Universidad de León

Facultad de Veterinaria

Departamento de Producción Animal

Programa de doctorado Ciencias Veterinarias y de los Alimentos

APLICACIÓN DE LA SECUENCIACIÓN MASIVA PARALELA PARA EL ANÁLISIS DE ALTA DEFINICIÓN DE REGIONES GENÓMICAS DE INTERÉS EN EL GANADO OVINO Y PARA EL ESTUDIO DE LA MICROBIOTA DE LA LECHE DE OVEJA

Application of massively parallel sequencing to the high-resolution analysis of relevant genomic regions in dairy sheep and the study the sheep milk microbiota

Cristina Esteban Blanco

León, marzo de 2020

Las investigaciones de esta Memoria de Tesis Doctoral han sido financiadas por el proyecto Protmilkoma (AGL-2015-66035-R) del Ministerio de Economía y Competitividad.

La autora de esta Memoria ha sido beneficiaria de una beca de posgrado correspondiente al Programa de Formación de Personal Investigador (FPI) del Ministerio de Ciencia e Innovación con referencia BES-2016-07-8080.

“If you don't believe in yourself, no one will do it for you”

Kobe Bryant

AGRADECIMIENTOS

Llegados a este punto y después de todo el trabajo que ha supuesto esta Tesis doctoral, me gustaría dedicar unas palabras a toda la gente, que directa o indirectamente me ha ayudado en este camino a lo largo de estos años.

En primer lugar, me gustaría agradecer a Juanjo y a Bea, mis directores de tesis, la oportunidad que me brindaron al depositar en mí su confianza para desarrollar este proyecto. Gracias por vuestra dedicación, esfuerzo, e incalculable ayuda durante todos estos años. Gracias por enseñarme a pensar desde otro punto de vista y comenzar mi formación como investigadora.

A Javier, por contagiarme su entusiasmo y darme la oportunidad de aprender de él durante dos meses maravillosos en el CNB. Muchas gracias también a Fernando, un gran descubrimiento tanto personal como profesional. Sin su ayuda, sin ellos, mi estancia Madrid no hubiese sido tan productiva. Agradecer también el tiempo vivido a mis compañeros del Lab 35.0.

A la Universidad de León y al Ministerio de Ciencia e Innovación por darme los medios para la realización de este trabajo.

A todos los profesores del Departamento de Producción Animal, por su ayuda siempre que lo he necesitado. A todos los PAS que han pasado por el laboratorio, con una mención especial a Elena, que siempre ha estado dispuesta a arrimar el hombro.

A todos mis compañer@s de Departamento, por todos los momentos compartidos, por sus palabras de ánimo en los momentos más complicados. Pero sobre todo a aquellos con los que he convivido más tiempo… Gracias Praveen, siempre dispuesto a ayudar con lo que podía, callado pero atento. Gracias Aroa, me siento afortunada de haber compartido esta experiencia contigo, has sido sin duda un gran apoyo, ayudándome a saltar piedritas que tú ya habías saltado antes, enseñándome a crecer en este campo de la mejora genética animal, pero sobre todo gracias por siempre estar ahí. Gracias Héctor, sin duda este trabajo sin ti no hubiese salido igual de bien, gracias por ser mi compañero de viaje investigador del día a día, con días buenos y días malos, creo que finalmente hemos sabido entendernos (puñito). Gracias Ro, por tu siempre positivo carácter, por

sacarme sonrisas cuando yo no podía, por todos los momentos vividos, incluidas las cañas despejadoras, gracias por ayudarme en estos meses finales.

A SCAYLE, por proporcionarme los recursos técnicos de supercomputación para el desarrollo de esta tesis. Pero, gracias a FCSCL por darme la oportunidad de tener contacto con el grupo de investigación en el que se he desarrollado este trabajo. Gracias a Carlos, que confió en mi hace ya 7 años y me dio la oportunidad de aprender de todo un auténtico especialista en supercomputación, gracias Jesús, por transmitirme todo tu conocimiento y pasión por este campo, por prender una chispa en mí que me llevó por este camino, ojalá los nuestros vuelvan a cruzarse. Gracias también a mi maestro jedi, Fanego, por instruirme en las artes de torear los acontecimientos de la vida. Gracias a todos los que forman y han formado SCAYLE y FCSCL (Jose, Ruth, Mariví, Álvaro, María, Pablo…), cada uno ha aportado su granito en mi desarrollo profesional. Gracias Ele, por estar en todos los cafés y cañas posibles.

A mis amig@s, l@s de León, l@s de Galicia, l@s de la uni, l@s del cole, los que están y los que ya no, por todos los momentos compartidos. En especial a Menchu, que siempre siempre siempre está ahí, lejos pero cerca. A mi equipo de basket, que lleva tres años ayudándome a despejar en cada entrenamiento, haciéndome reír día sí y día también. A *eufrasiers* que desde hace dos años se han convertido en uno de mis pilares más importantes, sin ellas, sin su apoyo, sin sus risas, sin sus bromas, sin sus chevechas, sin sus fonitos momentos no llegaría hasta aquí con los ánimos en perfecto estado. Gracias Gelo por todos los momentos vividos que, aunque se dice pronto, han sido muchísimos. Gracias a los benquerencian@s por aguantarme y hacerme feliz en los momentos de descanso y libertad.

A mis padres, Paco y Loli, que me han aguantado durante todos estos años. Gracias por guiarme y apoyarme para conseguir mis objetivos. Gracias por vuestras palabras de cariño incondicional y vuestros consejos, porque gracias a vosotros soy la persona que soy ahora.

A mi hermana, Marta, por ser mi mejor amiga, mi confidente, por apoyarme, guiarme, enseñarme y abrirme camino durante toda mi vida. Gracias por aportar tanto en mi vida. Gracias también a Manolo, que se ha convertido en familia durante estos años.

Y por último gracias a Cal, por tu cariño y comprensión a lo largo de este tiempo. Gracias por traer una ilusión enorme en mi vida después de un momento de oscuridad. Por no dejarme sola en los momentos más difíciles y estar siempre para sacarme una sonrisa, dejando a un lado todo lo demás. Por valorar todo lo que hago y creer en mí.

MUCHISIMAS GRACIAS A TODOS

ÍNDICE DE CONTENIDOS

1. Abreviaturas

16S rRNA	Gen que codifica para la subunidad 16S del RNA ribosómico bacteriano
ANCHE	Asociación Nacional de Criadores de Ganado Ovino Selecto de Raza Churra
ARG	Gen de resistencia a antibióticos
ASSAF.E	Asociación Nacional de Criadores de Ganado Ovino de raza Assaf
ASV	Variante de secuencias de amplicones
BAC	Cromosoma artificial bacteriano
BMPR-IB	Receptor de proteína morfogenética ósea tipo 1B
CCR	Regiones candidatas de convergencia
CI	Intervalo de confianza
CNV	Variación en el número de copias
COG	Base de datos de grupos ortólogos de proteínas
DNA	Ácido desoxirribonucleico
EIF2S2	Subunidad 2 del factor de iniciación de la traducción eucariota
FAOSTAT	Base de datos estadísticos corporativos de la Organización de las Naciones Unidas para la Alimentación y la Agricultura
GAS	Selección Asistida por Genes
GEO	Base de datos de expresión génica
GS	Selección Genómica
GWAS	Estudios de asociación a nivel genómico
ISGC	Consorcio Internacional para la genómica ovina
KEGG	Enciclopedia de genes y genomas Kyoto
LA	Análisis de ligamiento
LAB	Bacterias del ácido láctico

LALBA	Alfa-lactoalbúmina
LCORL	Correpresor de receptor nuclear dependiente de ligando
LD	Desequilibrio de ligamiento
LDLA	Combinación de LD y LA
lncRNA	RNA largo no codificante
MAF	Frecuencia del alelo menos frecuente
MAS	Selección Asistida por Marcadores
Mb	Megabases, 1.000.000 de pares de bases
MEGA	Grupo de investigación de Mejora Genética Animal
MHC	Complejo Mayor de Histocompatibilidad
MSTN	Miostatina
NCAPG	Subunidad G del complejo de la condensina
Ne	Tamaño efectivo de la población
NGS	Secuenciación de segunda generación
NPR2	Receptor 2 del péptido natriurético
OAR	Cromosoma ovino
OCDE	Organización para la Cooperación y el Desarrollo Económicos
ONT	Tecnologías de secuenciación de Oxford Nanopore
OTU	Unidad taxonómica operacional
pb	Pares de bases
PCR	Reacción en cadena de la polimerasa
PRNP	Proteína priónica
QTL	Locus/Loci con influencia sobre un carácter cuantitativo

QTN	Mutación causal de un QTL
RFI	Ingesta de alimento residual
RNA	Ácido ribonucleico
RNA-Seq	Secuenciación de RNA
rRNA	Ácido ribonucleico ribosómico
RT-PCR	Reacción en cadena de la polimerasa con transcriptasa inversa
RXFP2	Receptor 2 de la relaxina
SCC	Recuento de células somáticas
SCS	Logaritmo del recuento de células somáticas
SMRT	Secuenciación a tiempo real de una única molécula de DNA
SNP	Polimorfismo de un solo nucleótido
snRNA	RNA pequeño nuclear
SOCS2	Supresor de la señalización de citoquinas 2
SRA	Base de datos de secuencias de NGS
SSN	Mutación causal de huellas de selección
SV	Variantes estructurales
ULE	Universidad de León
WGR	Resecuenciación del genoma completo

2. Planteamiento y objetivos

La presente Tesis Doctoral se plantea gracias al auge, en los últimos años, de las tecnologías de secuenciación de segunda generación (NGS, *Next Generation Sequencing*) y su creciente aplicación en el campo de la genómica animal. A principios del siglo XXI, después de la finalización del proyecto de secuenciación del genoma humano, el abaratamiento de la secuenciación de genomas completos, introducida por las técnicas de secuenciación masiva paralela o NGS, impulsó la aparición de proyectos de secuenciación de genomas en las especies domésticas. Aunque los resultados iniciales implican solo el conocimiento de una parte de la secuencia del genoma de la especie en estudio; "lo que se conoce como "borrador", estos proyectos de secuenciación favorecieron el desarrollo de herramientas moleculares de gran utilidad en genómica animal, como los chips de polimorfismos de un solo nucleótido o chips de SNPs, que permiten explorar el genoma en busca de regiones asociadas a caracteres de interés económico y utilizar la información molecular en la selección genómica. Además, la posibilidad de secuenciar todo el genoma a un coste asequible ha revolucionado el estudio de la base genética de los caracteres productivos, al permitir identificar variantes genéticas a lo largo de todo el genoma con elevada fiabilidad.

Las herramientas derivadas de las metodologías NGS han proporcionado un conocimiento más amplio de la arquitectura molecular de caracteres complejos, como el crecimiento o la producción de leche. Algunas de las mutaciones detectadas en regiones genómicas asociadas a caracteres de interés económico e identificadas con estas herramientas pueden incluirse en programas de mejora con el fin de incrementar la respuesta a la selección. Sin embargo, el gran volumen de datos obtenidos a partir de estas nuevas tecnologías de secuenciación dificulta el manejo rápido y eficaz de dicha información, por lo que cada vez es más necesario el desarrollo, utilización y optimización de herramientas bioinformáticas para el manejo e interpretación de la gran cantidad de datos con los que se trabaja.

La presente Tesis Doctoral se ha realizado en el grupo de investigación de Mejora Genética Animal, perteneciente al Departamento de Producción Animal de la Universidad de León, conocido como MEGA-ULE. La actividad de este grupo de investigación ha sido históricamente la mejora genética del ganado ovino de leche, principalmente en una raza autóctona de la región de Castilla y León, la raza Churra. Este

grupo colabora activamente con la Asociación de Criadores de Raza Churra (ANCHE) en cuestiones de asesoramiento genético. Sin embargo, debido al aumento exponencial en los últimos años del censo de la raza Assaf, una raza foránea altamente productiva, el grupo de investigación ha establecido lazos con el Consorcio de Promoción del Ovino, la mayor cooperativa del sector ovino a nivel europeo. Gracias a esta colaboración, para la realización de los estudios incluidos en esta Tesis, hemos tenido acceso al material animal y a los datos de explotaciones comerciales no solo de raza Churra sino también de raza Assaf.

Una de las líneas objeto de estudio del grupo ha sido la búsqueda de genes con influencia sobre caracteres cuantitativos (QTL, *Quantitive Trait Loci*) relacionados, principalmente, con caracteres de interés económico en el ganado ovino de leche. Dentro de esta línea, los primeros estudios, utilizando marcadores microsatélites y análisis de ligamiento (LA, *Linkage Analysis*), detectaron regiones genómicas con influencia sobre caracteres de producción de leche (Gutierrez-Gil et al., 2009; García-Gámez et al., 2012b), de morfología corporal (Gutiérrez-Gil et al., 2011) y con el recuento de células somáticas, este último relacionado con el estado de salud de la ubre (Gutiérrez-Gil et al., 2007). Los estudios para la detección de genes asociados con rasgos de interés económico, una vez disponible el mapa físico del genoma ovino de referencia, aprovecharon la mayor densidad de marcadores proporcionada focalizándose en la utilización de herramientas moleculares como el chip de SNPs de media densidad ovino para detectar nuevos QTL y refinar la posición de los previamente identificados. Con esta mayor densidad de marcadores, además de los análisis de LA, o de LA combinado con desequilibrio de ligamiento (LDLA), ha sido posible realizar análisis de asociación a nivel de genoma completo (GWAS, *Genome-wide Association Study*).

Los chips de SNPs, en muchos casos en combinación con la estrategia de análisis GWAS, han demostrado ser una herramienta útil para identificar mutaciones responsables de enfermedades genéticas con una herencia mendeliana simple (Becker et al., 2010; Suárez-Vega et al., 2013), pero no suficiente para la identificación de mutaciones causales en relación a rasgos complejos. De todos los rasgos estudiados en este grupo de investigación solo se ha detectado una mutación causal con efecto directo sobre el contenido de proteína en la leche, en el gen que codifica para la alfa-lactoalbúmina

(*LALBA*) (García-Gámez et al., 2012a). Otros grupos de investigación han combinado este tipo de análisis con la secuenciación de segunda generación para incrementar la precisión de mapeo. Un ejemplo de éxito de esa combinación ha sido la identificación de la mutación causal de un QTL ovino que influye sobre la resistencia/susceptibilidad a la mastitis, localizada en el gen *SOCS2* (Rupp et al., 2015a). La densidad de marcadores proporcionada por los chips de SNPs también ha servido para el mapeo de huellas de selección, definidas como patrones característicos de variación en secuencias de DNA atribuibles a la selección natural o artificial para uno o varios caracteres objeto de selección. En genómica animal se han detectado huellas de selección analizando poblaciones con fenotipos divergentes en el cerdo (Rubin et al., 2012), la vaca (revisado por Gutierrez-Gil et al., 2015) y la oveja (Gutiérrez-Gil et al., 2014). La aparición de las tecnologías de secuenciación a nivel de genoma completo ha permitido evaluar la variación genómica dentro de las poblaciones para explorar huellas de selección (Moon et al., 2015) e identificar mutaciones responsables de las mismas.

Además, tenemos que destacar que las NGS también se han empleado en otros campos como la caracterización y estudio de los microbiomas en diferentes ecosistemas. Los microbiomas consisten en nichos ambientales o biológicos que contienen comunidades complejas de microorganismos (Cho and Blaser, 2012). Antes de la aparición de las tecnologías NGS, la microbiota se estudiaba fundamentalmente mediante procedimientos de cultivos desarrollados en microbiología a lo largo de los siglos XIX y XX. A partir de la introducción de las NGS se ha podido observar una “revolución” en el conocimiento de la microbiota al poder identificar microorganismos que no crecían con las metodologías clásicas de cultivo, por sus características fisiológicas, y aquellos que por estar en muy escasa concentración no se identificaban mediante métodos clásicos (Streit and Schmitz, 2004). Los microorganismos que constituyen el microbioma se denominan “microbiota”, que puede variar sustancialmente según el nicho ambiental donde se encuentren y el estado de salud del huésped. Entre los “ecosistemas” donde se ha producido un salto cualitativo en la identificación del microbioma se incluyen diferentes órganos y tejidos de animales, como por ejemplo; el tejido intestinal (Jiao et al., 2016), vaginal (Giannattasio-Ferraz et al., 2019) y ruminal (Tapio et al., 2017; Singh et al., 2019).

Las NGS han revolucionado nuestra comprensión acerca del papel que desempeñan las comunidades microbianas en el estado de salud de los huéspedes (Proctor, 2011). En el ganado vacuno de leche muchos estudios se han centrado en la caracterización de la microbiota de la glándula mamaria (Bhatt et al., 2012; Kuehn et al., 2013; Oikonomou et al., 2014; Addis et al., 2016). La secuenciación y el análisis de regiones hipervariables del gen que codifica para la subunidad 16S del RNA ribosómico bacteriano (16S rRNA) proporcionan un método relativamente rápido y rentable para evaluar la diversidad y la composición de las comunidades bacterianas y, por tanto, ofrece una metodología apropiada para explorar la evolución de las comunidades microbianas en el desarrollo de una enfermedad, como la mastitis en el ganado de leche (Oikonomou et al., 2012). Dada la falta de estudios en el ganado ovino, esta Tesis Doctoral pretende utilizar este método derivado de las tecnologías NGS para caracterizar la microbiota de leche de oveja y analizar su posible asociación con el carácter de resistencia a mastitis.

En función de lo expuesto, el objetivo general de esta Tesis Doctoral ha sido utilizar las tecnologías de secuenciación masiva paralela en el estudio de caracteres y aspectos de importancia económica en el ganado ovino lechero. Este objetivo general se ha desarrollado a través de la consecución de los siguientes objetivos concretos:

1. Utilización de la resecuenciación del genoma completo (WGR, *Whole Genome Resequencing*) en tríos segregantes para realizar análisis de alta resolución para la identificación de posibles mutaciones causales en regiones genómicas previamente identificadas en el ganado ovino como (i) huellas de selección, o como (ii) QTL con influencia sobre la resistencia a la mastitis.
2. Utilización de la secuenciación masiva paralela para caracterizar la microbiota de la leche de oveja y su posible asociación con caracteres de resistencia a la mastitis.
3.

3. Introducción general

3.1. Producción de leche de oveja

La producción de leche de oveja, junto con la de otros pequeños rumiantes lecheros, como las cabras, representa aproximadamente el 3,5% del total de la leche mundial. Las ovejas lecheras se encuentran principalmente alrededor de las regiones del Mediterráneo y el Mar negro (países del sur de Europa, Europa Central y Próximo Oriente). Normalmente, la leche es producida por pequeños productores utilizando razas locales que están altamente adaptadas al medio aunque, dependiendo de la raza, los sistemas de explotación que se utilizan varían de extensivos a intensivos (Carta et al., 2009).

El sector ovino a nivel mundial dispone de un efectivo de ~1.200 millones de cabezas (FAOSTAT, 2017). Los productos derivados de ovejas lecheras se dividen históricamente en leche, carne y lana, siendo la carne el producto más demandado por los consumidores, seguido de la leche.

En ovino, la importancia económica de la producción lechera no reside en el consumo directo de la leche, sino que ésta generalmente es utilizada para la elaboración de otros productos lácteos tradicionales y de alta calidad que forman parte de la dieta mediterránea, como pueden ser el queso y el yogurt (Willett et al., 1995). Esto contribuye notablemente al desarrollo económico y social de aquellas regiones donde se concentra la producción lechera de ovino, pudiendo llegar a favorecer la fijación de la población rural (Castel et al., 2011).

La producción de leche a partir del ganado ovino se ha caracterizado en los últimos años por la necesidad de incrementar la competitividad de las explotaciones obteniendo productos de mayor calidad y la obligación de cumplir con requerimientos de seguridad alimentaria. Todo ello ha llevado a desarrollar e implementar programas de mejora en muchas razas ovinas lecheras. La mejora del sector ovino de leche, acompañada de un aumento de la automatización de procesos dentro de la pequeña industria lechera, ha conseguido en los sistemas intensivos un aumento de la producción (Devendra, 2001).

3.1.1. Producción de leche de oveja en España

España tiene una base histórica con respecto a la industria ovina lechera. Tradicionalmente se han explotado razas autóctonas que permiten acceder a pastos

naturales y residuos de cosecha, favoreciendo el desarrollo social y económico de algunas regiones.

En el año 2017, España produjo un total de 2.255.120 ovejas de ordeño, lo que resultó en una producción de 544 millones de litros de leche. La Figura 1 muestra la producción lechera ovina por Comunidades Autónomas en España en ese mismo año. Como puede apreciarse, hay dos zonas que concentran la mayor producción, Castilla y León que produce un 55% de la leche total, y Castilla-La Mancha que aporta un 31%, de la producción nacional.

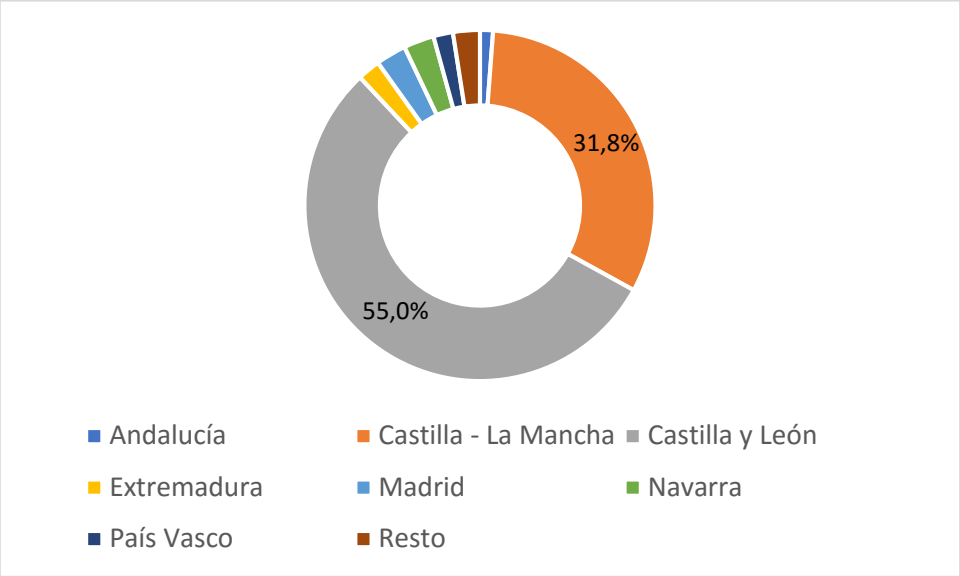


Figura 1. Distribución por Comunidades Autónomas de la producción de leche de oveja en España en el año 2017 (© Ministerio de Agricultura, Pesca y Alimentación)

España se posiciona como sexto país productor de leche de ovino a nivel mundial (http://faostat.fao.org/, 2017), y dado que, prácticamente toda la leche de oveja producida en nuestro país se utiliza en la elaboración de quesos que se consumen y elaboran en zonas muy limitadas (Martínez et al., 2011), España se convierte en el segundo productor de queso de oveja de la Unión Europea.

Tradicionalmente en España, la producción ovina ha estado ligada a la utilización y el aprovechamiento de razas autóctonas altamente adaptadas al clima de las regiones locales a las que pertenecen. Este es el caso de la raza Churra y la raza Manchega, asociadas respectivamente a las dos Comunidades más productoras, Castilla y León y Castilla-La Mancha (Ugarte et al., 2002). Sin embargo, en los últimos años, una gran parte

de la producción de leche de ovino procede de razas más productivas que los locales, razas foráneas como Assaf, Awassi y Lacaune que han encontrado un nicho en la industria láctea de nuestro país.

3.1.2. Importancia del sector ovino lechero en Castilla y León

Castilla y León ha sido y es una región rica en ganado ovino. Como se ha mencionado anteriormente, es la Comunidad Autónoma más productora de leche de oveja, aunque también tiene una alta producción de corderos lechales. Tradicionalmente se han utilizado las razas autóctonas, Churra y Castellana, en sistemas de explotación de pastoreo. No obstante, otras razas foráneas han desplazado a las anteriores debido a su alta aptitud lechera. Concretamente, la raza Assaf se ha convertido, junto con la raza Churra, en una de las dos razas más utilizadas en la comunidad de Castilla y León.

La raza Churra es una oveja rústica de las más primitivas de la península ibérica. Es una raza con elevado censo altamente productiva y, por lo tanto, una de las razas autóctonas española por antonomasia. La raza Assaf surge del cruce entre las razas Awassi (5/8) y Milschschaf (3/8). Se trata de una raza lechera especializada que se introdujo en España en el año 1977. Desde ese momento el número de animales se ha incrementado exponencialmente gracias a su capacidad de adaptación a las condiciones ambientales de Castilla y León y a sus características productivas. Animales con variable grado de mestizaje de la raza ovina Assaf conviven con rebaños de Churra en esta región (San Primitivo and De la Fuente, 2000).

El éxito y la alta producción de la empresa láctea ovina en Castilla y León en los últimos años reside en los cambios en los sistemas de producción que incluyen: mayor inversión en tecnología, mejoras en la alimentación, incremento del periodo de ordeño, mayor dimensión, control y dirección de la reproducción gracias a los programas de mejora.

3.1.3 Particularidades de la mejora genética del ganado ovino de leche

Los programas de mejora genética animal se basan en la identificación y selección de caracteres de interés económico que varían en función de la especie animal, de dónde y cuándo se apliquen dichos programas. La producción comercial de leche se utiliza principalmente para producir quesos y otros productos lácteos de alta calidad, representando más de dos tercios del ingreso total del sector ovino lechero (Miltiadou

et al., 2017). La producción de leche es el rasgo más importante de las ovejas lecheras, y su aumento se ha convertido en uno de los objetivos clave de la mejora genética. Los rasgos de producción de leche en ovejas son susceptibles de mejora con los programas tradicionales basados en datos genealógicos y fenotípicos, ya que son heredables con estimaciones de heredabilidad entre baja a moderada. Gracias a esto, ya en el último cuarto del siglo XX se establecieron programas de mejora en muchas poblaciones de ovejas (Barillet, 2007). Concretamente, el programa de cría de la raza Churra lo lleva a cabo la Asociación Nacional de Criadores de Ganado Ovino Selecto de Raza Churra (ANCHE), creada en el año 1973 con el objetivo principal de controlar la mejora genética para la producción de leche y corderos, además de controlar la genealogía de la población. Por otro lado, la Asociación Nacional de Criadores de Ganado Ovino de raza Assaf (ASSAF.E) se encarga de estas mismas cuestiones en la raza Assaf desde el año 2005.

El rendimiento de las producciones lecheras ovinas puede disminuir por varios factores, uno de los más importantes es la mastitis. La mastitis es una inflamación de la glándula mamaria que se caracteriza por cambios físicos, químicos y usualmente cambios bacteriológicos en la leche. La importancia de esta enfermedad principalmente se debe a las pérdidas económicas que representa en la cadena láctea debido a la reducción y descarte de la leche, el sacrificio temprano de animales y los costes en servicios veterinarios (Freitas et al., 2005; Davies et al., 2009). Normalmente las pérdidas económicas se estiman por la acción de la mastitis clínica. Sin embargo, la mastitis subclínica producida por la acción continua de microorganismos, en la mucosa, aunque sin signos clínicos o visibles de enfermedad, provoca pérdida progresiva del epitelio secretor durante uno o varios periodos de lactancia, lo que reduce la producción láctea y perjudica el crecimiento de la descendencia (Fthenakis and Jones, 1990; Saratsis et al., 1999; Sommerhäuser et al., 2003)

En el ganado ovino, la mastitis subclínica representa hasta el 95% de los casos de mastitis. Se estima que la prevalencia de esta enfermedad asciende hasta un 16-35% en Churra y 21-34% en Assaf (Las Heras et al., 1999; Gutiérrez-Gil et al., 2007). Se caracteriza por pasar desapercibida dificultando su detección debido a la ausencia de rasgos clínicos, pero conlleva cambios funcionales importantes y de comportamiento (Gougoulis et al.,

2008; Chiaradia et al., 2013). Dentro de los desafíos de la industria lechera de ovejas se encuentra el de ofrecer productos saludables a los consumidores y abordar el bienestar de los animales, lo que ha puesto de manifiesto la necesidad de incluir otros caracteres funcionales a los programas de mejora. Este es el caso del carácter recuento de células somáticas, que se define como un buen indicador de la resistencia frente a mastitis (Shook and Schutz, 1994).

3.2. Estrategias para la mejora genética en ganado ovino lechero

Clásicamente la mejora genética animal se ha basado en la identificación de animales con mayor valor genético seleccionándolos para ser los progenitores de la siguiente generación. Esta selección a "ciegas" ha incrementado los niveles productivos de muchas especies domésticas fijando, después de varias generaciones, los alelos "favorables" para el tipo de producción seleccionada. Sin embargo, el estudio de la base genética de los caracteres de interés económico en las especies domésticas ha evolucionado exponencialmente durante el siglo XX y hasta nuestros días.

Una pequeña parte de los caracteres de interés en especies de abasto están controlados por un único gen o por genes mayores (que explican una gran parte de la varianza), por ejemplo, el color de la capa, enfermedades mendelianas, etc. Sin embargo, la mayoría de los caracteres de interés económico y productivo en las especies domésticas como la oveja lechera son, en general, caracteres cuantitativos. Los caracteres cuantitativos muestran una distribución fenotípica continua y están influidos por múltiples genes con efecto pequeño y aditivo que se conocen como QTL, por el ambiente y por la interacción entre los factores genético y ambiental, lo que complica la tarea de identificación de las variantes de esos genes con efecto favorable para su utilización en mejora. A finales del siglo XX, el desarrollo de técnicas de mapeo genético permitió la identificación de regiones del genoma que contienen esos QTL, sobre todo aquellos que tienen un mayor efecto sobre el fenotipo, con el fin de poder seleccionar animales más productivos. A continuación, se realiza un breve repaso histórico de las herramientas utilizadas en el ganado ovino para la identificación de genes con influencia sobre caracteres de interés productivo.

3.2.1. Evolución de los estudios de la detección de genes con influencia sobre los caracteres productivos en la etapa pre-genómica

El fin último de las estrategias de detección de genes con influencia sobre caracteres de interés económico es la identificación de la mutación causal o QTN (*Quantitative Trait Nucleotide*), si se trata de una mutación responsable del efecto genético previamente identificado mediante el mapeo de QTL. Los primeros estudios de detección de genes con influencia sobre caracteres fenotípicos se basaron, principalmente, en una de estas dos estrategias: la estrategia del gen candidato y la del mapeo de genes. La primera consistía en el estudio detallado de un gen que, por su función fisiológica, se cree que puede tener un efecto sobre el carácter productivo de interés. Una vez seleccionado el gen de interés y tras la identificación de polimorfismos en su secuencia, generalmente en regiones codificantes, se hacía un estudio de asociación de esos polimorfismos con el carácter fenotípico. En las últimas décadas del siglo XX y principios del XXI, numerosos estudios utilizaron la estrategia del gen candidato en relación a los genes codificantes de las proteínas de la leche (Pirisi et al., 1999; Barillet et al., 2005; Moioli et al., 2007). Para superar la principal limitación de la estrategia del gen candidato, la necesidad de información previa sobre el gen objeto de estudio, se utilizó como estrategia alternativa el mapeo de QTL mediante estudios de ligamiento, cuyo objetivo era la identificación de regiones del genoma portadoras de genes responsables de la variación fenotípica observada. Estas dos estrategias se han utilizado tanto para el estudio de la arquitectura genética de caracteres simples, controlados por un único gen con un gran efecto, como para caracteres complejos, controlados por muchos genes de pequeño efecto. La identificación de los loci de interés se basa en la existencia de ligamiento entre los marcadores analizados y la mutación causal de un efecto dado. Los estudios de ligamiento se realizan sobre poblaciones con una estructura familiar específica, utilizando marcadores genéticos (polimorfismos de DNA) para los que se conoce su localización específica en el genoma.

En el ganado ovino, los primeros estudios de mapeo de QTL para caracteres de interés productivos se basaron en un tipo de marcadores genéticos denominados microsatélites, definidos por Jeffreys y colaboradores (1985) como repeticiones en tándem (*Variable Number Tandem Repeats*) de 1 a 10 pares de bases y que se distribuyen de forma más

o menos uniforme a lo largo del genoma de los mamíferos. Algunos ejemplos de la detección de genes mayores que controlan determinados caracteres a partir de los mapas de ligamiento de marcadores tipo microsatélites son el gen *MSTN* responsable del fenotipo "hipertrofia muscular" en bovino (Grobet et al., 1997) y en la raza ovina Texel (Clop et al., 2006), y el gen *BMPR-IB* responsable del fenotipo *Booroola* asociado a la alta prolificidad en la raza Merina (Mulsant et al., 2001).

Los mapas de ligamiento con marcadores microsatélites se utilizaron también para las primeras búsquedas de QTL con influencia sobre caracteres productivos. Los estudios que utilizan esta técnica abarcando todo el genoma, tomando como referencia los mapas de ligamiento publicados, como el de Maddox et al., (2001), se denominan barridos genómicos (*genome scans* en inglés). Concretamente, en la raza Churra se realizaron barridos genómicos utilizando microsatélites para la detección de QTL con efectos sobre caracteres lecheros y morfológicos utilizando un tipo concreto de diseño experimental, el diseño hija, adecuado a la estructura poblacional de la población objeto de estudio (Gutiérrez-Gil et al., 2007, 2008, 2009, 2011, Garcıa-Gámez et al., 2011). De hecho, uno de los puntos de partida importantes de esta tesis doctoral es la identificación de un QTL en el cromosoma 20, detectado a partir de un barrido genómico realizado con 181 marcadores microsatélites, con efectos sobre el carácter recuento de células somáticas, carácter indicador del estado sanitario de la ubre (Gutiérrez-Gil et al., 2007). Los resultados publicados a partir de los estudios de barridos genómicos desarrollados en especies domésticas para una amplio rango de caracteres de interés productivo se resumen en la base de datos AnimalQTLdb (https://www.animalgenome.org/cgi-bin/QTLdb/index), que un apartado específico para los QTL descritos en el ganado ovino, SheepQTLdb (https://www.animalgenome.org/cgi-bin/QTLdb/OA/index).

La finalidad de esos primeros estudios de identificación de QTL para caracteres de interés económico era la incorporación de información molecular a los programas de selección mediante la Selección Asistida por Marcadores (MAS, *Marker-Assisted Selection*), aproximación puesta en práctica en la primera década del siglo XXI en ganado vacuno Holstein de Francia (Boichard et al., 2006) y Alemania (Bennewitz et al., 2004). En la MAS se utiliza la información de marcadores, ligados a los QTL, que explican parte

de la varianza aditiva para el carácter de estudio además de información fenotípica y del pedigrí. Un tipo concreto de MAS es la Selección Asistida por Genes (GAS, *Gene-Assisted Selection*) que se basa en el uso de genotipos de uno o de un número reducido de genes para los que se conoce un efecto directo sobre el carácter a seleccionar. Un ejemplo práctico de GAS es la consideración del gen *PRNP* para el estudio de su genotipo en los programas de selección del ganado ovino en la Unión Europea (directiva CE2003/100/EC de la Comisión Europea 2003).

La detección de QTL utilizando marcadores microsatélites y mapas de ligamiento es una estrategia poco eficiente para la detección de mutaciones causales o QTN, ya que la baja densidad de marcadores utilizada tiene como consecuencia que la región genómica donde se acepta que se ha detectado un QTL es muy extensa y por lo tanto incluye muchos genes, dificultando la identificación del gen candidato que porta el QTN.

3.2.2. El genoma ovino y las herramientas de genotipado masivo

Una vez culminado y publicado el primer boceto del genoma humano (International Human Genome Sequencing Consortium, 2004), Schloss (2008) propuso un proyecto para reducir los costes totales de la secuenciación de un genoma completo dando paso a una nueva era de secuenciación, más barata y asequible para que los consorcios internacionales de otras especies pudieran también unirse a la vanguardia de la revolución genómica y tener su propio proyecto *Hapmap*. Las primeras especies ganaderas con proyecto de secuenciación genómica propio fueron la gallina y la vaca (Hillier et al., 2004; Elsik et al., 2009), especies domésticas cuyas repercusiones económicas eran tan altas como para justificar la inversión económica que debía hacerse para disponer de un borrador secuenciado del genoma. No fue hasta el año 2002 cuando se constituyó el ISGC (*International Sheep Genomics Consortium*), formado por científicos y agencias de financiación de muchos países con el objetivo de desarrollar recursos genómicos de ámbito público como la secuenciación completa del genoma ovino, en un proyecto llamado *SheepHapMap* (www.sheephapmap.org). El primer intento de obtener una secuencia del genoma ovino dio lugar a un genoma virtual. Para ello se utilizó la técnica de Sanger para secuenciar extremos de una genoteca de plásmidos BAC (*Bacterial Artificial Chromosome*) del genoma ovino de un macho de raza Texel, seguido de un mapeo frente a los genomas disponibles de especies

filogenéticamente cercanas a la oveja como el genoma bovino, canino y humano. Este alineamiento permitió la ordenación y creación de secuencias de mayor longitud por comparación de especies para su posterior reorganización utilizando la información posicional conocida de los marcadores en el mapa de ligamiento disponible en aquel momento (versión 4.6 del mapa australiano). Se obtuvo así, el conocido *Virtual Sheep Genome* (Dalrymple et al., 2007) que cubría el 76% del genoma de esta especie.

La rápida evolución de las técnicas de secuenciación a nivel genómico que, como se ha explicado anteriormente, produjo una disminución en los costes y tiempos de secuenciación, permitió que el ISGC publicara la secuencia de referencia de un genoma de oveja real en el año 2010, dejando a un lado el genoma virtual. A partir de este momento, ha habido varias versiones de la secuencia de referencia, y actualmente el proyecto del genoma ovino continúa en desarrollo. Los trabajos incluidos en esta Tesis Doctoral utilizan la versión del genoma de oveja Oar_v3.1, publicada en 2014 (Jiang et al., 2014). Se trata de la última versión del genoma ovino incluida en el repositorio Ensembl (https://www.ensembl.org/), en el que están anotados un total de 20.921 genes codificantes para 29,118 transcritos y más de 60 millones de variantes cortas (SNPs e indels) (consultado a 1 de febrero de 2020).

La información derivada de los proyectos de secuenciación y los proyectos *HapMap* (incluído el proyecto *SheepHapMap*) permitió la identificación, a lo largo de todo el genoma de la especie correspondiente, de miles de polimorfismos puntuales (polimorfismos de nucleótido simple, SNP; *Single Nucleotide Polymorphism*) (Nicholas and Hobbs, 2014). Generalmente los SNPs son bialélicos, característica que les permite ser utilizados para el genotipado a gran escala. Este hecho, unido a que aparecen uniformemente a lo largo del genoma, hace que hayan sido tan útiles como para desarrollar plataformas de genotipado de alto rendimiento que permiten el análisis de miles de SNPs a un coste muy razonable, llamados chips de SNPs. El primer chip de SNPs comercializado en el ganado ovino fue el *OvineSNP50 BeadChip*. Este chip comercial incluía más de 54.000 SNPs que se distribuyen uniformemente todo el genoma ovino. La selección de estos SNPs incluyó varios criterios, dentro de los cuales destacan: un filtro por MAF (*Minor Allele Frequency*), el recuento de alelos, las puntuaciones de calidad, la distancia entre SNPs y la ubicación en los cromosomas. Se consideró que estos

requisitos aportaban una excelente densidad de SNPs, ya que había una distancia promedio entre SNPs de 50,9 kb. Años más tarde empezaron a comercializarse chips de alta densidad, que incluían entre 600.000 y 800.000 SNPs. Estas plataformas de genotipado masivo están desarrolladas y optimizadas con las tecnologías de *Illumina Infinium* y *Affymetrix Axiom*, entre otras. Además, el desarrollo de esta tecnología y la evolución de las investigaciones en esta área han permitido el diseño desarrollo de chips de genotipado personalizados (*custom-made SNP chip).* La disponibilidad de chips de SNPs de media y alta densidad ha permitido la puesta en práctica de una variante de la MAS, denominada Selección Genómica (GS, *Genomic Selection*) (Meuwissen et al., 2001). Este método utiliza miles de marcadores distribuidos uniformemente por todo el genoma, asumiendo que todos los QTL que determinan un carácter estarán en desequilibrio de ligamiento con uno o varios marcadores utilizados y, por lo tanto, la suma de los efectos de los marcadores se utiliza para predecir el valor genético de los animales.

3.2.3. Estudios de los caracteres de interés productivo en la era post-genómica

El número de SNPs incluidos en un chip varía en función de la especie en estudio, aunque generalmente se pueden diferenciar chips de media y alta densidad (50-60K y 600-800K respectivamente). Por ejemplo, los chips comerciales de los que se dispone para vaca (50K, 800K) y oveja (50K, 700K) son de media y baja densidad, mientras que en la especie humana los *arrays* disponibles llegan a incluir hasta nueve millones de marcadores. Utilizando barridos genómicos con estas nuevas tecnologías la densidad de marcadores para el mapeo de genes de interés se incrementa de forma muy significativa y, por lo tanto, se reduce el tamaño de las regiones identificadas como QTL, aportando mayor precisión al mapeo (Mateescu, 2020). El análisis de estos datos puede estar basado, como en el caso de los microsatélites, en el LA pero también es posible combinar la información intrafamiliar explotada en los análisis LA con información de desequilibrio de ligamiento (LD, *Linkage Disequilibrium*), mediante lo que se conoce como análisis combinado LDLA.

Como ventaja derivada de la alta densidad de marcadores proporcionada por los chips de SNPs, es posible aproximar los genotipos de miles de marcadores SNPs con el fenotipo objeto de estudio mediante lo que se denomina estudio de asociación a nivel

genómico (GWAS), en el que no se requieren estructuras familiares concretas como las necesarias para los análisis tipo LA. Este tipo de análisis se ha utilizado para identificar QTNs en caracteres tanto mendelianos como cuantitativos. La microftalmia en la raza Texel (Becker et al., 2010) y la lisencenfalia con hipopasia cerebelar (Suárez-Vega et al., 2013) son ejemplos de estudios que han identificado la mutación causal de caracteres monogénicos en el ganado ovino utilizando estudios de GWAS. Por otro lado, el primer estudio de GWAS para caracteres cuantitativos de producción de leche de oveja fue realizado por Garcia-Gamez y colaboradores (2012). Se utilizó el chip ovino de 50K de Illumina para genotipar una población comercial de raza Churra con una estructura familiar de diseño hija. La asociación más significativa de este trabajo fue la identificada, en el cromosoma 3, entre un SNP localizado en el tercer intrón del gen de la alfa-lactoalbúmina (*LALBA*) y los caracteres de contenido proteico y graso de la leche. En base a análisis posteriores, este estudio sugirió una mutación tipo SNP del gen *LALBA*, determinante de un cambio aminoacídico como la posible mutación causal (QTN) del QTL inicialmente mapeado. Un estudio posterior basado en análisis LA y LDLA de los mismos caracteres y población, demostró, para este tipo de poblaciones con estructura familiar, las ventajas de explotar al mismo tiempo la información familiar y poblacional a través del análisis LDLA (García-Gámez et al., 2013). El método LDLA ha sido también el elegido para la identificación de QTL con influencia sobre caracteres de producción lechera en la oveja Sada (Usai et al., 2019).

Por otro lado, los chips de SNPs se han utilizado también de forma eficiente para la detección de huellas de selección, definidas como patrones característicos de variación en el genoma de las especies domésticas derivados de la selección artificial. A este respecto hay que tener en cuenta que el proceso de domesticación, iniciado hace unos 10.000 años (Hyams, 1972), seguido de los posteriores procesos de migración y selección de las distintas poblaciones (Fariello et al., 2014; Colli et al., 2018) explican la gran variabilidad genética ofrecida actualmente por las poblaciones de las diferentes especies domésticas. Todo ello, asociado con los avances en la cría de animales, ha permitido el desarrollo de razas locales, en algunos casos altamente especializadas para una cierta producción. Así, los análisis de detección de huellas de selección se basan en la detección de la alteración de las frecuencias alélicas provocados por el proceso de la

selección artificial ejercida por el hombre sobre la mutación causal del efecto fenotípico buscado, así como en la detección de un efecto de arrastre en las frecuencias alélicas de marcadores genéticos cercanos a dicha mutación causal (Kaplan et al., 1989). Por ejemplo, un indicador importante de la presencia de huellas de selección es la reducción local de la variabilidad genética en los genes causales afectados directamente por la selección y en las variantes SNP cercanas (Smith and Haigh, 2007). De esta manera, la densidad de mapeo aportada por los chips de SNPs ha permitido también identificar distintos patrones de variabilidad genética asociados a las huellas de selección (ej. reducción de la homocigosis, diferenciación genética, regiones de homocigosidad extendida, etc) en las distintas especies domésticas; vaca (Gutierrez-Gil et al., 2015; Naderi et al., 2020), cerdo (Rubin et al., 2012; Qin et al., 2020), cabra (Talenti et al., 2017; Alberto et al., 2018) y oveja (Kijas et al., 2012a; Gutiérrez-Gil et al., 2014; Gutierrez-Gil et al., 2015; Rochus et al., 2018).

3.3. La secuenciación de segunda generación

La aparición de la secuenciación masiva paralela, conocida también como secuenciación de segunda generación (NGS), ha supuesto un gran hito científico e histórico en el estudio del genoma. La NGS es considerada la segunda generación en lo que respecta a la secuenciación del DNA, considerando *Sanger* la primera generación (Dorado et al., 2019). La tecnología *Sanger* tiene una elevada precisión en la determinación de las secuencias de DNA, pero su rendimiento y escalabilidad son bajos. La secuenciación de segunda generación ha superado las limitaciones del método enzimático de Sanger al utilizar superficies de fijación de moléculas de DNA, que permiten la secuenciación en paralelo de millones de secuencias de DNA (Slatko et al., 2018). Esta paralelización de la secuenciación, junto con la miniaturización desarrollada por estas nuevas tecnologías ha permitido disminuir, por un lado, el tiempo necesario para secuenciar genomas enteros y, por otro, los costes de dicha secuenciación (Pettersson et al., 2009). Como se ha mencionado anteriormente, este abaratamiento de los costes de secuenciación ha permitido el estudio de la base genética a nivel de secuencia de caracteres complejos en especies domésticas minoritarias, como la oveja, gracias a la disponibilidad de un borrador de su genoma de referencia.

En las dos últimas décadas se han desarrollado diferentes tecnologías de secuenciación masiva caracterizadas principalmente por la longitud de los fragmentos secuenciados, que oscilan entre ~30 y 500 pb; Roche-454 (pirosecuenciación), Ion Torrent (secuenciación por semiconductores), Illumina/Solexa (secuenciación con terminadores reversibles) y SOLiD (secuenciación por ligamiento) (Mardis, 2017). Sin embargo, es la tecnología Illumina la que lidera en la actualidad el mercado de la secuenciación masiva paralela (van Dijk et al., 2014; Meera Krishna et al., 2019) y de tercera generación, como veremos más adelante en el apartado 3.4 de esta Tesis Doctoral. Este liderazgo se debe a la versatilidad de los instrumentos de Illumina que los hace ideales para una variedad de aplicaciones de secuenciación, entre las que se incluyen el mapeo frente a un genoma de referencia (resecuenciación), la secuenciación del transcriptoma, la detección de SNPs y la secuenciación utilizada para estudios metagenómicos y de meta-taxonomía. Así, instrumentos HiSeq son muy adecuados para analizar genomas grandes de animales y plantas, ideales el caso del presente trabajo para la secuenciación de genomas completos de ovejas (2.500 Mbp, https://www.ensembl.org/Ovis_aries/) y el estudio de transcriptomas de interés. El Illumina MiSeq produce menos lecturas, pero las longitudes de las mismas son significativamente más largas, lo que lo hace ideal para la secuenciación de genomas pequeños, como los bacterianos. El NovaSeq es el último instrumento de alto rendimiento de Illumina, muy parecido en prestaciones al HiSeq, pero diseñado para laboratorios que no se pueden permitir los costes de las secuenciaciones del HiSeq (Besser et al., 2018).

Los archivos de salida de los secuenciadores de segunda generación son archivos de texto de varios gigabytes de tamaño que contienen miles de millones de lecturas representadas como secuencias cortas de letras correspondientes a cada fragmento de DNA que ha sido analizado por el secuenciador. No solo se obtiene la secuencia concreta escrita en pares de bases de una lectura, sino que, asociada a ella, se tiene información de calidad para cada base, describiendo la probabilidad de error del secuenciador en cada posición correspondiente. Este tipo de datos sin procesar tiene un formato denominado FastQ (https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/), cuyo procesamiento requiere de una reducción masiva de sus dimensiones. Esta labor es llevada a cabo a través de la Bioinformática, una rama de la ciencia que utiliza métodos

computacionales para estudiar datos biológicos, como la estructura, función y evolución de genes, proteínas y genomas completos (Higgs and Attwood, 2013). El surgimiento de esta disciplina se remonta al año 1996 según varios autores (Robbins, 1996; Ouzounis and Valencia, 2003), momento en el que se pone de manifiesto que la Bioinformática tiene el potencial necesario para transformar la investigación biomédica (Altman, 1998; Kafatos, 1998). Muchas de las tareas llevadas a cabo por la Bioinformática son tareas simples que se podrían realizar fácilmente si los datos generados fuesen más pequeños, pero la dificultad de este análisis surge con la ingente cantidad de datos que deben ser procesados y la necesidad de repetir las mismas operaciones miles de millones de veces.

Las herramientas bioinformáticas son muy variadas en función del tipo de datos a analizar o el tipo de estudio a realizar; además, la disponibilidad de recursos informáticos también puede ser un limitante en la elección de estas herramientas. Se han desarrollado algoritmos bioinformáticos para la manipulación de secuencias cortas (Hatem et al., 2013), el ensamblado de novo (Sohn and Nam, 2016), la detección de variantes (McKenna et al., 2010), el alineamiento de datos de transcriptómica (Dobin et al., 2013).

Un hecho claro es que los avances de las tecnologías de secuenciación en los últimos años van de la mano del desarrollo de nuevas herramientas bioinformáticas (Chen et al., 2009; Ji, 2012; Akhtar et al., 2015). Este desarrollo combinado con técnicas computacionales ha favorecido la utilización de la NGS y la obtención de resultados en tiempos relativamente cortos de manera eficiente y eficaz, eliminando los problemas del esfuerzo humano que conllevaría realizar este tipo de estudios sin recurrir a esta nueva parte de la ciencia, y también haciendo posible que especialistas en ciencias de la vida integren estos tipos de análisis como parte de su rutina sin tener que depender de otros perfiles profesionales (ej. matemáticos, informáticos, etc). Sin embargo, en muchos casos, cuando los problemas bioinformáticos no son capaces de abarcarse con los equipos informáticos tradicionales, como ordenadores de sobremesa o incluso pequeños servidores de un laboratorio, se puede acudir al uso de la supercomputación como herramienta complementaria al análisis bioinformático (Díaz et al., 2011; Vega-Rodríguez and Santander-Jiménez, 2019). Tanto es así, que la bioinformática ha logrado adaptar las necesidades de cómputo al paralelismo de los centros de supercomputación,

de manera que es posible optimizar flujos de trabajo que consiguen utilizar, no solo todos los procesadores que ofrece un único servidor, si no también escalar proyectos de gran envergadura que consistan en el análisis de miles de muestras de secuenciación. De esta manera, se ha conseguido abordar el cúmulo de datos generados por estas nuevas tecnologías de secuenciación y extraer conclusiones a preguntas, que sin la bioinformática sería imposible responder. Cabe destacar, que las técnicas de secuenciación están en continuo desarrollo y producción de datos, por lo que conllevan necesariamente el reto de diseñar procesos que permitan que esta gran diversidad de datos sea útil a la comunidad científica. Debido a la importancia de la bioinformática en el desarrollo de esta Tesis Doctoral, se expone a continuación de forma general, los análisis llevados a cabo utilizando procedimientos bioinformáticos utilizados.

3.3.1. Análisis de datos de secuenciación de genomas completos

La resecuenciación del genoma completo (WGR) utilizando secuenciación masiva paralela tiene como objetivo determinar la secuencia completa del DNA del genoma de un organismo. Una de las aplicaciones más utilizadas de los estudios WGR es la identificación de variantes, ya sean de tipo SNP, inserciones y deleciones (*indels*), variantes estructurales (SV, *Structural Variants*) o variaciones en el número de copia (CNVs, *Copy Number Variations*). Para este tipo de análisis se requiere la secuencia de un genoma de referencia de alta calidad contra el que alinear las secuencias genómicas de los individuos de estudio para posteriormente detectar las variaciones de secuencia entre ese genoma de referencia y las muestras secuenciadas.

Uno de los puntos limitantes en este tipo de análisis es la profundidad de secuenciación, que tiene gran impacto no solo en los costes de la secuenciación sino también en los resultados biológicos del procesamiento de los datos, como la proporción de variantes raras y polimorfismos detectados (Rashkin et al., 2017), así como la precisión en el proceso de asignación alélica de los distintos marcadores (Ajay et al., 2011). Cuanta mayor profundidad de secuenciación, mayor confianza de que la variante reportada por el análisis bioinformático sea real y no un error de la secuenciación, así como de la fiabilidad del genotipo asignado para esa muestra. Además, la WGR proporciona información imparcial, de la secuencia completa del genoma, superando los sesgos de identificación de SNPs, con una potencia del ~99% para detectar variantes con una

frecuencia de población superior al 1% para la mayor parte del genoma (Abecasis et al., 2012). El flujo de análisis para la identificación, clasificación y anotación de variantes a partir de la secuencia completa de un genoma es complejo y requiere, no solo de la utilización de herramientas bioinformáticas estables, sino también de la implicación del investigador en la utilización de lenguajes de programación adicionales que sirvan para el procesamiento post-análisis de los datos (Goodwin et al., 2016; Pfeifer, 2017).

El potencial de los datos de WGR para aportar información sobre todas las variantes a lo largo del genoma ha revolucionado las investigaciones en genómica animal, al identificar en las muestras analizadas millones de polimorfismos a lo largo de todo el genoma (van El et al., 2013)(Boussaha et al., 2016). Esta información puede ser utilizada para identificar mutaciones causales de efectos fenotípicos en las especies domésticas. Por ejemplo, Hoff y colaboradores (2017), mediante el análisis de datos WGR de 109 animales de ganado Angus, reportaron una serie de mutaciones causales y haplotipos letales que afectaban directamente a la aptitud reproductiva de los animales, confirmando la utilidad de estos métodos para identificar variantes causales cuya información se podrían añadir a los métodos y programas de mejora existentes con el objetivo de hacerlos más eficiente. En el ganado ovino, por ejemplo, se ha utilizado recientemente la secuenciación masiva de datos de un trío segregante, combinada con datos de genotipado de media densidad, para identificar la mutación causal de un QTL del cromosoma 3 ovino con influencia sobre la susceptibilidad a la mastitis en el gen *SOCS2*, gen que codifica para la proteína supresora de la vía de señalización por citoquinas-2. La mutación causal identificada en ese gen, determinante de un cambio aminoacídico en la proteína (p.R96C), provoca la pérdida de la actividad funcional de la proteína *SOCS2* (Rupp et al., 2015a), lo que parece tener efectos antagónicos sobre la resistencia a mastitis y sobre el carácter crecimiento. En este caso, al estar el estudio basado en una población comercial con una estructura de diseño nieta, el trío segregante estaba formato por el abuelo Qq, y dos de sus hijos seleccionados en base a sus fases haplotípicas en la región del QTL en concordancia con valores genéticos extremos para el carácter, hijo QQ e hijo qq. Los datos de WGR también han sido en los últimos años utilizados. Comentar también la posibilidad de realizar análisis GWAS con

datos imputados a nivel de WGR, aproximación aplicada exitosamente en cerdos para el carácter número de vértebras lumbares (Yan et al., 2017).

En la última década, el coste asequible de las tecnologías de secuenciación también ha permitido el uso de la información de datos de WGR para incrementar la resolución del mapeo de huellas de selección a un coste relativamente bajo (Schlotterer et al., 2014), particularmente para organismos no modelo. Rubin y colaboradores (2010) presentaron un estudio pionero utilizando la técnica de resecuenciación del genoma completo de pools de muestras de DNA de diferentes líneas de pollos para identificar huellas de selección asociadas al proceso de domesticación en esta especie. Otro estudio similar en el cerdo, basado en la secuenciación de 55 genomas de 3 razas, también reveló numerosas huellas de selección en 24 poblaciones de cerdos de todo el mundo (Rubin et al., 2012). En ganado vacuno el *pooling* de muestras de DNA y la WGR han sido también utilizado para la identificación de variantes bajo presión de selección en el Zebú (Gir cattle) (Liao et al., 2013).

3.3.2. Análisis del transcriptoma

El microbioma se define como todos los microorganismos presentes en un ambiente particular. Desde que Robert Koch desarrolló la metodología para obtener cultivos bacterianos puros sobre medio sólido en 1881, la microbiología fue completamente dependiente de los cultivos a lo largo de un siglo. Este enfoque limitó el rango de organismos detectables favoreciendo a organismos aerobios de crecimiento fácil como *Escherichia* spp. Se estima que solo un 1% de los microorganismos observables en la naturaleza pueden cultivarse por técnicas estándar (Streit and Schmitz, 2004). A partir de la década de 1980 el estudio del mundo bacteriano, que antes estaba confinado a una pequeña minoría de especies que podían ser cultivadas en un laboratorio, se vio incrementado repentinamente con la utilización de métodos moleculares, que permiten detectar todas las bacterias presentes en la muestra. La reacción en cadena de la polimerasa (PCR por sus siglas en inglés) se ha utilizado ampliamente en ecología microbiana ya que sirve para amplificar copias de secuencias del gen 16S con el fin de estudiar las bacterias presentes en las muestras (Eckburg et al., 2005). Por otro lado, la secuenciación de Sanger también ha sido la técnica estándar durante muchos años para la secuenciación del DNA, sin embargo, su uso es limitado en ecología microbiana. La

aparición de la secuenciación masiva paralela ha facilitado el estudio de la diversidad del microbioma como un todo, los cambios en su composición bajo diferentes condiciones, el descubrimiento de nuevos microrganismos y las relaciones filogenéticas entre ellos (Bishop, 2014).

El análisis mediante NGS de la microbiota de una muestra consiste en la identificación de microorganismos a partir de lecturas cortas de secuenciación. En el caso de muestras que contengan más de una bacteria, por ejemplo muestras ecológicas del medio ambiente o muestras biológicas o de tejidos, el análisis NGS se realiza principalmente por medio de la técnica de DNA *barcoding*, que consiste en utilizar secuencias de fragmentos de genes conservados para identificar especies en base a la taxonomía conocida (Chakraborty et al., 2014). El *barcoding* se aplica utilizando una o varias regiones hipervariables del gen que codifica para la subunidad 16S del RNA ribosómico (16S rRNA) bacteriano. Este gen consiste en una secuencia que se encuentra en todas las especies bacterianas y tiene una longitud aproximada de 1.500 pares de bases. Se conserva evolutivamente pero tiene 9 regiones hipervariables (V1-V9) con un alto poder de discriminación (Vinje et al., 2014), por lo que se suelen utilizar una de esas regiones hipervariables, o la combinación de varias, para la secuenciación con el fin de la asignación taxonómica. Brevemente, los pasos básicos para hacer un estudio de meta-taxonomía se dividen en dos grupos; en primer lugar, la amplificación de la región hipervariable en cuestión del gen de la subunidad 16S del RNA ribosomal y la secuenciación de los productos amplificados, y, en segundo lugar, el análisis bioinformático de los amplicones 16S (Figura 2). Para llevar a cabo estos pasos es necesario el uso de la bioinformática como herramienta para extraer resultados fiables, por lo que se requiere un conocimiento específico previo de lenguajes de programación y de las herramientas bioinformáticas estándar disponibles. El análisis bioinformático tradicional incluye el control de calidad de las secuencias, la eliminación de secuencias quiméricas, la clusterización de las secuencias obtenidas en función de la similitud entre las mismas, y la asignación taxonómica para identificar la abundancia relativa de cada bacteria en la muestra (Logares et al., 2014; Jovel et al., 2016). Tradicionalmente las secuencias se agrupan en unidades taxonómicas operacionales (OTU) basadas en umbrales de similitud definidos arbitrariamente, comúnmente, las OTUs se han definido

con una similitud del 97% aunque en los últimos años, muchos autores consideran que este límite es demasiado bajo (Yarza et al., 2014). Como alternativa de mayor resolución que las OTUs, gracias a los recientes desarrollos metodológicos, surge las ASVs (amplicon sequence variant) que difieren en tan solo un nucleótido en su secuencia (Callahan et al., 2017).

La NGS ha llevado a una explosión de estudios focalizados en la comprensión de la composición y función de las poblaciones bacterianas en ambientes muy diversos (Project and Consortium, 2012; Yoon et al., 2015). En el ganado, muchos estudios se han centrado en el estudio de la microbiota del intestino, por ejemplo para estudiar la relación entre la microbiota intestinal y el peso corporal de los lechones destetados en un entorno comercial (Han et al., 2017), para definir un *core* en la microbiota intestinal porcina (Holman et al., 2017) o para hacer un meta-análisis de la composición de la microbiota del tracto gastrointestinal en ganado bovino (Holman and Gzyl, 2019). La investigación sobre la microbiota ruminal está siendo cada vez más importante en relación a la cría de rumiantes, sobre todo en ganado vacuno, ya que las comunidades microbianas y su expresión del genoma están relacionadas con rasgos importantes como el estado de salud (Zilber-Rosenberg and Rosenberg, 2008), la fermentación del alimento (Wilson et al., 2019), o emisiones de metano (Kamke et al., 2016; Vasta et al., 2019). Las diferencias en la composición de la microbiota también pueden utilizarse para predecir las diferencias entre rasgos complejos (Kamke et al., 2016). Concretamente en el ganado lechero, mayormente en vacuno, uno de los puntos objeto de estudio es la caracterización de la microbiota de la leche recolectada directamente de la glándula mamaria (Oikonomou et al., 2014; Addis et al., 2016; Oultram et al., 2017). También se ha estudiado, aunque en menor medida, la microbiota de la leche de búfala (Catozzi et al., 2017). Dada la importancia de la mastitis, como ya se ha mencionado anteriormente, en el ganado vacuno de leche muchos de estos estudios comparan la microbiota de la leche de muestras sanas con la microbiota de la leche de muestras con mastitis clínica y subclínica (Kuehn et al., 2013; Oultram et al., 2017).

