Universitat Autònoma de Barcelona

Genomic analysis of dairy and pigmentation traits in Murciano-Granadina goats

Dailu Guan

Universitat Autònoma de Barcelona

Departament de Ciència Animal i dels Aliments

Facultat de Veterinària

Centre de Recerca en Agrigenòmica (CRAG)

Departament de Genètica Animal

Genomic analysis of dairy and pigmentation traits in Murciano-Granadina goats

Doctoral thesis to obtain the Ph.D. degree in Animal Production of the Universitat Autònoma de Barcelona (UAB), September 2020.

Dailu Guan

Supervisor:

Dr. Marcel Amills

El Dr. Marcel Amills, professor agregat del Departament de Ciència Animal i dels Aliments de la Universitat Autònoma de Barcelona,

fa constar:

que el treball de recerca i la redacció de la memòria de la tesi doctoral titulada:

"Genomic analysis of dairy and pigmentation traits in Murciano-Granadina goats"

han estat realitzats sota la seva direcció per

Dailu Guan

i certifica:

que aquest treball s'ha dut a terme al Departament de Ciència Animal i del Aliments de la Facultat de Veterinària de la Universitat Autònoma de Barcelona i al Departament de Genètica Animal del Centre de Recerca en Agrigenòmica (CRAG), i es considera que la memòria resultant és apta per optar al grau de Doctor en Producció Animal per la Universitat Autònoma de Barcelona.

I perquè en quedi constància, signen aquest document el Setembre del 2020.

Dr. Marcel Amills Dailu Guan

This research was funded by the European Regional Development Fund (FEDER)/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación/Project Reference Grant: AGL2016-76108-R (2017-2019). We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016-2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain) as well as to the CERCA Programme/Generalitat de Catalunya. Dailu Guan was funded by a Ph.D. fellowship from the China Scholarship Council (CSC).

“路漫漫其修远兮

吾将上下而求索”

——屈原《离骚》

Contents

Summary

One of the main goals of this Thesis was to investigate the molecular basis of lactation in Murciano-Granadina goats from a transcriptomic perspective. Hence, we collected biopsies of mammary glands from seven Murciano-Granadina goats at three time points, i.e. 78 d (T1, early lactation), 216 d (T2, late lactation) and 285 d (T3, dry period) after parturition. By using a RNA-Seq approach, a differential expression analysis was carried out resulting in the identification of 1,654 differentially expressed (DE) genes (q-value ≤ 0.05 and absolute $\log_2$FC > 1.5). While the T1 vs. T2 contrast only yielded 42 DE genes, in the T1 vs. T3 and in the T2 vs. T3 contrasts 1,377 and 1,039 DE genes were detected, respectively. As expected, genes encoding milk protein components were significantly upregulated during lactation, i.e. *CSN1S1*, *CSN1S2*, *CSN2*, *CSN3*, *PAEP*, *LALBA*. Many of the DE genes were functionally linked to protein synthesis, lipid and carbohydrate metabolism, insulin signaling, calcium homeostasis, cell death, tissue remodeling and involution, as well as mammary immunity.

In the first study of this Thesis, we also aimed to uncover the genomic polymorphisms affecting milk yield and composition in Murciano-Granadina goats. The performance of a genome-wide association study (GWAS) with the GEMMA software made it possible to identify 24 quantitative trait loci (QTLs). Out of 24 significant associations, only three QTLs showed significant associations at the genome-wide level, i.e. QTL1 (chromosome 2, 130.72-131.01 Mb) for lactose percentage, QTL6 (chromosome 6, 78.90-93.48 Mb) for protein percentage and QTL17 (chromosome 17, 11.20 Mb) for both protein and dry matter percentages. By checking the overlapping of protein-coding genes detected in both the QTL mapping and the RNA-Seq analysis, we found 39 DE genes mapping to 14 genome-wide or chromosome-wide QTLs. The QTL6 region, which showed significant associations with protein, fat and dry matter

percentages, co-localized with casein genes which are upregulated during lactation. According to our results, the variability of the casein genes seems to be the main determinant of milk protein and fat content in Murciano-Granadina goats. We have also observed a low positional concordance between the GWAS signals detected in our study and those reported in French breeds, a finding that could be due to the existence of a remarkable degree of genetic heterogeneity or to technical factors.

The second paper aimed to address the question whether caprine casein polymorphisms emerged before (standing variation) or after (novel mutation) goat domestication. To this end, we collected 106 caprine whole-genome sequences from public databases and analyzed the distribution of single nucleotide polymorphisms (SNPs) mapping to the four casein genes (i.e. *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*) in bezoars and 4 groups of domestic goats from Europe, Africa, Far East, and Near East. A relevant fraction of casein SNPs were shared between domestic goats and bezoars i.e. from 36.1% (*CSN2*) to 55.1% (*CSN1S2*). Besides, more than 50% of the casein SNPs were shared by 2 or more domestic goat populations, and 18 to 44% are shared by all populations. This extensive sharing of polymorphisms in distant populations supports that they probably emerged before goat domestication and dispersal. The construction of haplotypes demonstrated that the majority of casein alleles present in domestic goats also segregated in the bezoar, e.g. A/B alleles of the *CSN1S1* and *CSN3* genes, and A allele of the *CSN2* gene. These alleles generally have been reported to have substantial effect on milk composition. We conclude that much of the extant diversity of the caprine casein locus is derived from standing variation which existed before goat domestication.

Another goal of the current Thesis was to characterize copy number variations (CNV) in Murciano-Granadina goats (paper III). To do so, we analyzed Goat SNP50 BeadChip data from 1,036 individuals. The use of the PennCNV and QuantiSNP software resulted in the discovery of 4,617 and 7,750 autosomal CNV, respectively. By applying the EnsembleCNV algorithm, these

CNV were assembled into 1,461 CNV regions (CNVR), of which 486 (33.3% of the total CNVR count) were consistently called by PennCNV and QuantiSNP and used in subsequent analyses. Our data highlights the existence of a considerable degree of structural variation in Murciano-Granadina goats. Probably, the use of large goat population allowed us to detect low frequency CNV which otherwise would have been missed.

