IDENTIFICACIÓN DE POLIMORFISMOS EN REGIONES GENÓMICAS CARACTERIZADAS COMO HUELLAS DE SELECCIÓN EN EL GANADO OVINO

Esteban-Blanco[1], C., Gutiérrez-Gil, B., Suárez-Vega, A., López-Iglesias, L.J. y Arranz, J.J.
[1]Dpto. de Producción Animal, Facultad de Veterinaria, Universidad de León, 24071 León.
cristina.esteban.blanco@gmail.com

INTRODUCCIÓN

En el ganado ovino, la selección para los fenotipos como el color de la capa, la conformación, etc., se inició hace aproximadamente 5.000 años. Esta selección dio lugar a cambios más rápidos que los causados por la selección natural y ha dejado huellas detectables en del genoma de las razas ovinas modernas. La selección artificial para un determinado carácter de interés no sólo aumenta la frecuencia de la mutación causal del efecto, sino que además produce una alteración de las frecuencias de los alelos de otros loci neutros para el carácter objeto de selección, pero en desequilibrio de ligamiento con la mutación causal, dando lugar a característicos patrones de las frecuencias alélicas en la región afectada por la selección conocidos como huellas de selección. En los últimos años, se han llevado a cabo numerosos estudios de cribado del genoma basados en el análisis de chips de SNPs con el objetivo de detectar huellas de selección en las distintas especies de animales. El reciente desarrollo de las nuevas tecnologías de secuenciación permite analizar en profundidad las regiones previamente detectadas como huellas de selección con objeto de identificar las posibles variantes de ADN que pudieran ser responsables del fenotipo seleccionado. Un trabajo previo de nuestro grupo ha identificado varias regiones como candidatas a ser huellas de selección en el ganado ovino tras el análisis de los genotipos generados con del Chip ovino de media densidad (Chip-50K) dentro del proyecto *SheepHapMap* para dos grupos de razas de ovejas. El primer grupo incluyó tres razas merinas, altamente especializadas en la producción de lana fina (*Australian Industry Merino, Australian Merino* y *Australian Poll Merino*), y el segundo grupo incluyó tres razas de lana basta (*Churra, Altamurana* y *C*hios). En base a la disponibilidad de datos de secuenciación del genoma completo de una de las razas incluidas en cada uno de los dos grupos, Churra y Merina, nos hemos planteado como objetivo del presente trabajo el análisis de alta resolución de la variabilidad genética de las huellas de selección detectadas en relación al grupo "Merino" con el fin de identificar las mutaciones que, en dichas regiones, presentan frecuencias alélicas más extremas entre la raza Churra y Merina, y que pudieran ser evaluadas como candidatas a explicar el efecto detectado en esas regiones.

MATERIAL Y MÉTODOS

Huellas de selección a estudiar*:* La detección de las huellas de selección se ha descrito anteriormente (Gutierrez-Gil et al., 2016) y se basó en el solapamiento de las regiones identificadas como huellas de selección al aplicar a los genotipos agrupados en los dos grupos considerados, "Merino" y "No-Merino", dos tipos de análisis: (i) un análisis de diferenciación genética basado en el parámetro F_{ST} definido por Weir y Cockerham (1984) obtenido al contrastar los genotipos de los dos grupos considerados y (ii) un análisis de identificación de regiones de heterozigosidad reducida basada en la estimación de la heterozigosidad observada (ObsHtz) en cada uno de los grupos en estudio. Para los análisis de localización de las regiones se ha utilizado la versión Oar_v3.1 del genoma ovino como referencia (http://www.ensembl.org/Ovis_aries/Info/Index). Considerando los valores más extremos de F_{ST} resultantes del contraste de los dos grupos y los más extremos de ObsHtz reducida en cada uno de los grupos, en base a los criterios aplicados por Gutierrez-Gil et al. (2014), se identificaron cinco regiones candidatas asociadas al grupo "Merino" (CR-Merino) y seis regiones candidatas asociadas al grupo "No-Merino" (CR-NoMerino). Dado que una de las regiones era común a los dos grupos, el presente estudio se centró en el estudio de la variabilidad genética de las cuatro regiones exclusivamente asociadas al grupo "Merino" localizadas en los siguientes intervalos genómicos: OAR6: 36,75-37,83 Mb, OAR11: 26,32-29,20 Mb, OAR16: 38,88-40,32 Mb y OAR25:7,36-7,69 Mb.