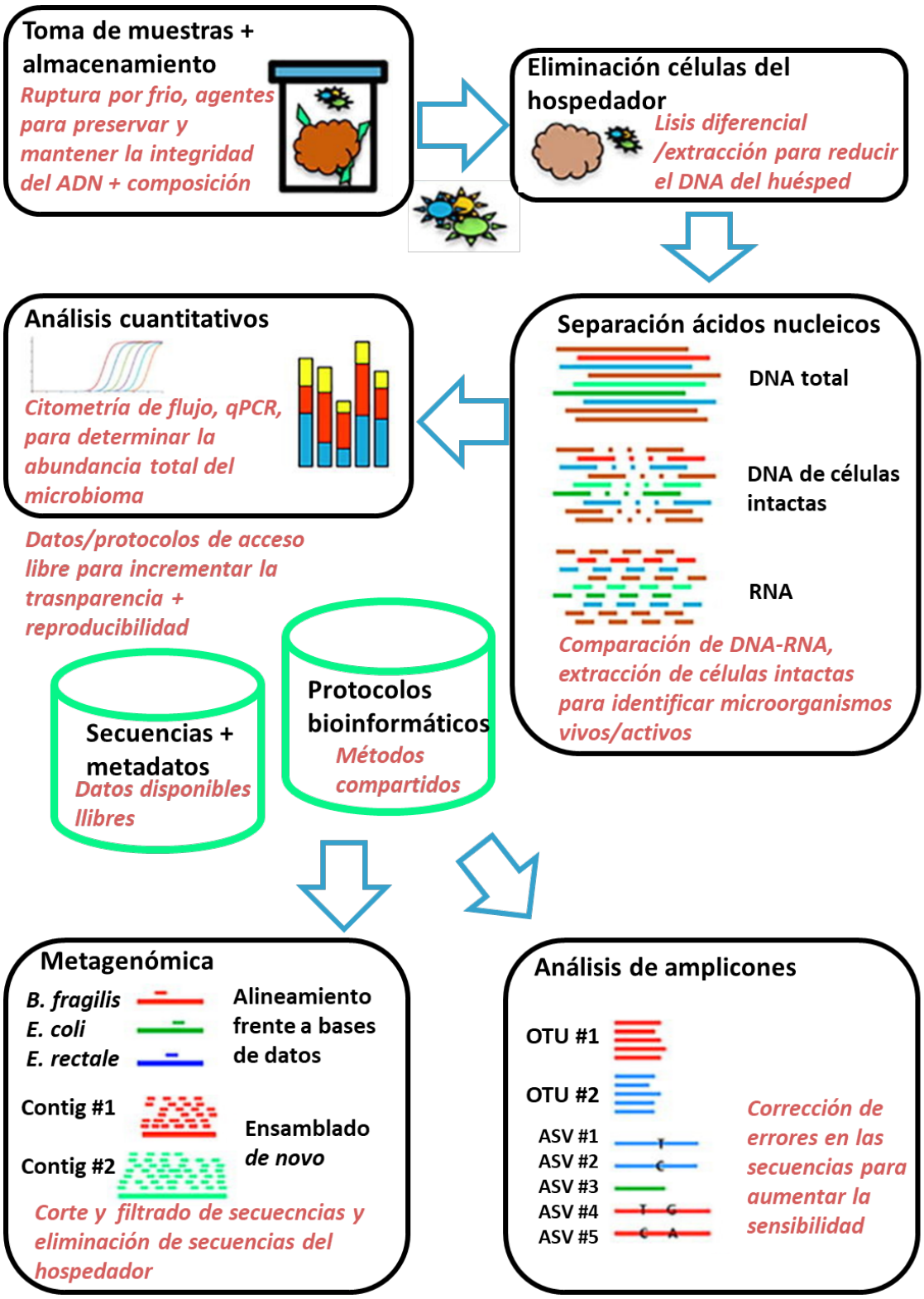


Figura 2. Recomendaciones para el análisis de la microbiota utilizando secuenciación masiva (fuente Fricker et al., 2019).

Por otro lado, es posible estudiar la totalidad de la información genética presente en una comunidad microbiana mediante la secuenciación integral de genomas completos utilizando secuenciación masiva paralela, sin los sesgos que pueden derivarse de utilizar

la PCR para amplificar el gen 16S rRNA (Acinas et al., 2005; Schloss et al., 2011). Se trata del denominado *shotgun process* que consisten en secuenciar por separado pequeños fragmentos de una muestra (o una comunidad). En contraste con los estudios basados en el gen 16S rRNA, en la aproximación metagenómica de alto rendimiento no se seleccionan y amplifican secuencias, sino que el DNA total de la muestra se corta en fragmentos de longitud definida que son secuenciados posteriormente. Este nuevo enfoque permite no solo obtener información sobre los genes codificados por el DNA bacteriano, sino también sobre los posibles roles de cada microorganismo en un determinado ambiente mediante el estudio de la función de los productos proteicos para los que codifican dichos genes. Esta asignación funcional está basada en la comparación de proteínas con función conocida descritas en bases de datos de referencia como las incluidas en la base de datos KEGG - *Kyoto Encyclopedia of Genes and Genomes* (http://www.kegg.jp/) y en la base de datos COG - *Clusters of Orthologous Groups of proteins* (Tatusov et al., 2003). Por lo tanto, en comparación con los resultados de la secuenciación del gen 16S rRNA, este tipo de secuenciación ofrece una visión más profunda y completa de la composición microbiana, así como de la estructura y las funciones metabólicas de esa microbiota presente en la muestra analizada (Gaeta et al., 2017).

3.4. La secuenciación de tercera generación

Las tecnologías de segunda generación, aun con el abaratamiento que han sufrido en los últimos años, son todavía relativamente caras y tienen una serie de requerimientos específicos. Por ello se ha intentado buscar alternativas más sencillas, rápidas y eficaces al menor coste posible, surgiendo como consecuencia de ello la secuenciación de tercera generación que se basa en la detección de una sola molécula donde la óptica del detector es lo suficientemente sensible para la lectura directa de ácidos nucleicos individuales (Pettersson et al., 2009; Schadt et al., 2010). Por lo tanto, en este caso no es necesaria la amplificación previa de fragmentos de DNA, como en el caso de la secuenciación de lecturas cortas (NGS), generándose lecturas más largas (~10-30 Kb) aunque se requiere una cantidad relativamente grande de DNA (entre 250-5.000 ng, en función de la tecnología) (Børsting and Morling, 2015). En general, la utilización de lecturas largas en el posterior análisis bioinformático mejora el ensamblaje *de novo*, el

ensamblaje de genomas con grandes extensiones de regiones repetidas y el análisis del transcriptoma (Marchet et al., 2018). Además se considera que estas tecnologías de secuenciación serán de gran relevancia en el campo de los estudios metagenómicos (Branton et al., 2009).

Actualmente hay dos plataformas de secuenciación de lecturas largas ampliamente utilizadas; el secuenciador de Pacific Biosiciences (PacBio) que se basa en la técnica SMRT (Single-Molecule Real-Time), y el secuenciador MinION de Oxford Nanopore Technologies (ONT), el primer secuenciador que utiliza tecnologías de nanoporos. La primera de estas plataformas utiliza la fluorescencia para secuenciar una única molécula de DNA a tiempo real, para lo que es necesario que cada nucleótido esté marcado con un fluoróforo incorporado por la polimerasa (Eid et al., 2009). Por otro lado, las tecnologías de ONT se basan en el paso de una corriente iónica a través de nanoporos, la secuencia de DNA se infiere a partir de los cambios en la corriente iónica medida a través de una membrana (Branton et al., 2009; Clarke et al., 2009; Jain et al., 2016). A pesar de los grandes beneficios que tienen estas tecnologías, como por ejemplo la enorme mejoría de la calidad del ensamblaje, las tasas de error observadas son mucho más altas a nivel de base que las lecturas cortas obtenidas con la tecnología Illumina (Ip et al., 2015). Para mejorar estas tasas de errores, muchos estudios realizan una corrección híbrida de las lecturas largas de nanoporos utilizando datos MiSeq (Illumina) complementarios, y de esta manera consiguen producir un ensamblaje *de novo* altamente continuo corrigiendo los datos de secuencias largas con secuencias cortas de Illumina (Goodwin et al., 2015; Tyson et al., 2018; Jung et al., 2019).

La secuenciación utilizando el aparato comercial MinION (Oxford Nanopore Technologies) se ha utilizado ampliamente para secuenciar genomas virales y bacterianos gracias a las altas calidades de los ensamblados de lecturas largas (Quick et al., 2016; Votintseva et al., 2017; Goldstein et al., 2019). Hasta hace poco, el rendimiento relativamente bajo de este instrumento ha limitado su utilización para el estudio de regiones específicas en el genoma humano. Sin embargo, los avances recientes que aumentan la precisión con la que se puede determinar la secuencia de DNA han aumentado este rendimiento produciendo secuencias con coberturas más altas y haciendo posible el estudio de genomas superiores (Minervini et al., 2016, 2017). Esta

tecnología está empezando a utilizarse como técnica de diagnóstico para la sospecha de enfermedades infecciosas, por ejemplo para la identificación y caracterización genética del enterovirus en vacuno (Beato et al., 2018), para la detección de patógenos causantes de enfermedades entéricas en cerdos (Theuns et al., 2018), o para el diagnóstico de la viruela aviar mediante secuenciación del genoma completo del avipoxvirus (Croville et al., 2018). Además, estudios del microbioma del rumen en el ganado bovino están revelando que la secuenciación ONT ofrece unos resultados similares a la secuenciación de Illumina, clasificando un mayor número de lecturas a nivel de especie, y por lo tanto puede posicionarse como una alternativa interesante para caracterizar microbiomas de interés en animales (Delgado et al., 2019b). En especies domésticas lecheras, como el búfalo, se ha puesto de manifiesto que la secuenciación utilizando la tecnología MinION para la identificación de patógenos potenciales y de bacterias resistentes en muestras de leche representa una de las futuras aplicaciones de esta tecnología (Catozzi et al., 2020).

4. Metodología

Esta Tesis Doctoral se presenta mediante la modalidad “compendio de publicaciones” por lo que la descripción de los materiales y métodos utilizados para llevar a cabo los análisis desarrollados en ella se describen en la sección de “Materiales y Métodos” de las publicaciones presentadas.

Las publicaciones se incluyen en la sección de “Resultados” de este mismo documento.

5. Resultados

Objetivo 1: Utilización de la secuenciación del genoma completo (WGS del inglés *Whole Genome Sequencing*) para el análisis de alta resolución en regiones genómicas de interés:

1.1. **Esteban-Blanco, C., Gutiérrez-Gil, B., Suárez-Vega, A., López-Iglesias, L.J. y Arranz, J.J**. Identificación de polimorfismos en regiones genómicas caracterizadas como huellas de selección en el ganado ovino. XVII Jornadas sobre Producción Animal, AIDA-ITEA, Zaragoza 30 y 31 de mayo de 2017.

1.2. **Gutiérrez-Gil, B., Esteban-Blanco, C., Wiener, P., Chitneedi, P.K., Suarez-Vega, A., and Arranz, J.J.**. 2017. High-resolution analysis of selection sweeps identified between fine-wool Merino and coarse-wool Churra sheep breeds. Genetics Selection Evolution. 49:81. doi:10.1186/s12711-017-0354-x.

1.3. **Gutiérrez-Gil. B.*, Esteban-Blanco. C.*, Suarez-Vega. A., and Arranz. J.J.**. 2018. Detection of quantitative trait loci and putative causal variants affecting somatic cell score in dairy sheep by using a 50K SNP-Chip and whole genome sequencing. Journal of Dairy Science. doi:10.3168/jds.2018-14736. (*****These two authors equally contributed to this work)

Objetivo 2: Utilización de la secuenciación masiva paralela para caracterizar la microbiota de la leche de oveja y su posible asociación con caracteres de resistencia a la mastitis.

2.1. **Esteban-Blanco, C., Gutiérrez-Gil, B., Marina-García, H., Linaje, B., Acedo A. y Arranz, J.J**. Estudio preliminar sobre la caracterización del microbioma de la glándula mamaria en ovejas assaf en lactación. XIX

Reunión Nacional de Mejora Genética Animal, León 14 y 15 de junio de 2018.

2.2. **Esteban-Blanco. C., Gutierrez-Gil. B., Puente-Sanchez. F., Marina. H., Tamames. J., Acedo. A., and Arranz. J.J.**. 2019. Microbiota characterization of sheep milk and its association with somatic cell count using 16s rRNA gene sequencing. Journal of Animal. Breeding and Genetics. doi:10.1111/jbg.12446.

2.3. **Esteban-Blanco, C., Gutiérrez-Gil, B., Marina, H., and Arranz, J.J..** Comparison of sheep milk microbiome in two dairy sheep breeds using 16S rRNA gene sequencing. *Artículo en preparación*.

2.4. **Esteban-Blanco C., Puente-Sánchez F., Gutiérrez-Gil B., Marina H., Tamames J., Arranz J.J.**. "Metagenomic de novo assembly of Corynebacterium bovis in lactating assaf sheep: a preliminary study". The 37th International Society for Animal Genetics Conference, Lleida 7-12 de julio de 2019.

2.5. **Esteban-Blanco, C., Gutiérrez-Gil, B., Marina, H., and Arranz, J.J..** First insights of shotgun metagenomics in sheep milk microbiota. *Artículo en preparación*.

La caracterización de la microbiota puede analizarse mediante otras tecnologías diferentes a las expuestas en el Objetivo 2. Se incluyen aquí dos aproximaciones para el estudio de los microorganismos que utilizan datos de secuenciación de segunda y tercera generación.

2.6. **Esteban-Blanco, C., Gutiérrez-Gil, B., Marina, H., Suárez-Vega, A. and Arranz, J.J..** Milk bacterial diversity between two sheep breeds using rnaseq dataset. *Artículo en preparación*.

2.7. **Esteban-Blanco, C., Gutiérrez-Gil, B., Marina, H., and Arranz, J.J..** Diversity and community composition of rumen microbiota in sheep using long reads from nanopore sequencing. *(enviado a 71st Annual Meeting of European Federation of Animal Science)*.

Resultado 1.1

Identificación de polimorfismos en regiones genómicas caracterizadas como huellas de selección en el ganado ovino

Cristina Esteban-Blanco, Beatriz Gutiérrez-Gil, Aroa Suárez-Vega, Luis Juan López-Iglesias, y Juan José Arranz

Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain;

XVII Jornadas sobre Producción Animal 2017, Tomo I (ed. Asociación Interprofesional para el Desarrollo Agrario) 555-557 (INO Reproducciones S.A.).

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.

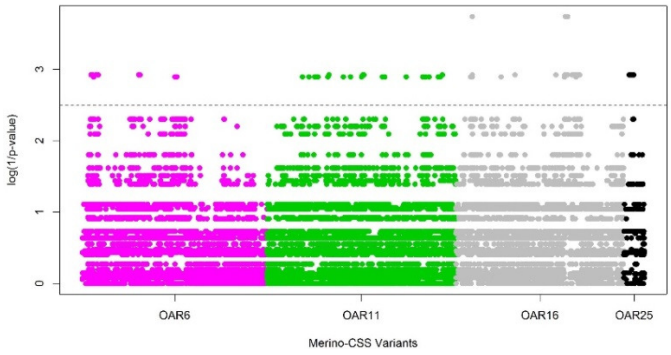


Figura 1. Identificación de variantes divergentes en cuatro regiones definidas previamente como huellas de selección en razas de Merino Australiano mediante el análisis de datos de secuenciación genómica de las razas Merino y Churra.

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.

Resultado 1.2

High-resolution analysis of selection sweeps identified between fine-wool Merino and coarse-wool Churra sheep breeds

B.Gutiérrez-Gil[1], C. Esteban-Blanco[1,2], P. Wiener[3], P. Krishna Chitneedi[1], A. Suárez-Vega & J.J. Arranz[1]

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain; [2]Fundación Centro Supercomputación de Castilla y León, Campus de Vegazana, León 24071, Spain; [3]Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush Cam- pus, Midlothian EH25 9RG, UK.

Genetics Selection Evolution 49(1), 81. https://doi.org/10.1186/s12711-017-0354-x

Gutiérrez-Gil *et al. Genet Sel Evol (2017) 49:81*
DOI 10.1186/s12711-017-0354-x

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

High-resolution analysis of selection sweeps identified between fine-wool Merino and coarse-wool Churra sheep breeds

Beatriz Gutiérrez-Gil[1*], Cristina Esteban-Blanco[1,2], Pamela Wiener[3], Praveen Krishna Chitneedi[1], Aroa Suarez-Vega[1] and Juan-Jose Arranz[1]

Abstract

Background: With the aim of identifying selection signals in three Merino sheep lines that are highly specialized for fine wool production (Australian Industry Merino, Australian Merino and Australian Poll Merino) and considering that these lines have been subjected to selection not only for wool traits but also for growth and carcass traits and parasite resistance, we contrasted the OvineSNP50 BeadChip (50 K-chip) pooled genotypes of these Merino lines with the genotypes of a coarse-wool breed, phylogenetically related breed, Spanish Churra dairy sheep. Genome re-sequencing datasets of the two breeds were analyzed to further explore the genetic variation of the regions initially identified as putative selection signals.

Results: Based on the 50 K-chip genotypes, we used the overlapping selection signals (SS) identified by four selection sweep mapping analyses (that detect genetic differentiation, reduced heterozygosity and patterns of haplotype diversity) to define 18 convergence candidate regions (CCR), five associated with positive selection in Australian Merino and the remainder indicating positive selection in Churra. Subsequent analysis of whole-genome sequences from 15 Churra and 13 Merino samples identified 142,400 genetic variants (139,745 bi-allelic SNPs and 2655 indels) within the 18 defined CCR. Annotation of 1291 variants that were significantly associated with breed identity between Churra and Merino samples identified 257 intragenic variants that caused 296 functional annotation variants, 275 of which were located across 31 coding genes. Among these, four synonymous and four missense variants (*NPR2_His847Arg*, *NCAPG_Ser585Phe*, *LCORL_Asp1214Glu* and *LCORL_Ile1441Leu*) were included.

Conclusions: Here, we report the mapping and genetic variation of 18 selection signatures that were identified between Australian Merino and Spanish Churra sheep breeds, which were validated by an additional contrast between Spanish Merino and Churra genotypes. Analysis of whole-genome sequencing datasets allowed us to identify divergent variants that may be viewed as candidates involved in the phenotypic differences for wool, growth and meat production/quality traits between the breeds analyzed. The four missense variants located in the *NPR2*, *NCAPG* and *LCORL* genes may be related to selection sweep regions previously identified and various QTL reported in sheep in relation to growth traits and carcass composition.

Background

Approximately 5000 years ago, humans began to select sheep for desired characteristics (e.g., coat color, horns, meat, wool) which resulted in the development of different breeds [1]. Initially, sheep were reared mainly for meat; later, specialization for 'secondary' products, such as wool, emerged [2–4]. Sheep that have been selected for secondary products appear to have replaced the more primitive domestic populations. Selection for such phenotypes has left detectable signatures of selection within the genome of modern sheep. Due to the very strong selection intensity involved in animal breeding,

*Correspondence: beatriz.gutierrez@unileon.es
[1] Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain
Full list of author information is available at the end of the article

these changes are expected to occur faster than those due to natural selection. Selection not only affects a favored mutation by rapidly increasing its frequency in the population, but it also produces a hitch-hiking effect of the frequency of neutral alleles at linked loci [5, 6]; these patterns in allele frequencies are known as selection signatures.

The signatures of selection in the genome, also known as selection sweeps, can be detected under the assumption that selection is locus-specific whereas other evolutionary forces such as random genetic drift, mutation and inbreeding, should be expressed genome-wide [7]. Hence, a variety of methods and statistics have been developed with the aim of identifying the selected loci at which allele frequencies have changed following a pattern that is consistent with positive selection. They can be based on between-population differentiation, reductions in local variability, deviations in the site frequency spectrum (SFS), and increases in linkage disequilibrium (LD) and extended haplotype structure [8–10]. Methods for detecting signatures of selection have historically been challenged by the confounding effects of demography; for example, recent population growth will result in an excess of rare variants compared to equilibrium expectation [11], and also recent and weak bottlenecks tend to mimic the effects of a selection sweep in several ways [12]. However, demographic events apply to the whole genome, whereas selective events affect different regions of the genome to various extents thanks to recombination [13]. This gives the possibility of distinguishing the two hypotheses by sampling several loci: a more or less common pattern is expected in the case of a bottleneck, while selective sweeps generate heterogeneity across loci [12]. Regions of low recombination may also produce an upward bias in the detection of signatures of selection, although apart from the bias issue, it should be noted that genuine selection sweeps in these regions will leave much stronger signals in regions of average recombination rate [14].

Another issue relates to the limitations of some selection mapping approaches; the standard approaches to detect signatures of selection consider "hard sweeps" where the new advantageous mutations spread rapidly to fixation, purging variation at linked sites as they spread. However, recent studies highlight the potential importance of 'soft sweeps', i.e., sweeps from standing variation, or sweeps in which multiple mutations start to sweep simultaneously at a single locus [15]. Soft sweeps, which are often related to adaptation, leave more subtle signatures in the genome (e.g. diversity is not necessarily reduced in the vicinity of the adaptive locus as with hard sweeps) and thus are more difficult to detect [16].

In past years, many genome screening studies based on high-density, genome-wide single nucleotide polymorphism (SNP) panels (i.e., SNP-chips) have been conducted with the goal of detecting signatures of selection in livestock species [17–19]. More recently, whole-genome re-sequencing has emerged as an economically feasible tool for assessing genomic variation within and among populations, and the large-scale information derived from the new sequencing technologies can be exploited to identify signatures of selection [20] or further explore previously detected signatures of selection.

The Sheep HapMap project, for which genotypes were generated from 3004 domestic sheep from 71 breeds using the Illumina OvineSNP50K BeadChip assay (50 K-chip), generated valuable information that can be used to perform analyses of signatures of selection in sheep [21]. Global analyses of genetic differentiation in the Sheep HapMap dataset identified several genomic regions that contained genes for coat pigmentation, skeletal morphology, body size, growth, and reproduction [21, 22]. Many of these regions were later confirmed by haplotype-based selection sweep mapping [23, 24]. Also based on this dataset, as well as additional information in some cases, signatures of selection have been reported in thin and fat tail sheep breeds [25] and in specialized European dairy sheep breeds [19]. Further selection mapping studies have identified signatures of selection related to resistance/susceptibility to gastrointestinal nematodes [26], adaptation to different ecoregions [27] or climate adaptation [28, 29]. Information from additional studies using the 50 K-chip to study the biodiversity of sheep breeds [30–32] can also help to extend our current knowledge on the ovine genomic regions that have been affected by human-driven selection.

In sheep, selection for wool traits has been extensively carried out for several centuries. Spanish Merino, which was developed since the late Middle Ages [33], appears to have originated during Roman times through the introduction of fine wool ewes from the Southern Italian region of Apulia into Spain and the later selection for white wool color through crosses with African rams imported by Arabs [34] at the beginning of the Middle Ages. Due to the value of their fiber, the Honourable Council of the Mesta strongly protected Merino flocks, and their exportation was strictly forbidden for several centuries. Removal of these restrictions in the eighteenth century led to the dispersal of Merino sheep to Eastern Europe, China, Australia and New Zealand [34]. The first Merino sheep were introduced from South Africa into Australia in 1797 [35]. In this country, intensive selective breeding has enhanced the already fine quality of the wool to produce Australian Merino wool, which

based on its long, fine fibers, enables the production of lighter and softer wool fabrics. Hence by 1870, Australian Merino wool industry was the global leader in both the quantity and quality of its wool production. However, in addition to wool, Australian Merino also plays an important role in lamb meat production. Over the last two decades, the Australian sheep meat industry has delivered large increases in lamb production and profitability, with genetic improvement in growth, leanness and muscling contributing substantially to these gains [36, 37]. Australian Merino flocks have also been selected for disease resistance, in particular, by focusing on gastrointestinal nematode parasites, flystrike (cutaneous myiasis) and footrot [38]. Specifically, the three Merino lines considered in this paper have been subjected to selection pressure to reduce susceptibility to parasites [21].

The goal of our work was the identification of regions of selection sweeps related to traits for which Australian fine-wool Merino breeds have been selected. Considering the Iberian origin of Merino breeds and the estimated divergence time between sheep breeds derived from the analysis of the Sheep HapMap project dataset [21], we analyzed genome-wide SNP information from three Australian Merino breeds that are highly specialized for the production of fine wool (Australian Industry Merino, Australian Merino and Australian Poll Merino) and the related coarse-wool breed, Spanish Churra. This is an autochthonous double-purpose breed of the northwest region of Castilla y León in Spain. Traditional Churra flocks are managed based on an intermediate level of dairy specialization (the dairy breeding program was started in 1986 [39]) with a variable fraction of the farm income derived from the sale of suckling lamb meat. The milk is used to produce cheese of high quality value, which is covered by a protected geographical indication (PGI) [40].

With the aim of further exploring the genetic variation in genomic regions that show evidence of selection, whole-genome sequence data from Churra and Australian Merino samples were subsequently analyzed. By identifying the SNPs within the regions of interest that exhibited the most extreme divergence in allele frequencies between the Churra and Merino datasets, this study provides a detailed survey of the genetic variation that underlies the identified regions of selection sweeps.

Methods

Mapping of selection sweeps

Breed selection

In order to identify selection sweeps related to fine-wool production, three Merino lines described as "with extreme fine wool" from the International Sheep Genomics (ISGC) dataset were included in this study. According to the divergence times estimated for the breeds included within the Sheep HapMap project, based on LD and haplotype sharing, these three Merino lines show a recent divergence time (0 to 80 generations) (Figure S10 and Fig. 3 from Kijas et al. [21]). With the aim of providing an appropriate comparison to identify signatures of selection related to wool production, we selected the Spanish Churra sheep breed, which is a coarse-wool breed related to Merino, as shown by the population analysis reported by Fariello et al. [22] in which these two breeds are grouped together within the defined South West European group. According to a haplotype-sharing analysis, Churra sheep show a short and consistent divergence time with each of the three Merino lines (160 to 240 years, see Figure S10 and Fig. 3 from Ref [21]), which supports our study design in which the three Australian sheep breeds selected are considered together against Churra sheep. In addition to wool characteristics, Merino and Churra sheep differ in other traits (see Additional file 1: Table S1) (Fig. 1). Briefly, adult animals of the Australian Merino lines are larger than Churra sheep individuals, whereas weight at birth is similar in the two breeds. Both breeds show white wool color although Churra sheep show characteristic black patches around the eyes, ears and the ends of legs. Note that because the two breeds differ in various traits, the signatures of selection identified here may be related not only to wool traits but also to other phenotypes for which the selection pressure performed in the two breeds differs. Considering the possibility that the geographical isolation and distance between the two studied populations could be a confounding effect for the identified selection sweeps, we performed additional validation analyses by contrasting Churra and Spanish Merino breeds.

Genotypes, quality control and analysis of population structure

We included in this work an initial subset of SNP genotypes for the ovine 50 K-chip that were generated within the framework of the Sheep HapMap project [21], and which are available upon request (http://www.sheephapmap.org/termsofaccess.php). The extracted subset included 332 samples from the Australian Industry Merino (n = 88), Australian Merino (n = 50), Australian Poll Merino (n = 98) and Spanish Churra (n = 96) breeds. In addition, 184 DNA samples of Churra sires included in the selection nucleus of the National Association of Spanish Churra Breeders were also genotyped with the same SNP array. These samples were extracted from semen samples following a classical phenol–chloroform DNA extraction protocol [41], and genotyped by an external laboratory service. The raw genotypes for the 54,241 SNPs included in the genotyping platform were

Fig. 1 Sheep breeds selected for this study, Australian Merino (left) and Spanish Churra (right). Original images taken from Wikipedia (https://commons.wikimedia.org/w/index.php?curid=12599612; https://commons.wikimedia.org/w/index.php?curid=12174588)

first analyzed with the GenomeStudio software (Illumina) (GenCall score for raw genotypes > 0.15) which was used to extract the genotypes in standard format for the Plink_v1.09 software [42].

The HapMap project samples had already been subjected to quality control (QC) filtering [21], resulting in 49,034 SNPs available for analysis. To join the two separate datasets, we first merged the new Churra dataset (n = 184; 54,241 SNPs) and the HapMap dataset (n = 332; 49,304 SNPs) based on the common SNPs. We then selected the SNPs that mapped, with positions based on sheep genome assembly Oar_v3.1 [43], on the ovine autosomes, resulting in 47,415 SNPs. This dataset was then subjected to the following filtering criteria: (1) individual call rate higher than 90% (two Churra individuals genotyped by our group were removed) and (2) marker call rate higher than 90% (28 SNPs removed due to missing genotype data). Hence, 514 individuals (Churra = 278 and Merino = 236) and 47,387 SNPs were available for further analyses.

To evaluate the genetic structure of the data and confirm the number of different genetic populations, the genotypes of the three Merino fine wool breeds and the genotypes of the Churra individuals were analysed by principal component analysis (PCA) of allele sharing (using smartpca, implemented in Eigensoft [44]), and ancestry estimation (Admixture software [45]). The results of these analyses identified two clearly distinct genetic populations, corresponding to the Merino group and Churra sheep, for details on these analyses and description of results (see Additional file 2 and Additional file 3: Figures S1, S2 and S3). Genotypes were pooled into a single Australian Merino dataset for the three Australian Merino populations.

In addition, as a further validation analysis, we compared the genotypes of 20 randomly chosen Churra samples from the Sheep HapMap dataset and the 20 Spanish Merino samples genotyped by Ciani et al. [34].

Identification of candidate regions under selection using individual analyses

Several analyses between the complete set of Spanish Churra and Australian Merino genotypes were performed to detect candidate regions that harbor signatures of selection. First, a genetic differentiation analysis was used to contrast the Australian Merino and Churra genotypes by calculating the unbiased estimate of Weir and Cockerham's F_{ST} [8] for each SNP, as described by Akey et al. [46]. In a second analysis, regions of reduced heterozygosity in the two groups were identified by estimating the observed heterozygosity (ObsHtz) for each SNP. For these two analyses, F_{ST} and ObsHtz values estimated for each SNP were each averaged across a sliding window of nine SNPs (e.g., F_{ST}_9SNPW). The size of the sliding window was based on a previous analysis by Gutierrez-Gil et al. [19] for a test control region encompassing the *myostatin* (*GDF-8*) gene, which is known to have been under selection in the Texel breed. The identification of candidate signatures of selection in each of the individual analyses was based on window estimates at the extreme of the empirical distributions, as suggested by Akey et al. [46] and has been used in a number of subsequent studies [18, 19, 47–49]. Specifically, we considered that a position carried a signature of selection if it was in the top 0.5th percent of the distributions for genetic differentiation (F_{ST}) or the bottom 0.5th percent for observed heterozygosity. The distribution of the physical sizes of windows based on the 9-SNP fixed-size criteria (average window size = 411.71 kb; average distance between central SNPs of consecutive windows = 51.53 kb) was found to be fairly narrow (98.28% of the windows were 200 to 600 kb long; only 0.74% of the windows were

longer than 1000 kb) and thus should provide reasonable estimates of local genomic diversity, in contrast to analyses based on a low-density chip [50].

As a complementary approach to map selection sweeps, we used the hapflk_v1.3 software (https://forge-dga.jouy.inra.fr/projects/hapflk), which implements the FLK [51] and hapFLK [23] tests. The FLK metric tests the neutrality of polymorphic markers by contrasting their allele frequencies in a set of populations against what is expected under a neutral evolution scenario. The hapFLK statistic extends the FLK test to account for the differences in haplotype frequencies between populations. This method has been shown to be robust with respect to bottlenecks and migration [23]. To run the hapflk analysis, the Reynolds' distances between the Churra and Merino populations were converted to a kinship matrix with an R script provided by the hapFLK developers (available at https://forge-dga.jouy.inra.fr/projects/hapflk/documents). Subsequently, by assuming 20 haplotype clusters in the LD model (-K 20; number of haplotype clusters determined by running a fastPHASE cross-validation analysis), the hapFLK statistics were later computed and averaged across 30 EM runs to fit the LD model (–nfit = 30). The standardization of the statistics using the corresponding python script provided with the software allowed the estimation of the associated *P* values from a standard normal distribution. To correct for multiple testing, we considered the threshold of the nominal *P* value as < 0.001 to identify the significant haplotypes, following previous studies using hapFLK analysis on the Sheep HapMap dataset [22, 24].

In addition, we used the *rehh* software [52] to perform an additional analysis based on the cross-population extended haplotype homozygosity (XP-EHH) test defined by Sabeti et al. [53]. This statistic compares the EHH profiles for bi-allelic SNPs between two populations and is defined, for a given allele, as the log of the ratio of the integrals of the EHH profiles between the two populations. The comparison between populations normalizes the effects of large-scale variation in recombination rates on haplotype diversity and has a high statistical power to detect sweeps that are close to fixation [53]. Alleles were designated at each locus as either minor ("1", "ancestral") or major ("2", "derived"), based on their allele frequency in the overall population. Positive and negative XP-EHH estimates indicated positive recent selection in Churra and Merino, respectively. Based on the *P* values supplied by *rehh*, and for consistency with the threshold previously used for hapFLK, we considered as significant those positions showing a *P* value less than 0.001.

For the four selection sweep mapping analyses, positions that showed evidence of selection (i.e. included in the top/bottom 0.5th percent of the corresponding distribution or showing a *P* value less than 0.001) and within 0.150 Mb of each other were considered to be the result of the same selection sweep and were labeled, depending on the analysis method, as F_{ST}-SS, Merino-ObsHtz-SS, Churra-ObsHtz-SS, hapFLK-SS and XP-EHH-SS. This criterion to connect identified SNPs into discrete regions was established based on an exploratory analysis of the extent of LD and the haplotype block structure of the Churra and Merino populations. Based on the results of LD analysis performed with Haploview_v4.2 [54] (for details see Additional file 4 and Additional file 5: Figure S4), and following Tang et al. [55], we initially considered regions of 50 kb (based on the fact that among the identified haplotype blocks, the proportion of blocks of size 50 kb or more was 43.65% and 50.59% in Merino and Churra, respectively) and extended these regions by 50 kb in both directions [based on the fact that the estimation of half-length decay in LD in the two breeds was around 50 kb (see Additional file 4)].

The four selection sweep mapping analyses described above were subsequently performed on the 20 Spanish Churra and 20 Spanish Merino samples selected for the validation analysis, using the same criteria to identify positions showing evidence of selection and to group these positions into selection sweeps.

Identification of shared regions across methods

Considering the results obtained in the Australian Merino versus Churra analyses, those regions showing an overlap between at least one of the two methods based on haplotype analysis (hapFLK and XP-EHH) and at least one of the two other considered methods (F_{ST} and ObsHz), were labeled as convergence candidate regions (CCR). The coordinates of the identified CCR were compared with previously reported ovine selective sweeps and previously described sheep QTL in these regions based on the Animal QTLdatabase [56]. An initial assessment of possible functional candidate genes that mapped within the identified CCR was performed using Ensembl BioMart [57] to extract the annotated genes for the relevant genomic intervals. The list of extracted genes was later contrasted with the list of 1255 genes provided by Gutierrez-Gil et al. [17], which are candidates for selection in cattle (and other livestock species) due to their known association with physical features (horns, stature, body size and coat color) or production traits (milk production, mastitis, and meat production/quality traits). This list was extended to include 148 candidate genes for wool production/quality, such as those related to hair follicle cycling (reviewed by Stenn and Paus [58]), or identified as associated with wool production/quality by genome-wide association studies (GWAS) [59] or differential expression analysis [60] (see Additional file 6: Table S2).

The same overlapping criteria were applied to the results from the validation analyses and the resulting convergence candidate regions were labelled as $CCR_{(Churra20\text{-}SpanishMerino20)}$.

High-resolution analysis of selection sweep regions

Whole-genome sequencing (WGSeq) data

WGSeq data for 13 Australian Merino and 15 Churra samples were analysed in this study. Below, we summarize the detailed description and source of the analyzed datasets, which are in Additional file 7: Table S3. Briefly, 13 of the Churra samples were sequenced by our research group and ANCHE (National Association of Breeders of Spanish Churra sheep). These samples included males from the selection nucleus of ANCHE with the largest number of daughters in the general commercial population of Spanish Churra dairy sheep. For these samples, the bam files of the reads mapping to the 18 CCR identified in this work are available in the sequence read archive (SRA) repository [24] within the Bioproject PRJNA395499. In addition, we included in our study publicly available WGSeq datasets from two different projects of the SRA repository: (a) sequencing data for two Churra and three Australian Merino samples were obtained from the "*Ovis aries* diversity study" (PRJNA160933), coordinated by the International Sheep Genomics Consortium as an extension of the Sheep HapMap project; and (b) WGSeq data for 10 Australian Merino samples generated within the "Australian CRC for Sheep Industry Innovation whole-genome sequence collection" project (PRJNA325682) carried out by the Sheep Commonwealth Government's Cooperative Research Centres (SheepCRC). All sequencing data were generated with paired-end Illumina technology (Illumina HiSeq 2000 and Hiseq 2500 sequencers).

Bioinformatics analysis

For the samples obtained from the SRA repository, the SRA-Toolkit [61] was used to convert the data to FASTQ format. Then, a common workflow was performed for all 28 WGSeq datasets. Following the criteria of Kijas et al. [62] for the identification of high-quality allelic variants within Run 1 of the Sheep genomes project (PRJEB14685 at the European Variant Archive, EVA), we performed the following five steps to identify allelic variants using GATK [63] and Samtools [64] software: (1) quality of the raw reads was assessed with the FastQC program [65]; (2) the low-quality reads were filtered with Trimmomatic [66] using filter options for paired end samples (-phred33, LEADING:5, TRAILING:5 SLIDINGWINDOW:4:20, MINLEN:36 ILLUMINACLIP: Trimmomatic-0.33/adapters/TruSeq 3-PE.fa:2:30:10); (3) alignment of samples against the reference genome OAR_v3.1 [43] with the Burrows-Wheeler aligner (BWA) [67] using the maximal exact matches (*mem*) mapping function; (4) data manipulation and preliminary statistical analysis using SAMtools [64, 68] (i.e. transformation of.*sam* files into.*bam* binary format and removal of non-mapped reads and the estimation of alignment statistics), the Picard program [69] (i.e. sorting reads, removal of duplicate reads and index building) and Genome Analysis ToolKit v3.3.0 (GATK) [63] (base quality score re-calibration and indel re-alignment); and (v) considering the reads that mapped to the 18 genomic intervals defined as CCR, a variant calling analysis of the 28 samples was done using two different algorithms: the Samtools mpileup [64, 68] analysis, using the default detection parameters, and the GATK HaplotypeCaller tool [63], using default parameters, as suggested in GATK Best Practices recommendations [70]. Using the snpSIFT software [71], filters were applied independently to each of the Samtools and GATK produced VCF files to remove lower quality variants (DP > 10, QUAL > 30, MQ > 30, QD > 5 and FS < 60). An intersect set for the 28 samples, containing those variants concordant between Samtools and GATK predictions, was extracted using BCFtools [68, 72] to produce the final VCF file.

Identification and study of divergent variability in the candidate regions

Among the variants localized in the targeted regions, we selected the SNPs that showed the most significant association with the breed identity between the Churra and Australian Merino samples. To select these SNPs, we first used the VCF-tools software [73] to filter only the SNPs that were detected in all variants and to convert the dataset into PLINK format [42]. Using the PLINK software, we first performed a quality control step on the raw genotypes by discarding SNPs and individuals with genotyping call rates lower than 90%, and SNPs with a minor allele frequency (MAF) lower than 0.01 (*--mind 0.1 --geno 0.1* --maf 0.01). We then performed a Chi square association test (using the *--assoc* option) to identify the SNPs that showed the most significant associations with breed identity and therefore that had the most extreme divergent allele frequencies between the compared populations (e.g. SNPs with genotype "11" in Churra and "22" in Merino, or vice versa). For those SNPs with significant Bonferroni-corrected *P* values, (considering the number of independent tests as the total number of tested SNPs considered), we performed a functional annotation analysis to assess the possible biological impact of the considered mutations using the Ensembl Variant Effect Predictor (VEP) software [74] (based on the annotated genes of the Oar_v3.1 reference genome). For the non-synonymous variants, the results of the SIFT software

analysis [75] regarding predicted effect on protein function were obtained from Ensembl. When the functional analysis assigned one of the divergent variants to a novel gene or pseudogene, we performed BLASTN searches (based on the ± 1.500 bp interval, centered at the SNP location) to identify orthologous genes in cattle (*Bos taurus*) and/or human (*Homo sapiens*) genomes. For one of the novel genes that harbored missense mutations, a BLASTN analysis was performed against the newly updated sheep reference genome (Oar_4.0) [76].

Results

Identification of candidate selection sweeps between Australian Merino and Churra sheep based on individual methods

Candidate regions identified by the genetic differentiation analysis

Of the F_{ST} values averaged in sliding windows of nine SNPs (Fig. 2a), the top 0.5% included 236 values, ranging from 0.119 to 0.325. Following the criteria previously described i.e. allowing a maximum gap of 0.150 Mb to define an F_{ST}-SS, 49 genomic regions, distributed over 17 autosomes, were considered as potential selection sweeps (F_{ST}-SS1 to F_{ST}-SS49; Fig. 2a and see Additional file 8: Table S4). The largest number of F_{ST}-SS was on chromosome 3 (OAR3, OAR for *Ovis aries*), where 12 signals were labelled as signatures of selection. The length of the labelled F_{ST}-SS regions varied from a single central tested SNP (including the averaged estimates of the corresponding 9-SNP window), for 17 regions, to one region involving 40 windows, spanning 1.699 Mb and including 40 SNPs (OAR2, F_{ST}-SS4).

Candidate regions identified based on reduced heterozygosity

Ninety-six Merino-ObsHtz-SS, distributed across 24 autosomes, were identified after grouping the bottom 0.5% values of the ObsHtz distribution (Fig. 2b; and see Additional file 9: Table S5). The largest candidate region identified by this analysis was located at the proximal end of OAR11 and spanned 1.120 Mb (Merino ObsHtz-SS58: 0.000012–1.120037 Mb). This region included information from 19 central tested SNPs while 45 of the Merino-ObsHtz-SS regions were defined by a single central tested SNP (including the averaged estimates of the corresponding 9-SNP window). When the same analysis was performed with the Churra genotypes, 72 genomic regions (over 23 autosomes) were identified based on the positions included in the bottom 0.5% of the ObsHtz distribution (Churra-ObsHtz-SS) (Fig. 2c; and see Additional file 10: Table S6). The largest of these regions, found on OAR2 (Churra-ObsHtz-SS5: 51.898–52.998 Mb), spanned 1.10 Mb and involved 28 9-SNP windows (i.e. 28 central tested SNPs) while 29 Churra-ObsHtz-SS were based on the averaged ObsHtz estimate assigned to single SNP position.

Candidate regions identified based on haplotype-based analyses

The hapFLK analysis identified seven significant regions (P value < 0.001), one located on OAR2 (hapFLK-SS-1) and the rest located on OAR3 (Fig. 3a; and see Additional file 11: Table S7). The longest selection sweep identified by this approach was hapFLK-SS-6, located on OAR3 (153.963–155.382 Mb). Two other candidate regions involved also an interval longer than 1 Mb: hapFLK-SS-1 (OAR2: 51.898–52.939 Mb) and hapFLK-SS-3 (OAR3: 151.088–152.393 Mb). The XP-EHH analysis identified 98 significant selection sweeps (P value < 0.001) (distributed over 12 autosomes) (Fig. 3b; and see Additional file 12: Table S8). Only six of them, located on OAR6, 11, 15 and 25, showed signatures of selection in the Merino group, whereas the remainder were identified in Churra. Seven of the regions detected by this analysis covered an interval longer than 1 Mb, with the longest selection sweep (XP-EHH-SS17) located on OAR3 (154.638–158.340 Mb). Of the significant regions identified by this analysis, 25 involved a single SNP position.

Convergence of results from the different analyses

Eighteen genomic regions were labelled as convergence candidate regions (CCR) based on the overlap of significant results based on at least one method of each of the two types of analyses performed (i.e. based on allele/genotype frequencies and on haplotype-based information). These regions were located on OAR2, 3, 6, 8, 10, 11, 15 and 25 (Table 1). The intervals included in these CCR ranged from 18.113 kb (CCR17 on OAR15) to 3.701264 Mb (CCR6 on OAR3). All the labeled convergence regions involved a significant result from the XP-EHH analysis, with a good concordance between the sign of the XP-EHH score and the regions showing a reduction in heterozygosity in Merino or Churra. Hence, five of the labeled CCR were related to positive selection in Merino and the other 13 were related to positive selection in Churra.

Fifty-two annotated genes were extracted from the five Merino-defined CCR regions (see Additional file 13: Table S9), whereas 83 genes were extracted from the Churra-CCR intervals (see Additional file 14: Table S10). By comparing these genes with our database of reference candidate genes, 15 unique genes of interest were identified based on their known association with traits targeted by selection, such as horns (*RXFP2*), stature (*LCORL* and *NCAPG*), hair follicle cycle and wool quality (*IFNG*, *DVL2*, and *TP53*), meat production/quality traits (*TPM2*, *CACNG2*, *PVALB*, *ACADVL*, *SLC2A4*, *CHRNB1*, and

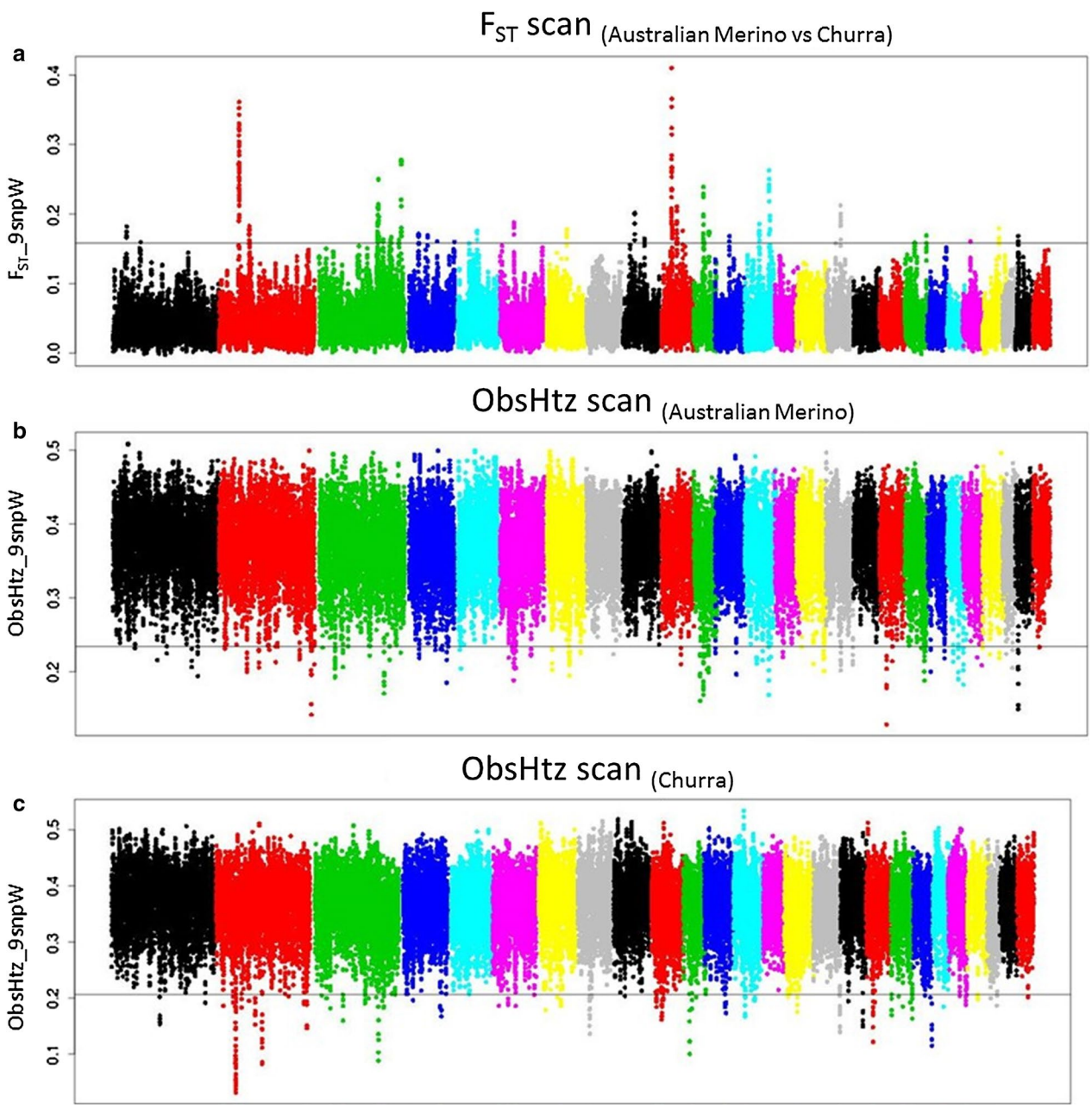


Fig. 2 Results of the genetic differentiation analysis (**a**) and of the analysis of reduced heterozygosity in the Australian Merino populations studied here (Australian Industry Merino, Australian Merino and Australian Poll Merino) (**b**) and Spanish Churra sheepAustralian Merino sheep (**c**). **a** F_{ST} values obtained across the whole-genome (averaged in sliding 9-SNP windows) when contrasting the 50 K-chip pooled genotypes of the Australian Merino and Spanish Churra sheep samples considered in this work. The horizontal line indicates the top 0.5th percent threshold of the F_{ST}-distribution. **b**, **c** Genome-wide distribution of observed heterozygosity values (averaged over a sliding 9-SNP window) for the pooled genotypes of the three Australian Merino populations (**b**) and Spanish Churra (**c**). The horizontal lines indicate the bottom 0.5th percent thresholds of the heterozygosity distributions

ATP1B2), or dairy traits (*IFNG*, *ABCG2*, *TP53*, *SPP1*, and *DVL2*) (see Additional file 14: Table S11). We found that the five Merino-related CCR and the 15 Churra-related CCR overlapped, respectively, with 103 and 84 previously reported genetic QTL or associations with phenotypic traits (see Additional file 16: Table S12). For each defined CCR, the correspondence with previously reported selection sweeps is indicated in Table 2.

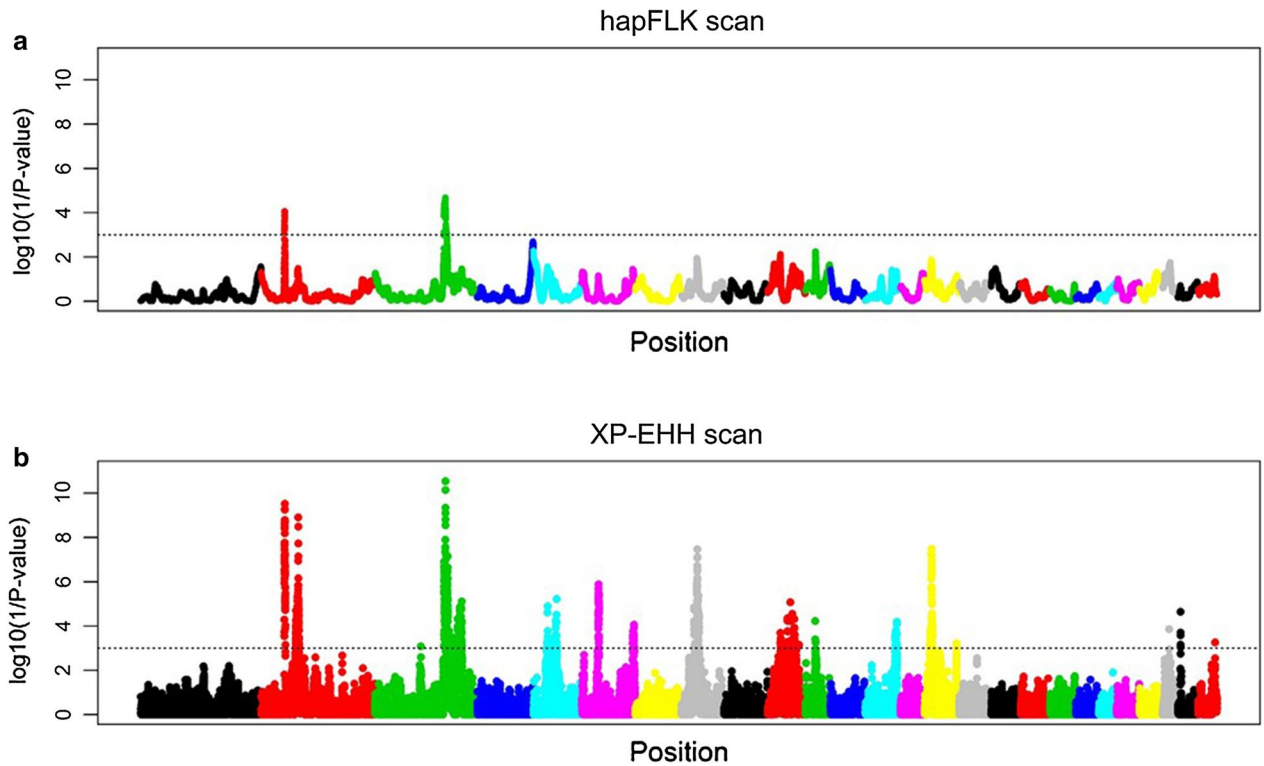


Fig. 3 Results of the selection sweep mapping analyses performed with the two haplotype-based methods used in this work, performed with the hapFLK (**a**) and the *rehh* (XP-EHH analysis) (**b**) software. Genome-wide distribution of the log (1/*P* value) obtained from each analysis are represented on the Y-axis. The horizontal lines represent the significance threshold considered ($P < 0.001$)

Selection mapping validation results

The results from the individual analyses performed for the samples included in the validation analysis are in Additional file 17: Table S13 and graphically represented in Additional file 18: Figures S5 and S6. The genetic differentiation analysis identified 39 candidate selection sweeps, whereas the scans looking for regions of reduced observed heterozygosity identified 97 and 68 candidate selection sweeps for Churra and Spanish Merino, respectively. The *hapflk* and the XP-EHH analyses identified, respectively, six and 76 significant genomic regions as potential selection signatures. Based on the selection signals identified in these individual analyses, and applying the same overlapping requirements than in the core analyses, we defined 18 CCR (labelled as CCR101 to CCR118, as shown in Additional file 19: Table S14). The correspondence between these CCR and those identified in the core analyses are also indicated in Table 3. In summary, seven of the Spanish Merino-Churra CCR were directly related with six of the CCR identified between Australian Merino and Churra breeds (those highlighted in blue), although some others were close to a previously identified CCR (e.g. CCR109 and CCR110 could be considered also related to CCR13). The core CCR that were clearly validated by this secondary analysis were CCR1, CCR3, CCR4 and CCR13 for Churra and CCR12 and CCR28 for Australian Merino. These validated CCR were, in general, those showing the most extreme XP-EHH estimates in the core analyses (e.g. all of them had an absolute XP-EHH estimate higher than 4.80, with the exception of CCR18, located on OAR25, which had an estimate of − 4.233).

Identification and annotation of divergent allelic variants in the identified CCR based on the analysis of whole-genome sequencing data

Results of the variant calling analysis

The maximum length of the reads obtained through the sequencing process was 100 bp. The whole-genome sequence datasets showed an average number of raw reads per sample of 318,377,494 paired reads. We obtained an average of 296,228,613 reads per sample that passed the quality control process. Per sample, the number of reads aligned to the reference genome varied between 112,607,669 and 513,347,097, with an average of 293,341,790 and an average of 2% unmapped reads per sample. The number of duplicates per sample

Table 1 Convergence regions identified in this study based on the overlapping of the results of the four mapping analyses performed to identify selection sweeps between Churra and Australian Merino breeds

CCR[a]	SS[b]	Chr[c]	CCR flanking markers	Start position (bp)	End position (bp)	XP-EHH value[d]
1	*XPEHH-SS1*	*2*	*OAR2_55248792.1- OAR2_57832237.1*	*51658967*	*53837176*	*6.297*
	F_{ST}-SS4	2	*OAR2_55493630_X.1- OAR2_57596413.1*	51898098	53597080	
	Churra-ObsHtz-SS5	2	*OAR2_55493630_X.1- OAR2_56828090.1*	51898098	52997998	
	hapflk-SS1	*2*	*OAR2_55493630_X.1- OAR2_56768579.1*	*51898098*	*52938537*	
	Churra-ObsHtz-SS6	2	*s18609.1- s53985.1*	53366034	53670410	
2	*XPEHH-SS7*	*2*	*OAR2_84010413.1- OAR2_84382185.1*	*78854385*	*79189919*	*4.571*
	F_{ST}-SS5	2	*OAR2_84182215.1- OAR2_84382185.1*	79017511	79189919	
3	*XPEHH-SS13*	*3*	*s59799.1- OAR3_162871753.1*	*151088496*	*152334140*	*5.232*
	F_{ST}-SS6	3	*OAR3_161831413.1- OAR3_162231144.1*	151512221	151778900	
	F_{ST}-SS7	3	*OAR3_162782289.1- OAR3_162794870.1*	152215311	152227684	
4	*hapflk-SS4*	*3*	*OAR3_163071695_X.1- OAR3_164185125.1*	*152544998*	*153519437*	
	XPEHH-SS14	*3*	*s59746.1- OAR3_164185125.1*	*152644200*	*153519437*	*6.651*
	F_{ST}-SS8	3	*OAR3_163342940.1- OAR3_163641518.1*	152795421	153090551	
	F_{ST}-SS9	3	*OAR3_164115875.1-OAR3_164185125.1*	153459890	153519437	
5	*XPEHH-SS16*	*3*	*s26177.1- OAR3_165200988.1*	*154006814*	*154402834*	*4.324*
	F_{ST}-SS10	3	*OAR3_164788310.1-OAR3_165324739.1*	154069702	154522600	
6	*XPEHH-SS17*	*3*	*OAR3_165450843.1-OAR3_169414477.1*	*154638280*	*158339544*	*5.409*
	F_{ST}-SS11	3	*OAR3_166034748.1- OAR3_166122747.1*	155167107	155252399	
7	*XPEHH-SS24*	*3*	*s07782.1- s67950.1*	*179815920*	*180128893*	*4.066*
	Churra-ObsHtz-SS23	3	*OAR3_193567675.1*	179832455		
8	Churra-ObsHtz-SS24	3	*OAR3_196791000.1- OAR3_196913312.1*	182778735	182916410	
	XPEHH-SS26	*3*	*OAR3_196880003.1- OAR3_196904777.1*	*182867529*	*182900674*	*3.373*
9	Churra-ObsHtz-SS25	3	*s67036.1- OAR3_197402139.1*	183347210	183368930	
	Merino-ObsHtz-SS22	3	*OAR3_197402139.1*	183368930		
	XPEHH-SS27	*3*	*OAR3_197402139.1- OAR3_197466728.1*	*183368930*	*183429797*	*4.061*
10	*XPEHH-SS31*	*3*	*OAR3_201886269.1- OAR3_202943170.1*	*187634152*	*188481721*	*4.323*
	F_{ST}-SS15	3	*OAR3_202741875.1*	188276666		
11	Merino-ObsHtz-SS35	6	*s73850.1-OAR6_40855809.1*	36461468	36655091	
	XPEHH-SS43	*6*	*s20660.1- s32980.1*	*36626596*	*36914376*	***− 4.211***
12	F_{ST}-SS24	6	*s17946.1*	37164263		
	XPEHH-SS44	*6*	*s17946.1- OAR14_57922732.1*	*37164263*	*38580198*	***− 4.837***
	Merino-ObsHtz-SS36	6	*OAR6_42247197.1*	37987281		
	Merino-ObsHtz-SS37	6	*OAR6_42484920_X.1*	38214088		
	Merino-ObsHtz-SS38	6	*OAR6_42743614.1- OAR6_42834740.1*	38417881	38481174	
13	*XPEHH-SS55*	*8*	*s50528.1- OAR8_36294417_X.1*	*32778561*	*33477406*	*4.846*
	Churra-ObsHtz-SS37	8	*OAR8_35694056.1- OAR8_35827974.1*	32849509	32979538	
14	*XPEHH-SS60*	*8*	*OAR8_39847976.1-s27049.1*	*37075040*	*37422641*	*4.194*
	Churra-ObsHtz-SS38	8	*OAR8_39977285.1- OAR8_40079017.1*	37211967	37313171	
15	*XPEHH-SS62*	*10*	*OAR10_29381795.1- OAR10_29448537.1*	*29344224*	*29415140*	*3.716*
	F_{ST}-SS29	10	*OAR10_29389966_X.1-OAR10_29737372.1*	29353089	29713193	
	Churra-ObsHtz-SS42	10	*OAR10_29511510.1-OAR10_29722772.1*	29476678	29688513	
16	Merino-ObsHtz-SS52	11	*OAR11_27752920.1- OAR11_28473036.1*	26512466	26939891	
	XPEHH-SS77	*11*	*s56248.1-s31301.1*	*26571629*	*26623188*	***− 3.459***

Table 1 continued

CCR[a]	SS[b]	Chr[c]	CCR flanking markers	Start position (bp)	End position (bp)	XP-EHH value[d]
17	Merino-ObsHtz-SS70	15	*s19862.1- s00941.1*	74618189	74636302	
	XPEHH-SS95	15	*s19862.1- s00941.1*	74618189	74636302	− **3.429**
18	*Merino-ObsHtz-SS95*	*25*	*s30024.1-s67158.1*	*7356301*	*7727709*	
	F_{ST}-SS49	25	*s31858.1-s44881.1*	7599609	7608913	
	XPEHH-SS97	*25*	*s44881.1- s74537.1*	*7608913*	*7821104*	− ***4.234***

After defining the selection signals identified by the different selection sweep mapping methods considered in our study, i.e. differentiation analysis (F_{ST}-SS), identification of regions of reduced heterozygosity (ObsHtz-SS) and haplotype-based selection mapping methods hapFLK and XEHPP analyses (hapFLK-SS) and XEHPP-SS), the corresponding intervals were compared and Convergence Candidate regions (CCR) were defined when at least one haplotype-based method showed coincidence with any of the two other analyses performed

[a] Convergence candidate regions defined based on the convergence of selection signals identified in this study

[b] Selection signals identified by the four analysis methods used in this study: the methods based on the estimation of F_{ST} and observed heterozygosity (ObsHtz) and the two methods based on haplotype analysis (hapFLK and XPEHH). Note that the signals identified by the haplotype-based methods are indicated in italics. It was necessary that at least overlapping of one significant haplotype-based SS (identified by the hapFLK or the XPEHH analyses) and one SS identified by any of the two other methods (FST or ObsHtz-based analyses) to label a region as a CCR

[c] Chromosome

[d] For the SS identified with the XP-EHH test, the most extreme XP-EHH estimate is provided. Note that positive and negative (negative highlighted in bold font) estimates indicate selection in the Churra and Merino populations, respectively

ranged from 4,818,140 to 43,064,209, with an average of 16,747,357. After alignment, focusing on the 18 target CCR intervals, the individual analyses with GATK and Samtools identified, initially, 194,413 and 196,128 genetic variants respectively (175,317 and 174,278, respectively, after applying the Snpshift filters). The intersection between the variants identified by the two different software programs showed 142,400 variants commonly identified for the 28 sheep genomes analyzed. All these variants, which included 139,745 biallelic SNPs and 2655 indels, were considered for further analyses.