In the set of 486 CNVR, we identified 78 gain, 353 loss and 55 gain/loss events. Their length (95.69 Mb) accounted for 3.9% of the goat autosomal genome (2,466.19 Mb), and the average size was estimated to be 196.89 kb, with sizes ranging from 2.0 kb to 11.1 Mb. Moreover, genes co-localizing with CNVR were functionally enriched in olfactory transduction, ABC transporters and embryo development. One of the most interesting copy number variable genes is *ASIP*, which encodes the agouti signaling protein that drives the synthesis of yellow/red pheomelanin. Increased copy number in the *ASIP* locus was associated with a white pigmentation in goats, so our finding that the *ASIP* CNVR was also segregating in the black/brown Murciano-Granadina goats was quite paradoxical. In the fourth paper, we quantified *ASIP* copy number in eight goat breeds with different pigmentation patterns. We observed an increased *ASIP* copy number not only in Saanen goats but also in brown/black Murciano-Granadina and brown/blond Malagueña goats. These results preclude the existence of a simple linear relationship between *ASIP* copy number and white pigmentation.

The final aim of the Thesis was to investigate the genomic architecture of coat color in Murciano-Granadina goats (paper V). We carried out a genome-wide association study comprising 387 black and 142 brown individuals. This analysis resulted in the identification of a single significant peak on chromosome 18, which contains the melanocortin 1 receptor (*MC1R*) gene. Sequencing of the *MC1R* coding region and genotyping experiments evidenced that the c.801C>G (p.Cys267Trp) polymorphism tightly segregates with coat color, indicating that

the inheritance of coat color in Murciano-Granadina goats is essentially monogenic.

Resumen

Uno de los principales objetivos de esta Tesis es el de investigar las bases moleculares de la lactación en cabras de la raza Murciano-Granadina desde una perspectiva transcriptómica. Por lo tanto, recolectamos biopsias de glándula mamaria de siete cabras Murciano-Granadinas en tres puntos temporales, ésto es 78 días (T1, lactación temprana), 216 días (T2, lactación tardía) y 285 días (T3, período seco) después del parto. Mediante el uso de la tecnología RNA-Seq, se llevó a cabo un análisis de expresión diferencial que dio como resultado la identificación de 1654 genes diferencialmente expresados (DE) (*q-valor* $\leq 0,05$ y $\log_2$FC absoluto> 1,5). El contraste T1 vs. T2 permitió identificar 42 genes DE, mientras que en los contrastes T1 vs. T3 y T2 vs. T3 se detectaron 1377 y 1039 genes DE, respectivamente. La expresión de RNA mensajero de los genes que codifican las proteínas lácteas (*CSN1S1*, *CSN1S2*, *CSN2*, *CSN3*, *PAEP* y *LALBA*) se incrementó significativamente durante la lactación. Muchos de los genes DE estaban relacionados funcionalmente con el metabolismo de las proteínas, los lípidos y los carbohidratos, la ruta de señalización de la insulina, la homeostasis del calcio, la muerte celular programada, la remodelación e involución de los tejidos, así como la inmunidad mamaria.

En el primer estudio de esta Tesis, también se planteó identificar los polimorfismos genéticos que están asociados a la producción y a la composición de la leche en cabras Murciano-Granadinas. La realización de un estudio de asociación del genoma completo (GWAS) con el software GEMMA permitió identificar 24 *quantitative trait loci* (QTLs). De las 24 asociaciones significativas, solo tres lo fueron a nivel genómico, ésto es QTL1 (cromosoma 2, 130,72-131,01 Mb) para el porcentaje de lactosa, QTL6 (cromosoma 6, 78,90-93,48 Mb) para el porcentaje de proteína y QTL17 (cromosoma 17, 11,20 Mb) tanto para los porcentajes de proteína como de materia seca. Al verificar la concordancia entre genes localizados dentro de QTLs y genes expresados diferencialmente,

encontramos 39 genes que presentaban DE en alguno de los contrastes analizados y que además estaban localizados dentro o en la cercanía de 14 QTLs significativos a nivel genómico o cromosómico. La región QTL6, que mostró asociaciones significativas con los porcentajes de proteína, grasa y materia seca, contiene los genes de las caseínas cuya expresión aumenta durante la lactación. Nuestros resultados indican que la variabilidad de los genes de las caseínas es el principal determinante del contenido de proteína y grasa de la leche en las cabras Murciano-Granadinas. También hemos observado una baja concordancia posicional entre las señales GWAS detectadas en nuestro estudio y las descritas en razas caprinas francesas, lo que podría deberse a la existencia de un notable grado de heterogeneidad genética o bien a factores técnicos.

El segundo artículo tenía como objetivo determinar si los polimorfismos de los genes de las caseínas caprinas surgieron antes o después de la domesticación de las cabras. Con este fin, recopilamos 106 secuencias genómicas caprinas procedentes de diversas bases de datos públicas y analizamos la segregación de polimorfismos nucleotídicos sencillos (SNPs) para los cuatro genes de las caseínas (*CSN1S1*, *CSN2*, *CSN1S2* y *CSN3*) en bezoares (los ancestros salvajes de las cabras domésticas) y 4 grupos de cabras domésticas originarias de Europa, África, Lejano Oriente y Oriente Próximo. Una fracción relevante de SNPs de las caseínas segregó tanto en cabras domésticas como en bezoares, es decir, del 36,1% (*CSN2*) al 55,1% (*CSN1S2*). Por otra parte, más del 50% de los SNPs de los genes de las caseínas fueron compartidos por 2 o más poblaciones de cabras domésticas, y del 18 al 44% fueron compartidos por todas las poblaciones. Estos resultados sugieren que una fracción importante de los SNPs de las caseínas ya estaba presente en el bezoar antes de su domesticación. Por otra parte, la reconstrucción de alelos a partir de datos SNP demostró que la mayoría de los alelos de las caseínas detectados en cabras domésticas también segregan en el bezoar, p.e. alelos A/B de los genes *CSN1S1* y *CSN3*, y el alelo A del gen *CSN2*. En diversas publicaciones se ha descrito que estos alelos muestran asociaciones significativas con la composición de la leche, y en algunos casos

también se ha determinado la existencia de relaciones causales. En definitiva, concluimos que una parte importante de la diversidad existente en los genes de las caseínas caprinas deriva de la variación genética que ya segregaba en el bezoar antes de su domesticación.

