y priorizar QTLs reportados por más de un estudio científico.

El análisis de sobreposición de anotaciones logró identificar 377 QTLs para DPR que han sido reportados por al menos dos estudios. De estos se eligieron 4 QTLs que contenían la mayor cantidad de reportes (≥3) y fueron considerados como validados en este trabajo. Los 4 QTLs identificados fueron posteriormente investigados para realizar la priorización de genes. Se asumió

Gráfica 1. Cariograma de la distribución de QTLs para DPR en el genoma bovino. La ubicación de los QTLs fue obtenida de la base de datos AnimalQTLdb.

que el número de reportes en estas regiones genómicas es un buen indicador de la presencia de genes que influyen en la DPR. La ubicación genómica de estos 4 QTLs se describe en el Cuadro 1. En resumen, las regiones genómicas validadas para DPR se localizan en los cromosomas 15, 18 y 25, y contienen QTLs reportados por Cochran *et al. (*2013), Parker *et al. (*2016), Ortega *et al*. (2016), Weller *et al. (*2018) y Jilang *et al. (*2019). La mayoría de estos QTLs fueron identificados en ganado Holstein a través de estudioso de GWAS.

Cabe mencionar que los QTLs validados para DPR tienen una longitud promedio de 2.6 Mb. Generalmente, a medida que los QTLs abarcan mayor espacio genómico tienden a presentar una mayor cantidad de genes; sin embargo, los 4 QTLs validados tienen un número variable de genes que va desde 13 hasta 211. Particularmente, los QTLs validados en el cromosoma 18 poseen una alta densidad génica (más de 100 genes). Precisamente, en este tipo de QTLs se requiere la aplicación de estrategias de priorización que ayuden a detectar los genes más relevantes para su posterior análisis mediante técnicas de mapeo fino. Otros QTLs con una menor densidad génica (como los del cromosoma 15 y 25) también pueden ser analizados mediante técnicas de priorización; sin embargo, en estos casos una simple búsqueda en la literatura podría ayudar a identificar los genes candidatos para el caracter de estudio (Cantor *et al.,* 2010; Bargsten *et al.,* 2014).

Además de su asociación con DPR, otro aspecto relevante de los QTLs validados, es que estas regiones genómicas también están asociadas a otras características de importancia en bovinos, por lo que se pueden considerar regiones pleiotrópicas. De acuerdo con la base de datos AnimalQTLdb, estos

Cuadro 1. QTLs validados para la característica de DPR identificados mediante el análisis de sobreposición.

Cromosoma	Inicio y final (pb)	Genes	Reportes	Características
15	74679526-76475224	17	3	37
18	49042269-52525163	211	4	47
18	61872508-64704565	146	3	51
25	14380150-16669292	13	4	38

QTLs se han asociado a 43 características fenotípicas en promedio. Por ejemplo, edad al primer parto, edad a la pubertad, edad al segundo parto, nivel de hormona Antimülleriana, respuesta inmune mediada por anticuerpos, contenido de ácido araquídico, entre otras (AnimalQTLdb, 2019). Diversos estudios han resaltado la importancia de investigar este tipo de QTLs, principalmente con el objetivo de identificar marcadores genéticos que ayuden a seleccionar y mejorar diferentes caracteres fenotípicos (Zhang *et al*., 2016, Fonseca *et al*., 2018) dada su influencia en las correlaciones genéticas entre características.

Matriz de datos y variables predictoras

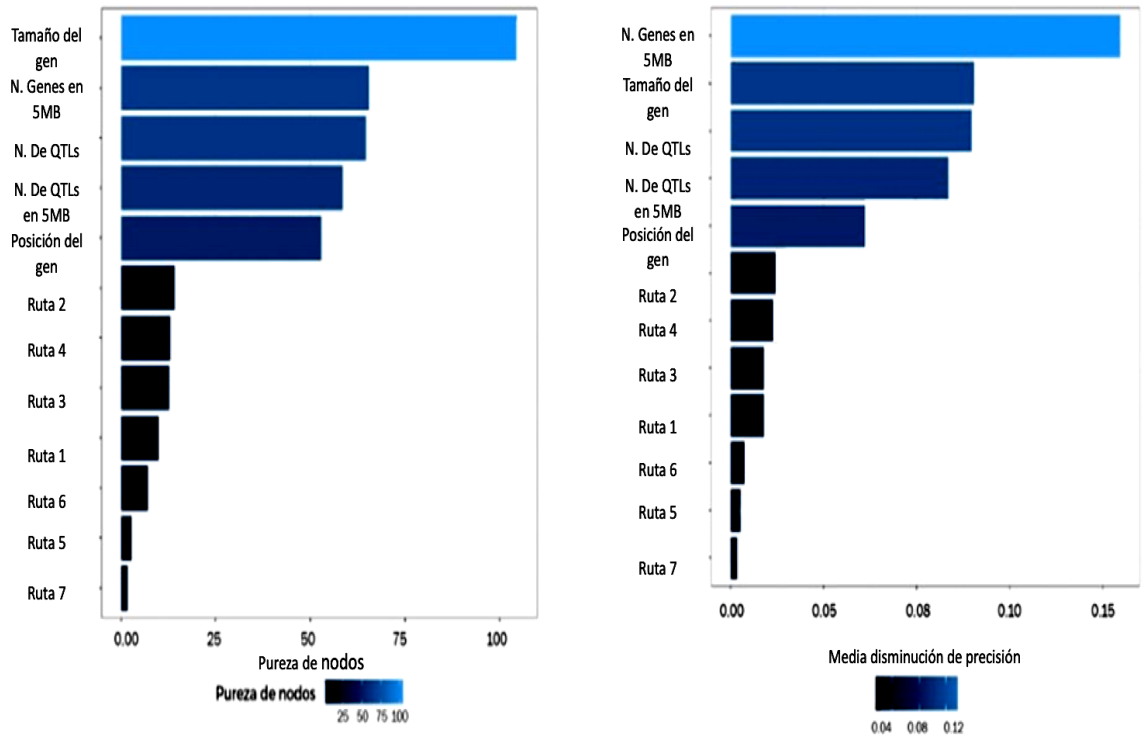
La matriz de datos construida para este análisis de priorización contiene 1,042 filas (genes) y 122 columnas (variables) más la variable de respuesta. Parte de esta matriz de datos fue utilizada como datos de entrenamiento (834 filas) y la otra parte (208 filas) como datos de prueba. La matriz de predicción estuvo representada por 387 genes pertenecientes a los 4 QTLs previamente descritos. Los identificadores de los genes presentes en cada una de estas matrices se presentan los Cuadros 1 y 2 del apéndice.

Respecto al análisis de las variables, la técnica de Random Forest no acepta valores ausentes, por lo que se eliminaron todas las variables de centralidad obtenidas del interactoma bovino, dado que una proporción importante de genes de bovino no cuentan con información de interacción con otras proteínas en la base de datos de STRING; un total de 2,273 genes de bovino no están presentes en el interactoma original de STRING (versión 11.0).

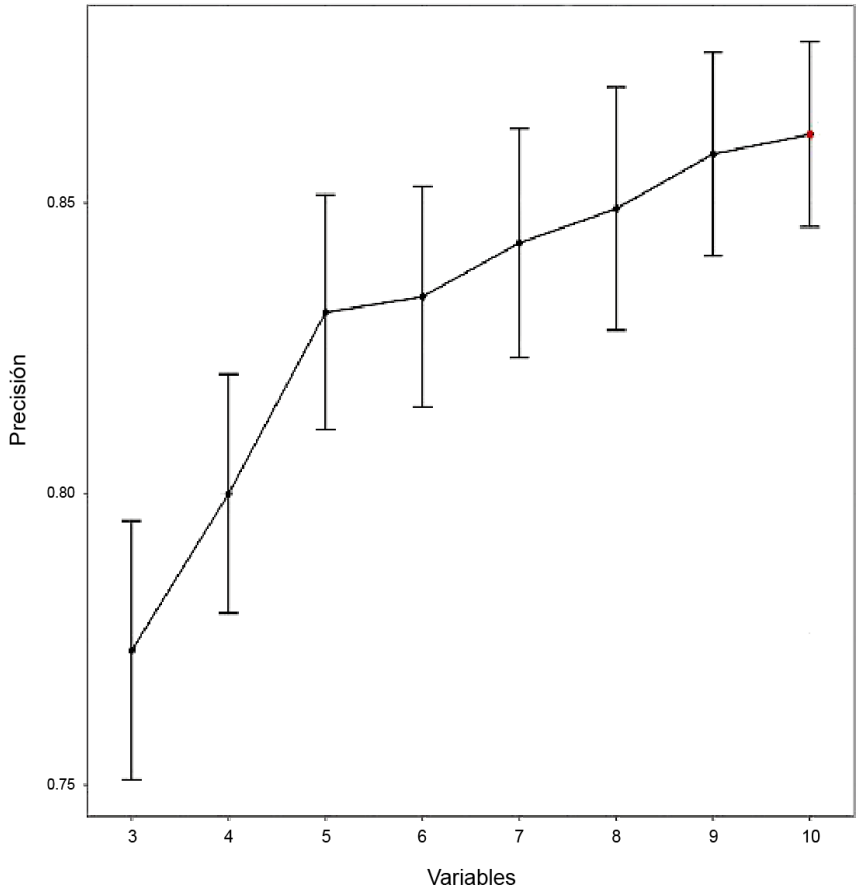
Además, tras filtrar el interactoma considerando sólo los genes con un puntaje de interacción de 700, se descartó la información de 3,458 proteínas más.

De acuerdo con las métricas de pureza de nodos y reducción de disminución de precisión (Gráfica 2), la jerarquía de las variables empezando por la más importante fue la siguiente: 1) tamaño del gen, 2) número de genes en 5 MB, 3) número de QTLs por gen, 4) número de QTLs en ventanas de 5 MB, 5) posición del gen dentro del cromosoma, 6) ruta 2, 7) ruta 4, 8) ruta 3, 9) ruta 1, 10) ruta 6, 11) ruta 5 y 12) ruta 7. En esta jerarquía de predictores, las variables genómicas (1, 2, y 5) calculadas a partir de la información de la base de datos de Ensembl estuvieron entre las más importantes, seguido de las variables (3 y 4) calculadas a partir de la información de la AnimalQTLdb. En los últimos lugares de importancia se encontraron las variables derivadas de las rutas metabólicas de KEGG.

La selección de variables mediante RandomForest también permite evaluar la exactitud de predicción respecto al número de variables consideradas en el modelo. De acuerdo con la Gráfica 3, la máxima precisión de predicción en este análisis fue de 87 % y se obtuvo con 10 predictores. Analizando la Gráfica 3 también es posible notar que incluso utilizando únicamente las primeras cinco variables de acuerdo con la jerarquía de importancia (tamaño del gen, número de genes en 5 MB, número de QTLs por gen, número de QTLs en ventanas de 5 MB y posición del gen dentro del cromosoma) se logra un 83 % en la exactitud de predicción. Aunque el modelo más completo proporciona la mayor exactitud de predicción, un modelo más simple con sólo 5 variables también posee una buena exactitud de predicción. En la literatura se han reportado modelos basados



Gráfica 2. Variables ordenadas de mayor a menor relevancia, según el análisis de RandomForest. Las métricas coinciden en que las variables más informativas son el tamaño del gen, el número de genes en 5 MB, número de QTLs, número de QTLs en 5 MB y posición del gen dentro del cromosoma.