Identification of the divergent SNPs in the CCR regions

All 28 samples and 139,244 variants, including SNPs and indels, passed the genotype QC filtering steps and were subsequently investigated by association analysis to compare Australian Merino and Churra breeds. Among these variants, 25,774 do not have an associated *rs* number. The results of this analysis for the tested SNPs are represented graphically in Fig. 3, where the X-axis shows the log (1/*P* value). Following a Bonferroni correction for the number of variants analyzed, 1291 variants (1282 SNPs and 9 indels) exceeded the 5% experiment-wise significance threshold ($P < 0.05/139{,}244 = P$ value < 0.000000359; log (1/*P* value) = 6.44). The distribution of these divergent variants over the considered chromosomes was as follows: 216 variants on OAR2, 117 on OAR3, 593 on OAR6, 316 on OAR8, 17 on OAR10, 2 on OAR11, 30 on OAR25, and none on OAR15 (Fig. 4). Considering the level of difference in allele frequencies (D) between the two breeds analyzed, which are in Additional file 20: Table S15, differences higher than 0.7 were shown by 79 variants (78 SNPs and 1 Indel) with a high frequency in Churra. Among them, the most extreme value of divergent allele frequency (D = − 0.8) were one SNP on OAR3 (*rs408539160*) and seven on OAR10 located in the interval 29.380–29.499 Mb within the *EEF1A1* gene (*ENSOARG00000011616*), with the exception of *rs421531355*, which is an intronic variant of the *RXFP2* gene. For the variants with a high frequency in Merino and low in Churra, 943 showed D values higher than 0.7; the most extreme of these D values were found for variants located on OAR3, 6 and 8 (see Additional file 20: Table S15).

The results of the functional annotation analysis showed that 1291 divergent variants included 257 intragenic variants causing 296 annotation variants, distributed across 31 protein coding genes, two pseudogenes (one of them identified as orthologous of bovine *TPI1*), one rRNA (*5S_rRNA*) and two snRNA (see Additional file 21: Table S16). Note that all the significant variants identified in the studied region of OAR25 were located in intergenic regions. The genetic variations included in protein coding genes resulted in 275 functional annotation variants classified as four missense variants, four synonymous variants, 199 intronic variants, 38 upstream gene variants, 29 downstream gene variants, and one variant in a 3′ UTR region.

Focusing on the variants located in exons of the above mentioned genes (see Additional file 21: Table S16), all of which were SNPs, we found the following: (1) on OAR2, a synonymous variant in the *PHF24* gene and a missense mutation in the *NPR2* gene, (2) on OAR6, three synonymous variants, in *PDK2*, *FAM184B* and *ENSOARG00000004249* (orthologous to human *LCORL*

Table 2 Correspondence of the 18 convergence candidate regions (CCR) identified as putative selection signals for Churra and Australian Merino sheep populations with previously reported signatures of selection

Present study		Other studies		
Region	**Genomic interval (Mb)**	**Correspondence with other studies Chr: peak marker/interval (Mb)**	**Putative candidate genes according to other studies**	**Population (target trait)**
CCR1	Chr2: 51.659–53.837	OAR2: 52.266–52.454		Zel-Lori Bakhtiri and HapMap dataset [25] (fat deposition)
		OAR2: 52.40 (peak SNP)		HapMap dataset [21]
		OAR2: 51.41–53.44		HapMap dataset [22]
		OAR2: 51.72–51.95		HapMap dataset [28] (climate adaptation)
		OAR2: 51.200–52.100; 52.100–52.900; 53.60–54.5800		Duolang sheep [27] (ecoregion adaptation)
CCR2	Chr2: 78.854–79.190			
CCR3	Chr3: 151.088–152.393	OAR3: 150.5–154.2	*HMGA2, WIF1*	Spanish breeds [30]
		OAR3: 151.42–156.93	*HMGA2*	HapMap dataset [22]
		OAR3: 152.68–154.679		HapMap dataset [19] (dairy specialization)
CCR4	Chr3: 152.545–153.519	OAR3: 150.5–154.2	*HMGA2, WIF1*	Spanish breeds [30]
		OAR3: 152.68–154.679		HapMap dataset [19] (dairy specialization)
CCR5	Chr3: 154.007–154.523	OAR3: 154.213 (peak SNP)	*HMGA2, MSRB3, LEMD3*	HapMap dataset [21]
		OAR3: 154.79–154.93		HapMap dataset [22]
		OAR3: 151.42–156.93		HapMap dataset [22]
		OAR3: 150.5–154.2		Spanish breeds [30]
		OAR3: 152.68–154.679		HapMap dataset [19] (dairy specialization)
CCR6	Chr3: 154.638–158.339	OAR3: 154.79–154.93		HapMap dataset [22]
CCR7	Chr3: 179.816–180.129			
CCR8	Chr3: 182.779–182.916	OAR3: 182.00–184.00		Duolang sheep [27] (ecoregion adaptation)
CCR9	Chr3: 183.347–183.430	OAR3: 182.00–184.00		Duolang sheep [27] (ecoregion adaptation)
CCR10	Chr3: 187.634–188.482			
CCR11	Chr6: 36.461–36.914	OAR6: 36.073 (peak SNP)	*ABCG2, NCAPG, PDK2*	HapMap dataset [21]
		OAR6: 34.71–39.12		HapMap dataset [22]
		OAR6: 36.63–36.8		HapMap dataset [28] (climate adaptation)
		OAR6: 36.200–36.500		Duolang sheep [27] (ecoregion adaptation
		OAR6: 30.367–41.863		HapMap dataset [19] (dairy specialization)
CCR12	Chr6: 37.164–38.580	OAR6: 34.71–39.12		
		OAR6: 37.2–38.0		HapMap dataset [22]
		OAR6: 37.40–37.60		HapMap dataset [24]
			LCORL, NCAPG	Small-tailed Han sheep [27] (ecoregion adaptation)
		OAR6: 30.367–41.863		HapMap dataset [19] (dairy specialization)
CCR13	Chr8: 32.779–33.477	OAR8: 32.159 (Peak SNP)	*BVES*	HapMap dataset [21]
CCR14	Chr8: 37.075–37.423			
CCR15	Chr10: 29.344–29.713	OAR10: 29.476 (peak SNP)	*RXFP2*	HapMap dataset [21]
		OAR10: 29.1–29.3		Spanish breeds [30]
		OAR10: 28.50–30.50		HapMap dataset [22]

Table 2 continued

Present study		Other studies		
Region	**Genomic interval (Mb)**	**Correspondence with other studies Chr: peak marker/interval (Mb)**	**Putative candidate genes according to other studies**	**Population (target trait)**
		OAR10: 28.71–29.00		HapMap dataset [28] (climate adaptation)
				HapMap dataset [24]
				HapMap dataset [24]
		OAR10: 27.1–31.2		Small-tailed Han sheep [27] (ecoregion adaptation
		OAR10: 29.1–31.9	*RXFP2, B3GALTL*	Duolang sheep [27] (ecoregion adaptation)
		OAR10: 29.40–29.700		
		OAR10: 29.50–29.400		
CCR16	Chr11: 26.512–26.940	OAR11: 24.18–38.74		HapMap dataset [22]
		OAR11: 26.8–29.9		Barki sheep versus temperate breeds (hot arid environment) [29]
				Small-tailed Han sheep [27] (ecoregion adaptation)
CCR17	Chr15: 74.618–74.636	OAR15: 72.774–74.55		HapMap dataset [19] (dairy specialization)
CCR18	Chr25: 7.356–7.821	OAR25: 7.517 (peak SNP)		HapMap dataset [21]
		OAR25: 7.400–7.600		Duolang sheep [27] (ecoregion adaptation)

by *BLASTN*), and three missense variants, in *NCAPG* and the inferred *LCORL* gene.

A full characterization of the identified missense variants is in Table 4. This is based on the eVEP annotation for the *NPR2* and *NCPAG* genes. For a proper assessment of the effects of the *LCORL* missense mutations, we aligned the interval including the two mutations identified in this gene against the most updated version of the sheep reference genome Oar_v4.0 [76] using a BLASTN search. Then, we identified the effect of the two *LCORL* missense mutations in the mRNA gene sequence (XM_015096407.1) and the corresponding protein sequence (XP_014951893.1; ligand-dependent nuclear receptor corepressor-like protein isoform X1). The prediction of the functional impact of the amino-acid changes of these mutations with the SIFT software [75] considered the mutations in the *NPR2* and *LCORL* genes as "tolerated", whereas the missense mutation located in the *NCAPG* gene, *NCAPG_Ser585Phe*, was classified as "deleterious" (score = 0.0) (Table 4).

Discussion

Based on the increasing affordable cost of next-generation sequencing technologies [77], the information derived from whole-genome resequencing offers increased detection power and a higher resolution to identify the genetic variants that underlie variation in traits of economic interest or linked to selection events in livestock populations [78–80]. In this work, we exploited the information from WGSeq datasets as a high-resolution step to investigate regions that were previously identified through the analysis of medium-density SNP panels in a representative sample of the population(s).

Identification of selection sweeps in Australian Merino and Churra sheep breeds

The putative selection sweep regions (referred herein as CCR) were determined by comparing 50 K-chip genotypes of three Australian Merino strains that are highly specialized for wool production with genotypes of the related, coarse-wool Spanish Churra dairy breed. In agreement with other authors [21, 22], our population structure analysis supported the use of these two groups of samples as appropriate for mapping selection sweeps because of their close phylogenic relationship but divergent phenotypic characteristics (see Additional file 1: Table S1). The contrasting features of the two breeds [e.g. white and fine wool, growth/carcass production, parasite resistance selection of Australian Merino; coarse wool, milk production/composition, dairy udder/body conformation, and characteristic black patches in specific body regions of Churra; (see Additional file 1: Table S1)] may help to identify the phenotypic targets of putative selection sweeps. Considering the possibility

Table 3 Correspondence between the convergence candidate regions (CCR) identified in the core analyses between Australian Merino and Churra breeds (labeled as CCR1 to CCR18), with the CCR identified in the validation analyses performed by contrasting a small dataset of Spanish Merino and Churra sheep genotypes (labeled as CCR101 to CCR118)

CCR AustralianMerino-Churra			CCR SpanishMerino-Churra		
Region	**Genomic region**	**Most extreme XPEHH value[a]**	**Region**	**Genomic region**	**Most extreme XPEHH value[a]**
CCR1	Chr2: 51.659–53.837	6.297	CCR101	Chr2: 51.530–53.798	4.282
CCR2	Chr2: 78.854–79.190	4.571			
CCR3	Chr3: 151.088–152.393	5.232	CCR102	Chr3: 151.433–152.055	3.648
CCR4	Chr3: 152.545–153.519	6.651	CCR103	Chr3: 152.855–152.861	3.560
CCR5	Chr3: 154.007–154.523	4.324			
CCR6	Chr3: 154.638–158.339	5.409			
CCR7	Chr3: 179.816–180.129	4.066			
CCR8	Chr3: 182.779–182.916	3.373			
CCR9	Chr3: 183.347–183.430	4.061			
CCR10	Chr3: 187.634–188.482	4.323			
			CCR104	Chr4: 30.499–30.929	− **4.131**
CCR11	Chr6: 36.461–36.914	− **4.211**			
CCR12	Chr6: 37.164–38.580	− **4.837**	CCR105	Chr6: 38.181–38.255	− **3.666**
			CCR106	Chr6: 38.429–38.617	− **4.256**
			CCR107	Chr8: 31.613–31.699	
CCR13	Chr8: 32.779–33.477	4.846	CCR108	Chr8: 32.364–32.597	
			CCR109	Chr8: 33.676–34.622	
			CCR110	Chr8: 34.791–35.740	
CCR14	Chr8: 37.075–37.423	4.194			
			CCR111	Chr8: 51.730–52.676	− **5.015**
			CCR112	Chr8: 52.997–54.352	− **4.599**
			CCR113	Chr8: 59.193–60.187	− **6.377**
CCR15	Chr10: 29.344–29.713	3.716			
			CCR114	Chr10: 51.490–52.154	− **4.592**
			CCR115	Chr10: 52.389–52.670	− **3.590**
CCR16	Chr11: 26.512–26.940	− **3.458**			
			CCR116	Chr15_ 37.553–37.776	4.543
			CCR117	Chr15: 38.783–38.943	3.734
CCR17	Chr15: 74.618–74.636	− **3.429**			
CCR18	Chr25: 7.356–7.821	− **4.234**	CCR118	Chr25: 7.356–7.970	− **3.361**

[a] For the CCR including a selection signal identified by the XP-EHH test, the most extreme XP-EHH estimate is provided. Positive and negative (highlighted in bold font) XP-EHH estimates indicate selection in the Churra and Merino populations, respectively

that the geographical isolation and distance between the two studied populations could be a confounding effect with respect of the signals evidenced we have, in addition, performed a validation analysis by contrasting an available dataset of Spanish Merino genotypes with the Churra sheep breed.

In addition, the high genetic diversity reported for the contrasting breeds [30, 81, 82] should be taken into account, which fits well with the known history of these breeds. In particular, the Australian Merinos have been reported as the most diverse sheep populations [81, 83, 84] since the foundation of this population involves contributions from different European, Asian and African breeds and, therefore, Australian Merino are a combination of strains of sheep rather than a single, ancient, homogenous breed. The Australian Merino lines considered in this study have some of the highest estimates for effective population size at 50 generations ago (average of $Ne = 868$; assuming four years per generation) [21]. The first historical references about Churra sheep date from the Middle Ages, approximately 800 years ago [85]. This breed shows a large influence from ovine populations brought in the Iberia Peninsula by the Celts [86]. Compared with other breeds, the estimated *Ne* at 50 generations ago for Churra is intermediate ($Ne = 600$) [21], and a steady decrease of this value until the start of the

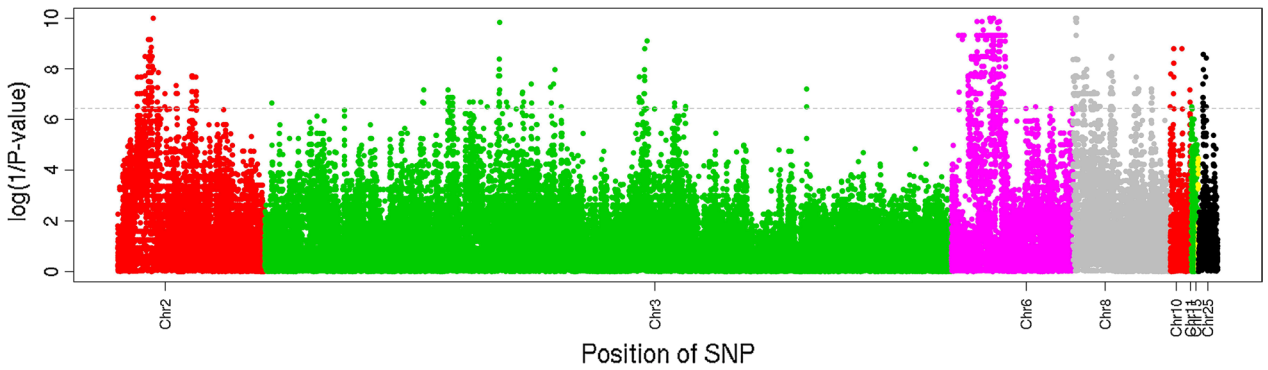


Fig. 4 Results of the association analysis performed for the 135,061 SNPs from the processing of 28 whole-genome sequencing samples of Churra and Australian Merino sheep breeds with the aim of identifying the markers with the most divergent allele frequencies between the two breeds compared. Genome-wide distribution of the log (1/*P* value) obtained from the association analysis with the breed identity are represented on the Y-axis. The horizontal line represents the significance threshold considered after a Bonferroni correction for multiple testing ($P < 0.05/139{,}244 = P$ value < 0.000000359; log (1/*P* value) = 6.44)

Table 4 Characterization of the three missense mutations identified in this study

Features	Missense mutations identified as divergent variants based on the analysis of whole-genome sequence datasets from Churra and Australian Merino samples			
SNP position (Oar_v3.1)	52,429,848	37,308,727	37,3557,21	37,356,400
Chromosome	2	6	6	6
dbSNP_ID	*rs160159505*	*rs159958168*	*rs419074913*	*rs159958380*
Gene	*NPR2*	*NCAPG*	*LCORL*[a]	*LCORL*[a]
Ref. (Texel Oar_v3.1) → Alt[b]	T → C	C → T	T → A	A → T
Position in CDS	c.2540	c.1754	c.4321[c]	c.3642[c]
Base pair substitution in CDS	T → C	C → T	A → T	T → A
Breed (mutant allele)[d]	Merino	Merino	Churra	Merino
Codon change	cAc → cGc	TCC → TTC	ATA → TTA	GAT → GAA
Amino acid change	Histitine (H) → Arginine (R)	Serine (S) → Phenilalanine (F)	Isoleucine (I) → Leucine (L)	Aspartate (D) → Glutamate (E)
Protein change	*NPR2_His847Arg*	*NCAPG_Ser585Phe*	*LCORL_Ile1441Leu*	*LCORL_Asp1214Glu*
Functional impact (ensemblVEP_Oarv3.1)	Moderate	Moderate	Moderate	Moderate
Functional impact (Polyphen-2)	Benign	Benign (score = 0.252; sensitivity: 0.91; specificity: 0.88)	Benign	Benign
Functional impact (SIFT_Oarv3.1)	Tolerated	Deleterious	Tolerated (low confidence)	Tolerated
Properties of wild aminoacid	Moderate hydropathy, charge "+"	Hydrophilic, polar, no charge	Hydrophobic, no charge	Hydrophilic, charge "−"
Properties of mutant aminoacid	Hydrophilic, charge "+"	Hydrophobic, apolar, no charge	Hydrophobic, no charge	Hydrophilic,, charge "−"
Churra genotypes	TT (15)	CC (14), CT (1)	AT (1), TT (14)	AA (14), AT (1)
Australian Merino genotypes	CC (9), TC (4)	TT (9), TC (3), CC (1)	AA (9), AT (3), TT (1)	TT (9), TA (3), AA (1)

[a] Mutation initially annotated within the *ENSOARG00000004249* novel gene (Oar_3.1). BLASTN analyses showed correspondence with the human *LCORL* gene and the ovine *LCOR* according to the most recent version of the sheep genome (Oar_v4.0)

[b] Ref. (Texel Oar_v3.1) → Alt: Reference and alternative alleles, respectively, identified in the analysis of the whole-genome sequence datasets

[c] Position of the SNP in the coding sequence based on the alignment of the sequence harboring the mutation to the annotation of the *LCORL* gene in the most recent version of the sheep genome (Oar_v4.0): NCBI Reference sequences: XM_015096407.1, XP_014951893.1 (ligand-dependent nuclear receptor corepressor-like protein isoform X1)

[d] Breed with the highest frequency for the mutant allele (regarding the wild protein sequence). Note that for SNP *rs419074913*, the Texel sheep of the reference genome harbors the mutant allele according the CDS and protein sequence

breeding program in 1986 suggests the absence of severe bottlenecks or other extreme demographic events [82, 87]. Selection sweep mapping methods that are based on haplotypes are known to be highly robust towards perturbations of the demographic model when compared with methods based on population subdivision or allele frequencies [9]. Hence, in our work, the labelling requirement of overlap between at least one haplotype-based method and the F_{ST}/ObsHtz methods may be seen as helping to limit the number of false positive results due to neutral (e.g. demographic) processes.

Comparing the results of the different analyses, we found discrepancies in the number of signals detected. For example, the hapfFLK analysis, which accounts for the relationship between populations and the LD pattern (haplotype diversity), only detected seven candidate signatures of selection, compared with 25 detected by the genetic differentiation approach, 96 and 77 regions of signatures of selection of reduced heterozygosity in Australian Merino and Churra, respectively, and 98 regions identified by the XP-EHH analysis. These differences can be explained by the fact that the different statistics used in selection sweep mapping do not capture the same patterns in the data, as previously pointed out by many authors [23, 88, 89]. The small number of significant regions detected with hapFLK compared with the other methods agrees with other studies that have analyzed the same datasets using hapFLK and other methods [30, 90]. As suggested by Manunza et al. [30], the multipoint linkage LD model implemented by hapFLK may explain the higher stringency of this method when compared with methods that do not consider haplotype structure (such as F_{ST} or ObsHtz). In contrast, XP-EHH analysis, which also makes use of haplotype information, detected a large number of selection sweeps (98) supporting several of the signatures of selection detected by the ObsHtz and F_{ST} analyses. Hence, the underlying model of this cross-population analysis for which the basic idea is to test if each site is homozygous in one population and polymorphic in the other population appears to fit efficiently for the Churra versus Australian Merino contrast undertaken in our study. Also the ObsHtz analysis identified a substantial number of candidate regions (96 Merino/77 Churra), which agrees with the fact that signatures of selection that are based on a reduction in genetic diversity persist for a longer period of time than signals based on haplotype structure and thus the former can detect older signatures of selection [88, 89]. The intermediate number of candidate regions detected based on population differentiation agrees with the intermediate position suggested for these methods by Sabeti et al. [88] regarding the time scale persistency for the different selection sweep mapping methods.

Similar discrepancies in the number of candidate selection sweeps identified by the different methods in the Spanish Merino and Churra analyses (39 with F_{ST}, 97 for ObsHtz-Churra, 68 for ObsHtz-SpanishMerino, 7 for hapflk and 76 with XP-EHH) (see Additional file 17: Table S13) prove that the differences observed in the number of signatures of selection in the Australian Merino versus Churra analyses were not due to the confounding effect of genetic drift and geographical isolation of the breeds analyzed. The fact that the six clearly confirmed CCR were those that had the most extreme XP-EHH estimates in the Australian Merino-Churra core analyses appears to suggest that the lack of confirmation of some other regions (e.g. CCR2 on OAR2, CCR16 on OAR11, CCR17 on OAR15) can be related to the lower power of detection of the validation analyses due to the limited number of samples analyzed (20 Spanish Merino and 20 Churra samples). Overall, we think that the validation strategy presented here supports the validity of the selection sweeps identified when contrasting Australian Merino and Churra sheep.

In addition, the combination of the four methods used in this work provides a comprehensive picture of the different types of selection sweeps present in the genomes of Churra and Australian Merino sheep. The requirement of overlap between regions identified by different methods to define a selection sweep increases the reliability of the 18 CCR reported here to result from genuine selection events. This is supported by the high level of positional correspondence of these regions with selection sweeps previously reported in sheep (Table 2).

Exploration of convergence candidate regions through WGSeq

In this study, we exploited WGSeq as a secondary step to provide a detailed study of the genetic variation within the regions previously identified as potential signatures of selection. As a technical issue and considering that paired-end sequencing is preferred over single-end sequencing, since it allows improved identification of duplicated reads and a better estimation of the fragment size distribution [91], it is worth clarifying that the workflow used in this study was based on Trimmomatic, which provides a flexible method to keep, in the analysis, the reads for which their paired read is filtered during the quality control filtering [66]. Also the reads for which their paired read was unmapped were included in the later variant calling analysis. To assess the impact that the use of singletons could have on the results, we repeated the variant calling analysis without considering singletons, and found a concordance level of 99.15%. This observation suggests that, for the variant calling analysis workflow applied in this study, which considers

the common variants identified by GATK and Samtools, the use of singletons does not have a negative impact on the quality of the variant calling analysis; however, we can also consider than using a simpler workflow that eliminates these singletons may be an efficient strategy for future studies.

Merino-related convergence candidate regions

Among the genomic regions showing positive selection in Merino, those located on OAR6, CCR11 (36.641–36.914 Mb) and CCR12 (37.164–38.580 Mb) showed substantial overlap with previously reported selection sweeps (Table 2) and QTL in sheep (see Additional file 16: Table S12). While most of the coincidences with selection sweeps reported in these regions are related to studies on the SheepHapMap dataset (Table 2), the study of Liu et al. [27] on adaptation to different ecoregions (regions where sheep are exposed to different climate, environment and feeding conditions) also identified the region OAR6: 36.200–36.500 Mb as a putative signature of selection in Duolang sheep. Furthermore, a large number of QTL/associations with production traits have been mapped within these two CCR in a population of Scottish Blackface lambs [92] (see Additional file 16: Table S12). Several of those effects were associated with carcass bone percentage and fat carcass traits, with some suggestion of muscle density effects. CCR11 and CCR12 also include QTL for growth traits [93, 94] and birth weight [95]. Most of these studies suggest *NCAPG* and *LCORL* as strong candidate genes for these effects, due to the reported associations of these loci with human stature [96] and body size in mammals [97–102]. In addition to these sheep studies, selection sweeps have been identified around the *NCAPG-LCORL* locus in dogs and pigs [49, 103].

In many cases, the intervals flanking the previously reported selection sweeps or the QTL reported in this OAR6 genomic region involve both CCR11 and CCR12. Considering the inherent inaccuracy of gene mapping, even when based on medium-density SNP arrays, and the large number of effects identified in that region, our approach to group individual signatures of selection based on the extent of LD of the studied breeds appears to identify two independent selection sweeps, which may help to differentiate the causal mutations that underlie the various QTL effects reported in this genomic region.

CCR11 contains the following annotated genes: *ABCG2*, *PKD2*, *SPP1*, *MEPE*, and *IBSP* (see Additional file 21: Table S16). Our high-resolution analysis based on sequence data showed that the CCR11 SNPs showing a significant association with breed identity between Churra and Australian Merino were located in three genes included in that interval: *PKD2*, *MEPE* and *IBSP*. Three intronic and one synonymous divergent variant were identified in the *PKD2* gene. In cattle, one SNP within this gene is significantly associated with hot carcass weight and intermuscular fat percentage [104]. The other divergent variants were located in non-coding regions of the *MEPE* and *IBSP* genes. *MEPE* is thought to play an inhibitory role in bone formation, and disruption of one of its alleles is known to increase bone mass in mouse [105]. No significant associations were identified with markers within *ABCG2* and *SPP1* which are functional candidate genes for milk production traits [106, 107].

CCR12 (OAR6: 37.164–38.580 Mb) includes four annotated genes, the major candidate genes *NCAPG* and *LCORL*, as well as *FAM184B*, which is highly expressed in skeletal muscle (http://www.proteinatlas.org/ENSG00000047662-FAM184B/tissue), and *DCAF16*. A considerable proportion of the intragenic variants (88/296) that show significant between-breed divergence were located within these four genes (see Additional file 21: Table S16), including two synonymous variants in *LCORL* and *FM184B* and three missense mutations in *NCAPG* and *LCORL*. Kühn et al. [108] suggested that the mechanism underlying the association between *NCAPG* and pre- and post-natal growth in several mammalian species may be related to the role of this gene in the modulation of growth and body tissue deposition by indirect effects on the nitric oxide (NO) pathway. In cattle, *NCAPG* is also suggested to be involved in early muscle development. [109]. An analysis in UniProt [110] shows the *NCAPG_Ser585Phe* amino acid substitution to affect the C-terminal, cysteine-rich domain of the protein, whereas a high level of across species conservation for this residue of the NCAPG protein (Fig. 5) was shown in a comparative analysis with Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). Further functional studies are needed to assess how this mutation may affect the protein function and possible effects on phenotype(s). The other gene for which divergent missense mutations were identified in this region is *LCORL*, which encodes a transcription factor that binds specific DNA elements and appears to function in spermatogenesis; polymorphisms in this gene have also been associated with skeletal frame size and human height, as previously discussed.

The results of the sequence analysis in our study support the hypothesis that the *NCAPG-LCORL* locus may be responsible for direct effects on sheep production traits, possibly including growth and/or carcass traits, which differ between Churra and Australian Merino. Studies that compared carcass characteristics of Churra and Spanish Merino at the same slaughter age showed that Merino has significantly higher hot carcass weight

Fig. 5 Multi-species alignment of the amino-acid sequence of the NCAPG protein across ruminant species, humans and mouse using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The serine amino acid affected by the *NCAPG_c.1754C > T* mutation shows a high conservation level in all species considered

and conformation score than Churra [111], and also higher carcass yield, total muscle and bone percentage [112]. Hence, the three missense mutations identified within the *NCAPG* and *LCORL* genes by our high-resolution study should be further studied to assess their potential role as causal mutations of the previously reported QTL effects within that genomic region (e.g. Matika et al. [92]). This is the first study that suggests specific mutations in this region that may influence production traits in sheep. As the single "deleterious" mutation identified in this region, the *NCAPG_Ser585Phe* protein variant should be considered as the top candidate to explain the CCR12 signature of selection or the potential selection sweep nucleotide (SSN). Analysis of this mutation in meat breeds where the mutation segregates will allow the verification of the true effect of this polymorphism on production traits. In order to obtain further information about the potential link between *NCAPG* and *LCORL* missense mutations and meat production traits, we extracted the genotypes for these three polymorphisms from the 453 samples analyzed in the "PRJEB14685" Project of the EVA repository (http://www.ebi.ac.uk/ena/data/view/PRJEB14685). As observed from these genotypes (see Additional file 22: Table S17), the Merino-related allele for the three mutations, is only present (in homozygous or heterozygous status) in "meat production" breeds.

Three other Merino-CCR regions, CCR16, CCR17 and CCR18, are located on OAR11, 15 and 25, respectively. CCR16 (OAR11) is the most interesting, from which six of the 42 annotated genes were highlighted by the survey performed against our database of candidate genes (see Additional file 15: Table S11). Among them, *TP53* and *DVL2*, are both candidates for wool production due

to their link to the hair follicle cycle [45]. However, the divergent intragenic variants identified in this region did not affect any of these wool-related candidate genes but included non-exonic variants within the *DLG4* and *ACADVL* genes. The *ACADVL* gene encodes the enzyme that catalyzes the first step of the mitochondrial fatty acid beta-oxidation pathway, which suggests that there may be differences in the fatty acid composition between Churra and Merino lamb meat.

A large proportion (7/9) of the previously reported QTL located within CCR18 (OAR25: 7.356–7.821 Mb) are wool-related QTL (see Additional file 16: Table S12). The study of Allain et al. [113], based on microsatellite markers, suggested the presence in this region of a locus with a major effect on fleece characteristics. These results were supported by another study based on the 50 K-chip in which genome-wise significant SNP associations for seven wool-related traits were reported in a nearby interval (OAR25: 6.1–8.2 Mb). Interestingly, a 2-kb insertion was identified at this location, which is a potential causal mutation for the absence of long, coarse hair in the birthcoat of the Romane breed [114]. This is a trait that is moderately genetically correlated with wool quality traits, with the woollier lambs showing a lower coefficient of variation, fewer fibers thicker than 30 μm and better wool quality compared with the other hairier lambs [115]. Within CCR18, 30 SNPs showed a significant difference in allele frequencies between Churra and Merino populations, although none of these SNPs were intragenic. Among the genes found in this region (see Additional file 13: Table S9), the novel gene *ENSOARG00000017989* shows correspondence with the bovine *EIF2S2* gene, which encodes the beta subunit of EIF-2 that functions in the early steps of protein synthesis. *EIF2S2* has been suggested as a novel candidate gene in relation to skin color in humans [116]. Because white wool color has been a major selection objective in Australian Merino, a possible link between the *EIF2S2* gene and the CCR18 Merino effect of the signature of selection should be investigated further.

Churra-related convergence candidate regions

With regard to the 15 Churra-associated CCR, nine overlapped with previously reported selection sweeps in sheep (Table 2). The region that showed correspondence with the largest number of studies (most based on SheepHapMap data) is CCR15 (OAR10: 29.334–29.713 Mb). This region includes the *RXFP2* gene, which is suggested to control the presence and size of horns in wild and domestic populations of sheep [21, 79, 117] and to be important for horn development in goats and cattle [118, 119]. *RXFP2* is a receptor for the relaxin and insulin-like factor 3 proteins and its effects on horn size and status appear to depend on its biochemical interactions with testosterone effects [120, 121]. This region was identified by XP-EHH and ObsHz analyses and overlapped with a signal from the F_{ST} analysis, which suggests positive selection in Churra (Table 1). Taking into account that selection for the polled phenotype has occurred in both Churra and all Australian Merino lines (not only the Australian Poll Merino), the detection of CCR15 as a Churra-related selection sweep region suggests that selection for the absence of horns is more recent (and thus more detectable) in Churra than in Merino. In fact, the high selection pressure for polledness in Churra started with the breeding program in 1985 [39], and prior to that time horned rams were preferred by Churra breeders (F. de la Fuente, personal communication). At present, about 90% of the Churra males are polled. There are practically no females with horns, and when horns are present they are rudimentary. For the OAR10 selection sweep, our high-resolution analysis identified 13 significant divergent intronic variants in the gene and one variant in the third intron of the *RXFP2* gene. A 1.8-kb insertion reported in the 3' UTR region of the ovine *RXFP2* gene, which includes two exons of the *EEF1A1* gene, has been suggested to explain the polled phenotype in some sheep breeds [122]. However, this insertion does not completely segregate with the horn status in breeds with a variable horn status in both sexes or with a sex-dependent horn status [123], which is the horn status in Churra sheep. An additional exploratory association analysis considering only the Churra and Australian Poll Merino SNPs located on OAR10 (5329 polymorphic SNPs) only identified one significant SNP, which was annotated as an intronic variant within the *EEF1A1* gene (at 29.380.801 bp; P value = 0.000006769). Additional research based on whole-genome sequencing data from breeds with a variable horn status may shed light on this complex phenotype.

Two of the other Churra-related CCR, CCR1 (OAR2: 51.659–53.837) and CCR4 (OAR3: 152.545–153.519), were the only CCR supported by the hapFLK analysis, which identified the smallest number of regions in our analyses. These two regions, which were also replicated when analyzing Spanish Merino and Churra genotypes, overlap with 11 and seven QTL, respectively, reported for a range of sheep production traits (see Additional file 16: Table S12). The CCR1 region on OAR2 was the only one detected by all four selection sweep mapping methods (Table 1) and overlaps with several previously reported signatures of selection (Table 2). The intragenic significant divergent variation identified in this region included a missense mutation in the *NPR2* gene, *rs160159505* (OAR2: 52,429,848 pb). This gene was previously suggested by other authors [21, 22] as a strong candidate

gene for ovine selection sweep effects reported in this region due to its major role in the regulation of skeletal growth regulation [124]. In humans, mutations in this gene are related to impaired skeletal growth [124, 125]. Because *rs160159505* was the mutation showing the highest significant association with breed identity within CCR1 (P value = 0.0000000001006) and although it was classified as "tolerated", further studies should assess the possible relationship of this SNP with growth- or size-related traits in sheep. CCR4, the other region identified by hapFLK, showed correspondence with a signature of selection that was previously reported in a study on ovine signatures of selection related to dairy specialization (also including the Churra breed) [19]. The *LALBA* (OAR3:137.390–137.392) gene, which was suggested to harbor a quantitative trait nucleotide (QTN) for QTL effects related to milk composition traits in Churra sheep [126], maps just outside the boundaries of this region. The divergent intragenic variants annotated within this region were in non-coding regions of the genes *HELB* and *IRAK3,* which were not highlighted by our candidate gene survey. The other Churra-related CCR are not discussed in detail, but we would like to mention that CCR3 (OAR3: 151.088–152.393 Mb) overlaps with three QTL for resistance to gastrointestinal infection (see Additional file 16: Table S12) and that our association analyses identified, within this interval, a single significant intergenic SNP (*rs421227322*) located in the upstream region containing three immune-related genes, *IL22*, *IL26* and *IFNG*. Also of possible relevance is the identification of three intronic variants within the well-defined region of CCR5 (OAR3:154.006–154.522 Mb), which show extreme allele frequencies between Churra and Merino within the genes *MSRB3* and *LEMD3*. Indeed, intronic variation within these two genes has been directly associated with a pleiotropic QTL reported in cattle for birth weight, calving ease direct, marbling and ribeye muscle area [127].

Conclusions

We have identified 18 putative selection sweeps by contrasting the Australian Merino and Spanish Churra sheep breeds. The phenotypes affected by these genetic effects may involve any trait for which selection was implemented in only one of the two compared breeds. The criteria used to define the selection candidate regions based on multiple mapping methods, together with a validation approach based on a comparison between Spanish Merino and Churra genotypes, support the validity of our mapping results and confirm the value of using selection mapping to detect genomic regions that influence the phenotypic variation of complex traits in livestock species. Our subsequent high-resolution study performed in the target CCR reveals promising candidate mutations to explain some of the identified selection events, including variants in the *RXFP2*, *NPR2*, *NCAPG* and *LCORL* genes, related to the presence of horns and skeletal growth. Further studies are necessary to confirm the possible direct effect of some of the mutations highlighted in this work on the phenotypic variation of traits of interest in sheep.

Additional files

Additional file 1. Phenotypic, production and reproductive traits of Spanish Churra and Australian Merino sheep breeds.

Additional file 2. Description of the population structure analyses performed with the 50K-chip genotypes of the samples considered in this study.

Additional file 3. Figure S1. Graphical representation of the principal component analysis (PCA) performed with Eigensoft for the final of 50K-Chip genotypes analysed in this study for 238 fine wool Merino [Australian Industry Merino (n = 88), Australian Merino (n = 50) and Australian Poll Merino (n = 98)] and 278 Spanish Churra individuals. **Figure S2.** Graphical representation of the results of the cross-validation approach performed with the Admixture software to determine the best K-value. **Figure S3.** Graphical representation of the proportion of membership of each of the analysed populations for K = 2 as obtained with the Admixture_v1.3 software.

Additional file 4. Description of the extent of the linkage disequilibrium and block structure based on the analysis of the 50K-chip genotypes of the samples considered in this study.

Additional file 5: Figure S4. Average linkage disequilibrium (LD) as a function of genomic distance between markers based on the Churra and Australian Merino 50K-Chip genotypes analyzed in this study. The LD values (y-axis), provided as D′ and r2, are plotted against inter-marker distance bins (x- axis). For each case, the total number of marker pairs were assigned according to their physical distance into 14 categories: < 10 Kb, 10-20 Kb, 20-40 Kb, 40-60 Kb, 60-100 Kb, 200-500 Kb, 0.5-1 Mb, 1-2 Mb, 2-5 Mb, 5-10 Mb, 10-20 Mb, 20-50 Mb or > 50 Mb.

Additional file 6: Table S2. List of candidate genes considered in relation to wool-related traits.

Additional file 7: Table S3. Whole-genome sequence datasets analyzed in this study.

Additional file 8: Table S4. Signatures of selection identified by the genetic differentiation analysis performed between the Australian Merino (Australian Industry Merino, Australian Merino and Australian Poll Merino) and the Churra populations analysed in this study.

Additional file 9: Table S5. Selection signals identified by the analysis of observed heterozygosity performed in the Australian Merino populations analysed in the present study.

Additional file 10: Table S6. Selection signals identified by the analysis of observed heterozygosity performed in the Churra population analysed in the present study.

Additional file 11: Table S7. Selection signals identified by the analysis performed with the hapFLK software (P-value < 0.001) for the Australian Merino and Churra samples analysed in the present study.

Additional file 12: Table S8. Signatures of selection identified by the cross-population extended haplotype homozygosity (XP-EHH) analysis performed between the Merino (Australian Industry Merino, Australian Merino and Australian Poll Merino) and the Churra populations analysed in this study (P-value < 0.001).

Additional file 13: Table S9. List of genes extracted from the five convergence candidate regions (CCR) identified in this study as positive selection genomic regions in the fine wool Australian Merino lines.

Additional file 14: Table S10. List of genes extracted from the five convergence candidate regions (CCR) identified in this study as positive selection genomic regions in the Spanish Churra breed.

Additional file 15: Table S11. List of positional candidate genes included in the identified convergence candidate regions (CCR) that were highlighted by our survey with a list of 1459 genes including candidate genes related to traits of interest in sheep.

Additional file 16: Table S12. Correspondence of the convergence candidate regions (CCR) identified in our study with previously published genetic effects (QTL and associations) with phenotypic traits of interest in sheep according to the AnimalQTLdb (http://www.animalgenome.org/cgi-bin/QTLdb/OA/index).

Additional file 17: Table S13. Results of the four selection sweep mapping analyses between Spanish Merino and Spanish Churra sheep as a validation procedure. The selection signals identified by genetic differentiation (F_{ST}), reduction of heterozygosity (ObsHtz), and haplotype-based methods (hapFLK andn XP-EHH) were labeled following the same criteria as for the core analyses between Australian Merino and Spanish Churra breeds.

Additional file 18. Figure S5. Graphical representation of the genetic differentiation analysis (a), and the analysis of reduced heterozygosity (b, c) when analysing the validation dataset "Spanish Merino [34] vs Spanish Churra". **Figure S6.** Graphical representation of the selection sweep mapping analyses performed with the two haplotype-based methods used in this work, performed with the hapFLK (a) and the rehh (XP-EHH analysis) (b) software, for the validation dataset considered in the present work (Spanish Merino [34] vs Spanish Churra sheep breeds).

Additional file 19: Table S14. Convergence candidate regions (CCR) of selection sweeps identified in the validation survey by contrasting genotypes of the Spanish Merino and Spanish Churra sheep breeds. The CCR were defined based on the overlapping between significant signatures of selection (SS) identified by the different individual analysis methods implemented in this study.

Additional file 20: Table S15. Allele frequencies for the genetic variants (SNPs and indels) identified from the analysis of 28 WGSeq datasets (15 Churras and 13 Australian Merino) within the 18 genomic regions identified as CCR in the present study.

Additional file 21: Table S16. Characterization of the intragenic SNP variations identified, through the processing of the WGSeq datasets considered, within each of the 18 convergence candidate regions (CCR) identified according to the annotation performed with the eVEP software (ensembl Variant Effect Predictor; for further information about the column field names see http://www.ensembl.org/info/docs/tools/vep/vep_formats.html).

Additional file 22: Table S17. Summary of the samples included in the "PRJEB14685" Project (high-quality variant calls from the Sheep genomes project - Run1) of the EVA repository (http://www.ebi.ac.uk/ena/data/view/PRJEB14685) carrying the Australian Merino-related allele for the three missense mutations identified in the present study within the candidate convergence region CCR12 (OAR6: 37164263-38580198 bp).

Authors' contributions

JJA and BGG conceived the study and designed the analysis design; BGG, PKC, CEB and ASV performed selection sweep mapping and bioinformatic analyses; PW and BGG developed selection mapping analysis scripts; BGG drafted the manuscript; JJ, PW and CEB supervised and edited the manuscript. All authors read and approved the final manuscript.

Author details

[1] Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain. [2] Fundación Centro Supercomputación de Castilla y León, Campus de Vegazana, León 24071, Spain. [3] Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK.

Acknowledgements

This work was supported by the AGL2015-66035-R project funded by the Spanish Ministry of Economy and Competitiveness (MINECO) and co-funded by the European Regional Development Fund. We thank the National Association of Spanish Churra Breeders for the close collaboration with our research group and the support for generating sequencing data of Churra genomes. B Gutiérrez-Gil is funded through the Spanish "Ramón y Cajal" Program (RYC-2012-10230) from MINECO. The support and availability to the computing facilities of the Foundation of Supercomputing Center of Castile and León (FCSCL) (http://www.fcsc.es) are greatly acknowledged. The ovine SNP50 K-chip HapMap dataset used in this work was provided by the International Sheep Genomics Consortium (ISGC) and obtained from http://www.sheephapmap.org in agreement with the ISGC Terms of Access. We are also grateful to the ISGC for the whole-genome sequencing datasets belonging to the project PRJNA160933 available at the SRA (https://www.ncbi.nlm.nih.gov/sra) that were analyzed in this study. In addition, we are grateful to the Australian Cooperative Research Centre for Sheep Industry Innovation for generating the whole-genome sequencing datasets included in project PRJNA325682 of the SRA and also included in this study.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 19 April 2016 Accepted: 19 October 2017
Published online: 07 November 2017

References

1. Maijala K. Genetic aspects of domestication, common breeds and their origin. In: Piper L, Ruvinsky A, editors. The genetics of sheep. Oxford: CABI; 1997. p. 539–64.
2. Larson G, Fuller DQ. The evolution of animal domestication. Annu Rev Ecol Evol Syst. 2014;45:115–36.
3. Clutton-Brock J. Domesticated animals from early times. London: Heinemann and British Museum (Natural History); 1981.
4. Fraser AF. Evolution of domesticated animals. London: Longman; 1985.
5. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 2007;89:391–403.
6. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics. 1989;123:887–99.
7. Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. Proc Biol Sci. 2011;278:3161–70.
8. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358–70.
9. Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010;464:587–91.
10. Wiener P, Pong-Wong R. A regression-based approach to selection mapping. J Hered. 2011;102:294–305.
11. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997;147:915–25.
12. Galtier N, Depaulis F, Barton NH. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics. 2000;155:981–7.
13. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116:153–9.
14. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007;8:857–68.
15. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010;20:R208–15.
16. Wilson BA, Petrov DA, Messer PW. Soft selective sweeps in complex demographic scenarios. Genetics. 2014;198:669–84.

17. Gutierrez-Gil B, Arranz JJ, Wiener P. An interpretive review of selective sweep studies in Bos taurus cattle populations: identification of unique and shared selection signals across breeds. Front Genet. 2015;6:167.
18. Wilkinson S, Lu ZH, Megens HJ, Archibald AL, Haley C, Jackson IJ, et al. Signatures of diversifying selection in European pig breeds. PLoS Genet. 2013;9:e1003453.
19. Gutierrez-Gil B, Arranz JJ, Pong-Wong R, Garcia-Gamez E, Kijas J, Wiener P. Application of selection mapping to identify genomic regions associated with dairy production in sheep. PLoS One. 2014;9:e94623.
20. Moon S, Kim TH, Lee KT, Kwak W, Lee T, Lee SW, et al. A genome-wide scan for signatures of directional selection in domesticated pigs. BMC Genomics. 2015;16:130.
21. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, et al. Genome-wide analysis of the World's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012;10:e1001258.
22. Fariello MI, Servin B, Tosser-Klopp G, Rupp R, Moreno C, International Sheep Genomics Consortium, et al. Selection signatures in worldwide sheep populations. PLoS One. 2014;9:e103813.
23. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193:929–41.
24. Kijas JW. Haplotype-based analysis of selective sweeps in sheep. Genome. 2014;57:433–7.
25. Moradi MH, Nejati-Javaremi A, Moradi-Shahrbabak M, Dodds KG, McEwan JC. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC Genet. 2012;13:10.
26. McRae KM, McEwan JC, Dodds KG, Gemmell NJ. Signatures of selection in sheep bred for resistance or susceptibility to gastrointestinal nematodes. BMC Genomics. 2014;15:637.
27. Liu Z, Ji Z, Wang G, Chao T, Hou L, Wang J. Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. BMC Genomics. 2016;17:863.
28. Lv FH, Agha S, Kantanen J, Colli L, Stucki S, Kijas JW, et al. Adaptations to climate-mediated selective pressures in sheep. Mol Biol Evol. 2014;31:3324–43.
29. Kim E-S, Elbeltagy AR, Aboul-Naga AM, Rischkowsky B, Sayre B, Mwacharo JM, et al. Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. Heredity (Edinb). 2016;116:255–64.
30. Manunza A, Cardoso TF, Noce A, Martínez A, Pons A, Bermejo LA, et al. Population structure of eleven Spanish ovine breeds and detection of selective sweeps with BayeScan and hapFLK. Sci Rep. 2016;6:27296.
31. Ciani E, Crepaldi P, Nicoloso L, Lasagna E, Sarti FM, Moioli B, et al. Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. Anim Genet. 2014;45:256–66.
32. Beynon SE, Slavov GT, Farré M, Sunduimijid B, Waddams K, Davies B, et al. Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. BMC Genet. 2015;16:65.
33. Diez-Tascon C, Littlejohn RP, Almeida PAR, Crawford AM. Genetic variation within the Merino sheep breed: analysis of closely related populations using microsatellites. Anim Genet. 2000;31:243–51.
34. Ciani E, Lasagna E, D'Andrea M, Alloggio I, Marroni F, Ceccobelli S, et al. Merino and Merino-derived sheep breeds: a genome-wide intercontinental study. Genet Sel Evol. 2015;47:64.
35. Lewis W, Balderstone S, Bowman J. Events that shaped Australia. London: New Holland Publishers; 2006.
36. Fogarty NM, Safari E, Taylor PJ, Murray W. Genetic parameters for meat quality and carcass traits and their correlation with wool traits in Australian Merino sheep. Aust J Agric Res. 2003;54:715–22.
37. Gardner GE, Williams A, Siddell J, Ball AJ, Mortimer S, Jacob RH, et al. Using Australian sheep breeding values to increase lean meat yield percentage. Anim Prod Sci. 2010;50:1098–106.
38. Raadsma HW, Gray GD, Woolaston RR. Breeding for disease resistance in Merino sheep in Australia. Rev Sci Tech. 1998;17:315–28.
39. de la Fuente LF, Fernández G, San Primitivo F. Breeding programme for the Spanish Churra sheep breed. Cahier Options Méditerranéennes. 1995;11:165–72.
40. Miguélez E, Zumalacárregui JM, Osorio MT, Figueira AC, Fonseca B, Mateo J. Quality traits of suckling-lamb meat covered by the protected geographical indication "Lechazo de Castilla y León" European quality label. Small Ruminant Res. 2008;77:65–70.
41. Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol:chloroform. CSH Protoc. 2006;2006:pii: pdb.prot4455.
42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
43. Sheep genome assembly v3.1. Available from: http://www.ensembl.org/Ovis_aries/Info/Index.
44. Patterson N, Price AL, Reich D. Population structure and Eigenanalysis. PLoS Genet. 2006;2:e190.
45. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
46. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 2002;12:1805–14.
47. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, et al. Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci USA. 2010;107:1160–5.
48. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7:e1002316.
49. Rubin CJ, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci USA. 2012;109:19529–36.
50. Stainton JJ, Haley CS, Charlesworth B, Kranis A, Watson K, Wiener P. Detecting signatures of selection in nine distinct lines of broiler chickens. Anim Genet. 2015;46:37–49.
51. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186:241–62.
52. Gautier M, Vitalis R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012;28:1176–7.
53. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913–8.
54. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
55. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. 2007;5:e171.
56. Hu ZL, Park CA, Wu XL, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 2013;41:D871–9.
57. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011;2011:bar030.
58. Stenn KS, Paus R. Controls of hair follicle cycling. Physiol Rev. 2001;81:449–94.
59. Wang Z, Zhang H, Yang H, Wang S, Rong E, Pei W, et al. Genome-wide association study for wool production traits in a Chinese Merino sheep population. PLoS One. 2014;9:e107101.
60. Liu N, Li H, Liu K, Yu J, Cheng M, De W, et al. Differential expression of genes and proteins associated with wool follicle cycling. Mol Biol Rep. 2014;41:5343–9.
61. SRA Toolkit documentation. Available from: http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software.
62. Kijas J, Brauning R, Clarke SM, McCulloch A, Cockett NE, Saunders G, et al. Launching SheepGenomesDB: 100 million variants from nearly 500 sheep genomes. J Anim Sci. 2016;94:S92–3.
63. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
64. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

65. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
66. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
67. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
68. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
69. Institute Broad. Picard tool, version 1.128. Available from: http://broadinstitute.github.io/picard/.
70. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
71. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3:35.
72. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32:1749–51.
73. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
74. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26:2069–70.
75. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
76. Sheep reference genome Oar_4.0. Available from: https://www.ncbi.nlm.nih.gov/genome?term=ovisaries.
77. Bai Y, Sartor M, Cavalcoli J. Current status and future perspectives for sequencing livestock genomes. J Anim Sci Biotechnol. 2012;3:8.
78. Boitard S, Boussaha M, Capitan A, Rocha D, Servin B. Uncovering adaptation from sequence data: Lessons from genome resequencing of four cattle breeds. Genetics. 2016;203:433–50.
79. Kardos M, Luikart G, Bunch R, Dewey S, Edwards W, McWilliam S, et al. Whole-genome resequencing uncovers molecular signatures of natural and sexual selection in wild bighorn sheep. Mol Ecol. 2015;24:5616–32.
80. Herrero-Medrano JM, Megens HJ, Groenen MAM, Bosse M, Pérez-Enciso M, Crooijmans RPMA. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. BMC Genomics. 2014;15:601.
81. Kijas JW, Porto-Neto L, Dominik S, Reverter A, Bunch R, McCulloch R, et al. Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Anim Genet. 2014;45:754–7.
82. Chitneedi PK, Arranz JJ, Suárez-Vega A, García-Gámez E, Gutiérrez-Gil B. Estimations of linkage disequilibrium, effective population size and ROH-based inbreeding coefficients in Spanish Churra sheep using imputed high-density SNP genotypes. Anim Genet. 2017;48:436–46.
83. Meadows JRS, Chan EKF, Kijas JW. Linkage disequilibrium compared between five populations of domestic sheep. BMC Genet. 2008;9:61.
84. Al-Mamun HA, Clark SA, Kwan P, Gondro C. Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. Genet Sel Evol. 2015;47:90.
85. Sánchez Belda A, Sánchez Trujillano MC. Razas ovinas españolas. Publicaciones de Extensión Agraria, Ministerio de Agricultura, Pesca y Alimentación; 1986.
86. Arranz JJ, Bayón Y, San Primitivo F. Genetic relationships among Spanish sheep using microsatellites. Anim Genet. 1998;29:435–40.
87. García-Gámez E, Sahana G, Gutiérrez-Gil B, Arranz J-J. Linkage disequilibrium and inbreeding estimation in Spanish Churra sheep. BMC Genet. 2012;13:43.
88. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312:1614–20.
89. González-Rodríguez A, Munilla S, Mouresan EF, Cañas-Álvarez JJ, Díaz C, Piedrafita J, et al. On the performance of tests for the detection of signatures of selection: a case study with the Spanish autochthonous beef cattle populations. Genet Sel Evol. 2016;48:81.
90. Gholami M, Reimer C, Erbe M, Preisinger R, Weigend A, Weigend S, et al. Genome scan for selection in structured layer chicken populations exploiting linkage disequilibrium information. PLoS One. 2015;10:e0130497.
91. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.
92. Matika O, Riggio V, Anselme-Moizan M, Law AS, Pong-Wong R, Archibald AL, et al. Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. Genet Sel Evol. 2016;48:11.
93. Raadsma HW, Thomson PC, Zenger KR, Cavanagh C, Lam MK, Jonas E. Mapping quantitative trait loci (QTL) in sheep. I. A new male framework linkage map and QTL for growth rate and body weight. Genet Sel Evol. 2009;41:34.
94. Cavanagh CR, Jonas E, Hobbs M, Thomson PC, Tammen I, Raadsma HW. Mapping quantitative trait loci (QTL) in sheep. III. QTL for carcass composition traits derived from CT scans and aligned with a meta-assembly for sheep and cattle carcass QTL. Genet Sel Evol. 2010;42:36.
95. Al-Mamun HA, Kwan P, Clark SA, Ferdosi MH, Tellam R, Gondro C. Genome-wide association study of body weight in Australian Merino sheep reveals an orthologous region on OAR6 to human and bovine genomic regions affecting height and weight. Genet Sel Evol. 2015;47:66.
96. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467:832–8.
97. Tetens J, Widmann P, Kühn C, Thaller G. A genome-wide association study indicates LCORL/NCAPG as a candidate locus for withers height in German Warmblood horses. Anim Genet. 2013;44:467–71.
98. Eberlein A, Takasuga A, Setoguchi K, Pfuhl R, Flisikowski K, Fries R, et al. Dissection of genetic factors modulating fetal growth in cattle indicates a substantial role of the *non-SMC condensin I complex, subunit G* (*NCAPG*) gene. Genetics. 2009;183:951–64.
99. Sahana G, Höglund JK, Guldbrandtsen B, Lund MS. Loci associated with adult stature also affect calf birth survival in cattle. BMC Genet. 2015;16:47.
100. Setoguchi K, Watanabe T, Weikard R, Albrecht E, Kühn C, Kinoshita A. The SNP c.1326T > G in the *non-SMC condensin I complex, subunit G* (*NCAPG*) gene encoding a p.Ile442Met variant is associated with an increase in body frame size at puberty in cattle. Anim Genet. 2011;42:650–5.
101. Setoguchi K, Furuta M, Hirano T, Nagao T, Watanabe T, Sugimoto Y. Cross-breed comparisons identified a critical 591-kb region for bovine carcass weight QTL (CW-2) on chromosome 6 and the Ile-442-Met substitution in *NCAPG* as a positional candidate. BMC Genet. 2009;10:43.
102. Lindholm-Perry AK, Sexten AK, Kuehn LA, Smith TPL, King DA, Shackelford SD, et al. Association, effects and validation of polymorphisms within the *NCAPG-LCORL* locus located on BTA6 with feed intake, gain, meat and carcass traits in beef cattle. BMC Genet. 2011;12:103.
103. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 2011;7:e1002316.
104. Abo-Ismail MK, Vander Voort G, Squires JJ, Swanson KC, Mandell IB, Liao X, et al. Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. BMC Genet. 2014;15:14.
105. Gowen LC, Petersen DN, Mansolf AL, Qi H, Stock JL, Tkalcevic GT, et al. Targeted disruption of the *osteoblast/osteocyte factor 45* gene (*OF45*) results in increased bone formation and bone mass. J Biol Chem. 2003;278:1998–2007.
106. Wei J, Geale PF, Sheehy PA, Williamson P. The impact of ABCG2 on bovine mammary epithelial cell proliferation. Anim Biotechnol. 2012;23:221–4.
107. Sheehy PA, Riley LG, Raadsma HW, Williamson P, Wynn PC. A functional genomics approach to evaluate candidate genes located in a QTL interval for milk production traits on BTA6. Anim Genet. 2009;40:492–8.
108. Kühn C, Weikard R, Widmann P. Metabolomics: a pathway for improved understanding of genetic modulation of mammalian growth and tissue deposition. In: Proceedings of the 10th world congress on genetics applied to livestock production: 17–22 August 2014; Vancouver. 2014.

109. Liu Y, Duan X, Chen S, He H, Liu X, Liu Y, et al. *NCAPG* is differentially expressed during longissimus muscle development and is associated with growth traits in Chinese Qinchuan beef cattle. Genet Mol Biol. 2015;38:450–6.
110. UniProt Consortium. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2008;36:D190–5.
111. Martínez-Cerezo S, Sañudo C, Panea B, Medel I, Delfa R, Sierra I, et al. Breed, slaughter weight and ageing time effects on physico-chemical characteristics of lamb meat. Meat Sci. 2005;69:325–33.
112. Campo MM, Olleta J, Sañudo C. Características de la carne de cordero con especial atención al Ternasco de Aragón. Agencia Aragonesa de Seguridad Alimentaria. 2008.
113. Allain D, Miari S, Usai MG, Barillet F, Sechi T, Sechi S, et al. SNP mapping of QTL affecting wool traits in a sheep backcross Sarda-Lacaune resource population. In: Proceedings of the 64th annual meeting of the European Federation of Animal Science: 26–30 August 2013; Nantes. 2013.
114. Cano M, Allain D, Foulquié D, Moreno C, Mulsant P, François D, et al. Fine mapping of birthcoat type in the Romane breed sheep. In: Proceedings of the 64th Annual Meeting of the European Federation of Animal Science: 26–30 August 2013; Nantes. 2013.
115. Olivier W, Olivier J, Greyling A. Quantifying the relationship between birth coat score and wool traits in Merino sheep. In: Proceedings of the 10th world conference on animal production: 23–28 November 2008; Cape Town. 2008.
116. Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, et al. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum Genet. 2015;134:823–35.
117. Johnston SE, McEwan J, Pickering NK, Kijas JW, Beraldi D, Pilkington JG, et al. Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol. 2011;20:2555–66.
118. Allais-Bonnet A, Grohs C, Medugorac I, Krebs S, Djari A, Graf A, et al. Novel insights into the bovine polled phenotype and horn ontogenesis in Bovidae. PLoS One. 2013;8:e63512.
119. Wiedemar N, Tetens J, Jagannathan V, Menoud A, Neuenschwander S, Bruggmann R, et al. Independent polled mutations leading to complex gene expression differences in cattle. PLoS One. 2014;9:e93435.
120. Yuan FP, Li X, Lin J, Schwabe C, Bullesbach EE, Rao CV, et al. The role of RXFP2 in mediating androgen-induced inguinoscrotal testis descent in LH receptor knockout mice. Reproduction. 2010;139:759–69.
121. Scott DJ, Rosengren KJ, Bathgate RAD. The different ligand-binding modes of relaxin family peptide receptors RXFP1 and RXFP2. Mol Endocrinol. 2012;26:1896–906.
122. Wiedemar N, Drögemüller C. A, 1.8-kb insertion in the 3′-UTR of RXFP2 is associated with polledness in sheep. Anim Genet. 2015;46:457–61.
123. Lühken G, Krebs S, Rothammer S, Küpper J, Mioč B, Russ I, et al. The 1.78-kb insertion in the 3′-untranslated region of *RXFP2* does not segregate with horn status in sheep breeds with variable horn status. Genet Sel Evol. 2016;48:78.
124. Bartels CF, Bükülmez H, Padayatti P, Rhee DK, van Ravenswaaij-Arts C, Pauli RM, et al. Mutations in the *transmembrane natriuretic peptide receptor NPR-B* impair skeletal growth and cause acromesomelic dysplasia, type Maroteaux. Am J Hum Genet. 2004;75:27–34.
125. Vasques GA, Amano N, Docko AJ, Funari MFA, Quedas EPS, Nishi MY, et al. Heterozygous mutations in *natriuretic peptide receptor-B* (*NPR2*) gene as a cause of short stature in patients initially classified as idiopathic short stature. J Clin Endocrinol Metab. 2013;98:E1636–44.
126. García-Gámez E, Gutiérrez-Gil B, Sahana G, Sánchez JP, Bayón Y, Arranz JJ. GWA analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the LALBA gene. PLoS One. 2012;7:e47782.
127. Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics. 2014;15:442.

Resultado 1.3

Detection of quantitative trait loci and putative causal variants affecting somatic cell score in dairy sheep by using a 50K SNP chip and whole-genome sequencing

B. Gutiérrez-Gil,[1,2] C. Esteban-Blanco,[1] A. Suarez-Vega, and J. J. Arranz

[1]These two authors equally contributed to this work.

Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain

Journal of Dairy Science, 101(10), 9072–9088. https://doi.org/10.3168/jds.2018-14736

J. Dairy Sci. 101:9072–9088
https://doi.org/10.3168/jds.2018-14736

Detection of quantitative trait loci and putative causal variants affecting somatic cell score in dairy sheep by using a 50K SNP chip and whole-genome sequencing

B. Gutiérrez-Gil,[1,2] C. Esteban-Blanco,[1] A. Suarez-Vega, and J. J. Arranz
Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain

ABSTRACT

This study presents a scan of the ovine genome to identify quantitative trait loci (QTL) influencing the somatic cell score (SCS), a classical indicator of subclinical mastitis in sheep, and a subsequent high-resolution analysis of one of the identified QTL regions based on the analysis of whole-genome sequence data sets. A half-sib commercial population of Churra sheep genotyped with a 50K SNP chip was analyzed using linkage analysis (LA) and combined linkage and linkage disequilibrium analysis (LDLA). By LA, 2 5% chromosome-wide significant QTL on OAR5 and OAR25 and one 5% genome-wide significant QTL on ovine chromosome 20 (OAR20) were detected, whereas 22 significant associations were identified by LDLA. Two of the associations detected by LDLA replicated LA-detected effects (OAR20, OAR25). We compared the detected associations with previously reported QTL in sheep and cattle, and functional candidate genes were identified within the estimated confidence intervals. We then performed a high-resolution analysis of the OAR20 QTL region, the most significant QTL region identified by LA that replicated a QTL previously described in Churra sheep for SCS using microsatellite markers. For that, 2 segregating trios of 2 segregating families for the OAR20 QTL (each including the *Qq* sire and 2 daughters, *QQ* and *qq*) were selected for whole-genome sequencing. The bioinformatic analysis of the 6 sequenced samples performed across the genomic interval considered (14.2–41.7 Mb) identified a total of 227,030 variants commonly identified by 2 independent software packages. For the 3 different concordance tests considered, due to discrepancies regarding the QTL peak in the segregating families, the list of mutations concordant with the QTL segregating pattern was processed to identify the variants identified in immune-related genes that show a moderate/high impact on the encoded protein function. Among a list of 85 missense variants concordant with the QTL segregation pattern that were within candidate immune-related genes, 13 variants distributed across 7 genes [*PKHD1*, *NOTCH4*, *AGER*, *ENSOARG00000009395* (*HLA-C*, *Homo sapiens*), *ENSOARG00000015002* (*HLA-B*, *H. sapiens*), *MOG*, and *ENSOARG00000018075* (*BoLA*, *Bos taurus*, orthologous to human *HLA-A*] were predicted to cause deleterious effects on protein function. Future studies should assess the possible associations of the candidate variants identified herein in commercial populations with indicator traits of udder inflammation (SCS, clinical mastitis).

Key words: mastitis, quantitative trait loci, single nucleotide polymorphism-chip, genomic sequencing, genetic marker

INTRODUCTION

In dairy species, the SCC of milk represents a predictive marker of the udder health and is widely used for evaluating milk quality. It also influences milk prices. An increased SCC is either the consequence of an inflammatory process due to the presence of an IMI or, under nonpathological conditions, due to physiological processes such as estrus or an advanced stage of lactation (Raynal-Ljutovac et al., 2007).

Subclinical mastitis constitutes one of the major problems influencing total productivity in dairy sheep. Therefore, resistance/susceptibility to this disease can be considered an important functional trait for the milk production sector. Because SCC provides a measurement of the level of defensive cells that migrate from blood to mammary gland as a response to infection (Gonzalo and Gaudioso, 1985), log-transformed SCC, known as the SCS, can be used as an indicator trait to achieve genetic improvement for mastitis resistance (Shook and Schutz, 1994). Although direct selection

Received March 12, 2018.
Accepted June 21, 2018.
[1] These two authors equally contributed to this work.
[2] Corresponding author: beatriz.gutierrez@unileon.es

for mastitis resistance has been implemented in dairy cattle for over 35 yr in Nordic countries (Østerås et al., 2007) and more recently in France (Govignon-Gion et al., 2016) and Canada (Jamrozik et al., 2013), most countries breed for mastitis resistance indirectly through SCS (Miglior et al., 2005). In dairy sheep, the SCS is considered a functional indicator trait of subclinical mastitis and is one of the factors influencing the price that farmers receive for the milk. In dairy sheep, reported heritability estimates of SCS range between 0.06 and 0.18 (Othmane et al., 2002; Rupp et al., 2003; Legarra and Ugarte, 2005). The SCS is included as a selection target in the breeding scheme of the French Lacaune breed (Barillet et al., 2006). In Churra sheep, although SCS is routinely recorded through the official recording control, the low number of rams under genetic evaluation makes consideration of this trait unfeasible when calculating the selection index. Indirect selection for subclinical mastitis resistance is performed in Churra sheep through the inclusion of udder morphology traits as selection objectives (de la Fuente et al., 1996), with the advantage that these traits are more heritable than SCS. The efficiency of this indirect selection to favor stabilization of SCS would be related to the expected genetic correlations between udder traits and mastitis resistance. Hence, in a Lacaune × Sarda backcross population, Casu et al. (2010) reported high genetic correlations between SCS and both udder attachment (measured as degree of suspension of the udder) and udder depth (−0.42 and −0.50, considering a scale of opposite sign for udder depth than in Churra sheep). These estimates suggest that selection for shallow udders, close to the abdominal wall, and udders with higher degree of suspension would be associated with a genetic response toward lower SCS. Although in Churra sheep genetic correlations have not been estimated between udder morphology traits and mastitis resistance, phenotypic correlations reported in this breed between SCS and udder depth and between SCS and teat size are low but positive (0.13 and 0.18 respectively), which can be explained by the higher frequency of trauma observed in very deep udders and larger teats than standard teat cups (Fernández et al., 1997).

Nevertheless, taking into account the direct influence of SCS on the price of milk, direct selection on this trait would be of great interest for breeders of Churra dairy sheep. As for other traits showing low heritability, marker- or gene-assisted selection would be a feasible strategy to improve resistance to subclinical mastitis not only in Churra sheep but also in other ovine populations devoted to milk production. Detecting genetic variants directly associated with the SCS trait could be exploited to increase the average resistance level of flocks to subclinical mastitis. Historically, the first attempts to identify genes related to this functional trait in dairy sheep populations were genome scans based on microsatellite markers aiming to identify QTL (reviewed by Arranz and Gutiérrez-Gil, 2012). In Churra sheep, an analysis of a half-sib population with 181 markers identified only one chromosome-wide significant QTL for SCS on sheep chromosome 20 (*Ovis aries* 20; **OAR20**). However, the low mapping resolution of this scan together with the differences in marker informativeness among the analyzed families limited the ability to identify reliable candidate genes for this QTL effect (Gutiérrez-Gil et al., 2007).

Presently, SNP chips of medium and high density in sheep provide a substantially improved mapping tool for the identification of QTL that directly control traits of economic interest. In addition, animal scientists currently have access to whole-genome sequence-based technologies, which increases their ability to detect and propose mutations as plausible causal mutations or quantitative trait nucleotides (Sellner et al., 2007).

In Churra sheep, a medium-density 50K SNP chip has been used by our research group to map QTL related to milk production traits (García-Gámez et al., 2012) and parasite resistance traits (Atlija et al., 2016). For milk traits, this chip greatly facilitated the identification of the putative causal mutation of a previously described QTL influencing milk protein percentage in Spanish Churra sheep (García-Gámez et al., 2012). In a more recent study, the combination of 50K SNP chip genotyping in a commercial population of French dairy sheep and the analysis of whole-genome sequencing information allowed the identification of the causal mutation of an OAR3 QTL influencing mastitis susceptibility in the *SOCS2* gene (Rupp et al., 2015). A custom-made 960-SNP DNA array has been recently used in the Greek Chios breed to confirm previously detected QTL in other sheep breeds and has suggested, for some of the regions, a conserved genetic architecture of mastitis resistance between distinct dairy sheep breeds (Banos et al., 2017).

The objectives of the present study were to (1) perform QTL mapping analyses for the SCS trait using a 50K SNP chip based on linkage (**LA**) and combined linkage and linkage disequilibrium (**LDLA**) analyses in the same commercial half-sib population of Churra dairy sheep analyzed for milk production traits by García-Gámez et al. (2012), and (2) exploit the whole-genome sequences of segregating trios to perform a high-resolution study of the genetic variation within the region harboring the most significant QTL detected and assess the polymorphisms showing concordance with the expected QTL genotypes as potential causal variants based on their biological relevance and the physiological effects of the affected gene.

MATERIALS AND METHODS

Resource Population, Genotypes, and Phenotypes

A commercial population of Spanish Churra dairy sheep including 1,598 ewes distributed in 16 half-sib families and the corresponding 16 sires was studied here. The average family size was 99 daughters per ram (ranging from 23 to 266 animals per half-sib family). Test-day SCS ($base_{10}$ logarithmic transformation) records for the ewes of this population were estimated based on the SCC provided by official milking records of the National Churra Breeders' Association (ANCHE). As the response variable for the QTL analysis, we used the yield deviations (**YD**) for the SCS trait estimated, as previously detailed by García-Gámez et al. (2012), with a multivariate animal repeatability model considering also other milk production traits measured through test-day records (milk yield, protein percentage, fat percentage, protein yield, fat yield). For the YD estimation, the raw phenotypic data were corrected for the environmental effects of herd test day, birth order, age of the ewe at parturition (as a covariate nested within birth order), number of born lambs, number of weeks of milk production of the ewe, and the ewe's permanent environmental effect.

To extract DNA, blood samples for the ewes and semen samples for the sires were collected by the official veterinarians of the breeders' association following standard animal welfare protocols. The whole population was genotyped with the Illumina Ovine SNP50 BeadChip (Illumina Inc., San Diego, CA). An initial control of raw genotypes was performed by applying a GenCall score greater than 0.15. Then, SNP order and genome positions were updated according to version Oar_v3.1 of the ovine genome assembly (http://www.ensembl.org/Ovis_aries/Info/Index) by considering a 1 cM to 1 Mb conversion rate. Only the SNP with known location on the ovine autosomes were considered. Following Anderson et al. (2010), we performed a 2-step quality control of the genotypes: per animal (call rate >90%) and per SNP (call rate >95%; minor allele frequency >0.05; correspondence with Hardy-Weinberg equilibrium: $P > 0.00001$). All the animals under study and a total of 43,613 autosomal SNP passed the quality control filters and were considered in the QTL mapping analysis.

QTL Mapping Analyses

Genome scans based on a classical LA and a combined LDLA procedure were performed for the SCS trait with the QTLMap software (Filangi et al., 2010), by testing the genome at 0.1 cM step intervals. For the 2 analyses, the half-sib structure of the studied population was indicated in the analysis with the *family = 1* option. A by-default haplotype size of 4 SNP was used for LDLA. Significance thresholds at the chromosome-wise significance level (P_c-value) were calculated through a total of 1,000 permutations (at 0.1 cM steps) for LA and 1,000 simulations (at 5 cM steps) for LDLA. For both analyses, genome-wise significance thresholds were calculated based on the chromosome-wise *P*-values by applying a Bonferroni correction for the 26 independent chromosomes under analysis (P_c-value < 0.0019, for a 5% genome-wise significance level). For the significant QTL that were detected by LA, likelihood ratio test (**LRT**) values were converted to logarithm odds ratio (**LOD**) values (Lander and Botstein, 1989), and confidence intervals (**CI**) for the QTL locations were estimated by the widely used 1-LOD drop-off method. The proportion of phenotypic variance that was explained by the QTL detected by LA was calculated based on the corresponding LOD values using the formula $\sigma_p = 1 - 10^{-\frac{2}{n} LOD}$ (Broman and Sen, 2009). In the LDLA, chromosomal regions that involved consecutive significant haplotype associations within a chromosome (allowing gaps no greater than 5 cM) were grouped as a significant LDLA interval and the remaining ones were considered as isolated significant haplotypes.

Comparison with Previously Reported QTL and Identification of Functional Candidate Genes Within Confidence Intervals

A systematic search for QTL and associations previously reported in sheep and the extraction of positional candidate genes was performed for the QTL regions detected by LA and LDLA. For each QTL detected, we considered the corresponding target genomic interval (**TGI**), which was defined as the genomic region based on the sheep reference genome assembly Oar_v3.1 that corresponded to either the estimated CI of the across-family LA analysis and the defined significant LDLA intervals or to a 250 kb-long interval centered on the significant LDLA-isolated haplotypes detected by LDLA. Once the TGI for each significant association was defined, we extracted the QTL annotated in the SheepQTL database (Hu et al., 2013) in relation to "Health traits" and "Udder" (one of the type traits within the "Exterior traits" category of the database hierarchy). Considering that most gene-mapping studies for mastitis resistance and udder morphology have been performed with dairy cattle populations, we searched for QTL described in the corresponding bovine orthologous regions to the TGI defined in our study based on LA and LDLA. We used *Liftover* (https://genome.ucsc

.edu/cgi-bin/hgLiftOver) to define the orthologous coordinates of the TGI in the *Bos taurus* UMD_3.1 assembly (BTA23:21961453–28742358) and extracted the QTL/associations related to mastitis resistance and udder morphology reported in cattle based on the Cattle database from AnimalQTLdb (Hu et al., 2013).

In addition, the positional candidate genes included within those intervals according to the reference genome were extracted with the BioMart web-based tool based on the Ensembl release 89 (http://www.ensembl.org/biomart/martview/). These positional candidate genes were assessed as putative functional candidate genes in relation to the immune response by performing a survey of a database of 5,029 unique immune-related genes, which was based on the IRIS (1,489 genes; Kelley et al., 2005) and ImmPort (4,677 genes) gene lists, both of which are available at http://www.innatedb.com/redirect.do?go=resourcesGeneLists.

Selection of Target Segregating Trios for Further Study of the Oar20 QTL

For the most significant QTL detected by LA on OAR20, which was also confirmed by the LDLA results, we selected 2 sires that, according to the within-family analyses, were segregating (*Qq*) for the OAR20 QTL. For each of these 2 sires, we identified 2 daughters with extreme divergent phenotypes for the SCS trait, in correspondence with the alternative homozygosity genotypes for markers included in the estimated QTL CI. The selection of the daughters was based on the phases obtained for each of the families for the markers included in the QTL within-family CI. For each trio, the sire and the offspring phases were obtained with the *out_phases* and *out_phases_offspring* options of QTLMap (Le Roy et al., 2013). To assign the QTL alleles, *Q* and *q*, to the corresponding paternal phase at the QTL regions, for each of the 2 families considered, the *out_pded* option of this software was used to estimate the transmission marginal probabilities for all of the animals of the 2 selected families for the QTL peak position in the within-family analysis. By considering the sign of the estimated effect for each sire and the phase of offspring showing extreme divergent inheritance probabilities (close or equal to zero or one) at the target QTL peak position, we could identify the parental haplotype associated with increased or decreased SCS. The haplotype for which a positive effect on the SCS trait was calculated in the regression analyses was denoted as *q* (increased susceptibility to mastitis), and the haplotype for which a negative effect on the trait was calculated was denoted as *Q* (increased resistance to mastitis). Based on this, for each selected family, we identified the daughters inheriting the *Q* and *q* allele at the corresponding target QTL position and ranked the list based on their individual SCS trait values (YD units). Later, focusing on the region of the within-family QTL peak, we selected homozygous daughters with extreme phenotypes consistent with the QTL effect. Hence, we identified, for each of the 2 selected families, a *QQ* daughter showing an extreme low SCS value and a *qq* daughter showing an extreme high SCS within the family SCS values.

Whole-Genome Sequencing and NGS-Variant Calling

Six DNA samples, 3 from each selected segregating family, were subjected to whole-genome sequencing by using the paired-end Illumina technology in an Illumina HiSeq 2000 sequencer. The bioinfomatic analysis workflow implemented for these samples was the same as that detailed by Gutiérrez-Gil et al. (2017). Briefly, the raw paired-end reads resulting from sequencing were subjected to quality assessment with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, the good quality reads filtered with Trimmomatic (Bolger et al., 2014) were mapped to the ovine reference genome Oar_v3.1 using the Burrows-Wheeler Aligner (BWA, *mem* mapping function; Li and Durbin, 2009). After that, SAMtools3 (Li et al., 2009) and PicardTools (Broad Institute, 2017) were used for diverse manipulations (see Gutiérrez-Gil et al., 2017, for details). Variant calling (SNP and InDEL) was carried out across the whole genome simultaneously for the 6 samples with 2 different software, the Genome Analysis Toolkit (GATK4, version 3.3.0, *HaplotypeCaller* tool) and Samtools (Li, 2011; *mpileup analysis*). After discarding low quality variants independently from each resulting VCF file using snpSIFT (Cingolani et al., 2012), an intersect set, containing those variants concordant between GATK and Samtools predictions, was extracted using BCFtools utilities (Li, 2011; Narasimhan et al., 2016) to produce a final VCF file. The reliability of whole-genome sequencing (**WGSeq**)-defined genotypes was determined by comparing with the previously analyzed 50K SNP chip genotypes.

Concordance Tests and Identification of Functional Variant Annotation

Considering the 6 samples sequenced at the whole-genome level, we performed concordance tests to filter the variants showing genotypic concordance with the considered QTL segregation pattern (*Qq* for the 2 sires, and the alternative homozygous genotypes for the daughters showing extreme phenotypes in agreement

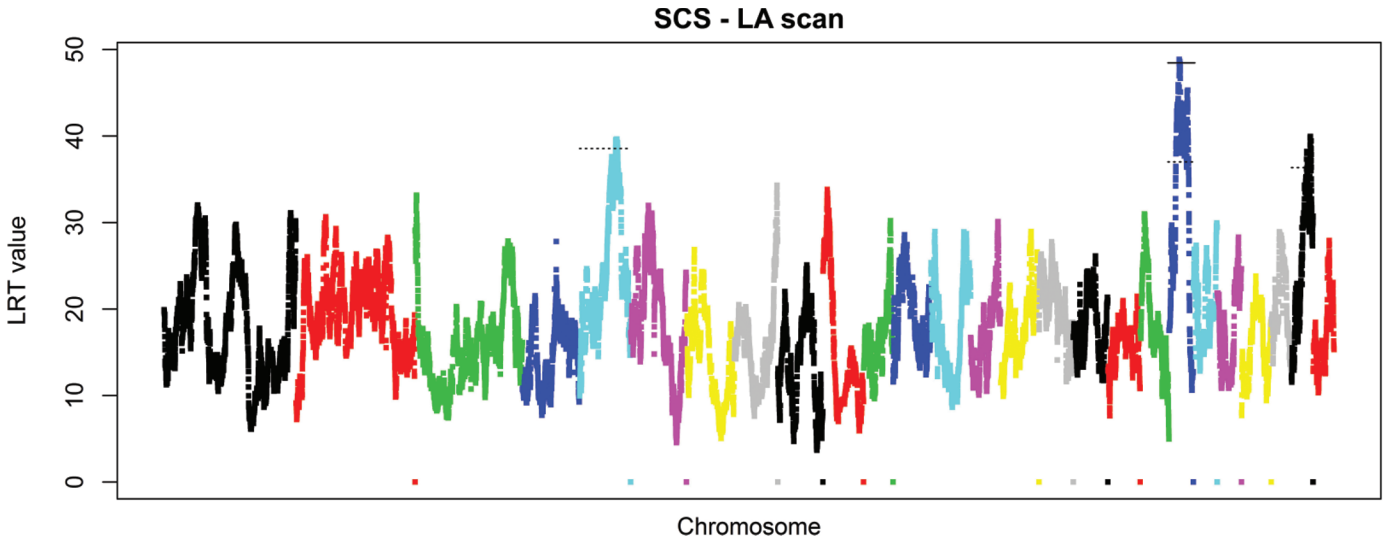


Figure 1. Results of the genome scan based on linkage analysis (LA) performed for the SCS trait studied in the present work. Likelihood ratio test (LRT) values obtained across the 26 ovine autosomes are represented. For the chromosomes (OAR) showing significant results, the dashed horizontal lines (OAR5, OAR20, OAR25) indicate the 5% chromosome-wise significance threshold, and the solid line (on OAR20) indicates the 5% genome-wise significance threshold. Color version available online.

with the chromosomal phase inherited from the sire, *QQ* or *qq*).

For the variants satisfying the QTL concordance genotype pattern, for each data set, functional annotation was performed with the Ensembl variant effect predictor (**eVEP**; McLaren et al., 2010) based on the Ovis_aries build 89 of Ensembl. Based on the functional annotation, we identified the genetic variants (indels and SNP) that were predicted to cause relevant biological effects (stop coding variant, missense variant, or frameshift variant). The classification of the variants as "deleterious" or "tolerated" by the SIFT tool was possible through the eVEP analysis. The known functions of the genes harboring the "deleterious" functional variants were assessed in relation to the immune response by performing a survey with a database of 5,029 immune-related genes considered as reference in this work. This database was based on the IRIS (1,489 genes; Kelley et al., 2005) and ImmPort (4,677 genes) gene lists, both of which are available at http://www.innatedb.com/redirect.do?go=resourcesGeneLists.

RESULTS

QTL Mapping Results

LA Results. The LA-based genome scan identified a genome-wide significant QTL on OAR20 (5% genome-wise threshold) and 2 chromosome-wide significant QTL on OAR5 (CI: 74.9–81.5 cM) and OAR25 (CI: 38.6–41.7 cM; Figure 1). A summary of the across-family analysis performed across the 26 ovine autosomes in our resource population is shown in Table 1, which also includes the results of the segregating families identified by the within-family analyses. The CI estimated for the OAR20 QTL (peak: 21.52 cM) included a 2.9 cM interval (20.9–23.8). For the 3 segregating families identified for this QTL, the magnitude of this QTL effect, expressed in phenotypic standard deviations, was 0.205 (family 1), 0.134 (family 2), and −0.192 (family 6). A single segregating family was identified for each of the chromosome-wide QTL identified on OAR5 and OAR25. For these QTL, the estimated effects were 0.384 and −0.405, respectively (Table 1). The estimated phenotypic variance explained by these QTL was 0.03 for the OAR20 QTL and approximately 0.024 for the 2 other mapped effects.

The results of the across-family analysis of OAR20 (Supplemental Figure S1; https://doi.org/10.3168/jds.2018-14736) show that although the maximum QTL peak was found at 21.52 cM, a secondary significant peak was identified at 38 cM. This profile is consistent with the results of the within-family analyses in which 2 families showed the maximum LRT and the corresponding estimated CI mapping within the first third of the chromosome and overlapping the across-family CI (family 1 at 18.82 cM and family 2 at 14.8 cM (Table 1; Supplemental Figure S2; https://doi.org/10.3168/jds.2018-14736), whereas the other segregating family (family 6) showed the maximum LRT value at 38.1 cM, overlapping with the secondary peak identified in the across-family QTL profile. For family 6, significant

Table 1. Significant QTL influencing SCS identified by the linkage analysis (LA) performed in the current study

	Across-family analysis			Within-family analysis			
OAR[1]	Position maximum LRT (cM)[2] [flanking markers]	Maximum LRT[3] (P-value)	CI[4] (cM)	Segregating families	Position maximum LRT (cM)[2] [flanking markers]	CI[4] (cM)	Size effect ± SE[5] (trait units; SD units)
OAR5	77.99 [OAR5_85726404-OAR5_85777507]	39.639 (P_c-value < 0.05)	74.9–81.5	Fam_5	65.4–66.8 cM [OAR5_71989039-OAR5_73467317.1]	60.4–97.8	0.092 ± 0.012 (0.384)
OAR20	21.52 [OAR20_22784700- OAR20_22851804]	48.848 (P_g-value < 0.05)	20.9–23.8	Fam_1	18.82 cM [OAR20_19701934-OAR20_19761881]	14.2–24.2	0.049 ± 0.006 (0.205)
				Fam_2	14.8 cM [s01331-OAR20_15691996]	12.6–15.3	−0.046 ± 0.004 (−0.192)
				Fam_6	38.12 cM [OAR20_41543083-OAR20_41618542]	19.9–41.7	0.032 ± 0.004 (0.134)
OAR25	39.48 [s10781- s37668]	39.939 (P_c-value < 0.05)	38.6–41.7	Fam_8	32.58 cM [s70694-OAR25_34067791]	27.4–36.9	−0.097 ± 0.015 (−0.405)

[1]OAR = ovine chromosome.

[2]Position of the chromosome (cM) at which the maximum likelihood ratio test (LRT) of the LA is reached in the analysis involving the 16 half-sib families included in this work (across-family analysis) or the individual analysis of the segregating families (those showing a P_c-value <0.05 in the within-family analysis), respectively. The flanking markers for that position are indicated.

[3]P_c-value: Chromosome-wise significance P-value established through 1,000 permutation analysis; P_g-value: genome-wise significance P-value established on the basis of the chromosome-wise significance P-values and considering the 26 autosomes analyzed.

[4]Confidence interval (cM) estimated from the position of the maximum LRT for the across-family analysis and the within-family analyses, respectively, following the 1-logarithm odds ratio-drop-off method (Lander and Botstein, 1989).

[5]Estimated size effect of the QTL identified in the within-family analysis expressed in trait units (yield deviations of SCS) and in phenotypic SD units of the trait (in parentheses).

LRT values were detected across most of the second half of the chromosome, resulting in a long estimated CI covering a 21.8 cM interval (19.9–41.7 cM).

LDLA Results. The LDLA scan identified a total of 21 significant QTL (5% chromosome-wide) distributed across 13 autosomes: OAR1, 2, 3, 8, 11, 13, 14, 17, 18, 19, 20, 22, and 25 (Table 2; Figure 2). Eight of the significant QTL involved consecutive significant haplotypes that defined a significant LDLA interval, whereas the other 13 associations were defined by isolated haplotypes. One of the 2 significant associations located on OAR25 (16.5–16.6 cM) reached the 5% genome-wide significant threshold. Two 5% chromosome-wise QTL located on OAR20 (22–28 cM) and OAR25 (32.5–35 cM) overlapped with the CI estimated for the significant QTL detected by LA on those chromosomes. As can be seen in Supplemental Figure S1 (https://doi.org/10.3168/jds.2018-14736), the LDLA statistical profile observed on OAR20 showed correspondence with that provided by LA, with the most significant region involving the 22 to 28 cM interval. A second significant position was detected by LDLA at the proximal end of the chromosome (5.4–5.5 cM), whereas the 39 cM position did not exceed the chromosome-wise significance threshold.

Correspondence with Sheep and Cattle QTL and Functional Candidate Genes

The correspondence between the TGI defined from the identified significant QTL in this work and previously reported QTL/associations annotated in the SheepQTLdb for "Health" and "Udder" traits are presented in Supplemental Table S1 (https://doi.org/10.3168/jds.2018-14736). The estimated CI for the QTL detected by LA overlapped with 6 QTL annotated for health traits in this database, 3 of which mapped within the OAR20 QTL and 3 of which mapped within the OAR25 QTL. The TGI intervals defined based on the LDLA significant associations collocated with a total of 28 reported health-related QTL/associations, 3 of which were common to the overlapping associations with the LA-defined TGI. Most of these health-related overlapping QTL had been identified in studies focused on parasite resistance, whereas only one QTL underlying mastitis susceptibility was identified, on OAR22

Table 2. Significant QTL influencing SCS identified by the combined linkage disequilibrium and linkage analysis (LDLA) performed in the current study

OAR[1]	Position of maximum LDLA significant associations[2] (P_c-value < 0.05) (cM)	Flanking markers for the maximum LDLA	Maximum LRT[3] (P_g-value)	Significant LDLA intervals[4] (cM)
1	259.05	[OAR1_280315444.1–OAR1_280355916.1]	78.54	—
2	83.1	[s47616.1–OAR2_88340779.1]	75.1	83.1–83.2
	140.36	[s25821.1–OAR2_149199138.1]	80.20	
	185.0	[s06128.1-s70629.1]	79.37	185–185.1
3	94.4	[OAR3_100129266.1–OAR3_100393157.1]	75.64	—
	184.4	[s10640.1–s37078.1]	75.57	—
	212.7	[s65581.1–s27933.1]	80.71	—
8	80.6	[s26350.1–OAR8_86871896.1]	72.81	—
11	56.78	[s68143.1–s02321.1]	71.341	56.8–56.9
13	71.82	[s64654.1–DU360920_246.1]	74.307	68.5–71.8
14	32.0	[OAR14_33291858.1–OAR14_33340378.1]	68.88	—
	43.28	[s54719.1–OAR14_45342785.1]	69.17	—
17	33.8	[OAR17_36829676.1–s46426.1]	74.50	—
	42.3	[OAR17_45742264.1–OAR17_45809081.1]	72.81	—
18	12.65	[OAR18_12576454.1–s09546.1]	75.048	—
19	26.17	[s00124.1–OAR19_27611685.1]	77.11	—
20	5.5	[OAR20_5532584.1–OAR20_5577528.1]	73.99	5.4–55
	23.52	[OAR20_24966073.1–s18014.1]	79.01	22–28
22	23.76	[s69443.1–DU467879_183.1]	69.67	—
25	16.58	[s37560.1–OAR25_17161978.1]	84.38	16.5–16.6
	32.5	[OAR25_33894798.1–OAR25_33941270.1]	(P_g-value < 0.05)	32.5–35
			77.04	

[1]OAR = ovine chromosome.

[2]Position of the chromosome (cM) where 5% chromosome-wise significant (P_c-value < 0.05) haplotype associations were identified by LDLA.

[3]P_g-value: The genome-wise *P*-value is indicated only for the haplotype associations exceeding the 5% genome-wise threshold established on the basis of the chromosome-wise significance *P*-values and considering the 26 autosomes analyzed.

[4]A significant LDLA interval (cM) was defined by clustering close-by significant (P_c-value < 0.05) LDLA associations on a chromosome (allowing gaps no greater than 5 cM).

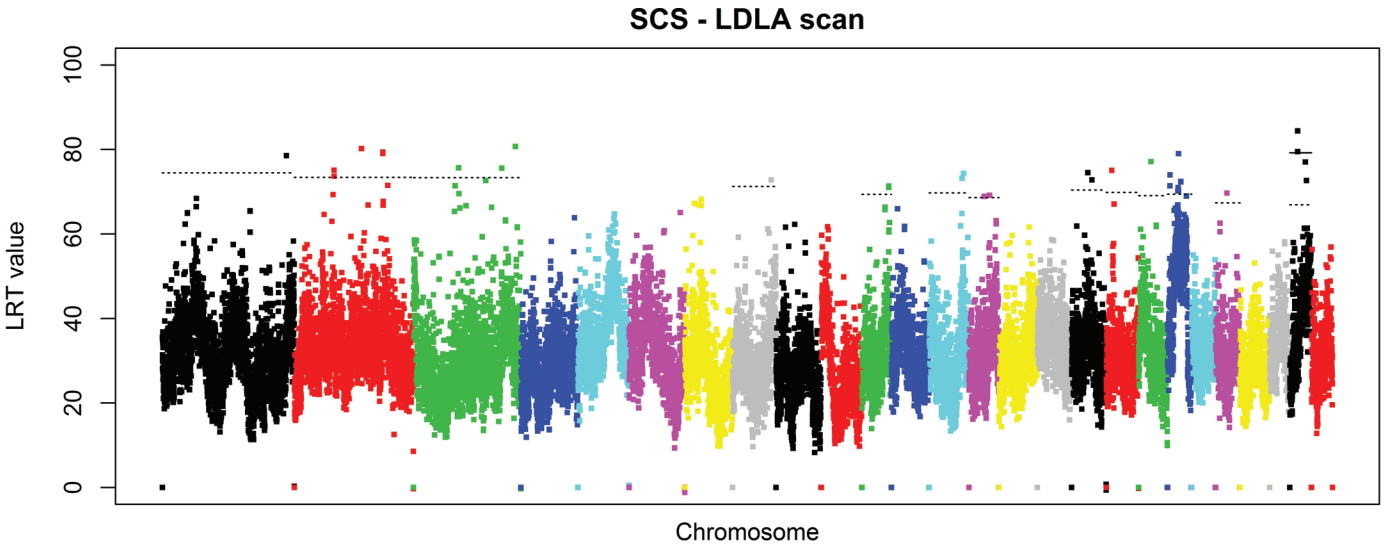


Figure 2. Results genome scan based on the combined linkage disequilibrium and linkage (LDLA) performed for the SCS trait studied in the present work. Likelihood ratio test (LRT) values obtained across the 26 ovine autosomes are represented. For the chromosomes (OAR) showing significant results, the dashed horizontal lines indicate the 5% chromosome-wise significance threshold and the solid line (on OAR25) indicates the 5% genome-wise significance threshold. Color version available online.

(Raadsma et al., 2009). Although parasite resistance traits are not directly related to mastitis resistance, we provide these correspondences (Supplemental Table S1; https://doi.org/10.3168/jds.2018-14736) considering that when a QTL for SCS overlaps with previously reported QTL for parasite resistance, this could serve as a hint to identify a gene or genes that are influencing the general immune response and that could have pleiotropic effects on disease resistance, from mastitis to parasite or other infection diseases. The region showing coincidence with the largest number of previously reported QTL for health-related traits was the LDLA significant association involving the TGI of 22 to 28 Mb on OAR20, although none of those coincident associations refer to mastitis resistance. Interestingly, the 22- to 28-Mb interval involves the major histocompatibility complex (**MHC**) class II region (*DQA*, *OVAR-DRB3*, *HLA_DRA* genes). In addition, looking in detail at studies recently reported in sheep for QTL influencing mastitis resistance related traits, we have seen that some of our significant LDLA associations were close to the reported associations although not always overlapped. For example, the significant association found on OAR19 in our study (at 26.17 Mb) is close to a large number of associations reported in Chios sheep for many different mastitis resistance traits (SCC, clinical mastitis, total viable bacterial count in milk, and so on) in the interval 26.442–28.134 Mb of that chromosome (Banos et al., 2017). Similarly, our results show vicinity to some SCC QTL reported in a commercial French dairy sheep population on OAR8 (82.60 Mb), OAR13 (70.89 Mb), OAR14 (39.40 Mb), and OAR19 (28.60 Mb; Rupp et al., 2015).

On the other hand, the TGI regions defined by our results did not show relevant overlapping with QTL influencing udder traits in sheep; only the OAR25 QTL detected by LA showed correspondence with 2 associations identified for the teat number trait (Supplemental Table S1; https://doi.org/10.3168/jds.2018-14736). As seen in Supplemental Table S2 (https://doi.org/10.3168/jds.2018-14736), our results showed correspondence with 19 bovine QTL for mastitis-related traits, and 6 of these coincidences were found within the OAR20 QTL as defined by LDLA, which corresponds to the region 21.96 to 28.74 Mb of bovine chromosome 23 (BTA23) where the bovine MHC is located (Supplemental Table S2). In addition, the associations reported in the present work for SCS overlapped with 21 bovine QTL previously reported for udder morphology traits. The udder traits affected by the larger number of these coincidences were teat length and udder depth, followed by teat placement and udder composite index (Supplemental Table S3; https://doi.org/10.3168/jds.2018-14736).

A total of 90 features, including 56 annotated protein-coding genes, 11 nonannotated protein-coding genes, and 1 pseudogene were identified by the Biomart extraction tool in the LA-defined TGI (Supplemental Table S4; https://doi.org/10.3168/jds.2018-14736). Within the LDLA-defined TGI, Biomart identified a total of 144 annotated protein-coding genes, 69 nonannotated protein-coding genes, and 1 pseudogene

(Supplemental Table S5; https://doi.org/10.3168/jds.2018-14736). A total of 74 genes included in the considered immune databases were identified within the statistically significant TIG, 16 and 58 within associations identified by LA and LDLA, respectively, with 5 genes detected within TGI defined by both analyses (Supplemental Table S6; https://doi.org/10.3168/jds.2018-14736). The OAR20 LDLA QTL at 22 to 28 Mb included the largest number of immune-related genes (47; Supplemental Table S6).

A Detailed Analysis of OAR20: Selection of Segregating Trio Samples for Whole-Genome Sequencing

The selection of the OAR20 QTL for a more detailed study based on whole-genome analysis was supported by the facts that it was the only genome-wide significant effect detected by LA and that it was also detected by LDLA. In addition, the TGI estimated for this QTL in the present study (20.9–23.8 Mb) overlaps with the QTL flanking interval (*BM1258-OLADBRPS*, which approximately corresponds to the region 20 to 34 Mb on the Oar_v3.1 reference genome) of one of the segregating families of the SCS QTL reported by Gutiérrez-Gil et al. (2007). Taking into account the 2 peaks identified in the across-family LRT profile obtained for OAR20 (Supplemental Figure S1; https://doi.org/10.3168/jds.2018-14736) and the within-family analysis results (Supplemental Figure S2, https://doi.org/10.3168/jds.2018-14736; Table 1), we selected family 1 (peak at 18.82 cM, close to the maximum across-family LRT at 21.52 cM) and family 6 (peak at 38.12 cM, overlapping with the secondary QTL peak identified in the across-family analysis at 38 cM) to further explore the genetic basis of the SCS QTL identified on OAR20.

Information regarding the selection of animals for whole-genome sequencing for the 2 segregating selected families is provided in Supplemental File S1 (https://doi.org/10.3168/jds.2018-14736). Briefly, based on the within-family phase segregating study, the daughters included in Trio 1 were 5481 (*QQ*, YD_SCS: −0.13; position 21/89 in the ranked list of *Q*-daughters by the lowest phenotype values) and 4404 (*qq*, YD_SCS: 0.259; position 8/80 in the ranked list of *q*-daughters by the highest phenotype values), referred hereafter as QQ1_5481 and qq1_4404. The daughters included in Trio 2 were 5594 (*QQ*, YD_SCS: −0.342; position 3/147 in the ranked list of daughters by the lowest phenotype values) and 4772 (*qq*, YD_SCS: 0.4356; position 6/117 in the ranked list of *q*-daughters by the highest phenotype values), hereafter referred to as QQ6_5594 and qq6_4772. A detailed characterization at the phenotypic and genotypic levels for the animals selected for WGSeq analysis is provided in Table 3.

Variant Survey Performed in the OAR20 QTL Region and Concordance Tests

The whole-genome data sets showed an average number of raw reads per sample of 496,352,108 paired reads. For the 6 sequenced samples, the sequencing depth across the genome ranged between 9.89× to 19.05×. In the OAR20 studied interval, the sequencing depth was very similar, ranging from 9.66× to 18.7× (average depth 15.25×). After the Trimmomatic quality trimming, ∼92% of the reads were aligned against the sheep reference genome (Oar_v3.1), and 99.40% of them could be mapped. Considering the TGI of the within-family analyses for family 1 (14.2–24.2 Mb) and family 6 (19.9–41.7 Mb), we performed a variant calling analysis in the OAR20 genomic interval 14.2 to 41.7 Mb. The number of variants identified across this whole genomic interval was 299,053 for Samtools and 301,954 by GATK. After applying the Snpshift filters, the number of genetic variants identified across that specific region were 279,048 with GATK and 268,568 with Samtools. A total of 227,030 variants were commonly identified by the 2 methods (so-called high-quality variants) across the whole considered interval. Next, various concordance tests were performed based on the genotypes of the 6 sequenced samples.

For the 2 sequenced trios, the initial concordance tests performed considered the TGI defined based on the CI estimated in the within-family analysis for family 1 (Trio 1-fam1) and family 6 (Trio2-fam6). We termed the one of the 2 concordance tests "Trio1-fam1-Region1," which examined the QTL concordance pattern in the 3 samples included in Trio 1 (Sire1_1444, QQ1_5481 and qq1_4404) across the region 14.2–24.2 Mb of OAR20, and the other "Trio2-fam6-Region2," which selected the concordant variants identified in the samples of Trio 2 (Sire6_2406, QQ6_5594, and qq6_4772) across the region 19.9 to 41.7 Mb (Table 4).

From the initial list of high-quality variants identified within the defined "Region1" and "Region2" (70.541 and 195.983 variants, respectively), the total number of concordant mutations identified in each of these concordance tests was 6,968 and 40,870 genetic variants, respectively. The annotation of these variants identified 1,646 and 17,897 intragenic variants (including SNP and Indels) for "Trio1-fam1-Region1" and "Trio2-fam6-Region2," respectively, resulting in 2,275 and 22,701 functional annotation variants, respectively (Supplemental Tables S7 and S8; https://doi.org/10.3168/jds.2018-14736).

Table 3. Characterization of the animals selected for whole-genome sequencing (WGSeq) analysis in relation to the high-resolution study of the ovine chromosome 20 (OAR20) QTL[1]

Family selected for WGSeq Sire identification (ID) Peak position (correspondent marker positions in the OAR20 map (total: 919 markers) Within-family CI	Animals selected for WGSeq	YD_SCS value (ranked position within family)	Marker flanking the homozygosity region with expected QTL phase at within-family-QTL peak position			Homozygosity region with expected phase (Mb)
			Marker position in the OAR20 map (total: 919 markers)	Marker 1 ID	Marker 2 ID	Based on the position of flanking markers
Trio 1–family 1 Sire 1_1444 Peak: 18 cM (position markers 234–343) CI: 14.2–24.2 cM	QQ1_5481	YD_SCS: −0.13457; position 21/89 lowest SCS value	89–373	OAR20_5752490.1	OAR20_21369424_X.1	5.749–20.377
	qq1_4404	YD_SCS: 0.25914; position 8/80 highest SCS value	335–562	OAR20_19423010.1	OAR20_34436224.1	18.544–31.339
Trio 2–family 6 Sire6_2406 Peak: 38 cM (position markers 682–683) CI: 19.9–41.7 cM	QQ6_5594	YD_SCS: −0.342; position 3/147 lowest SCS value	664–698	OAR20_40664275.1 0,37218	OAR20_42333924.1 0,388584	27.218–38.858
	qq6_4772	YD_SCS: 0.4356; position 6/117 highest SCS value	45–791	s03770.1	OAR20_46931094.1	3.502–43.238

[1]Animals from 2 segregating families were considered for the high-definition analysis. The selected animals showed extreme divergent values for the SCS trait in concordance with the alternative QTL inherited from the segregating sire (low ranked SCS value position for the daughter inheriting the *Q* allele associated with decreased SCS value; high-ranked SCS value position for the daughter inheriting the *q* allele associated with increased SCS value). In addition, the selection focused on the animals showing the corresponding QTL allele in homozygosity across the region considered as CI based on the corresponding within-family analysis. The markers flanking the homozygosity region and their position in the map for each animal around the target QTL are indicated.

Considering the possible bias that within-family LA may introduce in the estimation of QTL position, we defined a third concordance test focused on the significant interval defined by LDLA (22–28 Mb). This interval shows a 1.8 Mb of coincidence with the considered "Region1" and is completed included within the interval defined as "Region2." To confirm the samples that should be considered for this concordance test, we first confirmed the phase of the daughters of the 2 trios around the QTL peak based on LDLA (23.76 Mb). We observed that daughter QQ1_5481 from Trio 1 did not carry the expected chromosomal sire phase at this position (it showed the *q* phase instead). Furthermore, daughter QQ6_5594 showed the right *Q* phase from its sire, but its homozygous region was very short and just coincident with the LDLA peak (23.476–23.561). Within the LDLA significant interval (22–28 Mb), the 2 susceptible daughters from both trios were homozygous for the corresponding *q* paternal phase (see Table 3). Hence, the third concordance test, called "Trio1&2-Region3" considered 4 samples: the 2 segregating sires, Sire1_1444 and Sire6_2406, and the 2 susceptible daughters, qq1_4404 and qq6_4772. In this region, a total of 2,206 intragenic variants were concordant with the QTL segregation pattern and caused 2,620 annotation variants (Supplemental Table S9; https://doi.org/10.3168/jds.2018-14736).

From the 3 lists of concordant intragenic annotation variants (Supplemental Tables S7, S8, and S9; https://doi.org/10.3168/jds.2018-14736), those variants classified by the eVEP software as "moderate" or "high_impact" in affecting protein function were selected for further assessment (Supplemental Table S10; https://doi.org/10.3168/jds.2018-14736). For the mutations included in this list for which a gene symbol was not available from the sheep reference genome, we searched for orthologous genes in reference species (*Homo sapiens*, *Bos taurus*, *Mus musculus*) and considered those showing the highest *Query%id* parameter values, with a minimum of 60. In the list of concordant intragenic variants of medium/high impact presented in Supplemental Table S10, we highlight those causing a deleterious mutation based on the SIFT software results (highlighted in red font) and those included within genes considered as immune-related genes (highlighted in yellow cell background) based on our reference list. Among the 12 missense concordant mutations identified from the "Trio1-fam1-Region1" concordance test, all of them were of moderate impact, 2 were deleterious (within the *PGC* and *GUCA1B* genes), and 2 were within immune-related genes, *MDFI* and *CCND3* (Supplemental Table S10). For the concordance test "Trio2-fam6-Region2," apart from 2 high-impact variants (1 stop_gained, 1 frameshift_variant) identified within olfactory receptor genes (*ENSOARG00000009186* and *ENSOARG00000017255*), 191 missense variants of moderate impact were present, 38 of which were classified as deleterious. Among the 73 genes carrying missense mutations in this region, we found 16 genes related to the immune system, especially related to the MHC, directly or through their corresponding orthologous genes [*AGER*, *BTN1A1*, *BTNL2*, *JARID2*, *MOG*, *NOTCH4*, *PKHD1*, *PLA2G7*, *ENSOARG00000015646* (*HLA-DRB1*; *H. sapiens*), *ENSOARG00000015707* (*HLA-DQA1*; *H. sapiens*), *ENSOARG00000001254* (*TNXB*; *B. taurus*), *ENSOARG00000009395* (*HLA-C*; *H. sapiens*), *ENSOARG00000009868* (*MIC1*; *B. taurus*), *ENSOARG00000010572* (*BoLA*; *B. taurus*), *ENSOARG00000015002* (*HLA-B*; *H. sapiens*), *ENSOARG00000018075* (*BoLA*; *B. taurus*)]. From the deleterious mutations identified within this region, 10 were within some of the immune-linked genes previously mentioned: *PKHD1*, *NOTCH4*, *AGER*, *MOG*, and annotated genes orthologous to the human *HLA-A* and *HLA-B* and bovine *BoLA* genes (see Table 5).

Among the concordant variants identified based on the concordance test "Trio1&2-Region3," 2 stop_gained variants were localized in the novel gene *ENSOARG00000015427*, which has no orthologous genes in other genomes (Supplemental Table S10), and 61 missense variants, 16 of which were included in immune-related genes (*CDSN*, *GNL1*, *MDC1*, *PKHD1*, *RHAG*, *MIC1*, and *TNXB*; Supplemental Table S10). Four of the missense variants within region 3 were classified as deleterious, 2 of them in the immune-related *PKHD1* gene (*rs404196749* and *rs412600919*), 1 in *ENSOARG00000010006* (orthologous to the bovine *POU5F1* gene), and 1 in the novel gene *ENSOARG00000015427*.

Table 4. Concordance tests performed, with a summary of the samples and region considered in the 3 concordance tests performed in the present study

Concordance test	Samples considered	Region considered (Mb)
Trio1-fam1–Region1	Sire 1_1444 QQ1_5481 qq1_4404	14.2–24.2
Trio2-fam6–Region2	Sire6_2406 QQ6_5594 qq6_4772	19.9–41.7
Trio 1&2–Region3	Sire 1_1444 Sire6_2406 qq1_4404 qq6_4772	22–28

Table 5. List of missense deleterious mutations localized in the ovine chromosome 20 (OAR20) QTL regions that were included in immune-related genes and were concordant with the concordance tests performed in the present study

Concordance test	Variant class[1]	Chr	Post (bp)	dbSNP_ID	Ref	Alt	Gene symbol	Orthologous gene[2]	Functional effect (Oar_v3.1): ensemblVEP (SIFT)[3]	Ensembl gene identification	Exon in gene	Strand	Amino acid substitution	Codon change
Trio1–Region1														
Trio2 Region2	SNV	20	24068465	rs411503520	T	C	*PKHD1*		Moderate (deleterious; 0.01)	ENSOARG00000013288	51/69	−1	T/A	Acc/Gcc
	SNV	20	24143601	rs404196749	A	C	*PKHD1*		Moderate (deleterious; 0.02)	ENSOARG00000013288	39/69	−1	F/V	Ttc/Gtc
	SNV	20	24229321	rs603559656	G	A	*PKHD1*		Moderate (deleterious; 0.04)	ENSOARG00000013288	34/69	−1	T/M	aCg/aTg
	SNV	20	24272122	rs412600919	A	C	*PKHD1*		Moderate (deleterious; 0.02)	ENSOARG00000013288	10/69	−1	L/R	cTg/cGg
	SNV	20	26323922	rs162200998	C	A	*NOTCH4*		Moderate (deleterious; 0.01)	ENSOARG00000017704	29/33	1	P/H	cCt/cAt
	SNV	20	26336565	rs598146085	G	A	*AGER*		Moderate (deleterious; 0.01)	ENSOARG00000000355	02/11	1	A/T	Gcc/Acc
	SNV	20	26989615	—	G	C		*HLA-C* (*Homo sapiens*) 66.92%	Moderate (deleterious low confidence; 0.01)	ENSOARG00000009395	01/05	1	G/R	Gga/Cga
	SNV	20	27767118	—	C	A		*HLA-B* (*H. sapiens*) 69.02%	Moderate (deleterious low confidence; 0.01)	ENSOARG00000015002	06/08	−1	K/N	aaG/aaT
	SNV	20	28008206	rs161332389	G	A	*MOG*		Moderate (deleterious low confidence; 0.03)	ENSOARG00000016686	01/08	−1	P/L	cCg/cTg
	SNV	20	29598185	rs404652506	A	G		*BoLA* (*Bos taurus*) 62.61%	Moderate (deleterious low confidence; 0.02)	ENSOARG00000018075	04/08	−1	S/P	Tcc/Ccc
Trio1&2 Region3	SNV	20	24143601	rs404196749	A	C	*PKHD1*		Moderate (deleterious; 0.02)	ENSOARG00000013288	39/69	−1	F/V	Ttc/Gtc
	SNV	20	24272122	rs412600919	A	C	*PKHD1*		Moderate (deleterious; 0.02)	ENSOARG00000013288	12/69	−1	L/R	cTg/cGg

[1]Ensembl variation classification. SNV: single nucleotide variant, equivalent to SNP.

[2]For those genes without associated gene symbol, orthologous genes in the genomes of 3 reference species (*B. taurus*, *H. sapiens*, *Mus musculus*) were searched, and considered those showing the highest Query%id parameter value with a minimum of 60.

[3]Functional effect of the variant on the protein function predicted by the eVEP software (ensembl Variant Effect Predictor; for further information about the column field names see http://www.ensembl.org/info/docs/tools/vep/vep_formats.html). In parentheses, the functional prediction obtained with the SIFT software and the related prediction score is also indicated).

DISCUSSION

The present study mapped QTL underlying SCS, a classical indicator trait of mastitis resistance, in Spanish Churra dairy sheep. The most significant QTL, detected at the genome-wide level by LA (20.9–23.8 Mb), and confirmed by LDLA (22–28 Mb), was located on OAR20. This QTL was previously identified through a low-density microsatellite-based genome scan in a different commercial population of Churra sheep (Gutiérrez-Gil et al., 2007). Hence, the present study has replicated for this breed the presence of a QTL for mastitis resistance on OAR20. Whereas the initial scan presented by Gutiérrez-Gil et al. (2007) only identified the OAR20 QTL effect in the present study by using a different population and benefitting from increased marker accuracy, we identified 2 additional QTL at the chromosome-wide level by LA, on OAR5 and OAR25, together with 22 significant associations identified by LDLA, 2 of them replicating the LA-detected effects on OAR20 and OAR25. In addition to the classical LA analysis method, which is appropriate for the half-sib population analyzed here, the LDLA approach can be used to complete the global picture of segregating effects because the effects captured by these 2 analysis methods pinpoint associations with different features. For example, whereas in our design LA will only detect QTL if several sires are heterozygous at the same QTL (*Qq*), many marker trait associations that do not satisfy this assumption but have a genuine association at the population level can be detected by LDLA. This explains the larger number of significant associations identified by LDLA (22) than by LA (3) in the present study, similar to previous studies performed in the same commercial population regarding milk production or parasite resistance traits (Garcia-Gámez et al., 2013; Atlija et al., 2016). In any case, the within-family information provided by LA in a half-sib population like the one studied here is of major value to select the families that should be considered from the global population in further fine-mapping, high-resolution studies aiming at the identification of the causal mutation. Hence, the selection of trios for WGSeq presented in this work relies completely on the within-family information provided by the LA genome scan summarized in Table 1.

When examining the correspondence between the TGI defined by our analyses and QTL previously described in sheep (Supplemental Table S1; https://doi.org/10.3168/jds.2018-14736), several coincidences with general health traits are apparent, although only in one case (in relation to the OAR22 QTL) does the previously reported effect influence the SCS trait (Raadsma et al., 2009). Interestingly, we did not find any QTL in the region of the *SOCS2* gene (OAR3: 129,720,516–129,722,508 according Oar_v3.1), where a causal mutation for mastitis resistance has been reported in a French dairy sheep population (Rupp et al., 2015). A specific evaluation of the *SOCS2_p.R96C* mutation in the males of the Churra population studied in the present work showed that all of them where homozygous for the C base present in the reference sequence, and none of them were carriers of the T allele observed in susceptible animals in the French population.

Apart from the initial description in Churra sheep, no other study has reported a marker association with sheep mastitis resistance in the region suggested to harbor the OAR20 QTL by LA and LDLA (20.9–28 Mb), although this region does harbor QTL for other health-related traits (IgA level, *Haemonchus contortus* FEC, Maedi-Visna virus, ovine pulmonary adenocarcinoma, *Salmonella abortus ovis* susceptibilities). The low number of correspondences observed between the significant associations reported here with ovine QTL for udder traits may be explained by the limited number of QTL mapping studies on this species. Nevertheless, although our systematic search with the Search tool of SheepQTLdb did not report a direct overlapping with the QTL reported for udder morphology in Churra sheep (Gutiérrez-Gil et al., 2008), it should be taken into account that the flanking intervals of 2 QTL for udder depth reported in this breed would be close to the LDLA significant associations reported here on OAR14 and OAR20 (SheepQTLdb QTL ID: 13658 and 13660, respectively; flanking intervals: 53.5–60.2 Mb and 7.2 Mb, respectively).

The correspondence of the OAR20 QTL with several cattle QTL related to mastitis resistance described in the genomic region including the bovine MHC suggests that some genetic mechanisms controlling mastitis resistance might be shared between cattle and sheep. Several studies support the associations between MHC polymorphisms, especially in the *BoLA-DRB3.2* gene, on mastitis resistance in cattle (reviewed by Behl et al., 2012), whereas in sheep, few studies have focused on this potential relationship. Hereafter, the most important results from the initial QTL mapping analysis presented here are (1) the replication of the OAR20 QTL by using a medium density SNP chip and 2 different analysis approaches, LA and LDLA, and (2) the identification of a list of functional candidate genes related to immune traits included within the CI defined for the significant QTL identified by LA and LDLA (Supplemental Table S6; https://doi.org/10.3168/jds.2018-14736). Polymorphisms affecting these genes should be assessed as potential genetic markers in relation to increased SCS in dairy sheep. Comparing with results reported in goats, we did not find a relevant coincidence, as the sheep genomic regions on

OAR11 orthologous to the intervals harboring recently reported QTL on goat chromosome 19 for SCS and udder traits (Martin et al., 2018; Mucha et al., 2018; 26–28 Mb and 40–42 Mb) are not coincident with the significant LDLA association reported here on OAR11 (56.8–56.9 Mb). The scarce number of goat studies on this topic may explain the lack of coincidences. In addition, the correspondence of the sheep associations here detected for SCS with QTL/associations reported in cattle for udder traits, especially for teat length and udder depth, support the phenotypic and genetic correlations estimated between SCS and udder traits and the use of indirect selection based on udder traits to improve global mammary gland health in sheep.

In the second section of our study, we used next-generation sequencing to obtain a deeper insight into the OAR20 QTL region. Genome-wide association studies based on sequence data have shown high power to identify putative causative variants and stronger signals of association (Daetwyler et al., 2014; Höglund et al., 2014; Sahana et al., 2014). Both the LA and LDLA statistical profiles showed indications of more than one QTL peak, although they both showed the maximum significant values in close positions (21.52 and 23.52 cM for LA and LDLA, respectively). Because the sequencing analysis is based on segregating trios, we focused the analysis of genetic variability within the specific CI estimated for the 2 families showing the highest support for the OAR20 QTL, family 1 and family 6, although because the statistical profile of family 6 suggests that this family could be segregating for more than one QTL, we screened 2 different and partially overlapping regions of the chromosome (14.2–24.2 Mb and 19.9–41.7 Mb). The within-family discrepancies regarding the QTL peak position show a complex case of gene fine-mapping and the importance of performing a detailed study of the phase status in the considered animals. Based on our study of the phase status, we determined that the significant interval highlighted by the LDLA analysis could only be considered for a concordance test involving the 2 segregating sires and the 2 susceptible daughters, qq1_4404 and qq6_4472. Hence, considering the difficulties of leading with segregating families showing discrepancies in LA-estimated CI, we considered it appropriate to run the 3 described different concordance tests to filter the genetic variability identified across the considered intervals.

Within the 3 different intervals considered in these concordance tests and after applying the genotype concordance filtering considering the QTL segregation pattern, we focused on the genes that harbored intragenic variability and that showed functional effects on the encoded protein function, mainly missense variants. We acknowledge that noncoding or intergenic regions were not considered in the later steps of assessment of potential causal mutations. Although different studies have shown that noncoding variants may directly affect gene expression and protein abundance (Zappala and Montgomery, 2016; Igartua et al., 2017), and therefore represent impactful variation, we considered that focusing on the coding region was an appropriate initial approach to evaluate the high density variability study performed on this work. Among the genes harboring the potential functional relevant variants, we highlighted those related to immunity based on our reference candidate gene list (Supplemental Table S10, https://doi.org/10.3168/jds.2018-14736; a total of 85 variants highlighted in yellow). We recognize that any of these 85 missense variants concordant with the QTL segregation pattern (considering the 3 tests) and within a list of 23 immune-related genes (*CCND3*, *MDFI*, *PLA2G7*, *PKHD1*, *HLA-DRB1*, *HLA-DQA1*, *DQA*, *BTNL2*, *NOTCH4*, *AGER*, *TNXB*, *HLA-C*, *HLA-B*, *MIC1*, *CDSN*, *BoLA*, *MDC1*, *MOG*, *BTN1A1*, *JARID2*, *RHAG*, *MIC1*, and *GNL1*) could be considered as candidate variants to explain the studied OAR20 QTL effect. Among the 23 harboring genes of these mutations, obvious candidates are included in the ovine MHC (*HLA-DRB1*, *HLA-DQA1*, *DQA*, *HLa-B*, *HLA-C*, and *BoLA-HLA-A*). In cattle, bovine lymphocyte antigen (**BoLA**) has been associated with resistance to mastitis (Mallard et al., 1995; Rupp et al., 2007) and SCC (Sharif et al., 1998; Chu et al., 2012), with several studies supporting the direct association of these traits with the *DRB3* locus. In addition, in Holstein cattle, heterozygosity of the *BoLA-DQA1* gene has been associated with resistance to mastitis progression (Takeshima et al., 2008). The search for associations between the remaining filtered candidate genes and mastitis did not yield any notable results.

In an attempt to obtain further information from the functional annotation analysis, the list of 85 missense QTL concordant mutations within immune-related genes was further filtered to extract those mutations showing deleterious effects on protein function. We consider that among the listed variants, priority should be given to the 10 concordant variants identified through concordance test "Trio2-family6-Region2," which are distributed across 7 different genes [*PKHD1*, *NOTCH4*, *AGER*, *ENSOARG00000009395* (*HLA-C*, *H. sapiens*), *ENSOARG00000015002* (*HLA-B*, *H. sapiens*), *MOG*, and *ENSOARG00000018075* (*BoLA*, *B. taurus*, *orthologous to human HLA-A*)]. Two of these mutations, located on *PKHD1*, were also concordant with the concordance test "Trio1&2-Region3" (Table 5). However, the known biological function of the *PKHD1* gene, which encodes for the fibrocystin protein involved in the polycystic kidney and hepatic disease-1, and the lack of significant

expression of this gene in the mammary gland (Menezes et al., 2004) excluded this gene as a relevant candidate for our study. To confirm the expression of the remaining immune genes harboring deleterious variants in the sheep mammary gland, we estimated their expression levels from RNA-Seq data previously analyzed by our research group to study the dynamic transcriptome (at 4 time points, i.e., d 10, 50, 120, and 150 after lambing) of the sheep mammary gland (Suárez-Vega et al., 2015; see Supplemental Table S11; https://doi.org/10.3168/jds.2018-14736). This data set confirmed that, among the 7 considered genes, *NOTCH4* showed an average low level of gene expression (1.964 FPKM) across the studied time points and samples. In contrast, *ENSOARG00000009395* and *ENSOARG00000015002* (orthologous to the human *HLA-C* and *HLA-B* genes, respectively) showed moderate gene expression levels (46.914 and 389.569 average FPKM values, respectively). *HLA-B* showed the greatest variation, from 79.685 to 2017.630 FPKM Supplemental Table S11).

The *NOTCH4* gene encodes a member of the NOTCH family of proteins, which are transmembrane receptors that interact with membrane-bound ligands encoded by the Delta/Serrate/Jagged gene families. The Notch signaling pathway is an evolutionarily conserved intercellular signaling mechanism (reviewed by Callahan and Egan, 2004). A role of this protein has been suggested in relation to mammary gland development and mammary tumorigenesis, but no previous reports are available in relation to inflammation of the mammary gland. Regarding the proteins encoded by the *HLA-B* and *HLA-C* genes, these are class I MHC molecules whose function is to display peptide fragments of non-self-proteins from within the cell to cytotoxic T cells. The proteins encoded by these genes have been related to many disease traits, primarily with regard to autoimmune diseases and also with regard to human immunodeficiency virus (HIV) control or increased risk of developing Crohn's disease (Colmegna et al., 2004; Vince et al., 2016).

CONCLUSIONS

This gene mapping study for SCS in dairy sheep replicated a previously reported QTL for this trait on OAR20. Additionally, the use of a medium-high SNP chip and 2 complementary analytical approaches (LA and LDLA) allowed the identification of other genomic regions that might be related to ewe susceptibility to subclinical mastitis. For all of these regions, we present a list of positional and functional candidate genes that should be evaluated by future studies. In addition, for the replicated OAR20 QTL, we exploited the whole-genome sequencing of segregating trios to fully characterize the putative harboring genomic region. By performing high-resolution analyses and consecutive filtering strategies, the study provides a list of promising candidate genes and genetic variants that underlie the targeted QTL effect. Future studies should address the genotyping of the suggested variants in commercial populations and evaluate their associations with indicator traits of udder inflammation (SCS, clinical mastitis). Functional characterization experiments should then be designed to understand the biological mechanism controlling a fraction of the genetic variation that controls the complex trait of resistance to mastitis in sheep.