Otro objetivo de la Tesis consistía en caracterizar las variaciones del número de copias (CNV) en cabras Murciano-Granadinas (artículo 3). Así pues, analizamos los datos obtenidos a través del genotipado de 1036 cabras Murciano-Granadinas con el Goat SNP50 BeadChip. Mediante la utilización de las herramientas PennCNV y QuantiSNP, se identificaron 4617 y 7750 CNV autosómicos, respectivamente. Al aplicar el algoritmo EnsembleCNV, estas CNV se ensamblaron en 1461 regiones CNV (CNVR), de las cuales 486 (33,3% del recuento total de CNVR) fueron identificadas consistentemente por PennCNV y QuantiSNP y utilizadas en análisis posteriores. Nuestros datos indican la existencia de un grado considerable de variación estructural en las cabras de la raza Murciano-Granadina. Probablemente, el uso de una gran población de cabras nos permitió detectar CNV de baja frecuencia que, de otra manera, no hubieran sido identificadas. En el conjunto de 486 CNVR, identificamos 78 eventos de ganancia del número de copias, 353 de pérdida y 55 de ganancia/pérdida. Las CNVR cubrieron el 3,9% del genoma autosómico de la cabra (2466,19 Mb), y el tamaño medio de las CNVR fue de 196,89 kb, con un rango que oscilaba entre 2,0 kb y 11,1 Mb. Adicionalmente, los genes cuya posición coincidía con la de las CNVR estaban relacionados funcionalmente con la la transducción olfativa, los transportadores ABC y la desarrollo embrionario. Uno de los genes cuya posición coincidía con un CNVR fue el locus agouti signaling protein (*ASIP*), que codifica la proteína de señalización agouti que promueve la síntesis de feomelanina amarilla/roja. En algunas publicaciones, el aumento del número de copias del locus *ASIP* se asoció con una pigmentación blanca en el ganado caprino, por lo que nuestro hallazgo de que el *ASIP* CNVR también segrega en las cabras Murciano-Granadinas, que son negras o marrones, fue bastante paradójico. En el cuarto artículo, cuantificamos el número de copias

del gen *ASIP* en ocho razas de cabras con diferentes patrones de pigmentación. Observamos un aumento del número de copias del gen *ASIP* no solo en las cabras Saanen, que son blancas, sino también en cabras de las razas Murciano-Granadina y Malagueña (son marrones o rubias). Estos resultados no concuerdan con la existencia de una relación lineal simple entre el número de copias del gen *ASIP* y la pigmentación blanca.

El último objetivo de la Tesis consistió en investigar la arquitectura genómica del color de la capa en cabras Murciano-Granadinas (artículo 5). Llevamos a cabo un GWAS en el que se incluyeron datos de 387 individuos negros y 142 marrones. Este análisis dio como resultado la identificación de una asociación altamente significativa en el cromosoma 18, que contiene el gen del receptor de la melanocortina 1 (*MC1R*). La secuenciación de la región codificante del gen *MC1R* y los experimentos de genotipado evidenciaron que el polimorfismo c.801C> G (p.Cys267Trp) está asociado muy significativamente al color de la capa, lo que indica que la herencia del color de la capa en las cabras Murciano-Granadinas es esencialmente monogénica.

Resum

Un dels principals objectius d'aquesta Tesi és el d'investigar les bases moleculars de la lactació en cabres de la raça Murciano-Granadina des d'una perspectiva transcriptòmica. Amb aquesta finalitat, vam recollir biòpsies de glàndula mamària de set cabres Murciano-Granadines en tres punts temporals, és a dir 78 dies (T1, lactació primerenca), 216 dies (T2, lactació tardana) i 285 dies (T3, període sec) després del part. Mitjançant l'ús de la tecnologia RNA-Seq, es va dur a terme una anàlisi d'expressió diferencial que va donar com a resultat la identificació de 1654 gens diferencialment expressats (DE) (q-valor ≤ 0,05 i $\log_2$FC absolut > 1,5). El contrast T1 vs T2 va permetre detectar 42 gens DE, mentre que, en els contrastos T1 vs T3 i T2 vs. T3 es van detectar 1377 i 1039 gens DE, respectivament. L'expressió dels RNA missatgers que codifiquen les proteïnes làcties (*CSN1S1*, *CSN1S2*, *CSN2*, *CSN3*, *PAEP* i *LALBA*) es va incrementar significativament durant la lactació. Molts dels gens DE estaven relacionats funcionalment amb el metabolisme de les proteïnes, els lípids i els carbohidrats, la ruta de la senyalització de la insulina, l'homeòstasi del calci, la mort cel·lular programada, la remodelació i involució tisular, així com la immunitat mamària.