Análisis bioinformático: Se utilizaron datos de secuenciación del genoma completo generados por el *International Sheep Genomics Consortium* disponibles en el repositorio *Sequence Read Archive (SRA)* para dos individuos Churra (CHU1_SRR501848, CHU2_SRR501909) y tres Merino (MERA1_SRR501887, MER454_SRR501852,

MERC1_SRR501868), además se analizaron las secuencias del genoma completo de dos ovejas Churras (0890N0001, 0890N0004) secuenciadas por nuestro grupo de investigación. Para las muestras obtenidas del repositorio SRA, se utilizó el software SRA-Toolkit (http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software) para convertir los datos al formato FASTQ. Todas las muestras se sometieron a continuación al siguiente protocolo para identificar variantes alélicas: (i) evaluación del control de calidad de las lecturas con FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), ii) alineación de las muestras con el genoma de referencia OAR_v3.1 con Burrows-Wheeler (BWA) (Li & Durbin, 2009), (iii) manipulación de datos, análisis estadísticos y generación de ficheros indexados con SAMtools (Li et al., 2009) y Picard (http://broadinstitute.github.io/picard/) (iv) identificación de variantes siguiendo el flujo de trabajo recomendado por el software GATK (Genome Analysis Toolkit; McKenna et al., 2010) que incluye realineamiento, recalibración y la búsqueda de variantes con la función *HaplotypeCalle*r. Filtrado de variantes con snpSIFT (Cingolani et al., 2012) utilizando las siguientes opciones: DP> 10 & QUAL> 30 & MQ> 40 & QD> 5 & FS <60. Identificación y anotación de variantes puntuales (SNPs) con frecuencias alélicas extremas: Se realizó un estudio de la variabilidad divergente en las regiones candidatas seleccionando aquellos SNPs que muestran frecuencias alélicas más extremas entre las muestras de secuenciación genómica de Churra y Merina. Para ello, con el software VCFTools (Danecek et al., 2011), se seleccionaron las variantes identificadas en las regiones objeto de estudio. Posteriormente con el programa PLINK (Purcell et al., 2007) se realizó un control de calidad (QC) de los genotipos brutos (*--mind 0.1 --geno 0.1*) y se realizó un análisis de asociación Churra *versus* Merino. En base a los resultados de ese análisis se estableció un umbral para identificar las variantes con frecuencias alélicas más extremas entre las de anotación a las dos razas contrastadas. Para todas las variantes identificadas como divergentes se realizó un análisis de anotación utilizando la herramienta *Ensembl Variante Effect Predictor* (VEP) (McLaren et al., 2010). Considerando las variantes intragénicas se obtuvo la lista de genes que albergan los SNPs divergentes entre las dos razas, realizándose posteriormente un análisis de enriquecimiento funcional con la herramienta *WebGestalt* (Wang et al., 2013). Se consideraron estadísticamente significativos los términos con un valor $P\text{-}value_{adj} < 0{,}01$.

RESULTADOS Y DISCUSIÓN

Después del primer filtro para seleccionar las variantes dentro de las cuatro regiones candidatas, se identificaron un total de 70.626 variantes (SNPs e indels), de las cuales 62.015 pasaron los parámetros de control de calidad aplicados con snpSIFT. Finalmente se consideraron un total de 53.829 marcadores de SNP bi-alélicos para los análisis posteriores. Tras el QC, y en base a los valores *P-value* nominales obtenidos a partir de la prueba del chi-cuadrado del análisis de asociación realizado, se identificaron un total de 260 SNPs que exhibían las frecuencias de alelos más extremas entre Churra y Merino (*P-value* <0,00316) (Figura 1). El análisis de anotación funcional mostró que 167 de los SNPs divergentes se localizaron en regiones intergénicas, mientras que 93 de ellos son intragénicos y están incluidos en la secuencia de 14 genes anotados (*ADAMTS12, DHRS7C, GSG1L2, MYH1, MYH10, MYH13, NDEL1 NTN1 PPM1K, RXFP3, STX8, TMEM107, TTC23L, USP43*), un gen no caracterizado (*ENSOARG00000011486*), un pequeño ARN nucleolar (snoRNA) (*SNORD118*) y un lincRNA. El análisis con VEP mostró que los 93 marcadores intragénicos determinaban un total de 105 variantes funcionales, que se clasificaron como una variante sinónima (en el gen *MYH1*), 77 variantes intrónicas, 15 variantes *upstream*, 8 *downstream* y 3 variantes de regiones de splicing y una variante en la región 3'UTR. El análisis de enriquecimiento funcional realizado para la lista de 14 genes con variantes intragénicas identificó ocho términos significativos, principalmente relacionados con la fisiología muscular y del citoesqueleto ("hydrolase activity", "calmodulin binding", "actin binding", "motor activity" y "cytoskeletal protein binding"). En relación a la posible relación de los genes destacados por nuestro análisis con caracteres de interés económico hay que resaltar que en cerdos se ha visto que el gen *USP43,* por su participación en la degradación de las miofibrillas durante la conversión de músculo a carne, podría relacionarse con caracteres de calidad de carne (Huynh, 2013). En vacuno, el gen *PPM1K* se ha asociado con caracteres de crecimiento y caracteres de conformación grasa de la canal (Lu et al., 2007). Según nuestra revisión bibliográfica, ninguno de los genes considerados parece estar relacionado con características del color y/o de la lana. Los resultados de nuestro trabajo muestran que, en