ACKNOWLEDGMENTS

This research work was partially funded by the 3SR (Sustainable Solutions for Small Ruminants) project funded by the European Commission within the FP7 Programme (FP7-KBBE245140) and by the AGL2015-66035-R project, funded by the Spanish Ministry of Economy and Competitiveness (MINECO, Madrid, Spain) and co-funded by the European Regional Development Fund. We thank the National Association of Spanish Churra Breeders (ANCHE) for the close collaboration with our research group and the support for generating sequencing data of Churra genomes. B. Gutiérrez-Gil is funded through the Spanish "Ramón y Cajal" Programme (RYC-2012-10230) from MINECO. The support and availability to the computing facilities of the Foundation of Supercomputing Center of Castile and León (FCSCL, León, Spain; http://www.fcsc.es) are greatly acknowledged. C. Esteban-Blanco is funded by an FPU contract from MINECO (Ref. BES-2016-07-8080).

REFERENCES

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. 2010. Data quality control in genetic case-control association studies. Nat. Protoc. 5:1564–1573. https://doi.org/10.1038/nprot.2010.116.

Arranz, J. J., and B. Gutiérrez-Gil. 2012. Detection of QTL Underlying Milk Traits in Sheep: An Update. N. Chaiyab, ed. InTech, New York, NY.

Atlija, M., J.-J. Arranz, M. Martinez-Valladares, and B. Gutiérrez-Gil. 2016. Detection and replication of QTL underlying resistance to gastrointestinal nematodes in adult sheep using the ovine 50K SNP array. Genet. Sel. Evol. 48:4. https://doi.org/10.1186/s12711-016-0182-4.

Banos, G., G. Bramis, S. J. Bush, E. L. Clark, M. E. B. McCulloch, J. Smith, G. Schulze, G. Arsenos, D. A. Hume, and A. Psifidi. 2017. The genomic architecture of mastitis resistance in dairy sheep. BMC Genomics 18:624. https://doi.org/10.1186/s12864-017-3982-1.

Barillet, F., J. M. Astruc, and G. Lagriffoul. 2006. Taking into account functional traits in dairy sheep breeding programs through the French example. Vol 121, pages 57–64 in Proceeding of the EAAP. Wageningen, NLD. Academic Publishers., Kuopio, Finland.

Behl, J. D., N. K. Verma, N. Tyagi, P. Mishra, R. Behl, and B. K. Joshi. 2012. The major histocompatibility complex in bovines: A review. ISRN Vet. Sci. 2012:872710. https://doi.org/10.5402/2012/872710.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Broad Institute. 2017. Picard Tool, Version 1.128. Accessed Dec. 15, 2017. http://broadinstitute.github.io/picard/.

Broman, K., and S. Sen. 2009. A Guide to QTL Mapping with R. Springer, New York, NY.

Callahan, R., and S. E. Egan. 2004. Notch signaling in mammary development and oncogenesis. J. Mammary Gland Biol. Neoplasia 9:145–163. https://doi.org/10.1023/B:JOMG.0000037159.63644.81.

Casu, S., S. Sechi, S. L. Salaris, and A. Carta. 2010. Phenotypic and genetic relationships between udder morphology and udder health in dairy ewes. Small Rumin. Res. 88:77–83. https://doi.org/10.1016/J.SMALLRUMRES.2009.12.013.

Chu, M. X., S. C. Ye, L. Qiao, J. X. Wang, T. Feng, D. W. Huang, G. L. Cao, R. Di, L. Fang, and G. H. Chen. 2012. Polymorphism of exon 2 of BoLA-DRB3 gene and its relationship with somatic cell score in Beijing Holstein cows. Mol. Biol. Rep. 39:2909–2914. https://doi.org/10.1007/s11033-011-1052-3.

Cingolani, P., V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, and X. Lu. 2012. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front. Genet. 3:35. https://doi.org/10.3389/fgene.2012.00035.

Colmegna, I., R. Cuchacovich, and L. R. Espinoza. 2004. HLA-B27-associated reactive arthritis: Pathogenetic and clinical considerations. Clin. Microbiol. Rev. 17:348–369. https://doi.org/10.1128/CMR.17.2.348-369.2004.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865. https://doi.org/10.1038/ng.3034.

de la Fuente, L. F., G. Fernandez, and F. San Primitivo. 1996. A linear evaluation system for udder traits of dairy ewes. Livest. Prod. Sci. 45:171–178. https://doi.org/10.1016/0301-6226(96)00003-6.

Fernández, G., J. A. Baro, L. F. de la Fuente, and F. San Primitivo. 1997. Genetic parameters for linear udder traits of dairy ewes. J. Dairy Sci. 80:601–605. https://doi.org/10.3168/jds.S0022-0302(97)75976-9.

Filangi, O., C. Moreno, H. Gilbert, A. Legarra, P. Le-Roy, and J. M. Elsen. 2010. QTLMap, a software for QTL detection in outbreed populations. Pages 1–3 in 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany. Leipzig Gesellschaft für Tierzuchtwissenschaften 2010. Leipzig, Germany.

García-Gámez, E., B. Gutiérrez-Gil, G. Sahana, J.-P. Sánchez, Y. Bayón, and J.-J. Arranz. 2012. GWA Analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the LALBA gene. PLoS One 7:e47782.

Garcia-Gámez, E., B. Gutiérrez-Gil, A. Suarez-Vega, L. F. de la Fuente, and J. J. Arranz. 2013. Identification of quantitative trait loci underlying milk traits in Spanish dairy sheep using linkage plus combined linkage disequilibrium and linkage analysis approaches. J. Dairy Sci. 96:6059–6069. https://doi.org/10.3168/jds.2013-6824.

Gonzalo, C., and V. R. Gaudioso. 1985. Evolution des types cellulaires du lait de brebis (race Churra) en fonction des dénombrements cellulaires totaux pendant la traite mécanique et manuelle. Ann. Zootech. 34:257–264.

Govignon-Gion, A., R. Dassonneville, G. Baloche, and V. Ducrocq. 2016. Multiple trait genetic evaluation of clinical mastitis in three dairy cattle breeds. Animal 10:558–565. https://doi.org/10.1017/S1751731115002529.

Gutiérrez-Gil, B., M. F. El-Zarei, L. Alvarez, Y. Bayón, L. F. de la Fuente, F. San Primitivo, and J. J. Arranz. 2008. Quantitative trait loci underlying udder morphology traits in dairy sheep. J. Dairy Sci. 91:3672–3681. https://doi.org/10.3168/jds.2008-1111.

Gutiérrez-Gil, B., M. F. El-Zarei, Y. Bayón, L. Alvarez, L. F. de la Fuente, F. San Primitivo, and J. J. Arranz. 2007. Short communication: Detection of quantitative trait loci influencing somatic cell score in Spanish Churra sheep. J. Dairy Sci. 90:422–426.

Gutiérrez-Gil, B., C. Esteban-Blanco, P. Wiener, P. K. Chitneedi, A. Suarez-Vega, and J.-J. Arranz. 2017. High-resolution analysis of selection sweeps identified between fine-wool Merino and coarse-wool Churra sheep breeds. Genet. Sel. Evol. 49:81. https://doi.org/10.1186/s12711-017-0354-x.

Höglund, J. K., G. Sahana, R. Brøndum, B. Guldbrandtsen, B. Buitenhuis, and M. S. Lund. 2014. Fine mapping QTL for female fertility on BTA04 and BTA13 in dairy cattle using HD SNP and sequence data. BMC Genomics 15:790. https://doi.org/10.1186/1471-2164-15-790.

Hu, Z.-L., C. A. Park, X.-L. Wu, and J. M. Reecy. 2013. Animal QTLdb: An improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41:D871–D879. https://doi.org/10.1093/nar/gks1150.

Igartua, C., S. V. Mozaffari, D. L. Nicolae, and C. Ober. 2017. Rare non-coding variants are associated with plasma lipid traits in a founder population. Sci. Rep. 7:16415. https://doi.org/10.1038/s41598-017-16550-8.

Jamrozik, J., A. Koeck, F. Miglior, G. J. Kistemaker, F. S. Schenkel, D. F. Kelton, and B. J. Van Doormaal. 2013. Genetic and genomic evaluation of mastitis resistance in Canada. Interbull Bull. 0:43–51.

Kelley, J., B. de Bono, and J. Trowsdale. 2005. IRIS: A database surveying known human immune system genes. Genomics 85:503–511. https://doi.org/10.1016/j.ygeno.2005.01.009.

Lander, E. S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Le Roy, P., J. M. Elsen, H. Gilbert, C. Moreno, A. Legarra, and O. Filangi. 2013. QTLMap 0.9.6 User's guide 1–56. Accessed Nov. 21, 2016. https://forge-dga.jouy.inra.fr/attachments/download/2502/qtlmapV0.9.6.pdf.

Legarra, A., and E. Ugarte. 2005. Genetic parameters of udder traits, somatic cell score, and milk yield in Latxa sheep. J. Dairy Sci. 88:2238–2245. https://doi.org/10.3168/jds.S0022-0302(05)72899-X.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Mallard, B. A., K. E. Leslie, J. C. M. Dekkers, R. Hedge, M. Bauman, and M. J. Stear. 1995. Differences in bovine lymphocyte antigen associations between immune responsiveness and risk of disease following intramammary infection with *Staphylococcus aureus*. J. Dairy Sci. 78:1937–1944. https://doi.org/10.3168/jds.S0022-0302(95)76819-9.

Martin, P., I. Palhière, C. Maroteau, V. Clément, I. David, G. T. Klopp, and R. Rupp. 2018. Genome-wide association mapping for type and mammary health traits in French dairy goats identifies a pleiotropic region on chromosome 19 in the Saanen breed. J. Dairy Sci. https://doi.org/10.3168/jds.2017-13625.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants

with the Ensembl API and SNP effect predictor. Bioinformatics 26:2069–2070. https://doi.org/10.1093/bioinformatics/btq330.

Menezes, L. F. C., Y. Cai, Y. Nagasawa, A. M. G. Silva, M. L. Watkins, A. M. Da Silva, S. Somlo, L. M. Guay-Woodford, G. G. Germino, and L. F. Onuchic. 2004. Polyductin, the PKHD1 gene product, comprises isoforms expressed in plasma membrane, primary cilium, and cytoplasm. Kidney Int. 66:1345–1355. https://doi.org/10.1111/j.1523-1755.2004.00844.x.

Miglior, F., B. L. Muir, and B. J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 88:1255–1263. https://doi.org/10.3168/jds.S0022-0302(05)72792-2.

Mucha, S., R. Mrode, M. Coffey, M. Kizilaslan, S. Desire, and J. Conington. 2018. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. J. Dairy Sci. 101:2213–2225. https://doi.org/10.3168/jds.2017-12919.

Narasimhan, V., P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin. 2016. BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 32:1749–1751. https://doi.org/10.1093/bioinformatics/btw044.

Østerås, O., H. Solbu, A. O. Refsdal, T. Roalkvam, O. Filseth, and A. Minsaas. 2007. Results and evaluation of thirty years of health recordings in the Norwegian dairy cattle population. J. Dairy Sci. 90:4483–4497. https://doi.org/10.3168/jds.2007-0030.

Othmane, M. H., L. F. De La Fuente, J. A. Carriedo, and F. San Primitivo. 2002. Heritability and genetic correlations of test day milk yield and composition, individual laboratory cheese yield, and somatic cell count for dairy ewes. J. Dairy Sci. 85:2692–2698. https://doi.org/10.3168/jds.S0022-0302(02)74355-5.

Raadsma, H. W., E. Jonas, D. McGill, M. Hobbs, M. K. Lam, and P. C. Thomson. 2009. Mapping quantitative trait loci (QTL) in sheep. II. Meta-assembly and identification of novel QTL for milk production traits in sheep. Genet. Sel. Evol. 41:45. https://doi.org/10.1186/1297-9686-41-45.

Raynal-Ljutovac, K., A. Pirisi, R. de Crémoux, and C. Gonzalo. 2007. Somatic cells of goat and sheep milk: Analytical, sanitary, productive and technological aspects. Small Rumin. Res. 68:126–144. https://doi.org/10.1016/j.smallrumres.2006.09.012.

Rupp, R., A. Hernandez, and B. A. Mallard. 2007. Association of bovine leukocyte antigen (BoLA) DRB3.2 with Immune response, mastitis, and production and type traits in Canadian Holsteins. J. Dairy Sci. 90:1029–1038. https://doi.org/10.3168/jds.S0022-0302(07)71589-8.

Rupp, R., G. Lagriffoul, J. M. Astruc, and F. Barillet. 2003. Genetic parameters for milk somatic cell scores and relationships with production traits in French Lacaune dairy sheep. J. Dairy Sci. 86:1476–1481. https://doi.org/10.3168/jds.S0022-0302(03)73732-1.

Rupp, R., P. Senin, J. Sarry, C. Allain, C. Tasca, L. Ligat, D. Portes, F. Woloszyn, O. Bouchez, G. Tabouret, M. Lebastard, C. Caubet, G. Foucras, and G. Tosser-Klopp. 2015. A point mutation in suppressor of cytokine signalling 2 (Socs2) increases the susceptibility to inflammation of the mammary gland while associated with higher body weight and size and higher milk production in a sheep model. PLoS Genet. 11:e1005629. https://doi.org/10.1371/journal.pgen.1005629.

Sahana, G., B. Guldbrandtsen, B. Thomsen, L.-E. Holm, F. Panitz, R. F. Brøndum, C. Bendixen, and M. S. Lund. 2014. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. J. Dairy Sci. 97:7258–7275. https://doi.org/10.3168/jds.2014-8141.

Sellner, E. M., J. W. Kim, M. C. McClure, K. H. Taylor, R. D. Schnabel, and J. F. Taylor. 2007. Board-invited review: Applications of genomic information in livestock. J. Anim. Sci. 85:3148–3158. https://doi.org/10.2527/jas.2007-0291.

Sharif, S., B. A. Mallard, B. N. Wilkie, J. M. Sargeant, H. M. Scott, J. C. Dekkers, and K. E. Leslie. 1998. Associations of the bovine major histocompatibility complex DRB3 (BoLA-DRB3) alleles with occurrence of disease and milk somatic cell score in Canadian dairy cattle. Anim. Genet. 29:185–193.

Shook, G. E., and M. M. Schutz. 1994. Selection on somatic cell score to improve resistance to mastitis in the United States. J. Dairy Sci. 77:648–658. https://doi.org/10.3168/jds.S0022-0302(94)76995-2.

Suárez-Vega, A., B. Gutiérrez-Gil, C. Klopp, C. Robert-Granie, G. Tosser-Klopp, and J. J. Arranz. 2015. Characterization and Comparative analysis of the milk transcriptome in two dairy sheep breeds using RNA sequencing. Sci. Rep. 5:18399. https://doi.org/10.1038/srep18399.

Takeshima, S., Y. Matsumoto, J. Chen, T. Yoshida, H. Mukoyama, and Y. Aida. 2008. Evidence for cattle major histocompatibility complex (BoLA) class II *DQA1* gene heterozygote advantage against clinical mastitis caused by *Streptococci* and *Escherichia* species. Tissue Antigens 72:525–531. https://doi.org/10.1111/j.1399-0039.2008.01140.x.

Vince, N., H. Li, V. Ramsuran, V. Naranbhai, F.-M. Duh, B. P. Fairfax, B. Saleh, J. C. Knight, S. K. Anderson, and M. Carrington. 2016. HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. Am. J. Hum. Genet. 99:1353–1358. https://doi.org/10.1016/j.ajhg.2016.09.023.

Zappala, Z., and S. B. Montgomery. 2016. Non-coding loss-of-function variation in human genomes. Hum. Hered. 81:78–87. https://doi.org/10.1159/000447453.

Resultado 2.1

Estudio preliminar sobre la caracterización del microbioma de la glándula mamaria en ovejas assaf en lactación

C. Esteban-Blanco[1], B. Gutiérrez-Gil[1], H. Marina-García[1], B. Linaje[2], A. Acedo[3] y J.J. Arranz[1]

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain: [2]Consorcio de Promoción del Ovino, Ctra. Zamora-Palencia, km. 49, 49630 Villalpando, Zamora; [3]Biome Makers Inc., West Sacramento, CA, USA.

XIX Reunión Nacional de Mejora Genética Animal, León 14 y 15 de junio de 2018.

ESTUDIO PRELIMINAR SOBRE LA CARACTERIZACIÓN DEL MICROBIOMA DE LA GÁNDULA MAMARIA EN OVEJAS ASSAF EN LACTACIÓN

Esteban-Blanco[1], C., Gutiérrez-Gil[1], B., Marina-García[1], H., Linaje[2], B., Acedo[3] A. y Arranz[1], J.J.

[1]Dpto. de Producción Animal, Facultad de Veterinaria, Universidad de León, 24071; León
[2]Consorcio de Promoción del Ovino, Ctra. Zamora-Palencia, km. 49, 49630 Villalpando, Zamora; [3]Biome Makers, Paseo de Belén, 9, 47011 Valladolid

INTRODUCCIÓN

La leche es un fluido biológico complejo producido por las hembras de los mamíferos, específico de la especie y adaptado para satisfacer las necesidades nutricionales del recién nacido. Clásicamente ha sido considerado un producto estéril hasta el momento de su secreción y se creía que las bacterias presentes en leche provenían de la piel de la madre o la cavidad oral del neonato (A. et al., 2018). Actualmente se sabe que la leche materna contiene varios nutrientes que ayudan a crear el microambiente adecuado para el desarrollo y la maduración intestinal (Oftedal, 2002) y educa al sistema inmune confiriendo cierto grado de protección contra patógenos (Morrow and Rangel, 2004). Recientemente, varios estudios han revelado que el calostro y la leche materna son fuentes continuas de bacterias comensales, mutualistas y potencialmente probióticas para el lactante (Martin et al., 2003). El hecho de que los mismos géneros de bacterias puedan aislarse de la leche materna de hembras diferentes sugiere que su presencia en este sustrato es un evento común, aunque también existen diferencias importantes de unos animales a otros (Hunt et al., 2011). Habitualmente la identificación de microorganismos se ha realizado mediante aislamiento y análisis de pruebas morfológicas. Estos estudios se consideran ilimitados ya que muchas bacterias no se pueden cultivar y ofrecen poca información de la diversidad microbiana. Por esta razón, en los últimos años, el desarrollo de técnicas de secuenciación de ADN de alto rendimiento se está convirtiendo en una de las herramientas más utilizadas para estudiar el impacto de la comunidad de microorganismos, la llamada microbiota, en la salud humana y animal (Stubbendieck et al., 2016). Concretamente las técnicas basadas en la secuenciación del gen que codifica para la fracción 16S del RNA ribosómico permite una evaluación bastante completa de la biodiversidad de la muestra analizada. La microbiota de la leche materna está formada por una comunidad dinámica en la que existe un equilibrio complejo entre mutualista, comensales y organismos patógenos y cuya alteración puede influir en la calidad de los productos animales e incluso en el desarrollo de una enfermedad. En el caso concreto de las infecciones de la glándula mamaria, o mastitis en el ganado lechero, la relevancia de diferentes patógenos se conoce desde hace mucho tiempo pero el impacto de la compleja comunidad de microorganismos y su interacción en el desarrollo de la infección ha sido descrita sólo recientemente (Addis et al., 2016). Debido a los pocos estudios sobre la microbiota de la glándula mamaria en el ganado doméstico, el presente trabajo presenta un estudio preliminar de la biodiversidad bacteriana de la microbiota de la glándula mamaria de la oveja en lactación, en base a datos de secuenciación del gen 16S, y examinamos las diferencias entre la microbiota de grupos de ovejas con distintos valores de recuento de células somáticas (SCC), carácter indicador del estado de salud de la ubre.

MATERIAL Y MÉTODOS

Recogida de muestras: Para el presente estudio se obtuvieron dos muestras de leche, de 50 ml cada una, de un total de 49 ovejas lecheras de raza Assaf, de una granja de la provincia de Zamora perteneciente al Consorcio para la Promoción del Ovino (CPO). Ninguno de los animales muestreados presentaba signos clínicos de mastitis. Las muestras fueron obtenidas antes del ordeño de la mañana siguiendo la recomendación estándar del Consejo Nacional de Mastitis (Hogan, 1999); es decir, tras desinfectar las puntas de los pezones con alcohol etílico al 70%, descartando el primer chorro. Una de las muestras recogidas de cada animal se utilizó para obtener el recuento de células somáticas (SCC) y la otra para extracción del DNA. El SCC se utilizó para diferenciar las ovejas sanas de las que podrían presentar mastitis subclínica, siguiendo el umbral de 400.000 células/ml sugerido para las razas Assaf y Castellana por Gonzalez-Rodriguez et al. (1995).

Extracción de ADN, amplificación de 16S rRNA y secuenciación de alto rendimiento: Después de eliminar la grasa de la leche, las muestras se centrifugaron durante 10 minutos a 15000g

y el *pellet* obtenido se resuspendió en 0,5 ml de PBS a 4°C. El ADN se purificó mediante el uso de *Dneasy Powerlyzer powersoil kit (Qiagen)*. La amplificación de la región 16S V4 del genoma bacteriano se realizó utilizando primers de BiomeMakers® (Patente WO2017096385). La secuenciación 2 x 301bp paired-end utilizando el *Illumina MiSeq* (Illumina, San Diego, CA, USA) generó un promedio de 150.000 lecturas por muestra.

Análisis bioinformático y estadístico: Los datos de secuenciación fueron ensamblados usando FLASH (Magoč and Salzberg, 2011), combinando pares de lectura en la orientación "outie" con un solapamiento mínimo de 300pb. Se llevó a cabo un filtrado y recorte de los datos con prinseq-lite (Schmieder and Edwards, 2011). En base a la valoración de la calidad de las muestras realizada con FastQC (Andrews, 2010) se recortaron las 8 primeras bases de cada secuencia. El formato fastq a formato fasta, necesario para los pasos de eliminación de ruido y detección de quimeras realizado con el software USEARCH (Edgar, 2016) y utilizando la base de datos de *Greengenes*como referencia (DeSantis et al., 2006). El pipeline *Quantitative Insights Into Microbial Ecology* (*QIIME*) 1.9.1 (Caporaso et al., 2010) se utilizó para obtener unidades taxonómicas operacionales (*OTUs*), mediante una búsqueda *de novo* con un umbral de similitud del 97%, y utilizando un programa de agrupamiento bayesiano que delinea OTUs basándose en la distribución natural de los datos. Las secuencias representativas fueron alineadas a través del método PyNAST e insertadas en el árbol filogenético para la anotación taxonómica. La taxonomía fue asignada contra la versión 13.8 de *Greengenes*. Las diversidades alfa y beta entre muestras se estimaron con el paquete Vegan R (Oksanen et al., 2018). Finalmente, se utilizó el paquete ampvis2 (Andersen et al., 2018) para realizar un análisis multivariante considerando los grupos de muestras establecidos en función del fenotipo SCC.

RESULTADOS Y DISCUSIÓN

Según los recuentos de células somáticas, y siguiendo los criterios de Gonzalez-Rodriguez et al., (1995), entre las 49 muestras incluidas en el estudio se identificaron 36 muestras de ubres sanas (“Sanas”) y 13 muestras compatibles con mastitis subclínica (“MS”), que a su vez, se dividieron en dos grupos, tipo1 (“MS1”) (SCC> 4.000.000 células/ml) y tipo 2 (“MS2”) (400.000 células/ml < SCC < 2.000.000 células/ml). Considerando todas las muestras, la secuenciación de la región V4 del gen 16S de ARN ribosómico generó un total de 7,93 millones de lecturas brutas de 301 pb. Después del ensamblado y el filtrado de calidad se utilizaron un total de 6.387.130 secuencias de 289 pares de bases. Tras aplicar la profundidad de secuencia a 10.027, de acuerdo con los recuentos observados más bajos, se hallaron en total 44.068 OTUs. Todos los OTUs se agruparon en 43 géneros, 3 del dominio *Archea* y 40 del dominio *Bacteria*, siendo los más abundantes *Firmicutes* (60,51%), *Actinobacterias* (17,18%) y *Proteobacterias* (9.46%). A nivel de género se obtuvieron 473 observaciones. Con el fin de simplificar estos resultados se estableció una clasificación adicional en la que, por un lado, se definió el conjunto “Otros”, que agrupa aquellos géneros que aparecen en una proporción menor del 0.5%, y por otro lado se estableció el conjunto “Indefinidos”, que agrupa aquellos OTUs en los que el método no ha sido capaz de asignar un género concreto y solo ha definido hasta el nivel de familia. Así, el 30,25% del total de las secuencias se agrupó bajo el conjunto “Indefinidos” y el 10,22% bajo el conjunto “Otros”, obteniéndose 15 géneros abundantes entre los que destacan *Staphylococcus* (14,8%), *Corynebacterium* (11,8%), *Lactobacillus* (11,2%), *Alloiococcus* (4,6%) y *Streptococcus* (4%). Todas las muestras de leche analizadas revelaron una diversidad microbiana alta, independientemente de su recuento de SCC, apreciándose en las muestras del grupo MS1 un claro aumento del genero *Staphylococus* y *Streptococcus*, así como una reducción de la abundancia de géneros Indefinidos y Otros. La identificación del género *Staphylococcus* entre los más prevalentes del grupo MS1 concuerda con los resultados publicados en humano y bovino (Hunt et al., 2011; Kuehn et al., 2013). El análisis multivariante de coordenadas principales sobre la matriz de distancias mostró que las muestras sanas se discriminan fácilmente de las muestras del grupo MS1 (SCC> 4.000.000 células/ml) en función de sus perfiles de microbiota. Los análisis estadísticos no evidencian diferencias significativas grandes entre los tres grupos aquí considerados. Por ello, futuros estudios debieran basarse en un mayor número de muestras a analizar y, si es posible en abordar un análisis del metagenoma completo para poder detectar microorganismos (por ejemplo: hongos) que con el análisis del gen 16S no pueden ser identificado.

REFERENCIAS BIBLIOGRÁFICAS

• A. WP, H. HJ and M. MO 2018, *Journal of Applied Bacteriology* 46, 269–277 • Addis et al. 2016, *Molecular bioSystems* 12, 2359–2372 • Andersen et al. 2018, *bioRxiv* • Andrews 2010 • Caporaso et al. 2010, *Nature methods* 7, 335–336 • DeSantis et al. 2006, *Applied and environmental microbiology* 72, 5069–5072 • Edgar 2016, *bioRxiv* • Gonzalez-Rodriguez et al. 1995, *Journal of dairy science* 78, 2753–2759 • Heikkila et al. 2003, *Journal of applied microbiology* 95, 471–478 • Hogan 1999. *WI : National Mastitis Council* • Hunt et al. 2011, *PLOS ONE* 6, e21313 • Isaacs 2005, *The Journal of Nutrition* 135, 1286–1288 • Kuehn et al. 2013, *PLOS ONE* 8, e61959 • Magoč et al. 2011, *Bioinformatics* 27, 2957–2963 • Martin et al. 2003, *The Journal of pediatrics* 143, 754–758 • Morrow et al. 2004, *Seminars in pediatric infectious diseases* 15, 221–228 • Oftedal 2002, *Journal of mammary gland biology and neoplasia* • Oksanen et al. 2018, *Bioinformatics* 27, 863–864 • Stubbendieck et al. 2016, *Frontiers in microbiology* 7, 1234.

Agradecimientos: Este trabajo ha sido financiado por el Proyecto AGL-2015-66035-R del MINECO. C. Esteban-Blanco es beneficiaria de una beca FPI asociada al anterior proyecto (BES-2016-07-8080). B. Gutiérrez-Gil es investigadora contratada a través del programa "Ramón y Cajal" del MINECO (RYC-2021-10230).

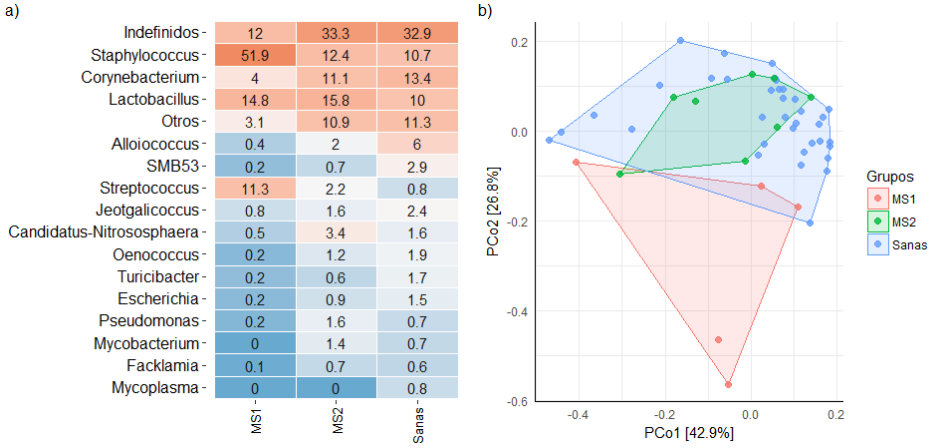


Figura 1. a) Comparación de la abundancia relativa de géneros dentro de los tres grupos clasificados en este estudio. ***b)*** Análisis de coordenadas principales considerando los tres grupos establecidos en función del SCC, basado en la matriz de distancias Bray-Curtis.

PRELIMINARY STUDY ON THE CHARACTERIZATION OF THE MAMMARY GLAND MICROBIOME IN LACTATING ASSAF SHEEP

ABSTRACT: The aim of this study was to use high-throughput sequencing of the 16S rRNA gene to describe the microbial diversity of ovine milk samples classified as derived from healthy and subclinical mastitis dairy ewes based on somatic cell counts (SCC). Milk samples from 49 Assaf sheep were analysed for SCC and used for extraction of bacterial DNA. Based on SCC, 36 samples were classified as "Healthy", and 13 as "Subclininal Mastitis", type 1 and type 2 (MS1, MS2). The region V4 of the 16S rRNA gene was individually amplified and sequenced for all the samples. *QIIME 1.9.1* software analysis was performed, and 44.068 operational taxonomic units (OTUs) were identified in total, distributed in 473 genera, although 436 genus show up in very little proportion (<0.1%) and were clustered under a group called "Others". The milk of sheep was dominated by *Staphylococcus*, accounting for 14.8%, followeb by *Corynebacterium* (11.8%), *Lactobacillus* (11,2%), *Alloiococcus* (4.6%), and *Streptococcus* (4%). The samples included in the MS1 group based on SCC showed a higher presence of *Staphylococus* y *Streptococcus*, and a decrease of the proportion of genera included the "Indefined" and "Others" groups, although significant differences were not found between groups. Future studies based on a larger number of samples and metagenome sequencing analysis may help to decipher the importance of microbioma in the sheep udder health.

Keywords: high-throughput sequencing, 16S rRNA gene, diversity, microbiome, OTUs.

Resultado 2.2

Microbiota characterization of sheep milk and its association with somatic cell count using 16s rRNA gene sequencing

C. Esteban-Blanco[1], B. Gutierrez-Gil[1], F. Puente-Sanchez[2], H. Marina[1], J. Tamames[2], A. Acedo[3], and J.J. Arranz[1]

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain; [2]2Departamento de Biología de Sistemas, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain; [3]Biome Makers Inc., West Sacramento, CA, USA.

Journal of Animal Breeding and Genetics, 137(1), 73–83.
https://doi.org/10.1111/jbg.12446

Received: 15 June 2019 | Revised: 13 September 2019 | Accepted: 20 September 2019

DOI: 10.1111/jbg.12446

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics WILEY

Microbiota characterization of sheep milk and its association with somatic cell count using 16s rRNA gene sequencing

Cristina Esteban-Blanco[1] | **Beatriz Gutiérrez-Gil**[1] | **Fernando Puente-Sánchez**[2] | **Héctor Marina**[1] | **Javier Tamames**[2] | **Alberto Acedo**[3] | **Juan José Arranz**[1]

[1]Facultad de Veterinaria, Departamento de Producción Animal, Universidad de León, León, Spain

[2]Departamento de Biología de Sistemas, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain

[3]Biome Makers Inc., West Sacramento, CA, USA

Correspondence
Juan José Arranz, Facultad de Veterinaria, Departamento de Producción Animal, Campus de Vegazana, s/n. Universidad de León, 24071, León, Spain.
Email: jjarrs@unileon.es

Funding information
Ministerio de Economía y Competitividad, Grant/Award Number: AGL2015-66035-R

Abstract

This work aimed to use 16S ribosomal RNA sequencing with the Illumina MiSeq platform to describe the milk microbiota from 50 healthy Assaf ewes. The global observed microbial community for clinically healthy milk samples analysed was complex and showed a vast diversity. The core microbiota of the sheep milk includes five genera: *Staphylococcus*, *Lactobacillus*, *Corynebacterium*, *Streptococcus* and *Escherichia/Shigella*. Although there are some differences, some of these genera are common with the microbiota core pattern of milk from other species, especially with dairy cows. The microbial composition of the studied samples, based on the definition of amplicon sequence variants, was analysed through a correlation network. A preliminary analysis by grouping the milk samples based on their somatic cell count (SCC), which is considered an indicator of subclinical mastitis (SM), showed certain differences for the core of the samples identified as SM. The differences in the microbiota diversity pattern among samples might also suggest that subclinical mastitis would be associated with the significant increase in some genera that are inhabitants of the mammary gland and a remarkable concomitant reduction in the microbial diversity. Additionally, we have also presented here a preliminary analysis to assess the impact of the sheep milk microbiome on SCC, as an indicator of subclinical mastitis. The results here reported provide a first characterization of the sheep milk microbiota and settle the basis for future studies in this field.

KEYWORDS

16S rRNA gene sequencing, amplicon sequence variants, cheese-making traits, dairy sheep, milk microbiota, somatic cell count

1 | INTRODUCTION

In livestock species, culture-independent molecular techniques, particularly those based on the study of 16S ribosomal RNA (rRNA) gene sequence, have been used to assess bacterial milk diversity. In dairy cows, many studies have been reported with the aim of characterizing the microbiota of cow's milk collected directly from mammary gland using next-generation sequencing technologies (Oikonomou et al., 2014; Oultram, Ganda, Boulding, Bicalho, & Oikonomou, 2017), and to a lesser extent, there are similar studies in water buffalo milk (Catozzi et al., 2017) and small ruminant's products (McInnis, Kalanetra, Mills, & Maga, 2015). In dairy cows, research on the diversity of the cow's milk core microbiota has been addressed mainly through the comparison between milk samples from healthy and mastitic cows because

of the economic importance of this disease affecting in dairy cattle herds worldwide (Hagnestam-Nielsen & Ostergaard, 2009). The core microbiota of healthy cow milk samples has been described by different authors (Kuehn et al., 2013; Oikonomou et al., 2014). This microbiota appears to be different among the different domestic dairy species (Catozzi et al., 2017). In dairy sheep, where milk is mainly used for the production of mature cheeses, some studies have analysed the microbiota composition of this milk product based on culture analysis (Goncalves et al., 2018). However, to our knowledge, no published research in this species has used 16S rRNA sequencing technologies to study sheep udder microbiota. Based on this, the present study aimed to characterize the microbial diversity of ovine milk samples derived from clinically healthy udders using high-throughput DNA sequencing of the 16S rRNA gene. Considering the economic relevance of subclinical mastitis in dairy sheep, we have also used the data generated in this study to examine possible differences in microbiota composition between samples with different levels of somatic cell count (SCC). Based on this, we present here an exploratory association analysis to assess the possible relationship between this indicator trait of subclinical mastitis and the members of the microbiota community identified in the samples.

2 | MATERIALS AND METHODS

2.1 | Sample collection and phenotypes

In total, 50 Assaf dairy ewes from a single flock (Zamora, Spain) and without clinical signs of mastitis were included in this study. For each animal, a total of 100 ml of milk from both mammary glands were collected into a single sterile container covered with sterile gauze to filter the milk and poured into two 50-ml sterile tubes (2 samples/ewe). Detailed information about the animals included in the study is given in Table S1 (birth date, lambing date, lambing number, age at lambing (year), sampling date and days in milk). All the milk samples were collected in a single day through the following procedure: the udder was carefully cleaned with sterile wet wipes, and then, the nipples were disinfected with 70% ethanol and rubbed with sterile gauze. During milk collection, the first streams were discarded. After sample collection, the two samples were immediately kept at 4°C, transported to the laboratory and processed for DNA extraction and SCC determination on the same day. The same conditions of transport are used for all the samples. Although all the ewes sampled in the present study showed healthy udders without signs of clinical mastitis, the SCC measurements of the milk samples analysed ranged between 27,000 and 26,915,000. To explore the possible differences in the core microbiota between milk samples with different SCC measurements, and following Gonzalez-Rodriguez, Gonzalo, San Primitivo, and Carmenes (1995), we initially classified the samples in two groups. Hence, 37 samples (74%) showing SCC < 400,000 cells/ml to enhance as healthy samples ("Healthy"), and 13 (26%) showing SCC > 400,000 cells/ml were considered as potential subclinical mastitis ("SM") samples (Table S1).

2.2 | 16S rRNA sequencing and bioinformatic analysis

After DNA extraction with the Dneasy Powerlyzer powersoil kit (Qiagen), the amplification of the 16S rRNA hypervariable V4 region was performed using BiomeMakers®custom primers (Patent WO2017096385). An average of 150,000 reads per sample was generated using 2 × 301bp paired-end sequencing with an Illumina MiSeq platform (Illumina, San Diego, CA, USA). In order to validate our procedure, a sterilized Milli-Q water sample (instead of gDNA) was included as a negative control in DNA extraction and downstream PCR amplification. Furthermore, a gDNA from pure bacteria was used as a positive control. The downstream PCR reactions showed no amplification for the negative controls, while an amplicon of 350 bp was obtained for the positive control. Both positive and negative controls were included in the pool sample mix, and they were sequenced in each Miseq run. The MiSEQ paired-end data were filtered using moira v1.3.2 (Puente-Sánchez, Aguirre, & Parro, 2016) by truncating the sequences to 250 base pairs before quality control. Amplicon sequence variants (ASVs) counts were determined from filtered sequence data using DADA2 pipeline, which differentiates sequencing error from real variants. ASVs are higher resolution analogues of the traditional operational taxonomic units (OTUs) down to the level of single-nucleotide differences over the sequenced gene region. The SILVA nr v.132 database (Quast et al., 2013) was used, and chloroplast and mitochondrial genome sequences were removed. In order to reduce the incidence of false positives, diversity analyses using the Shannon, Simpson and Chao1 indexes were performed in a randomly selected 7,000 sequences per sample. We used SparCC (Friedman & Alm, 2012) to determine correlations between ASVs, focusing on the relative abundance data. We used 1,000 bootstrap replicates to calculate nominal p values, and these values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure in R for a q value < 0.1. Then, correlations with an absolute value >0.6, p adjusted values <.05 and the corresponding attributes were imported into Cytoscape v.3.7.1 (Shannon et al., 2003) for the visualization of the network models.

2.3 | Exploratory association analysis with SCC

After examination of the phenotypic measurements available for analysis (Table S1), the SCC values were log-transformed

to obtain the somatic cell score (SCS). Then, considering those ASVs whose relative abundance accounted for more than 0.5% of the total sequences (a total of 27 ASVs), we performed further analyses to assess whether the milk microbiome contributes significantly to variation in SCC. For that, and considering the approaches suggested by some authors (Fu et al., 2015), we performed both a quantitative and a binomial analysis.

In the first of these analyses, each of the ASVs previously defined was considered as a fixed factor in the model. Also, the abundance level of each ASV was subjected to a Bayesian estimation, in order to replace the zeros counts by an estimated value and to a centred log-ratio (clr) transformation per individual. The following linear model was used:

$$y = AGE\_NB + NBL + DIM + ASV + e$$

where y is the vector, including the phenotypic SCC. In this model, we defined four fixed effects, including two factors (*AGE_NB* and *NBL*) and two covariates (*DIM* and *ASV*). In more detail, *AGE_NB* is the age at parturition combined with the number of births (12 levels), and *NBL* is the number of born lambs (2 levels, one or two lambs); *DIM* is the fixed effect of days in milk, and *ASV* is the abundance of each microbe after Bayes estimation and clr transformation. Finally, e is the vector of residual effects. The second binomial analysis performed for each of the three phenotypes tested the effect of each ASV (presence or absence; coded as a binary trait, 0 and 1) on the trait. For this purpose, we use the same model described above, adjusting for the same fixed factors, but with the difference that ASV presence/absence is considered as a factor instead of a covariate.

3 | RESULTS

3.1 | Taxonomic profile analysis and core microbiota

The sequencing of the V4 region of the 16S rRNA gene performed for the 50 sheep milk samples under study generated a total of 793 million raw reads. The length of each read was 301 bp. After trimming and quality control, a total of 4,200,253 sequences with 250 bp were used for the subsequent analyses (basic statistics in Table S2). After this quality filtering step, the DADA2 analysis performed for the 50 sheep milk samples analysed in this work identified 13,987 ASVs. The sampling depth was set to 7,000. This value, as shown in Figure S1, was appropriate to capture all bacterial diversity. From the initial number of ASVs identified, only 76 of them (71.2% of all of the sequences) showed a relative abundance higher than 0.1% (Table S3).

A total of 43 phyla were identified considering all the ASVs defined, two from the Archaea's domain and 41 from the Bacteria's domain (Table S4). Overall, the top most abundant taxa represented in the studied milk samples were *Firmicutes* (64.44%), *Actinobacteria* (14.25%), *Proteobacteria* (9.08%), *Acidobacteria* (2.7%), *Bacteroidetes* (2.3%), whereas *Thaumarchaeota* (*Archaea* domain), *Chloroflexi*, *Planctomycetes* and *Tenericutes* constituted minor phyla, each contributing less than 2% of total sequences (Figure 1a). Phyla, whose mean relative abundances accounted for more than 0.5% of the total sequences, were regarded as predominant bacterial phyla and accounted for 97.74% of sequences in all samples. The remaining bacteria phyla were considered into *Others* group. Bellow, at the genus level, a total of 988 were identified in this work. Many taxa were reported at a very low proportion (<0.5%) (Table S5) and were clustered under the "*Others*" label. We defined the "*Undefined*" label to include those taxa for which the bioinformatic analysis based on the 16S query database was unable to assign a specific genus and were only defined up to higher levels. "*Undefined*" bacteria and the taxa included in the "*Others*" label accounted for 13.62% and 18.32% of the total sequences. The 18 most predominant bacterial taxa accounted for over 68% of the sequences. According to our analyses, the microbiota of sheep udder milk samples was dominated by the genus *Staphylococcus*, accounting for 16.8%, followed by *Lactobacillus* (14.1%), *Corynebacterium* (8.8%), *Alloiococcus* (6.8%) and *Streptococcus* (4%) (Table 1; Figure 1b). Detail about genera in a lower proportion (e.g., *Romboutsia* (3%), *Jeotgalicoccus* (2.3%), *Mycobacterium* (0.6%) *Mycoplasma* (0.6%)) is shown in Table S5. Considering the asset of genera shared by all milk samples from the ewes studied in this work, the genera included in the core of microbiota of sheep milk included the most abundant genera in the general classification: *Staphylococcus*, *Lactobacillus*, *Corynebacterium*, *Streptococcus* and *Escherichia/Shigella* (Table 1), which are part of the topmost abundant phyla observed in this work.

3.2 | Differential microbiota composition between healthy and subclinical mastitis samples

A comparison between the core previously described for all the samples and the cores of bacterial taxa that would be previously defined if Healthy and SM samples would be considered independently is also provided in Table 1. We can see that the core defined for all the samples and that identified for the samples considered as "Healthy" include the same five genera, but with small variations in relative abundances. However, the core of the samples identified as "SM" includes,

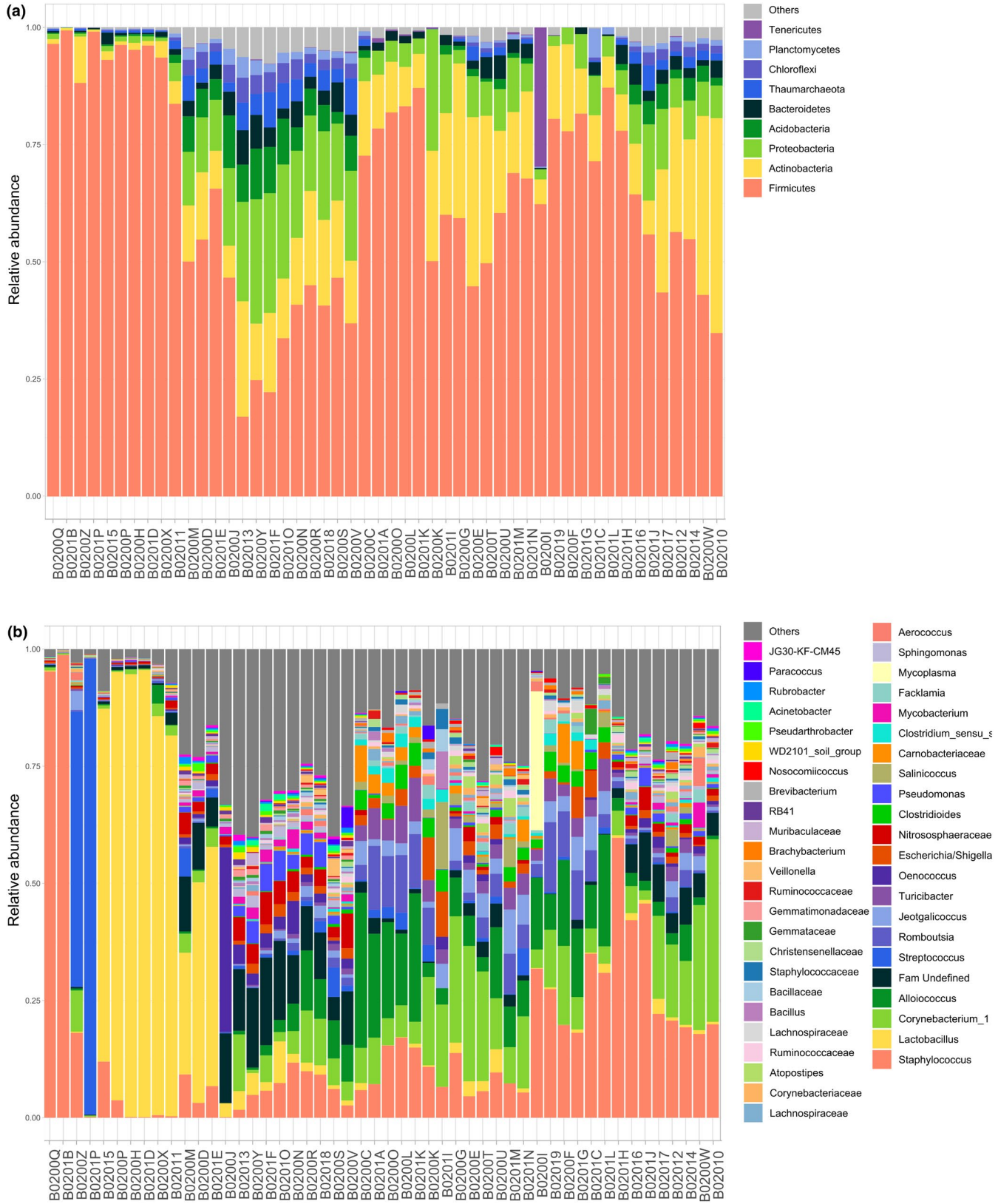


FIGURE 1 Taxonomic distribution of bacterial communities in sheep milk samples identified by 16S r RNA amplicon sequencing. Taxonomic clades were detected with an abundance >0.5%, (a) at the phylum level and (b) at the genus level. Each bar represents a subject and each coloured box a bacterial taxon. The height of a coloured box represents the relative abundance of that organism within the sample [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Characterization, at the genus level, of the core microbiota defined in this study for sheep milk samples. In addition to the core microbiota defined based on the complete set of samples analysed ($n = 50$), we indicate the core of the samples considered as "Healthy" ($n = 37$) and "Subclinical Mastitis (SM)" ($n = 13$) based on the SCC threshold considered (SCC > 400,000 cells/ml [Gonzalez-Rodriguez et al., 1995])

Core of all sheep samples ($n = 50$)	Core of Healthy milk samples ($n = 37$)	Core of SM milk samples ($n = 13$)
Corynebacterium (8.8%)	*Corynebacterium (9.7%)*	*Corynebacterium (6.4%)*
Escherichia/Shigella (1.7%)	*Escherichia/Shigella (2%)*	*Escherichia/Shigella (0.8%)*
Lactobacillus (14.1%)	*Lactobacillus (12.2%)*	*Lactobacillus (19.3%)*
Staphylococcus (16.8%)	*Staphylococcus (12.3%)*	*Staphylococcus (29.8%)*
Streptococcus (4.1%)	*Streptococcus (3.4%)*	*Streptococcus (5.8%)*
		Alloiococcus (2.04%)
		Clostridium_sensu_stricto_1 (0.2%)
		Jeotgalicoccus (1.4%)
		Pseudomonas (1.6%)
		Romboutsia (0.73%)
		Turicibacter (0.6%)

in addition to those five genera, six genera specific to samples with SCC > 400,000 cell/ml: *Alloiococcus* (2.04%), *Clostridium_sensu_stricto_1* (0.2%), *Jeotgalicoccus* (1.4%), *Pseudomonas* (1.6%), *Romboutsia* (0.73%) and *Turicibacter* (0.6%) (Table 1). When compared with the Healthy milk samples, milk from SM samples presented an increase in *Staphylococcus* (29.8%), *Lactobacillus* (19.3%) and *Streptococcus* (5.8%) and a decrease in *Corynebacterium* (6.4%) and *Escherichia/Shigella* (0.8%) (Table 1).

To explore bacterial species richness, the samples included in the SM group were divided between two different groups according the observed SCC levels: SM1 group (8 samples that showed SCC levels ranging from 400,000 and 2,000,000 cells/ml) and SM2 group (five samples with SCC > 4,000,000 cells/ml) (Table S6). In our study, SCC showed significant effects on the Shannon and Chao1 indexes, between Healthy-SM2 and SM1-SM2 samples ($p < .01$). For the Simpson index, SCC had significant effects between SM1 and SM2 groups and between Healthy and SM2 ($p < .05$) (Figure 2). Bacterial diversity was negatively correlated with SCC, suggesting that smaller groups of bacterial taxa dominate the microbiota of samples with high SCC.

In an attempt to better understand the overall structure of the milk microbiota, the analysis of the ASV interactions through a correlation network showed the presence of two different clusters (Figure 3). Cluster I includes 26 different ASVs, most of them belonging to *Firmicutes,* while Cluster II consists of 11 ASVs distributed in 4 phyla (see details in Figure 3a). Considering the distribution of the ASV related to the two genera with the highest abundances, it was interesting that *Staphylococcus* was only present in Cluster I, through four representative ASVs (ASV_7, ASV_8, ASV_17 and ASV_38), whereas *Lactobacillus* was only present in Cluster II, with two representatives ASVs, ASV_2 and ASV_11. These *Lactobacillus* ASVs were negatively correlated; some of the ASVs included in Cluster I (Figure 3a).

When the same correlation network was analysed considering the proportion of each ASV in the two groups defined based on the SCC (Figure 3b), none of the ASVs was exclusively associated with "Healthy" and "SM." However, it was observed that the ASVs included in Cluster II were, in general, present in a higher proportion of potential subclinical mastitis samples, whereas, alternatively, the ASVs of Cluster I was associated with a higher proportion of samples with low cell counts (categorized as Healthy samples).

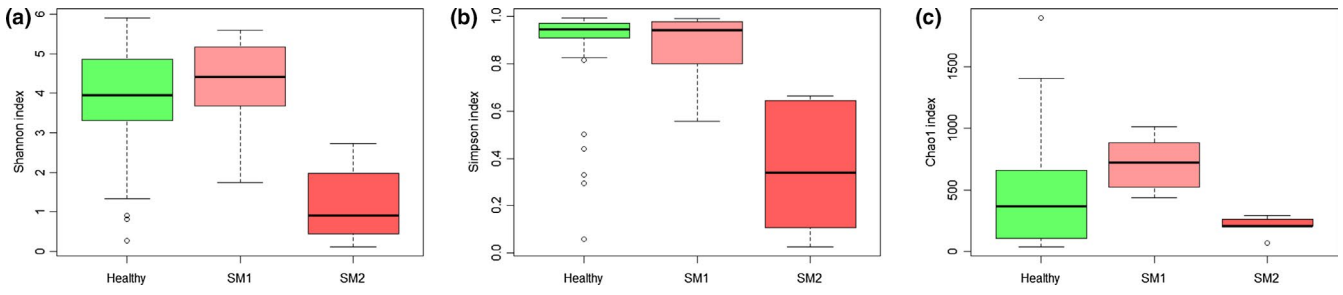


FIGURE 2 Changes in milk microbial diversity showed by groups: "Healthy," "SM1" and "SM2." (a) Bacterial alpha diversity determined by the Shannon index. (b) Bacterial alpha diversity determined by the Simpson index. (c) Bacterial alpha diversity determined by Chao1 index [Colour figure can be viewed at wileyonlinelibrary.com]

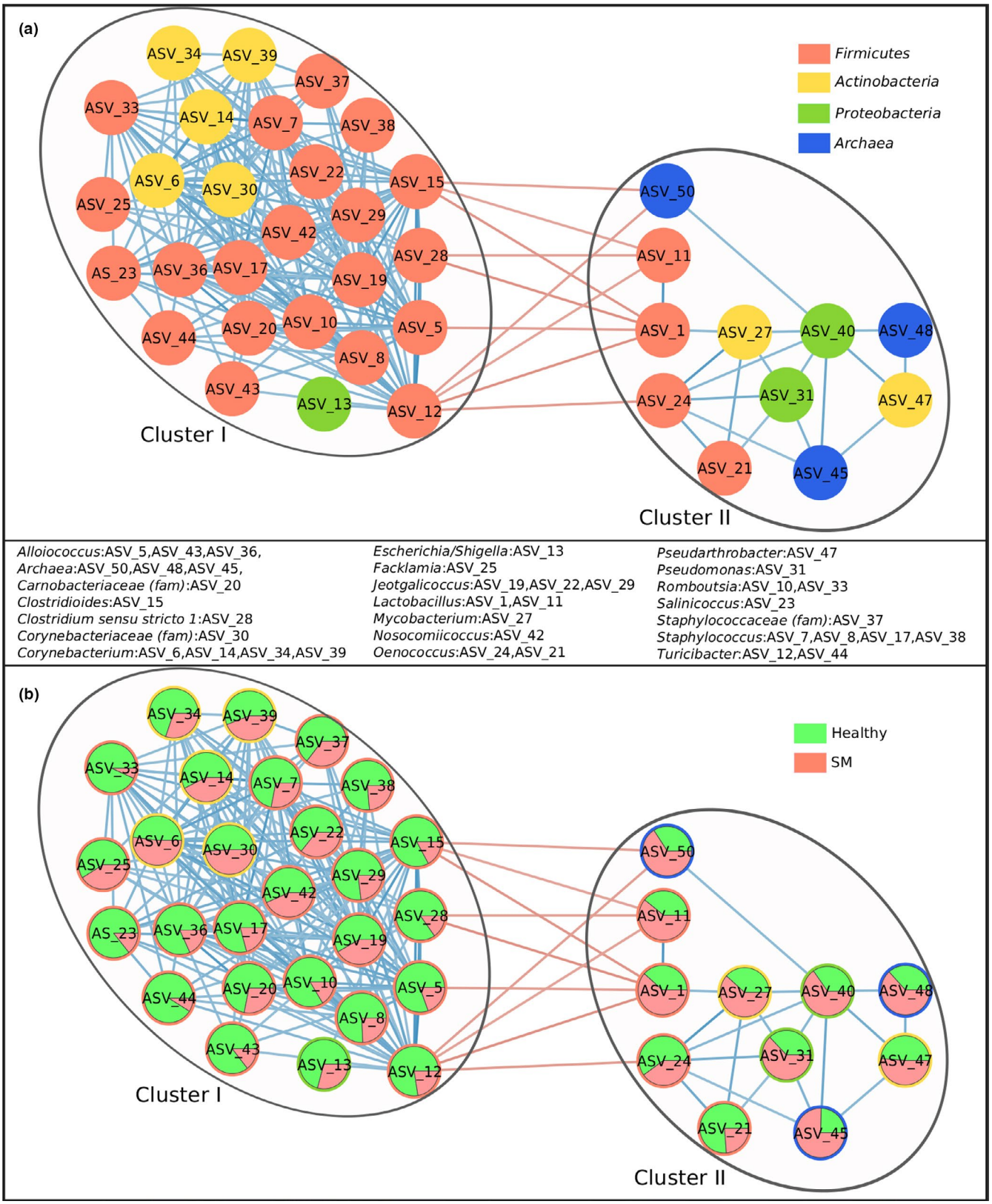


FIGURE 3 Correlation strengths of the abundant microbiota of the sheep milk represented as a correlation network. Different colours are used to indicate different phyla (a) and the abundance of each ASV in two studied groups in this work, "Healthy" and "SM" (b). The correlation coefficients were calculated with the software Sparse Correlations for Compositional data algorithm (SparCC). Red and blue lines in the network represent negative and positive correlations, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Results of the two, quantitative and binomial, association analyses performed in this study for the SCC trait. For the significant associations identified by each of the two analyses, the association estimate (Estimate), the standard error (Std.Error) and the significance level (*p* value) are indicated. Blank cells mean that the corresponding analyses did not identify any association with the studied trait and that ASV

		Bayesian analysis			Binomial analysis		
ASV	**Genus**	**Estimate**	**Std.Error**	***p* value**	**Estimate**	**Std.Error**	***p* value**
ASV_2	*Staphylococcus*	0.118407	0.047983	.0177			
ASV_5	*Alloiococcus*	−0.1556524	0.0550174	.00706			
ASV_8	*Staphylococcus*	−0.10852	0.040083	.00969	−0.767026	0.257626	.00476
ASV_14	*Corynebacterium*	−0.1219698	0.0602572	.0492			
ASV_15	*Clostridioides*	−0.150152	0.045213	.00184			
ASV_19	*Jeotgalicoccus*				−1.508082	0.570881	.0115
ASV_20	*Carnobacteriaceae*				−0.508906	0.240317	.04
ASV_22	*Jeotgalicoccus*	−0.138152	0.043677	.00286	−1.063515	0.316261	.00163
ASV_23	*Salinicoccus*	−0.103532	0.043429	.0216	−1.094039	0.359638	.00399

3.3 | Association of ASV with SCC

The quantitative analysis identified a total of seven ASVs significantly associated with SCC. Considering the sign of the effect estimate, six of these ASVs, belonging to six different taxa (*Alloiococcus, Staphylococcus, Corynebacterium*, etc.), were negatively associated with the trait, and therefore, higher abundances of these ASVs were related to a lower SCS value. Only the ASV_2, belonging *Staphylococcus*, was associated with an increase in the SCS value (Table 2). On the other hand, the binomial model for the SCS trait identified five significant associations for four different taxa (*Staphylococcus*, *Jeotgalicoccus*, *Salinicoccus*, *Carnobacteriaceae* family, etc.) all of them showing a negative effect estimate (Table 2). The significant associations for three of the ASVs, ASV_8 (*Staphylococcus*), ASV_22 (*Jeotgalicoccus*) and ASV_23 (*Salinicoccus*), were supported by the two analyses.

4 | DISCUSSION

We report here the first detailed characterization of milk microbiota in dairy ewes using high-throughput 16S rRNA gene sequencing in a commercial population of Assaf sheep. The present study is based on the analysis of the V4 hypervariable region of the 16S rRNA. Zhang et al. (2018) claimed that the V4 target region was one of the most powerful at capturing bacterial community data. Moreover, we provide a robust description of the bacterial diversity of sheep milk through the ASV inference, which allows the detection of intraspecies variation. At the phylum level, the more predominant phyla identified in our study, *Firmicutes*, *Actinobacteria* and *Proteobacteria*, have also been described as prevalent taxa in milk samples from other animal species (McInnis et al., 2015; Oikonomou et al., 2014). We provide a comparison of the core microbiota described here for sheep with that reported in other species, including human (Hunt et al., 2011), cow (reviewed by Derakhshani et al., 2018), buffalo (Catozzi et al., 2017) and goat (McInnis et al., 2015) (Table 3). By comparing the genera present in the core microbiota defined for sheep milk samples with that of other species, we can see, firstly, that *Staphylococcus*, *Streptococcus* and *Corynebacterium* are present in the core of the microbiota of the all the compared species, except buffalo and goat, respectively. Hence, our comparison table suggests that these genera are inhabitants of the milk in different species. However, they are also identified as mastitis-causing pathogens in dairy cows. With regard to the *Lactobacillus*, it appears to be shared only with the core defined for healthy cow milk samples by Derakhshani et al. (2018). Finally, *Escherichia/Shigella* is not included in the core microbiota of human, cow, buffalo or sow milk, but Zhang et al. (2017) reported that *Escherichia/Shigella* genus was present in milk from Saanen goats at the same proportion than in the present study.