En el primer estudi d'aquesta Tesi, també es va plantejar la identificació dels polimorfismes genètics que afecten la producció i la composició de la llet en cabres Murciano-Granadines. La realització d'un estudi d'associació del genoma complet (GWAS) amb el programari GEMMA va permetre identificar 24 *quantitative trait loci* (QTLs). De les 24 associacions significatives, només tres ho van ser a nivell genòmic: QTL1 (cromosoma 2, 130,72-131,01 Mb) per al percentatge de lactosa, QTL6 (cromosoma 6, 78,90-93,48 Mb) per al percentatge de proteïna i QTL17 (cromosoma 17, 11,20 Mb) tant pels percentatges de proteïna com de matèria seca. Al verificar la concordança entre els gens localitzats dins de QTLs i gens expressats diferencialment, vam trobar

39 gens que presentaven DE en algun dels contrastos analitzats i que a més estaven localitzats dintre o en la proximitat de 14 QTLs significatius a nivell genòmic o bé cromosòmic. La regió QTL6, que va mostrar associacions significatives amb els percentatges de proteïna, greix i matèria seca, conté els gens de les caseïnes, l'expressió dels quals augmenta durant la lactació. Els nostres resultats indiquen que la variabilitat dels gens de les caseïnes és el principal determinant del contingut de proteïna i greix de la llet a les cabres Murciano-Granadines. També hem observat una baixa concordança posicional entre els senyals GWAS detectats en el nostre estudi i els descrits en races franceses, la qual cosa podria atribuir-se a l'existència d'un notable grau d'heterogeneïtat genètica o bé a factors tècnics.

El segon article tenia com objectiu determinar si els polimorfismes dels gens de les caseïnes caprines van sorgir abans o després de la domesticació de les cabres. Amb aquesta finalitat, es van recopilar 106 seqüències genòmiques caprines procedents de diverses bases de dades públiques i es va analitzar la segregació de polimorfismes nucleotídics senzills (SNPs) localitzats en els quatre gens de les caseïnes (*CSN1S1, CSN2, CSN1S2* i *CSN3*) tant en bezoars (els ancestres salvatges de les cabres domèstiques) com en 4 grups de cabres domèstiques originàries d'Europa, Àfrica, Orient Llunyà i Pròxim Orient. Una fracció rellevant dels SNPs de les caseïnes va segregar tant en les cabres domèstiques com en els bezoars, és a dir, del 36,1% (*CSN2*) al 55,1% (*CSN1S2*). D'altra banda, més del 50% dels SNPs dels gens de les caseïnes van ser compartits per 2 o més poblacions de cabres domèstiques, i del 18 al 44% van ser compartits per totes les poblacions. Aquests resultats suggereixen que una fracció important dels SNPs de les caseïnes ja estava present en el bezoar abans de la seva domesticació. D'altra banda, la reconstrucció d'al·lels a partir de dades de SNP va demostrar que la majoria dels al·lels de les caseïnes detectats en cabres domèstiques també segreguen en el bezoar, p.e. al·lels A/B dels gens *CSN1S1* i *CSN3*, i al·lel A del gen *CSN2*. En diverses publicacions s'ha descrit que aquests al·lels mostren associacions significatives amb la composició de la

llet, i en alguns casos també s'ha determinat l'existència de relacions causals. En definitiva, vam concloure que una part important de la diversitat existent en els loci de les caseïnes caprines deriva de la variació genètica que ja segregava al bezoar abans de la seva domesticació.

Un altre objectiu de la tesi consistia a caracteritzar les variacions del nombre de còpies (CNV) en cabres Murciano-Granadines (article 3). Així doncs, vam analitzar les dades generades a través del genotipat de 1036 cabres Murciano-Granadines amb el Goat SNP50 BeadChip. Mitjançant la utilització de les eines PennCNV i QuantiSNP, es van identificar 4617 i 7750 CNV autosòmiques, respectivament. Al aplicar l'algoritme EnsembleCNV, aquestes CNV es van acoblar en 1461 regions CNV (CNVR), de les quals 486 (33,3% del recompte total de CNVR) van ser identificades consistentment per PennCNV i QuantiSNP i utilitzades en els anàlisis posteriors. Els nostres resultats indiquen l'existència d'un grau considerable de variació estructural en les cabres de la raça Murciano-Granadina. Probablement, l'ús d'una gran població de cabres va permetre detectar CNV de baixa freqüència que d'altra manera no haguessin estat identificades. En el conjunt de 486 CNVR, vam detectar 78 esdeveniments de guany del nombre de còpies, 353 de pèrdua i 55 de guany/pèrdua. Les CNVR van cobrir el 3,9% del genoma autosòmic de la cabra (2.466,19 Mb), i la longitud mitjana de les CNVR va ser de 196,89 kb, amb un rang de mida que oscil·lava entre 2,0 kb i 11, 1 Mb. A més, els gens la posició dels quals coincidia amb la de les CNVR estaven relacionats funcionalment amb la la transducció olfactiva, els transportadors ABC i la desenvolupament embrionari.

Un dels gens la posició del qual coincidia amb un CNVR va ser el locus agouti signaling protein (*ASIP*), que codifica la proteïna de senyalització agouti que promou la síntesi de feomelanina groga/vermella. En algunes publicacions, l'augment del nombre de còpies del locus *ASIP* es va associar amb una pigmentació blanca a les cabres domèstiques, així que la nostra troballa de que la CNVR que conté el gen *ASIP* també segrega en les cabres Murciano-Granadines, que són negres o marrons, va ser força paradoxal. En el quart article,

quantifiquem el nombre de còpies del gen *ASIP* en vuit races de cabres amb diferents patrons de pigmentació. Observem un augment del nombre de còpies del gen *ASIP* no només en les cabres Saanen, que són blanques, sinó també en cabres de les races Murciano-Granadina i Malaguenya (marrons o rosses). Aquests resultats no concorden amb l'existència d'una relació lineal simple entre el nombre de còpies del gen *ASIP* i la pigmentació blanca en cabres.

L'últim objectiu de la Tesi va consistir en investigar l'arquitectura genòmica del color de la capa en cabres de la raça Murciano-Granadina (article 5). Vam dur a terme un GWAS que abastava 387 individus negres i 142 marrons. Aquesta anàlisi va permetre la identificació d'una associació altament significativa en el cromosoma 18, que conté el gen de receptor de la melanocortina 1 (*MC1R*). La seqüenciació de la regió codificant del gen *MC1R* i els experiments de genotipat van evidenciar que el polimorfisme c.801C> G (p.Cys267Trp) està associat molt significativament amb el color de la capa, la qual cosa indica que l'herència del color de la capa en les cabres Murciano-Granadines és essencialment monogènica.