Gráfica 3. Precisión de la red neuronal con respecto a la cantidad de variables. Se obtiene una precisión de predicción de 87 % utilizando 10 variables. Las barras indican el intervalo de confianza al 95 % para la media de exactitud de predicción.

en técnicas de ML que incluyen desde pocos hasta muchos predictores. Por ejemplo, Deo *et al.* (2014) construyeron un modelo basado en ML denominado OPEN que incluyó más de 40,000 variables y que fue utilizado para predecir e incluso validar genes de enfermedades cardiacas en humanos. En contraste, Fu *et al. (*2020) utilizaron un modelo basado en ANN denominado ISwine que incluyó sólo 14 variables predictoras y fue creado para priorizar genes para caracteres de importancia en porcinos como la composición de ácidos grasos. Entre las variables utilizadas en el modelo ISwine se encuentran variables genómicas como el número de SNPs dentro o cerca de los genes, o bien dentro de exones o intrones; número de reportes de asociación de los genes de acuerdo con la literatura; así como variables de expresión de genes en tejidos específicos.

En este trabajo se procuró mantener el número de observaciones (genes) y se decidió descartar variables, aunque éstas pudieran contener información importante. Esto se realizó con el objetivo de aprovechar la lista tan reducida de genes causales conocidos en bovinos, así como la lista de genes no asociados a características cuantitativas obtenida a partir de la información de AnimalQTLdb.

Hasta ahora el modelo de la red neuronal para DPR sólo incluyó 10 variables predictoras, pero puede extenderse para incluir información de Gene Ontology, datos de expresión y de secuencias de los genes e incluso información de interacción de sitios del ADN. En cuanto a información a nivel de secuencias, las librerías de R denominadas rDNAse (Zhu y Dong, 2016) y protr (Xiao *et al*., 2015) proporcionan métodos para calcular variables a partir de secuencias nucleotídicas o aminoacídicas, respectivamente. Este tipo de variables han

resultado importantes para identificar y priorizar genes esenciales mediante técnicas de ML en la mosca de la fruta *Drosophila melanogaster* (Aromolaran *et al.,* 2020).

Por otro lado, las variables derivadas de la interacción entre proteínas (medidas de centralidad) podrían aportar información adicional a la red neuronal y mejorar la exactitud de predicción. Se ha reportado que en el interactoma de proteínas del humano este tipo de información biológica ayuda a discriminar entre genes mendelianos, genes esenciales y genes no relacionados con enfermedades (Spataro *et al.,* 2017). Particularmente, los genes mendelianos o de efecto mayor tienden a presentar valores intermedios en medidas de centralidad tales como cercanía (closeness) e intermediación (betweeness), esto con respecto a genes esenciales y genes no implicados en enfermedades o no esenciales (Spataro *et al.,* 2017). Por lo tanto, realizar estudios exploratorios y de construcción de redes de interacción de proteínas podría ayudar a representar todo el repertorio de genes del bovino e incluir variables del interactoma bovino. Algunas estrategias importantes para la construcción de interactomas han sido reportadas por Ashtiani *et al*. (2019) y Paredes-Sánchez *et al.* (2015). Estos autores utilizan y combinan la información de redes de interacción de proteínas de diferentes especies para reconstruir un nuevo y más completo interactoma de una especie de interés a través de sus genes ortólogos.

Validación de la red neuronal y optimización de parámetros

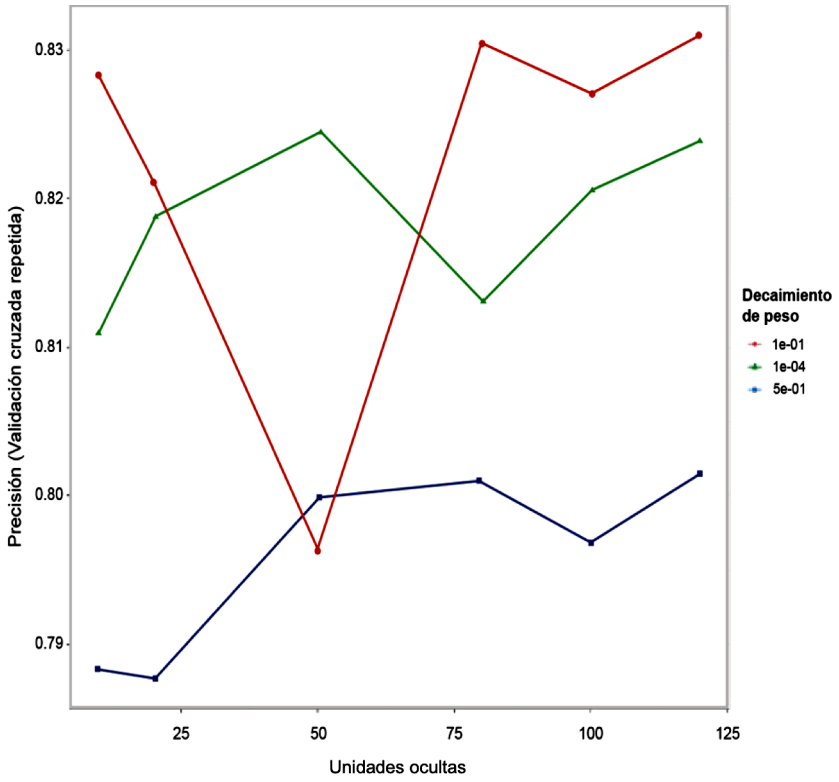
De acuerdo al análisis de validación cruzada y al proceso de optimización de hiperparámetros, la red neuronal alcanza su mejor desempeño con una

exactitud de predicción de 83 % utilizando 120 neuronas en la capa oculta y un valor de 0.1 para decay (Gráfica 4).

La exactitud de predicción alcanzada por el modelo, se podría definir como buena, ya que, para que un modelo predictivo sea útil, este debe de tener un porcentaje de aciertos superior a lo esperado por azar o a un determinado nivel basal (Joaquin, 2020). En problemas de clasificación, el nivel basal es el que se obtiene si se asignan todas las observaciones a la clase mayoritaria. En este caso la clase mayoritaria pertenece a los genes no causales, los cuales presentan una frecuencia del 67 %. Este porcentaje se obtendría si la red neuronal predijera que todos los genes son no causales, por lo que este es el porcentaje mínimo (nivel basal) que se tenía que superar.

Por otro lado, la exactitud de predicción también se corroboró al evaluar la predicción obtenida con la matriz de prueba; este análisis generó una exactitud del 80.77 %. Se considera que esta segunda evaluación permite calcular una exactitud de predicción más realista que la obtenida mediante validación cruzada (Joaquín, 2020). Esto se debe a que los genes en la matriz de prueba no participan durante el ajuste y entrenamiento del modelo, por lo que generalmente, este análisis adicional produce una exactitud más baja que la obtenida mediante validación cruzada.

Un modelo de priorización de genes basado en ML con una exactitud de predicción similar fue reportado por Fu *et al.* (2020), con una precisión del 73.4 % con validación cruzada y demostró un buen desempeño para priorizar genes para composición de ácidos grasos en porcinos. En otro estudio realizado por Manuk *et al.* (2020) obtuvieron una precisión del 78 % para identificar variaciones



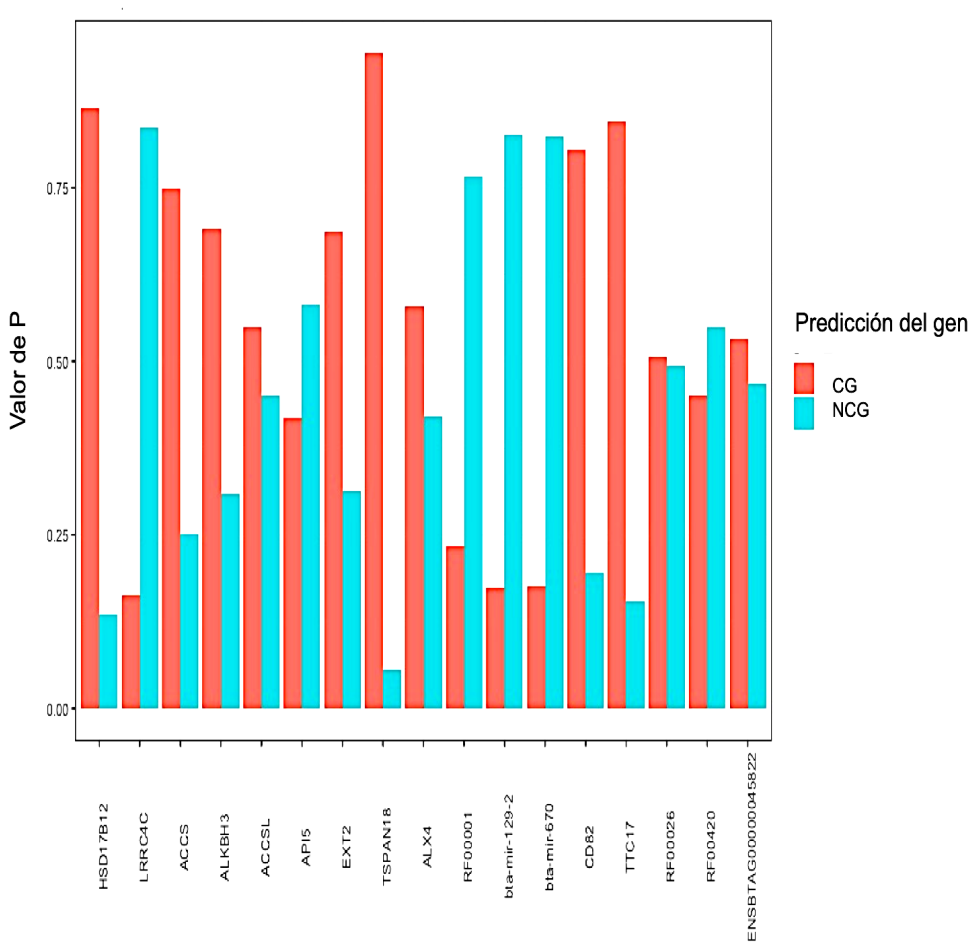
Gráfica 4. Evaluación de la exactitud de predicción de la red neuronal para diferentes valores de hiperparámetros: número de neuronas en la capa oculta (size = 10, 20, 50, 80, 100, 120) y el valor de regularización (decay = 0.0001, 0.1, 0.5). Se obtiene una exactitud de predicción de 83 % con un valor de size = 120 y decay = 0.1.

genéticas en bacterias aisladas de diferentes animales (cerdo, pollos de engorda, bovinos, patos y gallinas ponedoras). En general, los modelos de priorización mediante ANN se pueden ir mejorando, al integrar más variables biológicamente relevantes y una definición más precisa de la variable de respuesta.