In addition to the descriptive analysis of the sheep milk microbiota, we have presented herein a preliminary analysis to evaluate the possible relationship between the sheep milk microbiota and SCS, an indicator of subclinical mastitis. It should be taken into account that although the annual incidence of clinical mastitis in small ruminants is generally lower than 5%, the prevalence of subclinical mastitis has been estimated at 5%–30% or even higher (Contreras & Rodriguez, 2011). Hence, in the regular milking routine of dairy sheep flocks, the total production includes samples

TABLE 3 Bacterial populations, which are part of the core microbiota of milk, detected in milk samples from different species

Sheep (this study)	Human (Hunt et al., 2011)	Cow reviewed by (Derakhshani et al., 2018)	Buffalo (Catozzi et al., 2017)	Goat (McInnis et al., 2015)
Corynebacterium	*Bradyrhizobiaceae*	*Bacteroides*	*02d06*	*Agrobacterium*
Escherichia/Shigella	*Corynebacterium*	*Comamonas*	*5-7N15*	*Micrococcus*
Lactobacillus	*Propionibacterium*	*Corynebacterium*	*Acinetobacter*	*Phyllobacterium*
Staphylococcus	*Pseudomonas*	*Enterococcus*	*Aerococcus*	*Pseudomonas*
Streptococcus	*Ralstonia*	*Fusobacterium*	*Clostridium*	*Rhodococcus*
	Serratia	*Lachnospiraceae*	*Facklamia*	*Stenotrophomonas*
	Sphingomonas	*Lactobacillus*	*Micrococcus*	*Streptococcus*
	Staphylococcus	*Propionibacterium*	*Propionibacterium*	
	Streptococcus	*Pseudomonas*	*Pseudomona*	
		Ruminococcaceae	*Psychrobacter*	
		Staphylococcus	*SMB53*	
		Stenotrophomonas	*Solibacillus*	
		Streptococcus	*Staphylococcus*	
			Trichococcus	
			Turicibacter	

from a wide range of SCC values, including, in some cases, milk from animals suffering from subclinical mastitis. So, we have exploited the variability on SCC values observed in the milk samples considered in this study to perform a comparison of the core microbiota between the two major groups defined based on SCC values, "Healthy" (37 samples) and "SM" (13 samples) (Table 1). The increase in *Staphylococcus*, *Streptococcus* and *Lactobacillus* abundance reported for the SM samples appears to be associated with the decrease in the abundances of many other genera that we later observed for the samples with the highest SCC values (Figure 2b). *Staphylococcus aureus* is an opportunistic pathogen that causes a wide variety of infections in both humans and animals and is recognized as a major pathogen of the mammary gland responsible for clinical and subclinical intramammary infections in small ruminants (Bergonier, de Cremoux, Rupp, Lagriffoul, & Berthelot, 2003). *Streptococcus* genus is known as an important pathogen causing mastitis (Klaas & Zadoks, 2017). *Lactobacillus* is a genus reported as capable of inhibiting some major mastitis pathogens (Jara, Sanchez, Vera, Cofre, & Castro, 2011).

On the other hand, *Corynebacterium*, *Jeotgalicoccus and Escherichia/Shigella* showed lower abundances in the SM group when compared with the Healthy group. The identification of six additional genera specific of the SM group (*Clostridium sensu stricto 1, Pseudomonas, Romboutsia, Alloiococcus, Jeotgalicoccus* and *Turicibacter)* might be of interest about particular bacteria related to subclinical infection. Among this, *Clostridium perfringens* (a type of bacteria included in the Cluster I of *Clostridium sensu stricto*) has been identified as an aetiological agent of mastitis in other ruminants (Osman, El-Enbaawy, Ezzeldeen, & Hussein, 2009). *Pseudomonas aeruginosa* is a pathogenic species and has been reported to be among the most prevalent species cultured from bovine milk samples being also one of the bacteria most frequently associated with clinical or subclinical mastitis (Barkema et al., 1998). Also, *Alloiococcus* has been reported to be present in cow's raw milk by Doyle, Gleeson, O'Toole, and Cotter (2017). The exploratory comparison between the samples showing the intermediate and the highest levels of SCC (SM1 vs. SM2) showed a pronounced decrease in the microbiota diversity for the samples with SCC > 4,000,000 cell/ml. These results agree with the observations reported by several authors in cattle (Catozzi et al., 2017; Kuehn et al., 2013; Oikonomou et al., 2014), in where the increase in pathogenic bacteria is associated with a decrease in commensal and natural hosts of the considered tissue. However, we have to recognize that the validity of our results is limited due to the small number of samples considered, especially in the SM2 group.

This decrease in diversity was also reflected in the correlation network built based on the analysed data set. Hence, the cluster including the ASVs that, in general, showed higher abundances in SM samples, Cluster II, had less than half of the ASVs grouped in the alternative Cluster I associated with the Healthy samples (Figure 3b). Focusing on the two more abundant ASVs included in Cluster II, ASV_1 and ASV_11 are assigned to *Lactobacillus* genus. This result was in agreement with observations found when comparing the core between Healthy and SM samples, where *Lactobacillus* increased from 12.2% to 19.3%. However, none of the *Lactobacillus* ASVs showed a significant association with

the SCS trait in our study. Because previous studies have reported the *Lactobacillus* genera to be decreased in the milk of subclinical mastitis cows (Qiao et al., 2015) and to be capable of inhibiting some major mastitis pathogens (Jara et al., 2011), further studies should clarify the role of this genera in relation to the subclinical mastitis of dairy sheep.

Following the same criteria, the more abundant ASVs in Cluster I was associated with *Corynebacterium* (ASV_6) and *Staphylococcus* (ASV_7 and ASV_8). The first of this genera, also represented in Cluster I by other four ASVs (Figure 3a), had shown a decrease in abundance in SM samples when compared with Healthy samples (Table 1). Besides, one of these ASVs, ASV_14, was identified as negatively associated with the SCC trait in the Bayesian analysis here reported (Table 2). These results would disagree with the fact that *Corynebacterium spp.* are among the most frequently isolated potential pathogens associated with subclinical mastitis in dairy cows (Gonçalves et al., 2014) and dairy sheep (Fernández et al., 2001).

About the *Staphylococcus*, although this genus showed an increase in abundance in the SM samples compared with the Healthy samples and ASV_2 was the only one significantly associated with an increase in the SCS value in the Bayesian analysis, this specific ASV was not included in the correlation network. However, the association of ASV_8 with a decrease in SCS was supported by both the Bayesian and binomial association analyses. These results suggest that different *Staphylococcus* ASVs may have a different effect, protective or pathogenic, in the host.

It is important to note that due to the limited sample size of this study, the power of the association analyses performed here is low. As a consequence, the results reported here should be considered only as a first step to better understand the possible influence of the milk microbiota on an indicator of subclinical mastitis. Further analyses based on larger numbers of animals should be carried out to explore the composition of the sheep milk microbiota that could help to improve SCC trait.

5 | CONCLUSIONS

To our knowledge, the present study provides a first step towards the description of sheep milk microbiota using a culture-independent metagenomic approach based on the sequencing of the V4 hypervariable region of the 16S gene. Overall, the described microbial community for clinically healthy animals is complex and shows a vast diversity. Moreover, the core microbiota described for all the analysed samples was further explored comparing two groups of samples defined based on their SCC values. Although the number of samples per group was limited, the altered composition of the microbial community observed for the samples showing the highest SCC values suggests that subclinical mastitis is associated with the significant increase in some genera that are inhabitants of the mammary gland (mainly *Staphylococcus*) and a remarkable concomitant reduction in the microbial diversity. Also, we provide a first exploratory approximation to assess the influence of the bacterial composition of the sheep milk on an indicator of subclinical mastitis. Because of the limited sample size, these latter results must be regarded with caution. In any case, it would be interesting to confirm the results of this first study characterizing the sheep milk microbiota through the design of future studies with larger sample sizes or even exploiting shotgun metagenomic sequencing approaches.

ACKNOWLEDGMENTS

This work was developed under the framework of AGL-2015-66035-R project financed by the Spanish Ministry of Economy and Competitiveness (MINECO, Madrid, Spain) co-funded by the European Regional Development Fund. C. Esteban-Blanco is funded by an FPI from MINECO (Ref. BES-2016-07-8080). This research has made use of the high-performance computing resources of the Castilla y León Supercomputing Center (SCAYLE, www.scayle.es).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this article.

DATA AVAILABILITY STATEMENT

The data sets generated during and/or analysed during the current study are available as supplementary information.

ORCID

Cristina Esteban-Blanco https://orcid.org/0000-0002-2425-000X
Beatriz Gutiérrez-Gil https://orcid.org/0000-0001-7990-5723
Fernando Puente-Sánchez https://orcid.org/0000-0002-6341-3692
Héctor Marina https://orcid.org/0000-0001-9226-2902
Javier Tamames https://orcid.org/0000-0003-4547-8932
Juan José Arranz https://orcid.org/0000-0001-9058-131X

REFERENCES

Barkema, H. W., Schukken, Y. H., Lam, T., Beiboer, M. L., Wilmink, H., Benedictus, G., & Brand, A. (1998). Incidence of clinical mastitis in dairy herds grouped in three categories by bulk milk somatic

cell counts. *Journal of Dairy Science*, *81*(2), 411–419. https://doi.org/10.3168/jds.S0022-0302(98)75591-2

Bergonier, D., de Cremoux, R., Rupp, R., Lagriffoul, G., & Berthelot, X. (2003). Mastitis of dairy small ruminants. *Veterinary Research*, *34*(5), 689–716. https://doi.org/10.1051/vetres:2003030

Catozzi, C., Sanchez Bonastre, A., Francino, O., Lecchi, C., De Carlo, E., Vecchio, D., … Ceciliani, F. (2017). The microbiota of water buffalo milk during mastitis. *PLoS ONE*, *12*(9), e0184710. https://doi.org/10.1371/journal.pone.0184710

Contreras, G. A., & Rodriguez, J. M. (2011). Mastitis: Comparative etiology and epidemiology. *Journal of Mammary Gland Biology and Neoplasia*, *16*(4), 339–356. https://doi.org/10.1007/s10911-011-9234-0

Derakhshani, H., Fehr, K. B., Sepehri, S., Francoz, D., De Buck, J., Barkema, H. W., … Khafipour, E. (2018). Invited review: Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility. *Journal of Dairy Science*, *101*(12), 10605–10625. https://doi.org/10.3168/jds.2018-14860

Doyle, C. J., Gleeson, D., O'Toole, P. W., & Cotter, P. D. (2017). Impacts of seasonal housing and teat preparation on raw milk microbiota: A high-throughput sequencing study. *Applied and Environmental Microbiology*, *83*(2), e02694-16. https://doi.org/10.1128/AEM.02694-16

Fernández, E. P., Vela, A. I., Las Heras, A., Domínguez, L., Fernández-Garayzábal, J. F., & Moreno, M. A. (2001). Antimicrobial susceptibility of corynebacteria isolated from ewe's mastitis. *International Journal of Antimicrobial Agents*, *18*(6), 571–574. https://doi.org/10.1016/S0924-8579(01)00424-1

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, *8*(9), e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Fu, J., Bonder, M. J., Cenit, M. C., Tigchelaar, E. F., Maatman, A., Dekens, J. A. M., … Zhernakova, A. (2015). The gut microbiome contributes to a substantial proportion of the variation in blood lipids. *Circulation Research*, *117*(9), 817–824. https://doi.org/10.1161/CIRCRESAHA.115.306807

Gonçalves, J. L., Tomazi, T., Barreiro, J. R., Braga, P. A., Ferreira, C. R., Araújo-Junior, J. P., … dos Santos, M. V. (2014). Identification of *Corynebacterium* spp. isolated from bovine intramammary infections by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Veterinary Microbiology*, *173*(1–2), 147–151. https://doi.org/10.1016/j.vetmic.2014.06.028

Gonçalves, M. T. P., Benito, M. J., Córdoba, M. D. G., Egas, C., Merchán, A. V., Galván, A. I., & Ruiz-Moyano, S. (2018). Bacterial communities in serpa cheese by culture dependent techniques, 16S rRNA gene sequencing and high-throughput sequencing analysis. *Journal of Food Science*, *83*(5), 1333–1341. https://doi.org/10.1111/1750-3841.14141

Gonzalez-Rodriguez, M. C., Gonzalo, C., San Primitivo, F., & Carmenes, P. (1995). Relationship between somatic cell count and intramammary infection of the half udder in dairy ewes. *Journal of Dairy Science*, *78*(12), 2753–2759.

Hagnestam-Nielsen, C., & Ostergaard, S. (2009). Economic impact of clinical mastitis in a dairy herd assessed by stochastic simulation using different methods to model yield losses. *Animal*, *3*(2), 315–328. https://doi.org/10.1017/S1751731108003352

Hunt, K. M., Foster, J. A., Forney, L. J., Schütte, U. M. E., Beck, D. L., Abdo, Z., … McGuire, M. A. (2011). Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS One*, *6*(6), e21313. https://doi.org/10.1371/journal.pone.0021313

Jara, S., Sanchez, M., Vera, R., Cofre, J., & Castro, E. (2011). The inhibitory activity of *Lactobacillus* spp. isolated from breast milk on gastrointestinal pathogenic bacteria of nosocomial origin. *Anaerobe*, *17*(6), 474–477. https://doi.org/10.1016/j.anaerobe.2011.07.008

Klaas, I. C., & Zadoks, R. N. (2017). An update on environmental mastitis: Challenging perceptions. *Transboundary and Emerging Diseases*, *65*(S1), 166–185. https://doi.org/10.1111/tbed.12704

Kuehn, J. S., Gorden, P. J., Munro, D., Rong, R., Dong, Q., Plummer, P. J., … Phillips, G. J. (2013). Bacterial community profiling of milk samples as a means to understand culture-negative bovine clinical mastitis. *PLoS ONE*, *8*(4), e61959. https://doi.org/10.1371/journal.pone.0061959

McInnis, E. A., Kalanetra, K. M., Mills, D. A., & Maga, E. A. (2015). Analysis of raw goat milk microbiota: Impact of stage of lactation and lysozyme on microbial diversity. *Food Microbiology*, *46*, 121–131. https://doi.org/10.1016/j.fm.2014.07.021

Oikonomou, G., Bicalho, M. L., Meira, E., Rossi, R. E., Foditsch, C., Machado, V. S., … Bicalho, R. C. (2014). Microbiota of cow's milk; Distinguishing healthy, sub-clinically and clinically diseased quarters. *PLoS ONE*, *9*(1), e85904. https://doi.org/10.1371/journal.pone.00859044

Osman, K. M., El-Enbaawy, M. I., Ezzeldeen, N. A., & Hussein, H. M. G. (2009). Mastitis in dairy buffalo and cattle in Egypt due to *Clostridium perfringens*: Prevalence, incidence, risk factors and costs. *Revue Scientifique et Technique (International Office of Epizootics)*, *28*(3), 975–986. https://doi.org/10.20506/rst.28.3.1936.

Oultram, J. W. H., Ganda, E. K., Boulding, S. C., Bicalho, R. C., & Oikonomou, G. (2017). A metataxonomic approach could be considered for cattle clinical mastitis diagnostics. *Frontiers in Veterinary Science*, *4*, 36. https://doi.org/10.3389/fvets.2017.00036

Puente-Sánchez, F., Aguirre, J., & Parro, V. (2016). A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acids Research*, *44*(4), e40. https://doi.org/10.1093/nar/gkv1113

Qiao, J., Kwok, L., Zhang, J., Gao, P., Zheng, Y., Guo, Z., … Zhang, H. (2015). Reduction of *Lactobacillus* in the milks of cows with subclinical mastitis. *Beneficial Microbes*, *6*(4), 485–490. https://doi.org/10.3920/BM2014.0077

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Zhang, F., Wang, Z., Lei, F., Wang, B., Jiang, S., Peng, Q., … Shao, Y. (2017). Bacterial diversity in goat milk from the Guanzhong area of China. *Journal of Dairy Science*, *100*(10), 7812–7824. https://doi.org/10.3168/jds.2017-13244

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., … Lu, Z. (2018). Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *The Science of the Total Environment*, *618*, 1254–1267. https://doi.org/10.1016/j.scitotenv.2017.09.228

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Esteban-Blanco C, Gutiérrez-Gil B, Puente-Sánchez F, et al. Microbiota characterization of sheep milk and its association with somatic cell count using 16s rRNA gene sequencing. *J Anim Breed Genet*. 2020;137:73–83. https://doi.org/10.1111/jbg.12446

Resultado 2.3

Comparison of sheep milk microbiome in two dairy sheep breeds using 16S rRNA gene sequencing

C. Esteban-Blanco, B. Gutiérrez-Gil, A. Suárez-Vega, H. Marina, J.J. Arranz

Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain

Manuscript in preparation

ABSTRACT

The evolution of -omics technologies enables the sequencing of amplicons providing the capability to characterize cheaply the bacterial taxonomic profile in a large dataset. In this study, the main objective of this work was to access bacterial profiles of healthy milk samples using Next Generation Sequencing of amplicons from 16S rRNA gene to characterize the milk microbiome of the Churra breed and comparing the milk bacterial composition with Assaf breed. A total of 212 samples were collected from two Churra dairy farms following the same protocol of sampling, extraction and sequencing used in the Assaf study. The core microbiota of the sheep milk in Churra includes lesser genera (only two taxa: *Staphylococcus* and *Escherichia/Shigella*) than studies reported in other dairy species or, even, in Assaf sheep mil. Microbial populations were very different in the two breeds; low total diversity of the milk microbiota in Churra was detected against that obtained in the analysis of the ovine milk microbiota in Assaf. The breed category separates clearly samples based on their microbiota composition, thus allowing identification possible taxa breed-specific. The information reported here might be used to understand the complex issue of milk microbial composition.

Keywords: milk microbiota, Churra, Assaf, dairy sheep, 16S rRNA gene sequencing.

INTRODUCTION

The study of the microbial communities found in the milk has gained increasing interest in recent years. The classical concept of sterility of the mammary gland and the milk has been challenged by the results obtained in the last decade using bacterial DNA-based methodologies. The recent significant development in culture-independent techniques, especially using 16S rRNA gene sequencing, has led to new studies dedicated to understanding the composition, the diversity and the biological roles of the milk bacterial community in different dairy livestock species (Addis et al., 2016; Catozzi et al., 2017; Esteban-Blanco et al., 2019). The changes in the balance of the bacterial community and their mutualistic interactions with the host could have an impact on animal's health (Hooper et al., 2012). Hence, most studies on the milk microbiota of the dairy ruminant species have focused on these changes and the differences between the microbial composition of milk samples obtained from healthy udders and those suffering mastitis or local inflammation. These studies have also reported the milk core microbiota for the different species, which refers to all taxa commonly found across all the samples analysed in each study. Recent studies have shown that the milk microbiome could be more complex than expected and that the results can be biased depending on the different analysis approaches applied, on the analysed milk fraction, and also on the sequencing platform used for its study (Taponen et al., 2019; Oikonomou et al., 2020). In dairy cattle, the core microbiota has been shown to change when comparing different breeds (Cremonesi et al., 2018). So, the relationships between microbiota and udder health may be not applicable from one to other breed within the same species.

In dairy sheep, a previous study of our research group presented the first characterization of the milk core microbiota in this species by sequencing the 16S rRNA gene in milk samples from healthy ewes of the Spanish Assaf sheep. This is a highly specialized dairy

breed integrated in Spain since 1977, which today has the highest census of dairy sheep population in the region of Castilla y León. This study has provided a comparison of the sheep milk core microbiota with those reported in other species and characterized the microbiota of samples with different levels of somatic cell count (SCC), which is an indicator of the health status of the udder (Esteban-Blanco et al. 2019).

In the same geographical region of Castilla y León in Spain, Churra sheep is also exploited for milk production. This is a rustic, autochthonous breed from the region of Castilla y León with a milk selection scheme since 1986 (de la Fuente et al., 1995). The census of this breed has suffered an important decrease in the last years due to the higher milk production level of the Assaf breed. In the context, the aim of the present study is the characterization of the core milk microbiota in the Churra sheep breed, and the comparison of the microbial composition of milks samples from healthy ewes of these two sheep breeds to assess the influence of the breed factor on their bacterial, also considering different levels of SCC.

MATERIALS AND METHODS

Churra milk sampling and bioinformatic data analysis

In total, 212 milk samples from Churra ewes without clinical signs of mastitis were included in this study. The sampled animals belonged to two different flocks (n = 145 and 67, respectively for flock 1 and flock 2) from the region of Castilla y León (Spain) and each ewe was sampled once. Flock 1 had an intensive management system, whereas Flock 2 followed a semi-extensive management system, based on daily grazing. The sampling protocol used was the same that that described in our previous study on the Assaf breed milk microbiota (Esteban-Blanco et al., 2019). Briefly, 100 ml of milk were collected into two 50 ml sterile containers, one for the DNA extraction and the other one for measure

of the somatic cell count (SCC). Following Gonzalez-Rodriguez et al. (1995), a threshold to distinguish between healthy and subclinical mastitis ewes was set to 400,000 cells/ml. Hence, samples were distributed into two different groups based on SCC: "Healthy", those samples showing SCC < 400,000 cells/ml and "SM" (subclinical mastitis) samples with SCC > 400,000 cells/ml.

After sample collection, the samples were immediately kept at 4ºC and transported to the laboratory and processed for DNA extraction. The same conditions of transport were used for all the samples. In order to generate comparable sequencing data with the previous study in Assaf breed, all steps for DNA extraction was were carried out in the same laboratory and over the same conditions than those used by Esteban-Blanco et al., (2019). The hypervariable V4 region was amplified using BiomeMakers®custom primers (Patent WO2017096385) and the Illumina Miseq platform (Illumina, San Diego, CA, USA) was used to perform the sequencing process. The raw data generated were analyzed following the same bioinformatic pipeline described by Esteban-Blanco et al. (2019). Shortly, amplicon sequences variants (ASVs) were detected using the DADA2 pipeline and the SILVA nr v.132 database (Quast et al., 2013) was used to perform the taxonomic assignment. Microbiota diversity of Churra milk samples was measured with the Shannon index was calculated on the ASV rarefied data table. To explore bacterial diversity across Churra milk samples with different SCC levels, we followed the same approach that in our previous study by Esteban-Blanco et al. (2019) and considered within the SM group two different groups based on the SCC observed levels: "SM1" group (samples that showed SCC levels between 400,000 and 2,000,000 cells/ml) and "SM2" (samples with SCC > 2,000,000 cells/ml), whereas the "H" group included those samples with less than 400,000 cells/ml.

Comparative study between Churra and Assaf microbiota datasets

In addition, for a direct comparison study between the microbiota composition of the Assaf and the Churra milk samples, we merged the taxonomic assignment tables for 100 of the Churra milk samples here analysed, by performing a random selection of 50 Churra samples per flock, and for the 50 Assaf milk samples previously analysed by Esteban-Blanco et al. (2019). This combined Churra-Assaf dataset was later analyzed using a nonmetric multidimensional scaling (nMDS) ordination method based on Bray-Curtis distance to visualize the between-breed differences observed in the microbiota of the analyzed samples. Considering the genera with relative abundances higher than 0.1%, DESeq2 (Love et al., 2014) was used to identify the genera showing differences between breeds. Those genera were filtered using an adjusted p-value cutoff of 0.05 and a log2fold change higher than 1.5. An additional exploratory analysis was later perform with the aim of identifying a breed-specific bacterial profile of the analyzed milk samples. An statistical analysis based on a linear model including the breed and farm effects was performed to evaluate the effect of the breed on the microbiota profiles.

RESULTS

Microbiota and diversity profile in Churra Samples

The sequencing dataset of the V4 region of the 16s rRNA gene generated for the 212 Churra milk samples under study included a total of 16,3 million raw reads. The length of the raw reads was 301 bp. After size filtering, reads with a minimum of 250 bp were kept for the next steps. The quality control and chimera removal produced a total of 11,7 million of quality reads with an average of 55,000 reads per sample. The overall number of ASVs detected by the DADA2 analysis for the 212 Churra samples reached 2,519; only 142 ASVs of them showed a relative abundance higher than 0.1% (81.8% of all of the analysed sequences).

Once the ASV table was available, taxonomy was assigned using the Ribosomal Database Project (RDP) classifier (Wang et al., 2007) natively implemented in DADA2 and trained against the SILVA database. This analysis identified a total of 31 phyla for the 212 Churra milk samples, three from the Archaea's domain and 28 from the Bacteria's domain. In the present work, the top most abundant Phyla that accounted for 97.4% of the abundance in the dataset were *Firmicutes* (50,28%), *Proteobacteria* (25,5%), *Actinobacteria* (18,9%) and *Bacteroidetes* (2,6%). However, minor phyla, each contributing less than the *Bacteroidetes* abundance and high than 0.1%, accounted for the 2,2% of the total sequences; the most abundant minor phyla in this study were *Fusobacteria*, *Planctomycetes*, *Acidobacteria*, *Deinococcus-Thermus* (Figure 1a). At genus level, we defined the "*Undefined*" label to include those taxa for which the bioinformatic analysis was unable to assign a specific genus, and the "*Others*" label was also created to cluster genus with a proportion lower to 0.5%. The taxonomic assignment of data from the 212 Churra milk samples based on the 16S query database classified 16,9% and 5,9% of the taxa within the "*Others*" and "*Undefined*" labels respectively. The predominant genera in Churra sheep milk were *Staphylococcus* (20,29%), *Cutibacterium* (6,27%), *Corynebacterium* (4,34%), *Streptococcus* (4,1%), *Massilia* (3,5%) and *Bacillus* (3,2%). The core microbiota of Churra sheep milk described in this work, as the shared genera by all milk samples, included only two genera, *Staphylococcus* and *Escherichia/Shigella*. Setting the sampling depth to 19,221, the estimated Shannon index values showed a slight decreased in microbial diversity among the three groups of milk samples defined based on the SCC value, although without significant differences (Figure 2A).

Comparative study of the milk microbiota in Churra and Assaf breeds

The joint analysis of the beta-diversity performed for the Churra and Assaf datasets previously indicated (50 Assaf, 50 Churra from flock 1, 50 Churra from flock 2) with the NMDS ordination analysis revealed a clear separation among the two studied breeds.

Although the Churra milk samples belonged to two different flocks, this analysis did not show metagenome differences between the samples from different farms (Figure 3).

Later, the Churra and Assaf milk samples with a NMDS1 less than -2.2 for Assaf and between 0.4 and 1.6 for Churra, NMDS2 value ranging between 0.5 and -0.5 (35 for Churra and 20 for Assaf) were subjected to a later analysis to identify a breed-specific bacterial profile with DESeq2.

This analysis identified 252 differentially present genera between the milk microbiota of Churra and Assaf sheep. A total of 13 genera were up-represented in Churra milk and 239 were up-represented in Assaf milk. The 30 genera with the highest abundance and the largest changes in fold-changes between breeds are given in Table 2 and represented as a heatmap in Figure 4. The top 3 genera which are more expressed in Churra vs. Assaf were *Cutibacterium, Comamonas* and *Lawsonella.* Still, the higher values of fold change reported by DESeq2 from Assaf are for *Oenococcus*, *Lactobacillus* and *Ruminococcaceae_UCG-005*.

The linear regression analysis performed for the normalized relative abundances obtained after the DESeq2 analysis, considering breed and flock as fixed factors, revealed statistical differences (p-value < 0.01) between breeds in 28 of the 30 top abundant genera. Considering the most abundant genera (*Cutibacterium*, *Nosocomiicoccus*, *Aerococcus*, *Nocardioides*, *Candidatus_Nitrocosmicus*, *RB41*, *Pseudarthrobacter*, *Rubrobacter*, *Ruminococcaceae_UCG*-005, *Lactobacillus* and *Oenococcus*) the average of R-square for was 0.85. A graphical representation of these results is provided as a PCoA plot for the most significant genera (p-value < 0.01 and an R-squared > 0.7) (Figure 5). As it can be seen, the first principal component of the PCoA plot, which refers to the breed effect, explains 95.5% of the variance.

DISCUSSION

The study of the microbiota harboured in sheep milk is of interest for the dairy sheep production because of the potential relationship between the microbiota present in milk and the health status of the mammary gland. In addition, the milk microbiota is relevant in terms of biosafety as in some specific cases pathogenic bacteria may be present in this complex fluids. Also, taking into account that most of the sheep milk is addressed to the production of high quality cheeses, increased knowledge on sheep milk microbiota is encouraged by the known influence of the microbiota composition on the cheese-making process and on the flavour and texture of the different types of cheeses (Tilocca et al., 2020).

Although for many years, the research work related to the bacterial composition in different tissues was based on the study of the contribution of single or few microorganisms (Hiergeist et al., 2015), the analysis of the microbial communities present in sheep's milk has only been possible with the recent developments of massive parallel sequencing-based technologies, mainly through the analysis of the 16S rRNA gene.

To our knowledge, only one study previously reported by our research group has used this sequencing approach to characterize the sheep milk microbiota composition in a commercial population of Assaf breed (Esteban-Blanco et al., 2019). In the present work we have applied the same methodology approach, based on the sequencing of the V4 region of the 16S rRNA gene, to characterize the milk microbiota in dairy ewes of the Spanish Churra breed. The DADA2 analysis performed identified 142 ASVs that showed a relative abundance higher than 0.1% from a total of detected 2,519 ASVs. This value is lower than that previously reported in Assaf sheep by our group, which was 13,987 (Esteban-Blanco et al., 2019), although the entire biological diversity within samples was sufficiently captured. The accumulated abundance at phylum level revealed that the more

predominant phyla were *Firmicutes*, *Proteobacteria* and *Actinobacteria* (Table 1). These phyla have been stated also as prevalent taxa by other authors in different dairy livestock species such as dairy cattle, buffalo and sheep (Addis et al., 2016; Catozzi et al., 2017; Esteban-Blanco et al., 2019). However, at genus level, the number of the total genera included in the core microbiota reported in this work, *Staphylococcus* and *Escherichia/Shigella*, was lower than in the milk core microbiota in other species in where the core microbiota consists in more than seven different genera (Hunt et al., 2011; Catozzi et al., 2017; Derakhshani et al., 2018). The core microbiota reported here for Churra sheep also includes a lower number of genera than the five previously reported for the core microbiota of Assaf sheep milk: *Corynebacterium*, *Escherichia/Shigella*, *Lactobacillus*, *Staphylococcus* and *Streptococcus* (Esteban-Blanco et al. 2019). When comparing the results of this work with the microbiota characterization reported for Assaf milk (Esteban-Blanco et al.), we assume that no batch effect is present between the two datasets (Churra and Assaf samples) because the transport, extraction and sequencing were performed under the same conditions. Hence, our characterization of sheep milk microbiota in Assaf and Churra sheep breeds suggests that the milk microbiota is most diverse in Assaf than in Churra ewes.

Following the assessment presented in our previous work in Assaf breed in relation to the microbiota composition between samples with different SCC levels, we performed here a similar evaluation for Churra sheep, by dividing the 212 available Churra samples in the same groups previously defined according to the SCC levels, “Healthy” (165 samples with SCC < 400,000 cells/ml), “SM1” (33 samples with SCC > 400,000 cells/ml) and “SM2” (14 samples with SCC 2,000,000 cells/ml),). Interestingly, in contrast with the pronounced decrease observed in the microbiota diversity of Assaf milk samples with SCC > 4,000,000 cells/ml (Figure 2B), Churra milk samples with extreme values of SCC

did not show significant changes in the bacterial species richness (Figure 2A). Hence, it seems that the breed factor not only influence the milk core microbiota but also how SCC level influences the milk microbiota composition. Other potential explanation is that those breeds for which milk microbiota diversity is lower are less influenced by changes in SCC levels. In any case, to know if the different levels of basal microbiota diversity reported here between Assaf and Churra sheep breeds have any relationship with a potential higher resistance/susceptibility status of any of these two breeds to mastitis, further research will be needed.

In addition, our additional analyses with 50 Assaf milk samples and 100 Churra milk samples (50 of each studied farm), suggested that the farm factor did not influence the milk microbiota of Churra milk samples. Whereas the breed factor, at least in this particular case, appear to determine a clear distinction between the milk microbiota of Churra and Assaf milk samples. Moreover, the NMDS plot confirmed the higher microbiota diversity of Assaf milk samples, compared with Churra milk samples, previously suggested based on the Shannon index. To have a global view of the influence of how the breed factor influence sheep milk microbiota, it is obvious that more studies are needed on this regard. Overall, the results presented here highlight the complexity of the sheep milk microbiome, as other authors have already stayed in relation to the study of the milk microbiome in dairy cattle (Cremonesi et al., 2018). The differences observed in the microbiota profiles between breeds could be related to the protective role of a balanced microbiota and resistance to infections. Further analyses based on the correlation between SCC and the microbial composition in milk from different breeds should be carried out to explore the possible protection against pathogenic bacteria within breed.

A differential pattern of microbial composition abundances between the two breeds studied here were obtained with the DESeq2 analysis. Shortly, considering one of the genera for each group (with the highest value of fold change), we observed that *Cutibacterium* was more abundant in Churra; on the other hand, *Lactobacillus* and *Oenococcus* were more abundant in Assaf. In Churra milk samples, the presence of of *Cutibacterium* (formerly *Propionibacterium*) which is a common skin inhabitant (Brüggemann et al., 2004) may suggest a potential sample contamination. Moreover no other studies in the field have reported the presence of this species in milk samples. On the other hand, Oikonomou et al. (2014) claimed that geographical conditions and sampling sites have an impact on the microbiome detected in milk samples. These authors could differentiate samples from different farms based on their microbial profile. However, in this work the abundance of this genus is completely uniform across all 212 Churra samples collected across the two considered Churra flocks, which suggests that the sampling has minimized environmental contamination. Regarding to *Lactobacillus* that was recognized as a common microorganism in the core defined for healthy cow milk samples (Derakhshani et al., 2018) and also it is reported as a genus that may inhibit mastitis pathogens (Jara et al., 2011). Finally, *Oenococcus* is commonly used in food dairy industries (i.e. in the production of fermented milk and cheese factories) because is a lactic acid bacteria (McAuliffe et al., 2019). They usually inhabit nutrient-rich environments such as milk, meat, vegetable products and fermented drinks (Kandler, O. & Weiss, 1986).

In addition, this work attempts to identify, if any, the breed-specific species between Churra and Assaf milk samples. To do that, an exploratory statistical linear model was performed using the 30 most abundant. The adjusted R-squared of this model indicates variations in the microbial profiles explained by the breed. A high adjusted R-squared

(higher than 0.7) suggests that those genera might be breed-specific. This hypothesis was confirmed by the later PCA analysis performed (Figure 5), which showed that the variation in those genera was mainly explained by breed (95.5% of the total variance due to the breed factor), and showed the milk samples clearly separated into two different groups related to the two considered groups. These results could be a first step into the application of taxonomic information derived from the analysis of the microbiome in order to ensure traceability and quality labelling of dairy products.

CONCLUSIONS

By exploiting massive paralelle sequencing of the 16S rRNA gene, this study provides a first characterization of the milk microbiota in Spanish Churra sheep, a local double-aptitude breed in the northwest region of Castilla y León. Our results showed that Churra's milk microbiota show a much limited diversity than that previously reported in Spanish Assaf sheep, which is a highly specialized dairy sheep breed. Taking the opportunity to compare the microbiota characterization of milk samples from these two breeds, it seems that both general microbial diversity and microbioal taxonomy may different between different breeds. Moreover, we provide an exploratory study to assess the presence of some breed-specific microbial genera in the sheep milk samples analyzed. Overall, the present work provides a first step into our understanding of the interactions between sheep milk microorganisms and sheep breeds, or rearing systems, and also bewteen sheep milk microbiota and some issues of economical interesnt such as subclinical mastitis resistance and cheese-making efficiency, or the poptential of using information about sheep milk microbiota composition possible for milk traceability, and milk quality classification.

REFERENCES

Addis, M.F., A. Tanca, S. Uzzau, G. Oikonomou, R.C. Bicalho, and P. Moroni. 2016. The bovine milk microbiota: insights and perspectives from -omics studies.. Mol. Biosyst. 12:2359–2372. doi:10.1039/c6mb00217j.

Brüggemann, H., A. Henne, F. Hoster, H. Liesegang, A. Wiezer, A. Strittmatter, S. Hujer, P. Dürre, and G. Gottschalk. 2004. The complete genome sequence of Propionibacterium acnes, a commensal of human skin. Science (80-.). 305:671–673. doi:10.1126/science.1100330.

Catozzi, C., A. Sanchez Bonastre, O. Francino, C. Lecchi, E. De Carlo, D. Vecchio, A. Martucciello, P. Fraulo, V. Bronzo, A. Cusco, S. D'Andreano, and F. Ceciliani. 2017. The microbiota of water buffalo milk during mastitis.. PLoS One 12:e0184710. doi:10.1371/journal.pone.0184710.

Cremonesi, P., C. Ceccarani, G. Curone, M. Severgnini, C. Pollera, V. Bronzo, F. Riva, M.F. Addis, J. Filipe, M. Amadori, E. Trevisi, D. Vigo, P. Moroni, and B. Castiglioni. 2018. Milk microbiome diversity and bacterial group prevalence in a comparison between healthy Holstein Friesian and Rendena cows. PLoS One 13:e0205054–e0205054. doi:10.1371/journal.pone.0205054.

Derakhshani, H., K.B. Fehr, S. Sepehri, D. Francoz, J. De Buck, H.W. Barkema, J.C. Plaizier, and E. Khafipour. 2018. Invited review: Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility.. J. Dairy Sci. 101:10605–10625. doi:10.3168/jds.2018-14860.

Esteban-Blanco, C., B. Gutierrez-Gil, F. Puente-Sanchez, H. Marina, J. Tamames, A. Acedo, and J.J. Arranz. 2019. Microbiota characterization of sheep milk and its association with somatic cell count using 16s rRNA gene sequencing.. J. Anim.

Breed. Genet.. doi:10.1111/jbg.12446.

Gonzalez-Rodriguez, M.C., C. Gonzalo, F. San Primitivo, and P. Carmenes. 1995. Relationship between somatic cell count and intramammary infection of the half udder in dairy ewes.. J. Dairy Sci. 78:2753–2759.

Hiergeist, A., J. Gläsner, U. Reischl, and A. Gessner. 2015. Analyses of intestinal microbiota: culture versus sequencing. ILAR J. 56:228–240.

Hooper, L. V, D.R. Littman, and A.J. Macpherson. 2012. Interactions between the microbiota and the immune system.. Science 336:1268–1273. doi:10.1126/science.1223490.

Hunt, K.M., J.A. Foster, L.J. Forney, U.M.E. Schütte, D.L. Beck, Z. Abdo, L.K. Fox, J.E. Williams, M.K. McGuire, and M.A. McGuire. 2011. Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk. PLoS One 6:e21313.

Jara, S., M. Sanchez, R. Vera, J. Cofre, and E. Castro. 2011. The inhibitory activity of Lactobacillus spp. isolated from breast milk on gastrointestinal pathogenic bacteria of nosocomial origin.. Anaerobe 17:474–477. doi:10.1016/j.anaerobe.2011.07.008.

Kandler, O.; Weiss, N. 1986. Bergey's Manual of Systematic Bacteriology.

de la Fuente, L.-F., G. Fernández, and F. San Primitivo. 1995. Breeding programme for the Spnaish Churra sheep breed. Cah. Options Méditerranéennes 11:165–172.

Love, M.I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550. doi:10.1186/s13059-014-0550-8.

McAuliffe, O., K. Kilcawley, and E. Stefanovic. 2019. Symposium review: Genomic

investigations of flavor formation by dairy microbiota.. J. Dairy Sci. 102:909–922. doi:10.3168/jds.2018-15385.

McInnis, E.A., K.M. Kalanetra, D.A. Mills, and E.A. Maga. 2015. Analysis of raw goat milk microbiota: impact of stage of lactation and lysozyme on microbial diversity.. Food Microbiol. 46:121–131. doi:10.1016/j.fm.2014.07.021.

Oikonomou, G., M.F. Addis, C. Chassard, M.E.F. Nader-Macias, I. Grant, C. Delbès, C.I. Bogni, Y. Le Loir, and S. Even. 2020. Milk Microbiota: What Are We Exactly Talking About? . Front. Microbiol. 11:60.

Oikonomou, G., M.L. Bicalho, E. Meira, R.E. Rossi, C. Foditsch, V.S. Machado, A.G.V. Teixeira, C. Santisteban, Y.H. Schukken, and R.C. Bicalho. 2014. Microbiota of Cow's Milk; Distinguishing Healthy, Sub-Clinically and Clinically Diseased Quarters. PLoS One 9:e85904.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F.O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41:D590–D596. doi:10.1093/nar/gks1219.

Taponen, S., D. McGuinness, H. Hiitiö, H. Simojoki, R. Zadoks, and S. Pyörälä. 2019. Bovine milk microbiome: a more complex issue than expected. Vet. Res. 50:44. doi:10.1186/s13567-019-0662-y.

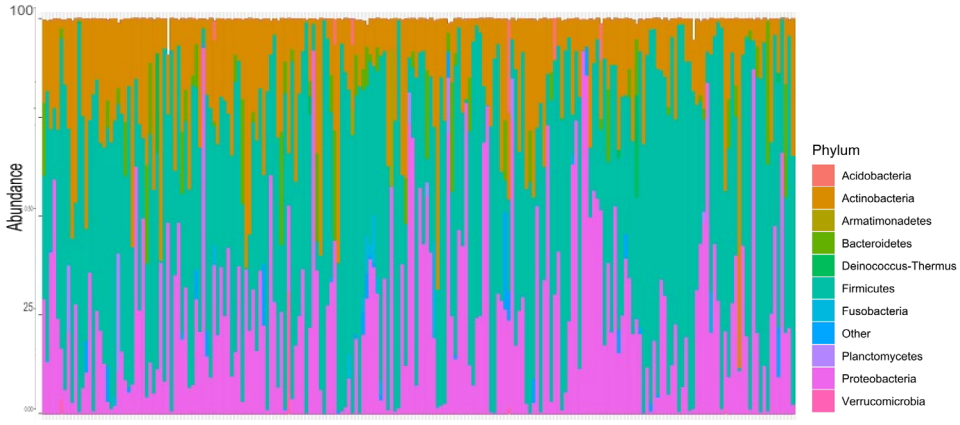
Tilocca, B., N. Costanzo, V.M. Morittu, A.A. Spina, A. Soggiu, D. Britti, P. Roncada, and C. Piras. 2020. Milk microbiota: Characterization methods and role in cheese production. J. Proteomics 210:103534. doi:https://doi.org/10.1016/j.jprot.2019.103534.

Wang, Q., G.M. Garrity, J.M. Tiedje, and J.R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.. Appl. Environ. Microbiol. 73:5261–5267. doi:10.1128/AEM.00062-07.

FIGURES

Figure 1. Taxonomic distribution of bacterial communities in Churra sheep milk samples identified using 16S rRNA gene sequencing at phylum level (A), and at genus level (B). Each bar represents a subject and each coloured box a bacterial taxon. The height of a coloured box represents the relative abundance of that organism within the sample**.**

A)



B)

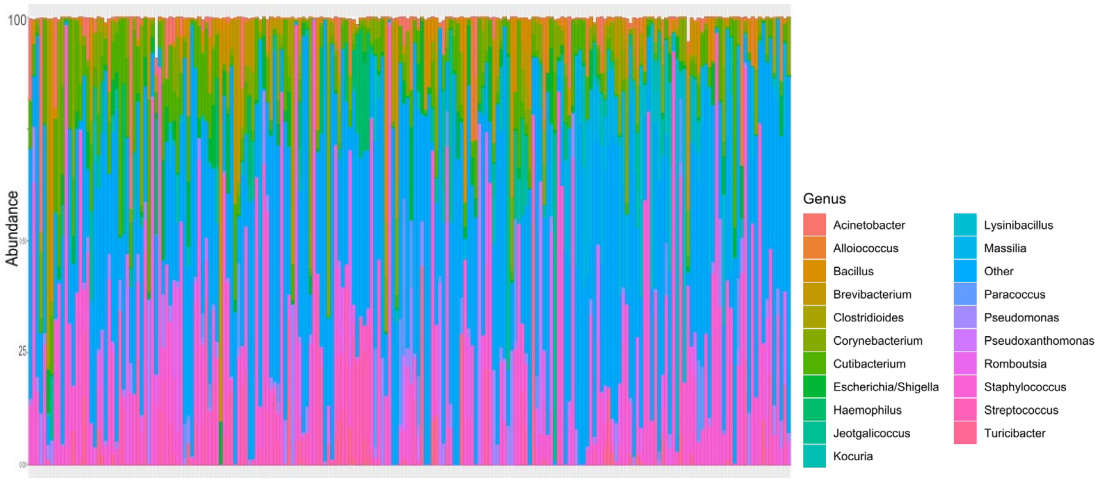


Figure 2. Bacterial alpha diversity determined by Shannon index for milk samples with different levels of somatic cell count (SCC). A) Milk microbial diversity in Churra (A) and Assaf (B) milk samples are shown for the three groups of samples defined in this work based on the SCC thresholds suggested by Gonzalez-Rodriguez et al., (1995): "Healthy", "SM1" (SCC > 400.000 cells/ml) and "SM2" " (SCC > 2,000,000 cells/ml).

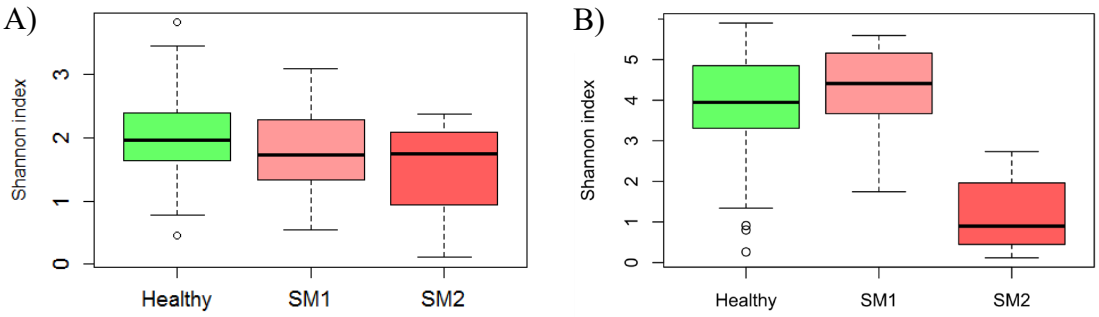


Figure 3. Similarity of bacterial communities within breed. Non-metric Multidimensional Scaling (nMDS) ordination plot of the microbiota results obtained for sheep milk samples of Churra and Assaf sheep breeds. The distance between the samples is based on similarity in ASVs composition of each sample calculated using the Bray-Curtis similarity index.

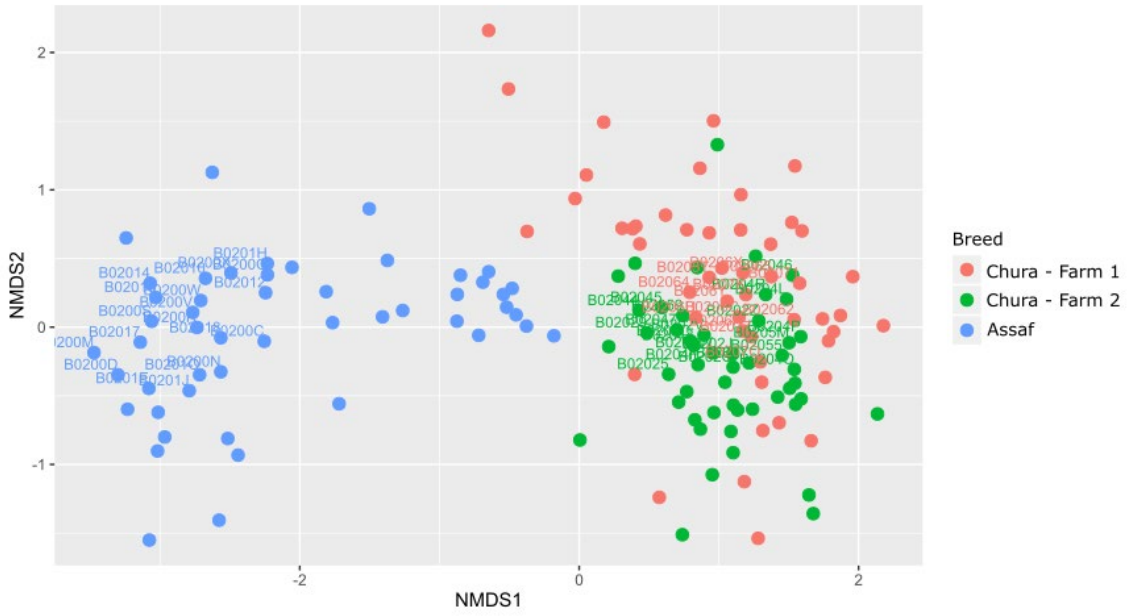


Figure 4. Heatmap of the top 30 genera that were differentially abundant between milk samples from Churra and Assaf sheep. Significantly different bacterial groups are shown at the right side of the heatmap and a dendrogram of the 55 samples is plotted at the bottom of the plot.

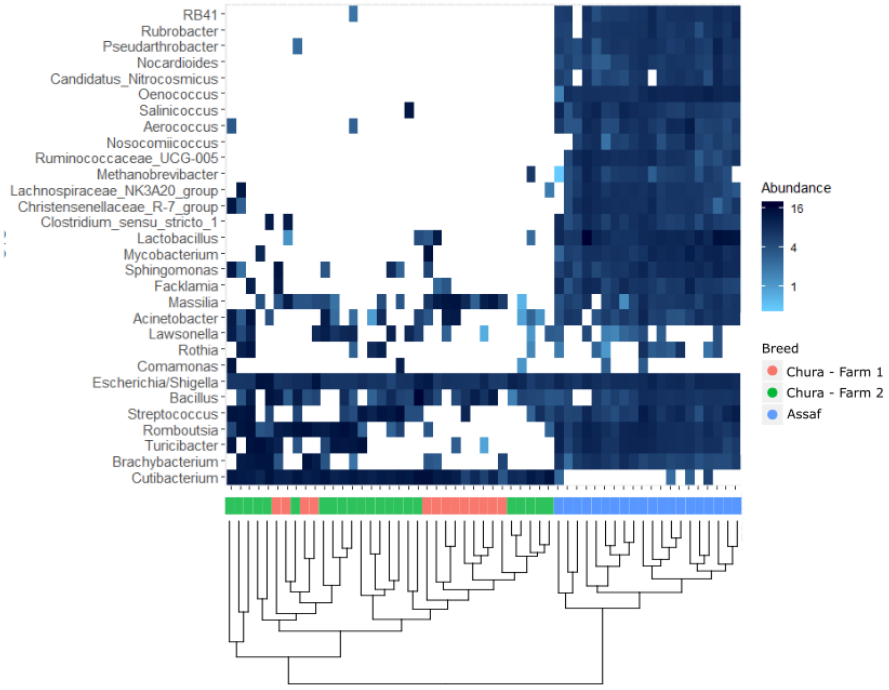
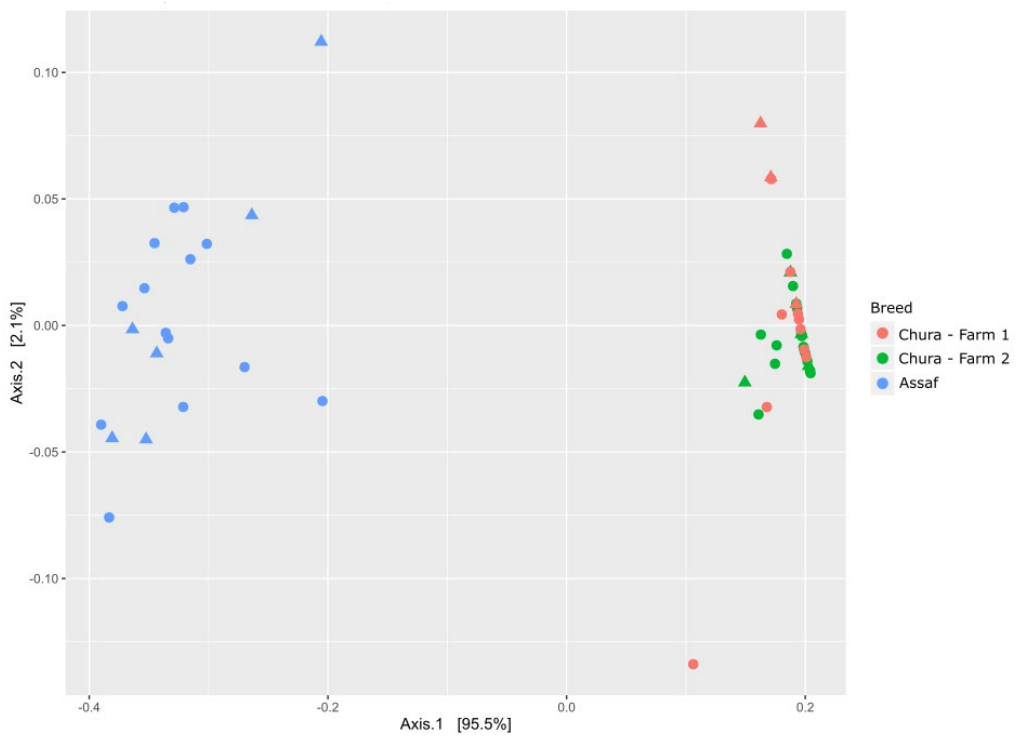


Figure 5. Milk microbiota variations across different breeds. Principal Component Analysis based on Bray-Curtis distance. The results represented here are for those genera showing high values of adjusted R-squared (>0.7) according to the implemented linear model analysis.



TABLES

Table 1. Relative microbial abundances for the top most abundant phyla and genera for the two sheep breeds studied in this work, Churra and Assaf.

Class	Churra	Assaf
At phylum level	*Firmicutes* (50,28%) *Proteobacteria* (25,5%) *Actinobacteria* (18,9%) *Bacteroidetes* (2,6%)	*Firmicutes* (64.44%) *Actinobacteria (14.25%)* *Proteobacteria (9.08%)* *Acidobacteria (2.7%)* *Bacteroidetes (2.3%)*
At genus level	*Staphylococcus*(20,29%) *Cutibacterium* (6,27%) *Corynebacterium* (4,34%) *Streptococcus* (4,1%) *Massilia* (3,5%) *Bacillus* (3,2%)	*Staphylococcus* (16.8%) *Lactobacillus* (14,1%) *Corynebacterium* (8.8%) *Alloiococcus* (6.8%) *Streptococcus* (4%) *Romboutsia* (3%)

TABLES

Table 2. Thirty differentially abundant genera detected with DESeq2 in the Churra *vs.* Assaf analysis contrast ($p_{adj} < 0.05$ and | log2foldChange |> 1.5)

Genus	baseMean	log2FoldChange	lfcSE	Stat	Pvalue	Padj
Oenococcus	171.2271	-10.946522	0.6976906	-15.689651	1.78E-55	5.88E-53
Lactobacillus	1491.1206	-10.270909	1.1379698	-9.025643	1.79E-19	2.56E-18
Ruminococcaceae_UCG-005	83.8875	-9.890707	0.8473458	-11.672575	1.76E-31	7.26E-30
Salinicoccus	78.77975	-9.51449	0.7853946	-12.11428	8.87E-34	4.88E-32
Rubrobacter	55.90108	-9.323135	0.7384876	-12.624632	1.54E-36	1.02E-34
Pseudarthrobacter	65.12921	-9.256819	0.7284744	-12.707131	5.40E-37	4.45E-35
RB41	56.6174	-9.058663	0.7117873	-12.726644	4.21E-37	4.45E-35
Candidatus_Nitrocosmicus	39.30031	-8.79551	0.8650444	-10.167698	2.76E-24	9.12E-23
Nocardioides	38.36621	-8.782761	0.7295485	-12.038626	2.23E-33	1.05E-31
Aerococcus	102.7299	-8.718169	0.9548231	-9.130664	6.81E-20	1.18E-18
Nosocomiicoccus	37.3631	-8.716161	0.9242138	-9.430892	4.07E-21	8.95E-20
Methanobrevibacter	39.81003	-7.996153	0.9616913	-8.314677	9.20E-17	9.79E-16
Lachnospiraceae_NK3A20_group	52.98835	-7.758957	0.9361113	-8.288498	1.15E-16	1.15E-15
Christensenellaceae_R-7_group	68.59827	-7.750322	1.0088776	-7.682123	1.56E-14	1.15E-13
Clostridium_sensu_stricto_1	54.79668	-7.280614	0.9231497	-7.886711	3.10E-15	2.56E-14
Facklamia	52.18613	-6.668425	0.9184782	-7.260298	3.86E-13	2.55E-12
Sphingomonas	118.18714	-4.341307	1.1141301	-3.896589	9.76E-05	2.46E-04
Mycobacterium	129.45354	-4.134871	1.3627183	-3.034282	2.41E-03	4.55E-03
Escherichia/Shigella	986.02125	1.730526	0.59067	2.929768	3.39E-03	5.99E-03

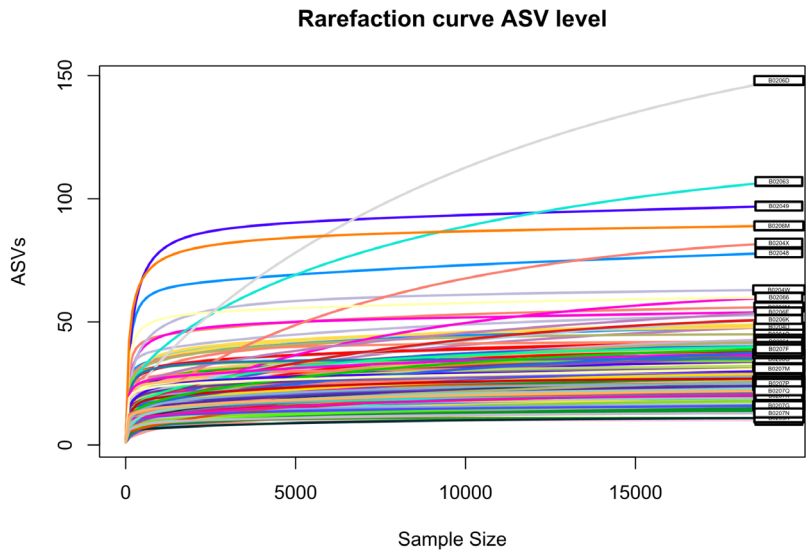
Streptococcus	3269.09037	3.204055	1.205768	2.657274	7.88E-03	1.31E-02
Acinetobacter	645.42554	3.216711	1.2253349	2.625169	8.66E-03	1.38E-02
Romboutsia	3590.56612	3.442382	1.148278	2.997865	2.72E-03	5.01E-03
Turicibacter	3042.20892	3.509678	1.3224483	2.653924	7.96E-03	1.31E-02
Bacillus	838.45589	3.535747	0.8895551	3.974736	7.05E-05	1.83E-04
Massilia	918.54036	3.790561	1.11041	3.413659	6.41E-04	1.35E-03
Brachybacterium	1190.2947	4.520218	1.3375254	3.379538	7.26E-04	1.49E-03
Rothia	82.76503	5.43691	2.066745	2.630663	8.52E-03	1.37E-02
Lawsonella	181.13017	5.92564	1.4271032	4.152216	3.29E-05	8.98E-05
Comamonas	42.29281	6.348138	3.0159017	2.104889	3.53E-02	4.66E-02
Cutibacterium	4394.48013	11.303143	0.7528862	15.013084	6.03E-51	9.95E-49

SUPPLEMENTARY FIGURES

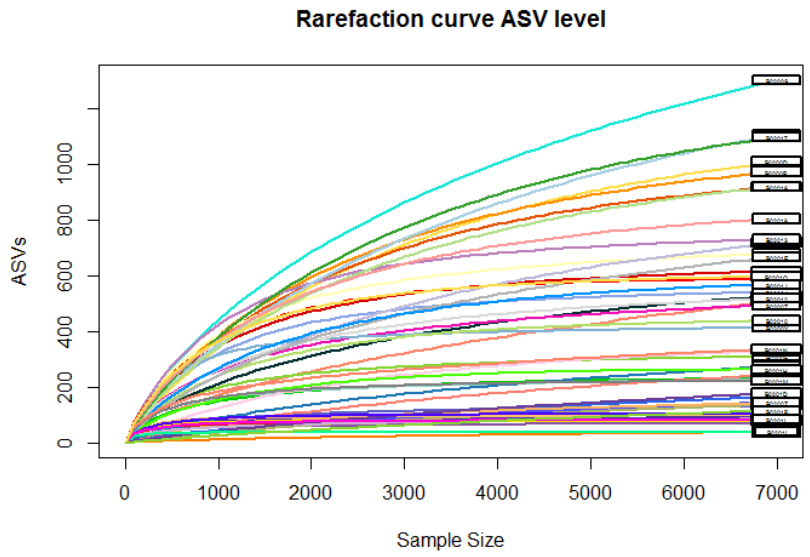
Supplementary Figure 1. Rarefaction analysis of the assessment of ASV coverage.

A) Captured microbial diversity in Churra breed. B) Captured microbial diversity in Assaf breed.

A)



B)



Resultado 2.4

Metagenomic de novo assembly of Corynebacterium bovis in lactating assaf sheep: a preliminary study

C. Esteban-Blanco, F. Puente-Sánchez, B. Gutiérrez-Gil, H. Marina, J. Tamames, J.J. Arranz

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain; [2]2Departamento de Biología de Sistemas, Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain

The 37th International Society for Animal Genetics Conference, Lleida 7-12 de julio de 2019

Metagenomic *de novo* assembly of *Corynebacterium bovis* in lactating assaf sheep: a preliminary study

C. Esteban-Blanco*[1], F. Puente-Sánchez[2], B. Gutiérrez-Gil[1], H. Marina[1], J. Tamames[2], J. J. Arranz[1]

[1]University of León, León, Castilla y León, Spain

[2]CNB-CSIC, Madrid, Madrid, Spain

The sheep milk microbiota is a complex community, which may have a major impact on host health and on the quality of the milk as a food product. High-throughput sequencing enables comprehensive microbial surveys with detection sensitivities higher than earlier molecular techniques as the 16S rRNA gene sequencing. In addition, the computational tools recently developed for metagenomic sequencing analysis attempt to classify the sequences present in a metagenomic dataset into different species. In this context, the aim of this work was to characterize the milk microbiota retrieving taxa present in ovine milk samples using the binning approach. In total, 14 Assaf dairy ewes from a single flock (Zamora, Spain) and without clinical signs of mastitis were included in this study. The samples were classified as derived from healthy and subclinical mastitis dairy ewes based on somatic cell counts (SCC). The SqueezeMeta software pipeline was used in this study for retrieving individual genomes and for analyzing the structure and functionality of microbiomes. After removing host sequence reads (contamination), SqueezeMeta assembled 11.5 million bacterial raw reads into 31,902 contigs. After taxonomic assignment of contigs, the predominant phyla were *Chordata*, *Actinobacteria*, *Firmicutes* and *Proteobacteria*. Some contigs were assigned to a "*Corynebacterium bovis*" bin. Interestingly, the contigs of 3 healthy samples (low SCC) make up almost the total of this bin. Our previous study based on 16S rRNA gene sequencing from 50 Assaf ewes reported that the *Corynebacterium* genus was also one of the most prevalent genera in the microbiota of the sheep mammary gland. In dairy cows, Corynebacteria are usually associated with low SCC milk although specifically, *Corynebacterium bovis* is frequently isolated from milk samples of infected mammary glands. Here, we provide a recovered *Corynebacterium bovis* strain genome from milk sheep samples with low SCC values, which are associated to a healthy mammary gland status.

Key words: sheep, metagenomics, high-throughput sequencing, animal health.

Resultado 2.5

Using shotgun metagenomics to characterize in sheep milk microbiota: a first insight of dairy sheep resistome

C. Esteban-Blanco[1], B. Gutiérrez-Gil[1], J. Tamames[2], F. Puente-Sánchez[2], H. Marina[1], J.J. Arranz[1].

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain; [2]Departamento de Biología de Sistemas, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain.