List of Tables

Chapter 1 General Introduction

Chapter 3 Papers and studies

Paper I

Paper II

Paper III

Paper IV

List of Figures

Chapter 1 General Introduction

Chapter 3 Papers and studies

Paper I

Paper II

Paper III

Paper IV

Paper V

List of publications included in the present Thesis

Guan D, Landi V, Luigi-Sierra MG, Delgado JV, Such X, Castelló A, Cabrera B, Mármol-Sánchez E, Fernández-Alvarez J, de la Torre Casañas JLR, Martínez A, Jordana J & Amills M. Analyzing the genomic and transcriptomic architecture of milk traits in Murciano-Granadina goats. *Journal of Animal Science and Biotechnology*. 2020; 11:35.

Guan D, Mármol-Sánchez E, Cardoso TF, Such X, Landi V, Tawari NR & Amills M. Genomic analysis of the origins of extant casein variation in goats. *Journal of Dairy Science*. 2019; 102:5230-5241.

Guan D, Castelló A, Landi V, Landi V, Luigi-Sierra MG, Álvarez JF, Cabrera B, Delgado JV, Such X, Jordana J & Amills M. A genome-wide analysis of copy number variation in Murciano-Granadina goats. *Genetics Selection Evolution.* 52:44.

Guan D, Castelló A, Luigi-Sierra MG, Landi V, Delgado JV, Martínez A & Amills M. Estimating the copy number of the agouti signaling protein (*ASIP*) gene in goat breeds with different color patterns. (submitted to Livestock Science).

Guan D, Martínez A, Luigi-Sierr MG, Delgado JV, Landi V, Castelló A, Álvarez JF, Such X, Jordana J & Amills M. Exploring the genomic architecture of coat color in Murciano-Granadina goats. (in preparation).

Chapter 1

General Introduction

1.1 The importance of the dairy goat sector in Spain

Goat milk production plays a key role in the economy of the developing world, being also an important asset in the food production system of high-income countries (Pulina et al., 2018, Miller and Lu, 2019). In Europe, goats yearly yielded 2.7 million tonnes of whole fresh milk throughout 2014 to 2018 according to the data released by the Food and Agriculture Organization of the United Nations database (FAOSTAT, http://www.fao.org/faostat). With 1.1 million tonnes per year on average (2014-2018), Southern European countries were the largest contributors. With regard to Spain, it produces around 486 thousand kilograms of whole fresh milk yearly, being the largest producer, followed by Greece (**Figure 1**). It is estimated that Spain currently has 2.66 million goats over 12 months of age, mainly distributed in Andalucía (991,844 heads), which is the main producing region, Castilla-La Mancha (410,162 heads), Extremadura (267,018 heads) and Región de Murcia (217,274 heads) (**Figure 1**). Females account for ~77.93% of the total census, and the majority of them (1.65 million) have at least one parity (**Figure 1**). Indeed, most of Spanish goats are bred for milk production, being Murciano-Granadina (MUG) and Malagueña the two most prominent dairy breeds (Sepe and Argüello, 2019).

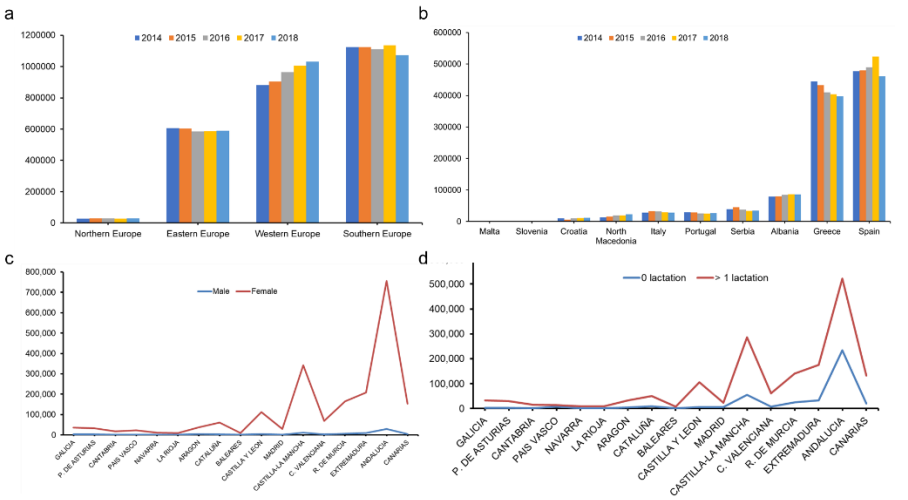

Figure 1. (**a**) The yield of whole fresh goat milk (tonnes) in Northern, Eastern, Western and Southern Europe in 2014-2018. (**b**) The yield of whole fresh goat milk in southern European countries in 2014-2018. Data presented in (a) and (b) have been retrieved from the Food and Agriculture Organization of the United Nations (FAOSTAT, http://www.fao.org, lastly accessed in April 8, 2020). (**c**) Census of goats and bucks in Spanish autonomous regions. (**d**) Number of female goats classified according to their lactation status in Spanish autonomous regions. The data of (c) and (d) have been retrieved from the Spanish Ministry of Agriculture, Fisheries and Food (https://www.mapa.gob.es, lasted accessed September 20, 2020).