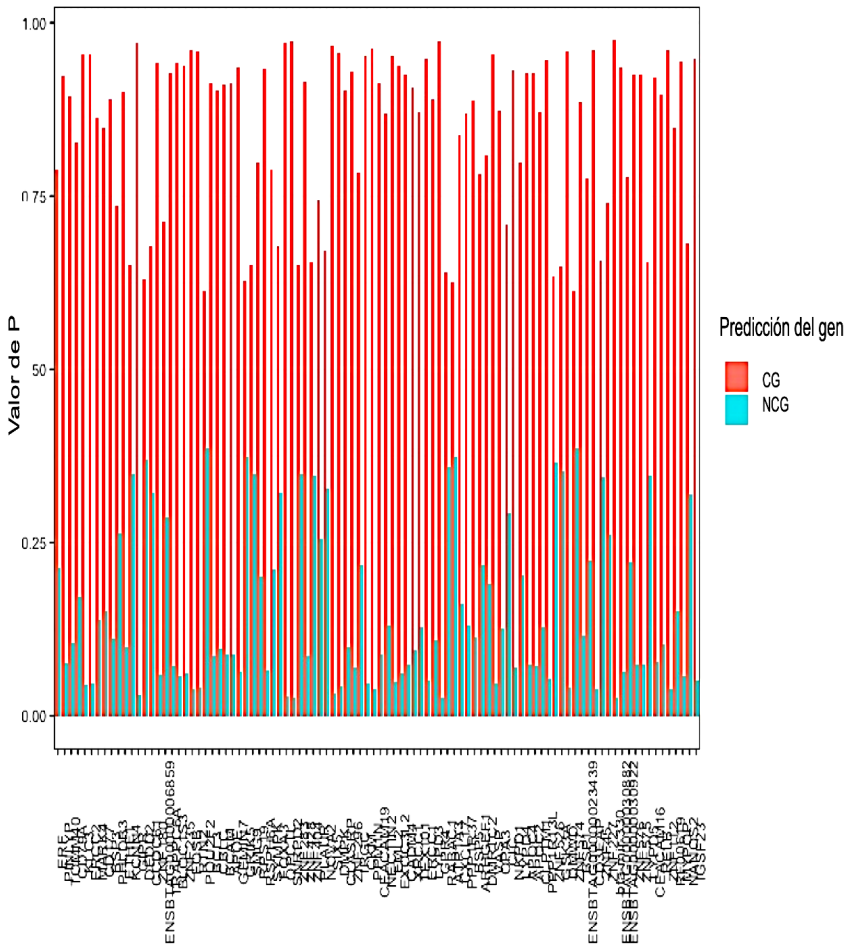
Es importante mencionar que los genes mendelianos reportados en bovinos no necesariamente se relacionan con DPR, lo que podría provocar inespecificidad en la red neuronal, es decir, los genes con una mayor probabilidad de ser causales de acuerdo con el modelo no necesariamente pudieran estar asociados con la DPR. Para reducir los efectos de la inespecificidad, los estudios de priorización suelen enfocarse en una lista reducida de genes ya sea porque presentan puntajes altos de probabilidad de estar asociados con el fenotipo (Fu *et al.,* 2020) o porque presentan anotaciones biológicas relacionadas con el caracter (Bargsten *et al.,* 2014). Por lo anterior, el análisis de priorización de genes suele estar acompañado de revisión de literatura e incluso de otros análisis estadísticos, por ejemplo, de sobrerrepresentación de procesos biológicos (Bargsten *et al.,* 2014).

Uso de la red neuronal para priorizar genes en QTLs asociados a DPR

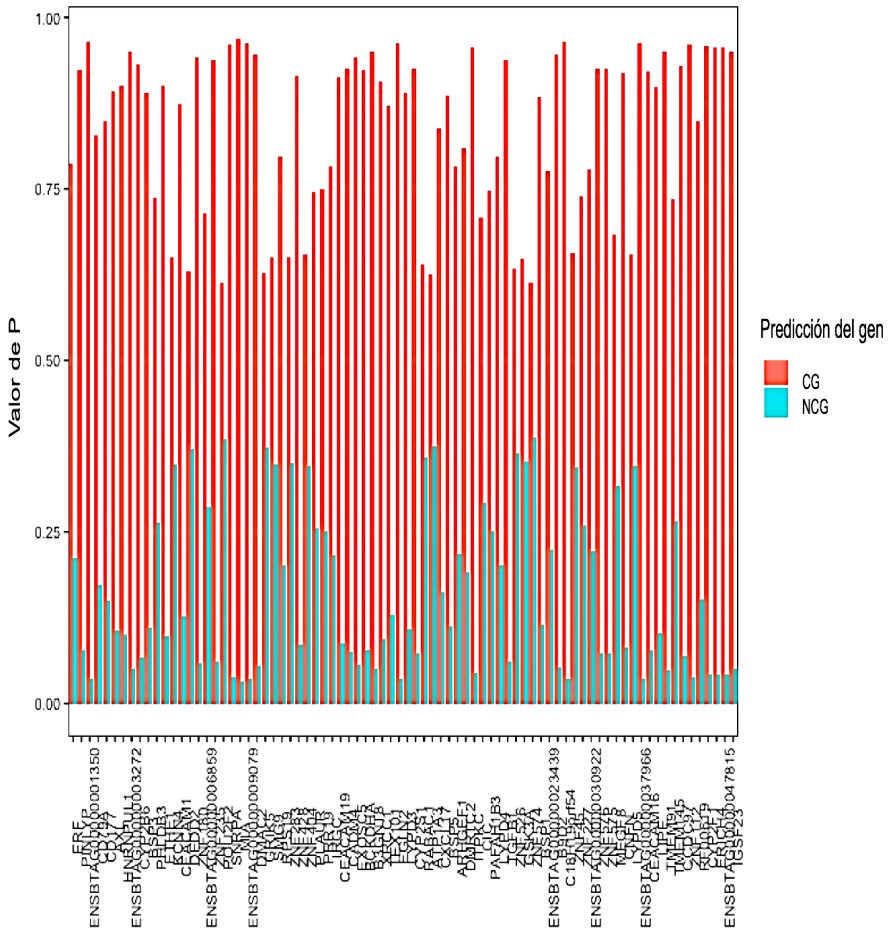
En las Gráficas 5-8 se muestran los valores de P para cada uno de los genes en los 4 QTLs investigados para DPR. Ya que se consideró un valor de $P \geq 0.95$ como referencia, sólo 29 genes de los 387 evaluados fueron priorizados por la red neuronal como causales para DPR. En el Cuadro 2 se muestra el valor de P, así como información posicional de estos genes.



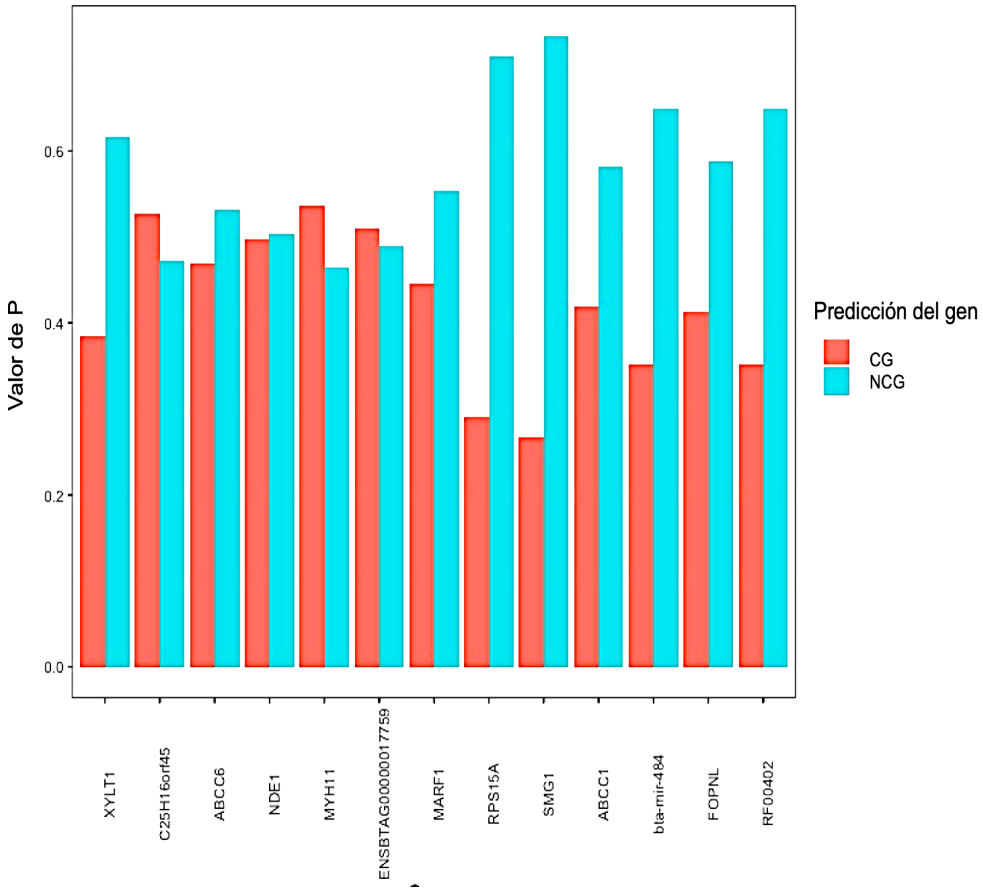
Gráfica 5. Priorización de genes en el QTL del cromosoma 15 en la posición 74679526 - 76475224 pb. Las barras en rojo indican la probabilidad de que el gen sea causal (CG) y en azul la probabilidad de no serlo (NCG).



Gráfica 6. Priorización de genes en el QTL del cromosoma 18 en la posición 49042269 - 52525163 pb. Las barras en rojo indican la probabilidad de que el gen sea causal (CG) y en azul la probabilidad de no serlo (NCG).



Gráfica 7. Priorización de genes en el QTL del cromosoma 18 en la posición 61872508 - 64704565 pb. Las barras en rojo indican la probabilidad de que el gen sea causal (CG) y en azul la probabilidad de no serlo (NCG).



Gráfica 8. Priorización de genes en el QTL del cromosoma 25 en la posición 14380150 - 16669292 pb. Las barras en rojo indican la probabilidad de que el gen sea causal (CG) y en azul la probabilidad de no serlo (NCG).

Cuadro 2. Genes con una P≥ 0.95 de ser causales para DPR en bovinos de acuerdo con la predicción de la red neuronal.

