UTILIZACIÓN DE LA SECUENCIACIÓN GENÓMICA PARA LA IDENTIFICACIÓN DE VARIANTES EN GENES DE LAS PROTEÍNAS DE LA LECHE EN EL GANADO OVINO

Marina, H., Gutiérrez-Gil, B., Esteban-Blanco, C., y Arranz[1], JJ.

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, España.

hmarg@unileon.es

INTRODUCCIÓN

La domesticación del ganado ovino se inició hace más de 10.000 años, produciendo cambios en el genoma ovino como consecuencia directa de la cría selectiva. Los primeros de estos cambios estuvieron relacionados con caracteres morfológicos como el color de la capa y la morfología de los cuernos (Kijas *et al.*, 2012), mientras que la selección para la producción de lana y leche fue posterior, hace aproximadamente 5.000 años (Chessa *et al.*, 2009). Actualmente, las razas ovinas utilizadas para la producción de leche se explotan principalmente en los países de la región mediterránea, destinándose su producción principalmente a la elaboración de quesos maduros de alta calidad (Selvaggi *et al.*, 2014). La fracción proteica de la leche está directamente relacionada con el rendimiento quesero y las propiedades tecnológicas de la leche. Las caseínas y las proteínas del lactosuero determinan el 5,5% de la composición total de la leche. Las caseínas, se clasifican en cuatro tipos: αs1-Cn (codificado por el gen *CSN1S1*), αs2-Cn (*CSN1S2*), β-Cn (*CSN2*), and κ-Cn (*CSN3*); y las proteínas del lactosuero son la alfa-lactoalbúmina (*LALBA*) y beta-lactoglobulina (*PAEP*) (Noce *et al.*, 2016). La gran especialización de la glándula mamaria determina que los genes de las caseínas constituyan un alto porcentaje (~ 70%) de los genes expresados en la glándula mamaria durante la lactación (Suárez-Vega *et al.*, 2015).

La identificación de variantes genéticas asociadas con caracteres de interés productivo en ganado ovino se inició con el estudio de genes candidatos, como los genes codificantes para las proteínas de la leche (Moioli, D'Andrea and Pilla, 2007). La disponibilidad, a partir de 2008, de un chip de SNPs de media densidad (50K-Chip) generó valiosa información sobre la variabilidad del genoma ovino a través del proyecto SheepHapMap (Kijas et al., 2012), y permitió estudios de mapeo de *Quantitative Traits Loci* (QTLs) y de huellas de selección de mayor precisión que los desarrollados hasta el momento. Uno de esos estudios, desarrollado por nuestro grupo de investigación en la raza Churra, identificó la primera mutación sugerida como causal de un QTL para un carácter de producción lechera en la oveja en el gen *LALBA* (García-Gámez et al. 2012). Más recientemente, el progresivo abaratamiento de la secuenciación completa del genoma (*Whole-genome sequencing*, WGSeq) ha proporcionado información de alta resolución en regiones previamente identificadas como portadoras de QTLs o huellas de selección (Rupp *et al.*, 2015; Gutiérrez-Gil *et al.*, 2017). El estudio de las variantes de los genes de las proteínas de la leche puede ser de gran interés para la identificación de marcadores genéticos a utilizar por la industria láctea, por su directa relación con las propiedades de coagulación de la leche y el rendimiento quesero. Así, partiendo de la gran cantidad de información generada mediante la secuenciación genómica, se ha planteado analizar, en el presente trabajo, datos de WGSeq para la identificación de variantes genéticas en seis genes codificantes de proteínas de la leche (los cuatros genes de las caseínas y los genes *PAEP* y *LALBA*) en seis razas ovinas españolas, incluyendo razas de diferente aptitud productiva.

MATERIAL Y MÉTODOS

Para el estudio de variabilidad genética se ha secuenciado el genoma completo (WGSeq) de un total de 21 individuos correspondientes a seis razas de ganado ovino doméstico presentes en la Península Ibérica: Assaf (n=5), Churra (n=6), Merino español (n=4), Castellana (n=2), Segureña (n=2) y Ojalada (n=2), siendo la primera de ellas la más especializada para la producción lechera, y la Churra de doble aptitud (lechera, cárnica). De estas muestras, 17 han sido secuenciadas por nuestro grupo de investigación y ANCHE (Asociación Nacional de Criadores de Ganado Ovino Selecto), y las cuatro restantes, pertenecientes a las razas Castellana y Ojalada, se han generado dentro de un proyecto coordinado por el Consorcio Internacional de Genómica Ovina (*International Sheep Genomics Consortium*, ISGC), como una extensión del Proyecto SheepHapMap (PRJNA160933), y se han obtenido a partir del