1.1.1 Main features of the Murciano-Granadina breed

Murciano-Granadina is the Spanish goat breed with the largest census (112,000 heads) among the 22 caprine breeds officially recognized by the Ministry of Agriculture, Fisheries and Food (https://www.mapa.gob.es/). Generally speaking, MUG goats display a straight/sub-concave profile and medium proportions (**Figure 2**). The height and weight of MUG goats display sex-dependent differences, i.e. male (77 cm, 65 kg) and female (70 cm, 50 kg, respectively). Two main coat colors, black and brown, are distinguished (**Figure 2**). Currently, this breed is distributed throughout Spain, but according to the

website of the Ministry of Agriculture, Fisheries and Food (https://www.mapa.gob.es), the largest populations can be found in Andalusia (37,808 heads) and Murcia (26,647 heads) (**Table 1**). Moreover, MUG goats have been also introduced in Morocco, Algeria, Greece, and South America due to their excellent adaptability to marginal environments with high temperatures and scarce food. Noteworthy, the MUG breed originated in mountainous areas of Murcia and Granada with extreme temperatures in summer and winter. Moreover, MUG goats are able to feed on agricultural by-products. Although MUG goats can be used for meat production, with an average daily gain of 166 g/day and 7-10 kg of carcass weight for kids, it is primarily devoted to milk production. MUG herds raised in semi-intensive farms are milked once a day and each goat, on average, makes 6 lactations (250 days per lactation approximately) throughout its productive life. Moreover, MUG goats can produce 530 kg of milk per lactation, with 5.6 % of fat and 3.6% of protein content, a feature which is favorable for cheese production (https://www.mapa.gob.es).

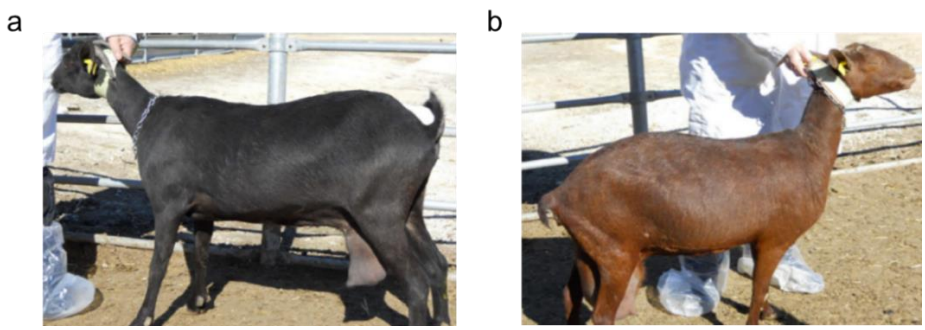

Figure 2. Pictures of black (**a**) and brown (**b**) Murciano-Granadina goats. These pictures were kindly provided Drs. Juan Manuel Serradilla (Universidad de Córdoba), Baltasar Urrutia (Instituto Murciano de Investigación y Desarrollo Agrario y Alimentario) and Juan Carrizosa (Instituto Murciano de Investigación y Desarrollo Agrario y Alimentario).

Table 1. Number of farms and MUG goats registered in the herd book throughout the Spanish geography

Autonomous community	Farm	Female	Male	Both sexes
Andalucía	86	36,379	1,429	37,808
Aragón	3	2,303	72	2,375
Cantabria	1	90	4	94
Castilla La Mancha	18	14,877	534	15,411
Castilla y León	12	7,898	369	8,267
Cataluña	8	2,326	71	2,397
Comunitat Valenciana	14	6,080	339	6,419
Extremadura	19	8,240	221	8,461
Madrid	2	4,272	113	4,385
Murcia	29	25,505	1,142	26,647
País Vasco	1	148	5	153
Total	193	108,118	4,299	112,417

These data have been retrieved from the Spanish Ministry of Agriculture, Fisheries and Food (https://www.mapa.gob.es, lasted accessed April 8, 2020).

About the historical origin of Murciano-Granadina goats, they are the result of admixing two different Murciano and Granadina populations. This event was motivated by a decision from the Spanish government to integrate these two closed populations into a single breed in 1999, which could facilitate having a higher census for intensive selection (Delgado et al., 2017). The existence of two divergent gene pools was detected with a panel of 25 microsatellites in 2010 (Martinez et al., 2010), but the Granadina component has been decreasing steadily and it is now nearing extinction (Delgado et al., 2017). The program of genetic improvement for MUG goats was established in 2012 with the aim of estimating breeding values for phenotypes of economic interest, mainly milk production and, more secondarily, milk composition. This program is being performed by the Asociación Nacional de Caprino de Raza Murciano-

Granadina (CAPRIGRAN), and the breeding goals for the MUG breed include the improvement of milk production and morphology. The former is achieved by taking into account four criteria: milk yield and protein and fat productions, which are measured in the autonomous milking control centers, and, on the other hand, casein genotypes (Delgado et al., 2017). Phenotype recording for morphology is achieved through the measurement of linear scores for body, udder and leg traits. In general, morphological criteria are complex and integrate a combination of different phenotypes: (1) Structure and capacity (height, thorax width, and angle of the rump); (2) Milk structure (angularity and bone quality); (3) Breast system (udder anterior insertion, middle suspensory ligament, udder width, udder depth, nipple insertion, and nipple diameter); and (4) Legs (rear view of the legs, lateral view of the legs, and mobility) (Delgado et al., 2017). The measurements of milk and morphological traits mentioned above are used for estimating breeding values by using the **B**est **L**inear **U**nbiased **P**rediction (BLUP) Animal Model implemented in the MTDFREML package (Boldman et al. 1993, Delgado et al., 2017). For measurements of milk production and quality, the statistical model takes into account as fixed effects the herd-year interaction, month of parturition and kidding size, doe age as a covariable, and random effects including the individual additive genotype effect on the trait and the environmental permanent effect (Delgado et al., 2017). For conformation traits, the univariate animal model considers several fixed effects including herd, year of qualification, month of qualification, and number of parturitions, doe days of lactation until qualification as a covariable and the random effect including the individual additive genotype effect on the trait (Delgado et al., 2017). By using these procedures, the breeding values of MUG individuals which are candidates to become breeders are estimated and selection is implemented. Two types of selection schemes are defined in the document "*Programa de mejora de la raza caprina Murciano Granadina*" (https://www.mapa.gob.es/es/ganaderia/temas/zootecnia/Programa%20de%20Mejora%20de%20la%20Raza%20Murciano_Granadina.%20Definitivo._tcm30