Valor de P	Identificador Ensembl	Nombre
0.9637964	ENSBTAG00000001350	NA
0.9608930	ENSBTAG00000009077	*SNRPA*
0.9626825	ENSBTAG00000009078	*MIA*
0.9626825	ENSBTAG00000009079	NA
0.9635026	ENSBTAG00000017996	*EGLN2*
0.9560170	ENSBTAG00000019465	*ITPKC*
0.9637299	ENSBTAG00000023453	*C18H19orf54*
0.9507760	ENSBTAG00000038711	*TMEM91*
0.9607225	ENSBTAG0000004020	*ZNF112*
0.9582339	ENSBTAG00000046805	*CYP2F1*
0.9568391	ENSBTAG00000047423	*ERICH4*
0.9544047	ENSBTAG00000002070	*KLC3*
0.9539523	ENSBTAG00000002072	*ERCC2*
0.9700948	ENSBTAG00000005527	*GIPR*
0.9612000	ENSBTAG00000008182	*FOSB*
0.9586540	ENSBTAG00000008185	*RTN2*
0.9707135	ENSBTAG00000012176	*QPCTL*
0.9734103	ENSBTAG00000012177	*SNRPD2*
0.9677627	ENSBTAG00000013346	*SIX5*
0.9573290	ENSBTAG00000013347	*DMPK*
0.9522987	ENSBTAG00000013921	*CKM*
0.9618580	ENSBTAG00000014187	*PPM1N*
0.9514046	ENSBTAG0000005764	*EML2*
0.9727692	ENSBTAG00000018398	*GPR4*
0.9534530	ENSBTAG00000019604	*VASP*
0.9595478	ENSBTAG00000021417	*DMWD*
0.9615483	ENSBTAG00000023601	*CD3EAP*
0.9745003	ENSBTAG00000029933	*bta-mir-330*
0.9607225	ENSBTAG00000040209	*ZNF112*

Los 29 genes predichos como causales para DPR se localizan en el cromosoma 18; para siete de ellos se encontró información biológica relevante con base en la revisión de literatura. El gen *FOSB y HSD17B12* han sido asociados con DPR en bovinos (Cocharan *et al.,* 2013). El resto de los genes discutidos están asociados a caracteres reproductivos de acuerdo con estudios en organismos modelos como el ratón y el humano (Sarkar *et al.,* 2004; Cerván-Martín *et al.,* 2020).

El gen *KLC3* (Cadena ligera de kinesina 3) de acuerdo con la predicción del modelo de ANN tiene una probabilidad de ser causal para DPR de un 95.44%. En humanos, se ha reportado que la baja expresión de este gen puede resultar en una función inadecuada de la pieza intermedia del espermatozoide, la cual tiene una acción importante en la motilidad, causando fenómenos como la oligozoozpermia y astenzoozpermia, afectando a su vez los parámetros del semen (Kargar-Dastjerdy *et al.,* 2016). En ratones, se ha observado que *KLC3* es la única quinesina de cadena ligera que se expresa en espermátidas. El alto nivel de expresión de un mutante de *KLC3* en ratones transgénicos también puede causar anomalías en la pieza intermedia y como resultado, se obtienen ratones subfértiles que tienen recuento reducido de espermatozoides con alta alteración de la motilidad (Zhang *et al.,* 2012).

Otro de los genes importantes fue *ERCC2* (reparación por escisión 2, subunidad de helicasa del complejo central TFIIH) del cual se sabe que es un componente de la vía NER y también un miembro integral del factor de transcripción basal BTF2/TFIIH. En humanos se ha reportado que la expresión reducida de *ERCC2* afectó significativamente los testículos, causando

azoospermia, en comparación con los testículos normales que muestran una expresión normal de *ERCC2* (Ji *et al.,* 2008). Así mismo, Vertika *et al.* (2019) encontraron una correlación de los niveles de expresión del gen *ERCC2* con infertilidad y azoospermia en humanos.

GIPR (Receptor de polipéptido inhibidor gástrico) afecta la fusión de los espermatozoides con el ovocito en ratones (Shimizu *et al*., 2017).

*FOSB (*Protooncogén FosB, subunidad del factor de transcripción AP-1). Cochran *et al.* (2013) utilizando el programa IPA encontraron cinco factores de transcripción (*HNF4A, TCF3, CTBO2, FOSB y SP100)* que están significativamente sobrerrepresentados como reguladores de los genes de DPR en bovinos de leche. También, se ha encontrado diferencialmente expresado en la pituitaria AFKP KO femeninas. Lo que sugiere que *FOSB* podría estar regulado por *ERL1* generando infertilidad en ratones (De Mess *et al.,* 2020).

SIX5 (SIX homeobox 5**)** fue reportado por Sarkar *et al.,* (2004) quienes realizaron un estudio en ratones, mencionan que la pérdida del gen produce esterilidad y pérdida progresiva de la masa testicular con la edad. Además, mencionan que la expresión reducida de *SIX5* causa problemas reproductivos y su expresión es un requisito estricto para la supervivencia de células espermatogénicas y para la espermiogénesis. En una población europea realizaron una evaluación de *loci* asociados a la fertilidad masculina con alteración espermatogénica grave donde también encontraron que *SIX5* influía significativamente (Cerván-Martín *et al.,* 2020).

*DMPK (*Proteína Quinasa de la Distrofia Miotonica) tiene una repetición de CTG que disminuye el volumen del esperma y la motilidad progresiva en los

espermatozoides de humanos (Puy *et al.* 2020). Kim *et al. (*2012) indicaron que la mutación de *DMPK* y *DMWD* causa distrofia miotónica 1 (DM1), que es un trastorno multisistémico dominante caracterizado por defectos endócrinos que incluyen atrofia testicular y tubular, oligospermia y azoospermia, así como niveles elevados de hormona folículo estimulante. Se sabe que este gen se expresa en niveles altos en los testículos y secundariamente en el cerebro y a niveles bajos en otros tejidos. Buckley *et al*. (2016) también reportaron que mutaciones en *DMPK*, así como en *DMWD, SIX5, BHMG1* y *RSPH6A* podrían contribuir a los problemas reproductivos masculinos en humanos.

Si se toma un umbral de probabilidad menos estricto para declarar a un gen como causal, por ejemplo, del 85 %, en el QTL del cromosoma 15 también destacan algunos genes relacionados a características reproductivas y uno relacionado a DPR. Por ejemplo, *TSPAN18 (*tetraspanina 18*)* que participa en la apoptosis (vía de señalización del calcio) (Zhang *et al.,* 2015), en ratones se sabe que se expresa en sistemas reproductivos femeninos (ovario) y masculinos (epidídimo, testículo, tubos seminíferos y glándulas accesorias) (Bult *et al.,* 2019).

El gen *HSD17B12* de la familia de hidroxiesteroide (17β) deshidrogenasa, fue reportado por Cochran *et al.* (2013) donde mencionan 5 vías de genes sobrerrepresentadas con SNPs asociados con DPR, de las cuales dos mostraron la participación de *HSD17B12*, que fueron la biosíntesis de estrógenos y señalización de cáncer de mama dependiente de estrógenos. Utilizando el software IPA también construyeron 4 redes de genes relacionados con DPR. La más relevante incluía 16 genes donde también se encontró *HSD17B12.* Se ha encontrado que se expresa en los ovocitos, células de la teca interna y granulosa,

cuerpo lúteo, epitelio superficial de ovarios en humanos y ratón, y se ha involucrado en la vía de síntesis de prostaglandinas, la función ovárica y la regulación de la fertilidad en mujeres. *HSD17B12* codifica una enzima que realiza la conversión de estrona en estradiol, y participa en la síntesis de ácido araquidónico (Kemiläinen *et al.,* 2016).

CONCLUSIONES Y RECOMENDACIONES

En este trabajo se logró integrar la información de diferentes repositorios (KEGG, OMIA, ENSEMBL, AnimalQTLdb) con lo cual se construyó un modelo de red neuronal para priorizar genes causales para DPR con una exactitud del 83 %. La estrategia de priorización de genes mediante redes neuronales resultó eficaz para el estudio de QTLs en bovinos y puede ser aplicada al estudio de cualquier característica y especie de ganado aprovechando el continuo crecimiento de la información en bases de datos.

Con este modelo se logró priorizar un total de 29 genes de los 387 localizados en los 4 QTLs investigados. De éstos, *KLC3, ERCC2, GIPR, FOSB, SIX5, DMPK, DMWD, HSD17B12* mostraron información biológica relevante de acuerdo con el análisis comparativo con diferentes bases de datos. Incluso la revisión de literatura respalda la asociación de algunos de ellos con DPR o con caracteres reproductivos relacionados.

Previo a la validación de los genes aquí identificados se recomienda probar nuevamente la red neuronal alimentando la matriz de entrenamiento con los genes para DPR reportados en la literatura, los cuales no fueron considerados en el presente estudio. Esto podría ayudar a tener una red neuronal con mayor especificidad para el carácter de interés.

También se recomienda considerar la información de otras bases de datos (por ejemplo, Gene Ontology, GEO y de secuencias) para tener una mayor cantidad de variables y que el estudio sea cada vez más integrativo. El análisis de cada base de datos implicará realizar análisis de minería de datos y la obtención de variables informativas.

Para tener una mayor cantidad de genes representados en la matriz de datos, se recomienda utilizar los genes ortólogos de humano o ratón, ya que son de los mamíferos más estudiados y por consiguiente sus genes tienen un mayor número de anotaciones en las bases de datos. Además, esto permitiría comparar los resultados con herramientas de priorización ya estandarizadas en estas especies.

LITERATURA CITADA

Abiodun, O. I., A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed y H. Arshad. 2018. State of the art in artificial neural network applications: a survey. Heliyon, 4:11.

Agatonovic-Kustrin, S y R. Beresford. 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J. Pharm. Biomed. Anal. J Pharmaceut. Biomed. 22:717-727.

Angermueller, C., T. Pärnamaa, L Parts, O. Stegle. 2016. Deep learning for computational biology. Mol Syst Biol. 12:878.

Aromolaran, O., T. Beder, M. Oswald, J. Oyelade, E.I Adebiyi y R. Koenig. 2020. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. Comput. Struct. Biotechnol. J. 18:612-621.

Ashtiani, M., M. Mirzaie, Mehdi y M. Jafari. 2018. CINNA: an R/CRAN package to decipher Central Informative Nodes in Network Analysis. Bioinformatics (Oxford, England). 35:10-1093.

Ashtiani, M., P. Nickchi, S. Jahangiri-Tazehkand, A. Safari, M. Mirzaie y M. Jafari. 2019. IMMAN: an R/Bioconductor package for Interolog protein network reconstruction, mapping and mining analysis. BMC Bioinformatics. 20:73.

Bargsten, J.W., J. P. Nap, J. F. Sanchez-Perez y A. D. J. Dijk. 2014. Prioritization of candidate genes in QTL regions based on associations between traits and biological processes. BMC Plant Biol. 14:330.

Bo, N. 2009. Practical cattle breeding in the future – commercialised or co-operative, across borderlines between countries and organisations. Vikinggenetics. 16:256-258.

Bono, H. 2020. All of gene expression (AOE): An integrated index for public gene expression databases. PLoS ONE. 15:1.