Manuscript in preparation

Resultado 2.6

Comparison of milk bacterial diversity between two sheep breeds using RNA-Seq datasets

C. Esteban-Blanco, B. Gutiérrez-Gil, H. Marina, A. Suárez-Vega and J.J. Arranz

Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León,

Campus de Vegazana s/n, León 24071, Spain

Short communication in preparation

COMPARISON OF MILK BACTERIAL DIVERSITY BETWEEN TWO SHEEP BREEDS USING RNA-SEQ DATASETS

Esteban-Blanco, C., Gutiérrez-Gil, B., Marina, H., Suárez-Vega, A. and Arranz, J.J.
Dpto. de Producción Animal, Facultad de Veterinaria, Universidad de León, 24071; León

INTRODUCCIÓN

The development of Next-generation sequencing (NGS) has enabled the study of the microbiome of different environments, species and tissues (Ley et al., 2006; Cabrera-Rubio et al., 2012; Oikonomou et al., 2014; Fu et al., 2015; Catozzi et al., 2017). In the last years, microbial diversity and its relation with host health and other host phenotypes have been analysed by studying the entire genomic content of the microbiota. In the last years, metataxonomic approaches have focused on the use of two different sequencing data; 16S rRNA gene sequencing (sequencing specific amplified products from the 16S rRNA gene) and shotgun DNA sequencing (random sequencing across entire genomes). Both are now well-stablished and robust methods used for analysing the microbiome, although there are clear differences between these two approaches. For instance, the 16S rRNA amplicon-based sequencing method is a domain restricted to bacteria and archaea while shotgun metagenomics could infer metabolic pathways from the whole genome sequences of bacterial communities, archaea, and also of viruses and fungi (Norman et al., 2015; Donovan et al., 2018). In livestock, the studies to understand bacterial population dynamics have especially targeted on 16S rRNA sequencing because it is considered simple and cost-effective above shotgun DNA metagenomics. Nevertheless, the study of the host functional genome has generally used other popular omic-approaches, such as transcriptomics, proteomics and metabolomics, which promise highly detailed information (Deusch et al., 2015). Specifically in sheep, the application of ribosomal depleted shotgun RNA sequencing (RNA-Seq) allowed the identification of genes differentially expressed throughout lactation (Suárez-Vega et al., 2016). The assembly of RNA-Seq reads could also reveal bacterial sequences improving metatranscriptome functional annotation (Celaj et al., 2014) and providing an opportunity to use previously generated RNA-Seq dataset for metataxonomic studiess. Furthermore, this kind of approach could also identify viruses and fungi microorganisms and would provide an alternative to the shotgun metagenomic approach. The milk microbiota in sheep has recently been studied using 16S ribosomal RNA sequencing with the Illumina MiSeq platform in a single flock of a Spanish Assaf, a highly productive sheep breed reared in the northwest of Spain (Esteban-Blanco et al., 2019). Because milk microbiota may be more complex than expected, the main objective of this research was to gain a greater understanding of the bacterial dynamic and composition of sheep milk. For that, we examine here the complex mixture of microorganisms and viruses present in milk samples of two sheep breeds reared in the northwest of Spain, Spanish Churra (autochthonous breed of double aptitude) and Assaf (a highly specialized dairy breed), through the analysis of RNA-Seq datasets previously generated from milk somatic cells (MSCs).

MATERIAL AND METHODS

Animals and sampling: For this study, RNA-Seq reads of MSCs with accession number GSE74825 were downloaded from the Gene Expression Omnibus (GEO) database. These RNA-Seq samples had been generated in a previous study where milk samples (50 ml) had been collected from eight healthy sheep of two breeds, four Assaf and four Churra ewes (Suarez-Vega et al., 2016). As described in the related manuscript to those datasets, milk samples from . Milk samples were collected on days 10 (D10), 50 (D50), 120 (D120) and 150 (D150) after lambing had been used for somatic cell RNA extraction and subsequent sequencing with an Illumina Hi-Seq 200 sequencer (Fasteris SA, Plan-les-Ouates, Switzerland). In order to simplify microbial detection and to allow comparison with previously

reported results, the microbiome analysis implemented here has only been applied to the D50 samples generated by Suarez-Vega et al. (2016).

Bioinformatic analyses: The FastQC software (Andrews, 2010) was used to evaluate the quality of the reads resulted from the RNA sequencing. Raw sequences were aligned against the ovine genome assembly v.3.1. (Oar_v3.1 International Sheep Genome Consortium. Ovis aries Oar_v3.1, INSDC Assembly. Ensembl database. 2012) with the aim to remove host RNA contamination and retain only microbial and virus sequences for further metagenomic steps. After host removal, we performed a direct analyses of sequences reads to explore taxonomic composition of the microbiomes in the 8 analyzed samples using a bioinformatic-tool developed for this end (Tamames and Puente-Sanchez, 2018). Shortly, this approach reported abundance of taxa and functions based on the homologies against GenBank NR databases (Kanehisa et al., 2016) using Diamond (Buchfink et al., 2014) and being processed with the LCA algorithm. After taxonomic assignment, a 20% prevalence filter and a minimum of five read per taxa were used to remove low-prevalence genera. We applied the centered log-ratio (clr) transformation, recommended for compositional data (Nawrocki et al., 2009), on the relative abundance taxa matrix in order to perform a Principal Component Analysis (PCoA). The rarefy function of the Vegan package (R) was used to fix sample-size for diversity analyses between the two breed groups considered in this study, Churra and Assaf. Finally, we used negative binomial generalized lineal models, as implemented in the DESeq2 package (Oksanen et al., 2018), exploring differentially abundant taxa (DAT) between the two breeds at genus level. Raw counts of the taxa matrix were normalized for differences in sequencing depth between samples. The DAT were defined as those genera that had an absolute log2-fold change > 1.5 and an adjusted p-value (padj-value) < 0.01.

RESULTS AND DISCUSSION

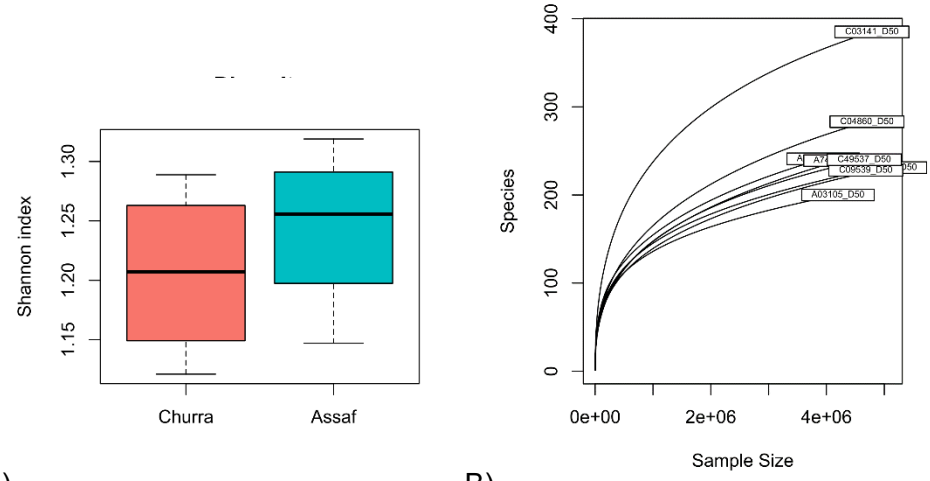
The raw Illumina sequencing dataset analysed here for of eight MSC RNA samples included a total of 317 million paired-end reads of 300 bp. The number of reads per sample ranged between 29 and 46 million raw reads, with an average of 39 million raw reads. More than 29 million raw reads were aligned against the sheep reference genome, accounting for an average percentage of mapped reads higher than 93%. This result was higher than RNA-Seq studies in dairy cattle where only 65% of the total of the reads were mapped to the bovine reference genome (Wickramasinghe et al., 2012). After host removal reads, the total of unmapped reads (22 million reads) was used for the subsequent analysis steps. An average of 16 million reads had hit to the NR database, however 2 million of those reads were taxonomically classified to the Bacteria Kingdom. This fact is in concordance with other studies where the extraction of taxonomic information from RNA-Seq data might be compounded by the abundant host reads (Cox et al., 2017). In addition, more than 50% of the total reads were not taxonomic assignned (labelled "*Unclassified*"), which highlights that this approach using RNA-Seq data might not be and efficient method to explore microbial composition. For the remainder sequences, the most dominant taxa at phylum level was *Chordata*, which suggests that despite the removal of host RNA performed, remnants of host reads can still be found in the dataset, followed by *Firmicutes* and *Proteobacteria*. Other studies in different livestock species reported that the three most abundant phyla in milk were *Firmicutes*, *Proteobacteria* and *Actinobacteria*. Interestingly, the present study revealed that the phylum *Microviridae* was evenly present across all the samples. This phylum has been identified in the gut virome of mammals (Wang et al., 2019), and also in the rumen virome of domestic caprids (Namonyo et al., 2018). Focusing on non-Eukaryotic genera, the two most abundant taxa at genus level were *Staphylococcus* and *Enterococcus*. Both are generally identified in the milk microbiome in several species (Cabrera-Rubio et al., 2012; Addis et al., 2016; Bonsaglia et al., 2017; Esteban-Blanco et al., 2019). Bacterial diversity was explored between the two breeds based on the Shannon index. Low diversity values were observed among samples. Moreover, as shown in **Figure 1A**, no significant differences between Churra and Assaf breeds were found. This results disagrees with a previous work of our group where a vast decrease in bacterial

diversity was observed in Churra compared with Assaf using data from the V4 hypervariable region of 16S rRNA gene (Esteban-Blanco, et al., *unpublished work*). In that study we claimed that samples of the two breeds could be distinguish based on their microbial composition (Bray-Curtis distance), whereas the RNA-Seq-based approach here presented cannot be used to discriminate samples from these two different breeds. The observed discrepancies between these two analyses could be due to the low efficiency of the RNA-Seq approach to identify microbial composition. **Figure 1B** illustrated that this method is not enough to capture the entire diversity among samples. Finally, DESeq2 only identified one significant DAT *Lentivirus* (| logFC | > 25), which was in higher abundance levels in Assaf. It should be noted that the *Lentivirus* genus includes the Visna/maedi virus which has a major impact and causes direct loses in sheep production (Minguijón et al., 2015). In the Castilla y León region of Spain, numerous cases of visna have been diagnosed, mostly in the Assaf breed (Benavides et al., 2006), which agrees with our results. Future studies based on a large sample size and with different sequencing approaches could help understanding the complex microbiota of sheep milk. In addition, the RNA-Seq dataset here analysed could be used to perform a complete study of the sheep milk virome.

REFERENCES

• Addis et al. 2016. Molecular Biosystems 12:2359–2372. • Andrews 2010 Accessed. http://www.bioinformatics.babraham.ac.uk/projects/fastqc • Benavides et al. 2006. Veterinary Record 158:230–235. • Bonsaglia et al. 2017. Scientific Reports, 7:8067. • Buchfink et al. 2014 Nature Methods 12:59. • Cabrera-Rubio et al. 2012. Journal of Clinical Nutrition 96:544–551. • Catozzi et al. 2017. PloS One 12:e0184710. • Celaj et al. 2014. Microbiome 2:39. • Cox et al. 2017. Microbiome 5:7. • Deusch et al. Computational and Structural Biotechnology Journal 13:55–63. • Gonzalez et al. 2018. PloS One 13:e0192898–e0192898. • Esteban-Blanco et al. 2019. Journal of Animal Breeding and Genetics 137(1), 73–83. • Esteban-Banco, C. et al., unpublished work. • Fu et al. 2015. Circulation Research 117:817–824. • Kanehisa et al. 2016. Nucleic Acids Research 44:D457-62. • Ley et al. 2006. Cell 124:837–848. • Minguijón et al. 2015. Veterinary Microbiology 181:75–89. • Namonyo et al. 2018. Archives of Virology 163:3415–3419. • Nawrocki et al. 2009. Bioinformatics 25:1335–1337. • Norman et al. 2015. Cell 160:447–460. • Oikonomou et al. 2014. PLoS One 9:e85904. • Suárez-Vega, et al. 2016. Scientific Data 3:160051. • Tamames, J., and F. Puente-Sanchez. 2018. Frontiers in Microbiology 9:3349. • Wang et al. 2019. Virus Evolution 5. • Wickramasinghe et al. 2012. BMC Genomics 13:45.

Acknowledgments: This work was developed under the framework of AGL-2015-66035-R project financed by the Spanish Ministry of Economy and Competitiveness (MINECO, Madrid, Spain) and co-funded by the European Regional Development Fund. C. Esteban-Blanco is funded by an FPI from MINECO (Ref. BES-2016-07-8080). This research has made use of the high-performance computing resources of the Castilla y León Supercomputing Center (SCAYLE,www.scayle.es).



A) B)

Figure 1. A) Changes in milk microbial diversity using Shannon index showed by the breeds: Churra and Assaf breeds. B) RNA-Seq library rarefaction curves in milk microbiota from healthy ewes.

MILK BACTERIAL DIVERSITY BETWEEN TWO SHEEP BREEDS USING RNA-SEQ DATA

ABSTRACT: The present study explores bacterial composition and diversity of milk in two dairy sheep breeds, Chura and Assaf. The aim of this work was to retrieve data from another RNA-Seq research to identify microorganisms present in sheep milk. Milk samples from eight ewes, four Churra and four Assaf, collected on day 50 of lactation and used for RNA extraction and sequencing were analyzed. After removing host RNA, a total of 22 million reads were used to perform the taxonomic assignment. We verify that the milk RNA-Seq dataset could be used to identify different taxa between samples but it was not enough to capture the entire bacterial diversity among samples. We reported some problems with this kind of approach to determine the microbiota of sheep milk. The results obtained here could not identify alpha diversity differences between milk samples from Churra and Assaf sheep breeds. However, we observed that *Lentivirus*, which is a genus including the ovine maedi-visna virus, showed a significant higher abundance in Assaf milk samples when compared with Churra milk samples.

Keywords: microbiota, RNA-Seq, sheep, maedi/visna.

Resultado 2.7

Diversity and community composition of rumen microbiota in sheep using long reads from nanopore sequencing

C. Esteban-Blanco[1], P.G. Toral[2], C. Fernández-Díez[2], A. Suarez-Vega[1], B. Gutiérrez-Gil[1], O. González-Recio[3,] G. Hervás[2], H. Marina[1], P. Frutos[2], J.J. Arranz[1]

[1]Departamento de Producción Animal, Facultad de Veterinaria, Universidad de León, Campus de Vegazana s/n, León 24071, Spain; [2]IGM (CSIC-Univ. León). León, Spain; [3]Departamento de Mejora Genética Animal, INIA, Madrid. Spain

Short communication in preparation

6. Discusión general

El estudio de la base genética y molecular de caracteres de importancia económica en el ganado ovino se ha visto favorecido, en las últimas décadas, gracias al desarrollo de herramientas genómicas que permiten obtener información genotípica de media y alta densidad (chips de SNPs). A su vez, estas herramientas genómicas han sido el resultado del abaratamiento de las NGS que permitió obtener un genoma de referencia de diferentes especies domésticas, incluida la oveja. Este abaratamiento supuso también un hito en la investigación genómica de las especies domésticas ya que se incrementaron, de manera exponencial, los estudios sobre la base genética de caracteres productivos o funcionales que utilizaban la secuenciación NGS. La WGR en los animales superiores permite obtener información en todas las posiciones del genoma, y no solo la de los marcadores incluidos en los chips de SNPs, lo que hace que sea una técnica muy útil para rastrear mutaciones causales que influyen en los fenotipos de interés. Por otro lado, las NGS han transformado el estudio de los microbiomas, definidos como el conjunto de todos los microorganismos presentes en un ambiente concreto, tal y como se refleja en el aumento exponencial de publicaciones relacionadas con la microbiota en aspectos de salud, biología humana, animal y vegetal. Estos estudios permiten incluso la caracterización de microorganismos previamente no cultivados mediante dos aproximaciones principales: la secuenciación de amplicones (16S rRNA) y la metagenómica.

El desarrollo de las técnicas de secuenciación masiva va acompañado, de manera inherente, al desarrollo de herramientas capaces de analizar la enorme cantidad de los datos de secuenciación generados. En el momento que surgen las NGS se desarrolla de forma exponencial un campo científico interdisciplinario, la bioinformática, que tiene como objetivo explorar grandes volúmenes de datos complejos de origen biológico mediante la combinación de ciencia computacional, estadística, matemáticas, reconocimiento de patrones, aprendizaje automático y enfoques iterativos, entre otros. Esta Tesis Doctoral surge en un momento en el que el grupo de investigación donde desarrollo mi trabajo se plantea la utilización, el desarrollo y la optimización de herramientas bioinformáticas que permitan utilizar la información de las NGS en el análisis de los fenotipos de interés en el ganado ovino de leche.

En este apartado se presenta la discusión general de los principales resultados obtenidos en la presente Tesis Doctoral. Para ello se sigue el guion presentado en el primer apartado de este documento donde se detallan los objetivos planteados. El primer objetivo se centra en la aplicación de un flujo de análisis bioinformático para datos de WGS mientras que el segundo se centra en la utilización datos de secuenciación para explorar la diversidad microbiana de la leche de oveja.

6.1. Utilización de la secuenciación masiva paralela para el estudio de alta definición de regiones con genes de interés económico en el ganado ovino

La primera tarea realizada en esta Tesis doctoral fue la optimización de un flujo de análisis bioinformático (*pipeline*) para la identificación de polimorfismos a lo largo de todo el genoma utilizando datos de secuenciación masiva paralela. Se trata de un flujo de análisis automático optimizado para su utilización sobre una arquitectura de un clúster de supercomputación. La secuenciación masiva de genomas completos en animales superiores produce cantidades muy grandes de datos en formato de texto plano, pudiendo llegar incluso a más de 100GB de datos por muestra. Los tiempos de ejecución disminuyen exponencialmente al utilizar los recursos de un clúster de supercomputación, en el que se puede disponer de miles de núcleos de procesamiento paralelo, para llevar a cabo tareas complejas mediante algoritmos bioinformáticos.

En primer lugar, el *pipeline* se aplicó para llevar a cabo un análisis de alta densidad de marcadores en regiones definidas como huellas de selección relacionadas con un carácter de interés económico en el ganado ovino, la producción de lana. En genómica animal los métodos empleados para la detección de huellas de selección han evolucionado en los últimos años, desde la utilización de microsatélites (Pollinger et al., 2005) hasta el aprovechamiento de herramientas genómicas, como los chips de DNA (Yuan et al., 2017), derivadas de la aparición de las tecnologías de secuenciación de segunda y tercera generación. La alta cobertura de los genomas proporcionada por los chips de DNA es útil para la detección de huellas de selección en el ganado ovino (Kijas et al., 2012b). Utilizando esta metodología, el trabajo 1.1 presenta un análisis preliminar para la identificación de huellas de selección que identifica 5 regiones candidatas a ser huellas de selección para la producción de lana entre dos grupos de razas de ganado

ovino; Merino de lana fina (Australian Industry Merino, Australian Merino y Australian Poll Merino), y No-Merino de lana basta (Churra, Altamurana y Chios). El análisis de alta definición en estas 5 regiones a partir de WGR de 7 genomas individuales (4 Churra y 3 Merino Australiano), identificó 93 marcadores que determinaban un total de 105 variantes funcionales con frecuencias alélicas extremas entre las razas Churra y Merina, que presentan un fenotipo extremo para este carácter (resultado 1.1). Siguiendo este planteamiento, en el trabajo 1.2 nuestro grupo de investigación realizó análisis más exhaustivos para la identificación de huellas de selección entre la razas Churra y varias líneas de Merino australiano en base a los genotipos del chip e SNPs ovino de media densidad (50KChip), y teniendo en cuenta en el planteamiento de los análisis realizados que las regiones identificadas como huellas de selección podrían afectar a cualquier tipo de carácter para los cuales estas dos razas diferían, incluyendo en esos caracteres divergentes, además de las características de la lana, otros como la velocidad de crecimiento, el tamaño corporal, etc. De esta manera, y en base a tres métodos de mapeo de huellas de selección diferentes se identificaron 18 regiones candidatas de convergencia (CCR) a ser huellas de selección (3 para Merino y 15 para Churra). La coincidencia de varias de esas regiones con huellas de selección previamente descritas, con QTL sobre caracteres productivos y en algunos casos, con claros genes candidatos, por ej. *NCAPG*, *LCORL* y *EIF2S2*, para huellas de selección en Merino, y *RXFP2* y *NPR2* en relación a dos huellas de selección en Churra, apoyó la validez de los resultados identificados. Además, en el trabajo 1.2 el análisis de alta definición de las regiones identificadas como huellas de selección, se basó en la información derivada del análisis de la secuencia del genoma completo de 28 individuos, secuenciando por un lado, 13 ovejas de raza Churra y por otro, descargando 15 genomas de Merino Australiano disponibles en el repositorio SRA (https://www.ncbi.nlm.nih.gov/sra). La secuencia de los genomas de estos 28 animales se utilizó para estudiar detalladamente la distribución de la variación genética en las 18 regiones. Gracias a la información obtenida a través de los datos de WGR, el estudio de alta definición identificó 1.291 variantes con frecuencias divergentes entre las dos razas analizadas. De ellas, 257 eran variantes intragénicas distribuidas en: 31 genes que codifican para proteínas (275 variantes funcionales, 199 intrónicas), dos pseudogenes (uno de ellos identificado como ortólogo de TPI1 bovino), un rRNA (5S_rRNA) y dos snRNA. Después de estudiar en profundidad

el papel funcional y el posible impacto biológico de todas las mutaciones intragénicas significativamente asociados con la identidad racial localizadas dentro de las 18 huellas de selección detectadas, se identificaron cuatro mutaciones determinantes de un cambio de aminoácido (*missense*), una ellas en el gen *NPR2,* localizado en una huella de selección en el cromosoma (OAR) 2, asociada a la raza Churra, y las otras tres mutaciones en los genes *NCAPG* y *LCORL,* localizados dentro de una de las huellas de selección localizada en OAR6 y asociada al Merino Australiano. De estas mutaciones, solo la mutación *NCAPG_Ser585Phe* fue inferida, a través del análisis de anotación funcional, como deletérea. En base a esto, al alto nivel de conservación filogenética del residuo aminoacídico afectado por esta mutación, y a los conocidos efectos directos del gen *NCAPG* sobre el crecimiento corporal en mamíferos (Eberlein et al., 2009; Sahana et al., 2015), esta mutación fue propuesta en este trabajo como posible SSN, es decir, como la posible mutación causal de la huellas de selección previamente identificada en la región Chr6: 37.164–38.580 Mb. El gen *NCAPG* se encuentra en la misma región genómica que el gen *LCORL*, ambos relacionados con caracteres de crecimiento, tamaño, peso y altura en ganado vacuno (Eberlein et al., 2009), en caballos (Metzger et al., 2013) y también en el ganado ovino (Rochus et al., 2018). Estudios recientes en el ganado vacuno identifican mutaciones asociadas a huellas de selección para el tamaño corporal en los genes *LCORL* y *NCAPG* (Chen et al., 2020). En cabras, incluso, se ha propuesto utilizar una variante en el gen *LCORL* que aumenta el tamaño corporal para mejorar programas de reproducción y selección (Saif et al., 2020). Todos estos estudios apoyan que la región *NCAPG-LCORL* es muy importante en relación a efectos fenotípicos directos sobre caracteres de tamaño y crecimiento en las diferentes especies domésticas, y futuros estudios debieran confirmar el posible efecto causal propuesto en el trabajo 1.2 para la mutación *NCAPG_Ser585Phe* en ganado ovino, por ejemplo a través de su genotipado en poblaciones comerciales con datos de morfología corporal y medidas de estatura o peso al nacimiento.

Por otra parte, aunque los genes mencionados son interesantes para estudiar la divergencia selectiva entre las razas comparadas, los resultados obtenidos no proporcionan una clara información sobre genes directamente relacionados con el carácter de estudio inicial, la producción de lana. Un estudio reciente que explora más

a fondo las huellas de selección comparando seis razas de ovejas de origen merino y, cinco no merino, históricamente no seleccionadas para la calidad de la lana, ha detectado nuevas huellas de selección, superpuestas además con QTL previamente descritos para caracteres de la lana en los cromosomas OAR17 y OAR18 (Megdiche et al., 2019). Nuestro estudio también identificó regiones candidatas a ser huellas de selección para este carácter en estos dos cromosomas pero ninguna coincide ni se superpone con las regiones descritas en el trabajo de Megdiche et al. (2019), que fueron detectadas utilizando un método menos habitual basado en el análisis de "ancestros locales en poblaciones cruzadas" (Sankararaman et al., 2008). Las diferencias y la no-convergencia de los resultados hace pensar que las metodologías utilizadas para identificar huellas de selección en estos dos trabajos no tiene resolución suficiente para detectar todas las huellas de selección presentes en el genoma ovino. En otras especies como el ganado vacuno de raza Holstein, se sugiere que para detectar huellas de selección en todas las regiones del genoma se requiere genotipar a los individuos con chips de por lo menos 150.000 SNPs repartidos uniformemente a lo largo de todo el genoma (Barendse et al., 2009). En los últimos años, la información de polimorfismos obtenida a partir de WGR, ha demostrado ser eficaz para la detección de huellas asociadas a la domesticación, adaptación, caracteres de producción y reproducción en varias especies de ganado, como la cabra (Guo et al., 2018), la oveja (Li et al., 2019), el búfalo (Luo et al., 2020), el cerdo (Li et al., 2020) y el ganado vacuno (Weldenegodguad et al., 2019). Cuando se propusieron los objetivos de los trabajos 1.1 y 1.2, la cantidad de genomas disponibles era mucho menor que en la actualidad y, por lo tanto, no disponíamos de un número suficiente de genomas para plantear la detección de huellas de selección a partir de datos de WGR. Hoy en día, nuestro grupo de investigación ha secuenciado más de 40 genomas de individuos de raza Churra y, en el repositorio público SRA (https://www.ncbi.nlm.nih.gov/sra), hay disponibles más de 130 genomas de Merino Australiano. Por lo tanto, se podrían plantear futuros estudios de verificación del mapeo de huellas de selección aquí descrito, utilizando para ello directamente la información WGR disponible con el fin de identificar nuevas regiones como candidatas a ser huellas de selección o, en otros casos, refinar el intervalo de las previamente identificadas, todo ello con el fin de ampliar el conocimiento que hoy tenemos sobre las huellas que la selección ha determinado, o determina, en el genoma ovino.

En segundo lugar, se llevó a cabo la búsqueda de variantes en todo el genoma mediante la utilización del flujo de análisis bioinformático optimizado en esta Tesis Doctoral sobre tríos segregantes para caracterizar la base genética de un QTL con efectos sobre la resistencia a la mastitis en el ganado ovino (trabajo 1.3). Inicialmente se realizó un genotipado con un chip de 50K SNPs en una población de ovejas de raza Churra con registros para el carácter SCS (del inglés *Somatic Cell Score*, logaritmo del recuento de células somáticas), clásico indicador del estado sanitario de la glándula mamaria. De todos los QTL identificados nos centramos en uno localizado en el cromosoma OAR20, por ser el más significativo y porque replica resultados anteriores obtenidos por nuestro grupo de investigación en una población independiente de la misma raza (Gutiérrez Gil et al., 2007). Para este QTL localizado en OAR20 se identificaron, mediante análisis LA, dos familias segregantes pero que diferían en la región estimada como intervalo de confianza más probable en base a los respectivos análisis intrafamiliares. La estrategia seguida en esta Tesis Doctoral ha sido seleccionar dos tríos de animales compuestos por el padre, cabeza de cada familia segregante, y dos de sus hijas con fenotipos divergentes para el carácter SCS y portadoras en homocigosis de los haplotipos alternativos del padre para los marcadores incluidos en el intervalo de confianza considerado, con objeto de (re)secuenciar su genoma. La hipótesis de partida de esta aproximación es que los padres son heterocigotos para el carácter (Qq) y cada una de las hijas tiene alta probabilidad de ser homocigota para cada uno de los alelos alternativos del QTL, es decir, QQ y qq, respectivamente. Aunque disponíamos del genoma de cada individuo con una redundancia media de 15X, el intervalo genómico que se analizó para la detección de variantes con alta resolución fue el correspondiente a la combinación de los intervalos de confianza, parcialmente superpuestos, de las dos familias, con una extensión de 27,5 MB, en la región comprendida entre 14,2 y 41,7Mb del OAR20. En esta región se detectaron 227.030 variantes de las que solo 47.838 fueron concordantes con el patrón de segregación del QTL en estudio, es decir, heterocigotas en los padres y homocigotas para los alelos alternativos en cada una de las hijas. Para las tres pruebas de concordancia realizadas, y dado el elevado número de variantes concordantes detectadas, decidimos centrarnos en valorar el posible papel a nivel funcional de las variantes localizadas en los exones, sin entrar a valorar variantes en regiones no codificantes o intergénicas, a pesar de que cada vez más estudios están demostrando

que estas variantes pueden afectar a la expresión génica (Igartua et al., 2017). De esta manera, como aproximación inicial, realizamos la anotación funcional de las variantes localizadas en regiones codificantes para intentar identificar alguna posible mutación que fuese buena candidata a ser el QTN del QTL en estudio. Siguiendo este razonamiento, se seleccionaron, como mutaciones de interés, 85 variantes concordantes con el QTL, que producían cambio de aminoácido (*missense*) con efecto deletéreo, y que se localizan en genes relacionados con la inmunidad, en un total de 23 genes concretamente. Estas mutaciones son, desde nuestra aproximación, potencialmente candidatas para explicar el efecto del QTL para SCS localizado en OAR20, es decir, posibles QTN. A pesar de estar todos los genes portadores e estas mutaciones relacionados con la respuesta inmune los resultados generados hasta el momento no nos permitieron identificar un claro gen candidato funcional que pudiera ser responsable directo del cambio en el fenotipo, es decir, de la posible resistencia/susceptibilidad a la mastitis. No podemos olvidar que la resistencia a la mastitis es un carácter cuantitativo y como tal, es un carácter complejo de naturaleza poligénica (Rupp and Boichard, 2003), por lo que las mutaciones aquí descritas deberían analizarse en conjunto y no tratar de explicar un alto porcentaje de la variación en el fenotipo solo con un QTN, ya que esta variación puede ser la combinación de más de un QTN (Glazier et al., 2002).

Otros autores consiguieron identificar en una población comercial de ganado ovino de raza Lacaune, utilizando la misma metodología descrita en el trabajo 1.3 (detección de QTL para el recuento de células somáticas mediante estudios de GWAS seguido de un mapeo fino en regiones concretas con WGR de un trío segregante, una mutación candidata en el gen *SOCS2* (supresor altamente conservado del gen de señalización de citoquinas 2), que explica el efecto de un QTL detectado en OAR3 asociado a una inflamación crónica de la glándula mamaria(Rupp et al., 2015b). Además, estudios de validación posteriores para esta mutación han confirmado el efecto pleiotrópico de la mutación en *SOCS2* para el crecimiento corporal y para la producción de leche (Oget et al., 2019). Estos trabajos demuestran que la metodología utilizada es válida para la detección de mutaciones causales o QTN que explican el efecto fenotípico de un QTL detectado (Rupp et al., 2015b; Oliveira Júnior et al., 2019). No obstante, nuestro trabajo

no replica el QTL del OAR3 detectado en la raza Laucane y Rupp y colaboradores (2015) tampoco detectan el QTL del OAR20 de nuestro trabajo, lo que indica que ambos QTL son principalmente específicos de la población dónde se han detectado y, por lo tanto, es difícil utilizar estos resultados en selección asistida por marcadores, ya que es posible que haya QTN que segregan solo en algunas razas. Hay que tener en cuenta, además, que el pequeño efecto de los QTN hace que sea muy complicado identificar, sin género de dudas, la mutación responsable de un determinado efecto. En este sentido, algunos autores han recopilado los requisitos más importantes que debe cumplir una mutación para ser considerada QTN (Ron y Weller, 2007) y hasta el momento muy pocas mutaciones han verificado la mayoría de los requisitos indicados.

Aunque esta metodología basada en explotar información de datos WGR para identificar mutaciones potencialmente causales ha demostrado ser útil para la detección de QTN en el ganado ovino, una de sus principales limitaciones en especies animales es que la anotación de los genomas aún está en una etapa temprana de desarrollo (Murdoch, 2019), y puede ser que la mutación causal se encuentre en un elemento regulador no anotado o en algún gen no identificado en la región. De hecho, explorando solo regiones exónicas en el mapeo fino con WGR de tríos segregantes se omiten variantes en intrones y regiones intergénicas, y aunque éstas explican una baja proporción de la variancia genética para caracteres complejos (Koufariotis et al., 2014; Wang et al., 2020), se ha demostrado que las variantes encontradas dentro de intrones pueden afectar a los fenotipos de los mamíferos (Wu et al., 2012; Nietfeld et al., 2020; Verdura et al., 2020; Wang et al., 2020). La secuenciación de genomas es el enfoque más completo para medir la variación de regiones codificantes y no codificantes, sin embargo, su aplicación para estudios de ligamiento y asociación está limitada debido al coste de secuenciar muchos individuos. Una alternativa efectiva a este problema sería combinar la imputación de genotipos con la secuenciación dirigida de genes candidatos (Martínez-Bueno and Alarcón-Riquelme, 2019).

Asimismo, la imputación de genotipos a la densidad de WGS ha emergido en los últimos años como un técnica para aumentar la precisión de las predicciones de valores genómicos en especies domésticas, concepto en el que se basa la GS (Meuwissen et al., 2001). Inicialmente se utilizaron marcadores aleatorios a lo largo de todo el genoma y

gracias al desarrollo de herramientas genómicas como los chips de SNPs y la disminución de los precios de genotipado, la GS se está convirtiendo en el procedimiento estándar para las evaluaciones genéticas en las especies lecheras como la vaca (Strandén et al., 2019; Peñagaricano, 2020) y la oveja (Duchemin et al., 2012; Cesarani et al., 2019; Gootwine, 2020). Todo lo expuesto, unido a que la precisión de la imputación de genotipos en razas de ganado ovino es alta (Bolormaa et al., 2019), se propone que estudios futuros contemplen la posibilidad de utilizar los datos de genotipado de la población del trabajo 1.3 e imputar a la densidad de WGS. De esta manera se podrían realizar estudios de ligamiento y GWAS para los 1680 animales, inicialmente genotipados para el chip 50K de SNPs, en base a genotipos imputados a la densidad de WGS, con el objetivo de incrementar la posibilidad de identificar mutaciones que sean claras candidatas a explicar los efectos QTL previamente identificados para el carácter susceptibilidad a la mastitis y así poder esclarecer la complejidad de este carácter de interés en el ganado ovino.

Hasta este punto, se han descrito las metodologías utilizadas en esta Tesis Doctoral, sus limitaciones y alternativas, para la búsqueda de posibles mutaciones causales o QTN putativos.

6.2. Utilización de la secuenciación masiva paralela para caracterizar la microbiota de la leche de oveja y su posible asociación con caracteres de resistencia a la mastitis

Para llevar a cabo el segundo objetivo de la presente Tesis Doctoral se ha utilizado principalmente la secuenciación de algunas de las regiones hipervariables del gen que codifica para la subunidad 16S del RNA ribosómico bacteriano, comúnmente denominada secuenciación 16S rRNA. La plataforma seleccionada para analizar muestras de leche fue Illumina MiSeq, que comparada con otros instrumentos de secuenciación, requiere menos DNA, tiene menor tasa de error con una calidad de lectura alta y costes relativamente bajos (Caporaso et al., 2012; Quail et al., 2012). Adicionalmente, se ha explorado la composición taxonómica de la leche utilizando secuenciación metagenómica de alto rendimiento (en inglés *shotgun metagenomics*) que proporciona, en principio, mayor resolución taxonómica, información directa sobre vías funcionales, puede identificar secuencias de todos los microorganismos de la

7. Conclusiones

PRIMERA,

El análisis de un chip de SNPs de media densidad permitió la identificación de 18 huellas de selección al comparar la raza Churra con el Merino Australiano. El posterior estudio de alta resolución con datos de secuenciación de genomas completos puso de manifiesto al polimorfismo *NCAPG_Ser585Phe* como posible mutación causal o SSN (*selection sweep nucleotide*) de una de las huellas de selección identificadas.

SEGUNDA,

La aproximación de secuenciación de tríos segregantes, formados por un padre heterocigoto y dos hijas con valores genéticos extremos para el carácter recuento de células somáticas y haplotipos homocigotos alternos, ha hecho posible la identificación de una serie de variantes genómicas candidatas a explicar un QTL para este carácter, previamente descrito en el cromosoma 20 ovino.

TERCERA,

La secuenciación masiva de la región hipervariable V4 del gen que codifica para la subunidad 16S del RNA ribosómico bacteriano ha permitido, por mi primera vez, la caracterización de la microbiota de la leche de ovejas fenotípicamente sanas. El estudio de animales de las razas Churra y Assaf ha revelado que el factor “raza” tiene una importancia significativa en la composición de la microbiota.

CUARTA,

La optimización y utilización de flujos de análisis bioinformáticos ha ofrecido un procedimiento adecuado para profundizar en el conocimiento sobre la base genética del carácter “resistencia a la mastitis subclínica”, de gran interés económico en el ganado ovino. Por un lado, mediante la identificación de mutaciones potencialmente causales de QTL y por otro, mediante la evaluación de la posible asociación de la microbiota de la leche con el recuento de células somáticas.

8. Resumen

El planteamiento de la presente Tesis Doctoral se realiza gracias al desarrollo de las tecnologías de secuenciación de segunda generación (NGS del inglés *next generation sequencing*) en los últimos años y su aplicación en investigación genómica animal. El avance de diferentes áreas, entre las que se incluyen la genética, la biotecnología, la biología molecular, la ingeniería electrónica y la informática, ha permitido desarrollar herramientas que han servido para estudiar ampliamente la compleja arquitectura genética de caracteres de interés económico en las especies de animales domésticos. El objetivo general de esta Tesis Doctoral es el aprovechamiento y la utilización de las NGS para estudiar caracteres complejos y algunos fenotipos de importancia económica en el ganado ovino lechero. Esta Tesis se ha desarrollado en el grupo de Mejora Genética Animal (MEGA) de la Universidad de León (ULE), grupo principalmente centrado en la mejora de la producción lechera en el ganado ovino, mediante el estudio de parámetros genéticos y de factores ambientales, de variantes genómicas que influyen sobre la variabilidad tanto de caracteres relacionados con la producción y la composición lechera, como con caracteres funcionales como la morfología o la resistencia a enfermedades.

En función de lo expuesto, esta Tesis presenta varios trabajos de investigación en los que se han utilizado diferentes tecnologías NGS y flujos de análisis bioinformáticos para el estudio genómico de caracteres de interés en el ganado ovino lechero. En primer lugar, se ha utilizado la resecuenciación de genomas completos (WGR; *whole genome resequencing*) para el estudio de alta resolución de varias regiones del genoma ovino identificadas previamente como huellas de selección, así como de una región identificada como un QTL (del inglés, *quantitavie trait locus*) con influencia sobre la resistencia a la mastitis. En ambos casos, estos estudios de alta resolución han tenido por objetivo intentar descifrar la base genética de regiones genómicas con un efecto directo sobre fenotipos de interés. En segundo lugar, se utilizaron otras dos estrategias de NGS, la secuenciación de regiones hipervariables del gen 16S rRNA microbiano y la secuenciación metagenómica de alto rendimiento para caracterizar por primera vez la microbiota de la leche de oveja y valorar su posible asociación con el carácter resistencia a la mastitis. Adicionalmente, se han explorado otros tipos de metodologías de secuenciación, como la secuenciación masiva paralela de RNA (RNA-Seq) y la secuenciación de lecturas largas mediante nanoporos (ONT del inglés, *Oxford Nanopore*

Technologies), para caracterizar la microbiota de leche y de rumen de oveja, respectivamente.

En la primera etapa de la presente Tesis Doctoral, se optimizó un flujo de análisis bioinformático que utiliza datos WGR para la detección de variantes génicas. Gracias a este flujo de análisis se identificaron polimorfismos en regiones definidas como huellas de selección entre dos razas ovinas, la raza Churra y el Merino Australiano. De todas las variantes genéticas identificadas en 18 regiones candidatas a ser huellas de selección, la anotación funcional *in silico* realizada destacó, entre las 296 variantes funcionales intragénicas asociadas con la identidad racial entre muestras de Churra y Merino, una única variante con un potencial efecto funcional deletéreo, la mutación *NCAPG_Ser585Phe.* Esta mutación fue catalogada como posible mutación causal de una de las huellas de selección (SSN, del inglés *selection sweep nucleotide*), en concreto de la huella de selección localizada en el cromosoma 6 ovino en la región 37,1–38,5 Mb, y descrita previamente por otros autores. Por otro lado, se aplicó el flujo de análisis bioinformático descrito anteriormente para la detección de variantes sobre datos WGR de dos tríos segregantes para un QTL localizado en el cromosoma 20 y relacionado con la resistencia a la mastitis en el ganado ovino. Los tríos de animales estaban compuestos por el padre, cabeza de cada familia segregante, y dos de sus hijas con valores genéticos extremos para el carácter recuento de células somáticas (SCS del inglés, *somatic cell score*) y con haplotipos homocigotos alternos para los marcadores incluidos en el intervalo de confianza estimado para el QTL. Gracias al análisis de alta definición realizado en dicho intervalo (14,2-41,7 Mb) se identificaron varios polimorfismos que podrían considerarse en futuros estudios de confirmación como posibles mutaciones relacionadas con la resistencia/susceptibilidad a la mastitis en el ganado ovino.

En la segunda etapa de esta Tesis Doctoral se llevó a cabo la caracterización de la microbiota de la leche de oveja de raza Assaf utilizando dos estrategias: la primera, la secuenciación de la región hipervariable V4 del gen que codifica para la subunidad 16S del RNA ribosómico bacteriano, y la segunda, la secuenciación metagenómica de alto rendimiento. Se identificaron los microorganismos que forman parte del núcleo central del microbioma de la leche de oveja en la raza Assaf: *Staphylococcus*, *Corynebacterium*, *Streptococcus, Lactobacillus* y *Escherichia*/*Shigella*. Además, se observó una disminución

drástica de la diversidad bacteriana en las muestras de leche con altos valores para el recuento de células somáticas (SCC del inglés, *somatic cell count*) que se consideran un indicador de mastitis subclínica. Aunque en el ganado ovino se necesitan más estudios en este campo, estos resultados indican que el aumento de bacterias patógenas asociadas a procesos inflamatorios subclínicos de la ubre podría estar asociado con una disminución de los microorganismos comensales de dicho órgano. Asimismo, en esta Tesis Doctoral se ha comparado la microbiota de la leche de oveja de raza Assaf con la de raza Churra y se ha observado que la primera es mucho más diversa que la segunda lo que puede sugerir un posible papel de la raza sobre la variación en la composición de la microbiota de la leche de oveja. Cabe destacar que, de forma complementaria a estos estudios sobre la microbiota de la leche ovina, esta memoria presenta también una aproximación para explorar el *resistoma* del ecosistema mamario de la oveja basada en datos de secuenciación de alto rendimiento en muestras de leche de ovejas de raza Assaf. Este estudio ha identificado la presencia, en las muestras analizadas, de genes bacterianos que confieren resistencia a tres tipos de antibióticos, tetraciclinas, aminoglucósidos y betalactámicos, todos ellos de utilización mayoritaria en explotaciones comerciales ovinas de manejo intensivo, como el utilizado en la raza Assaf. Además, el estudio de la microbiota de la leche de oveja utilizando RNA-Seq ha servido como paso inicial para estudiar el *viroma* de la leche de oveja. Este análisis ha permitido identificar el género *Lentivirus*, dentro del cual estaría el agente causante del maedi-visna, con una abundancia diferencial incrementada en muestras de leche de ovejas Assaf, comparadas con las de ovejas de raza Churra. Finalmente, la tecnología de secuenciación de tercera generación (ONT) ha servido para caracterizar taxonómica y funcionalmente la microbiota de muestras de rumen en el ganado ovino.

Todos los trabajos incluidos en esta Tesis Doctoral proporcionan una visión global de la gran utilidad del uso de las tecnologías de secuenciación para profundizar en distintos aspectos relacionados con la genómica animal. La información derivada de estos trabajos servirá de base para el desarrollo de futuros proyectos relacionados con el conocimiento -ómico de los caracteres productivos de interés económico en el ganado ovino, además de justificar el planteamiento de estudios más específicos que utilicen los

resultados aquí presentados como punto de partida para su confirmación en otras poblaciones ovinas.

9. Summary

The proposal of this PhD Thesis settles on the large advances that Next Generation Sequencing (NGS) techniques have shown in the last years and on their application in animal genomic research. The progress in different areas, such as genetics, biotechnology, molecular biology, electronic engineering, and computing, has allowed the development of tools to study the complex architecture of traits of economic interest traits in livestock. The general aim of this PhD Thesis was the use of NGS technologies to study complex traits and some economically important phenotypes in dairy sheep. This Thesis has been developed within the research group of Animal Breeding of the University of León, (MEGA-ULE). The research activity of this group is focused on the improvement of dairy sheep, through the study of genetic parameters and environmental factors, genomic variants that influence the variability of traits related to milk production and dairy composition, and functional traits such as morphology or disease resistance.

Taking all this into account, this Thesis includes several research studies using different NGS technologies and bioinformatic approaches for the genomic analysis of important traits in dairy sheep. Firstly, the Whole Genome Resequencing (WGR) approach for the high-resolution study of several regions of the sheep genome previously identified as putative selection signals, as well as one region already defined as a Quantitative Trait Locus (QTL) underlying mastitis resistance. The objective of both high-resolution studies was to decipher the genetic basis of genomic regions that have a direct effect on the phenotype. Secondly, two other NGS sequencing approaches, the sequencing of the hypervariable V4 region of the 16S rRNA gene and shotgun metagenomic sequencing, were used to characterise, for the first time, the microbiota of sheep milk and to assess its potential association with the trait resistance to mastitis. Besides, other two different sequencing methodologies were explored, the massive parallel RNA sequencing (RNA-Seq) and the long reads sequencing from Oxford Nanopore Technologies (ONT), to characterise the sheep milk and rumen microbiota respectively.

In the first stage of this PhD Thesis, a WGR bioinformatic analysis pipeline was optimised to identify genomic variants. This pipeline allowed the identification of polymorphisms in regions previously defined as selective signatures between Churra and fine-wool Australian Merino breeds. The variant calling and functional annotation analyses

performed for the 18 regions defined as selection sweeps highlighted, among the 296 intragenic variants associated with the breed identity, one mutation with a potential deleterious functional effect, the *NCAPG_Ser585Phe* variant. This mutation was suggested to be the putative selection sweep nucleotide (SSN) of one of the detected selection signatures, located on ovine chromosome 6, at 37,1-38,5 Mb, and which had previously been identified by other authors. On the other hand, the same bioinformatic analysis pipeline was applied to detect mutations in WGR data of two segregating trios (one sire and two daughters) for one QTL located on sheep chromosome 20 influencing mastitis resistance in sheep. Each segregating trio included the corresponding sire, and two daughters with extreme divergent phenotypes for the somatic cell score trait (SCS), in correspondence with the alternative homozygosity haplotypes of markers included in the estimated confident interval QTL. The high-resolution analyses performed on such interval for the studied QTL (14.2-41.7 Mb) identified a list of polymorphisms that could be considered by future confirmation studies as potential mutations related to the resistance/susceptibly mastitis in sheep.

The second stage of this Thesis was focused on the characterisation of Assaf sheep milk microbiota using two different approaches: firstly, the sequencing of the V4 region of the 16S rRNA gene and secondly, a metagenomic shotgun sequencing approach. Based on this, the core microbiota of milk samples was described for the first time for the Assaf breed as including the following genera: *Staphylococcus*, *Corynebacterium*, *Streptococcus, Lactobacillus* and *Escherichia*/*Shigella*. Moreover, a remarkable concomitant reduction of microbial diversity was observed for samples with high SCC values. Although further studies in this field are required, the results reported here suggest that the increase in pathogenic bacteria associated to subclinical inflammation of the mammary gland might be associated with a decrease in commensal and natural hosts of this organ. Also, this PhD Thesis provides the comparison between the milk microbiota of the Assaf and Churra sheep breeds. This contrast has shown that the Assaf milk microbiota is more diverse than that of the Churra breed, which may suggest that the breed factor has a potential role on the variation of the sheep milk microbiota.

Additionally, one of the studies included in this Thesis describes an analysis approach to explore the *resistome* of the sheep mammary gland based on a shotgun metagenomic

sequencing dataset generated for Assaf milk samples. The most prevalent antibiotic resistance genes (ARGs) identified for the analysed samples were resistant genes for tetracyclines, aminoglycoside and beta-lactam. All of these antibiotics are commonly used in sheep commercial flocks based on an intensive management system, such as the one routinely used for Spanish Assaf sheep. In addition, the analysis of RNA-Seq datasets from milk samples has been a first step to explore the *virome* of sheep milk. This analysis has identified the *Lentivirus* genus, which includes the microorganism causing maedi-visna disease, with a higher differential abundance in milk samples of the Assaf breed comparing with milk samples of the Churra breed. Finally, ONT technologies have been exploited to taxonomically and functionally characterise the rumen microbiota in sheep.

All the research works included in this PhD Thesis provide a global view of the usefulness of sequencing technologies to go in-depth concerning different issues related to animal genomics. The information resulting from these studies will settle the basis for future projects related to the *-omic* knowledge of traits of economic interest in sheep. Also, this information justifies the proposal of further specific studies considering the results presented here as a starting point to be confirmed in other sheep populations.

10. Bibliografía

Abecasis, G.R., A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, and G.A. McVean. 2012. An integrated map of genetic variation from 1,092 human genomes.. Nature 491:56–65. doi:10.1038/nature11632.

Acinas, S.G., R. Sarma-Rupavtarm, V. Klepac-Ceraj, and M.F. Polz. 2005. PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Appl. Environ. Microbiol. 71:8966 LP-8969. doi:10.1128/AEM.71.12.8966-8969.2005.

Addis, M.F., A. Tanca, S. Uzzau, G. Oikonomou, R.C. Bicalho, and P. Moroni. 2016. The bovine milk microbiota: insights and perspectives from -omics studies.. Mol. Biosyst. 12:2359–2372. doi:10.1039/c6mb00217j.

Ajay, S.S., S.C.J. Parker, H.O. Abaan, K.V.F. Fajardo, and E.H. Margulies. 2011. Accurate and comprehensive sequencing of personal genomes.. Genome Res. 21:1498–1505. doi:10.1101/gr.123638.111.

Akhtar, M.M., L. Micolucci, M.S. Islam, F. Olivieri, and A.D. Procopio. 2015. Bioinformatic tools for microRNA dissection. Nucleic Acids Res. 44:24–44. doi:10.1093/nar/gkv1221.

Alberto, F.J., F. Boyer, P. Orozco-terWengel, I. Streeter, B. Servin, P. de Villemereuil, B. Benjelloun, P. Librado, F. Biscarini, L. Colli, M. Barbato, W. Zamani, A. Alberti, S. Engelen, A. Stella, S. Joost, P. Ajmone-Marsan, R. Negrini, L. Orlando, H.R. Rezaei, S. Naderi, L. Clarke, P. Flicek, P. Wincker, E. Coissac, J. Kijas, G. Tosser-Klopp, A. Chikhi, M.W. Bruford, P. Taberlet, and F. Pompanon. 2018. Convergent genomic signatures of domestication in sheep and goats. Nat. Commun. 9:813. doi:10.1038/s41467-018-03206-y.

Altman, R.B. 1998. Bioinformatics in support of molecular medicine. Proceedings. AMIA Symp. 53–61.

Barendse, W., B.E. Harrison, R.J. Bunch, M.B. Thomas, and L.B. Turner. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC Genomics 10:178. doi:10.1186/1471-2164-10-178.

Barillet, F. 2007. Genetic improvement for dairy production in sheep and goats. Small Rumin. Res. 70:60–75. doi:https://doi.org/10.1016/j.smallrumres.2007.01.004.

Barillet, F., J.-J. Arranz, and A. Carta. 2005. Mapping quantitative trait loci for milk production and genetic polymorphisms of milk proteins in dairy sheep. Genet. Sel. Evol. 37:S109–S123.

Beato, M.S., M. Marcacci, E. Schiavon, L. Bertocchi, M. Di Domenico, A. Peserico, M. Mion, G. Zaccaria, L. Cavicchio, I. Mangone, E. Soranzo, C. Patavino, C. Cammà, and A. Lorusso. 2018. Identification and genetic characterization of bovine enterovirus by combination of two next generation sequencing platforms. J. Virol. Methods 260:21–25. doi:https://doi.org/10.1016/j.jviromet.2018.07.002.

Beaumont, M., J.K. Goodrich, M.A. Jackson, I. Yet, E.R. Davenport, S. Vieira-Silva, J. Debelius, T. Pallister, M. Mangino, and J. Raes. 2016. Heritable components of the

human fecal microbiome are associated with visceral fat. Genome Biol. 17:189.

Becker, D., J. Tetens, A. Brunner, D. Bürstel, M. Ganter, J. Kijas, for the I.S.G. Consortium, and C. Drögemüller. 2010. Microphthalmia in Texel Sheep Is Associated with a Missense Mutation in the Paired-Like Homeodomain 3 (PITX3) Gene. PLoS One 5:e8689.

Becker, K., C. Heilmann, and G. Peters. 2014. Coagulase-Negative Staphylococci. Clin. Microbiol. Rev. 27:870 LP-926. doi:10.1128/CMR.00109-13.

Bennewitz, J., N. Reinsch, F. Reinhardt, Z. Liu, and E. Kalm. 2004. Top down preselection using marker assisted estimates of breeding values in dairy cattle. J. Anim. Breed. Genet. 121:307–318. doi:10.1111/j.1439-0388.2004.00467.x.

Bergonier, D., and X. Berthelot. 2003. New advances in epizootiology and control of ewe mastitis. Livest. Prod. Sci. 79:1–16. doi:https://doi.org/10.1016/S0301-6226(02)00145-8.

Besser, J., H.A. Carleton, P. Gerner-Smidt, R.L. Lindsey, and E. Trees. 2018. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin. Microbiol. Infect. 24:335–341. doi:https://doi.org/10.1016/j.cmi.2017.10.013.

Bexiga, R., M.T. Koskinen, J. Holopainen, C. Carneiro, H. Pereira, K.A. Ellis, and C.L. Vilela. 2011. Diagnosis of intramammary infection in samples yielding negative results or minor pathogens in conventional bacterial culturing.. J. Dairy Res. 78:49–55. doi:10.1017/S0022029910000725.

Bhatt, V.D., V.B. Ahir, P.G. Koringa, S.J. Jakhesara, D.N. Rank, D.S. Nauriyal, A.P. Kunjadia, and C.G. Joshi. 2012. Milk microbiome signatures of subclinical mastitis-affected cattle analysed by shotgun sequencing.. J. Appl. Microbiol. 112:639–650. doi:10.1111/j.1365-2672.2012.05244.x.

Bishop, O.T. 2014. Bioinformatics and Data Analysis in Microbiology. Caister Academic Press.

Boichard, D., S. Fritz, M.-N. Rossignol, F. Guillaume, J.J. Colleau, and T. Druet. 2006. Implementation of marker-assisted selection: practical lessons from dairy cattle. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Commun 11–22.

Bolormaa, S., A.J. Chamberlain, M. Khansefid, P. Stothard, A.A. Swan, B. Mason, C.P. Prowse-Wilkins, N. Duijvesteijn, N. Moghaddar, J.H. van der Werf, H.D. Daetwyler, and I.M. MacLeod. 2019. Accuracy of imputation to whole-genome sequence in sheep. Genet. Sel. Evol. 51:1. doi:10.1186/s12711-018-0443-5.

Bonder, M.J., A. Kurilshikov, E.F. Tigchelaar, Z. Mujagic, F. Imhann, A.V. Vila, P. Deelen, T. Vatanen, M. Schirmer, S.P. Smeekens, D. V Zhernakova, S.A. Jankipersadsing, M. Jaeger, M. Oosting, M.C. Cenit, A.A.M. Masclee, M.A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R.K. Weersma, L. Franke, M.H. Hofker, R.J. Xavier, D. Jonkers, M.G. Netea, C. Wijmenga, J. Fu, and A. Zhernakova. 2016. The effect of host genetics on the gut microbiome. Nat. Genet. 48:1407–1412. doi:10.1038/ng.3663.

Børsting, C., and N. Morling. 2015. Next generation sequencing and its applications in forensic genetics. Forensic Sci. Int. Genet. 18:78–89. doi:https://doi.org/10.1016/j.fsigen.2015.02.002.

Boussaha, M., P. Michot, R. Letaief, C. Hozé, S. Fritz, C. Grohs, D. Esquerré, A. Duchesne, R. Philippe, V. Blanquet, F. Phocas, S. Floriot, D. Rocha, C. Klopp, A. Capitan, and D. Boichard. 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole-genome sequences. Genet. Sel. Evol. 48:87. doi:10.1186/s12711-016-0268-z.

Branton, D., D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S.B. Jovanovich, P.S. Krstic, S. Lindsay, X.S. Ling, C.H. Mastrangelo, A. Meller, J.S. Oliver, Y. V Pershin, J.M. Ramsey, R. Riehn, G. V Soni, V.T.- Cossa, M. Wanunu, M. Wiggin, and J.A. Schloss. 2009. The potential and challenges of nanopore sequencing. Co-Published with Macmillan Publishers Ltd, UK.

Callahan, B.J., P.J. McMurdie, and S.P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. Isme J. 11:2639.

Caporaso, J.G., C.L. Lauber, W.A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S.M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J.A. Gilbert, G. Smith, and R. Knight. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 6:1621–1624. doi:10.1038/ismej.2012.8.

Carta, A., S. Casu, and S. Salaris. 2009. Invited review: Current state of genetic improvement in dairy sheep. J. Dairy Sci. 92:5814–5833. doi:https://doi.org/10.3168/jds.2009-2479.

Castel, J.M., Y. Mena, F.A. Ruiz, J. Camúñez-Ruiz, and M. Sánchez-Rodríguez. 2011. Changes occurring in dairy goat production systems in less favoured areas of Spain. Small Rumin. Res. 96:83–92. doi:https://doi.org/10.1016/j.smallrumres.2011.01.002.

Catozzi, C., F. Ceciliani, C. Lecchi, A. Talenti, D. Vecchio, E. De Carlo, C. Grassi, A. Sánchez, O. Francino, and A. Cuscó. 2020. Short communication: Milk microbiota profiling on water buffalo with full-length 16S rRNA using nanopore sequencing. J. Dairy Sci.. doi:https://doi.org/10.3168/jds.2019-17359.

Catozzi, C., A. Sanchez Bonastre, O. Francino, C. Lecchi, E. De Carlo, D. Vecchio, A. Martucciello, P. Fraulo, V. Bronzo, A. Cusco, S. D'Andreano, and F. Ceciliani. 2017. The microbiota of water buffalo milk during mastitis.. PLoS One 12:e0184710. doi:10.1371/journal.pone.0184710.

Cesarani, A., G. Gaspa, F. Correddu, M. Cellesi, C. Dimauro, and N.P.P. Macciotta. 2019. Genomic selection of milk fatty acid composition in Sarda dairy sheep: Effect of different phenotypes and relationship matrices on heritability and breeding value accuracy. J. Dairy Sci. 102:3189–3203. doi:https://doi.org/10.3168/jds.2018-15333.

Chakraborty, C., C.G.P. Doss, B.C. Patra, and S. Bandyopadhyay. 2014. DNA barcoding

to map the microbial communities: current advances and future directions.. Appl. Microbiol. Biotechnol. 98:3425–3436. doi:10.1007/s00253-014-5550-9.

Chen, Q., J. Zhan, J. Wang, K. Qu, F. Zhang, J. Shen, P. Jia, Q. Ning, J. Zhang, N. Chen, H. Chen, B. Huang, and C. Lei. 2020. Whole-genome analyses identify loci and selective signals associated with body size in cattle. J. Anim. Sci.. doi:10.1093/jas/skaa068.

Chen, X., E. Jorgenson, and S.T. Cheung. 2009. New tools for functional genomic analysis. Drug Discov. Today 14:754–760. doi:https://doi.org/10.1016/j.drudis.2009.05.005.

Chiaradia, E., A. Valiani, M. Tartaglia, F. Scoppetta, G. Renzone, S. Arena, L. Avellini, S. Benda, A. Gaiti, and A. Scaloni. 2013. Ovine subclinical mastitis: Proteomic analysis of whey and milk fat globules unveils putative diagnostic biomarkers in milk. J. Proteomics 83:144–159. doi:https://doi.org/10.1016/j.jprot.2013.03.017.

Cho, I., and M.J. Blaser. 2012. The human microbiome: at the interface of health and disease.. Nat. Rev. Genet. 13:260–270. doi:10.1038/nrg3182.

Clarke, J., H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. 4:265–270. doi:10.1038/nnano.2009.12.

Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoir, B. Bibe, J. Bouix, F. Caiment, J.-M. Elsen, F. Eychenne, C. Larzul, E. Laville, F. Meish, D. Milenkovic, J. Tobin, C. Charlier, and M. Georges. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep.. Nat. Genet. 38:813–818. doi:10.1038/ng1810.

Colli, L., M. Milanesi, A. Talenti, F. Bertolini, M. Chen, A. Crisà, K.G. Daly, M. Del Corvo, B. Guldbrandtsen, J.A. Lenstra, B.D. Rosen, E. Vajana, G. Catillo, S. Joost, E.L. Nicolazzi, E. Rochat, M.F. Rothschild, B. Servin, T.S. Sonstegard, R. Steri, C.P. Van Tassell, P. Ajmone-Marsan, P. Crepaldi, A. Stella, and the A. Consortium. 2018. Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. Genet. Sel. Evol. 50:58. doi:10.1186/s12711-018-0422-x.

Cremonesi, P., C. Ceccarani, G. Curone, M. Severgnini, C. Pollera, V. Bronzo, F. Riva, M.F. Addis, J. Filipe, M. Amadori, E. Trevisi, D. Vigo, P. Moroni, and B. Castiglioni. 2018. Milk microbiome diversity and bacterial group prevalence in a comparison between healthy Holstein Friesian and Rendena cows. PLoS One 13:e0205054–e0205054. doi:10.1371/journal.pone.0205054.