repositorio público *Sequence Read Archive* (SRA). Todas las secuencias han sido generadas con la tecnología paired-end de Illumina (Illumina HiSeq 2000 y Hiseq 2500). Las muestras obtenidas del repositorio público SRA (OJA4_SRR501900 y OJA5_SRR501911, CAS1_SRR501904 y CAS3_SRR501883) se transformaron de formato SRA a formato FASTQ con el software *SRA-Toolkit* (disponible en: http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software). Tras este paso todas las muestras se sometieron al siguiente protocolo de análisis: (i) evaluación de la calidad de las lecturas brutas con el programa FastQC (Andrews, 2010. disponible en: http://www.bioinformatics.babraham.ac.uk/projects/fastqc); (ii) filtrado de las lecturas de baja calidad mediante el programa *Trimmomatic* (Bolger, Lohse and Usadel, 2014); (iii) alineación de las muestras frente al genoma de referencia OAR_v3.1, utilizando el programa *Burrows-Wheeler Aligner* aplicando el algoritmo *mem* (Li and Durbin, 2009); (iv) manipulación de datos y generación de estadísticas preliminares realizada con el programa SAMtools (Li *et al.*, 2009), y eliminación de lecturas duplicadas con el programa Picard (disponible en: http://broadinstitute.github.io/picard/); (v) identificación de variantes de forma independiente con el programa *Genome Analysis Toolkit* (GATK; opción *haplotypecaller*) (McKenna *et al.*, 2010) y con el programa *Samtools* (opción *mpileup*) (Li *et al.*, 2009); (vi) filtrado de variantes identificadas con ambos software, con el programa *snpSIFT* (Cingolani *et al.*, 2012); (vii) selección de las variantes comunes identificadas por ambos programas con *BCFtools* (Li, 2011); (viii) extracción de variantes en genes candidatas, anotación y predicción del efecto de las variantes con los programas *snpEff* (Cingolani *et al.*, 2012) y *VEP* (McLaren *et al.*, 2010).

RESULTADOS Y DISCUSIÓN

Las lecturas generadas a través del proceso de secuenciación tuvieron una longitud máxima de 100 pares de bases (pb). El promedio de las lecturas brutas obtenidas por muestra fue de 150.535.606, de las cuales 87,52% de las lecturas pasaron el control de calidad realizado con *Trimmomatic*. Como promedio el número de las lecturas mapeadas frente al genoma de referencia fue de 264.144.270 por muestra. El total de variantes identificadas en todo el genoma de forma común por los programas *Samtools* y *GATK*, a partir del análisis de todas las muestras, fue de 31.558.572, mostrándose su distribución por Gigabase a lo largo de los 26 autosomas ovinos en la Figura 1. Tras realizar el filtrado por regiones candidatas, las variantes identificadas en los intervalos de los seis genes en estudio fueron un total de 547, incluyendo 531 SNPs, 13 delecciones y tres inserciones. La anotación de esas variantes mostró un total de 626 consecuencias funcionales (588 intrónicas, 18 exónicas, 12 en regiones de splicing y ocho en regiones 5'UTR). Los resultados del *snpEff* indican además una ratio de una variante cada 81 pb en todo el genoma. Esta ratio fue de una variante cada 623.975 pb en las regiones codificantes de los genes candidatos considerados.

De las variantes localizadas en exones de los genes candidatos, diez de ellas dieron lugar a un cambio de aminoácido en la secuencia proteica (*missense)* y ocho fueron sinónimas (Tabla 1). Según su impacto funcional sobre la correspondiente proteína, 10 de las mutaciones exónicas fueron clasificadas de impacto moderado y 8 de bajo impacto. Cuatro de las variantes *missense* se clasificaron como deletéreas y seis como toleradas, según la clasificación *SIFT* de *Ensembl*. Finalmente, en el gen *CSN3* no se ha encontró ninguna variante determinante de un cambio de aminoácido, apoyando las observaciones de estudios clásicos sobre un alto grado de conservación de la secuencia de este gen en la especia ovina.