Buckley, L., M. Lacey y M. Ehrlich, M. 2016. Epigenetics of the myotonic dystrophy-associated DMPK gene neighborhood. Epigenomics. 8:13–31.

Bult, C. J., J. A. Blake, C. L. Smith, J. A. Kadin y J. E. Richardson. 2019. the Mouse Genome Database Group. 2019. Mouse Genome Database (MGD). Nucleic Acids Res. 47:801–806.

Cantor, R. M., K. Lange y J. S. Sinsheimer. 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American journal of human genetics. 86:6–22.

Carmo, A. S., M. Santos, L. Braga y A. Mascioli. 2019. PSVIII-35 The impact of daughter pregnancy rate on the reproductive efficiency of Brazilian dairy herds. J AniM Sci. 97:262.

Cerri, R. L., I. M. Thompson, I. H. Kim, A. D. Ealy, P. J. Hansen, C. R. Staples, J. L. Li, J. E. Santos y W. W. Thatcher.2012. Effects of lactation and gestation on gene expression in the endometrium of Holstein cows on day 17 of the oestrous cycle or gestation. J Dairy Sci. 95:5657-5675.

Cerván-Martín, M., L. Bossini-Castillo, R. Rivera-Egea, N. Garrido, S. Luján, G. Romeu, S. Santos-Ribeiro, I. Group, L. C. Group, J. A. Castilla, M. C. Gonzalvo, A. Clavero, F. J. Vicente, A. Guzmán-Jiménez, C. Costa, I. Llinares-Burguet, C. Khantham, M. Burgos, F. J. Barrionuevo, R. Jiménez, J. Sánchez-Curbelo, O. López-Rodrigo, M. F. Peraza, I. Pereira-Caetano, P. I. Marqués, F. Carvalho, A.

Barros, L. Bassas, S. Seixas, J. Gonçalves, S. Larriba, A. M. Lopes, R. J. Palomino-Morales, F. D. Carmona. 2020. Evaluation of Male Fertility-Associated Loci in a European Population of Patients with Severe Spermatogenic Impairment. J. Pers. Med. 11: 22.

Chebel, R.C y E.S. Ribeiro. 2016. Reproductive systems of North American dairy cattle herds. Vet Clin North Am Food Anim Pract. 32:267-284.

Cochran, S. D., J. B. Cole, D. J. Null y P. J. Hansen. 2013. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. 14:49.

Cochran, S.D., J. B. Cole, D. J. Null y P. J. Hansen. 2013. Discovery of single nucleotide polymorphisms in candidate genes associated with fertility and production traits in Holstein cattle. BMC Genet. 14:49.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. Van Tassell, J. Yang, S. Wang, L. K. Matukumalli y Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics. 12:408.

Dave, V.S. y K. Dutta. 2014. Neural Network-Based Models for Software Stress Estimation: A Review. Artif. Intell. Rev. 20: 295-307.

De Mees, C., L. Jean-Francois, J. Bakker, J. Smitz, B. Hennuy, P. Van Vooren, P. Gabant, J. Szpirer y C. Szpirer. 2020. Alpha-Fetoprotein Controls Female Fertility and Prenatal Development of the Gonadotropin-Releasing Hormone Pathway through an Antiestrogenic Action. Molecular and Cell Biology. 26:5.

Deo, R. C., G. Musso, M. Tasan, P. Tang, A. Poon, C. Yuan, J. F. Felix, R. S. Vasan, R. Beroukhim, T. De Marco, P. Y. Kwok, C. A. MacRae y F P. Roth. 2014. Prioritizing causal disease genes using unbiased genomic features. Genome Biol. 15:534.

Durinck, S., P. Spellman, E. Birney, W. Huber W. 2009. Identifier mapping for the integration of genomic data sets with the R / Bioconductor biomaRt package. Nature Protocols. 4:1184-1191.

Ensembl versión 104 en: https://www.ensembl.org/index.html consultado Agosto 2020.

Fair, D. A., N. U. Dosenbach, J. A. Church, A. L. Cohen, S. Brahmbhatt, F. M. Miezin, D. M. Barch, M. E. Raichle, S. E. Petersen y B. L. Schlaggar BL. 2007. Development of distinct control networks through segregation and integration. Proc Natl Acad Sci U. S. A. 104:13507-12.

Fonseca, P., A. Suárez-Vega y A. Cánovas. 2020. Weighted Gene Correlation Network Meta-Analysis Reveals Functional Candidate Genes Associated with High- and Sub-Fertile Reproductive Performance in Beef Cattle. Genes(Basel). 11:543.

Fonseca, P., F. C. Dos Santos, S. Lam, A. Suárez-Vega, F. Miglior, F. S. Schenkel, L. Diniz, S. Id-Lahoucine, C. M.arvalho y A. Cánovas. 2018. Genetic mechanisms underlying spermatic and testicular traits within and among cattle breeds: systematic review and prioritization of GWAS results. J Dairy Sci. 12:4978–4999.

Fu, Y., J. Xu, Z. Tang, L. Wang, D. Yin, Y. Fan, D. Zhang, F. Deng, Y. Zhang, H. Zhang, H. Wang, W. Xing, L. Yin, S. Zhu, M. Zhu, M. Yu, X. Li, X. Liu, X. Yuan y S. Zhao. 2020. Gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. Commun Biol. 10:1:502.

Gaddis, K. L. P., D. J. Null y J. B. Cole. 2016. Explorations in genome-wide association studies and network analysis with dairy cattle fertility traits. J. Dairy Sci. 99:6420-6435.

Goddard, M. E., B. J. Hayes y T. H. Meuwissen. 2010. Genomic selection in livestock populations. Genet Res. 92:413-21.

Gutiérrez-Reinoso, M. A., P. M. Aponte y M. García-Herreros. 2021. Genomic analysis, advances and future perspectives in the selection of dairy cattle: a review. Animals. 11:1- 49.

Hagen, D. E., D. R. Unni, A. Tayal, B. G. W. urns y C. G. Elsik. 2018. Bovine Genome Database: Tools for Mining the *Bos taurus* Genome. Methods in molecular biology. 1757:211–249.

Hassani-Pak, K. y C. Rawlings. 2017. Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. Journal of integrative bioinformatics. 14:1-19.

Hernández-Ramos, P., A. M. Vivar-Quintana, I. Revilla, M.I. González-Martín, M. Hernández-Jiménez y I. Martínez-Martín. 2020. Prediction of Sensory Parameters of Cured Ham: A Study of the Viability of the Use of NIR Spectroscopy and Artificial Neural Networks. Sensors (Basel). 20:5624.

Hoglund, J. K., G. Sahana, B. Guldbrandtsen y M. S. Lund. 2014. Validation of associations for female fertility traits in Nordic Holstein, Nordic Red and Jersey dairy cattle. BMC Genet. 15:1-8.

Hu, Z. L., C. A. Park y J. M. Reecy. 2019. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic acids research. 47:701-710.

Hulsegge, I., H. Woelders, M. Smits, D. Schokker, L. Jiang y P. Sorensen. 2013. Prioritization of candidate genes for cattle reproductive traits, based on protein-protein interactions, gene expression, and text-mining. Physiological genomics. 45:400–406.

Igraph paquete de R versión 1.2.6 en : https://igraph.org/r/ consultado Agosto 2020.

Ji, G., A. Gu, Y. Xia, C. Lu, J. Liang, S. Wang, J. Ma, Y. Peng y X. Wang. 2008. *ERCC1* and *ERCC2* polymorphisms and risk of idiopathic azoospermia in a Chinese population. Reprod Biomed. 17:36-41.

Jiang, L., P. Sorensen, B. Thomsen, S. M. Edwards, A. Skarman, C. M. Røntved, M. S. Lund y C. T. Workman. 2012. Gene prioritization for livestock diseases by data integration. Physiological genomics. 44:305–317.

Joaquin. A. R. 2019. Decision trees, random forest, gradient boosting and C5.0 available under a Attribution 4.0 International en: https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_C50.html consultado agosto 2020

Joaquin. A. R. 2019. Machine Learning con R y caret en: https://www.cienciadedatos.net/documentos/41_machine_learning_con_r_y_caret#Creaci%C3%B3n_de_un_modelo_predictivo consultado agosto 2020.

Junior, J.L., A. J. D. Neto, L. B. Neto, F. J. C. P. Soeiro., C. C. Santana y H.F.C Velho. 2011. Application of artificial neural networks and hybrid methods in the solution of inverse problems. Artificial Neural Networks - Application. In Tech. April. 11.

Karami, K., S. Zerehdaran, A. Javadmanesh, M. M. Shariati y H. Fallahi. 2019. Characterization of bovine (*Bos taurus*) imprinted genes from genomic to amino acid attributes by data mining approaches. PloS one. 14:6.

Kargar- Dastjerdy, P., M. Tavalaee, M. Salehi, M. Falahati, T. Izadi, M. H. Nasr. 2016. Altered expression of *KLC3* may affect semen parameters. J Reprod Biomed. 14:15-22.

Kargo, M., L. Hjorto, M. Toivonen, J. A. Eriksson, G. P. Amand, J. Pedersen. 2014. Economic basis for the nordic total merit index. J. Dairy Sci. 97:7879-7888.

KEGG versión 2019 en: https://www.genome.jp/kegg/ consultado en septiembre 2020.

Keith, J., V. Vassilev-Galindo, B. Cheng, Bingqing, S. Chmiela, G. Stefan, M. Michael y T. Klaus-Robert. 2012. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. Chemical Reviews. 121:9816-9872.

Kim, W. B., J. Y. Jeong, S. W. Doo, W. J. Yang, Y. S. Song, S. R. Lee, J. W. Park y D. W. Kim. 2012. Myotonic dystrophy type 1 presenting as male infertility. Korean J Urol. 53:134-6.

Kiser, J. N., E. M. Keuter, C. M. Seabury, M. Neupane, J. Morales, J. Dalton, G. W. Burns, T. E. Spenser y H. L. Neibergs. 2019. Validation of 46 loci associated with female fertility traits in cattle. BMC Genomics. 20:576.

Klasberg, S., T. Bitard-Feildel y L. Mallet. 2016. Computational Identification of Novel Genes: Current and Future Perspectives. Bioinform Biol Insights. 10:121-31.

Kommadath, A., H. Woelders, M. Beerda, H. Mulder, A. Wit, R. Veerkamp, M. Pas y M. Smit. 2011. Gene expression patterns in four brain areas associate with quantitative measure of estrous behavior in dairy cows. BMC genomics. 12:200.

Koyama, K. y T. Takahashi. 2020. Relationship between sire predicted transmitting ability for daughter pregnancy rate and daughter's reproductive performance and milk production in Japanese dairy herds. The Journal of reproduction and development. 66:445–452.

Kemiläinen, K., M. Adam, J. Mäki-Jouppila, P. Damdimopoulou, A. E. Damdimopoulos, J. Kere, O. Hovatta, T. D. Laajala, T. Aittokallio, J. Adamski, H. Ryberg, C. Ohlsson, L. Strauss y M. Poutanen. 2016. The Hydroxysteroid (17β) Dehydrogenase Family Gene *HSD17B12* Is Involved in the Prostaglandin Synthesis Pathway, the Ovarian Function, and Regulation of Fertility. Endocrinology. 157:3719–3730.

Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan y V. Carey. 2013. Software to calculate and annotate genomic ranges. PLoS Comput. Biol. 9(8):e1003118.

Le, N. Q. K., T. D. Do, T. N. K. Hung, L. H. T. Lam, T. T. Huynh y N. T. 2020. A computational framework based on ensemble deep neural networks for essential genes identification. International journal of molecular sciences. 2123:1-16.

Liaw, A. y M. Wiener. 2007. Classification and Regression by RandomForest. Forest. 23.

Lin, F., E. Z. Lazarus y S. Y. Rhee. 2020. QTG-Finder2: A Generalized Machine-Learning Algorithm for Prioritizing QTL Causal Genes in Plants. Genes-Genomes- Genetics. 7:2411-2421.

Lin, H., K. A. Hargreaves, R. Li, J. L. Reiter, Y. Wang, M. Mort, D. N. Cooper, Y. Zhou, C. Zhang, M. T. Eadon, M. E. Dolan, J. Ipe, T. C. Skaar y Y. Liu. 2020. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. Genome biology. 20:254.

Liu, Z., J. Jaitner, F. Reinhardt, E. Pasman, S. Rensing, R. Reents. 2009. Genetic evaluation of fertility traits of dairy cattle using a multiple-trait animal model. J Dairy Sci. 91:4333–4343.

López-López, M., A. López, T. Sainz, A. Rosales. 2005. What do you know about Genomics? J. Pharm. Sci. 36:42-44.

Lucy, M. C. 2001. Reproductive Loss in High-Production Dairy Cattle: Where Will It End?

Martínez, C.A., C. Manrique y M. Elzo. 2012. Cattle genetic evaluation: a historical perception. Rev Colom Cienc Pecua. 25:293:311.

Michelizzi, V.N., X. Wu, M.V. Dodson, J.J. Michal, J. Zambrano-Varon, D.J. McLean y Z. Jiang. 2011. An overview of 54,001 single nucleotide polymorphisms (SNPs) in the Illumina BovineSNP50 BeadChip and their transferability to water buffalo. Int. J. Biol. Sci. 7:18-27.

Miglior, F., B. L. Muir y B. J. Van Doormaal. 2005. Selection indices in Holstein cattle from various countries. J Dairy Sci. 88:1255-1263.

Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray y P. M. Visscher. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS genetics.11:1-4.

Munck, N., P. Njage, P. Leekitcharoenphon, E. Litrup y T. Hald. 2020. Application of Whole-Genome Sequences and Machine Learning in Source Attribution of *Salmonella Typhimurium*. Risk Anal. 40:1693–1705.

Nguyen, L. T., A. Reverter, A. Cánovas, B. Venus, S. T. Anderson, A. Islas-Trejo, M. M. Dias, N. F. Crawford, S. A. Lehnert, J. F. Medrano, M. G. Thomas, S.S. Moore y M. R. S. Fortes. 2018. *STAT6*, *PBX2*, and *PBRM1* Emerge as Predicted Regulators of 452 Differentially Expressed Genes Associated With Puberty in Brahman Heifers. Front Genet. 20:9:87.

Nicholas, F.W. 2003. Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. Nucleic Acids Res. 1:275-7.

Norman, H. D., J. R Wright, S. M. Hubbard, R.H. Miller, J.L. y Hutchison. 2009. Reproductive status of Holstein and Jersey cows in the United States. J Dairy Sci. 92:3517–3528.

Oludare I., A. Abiodun, J. Abiodun, E. Omolara, K. V. Dada, N. A. Mohamed y H Arshad. 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon. 4:11.

Orgogozo, V., B. Morizot, y A. Martin. 2015. The differential view of genotype-phenotype relationships. Frontiers in genetics. 6:179.

Paredes-Sánchez, F. A., A. M. Sifuentes-Rincón, A. Segura Cabrera, A. García, G. M. Parra y P. Ambríz. 2015. Associations of SNPs located at candidate genes to bovine growth traits, prioritized with an interaction networks construction approach. BMC Genet. 16:91.

Patel, J.L y R. K. Goyal. 2007. Applications of artificial neural networks in medical science. Curr. Clin. Pharmacol. 23:217–226.

Phanther classification system versión 16. en: http://pantherdb.org/panther/summaryStats.jsp consultado 6 septiembre 2021.

Pritchard, T., M. Coffey, R. Mrode y E. Wall. 2013. Genetic parameters for production, health, fertility and longevity traits in dairy cows. Animal. 34–46.

Pryce, J. E. y R. F. Veerkamp. 2001. Incorporation of fertility indices in genetic improvement programs. BSAP Occas Publ. 26:237–249.

Puy, V., A. Mayeur, A. Levy, L. Hesters, J. Raad, S. Monnot, J. Steffann y N. Frydman. 2020. CTG Expansion in the *DMPK* Gene: Semen Quality Assessment and Outcome of Preimplantation Genetic Diagnosis. J Clin Endocrinol Metab. 105:3.

Salih, H y D. L. Adelson. 2009. QTL global meta-analysis: are trait determining genes clustered? BMC Genomics. 10:184

Sarkar, P. S., S. Paul, J. Han y S. Reddy. 2004. *SIX5* is required for spermatogenic cell survival and spermiogenesis. Hum Mol Genet. 13:1421-31.

Sarker, I. H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN comput Sci. 2:160.

Sevilla, S. D. 2007. Metodología de los estudios de asociación genética. Insuficiencia cardiaca. 2:111-114.

Shamimuzzaman, M. D., J. Justin, L. Tourneau, D. R. Unni, C. M. Diesh, D. A Triant, A. T. Walsh, A. Tayal, G. C. Conant, D. E. Hagen y C. G Elsik. 2020. Bovine Genome Database: new annotation tools for a new reference genome. Nucleic Acids Research. 48:676–681.

Shao, B., H. Sun, M. J. Ahmad, N. Ghanem, H. Abdel-Shafy, C. Du, T. Deng, S. Mansoor, Y. Zhou, Y. Yang, S. Zhang, L. Yang y G. Hua. 2021. Genetic Features of Reproductive Traits in Bovine and Buffalo: Lessons From Bovine to Buffalo. Front. Genet. 12:617-128.

Sharma, A., S. J. Lee, C.G. Dang, P. Sudrajad, y H.A Cheol. 2015. Stories and Challenges of Genome Wide Association Studies in Livestock - A Review. J. Anim. Sci. 28:1371-1379.

Snelling, W.M., R. A. Cushman, J. W. Keele, C. Maltecca, M. G. Thomas, M. R. Fortes y A. Reverter. 2013. Breeding and Genetics Symposium: networks and pathways to guide genomic selection. J Anim Sci. 91:537-552.

Spataro, N., J. A. Rodríguez, A. Navarro y E. Bosch. 2017. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Hum. Mol. Genet. 26:489–500.

String versión 11.5 en: https://string-db.org/ consultado agosto 2020.

Tellam, R. L., D. G. Lemay, T. Van, C. P. Tassell, H. A. Lewin, K. C. Worley y G. C. Elsik. 2009. Unlocking the bovine genome. BMC Genomics. 193.

Tian, D., S. Wenlock, M. Kabir, G. Tzotzos, A. J. Doig y K. E. Hentges. 2018. Identifying mouse developmental essential genes using machine learning. Dis Model Mech. 11.

Tickle, A. B., R. Andrews, M. Golea y J. Diederich. 1998. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. IEEE Trans Neural Netw. 9:1057-68.

Tranchevent, L. C., A. Ardeshirdavani, S. ElShal, D. Alcaide, J. Aerts, D. Auboeuf y Y. Moreau. 2016. Candidate gene prioritization with Endeavour. Nucleic acids research. 44:117–121.

Tu, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 49:1225-31.

VanRaden, P. M. 2004. Invited Review: Selection on Net Merit to Improve Lifetime Profit. J. Dairy Sci. 87:3125-3131.

VanRaden, P. M. 2003. Longevity and fertility trait definitions compared in theory and simulation. Interbull Bull. 30:43-46.

Vertika, S., K. M. Sandeep, D. V. S. Sudhakar, A. Chakraborty, S. Trivedi, G. Gupta, K. Thangaraj, S. Rajander y K. Singh. 2019. SNPs in *ERCC1*, *ERCC2*, and *XRCC1* genes of the DNA repair pathway and risk of male infertility in the Asian populations: association study, meta-analysis, and trial sequential analysis. J Assist Reprod Genet. 36:79–90.

Von-Lilienfeld, O. A., K. R. Müller y A. Tkatchenko. 2020. Exploring chemical compound space with quantum-based machine learning. Nat Rev Chem. 4:347–358.

Weller, J. I., D. M. Bickhart, G. R. Wiggans, M. E. Tooker, J. R. O'Connell, J. Jiang, M. Ron, y P. M. VanRaden. 2018. Determination of quantitative trait nucleotides by concordance analysis between quantitative trait loci and marker genotypes of US Holsteins. Dairy Sci. 101:9089–9107.

Wiggans, G. R., P. M. Vanraden y T. A. Cooper. 2011. The Genomic Evaluation System in the United States: Past, Present, Future. J Dairy Sci. 94: 3202-3211.

Wu, G.D., J. Chen, C. Hoffmann, K. Bittinger, Y. Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman y J. D. Lewis. 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science. 334:105-8.

Xiao, N., C. Dong-Sheng, Z. Min-Feng y X. Qing-Song. 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 31:1857–1859.

Yandell, M. y D. Ence. 2012. A Beginner's Guide to Eukaryotic Genome Annotation. Nat Rev Genet.13:329-42.

Zhang, Y. E., P. Landback, M. Vibranovski, M. Long. 2012 New genes expressed in human brains: implications for annotating evolving genomes. Bioassays. 34:982–91.

Zhang, Y., O. Young, C. Min, S. Saadi, H. Thundathil, J.Hoorn y A. Frans. 2012. *KLC3* is involved in sperm tail midpiece formation and sperm function. Developmental biology. 366:101-10.

Zhang, Z., Z. Wang, Y. Yang y J. Zhao. 2016. Identification of pleiotropic genes and sets underlying growth and immunity traits: a case study on Meishan pigs. Animal. 10: 550-557.

Zhu, M., y J. Dong. 2016. rDNAse: R package for generating various numerical representation schemes of DNA sequences.

Zimin, A.V., A. L. Delcher, L. Florea, D. R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke y S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol.10:1-10.

Zhang, B., D. X. Li, N. L u, Q. R. Fan, W. H. L i y Z. F. Feng. 2015. Lack of association between the *TSPAN18* gene and schizophrenia based on new data from Han chinese and a meta-analysis. Int. J. Mol. Sci. 16(6):11864-72.

APÉNDICE

Cuadro 1. Genes de Entrenamiento para la Red Neuronal.