-114417.pdf). **Intra-herd selection** aims to select the mothers of the bucks which will become the future breeders. In this case, an Individual Multi-trait Selection Index is used to select the best females (less than 10% of population). This process is made once a year and it provides the males used for natural reproduction or artificial insemination. Subsequently, the male offspring of the mating between these selected females and elite bucks undergoes two cycles (at 3 and 7 months of age) of phenotypic selection. Individuals overcoming this initial selection are sent to a testing station where they are trained for sperm collection. At this point, it begins the process of **between-herd selection**, where the male candidates are genetically evaluated, with a BLUP animal model, on the basis of phenotypic information provided by at least 80 daughters distributed in at least three farms. Of course, in the testing station the absence of observable defects, and the reproductive and growth ability of the candidates are also evaluated. The best individuals are integrated in the panel of elite breeders, and then they are genotyped for the casein loci to obtain a complementary view of their improving potential for dairy traits. These efforts have resulted in approximately 108,118 females and 4,299 males recorded in the Genealogical Book of the MUG breed up to now (**Table 1**). The availability of high throughput sequencing and genotyping methods in recent years makes it possible to improve animal productivity by using genomic selection (Hayes et al., 2009, Hayes et al., 2013, Rupp et al., 2016). However, the genetic architecture of dairy and other relevant traits in MUG goats remains largely unknown, and from an economic point of view genomic selection might not be cost effective although the price of the SNP chips is decreasing each year. These factors limit the application of genomic selection technologies to the improvement of the MUG breed.

1.2 Genomic technologies: theoretical basis and applications

1.2.1 RNA-Seq can be used to massively sequence transcriptomes

In 2007, Emrich and colleagues reported for the first time an approach that significantly increased the sequencing throughput of the transcriptome by coupling laser capture microdissection (LCM) and 454 sequencing technologies (Emrich et al., 2007). The subsequent development of powerful sequencing platforms such as Roche 454 (Life Science), SOLiD22 (Applied Biosystems), and Illumina sequencers enhanced the performance of transcriptomic profiling experiments based on RNA sequencing (RNA-Seq) (Goodwin et al., 2016, Stark et al., 2019). RNA-Seq is able to profile the abundance and structure of all transcripts in a given sample, being widely used to construct atlas of gene expression as well as to detect differential gene expression in samples subjected to different experimental conditions (Emrich et al., 2007, Conesa et al., 2016, Stark et al., 2019). Other applications are related with the analysis of alternative splicing, co-regulation between genes, allele-specific expression and the identification of variants (Han et al., 2015). The RNA-Seq technology possesses important advantages over existing technologies, such as tiling microarray and cDNA sequencing (Wang et al., 2009). On the one hand, RNA-Seq is able to make use of small amounts of RNA samples with a relatively low cost, and it can be used in any organism and tissue (Wang et al., 2009, Han et al., 2015, Stark et al., 2019). On the other hand, RNA-Seq can measure gene expression in a much broader range than microarrays and, needless to say, it can be implemented to a much larger scale than quantitative polymerase chain reaction (qPCR) (Wang et al., 2009, Git et al., 2010). Therefore, RNA-Seq technology is employed as a routine tool in molecular biology to gain new insights into the physiology of many simple and complex traits (Han et al., 2015, Stark et al., 2019).

A typical workflow of a RNA-Seq experiment involves RNA preparation, cDNA synthesis and library construction, a procedure that includes fragmentation and adaptor ligation, and finally sequencing in a high throughput

platform (**Figure 3**). The initial step is to isolate RNA from either tissues or cells, being very important to take into account which type of RNA (basically short or long) needs to be sequenced. Indeed, different kinds of RNA require the use of different extracting methods in order to ensure that they can be adequately profiled (Hammerle-Fickinger et al., 2009). The sequencing methodology is different depending on the technology under consideration, so we will refer to Illumina sequencing platforms which have been used in the experimental work of the current thesis. After RNA purification, single stranded RNA is converted to cDNA by reverse transcription. Notably, third generation sequencing techniques, such as Nanopore, make it possible to sequence RNA molecules without the need of reverse transcription, thus improving many biases and artifacts resulting from this step (Ozsolak and Milos, 2011). After reverse transcription, cDNA needs to be fragmented into short sized molecules by using either enzymatic (e.g. transposase tagmentation reactions, non-specific endonuclease cocktails), or physical (e.g. sonication, acoustic shearing) methods (Kumar et al., 2012, Head et al., 2014). Subsequently, short oligonucleotide adaptors are ligated to either the 5'- or 3'-end of the cDNA fragments (Wang et al., 2009, Head et al., 2014, Hrdlickova et al., 2017, Stark et al., 2019). The quality of the resulting library is evaluated and, if satisfactory, it is loaded onto the flow cells of the sequencer machine (**Figure 3**). The Illumina sequencing method is essentially based on the principle of "sequencing by synthesis (SBS)". With this technology, each molecule is clonally amplified by bridge PCR and subsequently sequenced by the cyclic addition of fluorescently labeled deoxynucleoside triphosphate (dNTP) reversible terminators. After the addition of each nucleotide, the clusters are excited by a light source and a characteristic fluorescence is emitted and recorded making it possible to infer the sequence of each DNA molecule (Shendure and Ji, 2008, Fuller et al., 2009). The sequencing bias is largely controlled because four types of dNTP are simultaneously present and they naturally compete to hybridize to their complementary base in each sequencing cycle (Shendure and Ji, 2008, Fuller et al., 2009).