Croville, G., G. Le Loc'h, C. Zanchetta, M. Manno, C. Camus-Bouclainville, C. Klopp, M. Delverdier, M.-N. Lucas, C. Donnadieu, M. Delpont, and J.-L. Guérin. 2018. Rapid whole-genome based typing and surveillance of avipoxviruses using nanopore sequencing. J. Virol. Methods 261:34–39. doi:https://doi.org/10.1016/j.jviromet.2018.08.003.

Dalrymple, B.P., E.F. Kirkness, M. Nefedov, S. McWilliam, A. Ratnakumar, W. Barris, S. Zhao, J. Shetty, J.F. Maddox, M. O'Grady, F. Nicholas, A.M. Crawford, T. Smith, P.J.

de Jong, J. McEwan, V.H. Oddy, N.E. Cockett, and the I.S.G. Consortium. 2007. Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. Genome Biol. 8:R152. doi:10.1186/gb-2007-8-7-r152.

Davies, G., S. Genini, S.C. Bishop, and E. Giuffra. 2009. An assessment of opportunities to dissect host genetic variation in resistance to infectious diseases in livestock.. Animal 3:415–436. doi:10.1017/S1751731108003522.

Delgado, B., A. Bach, I. Guasch, C. González, G. Elcoso, J.E. Pryce, and O. Gonzalez-Recio. 2019a. Whole rumen metagenome sequencing allows classifying and predicting feed efficiency and intake levels in cattle. Sci. Rep. 9:11. doi:10.1038/s41598-018-36673-w.

Delgado, B., M. Serrano, C. González, A. Bach, and O. Gonzalez-Recio. 2019b. Long reads from Nanopore sequencing as a tool for animal microbiome studies. Peer Rev.

Derakhshani, H., K.B. Fehr, S. Sepehri, D. Francoz, J. De Buck, H.W. Barkema, J.C. Plaizier, and E. Khafipour. 2018. Invited review: Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility.. J. Dairy Sci. 101:10605–10625. doi:10.3168/jds.2018-14860.

Devendra, C. 2001. Small Ruminants: Imperatives for Productivity Enhancement Improved Livelihoods and Rural Growth - A Review. Asian-Australas J Anim Sci 14:1483–1496. doi:10.5713/ajas.2001.1483.

Díaz, D., F.J. Esteban, P. Hernández, J.A. Caballero, G. Dorado, and S. Gálvez. 2011. Parallelizing and optimizing a bioinformatics pairwise sequence alignment algorithm for many-core architecture. Parallel Comput. 37:244–259. doi:https://doi.org/10.1016/j.parco.2011.03.003.

van Dijk, E.L., H. Auger, Y. Jaszczyszyn, and C. Thermes. 2014. Ten years of next-generation sequencing technology.. Trends Genet. 30:418–426. doi:10.1016/j.tig.2014.07.001.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. doi:10.1093/bioinformatics/bts635.

Dorado, G., S. Gálvez, H. Budak, T. Unver, and P. Hernández. 2019. Nucleic-Acid Sequencing. R.B.T.-E. of B.E. Narayan, ed. Elsevier, Oxford.

Duchemin, S.I., C. Colombani, A. Legarra, G. Baloche, H. Larroque, J.-M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. 2012. Genomic selection in the French Lacaune dairy sheep breed. J. Dairy Sci. 95:2723–2733. doi:https://doi.org/10.3168/jds.2011-4980.

Eberlein, A., A. Takasuga, K. Setoguchi, R. Pfuhl, K. Flisikowski, R. Fries, N. Klopp, R. Fürbass, R. Weikard, and C. Kühn. 2009. Dissection of genetic factors modulating fetal growth in cattle indicates a substantial role of the non-SMC condensin I complex, subunit G (NCAPG) gene.. Genetics 183:951–64. doi:10.1534/genetics.109.106476.

Eckburg, P.B., E.M. Bik, C.N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S.R. Gill, K.E. Nelson, and D.A. Relman. 2005. Diversity of the human intestinal microbial flora.. Science 308:1635–1638. doi:10.1126/science.1110591.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, and B. Bettman. 2009. Real-time DNA sequencing from single polymerase molecules. Science (80-.). 323:133–138.

van El, C.G., M.C. Cornel, P. Borry, R.J. Hastings, F. Fellmann, S. V Hodgson, H.C. Howard, A. Cambon-Thomsen, B.M. Knoppers, H. Meijers-Heijboer, H. Scheffer, L. Tranebjaerg, W. Dondorp, and G.M.W.R. de Wert. 2013. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics.. Eur. J. Hum. Genet. 21:580–584. doi:10.1038/ejhg.2013.46.

Elsik, C.G., R.L. Tellam, K.C. Worley, R.A. Gibbs, D.M. Muzny, G.M. Weinstock, D.L. Adelson, E.E. Eichler, L. Elnitski, R. Guigo, D.L. Hamernik, S.M. Kappes, H.A. Lewin, D.J. Lynn, F.W. Nicholas, A. Reymond, M. Rijnkels, L.C. Skow, E.M. Zdobnov, L. Schook, J. Womack, T. Alioto, S.E. Antonarakis, A. Astashyn, C.E. Chapple, H.-C. Chen, J. Chrast, F. Camara, O. Ermolaeva, C.N. Henrichsen, W. Hlavina, Y. Kapustin, B. Kiryutin, P. Kitts, F. Kokocinski, M. Landrum, D. Maglott, K. Pruitt, V. Sapojnikov, S.M. Searle, V. Solovyev, A. Souvorov, C. Ucla, C. Wyss, J.M. Anzola, D. Gerlach, E. Elhaik, D. Graur, J.T. Reese, R.C. Edgar, J.C. McEwan, G.M. Payne, J.M. Raison, T. Junier, E. V Kriventseva, E. Eyras, M. Plass, R. Donthu, D.M. Larkin, J. Reecy, M.Q. Yang, L. Chen, Z. Cheng, C.G. Chitko-McKown, G.E. Liu, L.K. Matukumalli, J. Song, B. Zhu, D.G. Bradley, F.S.L. Brinkman, L.P.L. Lau, M.D. Whiteside, A. Walker, T.T. Wheeler, T. Casey, J.B. German, D.G. Lemay, N.J. Maqbool, A.J. Molenaar, S. Seo, P. Stothard, C.L. Baldwin, R. Baxter, C.L. Brinkmeyer-Langford, W.C. Brown, C.P. Childers, T. Connelley, S.A. Ellis, K. Fritz, E.J. Glass, C.T.A. Herzig, A. Iivanainen, K.K. Lahmers, A.K. Bennett, C.M. Dickens, J.G.R. Gilbert, D.E. Hagen, H. Salih, J. Aerts, A.R. Caetano, B. Dalrymple, J.F. Garcia, C.A. Gill, S.G. Hiendleder, E. Memili, D. Spurlock, J.L. Williams, L. Alexander, M.J. Brownstein, L. Guan, R.A. Holt, S.J.M. Jones, M.A. Marra, R. Moore, S.S. Moore, A. Roberts, M. Taniguchi, R.C. Waterman, J. Chacko, M.M. Chandrabose, A. Cree, M.D. Dao, H.H. Dinh, R.A. Gabisi, S. Hines, J. Hume, S.N. Jhangiani, V. Joshi, C.L. Kovar, L.R. Lewis, Y.-S. Liu, J. Lopez, M.B. Morgan, N.B. Nguyen, G.O. Okwuonu, S.J. Ruiz, J. Santibanez, R.A. Wright, C. Buhay, Y. Ding, S. Dugan-Rocha, J. Herdandez, M. Holder, A. Sabo, A. Egan, J. Goodell, K. Wilczek-Boney, G.R. Fowler, M.E. Hitchens, R.J. Lozado, C. Moen, D. Steffen, J.T. Warren, J. Zhang, R. Chiu, J.E. Schein, K.J. Durbin, P. Havlak, H. Jiang, Y. Liu, X. Qin, Y. Ren, Y. Shen, H. Song, S.N. Bell, C. Davis, A.J. Johnson, S. Lee, L. V Nazareth, B.M. Patel, L.-L. Pu, S. Vattathil, R.L.J. Williams, S. Curry, C. Hamilton, E. Sodergren, D.A. Wheeler, W. Barris, G.L. Bennett, A. Eggen, R.D. Green, G.P. Harhay, M. Hobbs, O. Jann, J.W. Keele, M.P. Kent, S. Lien, S.D. McKay, S. McWilliam, A. Ratnakumar, R.D. Schnabel, T. Smith, W.M. Snelling, T.S. Sonstegard, R.T. Stone, Y. Sugimoto, A. Takasuga, J.F. Taylor, C.P. Van Tassell, M.D. Macneil, A.R.R. Abatepaulo, C.A. Abbey, V. Ahola, I.G. Almeida, A.F. Amadio, E. Anatriello, S.M. Bahadue, F.H. Biase, C.R. Boldt, J.A. Carroll, W.A. Carvalho, E.P. Cervelatti, E. Chacko, J.E. Chapin, Y. Cheng, J. Choi, A.J. Colley, T.A. de Campos, M. De Donato, I.K.F. de M. Santos, C.J.F. de Oliveira, H. Deobald, E. Devinoy, K.E. Donohue, P. Dovc, A. Eberlein, C.J. Fitzsimmons, A.M.

Franzin, G.R. Garcia, S. Genini, C.J. Gladney, J.R. Grant, M.L. Greaser, J.A. Green, D.L. Hadsell, H.A. Hakimov, R. Halgren, J.L. Harrow, E.A. Hart, N. Hastings, M. Hernandez, Z.-L. Hu, A. Ingham, T. Iso-Touru, C. Jamis, K. Jensen, D. Kapetis, T. Kerr, S.S. Khalil, H. Khatib, D. Kolbehdari, C.G. Kumar, D. Kumar, R. Leach, J.C.-M. Lee, C. Li, K.M. Logan, R. Malinverni, E. Marques, W.F. Martin, N.F. Martins, S.R. Maruyama, R. Mazza, K.L. McLean, J.F. Medrano, B.T. Moreno, D.D. More, C.T. Muntean, H.P. Nandakumar, M.F.G. Nogueira, I. Olsaker, S.D. Pant, F. Panzitta, R.C.P. Pastor, M.A. Poli, N. Poslusny, S. Rachagani, S. Ranganathan, A. Razpet, P.K. Riggs, G. Rincon, N. Rodriguez-Osorio, S.L. Rodriguez-Zas, N.E. Romero, A. Rosenwald, L. Sando, S.M. Schmutz, L. Shen, L. Sherman, B.R. Southey, Y.S. Lutzow, J. V Sweedler, I. Tammen, B.P.V.L. Telugu, J.M. Urbanski, Y.T. Utsunomiya, C.P. Verschoor, A.J. Waardenberg, Z. Wang, R. Ward, R. Weikard, T.H.J. Welsh, S.N. White, L.G. Wilming, K.R. Wunderlich, J. Yang, and F.-Q. Zhao. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution.. Science 324:522–528. doi:10.1126/science.1169588.

Endres, C.M., Í.M.S. de Castro, L.D. Trevisol, M.B. Mann, A.P.M. Varela, A.P.G. Frazzon, F.Q. Mayer, and J. frazzon. 2019. Molecular characterization of bacterial communities in sheep cheese through 16S rRNA gene sequencing. bioRxiv 753053. doi:10.1101/753053.

Falentin, H., L. Rault, A. Nicolas, D.S. Bouchard, J. Lassalas, P. Lamberton, J.-M. Aubry, P.-G. Marnet, Y. Le Loir, and S. Even. 2016. Bovine Teat Microbiome Analysis Revealed Reduced Alpha Diversity and Significant Changes in Taxonomic Profiles in Quarters with a History of Mastitis . Front. Microbiol. 7:480.

Fariello, M.-I., B. Servin, G. Tosser-Klopp, R. Rupp, C. Moreno, I.S.G. Consortium, M.S. Cristobal, and S. Boitard. 2014. Selection Signatures in Worldwide Sheep Populations. PLoS One 9:e103813.

Freitas, M.F.L., J.W. Pinheiro Júnior, T.L.M. Stamford, S.S. de A. Rabelo, D.R. da Silva, V.M. da Silveira Filho, F.G.B. Santos, M.J. de Sena, and R.A. Mota. 2005. Perfil de sensibilidade antimicrobiana in vitro de Staphylococcus coagulase positivos isolados de leite de vacas com mastite no agreste do estado de Pernambuco. Biológico, São Paulo 72:171–177.

Fricker, A.M., D. Podlesny, and W.F. Fricke. 2019. What is new and relevant for sequencing-based microbiome research? A mini-review. J. Adv. Res. 19:105–112. doi:10.1016/j.jare.2019.03.006.

Fthenakis, G.C., and J.E.T. Jones. 1990. The effect of experimentally induced subclinical mastitis on milk yield of ewes and on the growth of lambs. Br. Vet. J. 146:43–49. doi:https://doi.org/10.1016/0007-1935(90)90075-E.

Gaeta, N.C., S.F. Lima, A.G. Teixeira, E.K. Ganda, G. Oikonomou, L. Gregory, and R.C. Bicalho. 2017. Deciphering upper respiratory tract microbiota complexity in healthy calves and calves that develop respiratory disease using shotgun metagenomics. J. Dairy Sci. 100:1445–1458. doi:https://doi.org/10.3168/jds.2016-11522.

Ganda, E.K., R.S. Bisinotto, S.F. Lima, K. Kronauer, D.H. Decter, G. Oikonomou, Y.H.

Schukken, and R.C. Bicalho. 2016. Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. Sci. Rep. 6:37565. doi:10.1038/srep37565.

García-Gámez, E., B. Gutiérrez-Gil, G. Sahana, J.P. Sánchez, Y. Bayon, and J.J. Arranz. 2012a. GWA analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the LALBA gene. PLoS One 7:e47782. doi:10.1371/journal.pone.0047782 [doi].

García-Gámez, E., B. Gutierrez-Gil, J.P. Sanchez, and J.J. Arranz. 2012b. Replication and refinement of a quantitative trait locus influencing milk protein percentage on ovine chromosome 3. Anim. Genet. 43:636–641. doi:10.1111/j.1365-2052.2011.02294.x [doi].

García-Gámez, E., B. Gutiérrez-Gil, A. Suárez-Vega, L.F. de la Fuente, and J.J. Arranz. 2013. Identification of quantitative trait loci underlying milk traits in Spanish dairy sheep using linkage plus combined linkage disequilibrium and linkage analysis approaches. J. Dairy Sci. 96:6059–6069. doi:10.3168/jds.2013-6824 [doi].

Giannattasio-Ferraz, S., M. Laguardia-Nascimento, M.R. Gasparini, L.R. Leite, F.M.G. Araujo, A.C. de Matos Salim, A.P. de Oliveira, J.R. Nicoli, G.C. de Oliveira, F.G. da Fonseca, and E.F. Barbosa-Stancioli. 2019. A common vaginal microbiota composition among breeds of Bos taurus indicus (Gyr and Nellore). Brazilian J. Microbiol. 50:1115–1124. doi:10.1007/s42770-019-00120-3.

Glazier, A.M., J.H. Nadeau, and T.J. Aitman. 2002. Finding genes that underlie complex traits.. Science 298:2345–2349. doi:10.1126/science.1076641.

Goldstein, S., L. Beka, J. Graf, and J.L. Klassen. 2019. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing.. BMC Genomics 20:23. doi:10.1186/s12864-018-5381-7.

Goodwin, S., J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M.C. Schatz, and W.R. McCombie. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome.. Genome Res. 25:1750–1756. doi:10.1101/gr.191395.115.

Goodwin, S., J.D. McPherson, and W.R. McCombie. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17:333–351. doi:10.1038/nrg.2016.49.

Gootwine, E. 2020. Chapter 10 - Genetics and breeding of sheep and goats. F.W. Bazer, G.C. Lamb, and G.B.T.-A.A. Wu, ed. Academic Press.

Gougoulis, D.A., I. Kyriazakis, N. Papaioannou, E. Papadopoulos, I.A. Taitzoglou, and G.C. Fthenakis. 2008. Subclinical mastitis changes the patterns of maternal–offspring behaviour in dairy sheep. Vet. J. 176:378–384. doi:https://doi.org/10.1016/j.tvjl.2007.02.024.

Grobet, L., L.J. Martin, D. Poncelet, D. Pirottin, B. Brouwers, J. Riquet, A. Schoeberlein, S. Dunner, F. Menissier, J. Massabanda, R. Fries, R. Hanset, and M. Georges. 1997. A deletion in the bovine myostatin gene causes the double-muscled phenotype in

cattle.. Nat. Genet. 17:71–74. doi:10.1038/ng0997-71.

Guo, J., H. Tao, P. Li, L. Li, T. Zhong, L. Wang, J. Ma, X. Chen, T. Song, and H. Zhang. 2018. Whole-genome sequencing reveals selection signatures associated with important traits in six goat breeds. Sci. Rep. 8:10405. doi:10.1038/s41598-018-28719-w.

Gutiérrez-Gil, B., L. Alvarez, L.F. de la Fuente, J.P. Sanchez, F. San Primitivo, and J.J. Arranz. 2011. A genome scan for quantitative trait loci affecting body conformation traits in Spanish Churra dairy sheep. J. Dairy Sci. 94:4119–4128. doi:https://doi.org/10.3168/jds.2010-4027.

Gutiérrez-Gil, B., J.J. Arranz, R. Pong-Wong, E. García-Gámez, J. Kijas, and P. Wiener. 2014. Application of selection mapping to identify genomic regions associated with dairy production in sheep. PLoS One 9:e94623. doi:10.1371/journal.pone.0094623.

Gutierrez-Gil, B., J.J. Arranz, and P. Wiener. 2015. An interpretive review of selective sweep studies in Bos taurus cattle populations: Identification of unique and shared selection signals across breeds. Front. Genet. 6:167. doi:10.3389/fgene.2015.00167.

Gutierrez-Gil, B., M.F. El-Zarei, L. Alvarez, Y. Bayon, L.F. de la Fuente, F. San Primitivo, and J.J. Arranz. 2008. Quantitative trait loci underlying udder morphology traits in dairy sheep.. J. Dairy Sci. 91:3672–3681. doi:10.3168/jds.2008-1111.

Gutierrez-Gil, B., M.F. El-Zarei, L. Alvarez, Y. Bayon, L.F. de la Fuente, F. San Primitivo, and J.J. Arranz. 2009. Quantitative trait loci underlying milk production traits in sheep. Anim. Genet. 40:423–434. doi:10.1111/j.1365-2052.2009.01856.x [doi].

Gutiérrez-Gil, B., M.F. El-Zarei, Y. Bayón, L. Álvarez, L.F. de la Fuente, F.S. Primitivo, and J.J. Arranz. 2007. Short Communication: Detection of Quantitative Trait Loci Influencing Somatic Cell Score in Spanish Churra Sheep. J. Dairy Sci. 90:422–426. doi:10.3168/jds.S0022-0302(07)72643-7.

Hall, A.B., A.C. Tolonen, and R.J. Xavier. 2017. Human genetic variation and the gut microbiome in disease. Nat. Rev. Genet. 18:690–699. doi:10.1038/nrg.2017.63.

Han, G.G., J.-Y. Lee, G.-D. Jin, J. Park, Y.H. Choi, B.J. Chae, E.B. Kim, and Y.-J. Choi. 2017. Evaluating the association between body weight and the intestinal microbiota of weaned piglets via 16S rRNA sequencing. Appl. Microbiol. Biotechnol. 101:5903–5911. doi:10.1007/s00253-017-8304-7.

Hatem, A., D. Bozdag, A.E. Toland, and U. V Catalyurek. 2013. Benchmarking short sequence mapping tools.. BMC Bioinformatics 14:184. doi:10.1186/1471-2105-14-184.

Las Heras, A., L. Domínguez, and J.F. Fernández-Garayzábal. 1999. Prevalence and aetiology of subclinical mastitis in dairy ewes of the Madrid region. Small Rumin. Res. 32:21–29. doi:https://doi.org/10.1016/S0921-4488(98)00152-7.

Heravi, F.S., M. Zakrzewski, K. Vickery, and H. Hu. 2020. Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples. J. Microbiol.

Methods 170:105856. doi:https://doi.org/10.1016/j.mimet.2020.105856.

Hiergeist, A., J. Gläsner, U. Reischl, and A. Gessner. 2015. Analyses of intestinal microbiota: culture versus sequencing. ILAR J. 56:228–240.

Higgs, P.G., and T.K. Attwood. 2013. Bioinformatics and Molecular Evolution. John Wiley & Sons.

Hillier, L.W., W. Miller, E. Birney, W. Warren, R.C. Hardison, C.P. Ponting, P. Bork, D.W. Burt, M.A.M. Groenen, M.E. Delany, J.B. Dodgson, A.T. Chinwalla, P.F. Cliften, S.W. Clifton, K.D. Delehaunty, C. Fronick, R.S. Fulton, T.A. Graves, C. Kremitzki, D. Layman, V. Magrini, J.D. McPherson, T.L. Miner, P. Minx, W.E. Nash, M.N. Nhan, J.O. Nelson, L.G. Oddy, C.S. Pohl, J. Randall-Maher, S.M. Smith, J.W. Wallis, S.-P. Yang, M.N. Romanov, C.M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H.G. Tempest, L. Robertson, J.S. Masabanda, D.K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R.P.M. Crooijmans, J. Aerts, J.J. van der Poel, H. Ellegren, R.B. Caldwell, S.J. Hubbard, D. V Grafham, A.M. Kierzek, S.R. McLaren, I.M. Overton, H. Arakawa, K.J. Beattie, Y. Bezzubov, P.E. Boardman, J.K. Bonfield, M.D.R. Croning, R.M. Davies, M.D. Francis, S.J. Humphray, C.E. Scott, R.G. Taylor, C. Tickle, W.R.A. Brown, J. Rogers, J.-M. Buerstedde, S.A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M.M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G.K.-S. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P.J. deJong, L. Goodstadt, C. Webber, N.J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E.M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D.C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R.S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M.M. Hoffman, J. Severin, S.M.J. Searle, A.S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D.D. Shteynberg, M.R. Brent, J.M. Bye, E.J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A.G. Hatzigeorgiou, A.H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M.T. Webster, O. Pourquie, A. Reymond, C. Ucla, S.E. Antonarakis, M. Long, J.J. Emerson, E. Betrán, I. Dupanloup, H. Kaessmann, A.S. Hinrichs, G. Bejerano, T.S. Furey, R.A. Harte, B. Raney, A. Siepel, W.J. Kent, D. Haussler, E. Eyras, R. Castelo, J.F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P.A. Pevzner, A. Smit, L.A. Fulton, E.R. Mardis, R.K. Wilson, I.C.G.S. Consortium, O. coordination:, sequence and assembly: Genome fingerprint map, Mapping:, cDNA sequencing:, O. sequencing and libraries:, A. and annotation:, and P. management: 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716. doi:10.1038/nature03154.

Hoff, J.L., J.E. Decker, R.D. Schnabel, and J.F. Taylor. 2017. Candidate lethal haplotypes and causal mutations in Angus cattle. BMC Genomics 18:799. doi:10.1186/s12864-017-4196-2.

Holman, D.B., B.W. Brunelle, J. Trachsel, and H.K. Allen. 2017. Meta-analysis To Define a Core Microbiota in the Swine Gut. mSystems 2:e00004-17. doi:10.1128/mSystems.00004-17.

Holman, D.B., and K.E. Gzyl. 2019. A meta-analysis of the bovine gastrointestinal tract

microbiota. FEMS Microbiol. Ecol. 95. doi:10.1093/femsec/fiz072.

Hunt, K.M., J.A. Foster, L.J. Forney, U.M.E. Schütte, D.L. Beck, Z. Abdo, L.K. Fox, J.E. Williams, M.K. McGuire, and M.A. McGuire. 2011. Characterization of the Diversity and Temporal Stability of Bacterial Communities in Human Milk. PLoS One 6:e21313.

Hyams, E. 1972. Animals in the Service of Man. 10000 Years of Domestication.

Igartua, C., S. V Mozaffari, D.L. Nicolae, and C. Ober. 2017. Rare non-coding variants are associated with plasma lipid traits in a founder population. Sci. Rep. 7:16415. doi:10.1038/s41598-017-16550-8.

Imhann, F., A.V. Vila, M.J. Bonder, J. Fu, D. Gevers, M.C. Visschedijk, L.M. Spekhorst, R. Alberts, L. Franke, and H.M. van Dullemen. 2018. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut 67:108–119.

Ip, C.L.C., M. Loose, J.R. Tyson, M. de Cesare, B.L. Brown, M. Jain, R.M. Leggett, D.A. Eccles, V. Zalunin, J.M. Urban, P. Piazza, R.J. Bowden, B. Paten, S. Mwaigwisya, E.M. Batty, J.T. Simpson, T.P. Snutch, E. Birney, D. Buck, S. Goodwin, H.J. Jansen, J. O'Grady, H.E. Olsen, and M.A. and R. Consortium. 2015. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research 4:1075. doi:10.12688/f1000research.7201.1.

Jain, M., H.E. Olsen, B. Paten, and M. Akeson. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17:239. doi:10.1186/s13059-016-1103-0.

Jara, S., M. Sanchez, R. Vera, J. Cofre, and E. Castro. 2011. The inhibitory activity of Lactobacillus spp. isolated from breast milk on gastrointestinal pathogenic bacteria of nosocomial origin.. Anaerobe 17:474–477. doi:10.1016/j.anaerobe.2011.07.008.

Jeffreys, A.J., V. Wilson, and S.L. Thein. 1985. Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73. doi:10.1038/314067a0.

Ji, H.P. 2012. Improving bioinformatic pipelines for exome variant calling. Genome Med. 4:7. doi:10.1186/gm306.

Jiang, Y., M. Xie, W. Chen, R. Talbot, J.F. Maddox, T. Faraut, C. Wu, D.M. Muzny, Y. Li, W. Zhang, J.A. Stanton, R. Brauning, W.C. Barris, T. Hourlier, B.L. Aken, S.M. Searle, D.L. Adelson, C. Bian, G.R. Cam, Y. Chen, S. Cheng, U. DeSilva, K. Dixen, Y. Dong, G. Fan, I.R. Franklin, S. Fu, P. Fuentes-Utrilla, R. Guan, M.A. Highland, M.E. Holder, G. Huang, A.B. Ingham, S.N. Jhangiani, D. Kalra, C.L. Kovar, S.L. Lee, W. Liu, X. Liu, C. Lu, T. Lv, T. Mathew, S. McWilliam, M. Menzies, S. Pan, D. Robelin, B. Servin, D. Townley, W. Wang, B. Wei, S.N. White, X. Yang, C. Ye, Y. Yue, P. Zeng, Q. Zhou, J.B. Hansen, K. Kristiansen, R.A. Gibbs, P. Flicek, C.C. Warkup, H.E. Jones, V.H. Oddy, F.W. Nicholas, J.C. McEwan, J.W. Kijas, J. Wang, K.C. Worley, A.L. Archibald, N. Cockett, X. Xu, W. Wang, and B.P. Dalrymple. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344:1168–1173. doi:10.1126/science.1252806 [doi].

Jiao, J., J. Wu, C. Zhou, S. Tang, M. Wang, and Z. Tan. 2016. Composition of Ileal Bacterial Community in Grazing Goats Varies across Non-rumination, Transition and Rumination Stages of Life . Front. Microbiol. 7:1364.

Jimenez, E., J. de Andres, M. Manrique, P. Pareja-Tobes, R. Tobes, J.F. Martinez-Blanch, F.M. Codoner, D. Ramon, L. Fernandez, and J.M. Rodriguez. 2015. Metagenomic Analysis of Milk of Healthy and Mastitis-Suffering Women.. J. Hum. Lact. 31:406–415. doi:10.1177/0890334415585078.

Jovel, J., J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A.L. Mason, K.L. Madsen, and G.K.-S. Wong. 2016. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics.. Front. Microbiol. 7:459. doi:10.3389/fmicb.2016.00459.

Jung, H., C. Winefield, A. Bombarely, P. Prentis, and P. Waterhouse. 2019. Tools and strategies for long-read sequencing and de novo assembly of plant genomes. Trends Plant Sci.

Kafatos, F.C. 1998. Challenges for European biology.. Science 280:1327. doi:10.1126/science.280.5368.1327a.

Kamimura, B.A., L. Cabral, M.F. Noronha, R.C. Baptista, H.M. Nascimento, and A.S. Sant'Ana. 2020. Amplicon sequencing reveals the bacterial diversity in milk, dairy premises and Serra da Canastra artisanal cheeses produced by three different farms. Food Microbiol. 89:103453. doi:https://doi.org/10.1016/j.fm.2020.103453.

Kamke, J., S. Kittelmann, P. Soni, Y. Li, M. Tavendale, S. Ganesh, P.H. Janssen, W. Shi, J. Froula, E.M. Rubin, and G.T. Attwood. 2016. Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised by lactic acid formation and utilisation. Microbiome 4:56. doi:10.1186/s40168-016-0201-2.

Kaniyamattam, K., A. De Vries, L.W. Tauer, and Y.T. Gröhn. 2020. Economics of reducing antibiotic usage for clinical mastitis and metritis through genomic selection. J. Dairy Sci. 103:473–491. doi:https://doi.org/10.3168/jds.2018-15817.

Kaplan, N.L., R.R. Hudson, and C.H. Langley. 1989. The "hitchhiking effect" revisited.. Genetics 123:887–99.

Kijas, J.W., J.A. Lenstra, B. Hayes, S. Boitard, L.R. Porto Neto, M. San Cristobal, B. Servin, R. McCulloch, V. Whan, K. Gietzen, S. Paiva, W. Barendse, E. Ciani, H. Raadsma, J. McEwan, and B. Dalrymple. 2012a. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection.. PLoS Biol. 10:e1001258. doi:10.1371/journal.pbio.1001258.

Kijas, J.W., J.A. Lenstra, B. Hayes, S. Boitard, L.R. Porto Neto, M. San Cristobal, B. Servin, R. McCulloch, V. Whan, K. Gietzen, S. Paiva, W. Barendse, E. Ciani, H. Raadsma, J. McEwan, B. Dalrymple, and other members of the I.S.G. Consortium. 2012b. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLOS Biol. 10:e1001258.

Klaas, I.C., and R.N. Zadoks. 2017. An update on environmental mastitis: Challenging

perceptions. Transbound. Emerg. Dis. 65:166–185. doi:10.1111/tbed.12704.

Koufariotis, L., Y.-P.P. Chen, S. Bolormaa, and B.J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436. doi:10.1186/1471-2164-15-436.

Kuehn, J.S., P.J. Gorden, D. Munro, R. Rong, Q. Dong, P.J. Plummer, C. Wang, and G.J. Phillips. 2013. Bacterial Community Profiling of Milk Samples as a Means to Understand Culture-Negative Bovine Clinical Mastitis. PLoS One 8:e61959.

Li, C., M. Li, X. Li, W. Ni, Y. Xu, R. Yao, B. Wei, M. Zhang, H. Li, Y. Zhao, L. Liu, Y. Ullah, Y. Jiang, and S. Hu. 2019. Whole-Genome Resequencing Reveals Loci Associated With Thoracic Vertebrae Number in Sheep . Front. Genet. 10:674.

Li, X., J. Ye, X. Han, R. Qiao, X. Li, G. Lv, and K. Wang. 2020. Whole-genome sequencing identifies potential candidate genes for reproductive traits in pigs. Genomics 112:199–206. doi:https://doi.org/10.1016/j.ygeno.2019.01.014.

Liao, X., F. Peng, S. Forni, D. McLaren, G. Plastow, and P. Stothard. 2013. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection.. Genome 56:592–598. doi:10.1139/gen-2013-0082.

Lima, S.F., M.L. de S. Bicalho, and R.C. Bicalho. 2018. Evaluation of milk sample fractions for characterization of milk microbiota from healthy and clinical mastitis cows. PLoS One 13:e0193671–e0193671. doi:10.1371/journal.pone.0193671.

Logares, R., S. Sunagawa, G. Salazar, F.M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, J. Raes, J. Poulain, O. Jaillon, P. Wincker, S. Kandels-Lewis, E. Karsenti, P. Bork, and S.G. Acinas. 2014. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities.. Environ. Microbiol. 16:2659–2671. doi:10.1111/1462-2920.12250.

Lollai, S.A., M. Zicchedu, I. Duprè, and D. Piras. 2016. Characterization of resistance to tetracyclines and aminoglycosides of sheep mastitis pathogens: study of the effect of gene content on resistance. J. Appl. Microbiol. 121:941–951. doi:10.1111/jam.13229.

Luo, X., Y. Zhou, B. Zhang, Y. Zhang, X. Wang, T. Feng, Z. Li, K. Cui, Z. Wang, C. Luo, H. Li, Y. Deng, F. Lu, J. Han, Y. Miao, H. Mao, X. Yi, C. Ai, S. Wu, A. Li, Z. Wu, Z. Zhuo, D. Da Giang, B. Mitra, M.F. Vahidi, S. Mansoor, S.A. Al-Bayatti, E.M. Sari, N.A. Gorkhali, S. Prastowo, L. Shafique, G. Ye, Q. Qian, B. Chen, D. Shi, J. Ruan, and Q. Liu. 2020. Understanding divergent domestication traits from the whole-genome sequencing of swamp and river buffalo populations. Natl. Sci. Rev.. doi:10.1093/nsr/nwaa024.

Maddox, J.F., K.P. Davies, A.M. Crawford, D.J. Hulme, D. Vaiman, E.P. Cribiu, B.A. Freking, K.J. Beh, N.E. Cockett, and N. Kang. 2001. An enhanced linkage map of the sheep genome comprising more than 1000 loci. Genome Res. 11:1275–1289.

Marchet, C., L. Lecompte, C. Da Silva, C. Cruaud, J.-M. Aury, J. Nicolas, and P. Peterlongo. 2018. De novo clustering of long reads by gene from transcriptomics

data. Nucleic Acids Res. 47:e2–e2. doi:10.1093/nar/gky834.

Mardis, E.R. 2017. DNA sequencing technologies: 2006–2016. Nat. Protoc. 12:213–218. doi:10.1038/nprot.2016.182.

Marotz, C.A., J.G. Sanders, C. Zuniga, L.S. Zaramela, R. Knight, and K. Zengler. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome 6:42. doi:10.1186/s40168-018-0426-3.

Martínez-Bueno, M., and M.E. Alarcón-Riquelme. 2019. Exploring Impact of Rare Variation in Systemic Lupus Erythematosus by a Genome Wide Imputation Approach . Front. Immunol. 10:258.

Martínez, S., I. Franco, and J. Carballo. 2011. Spanish goat and sheep milk cheeses. Small Rumin. Res. 101:41–54. doi:https://doi.org/10.1016/j.smallrumres.2011.09.024.

Mateescu, R.G. 2020. Chapter 2 - Genetics and breeding of beef cattle. F.W. Bazer, G.C. Lamb, and G.B.T.-A.A. Wu, ed. Academic Press.

Mathema, B., J.R. Mediavilla, L. Chen, and B.N. Kreiswirth. 2009. Evolution and taxonomy of staphylococci. Staphylococci Hum. Dis. 31–64.

McInnis, E.A., K.M. Kalanetra, D.A. Mills, and E.A. Maga. 2015. Analysis of raw goat milk microbiota: impact of stage of lactation and lysozyme on microbial diversity.. Food Microbiol. 46:121–131. doi:10.1016/j.fm.2014.07.021.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. doi:10.1101/gr.107524.110 [doi].

Meera Krishna, B., M.A. Khan, and S.T. Khan. 2019. Next-Generation Sequencing (NGS) Platforms: An Exciting Era of Genome Sequence Analysis BT - Microbial Genomics in Sustainable Agroecosystems: Volume 2. V. Tripathi, P. Kumar, P. Tripathi, A. Kishore, and M. Kamle, ed. Springer Singapore, Singapore.

Megdiche, S., S. Mastrangelo, M. Ben Hamouda, J.A. Lenstra, and E. Ciani. 2019. A Combined Multi-Cohort Approach Reveals Novel and Known Genome-Wide Selection Signatures for Wool Traits in Merino and Merino-Derived Sheep Breeds . Front. Genet. 10:1025.

Metzger, J., R. Schrimpf, U. Philipp, and O. Distl. 2013. Expression levels of LCORL are associated with body size in horses. PLoS One 8:e56497–e56497. doi:10.1371/journal.pone.0056497.

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

Miltiadou, D., A.L. Hager-Theodorides, S. Symeou, C. Constantinou, A. Psifidi, G. Banos, and O. Tzamaloukas. 2017. Variants in the 3’ untranslated region of the ovine acetyl-coenzyme A acyltransferase 2 gene are associated with dairy traits and exhibit differential allelic expression.. J. Dairy Sci. 100:6285–6297.

doi:10.3168/jds.2016-12326.

Minervini, C.F., C. Cumbo, P. Orsini, L. Anelli, A. Zagaria, L. Impera, N. Coccaro, C. Brunetti, A. Minervini, P. Casieri, G. Tota, A. Russo Rossi, G. Specchia, and F. Albano. 2017. Mutational analysis in BCR-ABL1 positive leukemia by deep sequencing based on nanopore MinION technology. Exp. Mol. Pathol. 103:33–37. doi:https://doi.org/10.1016/j.yexmp.2017.06.007.

Minervini, C.F., C. Cumbo, P. Orsini, C. Brunetti, L. Anelli, A. Zagaria, A. Minervini, P. Casieri, N. Coccaro, G. Tota, L. Impera, A. Giordano, G. Specchia, and F. Albano. 2016. TP53 gene mutation analysis in chronic lymphocytic leukemia by nanopore MinION sequencing. Diagn. Pathol. 11:96. doi:10.1186/s13000-016-0550-y.

Moioli, B., M. D'Andrea, and F. Pilla. 2007. Candidate genes affecting sheep and goat milk quality. Small Rumin. Res. 68:179–192. doi:https://doi.org/10.1016/j.smallrumres.2006.09.008.

Moon, S., T.-H. Kim, K.-T. Lee, W. Kwak, T. Lee, S.-W. Lee, M.-J. Kim, K. Cho, N. Kim, W.-H. Chung, S. Sung, T. Park, S. Cho, M.A. Groenen, R. Nielsen, Y. Kim, and H. Kim. 2015. A genome-wide scan for signatures of directional selection in domesticated pigs.. BMC Genomics 16:130. doi:10.1186/s12864-015-1330-x.

Mulsant, P., F. Lecerf, S. Fabre, L. Schibler, P. Monget, I. Lanneluc, C. Pisselet, J. Riquet, D. Monniaux, I. Callebaut, E. Cribiu, J. Thimonier, J. Teyssier, L. Bodin, Y. Cognie, N. Chitour, and J.M. Elsen. 2001. Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes.. Proc. Natl. Acad. Sci. U. S. A. 98:5104–5109. doi:10.1073/pnas.091577598.

Murdoch, B.M. 2019. The functional annotation of the sheep genome project. J. Anim. Sci. 97:16. doi:10.1093/jas/skz122.029.

Naderi, S., M.H. Moradi, M. Farhadian, T. Yin, M. Jaeger, C. Scheper, P. Korkuc, G.A. Brockmann, S. König, and K. May. 2020. Assessing selection signatures within and between selected lines of dual-purpose black and white and German Holstein cattle. Anim. Genet. n/a. doi:10.1111/age.12925.

Nicholas, F.W., and M. Hobbs. 2014. Mutation discovery for Mendelian traits in non-laboratory animals: a review of achievements up to 2012.. Anim. Genet. 45:157–170. doi:10.1111/age.12103.

Nietfeld, F., D. Höltig, H. Willems, P. Valentin-Weigand, C. Wurmser, K.-H. Waldmann, R. Fries, and G. Reiner. 2020. Candidate genes and gene markers for the resistance to porcine pleuropneumonia. Mamm. Genome 31:54–67. doi:10.1007/s00335-019-09825-0.

Norman, H.D., R.H. Miller, J.R. Wright, and G.R. Wiggans. 2000. Herd and State Means for Somatic Cell Count from Dairy Herd Improvement. J. Dairy Sci. 83:2782–2788. doi:10.3168/jds.S0022-0302(00)75175-7.

OCDE. 2018. Stemming the Superbug Tide: Just a Few Dollars More. OECD publishing.

Oget, C., C. Allain, D. Portes, G. Foucras, A. Stella, J.-M. Astruc, J. Sarry, G. Tosser-Klopp, and R. Rupp. 2019. A validation study of loci associated with mastitis resistance in

two French dairy sheep breeds. Genet. Sel. Evol. 51:5. doi:10.1186/s12711-019-0448-8.

Oikonomou, G., M.F. Addis, C. Chassard, M.E.F. Nader-Macias, I. Grant, C. Delbès, C.I. Bogni, Y. Le Loir, and S. Even. 2020. Milk Microbiota: What Are We Exactly Talking About? . Front. Microbiol. 11:60.

Oikonomou, G., M.L. Bicalho, E. Meira, R.E. Rossi, C. Foditsch, V.S. Machado, A.G.V. Teixeira, C. Santisteban, Y.H. Schukken, and R.C. Bicalho. 2014. Microbiota of Cow's Milk; Distinguishing Healthy, Sub-Clinically and Clinically Diseased Quarters. PLoS One 9:e85904.

Oikonomou, G., V.S. Machado, C. Santisteban, Y.H. Schukken, and R.C. Bicalho. 2012. Microbial Diversity of Bovine Mastitic Milk as Described by Pyrosequencing of Metagenomic 16s rDNA. PLoS One 7:e47671.

Oliveira Júnior, G.A., D.J.A. Santos, A.S.M. Cesar, S.A. Boison, R. V Ventura, B.C. Perez, J.F. Garcia, J.B.S. Ferraz, and D.J. Garrick. 2019. Fine mapping of genomic regions associated with female fertility in Nellore beef cattle based on sequence variants from segregating sires. J. Anim. Sci. Biotechnol. 10:97. doi:10.1186/s40104-019-0403-0.

Oultram, J.W.H., E.K. Ganda, S.C. Boulding, R.C. Bicalho, and G. Oikonomou. 2017. A Metataxonomic Approach Could Be Considered for Cattle Clinical Mastitis Diagnostics.. Front. Vet. Sci. 4:36. doi:10.3389/fvets.2017.00036.

Ouzounis, C.A., and A. Valencia. 2003. Early bioinformatics: the birth of a discipline—a personal view. Bioinformatics 19:2176–2190. doi:10.1093/bioinformatics/btg309.

Pang, M., X. Xie, H. Bao, L. Sun, T. He, H. Zhao, Y. Zhou, L. Zhang, H. Zhang, R. Wei, K. Xie, and R. Wang. 2018. Insights Into the Bovine Milk Microbiota in Dairy Farms With Different Incidence Rates of Subclinical Mastitis . Front. Microbiol. 9:2379.

Pankey, J.W., S.C. Nickerson, R.L. Boddie, and J.S. Hogan. 1985. Effects of Corynebacterium bovis Infection on Susceptibility to Major Mastitis Pathogens. J. Dairy Sci. 68:2684–2693. doi:https://doi.org/10.3168/jds.S0022-0302(85)81153-X.

Peñagaricano, F. 2020. Chapter 6 - Genetics and genomics of dairy cattle. F.W. Bazer, G.C. Lamb, and G.B.T.-A.A. Wu, ed. Academic Press.

Pettersson, E., J. Lundeberg, and A. Ahmadian. 2009. Generations of sequencing technologies.. Genomics 93:105–111. doi:10.1016/j.ygeno.2008.10.003.

Pfeifer, S.P. 2017. From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb). 118:111–124. doi:10.1038/hdy.2016.102.

Pirisi, A., G. Piredda, C.M. Papoff, R. Di Salvo, S. Pintus, G. Garro, P. Ferranti, and L. Chianese. 1999. Effects of sheep αs1-casein CC, CD and DD genotypes on milk composition and cheesemaking properties. J. Dairy Res. 66:409–419. doi:DOI: 10.1017/S0022029999003635.

Pollinger, J.P., C.D. Bustamante, A. Fledel-Alon, S. Schmutz, M.M. Gray, and R.K. Wayne. 2005. Selective sweep mapping of genes with large phenotypic effects..

Genome Res. 15:1809–1819. doi:10.1101/gr.4374505.

Proctor, L.M. 2011. The Human Microbiome Project in 2011 and beyond.. Cell Host Microbe 10:287–291. doi:10.1016/j.chom.2011.10.001.

Project, T.H.M., and Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207.

Qin, M., C. Li, Z. Li, W. Chen, and Y. Zeng. 2020. Genetic Diversities and Differentially Selected Regions Between Shandong Indigenous Pig Breeds and Western Pig Breeds . Front. Genet. 10:1351.

Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.. BMC Genomics 13:341. doi:10.1186/1471-2164-13-341.

Quick, J., N.J. Loman, S. Duraffour, J.T. Simpson, E. Severi, L. Cowley, J.A. Bore, R. Koundouno, G. Dudas, and A. Mikhail. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232.

Quince, C., A.W. Walker, J.T. Simpson, N.J. Loman, and N. Segata. 2017. Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35:833–844. doi:10.1038/nbt.3935.

Rashkin, S., G. Jun, S. Chen, and G.R. Abecasis. 2017. Optimal sequencing strategies for identifying disease-associated singletons.. PLoS Genet. 13:e1006811. doi:10.1371/journal.pgen.1006811.

Raynal-Ljutovac, K., A. Pirisi, R. de Crémoux, and C. Gonzalo. 2007. Somatic cells of goat and sheep milk: Analytical, sanitary, productive and technological aspects. Small Rumin. Res. 68:126–144. doi:https://doi.org/10.1016/j.smallrumres.2006.09.012.

Robbins, R.J. 1996. Bioinformatics: Essential Infrastructure for Global Biology1. J. Comput. Biol. 3:465–478.

Rochus, C.M., F. Tortereau, F. Plisson-Petit, G. Restoux, C. Moreno-Romieux, G. Tosser-Klopp, and B. Servin. 2018. Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. BMC Genomics 19:71. doi:10.1186/s12864-018-4447-x.

Ron, M., and J.I. Weller. 2007. From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. Anim. Genet. 38:429–439. doi:10.1111/j.1365-2052.2007.01640.x.

Ross, E.M., P.J. Moate, L.C. Marett, B.G. Cocks, and B.J. Hayes. 2013. Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. PLoS One 8.

Rubin, C.-J., H.-J. Megens, A. Martinez Barrio, K. Maqbool, S. Sayyab, D. Schwochow, C. Wang, Ö. Carlborg, P. Jern, C.B. Jørgensen, A.L. Archibald, M. Fredholm, M.A.M. Groenen, and L. Andersson. 2012. Strong signatures of selection in the domestic

pig genome.. Proc. Natl. Acad. Sci. U. S. A. 109:19529–36. doi:10.1073/pnas.1217149109.

Rubin, C.-J., M.C. Zody, J. Eriksson, J.R.S. Meadows, E. Sherwood, M.T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication.. Nature 464:587–91. doi:10.1038/nature08832.

Rupp, R., and D. Boichard. 2003. Genetics of resistance to mastitis in dairy cattle. Vet. Res. 34:671–688.

Rupp, R., P. Senin, J. Sarry, C. Allain, C. Tasca, L. Ligat, D. Portes, F. Woloszyn, O. Bouchez, G. Tabouret, M. Lebastard, C. Caubet, G. Foucras, and G. Tosser-Klopp. 2015a. A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model. PLoS Genet. 11:e1005629. doi:10.1371/journal.pgen.1005629 [doi].

Rupp, R., P. Senin, J. Sarry, C. Allain, C. Tasca, L. Ligat, D. Portes, F. Woloszyn, O. Bouchez, G. Tabouret, M. Lebastard, C. Caubet, G. Foucras, and G. Tosser-Klopp. 2015b. A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model.. PLoS Genet. 11:e1005629. doi:10.1371/journal.pgen.1005629.

Sahana, G., J.K. Höglund, B. Guldbrandtsen, and M.S. Lund. 2015. Loci associated with adult stature also affect calf birth survival in cattle. BMC Genet. 16:47. doi:10.1186/s12863-015-0202-3.

Saif, R., J. Henkel, V. Jagannathan, C. Drögemüller, C. Flury, and T. Leeb. 2020. The LCORL Locus is under Selection in Large-Sized Pakistani Goat Breeds. Genes (Basel). 11:168.

Salmela, L., R. Walve, E. Rivals, and E. Ukkonen. 2016. Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics 33:799–806. doi:10.1093/bioinformatics/btw321.

San Primitivo, F., and L.F. De la Fuente. 2000. Situación actual de la oveja de raza Churra. Arch. Zootec. 49:161–165.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin. 2008. Estimating Local Ancestry in Admixed Populations. Am. J. Hum. Genet. 82:290–303. doi:https://doi.org/10.1016/j.ajhg.2007.09.022.

Saratsis, P., C. Alexopoulos, A. Tzora, and G.C. Fthenakis. 1999. The effect of experimentally induced subclinical mastitis on the milk yield of dairy ewes. Small Rumin. Res. 32:205–209. doi:https://doi.org/10.1016/S0921-4488(98)00189-8.

Sasson, G., S.K. Ben-Shabat, E. Seroussi, A. Doron-Faigenboim, N. Shterzer, S. Yaacoby, M.E.B. Miller, B.A. White, E. Halperin, and I. Mizrahi. 2017. Heritable bovine rumen bacteria are phylogenetically related and correlated with the cow's

capacity to harvest energy from its feed. MBio 8:e00703-17.

Schadt, E.E., S. Turner, and A. Kasarskis. 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19:R227–R240. doi:10.1093/hmg/ddq416.

Schloss, J.A. 2008. How to get genomes at one ten-thousandth the cost.. Nat. Biotechnol. 26:1113–1115. doi:10.1038/nbt1008-1113.

Schloss, P.D., D. Gevers, and S.L. Westcott. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.. PLoS One 6:e27310. doi:10.1371/journal.pone.0027310.

Schlotterer, C., R. Tobler, R. Kofler, and V. Nolte. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding.. Nat. Rev. Genet. 15:749–763. doi:10.1038/nrg3803.

Schukken, Y.H., R.N. González, L.L. Tikofsky, H.F. Schulte, C.G. Santisteban, F.L. Welcome, G.J. Bennett, M.J. Zurakowski, and R.N. Zadoks. 2009. CNS mastitis: Nothing to worry about?. Vet. Microbiol. 134:9–14. doi:https://doi.org/10.1016/j.vetmic.2008.09.014.

Shook, G.E., and M.M. Schutz. 1994. Selection on somatic cell score to improve resistance to mastitis in the United States.. J. Dairy Sci. 77:648–658. doi:10.3168/jds.S0022-0302(94)76995-2.

Singh, B., G. Mal, S.K. Gautam, and M. Mukesh. 2019. Gut/Rumen Microbiome—A Livestock and Industrial Perspective BT - Advances in Animal Biotechnology. B. Singh, G. Mal, S.K. Gautam, and M. Mukesh, ed. Springer International Publishing, Cham.

Slatko, B.E., A.F. Gardner, and F.M. Ausubel. 2018. Overview of Next-Generation Sequencing Technologies. Curr. Protoc. Mol. Biol. 122:e59. doi:10.1002/cpmb.59.

Smith, J.M., and J. Haigh. 2007. The hitch-hiking effect of a favourable gene.. Genet. Res. 89:391–403. doi:10.1017/S0016672308009579.

Sohn, J., and J.-W. Nam. 2016. The present and future of de novo whole-genome assembly. Brief. Bioinform. 19:23–40. doi:10.1093/bib/bbw096.

Sommerhäuser, J., B. Kloppert, W. Wolter, M. Zschöck, A. Sobiraj, and K. Failing. 2003. The epidemiology of Staphylococcus aureus infections from subclinical mastitis in dairy cows during a control programme. Vet. Microbiol. 96:91–102. doi:https://doi.org/10.1016/S0378-1135(03)00204-9.

Strandén, I., J. Kantanen, I.-R.M. Russo, P. Orozco-terWengel, M.W. Bruford, and the C. Consortium. 2019. Genomic selection strategies for breeding adaptation and production in dairy cattle under climate change. Heredity (Edinb). 123:307–317. doi:10.1038/s41437-019-0207-1.

Streit, W.R., and R.A. Schmitz. 2004. Metagenomics--the key to the uncultured microbes.. Curr. Opin. Microbiol. 7:492–498. doi:10.1016/j.mib.2004.08.002.

Suárez-Vega, A., B. Gutiérrez-Gil, I. Cuchillo-Ibáñez, J. Sáez-Valero, V. Pérez, E. García-Gámez, J. Benavides, and J.J. Arranz. 2013. Identification of a 31-bp Deletion in

the RELN Gene Causing Lissencephaly with Cerebellar Hypoplasia in Sheep. PLoS One 8:e81072.

Suarez-Vega, A., B. Gutierrez-Gil, C. Klopp, G. Tosser-Klopp, and J.J. Arranz. 2016. Comprehensive RNA-Seq profiling to evaluate lactating sheep mammary gland transcriptome. Sci. Data. doi:10.1038/sdata.2016.51.

Sweeney, T.E., and J.M. Morton. 2013. The Human Gut Microbiome: A Review of the Effect of Obesity and Surgically Induced Weight Loss. JAMA Surg. 148:563–569. doi:10.1001/jamasurg.2013.5.

Talenti, A., F. Bertolini, G. Pagnacco, F. Pilla, P. Ajmone-Marsan, M.F. Rothschild, P. Crepaldi, and T.I.G. Consortium. 2017. The Valdostana goat: a genome-wide investigation of the distinctiveness of its selective sweep regions. Mamm. Genome 28:114–128. doi:10.1007/s00335-017-9678-7.

Tapio, I., T.J. Snelling, F. Strozzi, and R.J. Wallace. 2017. The ruminal microbiome associated with methane emissions from ruminant livestock. J. Anim. Sci. Biotechnol. 8:7. doi:10.1186/s40104-017-0141-0.

Taponen, S., D. McGuinness, H. Hiitiö, H. Simojoki, R. Zadoks, and S. Pyörälä. 2019. Bovine milk microbiome: a more complex issue than expected. Vet. Res. 50:44. doi:10.1186/s13567-019-0662-y.

Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E. V Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A. V Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41. doi:10.1186/1471-2105-4-41.

Theuns, S., B. Vanmechelen, Q. Bernaert, W. Deboutte, M. Vandenhole, L. Beller, J. Matthijnssens, P. Maes, and H.J. Nauwynck. 2018. Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. Sci. Rep. 8:9830. doi:10.1038/s41598-018-28180-9.

Tremblay, J., K. Singh, A. Fern, E.S. Kirton, S. He, T. Woyke, J. Lee, F. Chen, J.L. Dangl, and S.G. Tringe. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front. Microbiol. 6:771.

Tyson, J.R., N.J. O'Neil, M. Jain, H.E. Olsen, P. Hieter, and T.P. Snutch. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. Genome Res. 28:266–274.

Ugarte, E., M. Serrano, L.F. De la Fuente, M.D. Pérez-Guzmán, L. Alfonso, and J.P. Gutiérrez. 2002. Situación actual de los programas de mejora genética en ovino de leche. ITEA 98:102–117.

Usai, M.G., S. Casu, T. Sechi, S.L. Salaris, S. Miari, S. Sechi, P. Carta, and A. Carta. 2019. Mapping genomic regions affecting milk traits in Sarda sheep by using the OvineSNP50 Beadchip and principal components to perform combined linkage and linkage disequilibrium analysis. Genet. Sel. Evol. 51:65. doi:10.1186/s12711-

019-0508-0.

Vasta, V., M. Daghio, A. Cappucci, A. Buccioni, A. Serra, C. Viti, and M. Mele. 2019. Invited review: Plant polyphenols and rumen microbiota responsible for fatty acid biohydrogenation, fiber digestion, and methane emission: Experimental evidence and methodological approaches. J. Dairy Sci. 102:3781–3804. doi:https://doi.org/10.3168/jds.2018-14985.

Vega-Rodríguez, M.A., and S. Santander-Jiménez. 2019. Parallel computing in bioinformatics: a view from high-performance, heterogeneous, and cloud computing. J. Supercomput. 75:3369–3373. doi:10.1007/s11227-019-02934-2.

Verdura, E., A. Schlüter, G. Fernández-Eulate, R. Ramos-Martín, M. Zulaica, L. Planas-Serra, M. Ruiz, S. Fourcade, C. Casasnovas, A. López de Munain, and A. Pujol. 2020. A deep intronic splice variant advises reexamination of presumably dominant SPG7 Cases. Ann. Clin. Transl. Neurol. 7:105–111. doi:10.1002/acn3.50967.

Vinje, H., T. Almøy, K.H. Liland, and L. Snipen. 2014. A systematic search for discriminating sites in the 16S ribosomal RNA gene. Microb. Inform. Exp. 4:2. doi:10.1186/2042-5783-4-2.

Votintseva, A.A., P. Bradley, L. Pankhurst, C. del Ojo Elias, M. Loose, K. Nilgiriwala, A. Chatterjee, E.G. Smith, N. Sanderson, and T.M. Walker. 2017. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J. Clin. Microbiol. 55:1285–1298.

Walsh, A.M., F. Crispie, O. O'Sullivan, L. Finnegan, M.J. Claesson, and P.D. Cotter. 2018. Species classifier choice is a key consideration when analysing low-complexity food microbiome data. Microbiome 6:50.

Wang, Y., F. Zhang, R. Mukiibi, L. Chen, M. Vinsky, G. Plastow, J. Basarab, P. Stothard, and C. Li. 2020. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: carcass merit traits. BMC Genomics 21:38. doi:10.1186/s12864-019-6273-1.

Watts, J.L., D.E. Lowery, J.F. Teel, C. Ditto, J.-S. Horng, and S. Rossbach. 2001. Phylogenetic Studies on Corynebacterium bovis Isolated from Bovine Mammary Glands. J. Dairy Sci. 84:2419–2423. doi:https://doi.org/10.3168/jds.S0022-0302(01)74691-7.

Waugh, M., K. Briggs, D. Gunn, M. Gibeault, S. King, Q. Ingram, A.M. Jimenez, S. Berryman, D. Lomovtsev, L. Andrzejewski, and V. Tabard-Cossa. 2020. Solid-state nanopore fabrication by automated controlled breakdown. Nat. Protoc. 15:122–143. doi:10.1038/s41596-019-0255-2.

Weldenegodguad, M., R. Popov, K. Pokharel, I. Ammosov, Y. Ming, Z. Ivanova, and J. Kantanen. 2019. Whole-Genome Sequencing of Three Native Cattle Breeds Originating From the Northernmost Cattle Farming Regions . Front. Genet. 9:728.

WHO. 2014. Antimicrobial Resistance: Global Report on Surveillance. World Health

Organization.

Wick, R.R., L.M. Judd, and K.E. Holt. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20:129. doi:10.1186/s13059-019-1727-y.

Willett, W.C., F. Sacks, A. Trichopoulou, G. Drescher, A. Ferro-Luzzi, E. Helsing, and D. Trichopoulos. 1995. Mediterranean diet pyramid: a cultural model for healthy eating. Am. J. Clin. Nutr. 61:1402S–1406S. doi:10.1093/ajcn/61.6.1402S.

Wilson, R., K. Østbye, I.L. Angell, and K. Rudi. 2019. Association between diet and rumen microbiota in wild roe deer. FEMS Microbiol. Lett. 366. doi:10.1093/femsle/fnz060.

Wu, Y., J.S. Pi, A.L. Pan, Y.J. Pu, J.P. Du, J. Shen, Z.H. Liang, and J.R. Zhang. 2012. An SNP in the MyoD1 Gene Intron 2 Associated with Growth and Carcass Traits in Three Duck Populations. Biochem. Genet. 50:898–907. doi:10.1007/s10528-012-9530-4.

Xia, Y., A.-D. Li, Y. Deng, X.-T. Jiang, L.-G. Li, and T. Zhang. 2017. MinION nanopore sequencing enables correlation between resistome phenotype and genotype of coliform bacteria in municipal sewage. Front. Microbiol. 8:2105.

Yan, G., R. Qiao, F. Zhang, W. Xin, S. Xiao, T. Huang, Z. Zhang, and L. Huang. 2017. Imputation-Based Whole-Genome Sequence Association Study Rediscovered the Missing QTL for Lumbar Number in Sutai Pigs. Sci. Rep. 7:615. doi:10.1038/s41598-017-00729-0.

Yang, J.-J., T.-W. Chang, Y. Jiang, H.-J. Kao, B.-H. Chiou, M.-S. Kao, and C.-M. Huang. 2018. Commensal Staphylococcus aureus Provokes Immunity to Protect against Skin Infection of Methicillin-Resistant Staphylococcus aureus. Int. J. Mol. Sci. 19:1290. doi:10.3390/ijms19051290.

Yang, X., N.R. Noyes, E. Doster, J.N. Martin, L.M. Linke, R.J. Magnuson, H. Yang, I. Geornaras, D.R. Woerner, and K.L. Jones. 2016. Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain. Appl. Environ. Microbiol. 82:2433–2443.

Yarza, P., P. Yilmaz, E. Pruesse, F.O. Glockner, W. Ludwig, K.-H. Schleifer, W.B. Whitman, J. Euzeby, R. Amann, and R. Rossello-Mora. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences.. Nat. Rev. Microbiol. 12:635–645. doi:10.1038/nrmicro3330.

Yip, C.C.-Y., W.-M. Chan, J.D. Ip, C.W.-M. Seng, K.-H. Leung, R.W.-S. Poon, A.C.-K. Ng, W.-L. Wu, H. Zhao, K.-H. Chan, G.K.-H. Siu, T.T.-L. Ng, V.C.-C. Cheng, K.-H. Kok, K.-Y. Yuen, and K.K.-W. To. 2020. Nanopore sequencing reveals novel targets for the diagnosis and surveillance of human and avian influenza A virus. J. Clin. Microbiol. JCM.02127-19. doi:10.1128/JCM.02127-19.

Yoon, S.S., E.-K. Kim, and W.-J. Lee. 2015. Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism.. Curr. Opin. Microbiol. 24:38–46. doi:10.1016/j.mib.2015.01.007.

Yuan, Z., E. Liu, Z. Liu, J.W. Kijas, C. Zhu, S. Hu, X. Ma, L. Zhang, L. Du, H. Wang, and C.

Wei. 2017. Selection signature analysis reveals genes associated with tail type in Chinese indigenous sheep. Anim. Genet. 48:55–66. doi:10.1111/age.12477.

Zhou, B., C. Wang, Q. Zhao, Y. Wang, M. Huo, J. Wang, and S. Wang. 2016. Prevalence and dissemination of antibiotic resistance genes and coselection of heavy metals in Chinese dairy farms. J. Hazard. Mater. 320:10–17. doi:https://doi.org/10.1016/j.jhazmat.2016.08.007.

Zilber-Rosenberg, I., and E. Rosenberg. 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution.. FEMS Microbiol. Rev. 32:723–735. doi:10.1111/j.1574-6976.2008.00123.x.