ENSBTAG00000023652, ENSBTAG00000040393, ENSBTAG00000018777,
ENSBTAG00000005122, ENSBTAG00000017490, ENSBTAG00000014461,
ENSBTAG00000017060, ENSBTAG00000010658, ENSBTAG00000018936,
ENSBTAG00000017442, ENSBTAG00000000898, ENSBTAG00000020199,
ENSBTAG00000033983, ENSBTAG00000025822, ENSBTAG00000001601,
ENSBTAG00000001578, ENSBTAG00000018419, ENSBTAG00000011934,
ENSBTAG00000014824, ENSBTAG00000019079, ENSBTAG00000006784,
ENSBTAG00000005268, ENSBTAG00000032068, ENSBTAG00000005269,
ENSBTAG00000014890, ENSBTAG00000025644, ENSBTAG00000027991,
ENSBTAG00000021181, ENSBTAG00000003108, ENSBTAG00000010531,
ENSBTAG00000021102, ENSBTAG00000011643, ENSBTAG00000031347,
ENSBTAG00000007486, ENSBTAG00000007073, ENSBTAG00000007284,
ENSBTAG00000016385, ENSBTAG00000019971, ENSBTAG00000008338,
ENSBTAG00000008056, ENSBTAG00000002758, ENSBTAG00000004281,
ENSBTAG00000013854, ENSBTAG00000022564, ENSBTAG00000039289,
ENSBTAG00000021368, ENSBTAG00000037453, ENSBTAG00000002683,
ENSBTAG00000008436, ENSBTAG00000017475, ENSBTAG00000047223,
ENSBTAG00000001936, ENSBTAG00000013301, ENSBTAG00000013303,
ENSBTAG00000008291, ENSBTAG00000008938, ENSBTAG00000014600,
ENSBTAG00000019145, ENSBTAG00000005287, ENSBTAG00000001299,
ENSBTAG00000010709, ENSBTAG00000019246, ENSBTAG00000010419,
ENSBTAG00000010889, ENSBTAG00000002302, ENSBTAG00000000087,
ENSBTAG00000007148, ENSBTAG00000016267, ENSBTAG00000017714,
ENSBTAG00000009876, ENSBTAG00000006984, ENSBTAG00000038496,
ENSBTAG00000023177, ENSBTAG00000040409, ENSBTAG00000039995,
ENSBTAG00000008453, ENSBTAG00000017722, ENSBTAG00000013275,
ENSBTAG00000013411, ENSBTAG00000004362, ENSBTAG00000013298,
ENSBTAG00000015086, ENSBTAG00000005397, ENSBTAG00000038171,
ENSBTAG00000006745, ENSBTAG00000001638, ENSBTAG00000022120,
ENSBTAG00000013326, ENSBTAG00000003092, ENSBTAG00000016675,
ENSBTAG00000008743, ENSBTAG00000006608, ENSBTAG00000013973,
ENSBTAG00000003234, ENSBTAG00000006208, ENSBTAG00000019267,
ENSBTAG00000005685, ENSBTAG00000006396, ENSBTAG00000015917,
ENSBTAG00000003871, ENSBTAG00000013125, ENSBTAG00000030882,
ENSBTAG00000014583, ENSBTAG00000020872, ENSBTAG00000038735,
ENSBTAG00000013099, ENSBTAG00000007436, ENSBTAG00000005534,
ENSBTAG00000013392, ENSBTAG00000012103, ENSBTAG00000021113,
ENSBTAG00000039161, ENSBTAG00000010576, ENSBTAG00000010184,
ENSBTAG00000016253, ENSBTAG00000002726, ENSBTAG00000008151,
ENSBTAG00000001937, ENSBTAG00000004736, ENSBTAG00000001061,
ENSBTAG00000004216, ENSBTAG00000004241, ENSBTAG00000014447,
ENSBTAG00000011196, ENSBTAG00000011193, ENSBTAG00000007153,
ENSBTAG00000017457, ENSBTAG00000009995, ENSBTAG00000032705,
ENSBTAG00000014239, ENSBTAG00000010989, ENSBTAG00000016149,

ENSBTAG00000015478, ENSBTAG00000021527, ENSBTAG00000000085,
ENSBTAG00000018843, ENSBTAG00000004063, ENSBTAG00000021717,
ENSBTAG00000006270, ENSBTAG00000017636, ENSBTAG00000014547,
ENSBTAG00000011628, ENSBTAG00000007131, ENSBTAG00000020455,
ENSBTAG00000021724, ENSBTAG00000010026, ENSBTAG00000008719,
ENSBTAG00000020645, ENSBTAG00000009586, ENSBTAG00000004514,
ENSBTAG00000013245, ENSBTAG00000010007, ENSBTAG00000000778,
ENSBTAG00000019762, ENSBTAG00000047039, ENSBTAG00000037533,
ENSBTAG00000046158, ENSBTAG00000023026, ENSBTAG00000023198,
ENSBTAG00000004729, ENSBTAG00000016629, ENSBTAG00000014453,
ENSBTAG00000020921, ENSBTAG00000016156, ENSBTAG00000012927,
ENSBTAG00000013669, ENSBTAG00000047238, ENSBTAG00000014465,
ENSBTAG00000012201, ENSBTAG00000047379, ENSBTAG00000040014,
ENSBTAG00000007049, ENSBTAG00000005928, ENSBTAG00000003636,
ENSBTAG00000015450, ENSBTAG00000037795, ENSBTAG00000024700,
ENSBTAG00000012697, ENSBTAG00000023939, ENSBTAG00000014335,
ENSBTAG00000004330, ENSBTAG00000018365, ENSBTAG00000009501,
ENSBTAG00000003572, ENSBTAG00000001244, ENSBTAG00000045717,
ENSBTAG00000033398, ENSBTAG00000012380, ENSBTAG00000012667,
ENSBTAG00000005947, ENSBTAG00000023032, ENSBTAG00000007876,
ENSBTAG00000008683, ENSBTAG00000012510, ENSBTAG00000005069,
ENSBTAG00000005772, ENSBTAG00000012054, ENSBTAG00000013003,
ENSBTAG00000046644, ENSBTAG00000017852, ENSBTAG00000005976,
ENSBTAG00000017056, ENSBTAG00000046797, ENSBTAG00000013016,
ENSBTAG00000011953, ENSBTAG00000004037, ENSBTAG00000003774,
ENSBTAG00000004688, ENSBTAG00000002979, ENSBTAG00000000497,
ENSBTAG00000026181, ENSBTAG00000001992, ENSBTAG00000002714,
ENSBTAG00000001908, ENSBTAG00000009520, ENSBTAG00000012653,
ENSBTAG00000032288, ENSBTAG00000008895, ENSBTAG00000001658,
ENSBTAG00000021761, ENSBTAG00000007829, ENSBTAG00000005623,
ENSBTAG00000005464, ENSBTAG00000000286, ENSBTAG00000016906,
ENSBTAG00000004021, ENSBTAG00000011082, ENSBTAG00000002313,
ENSBTAG00000009778, ENSBTAG00000019603, ENSBTAG00000018137,
ENSBTAG00000019741, ENSBTAG00000008612, ENSBTAG00000004840,
ENSBTAG00000011274, ENSBTAG00000019782, ENSBTAG00000014731,
ENSBTAG00000012265, ENSBTAG00000026501, ENSBTAG00000019574,
ENSBTAG00000004943, ENSBTAG00000035319, ENSBTAG00000014463,
ENSBTAG00000034501, ENSBTAG00000015729, ENSBTAG00000006991,
ENSBTAG00000034139, ENSBTAG00000016007, ENSBTAG00000000646,
ENSBTAG00000005773, ENSBTAG00000000438, ENSBTAG00000039362,
ENSBTAG00000032350, ENSBTAG00000037613, ENSBTAG00000039991,
ENSBTAG00000022205, ENSBTAG00000003523, ENSBTAG00000039647,
ENSBTAG00000015047, ENSBTAG00000004040, ENSBTAG00000011952,
ENSBTAG00000020048, ENSBTAG00000004709, ENSBTAG00000002350,
ENSBTAG00000033278, ENSBTAG00000006642, ENSBTAG00000001793,
ENSBTAG00000024450, ENSBTAG00000031387, ENSBTAG00000009646,
ENSBTAG00000018872, ENSBTAG00000048122, ENSBTAG00000005293,

ENSBTAG00000010069, ENSBTAG00000020087, ENSBTAG00000005574, ENSBTAG00000000164, ENSBTAG00000019554, ENSBTAG00000009733, ENSBTAG00000025760, ENSBTAG00000015358, ENSBTAG00000016138, ENSBTAG00000001271, ENSBTAG00000020079, ENSBTAG00000021741, ENSBTAG00000004003, ENSBTAG00000002015, ENSBTAG00000000894, ENSBTAG00000005039, ENSBTAG00000009287, ENSBTAG00000016097, ENSBTAG00000015769, ENSBTAG00000017639, ENSBTAG00000019852, ENSBTAG00000020789,ITSN1, GART, URB1, RF00026, SAMSN1, RBM11, LIPI, GBE1, CMSS1, FILIP1L, PHLDB2, ENSBTAG00000038422, OSTN, CLDN16, LIPH, NLGN1, SLC2A2, RF00600, ITGB2, COLQ, LVRN, ARSB, SERINC5, CLN6, HEXA, MYH7, MYH6, DPH6, PPIB, RF01241, MYO5C, CYP19A1, GALK2, COPS2, FBN1, SLC12A1, KIAA0586, KCNH5, PIGH, NPC2, LTBP2, GALC, SPAST, SLC3A1, MSH6, EDAR, POLR1B, RPIA, CAD, APOB, GDF7, GGTA1, ASS1, TTF1, PAEP, SNAPC4, ABO, MED22, COL5A1, SUGT1, RNASEH2B, ATP7B, FLT3, URAD, RF01161, PCDH9, EDNRB, UGGT2, MBNL2, EFNB2, ARGLU1, FLRT3, ARMC3, KIAA1217, DCLRE1C, ENSBTAG00000027444, PRNP, ATRN, EDN3, KIF3B, ASIP, CEP250, ENSBTAG00000045579, SLC39A4, MROH1, RF00426, WDR97, MAF1, SHARPIN, CYC1, GPAA1, EXOSC4, OPLAH, ENSBTAG00000015040, SPATC1, GRINA, PLEC, TG, CHD7, DPYS, ENSBTAG00000044190, RF00026, GRIA4, TRPC6, CWC15, BCO2, HMBS, CYP2R1, SMPD1, TRPC2, bta-mir-2317, RF00001, LRRC4C, LRP4, DDB2, MYBPC3, ENSBTAG00000024647, RF00403, ENSBTAG00000008449, FMO3, FMO4, FASLG, PLOD1, TNNT2, AGRN, TNR, LAMC2, EDNRA, HHIP, RF00026, TBX3, TBX5, TRPV4, LIMK2, MC1R, ESRP2, POP4, bta-mir-2901, CCNE1, URI1, SLC7A9, SPTBN4, BCKDHA, PPP1R13L, OPA3, PNKP, CA10, ENSBTAG00000047948, TADA2A, AIPL1, KIF1C, CHRNE, CHRNB1, HES7, PFAS, COL1A1, ITGA3, DLX3, B4GALNT2, SP6, RAPGEFL1, KRT25, KRT14, WNK4, SLC4A1, ITGA2B, MRC2, GH1, SCN4A, GAA, ITGB4, OCA2, MSTN, PDE1A, HOXD3, ACVR1, RND3, ZEB2, ENSBTAG00000024632, THSD7B, DPP10, PLCL1, ABCA12, PRKAG3, IHH, PAX3, TMEM39B, KHDRBS1, SLC9A1, ENSBTAG00000047333, HEXB, GHR, SPEF2, PRLR, ENSBTAG00000048048, ACAN, ADAMTSL3, TRPM1, FAM189A1, CYP11A1, DLK1, PLD4, ENSBTAG00000048268, ENSBTAG00000001213, ENSBTAG00000046076, ENSBTAG00000045666, MRPS24, URGCP, UBE2D4, DBNL, PGAM2, GLB1, MLH1, MITF, LRIG1, COL7A1, RAF1, PPARD, MOCS1, HSPA1A, ENSBTAG00000003892, PRL, ENSBTAG00000039099, NETO1, ENSBTAG00000044087, TPGS2, FHOD3, MOCOS, LAMA3, NPC1, SETBP1, KDSR, CLCN7, ATP2A1, GUSB, ENSBTAG00000035282, ENSBTAG00000046109, UBE2D1, PDE6C, PITX3, STN1, NHLRC2, FGFR2, XKR5, TAP, SPAG11B, ZNF705A, CENPU, F11, FGF20, IRF2BP2, B3GALNT2, LYST, NID1, TMEM26, HEPHL1, DEUP1, TYR, PYGM, COPA, GON4L, PKLR, GBA, TRIM33, RPAP2, SYDE2, RABGGTB, CRYZ, RNF11, MUTYH, UROD, CSMD2, ACKR3, COL6A3, MLPH, VSTM2A, DGKB, RF00026, TWIST1, RELN, PNPLA8, FOXP2, PPP1R3A, HOXA1, NUDCD3, CAMK2B, YKT6, GCK, MYL7, POU6F2, RF00405, VPS41, CADPS2, SLC13A1, ZNF800, RF00100, CLCN1, SLC4A2, SHH, LMBR1,