las cuatro regiones de huellas de selección de ganado Merino, los genes que contienen las variantes con frecuencias alélicas más extremas, comparándolas con la raza Churra, están relacionados con el crecimiento y caracteres de calidad de la carne. Esto concuerda con las conocidas diferencias en conformación y características de la carne de estas dos razas contrastadas. Futuros estudios debieran evaluar las posibles asociaciones de estos genes con caracteres de interés económico en ganado ovino.

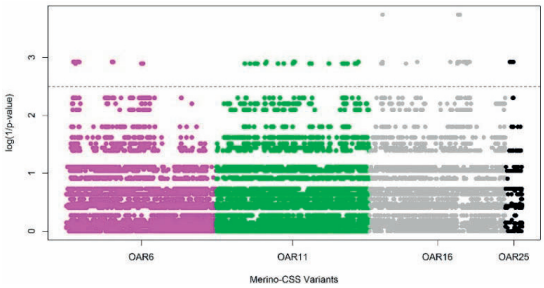

Figura 1. Identificación de variantes divergentes en cuatro regiones definidas previamente como huellas de selección en razas de Merino Australiano mediante el análisis de datos de secuenciación genómica de las razas Merino y Churra.

REFERENCIAS BIBLIOGRÁFICAS

• Cingolani, P. et al. 2012. Front. Genet. 3:35 • Danecek, P. et al. 2011. Bioinformatics 27: 2156-2158. • Gutierrez-Gil, B. et al. 2016. Proc. of the 67th EAAP meeting. Belfast. • Gutierrez-Gil, B. et al. 2014. PLoS One.9:e94623. • Huynh, T.P.L. 2013. Doctoral Thesis. Universität Bonn. • Kijas, J. et al. 2012. PLoS Biol.10: e1001258. • Li, H. & Durbin, R. 2009. Bioinformatics 25: 1754-1760. • Li, H. et al. 2009. Bioinformatics 25:2078-2079. • Lu, G. et al. 2007. Genes Dev. 21: 784-796.• Mckenna, A. et al. 2010. Genome Res. 20: 1297-1303. • Mclaren, W. et al. 2010. Bioinformatics 26: 2069-2070. • Purcell, S. et al. 2007. Hum. Genet. 81: 559-575. • Wang, J. et al. 2013. Nucleic Acids Res. 41 (W1): W77-W83. • Weir, B.S. & Cockerham, C.C. 1984. Evolution (N. Y) 38: 1358-1370.

Agradecimientos: Trabajo financiado por el proyecto AGL2015-66035-R del Ministerio de Economía y Competitividad España (MINECO). B. Gutiérrez-Gil es investigadora contratada del programa “Ramón y Cajal” del MINECO (RYC-2012-10230).

VARTIANT INDENTIFICATION IN GENOMIC REGIONS BETWEEN TWO GROUPS OF SHEEP DIVERGENTLY SELECTED FOR PRODUCTION TRAITS

ABSTRACT: The aim of this study was to use whole genome sequencing (WGS) to characterize the genetic variation of four candidate regions (CR) previously identified as selection signals (SS) associated with Merino breeds based on the comparison of the 50K-Chip genotypes for a group of three Merino sheep breeds highly specialized for fine wool production (Australian Industry Merino, Australian Merino and Australian Poll Merino) and three coarse wool breeds (Churra, Altamurana and Chios). Here, WGS datasets for Merino and Churra samples were analysed to identify SNP variants showing the most extreme allele frequencies between these two breeds in the four selected regions. From the total of variants identified in these four regions (70,626 SNPs e indels) a total of 53,829 SNPs were selected for later analyses. An association analysis was used to detect 260 SNPs showing the most divergent allele frequencies between the two breeds. Most of the genes harbouring the divergently selected intragenic SNPs within the four studied regions were related to muscle physiology. Future studies should assess the putative associations of the promising candidates identified herein with traits of economic interest in sheep.

Keywords: massive sequencing, whole genome sequence, sheep, divergent breeds.