REFERENCIAS BIBLIOGRÁFICAS

Bolger et al. 2014. Bioinformatics. 30(15), pp. 2114–2120. • Chessa et al. 2009. Science. 324(5926), pp. 532–6. • Cingolani et al. 2012. Frontiers in Genetics. 3, p. 35. • Gutiérrez-Gil et al. (2017). Genetics Selection Evolution. 49(1), p. 81. • Kijas et al. (2012). PLoS Biology. 10(2), p. e1001258. • Li. 2011. Bioinformatics, 27(21), pp. 2987–2993. • McKenna et al. 2010. Genome research. 20(9), pp. 1297–303. • McLaren et al. 2010. Bioinformatics. 26(16), pp. 2069–2070. • Moioli et al. 2007. Small Ruminant Research. 68(1–2), pp. 179–192. • Noce et al. 2016. Animal Genetics. 47(6), pp. 717–726. • Rupp et al. 2015. PLOS Genetics. 11(12), p. e1005629. • Selvaggi et al. 2014. Journal of the Science of Food and Agriculture. 94(15), pp. 3090–3099. • Suárez-Vega et al. 2015. Scientific reports. 5, p. 18399.

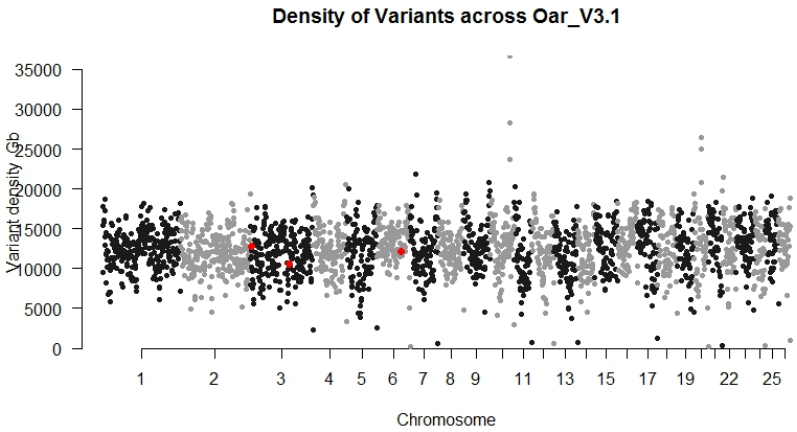

Figura 1. Densidad de variantes por Gigabase identificadas en las 21 muestras secuenciadas a lo largo de los 26 cromosomas autosómicos ovinos. En rojo se resaltan las regiones correspondientes a los seis genes candidatos considerados *(CSN1S, CSN1S2, CSN2, CSN3, PAEP, LALBA)*.

Gen	Posición	Cambio de aminoácido	Impacto funcional (precisión) (*SIFT*)
CSN1S1	c.626C>T	p.Thr209Ile	tolerada (0.06)
CSN1S2	c.268G>T	p.Asp90Tyr	deletérea (0)
	c.358A>G	p.Ile120Val	tolerada (0.71)
	c.527G>A	p.Arg176His	tolerada (0.38)
CSN2	c.595A>G	p.Met199Val	deletérea (0.15)
	c.634C>A	p.Leu212Ile	tolerada (0.2)
PAEP	c.106C>T	p.His36Tyr	tolerada (0.15)
	c.500A>T	p.Gln167Leu	deletérea (0)
	c.500C>T	p.Gln167Arg	deletérea (0)
LALBA	c.80T>C	p.Val27Ala	deletérea (0.02)

Tabla 1. Variantes *missense* identificadas en los genes candidatos considerados en estudio mediante el análisis de secuencias genómicas de 21 individuos pertenecientes a seis razas ovinas presentes en la Península Ibérica.

Agradecimientos: Esta investigación se ha financiado gracias al proyecto AGL2015-66035-R del MINECO, cofinanciado por el *European Regional Development Fund*. H.M. es beneficiario de una beca del Programa F.P.U. del Ministerio de Educación, Cultura y Deporte (Ref. FPU16/01161). B.G.G. disfruta de un contrato del programa Ramón y Cajal (Ref. RYC-2012-10230) del MINECO.

GENOME SEQUENCING FOR THE IDENTIFICATION OF VARIANTS IN GENES CODING FOR THE SHEEP MILK PROTEINS

ABSTRACT: The present study summarizes the identification of variants in candidate genes encoding sheep milk proteins based on the bioinformatic analysis of whole genome sequence datasets obtained from a total of 21 animals of six sheep breeds reared in Spain. Hence, considering the ovine genes encoding for caseins (*CSN1S, CSN1S2, CSN2, CSN3*) and for the milk´s whey proteins (*PAEP, LALBA*), a total of 547 variants were identified. Among them ten were missense variants classified with functional moderate impact. Variants in these selected genes could have important effects on the composition of milk and could explain differences on traits related to the cheese yield and the milk technological properties of the different breeds.

Keywords: whole genome sequence, Spanish sheep, gene, milk protein, livestock.