ALX1, CEP290, KITLG, EPYC, PLXNC1, CEP83, AMHR2, KRT5, KRT75, COL2A1, KIF21A, MSRB3, TAC3, SUOX, PMEL, MYBPC1, GNPTAB, TMTC1, OVCH1, ERGIC2, FAR2, SOX5, SLCO1B3, CD163, GNB3, CD4, VWF, IL17RA, SOX10, CYB5R3, BBS7, PDE5A, FABP2, C6H4orf3, USP53, RF00001, MYOZ2, SYNPO2, MANBA, UGDH, KIT, CSN2, ADAMTS3, SOWAHB, FGF5, DMP1, EVC2, HMX1, FAM193A, RNF4, ZFYVE28, HAUS3, POLN, PDE6B, IDUA, FGFR3, CPAMD8, ENSBTAG00000008335, OR10H1, ENSBTAG00000006216, RTBDN, ZNF266, LDLR, RANBP3, CAPS, VMAC, NDUFA11, F12, B4GALT7, CSNK1G2, GM2A, SLC36A1, GLRA1, CCNH, LIX1, EFNA5, SLC25A46, ANXA10, ENSBTAG00000035768, ENSBTAG00000025954, ENSBTAG00000008678, ENSBTAG00000039873, CTSB, FDFT1, NEIL2, IFNE, ENSBTAG00000025928, IFN-TAU, KLHL9, IFNAC, ENSBTAG00000039675, ENSBTAG00000046967, MPDZ, TYRP1, GRHPR, STX17, LPL, HR, ENSBTAG00000001219, GULO, AGTPBP1, PTPDC1, ROR2, RNF20, ABCA1, RF00001, ENSBTAG00000036273, PTP4A1, PHF3, SMAP1, LCA5, ENSBTAG00000023792, SEC63, ENSBTAG00000030490, SAMD3, TBXT, TENM1, ENSBTAG00000022399, ENSBTAG00000047994, ENSBTAG00000048126, bta-mir-363, bta-mir-18b, bta-mir-106a, CCDC160, PHF6, HPRT1, MIR450B, bta-mir-450a-1, bta-mir-450a-2, bta-mir-542, bta-mir-424, PLAC1, FAM122B, RF01233, RF01233, RF00026, MOSPD1, RTL8C, F9, ENSBTAG00000048281, NSDHL, CMC4, MTCP1, F8, RF00402, ENSBTAG00000033167, MPP1, DKC1, RF00417, RF00340, GAB3, ENSBTAG00000020185, HAUS7, BGN, CCNQ, DUSP9, PABPC5, ENSBTAG00000016019, ENSBTAG00000009457, ENSBTAG00000020390, TMEM35A, CENPI, DRP2, PLP1, PRPS1, RF00006, ENSBTAG00000048086, VSIG1, PSMD10, ATG4A, RF00026, COL4A5, RF00026, RF00012, IRS4, ENSBTAG00000046820, GUCY2F, NXT2, KCNE5, ACSL4, ENSBTAG00000046463, PLS3, RF00566, CYSLTR1, ENSBTAG00000035882, ENSBTAG00000026025, UPRT, ENSBTAG00000017241, PHKA1, HDAC8, RF00026, CITED1, ERCC6L, PIN4, NHSL2, EDA, bta-mir-532, bta-mir-502a-2, bta-mir-502b, bta-mir-500, bta-mir-660, CLCN5, AKAP4, MAGED4B, RF00614, ZC3H12B, ZC4H2, ENSBTAG00000045736, MTMR8, RF00026, ENSBTAG00000046004, bta-mir-222, bta-mir-221, CXHXorf36, KDM6A, DUSP21, PPP1R2C, MED14, CXHXorf38, MPC1L, ATP6AP2, ENSBTAG00000008248, ENSBTAG00000006383, ENSBTAG00000006380, ZFX, EIF2S3, KLHL15, CXHXorf58, APOO, SAT1, ACOT9, PRDX4, ENSBTAG00000036343, MBTPS2, GPM6B, OFD1, TCEANC, FRMPD4, bta-mir-584-8

Cuadro 2. Genes de Prueba para la Red Neuronal.

ENSBTAG00000012467, ENSBTAG00000010658, ENSBTAG00000014824, ENSBTAG00000006784, ENSBTAG00000005268, ENSBTAG00000005269, ENSBTAG00000025644, ENSBTAG00000010531, ENSBTAG00000001575, ENSBTAG00000022564, ENSBTAG00000037453, ENSBTAG00000001936, ENSBTAG00000019246, ENSBTAG00000016267, ENSBTAG00000023177, ENSBTAG00000012808, ENSBTAG00000013298, ENSBTAG00000005397, ENSBTAG00000000070, ENSBTAG00000038171, ENSBTAG00000003068, ENSBTAG00000013326, ENSBTAG00000013973, ENSBTAG00000003871, ENSBTAG00000012103, ENSBTAG00000003950, ENSBTAG00000004241, ENSBTAG00000009995, ENSBTAG00000014239, ENSBTAG00000010989, ENSBTAG00000014177, ENSBTAG00000018843, ENSBTAG00000007131, ENSBTAG00000020455, ENSBTAG00000008719, ENSBTAG00000020645, ENSBTAG00000020783, ENSBTAG00000046158, ENSBTAG00000023026, ENSBTAG00000004729, ENSBTAG00000016629, ENSBTAG00000005450, ENSBTAG00000037795, ENSBTAG00000023939, ENSBTAG00000010109, ENSBTAG00000007876, ENSBTAG00000012510, ENSBTAG00000046644, ENSBTAG00000017056, ENSBTAG00000011953, ENSBTAG00000019011, ENSBTAG00000001992, ENSBTAG00000002714, ENSBTAG00000004021, ENSBTAG00000002313, ENSBTAG00000004840, ENSBTAG00000019782, ENSBTAG00000026502, ENSBTAG00000034139, ENSBTAG00000016007, ENSBTAG00000000646, ENSBTAG00000005773, ENSBTAG00000048013, ENSBTAG00000039362, ENSBTAG00000037613, ENSBTAG00000039647, ENSBTAG00000015047, ENSBTAG00000004040, ENSBTAG00000011952, ENSBTAG00000006642, ENSBTAG00000009646, ENSBTAG00000048122, ENSBTAG00000008759, ENSBTAG00000010069, ENSBTAG00000009733, ENSBTAG00000025760, ENSBTAG00000015358, ENSBTAG00000014430, ENSBTAG00000001271, ENSBTAG00000020079, ENSBTAG00000004003, ENSBTAG00000015817, ENSBTAG00000009287, ENSBTAG00000020789, GART, RF00026, RBM11, FILIP1L, ITGB2, COLQ, HEXA, DPH6, MYO5A, DUOX2, KIAA0586, ASS1, PAEP, MED22, COL5A1, RNASEH2B, ATP7B, CLN5, UGGT2, MBNL2, ENSBTAG00000027444, KIF3B, HGH1, GPAA1, EXOSC4, WWP1, RF00001, DDB2, RF00403, RF00026, UBE3B, BCKDHA, PFAS, COL1A1, DPP10"ABCA12, IHH,ENSBTAG00000047333, SLC45A2, FAM189A1, CYP11A1, "ENSBTAG00000046076, SLC25A26, PPARD, MOCS1, PRL, FHOD3, SETBP1, ATP2A1, CCDC189, IPMK, GFRA1, B3GALNT2, LYST, RAB38, COPA, RABGGTB, DGKB, YKT6, GCK, MYL7, ZNF800, CEP290, KITLG, HOXC13, KRT71, GNS"TAC3, TMEM263, BBS7, FABP2, RF00001, MYOZ2, SYNPO2, ZFYVE28, HAUS3, MAN2B1, ZNF266, LDLR, ENSBTAG00000024041, CSNK1G2, EFNA5, ENSBTAG00000039873, CTSB, IFNE, ENSBTAG00000034294, VLDLR, ENSBTAG00000001219, PTPDC1, ENSBTAG00000036273, PTP4A1, bta-mir-106a, ENSBTAG00000033445, MPP1"RF00340, HAUS7, ENSBTAG00000033749, ENSBTAG00000016019, XKRX, CENPI, TSC22D3, PRPS1, RF00026, RF00026, ENSBTAG00000046463, PLS3ABCB7, ENSBTAG00000017241, PHKA1, RPS4X, NHSL2, EDA, bta-mir-660, MAGED4B, ZC3H12B, ZC4H2,

ENSBTAG00000046004, bta-mir-221, DUSP21, ATP6AP2, ENSBTAG00000008248, ENSBTAG00000006380, RAB9A.