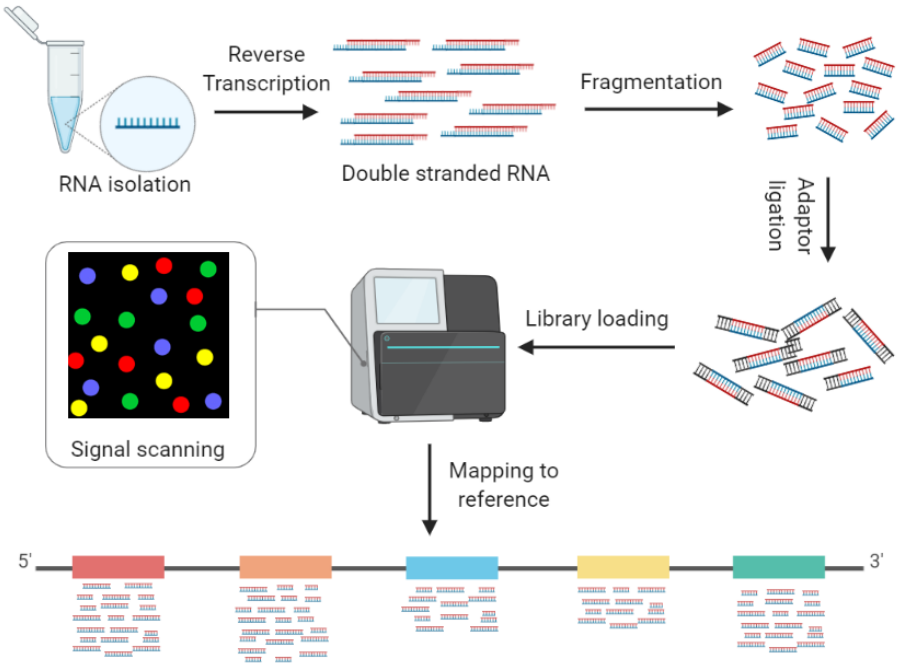

Figure 3. Main steps in a RNA-Seq analysis. First the RNA is extracted and reversely transcribed to cDNA. Subsequently, cDNA is fragmented and ligated to adaptors and a library is built and loaded into the sequencer. Once data is generated, they are subjected to bioinformatics analysis. Please see the text for additional details. This picture has been created with BioRender.com.

Once the transcriptome has been sequenced (**Figure 3**), sequence data needs to be analyzed with *in silico* bioinformatics methods (Conesa et al., 2016). Versatile tools have been developed for quality control, adaptor removal, alignment against a reference genome, and transcript assembly and quantification (Conesa et al., 2016). FastQC is commonly used to evaluate sequencing quality, (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), and several tools

for trimming reads, such as FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit), Trimmomatic (Bolger et al., 2014), and Cutadapt (Martin, 2011) have been implemented. It is worthwhile to mention that Trimmomatic is especially suitable for Illumina sequencing data (Bolger et al., 2014), and trim_galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) is also a convenient pipeline that includes Cutadapt and FastQC, which can be consistently used for both quality control and adaptor trimming. Over sixty different computational approaches have been developed to align reads to a reference genome (Fonseca et al., 2012). However, the preference for using one read mapper or another greatly depends on the sequencing platform, data type and size (Fonseca et al., 2012, Yorukoglu et al., 2016). So far there are four broadly used mappers for RNA-Seq data, i.e. Bowtie (Langmead, 2010), STAR (Dobin et al., 2012), TopHat (Trapnell et al., 2009) and HISAT (Kim et al., 2015). Bowtie, which was developed by Langmead et al. in 2009, leverages the ability of both Burrows-Wheeler indexing (Burrows and Wheeler, 1994) and full-text minute indexing (Ferragina and Manzini, 2000) to achieve ultrafast and memory-efficient alignment (Langmead et al., 2009). The newly extended Bowtie 2 tool allows gapped alignment, thus improving speed, sensitivity and accuracy (Langmead and Salzberg, 2012). The TopHat pipeline is able to identify splice junctions without the dependence of known splice sites based on aligned reads resulting from the Bowtie package (Trapnell et al., 2009, Langmead, 2010, Trapnell et al., 2012). A new version of this program (TopHat2) has been developed by achieving several enhancements, such as allowing for variable-length INDEL (insertion or deletion) with respect to the reference genome, and increased sensitivity and accuracy (Kim et al., 2013). Likewise, the performance of the HISAT aligner (Kim et al., 2015) was improved in 2019, and now it can map both DNA and RNA sequencing reads by using a graph Ferragina Manzini index which reduces the impact of variant positions on mappability (Kim et al., 2019). In this way, HISAT overperforms the STAR software package in handling

spliced sequences and coping with complex RNA sequences, such as those of chimeric and circular RNAs (Dobin et al., 2012, Dobin and Gingeras, 2015). In the step of RNA quantification, several metrics such as RPKM (reads kilobase per million reads), FPKM (fragments per kilobase per million reads), TPM (transcripts per million) are used. These quantitative metrics correct the estimates of gene expression for parameters such as sequencing depth and gene length, while the appropriate normalization of raw read counts eliminates additional biases (Conesa et al., 2016). There are several tools that were developed for counting RPKM, FPKM and TPM metrics, such as Kallisto (Bray et al., 2016), Cufflinks (Trapnell et al., 2010) and StringTie (Pertea et al., 2015). Others directly summarize raw reads, such as featureCounts (Liao et al., 2014) and HTSeq-count (Anders et al., 2015). In differential expression analysis, the expression of genes needs to be compared across two or more experimental conditions. Several methods implement statistical tools that correct different sources of bias before undertaking the differential expression analysis. For instance, edgeR (Robinson et al., 2009) introduces a Poisson model that corrects both technical and biological variability, while DESeq2 (Love et al., 2014) assumes a negative binomial distribution for gene expression estimates by fitting a generalized linear model for each gene on the basis of data from the read count matrix (Love et al., 2014, Conesa et al., 2016). Several other packages for RNA-Seq differential expression analysis have been comprehensively compared and reviewed in Oshlack et al. (2010), Seyednasrollah et al. (2013), and Conesa et al. (2016). In order to achieve the fast and effective analysis of RNA-Seq data, software packages can be combined into a single pipeline. For instance, the coupling of HISAT (Kim et al., 2015), StringTie (Pertea et al., 2015) and Ballgown (Frazee et al., 2015), allows users to analyze RNA-Seq data in a straightforward manner (Pertea et al., 2016). Another pipeline incorporating TopHat (Trapnell et al., 2009, Kim et al., 2013) and Cufflinks (Trapnell et al., 2010) was also widely used in previous years (Trapnell et al., 2012).

1.2.2 The development of high throughput SNP genotyping methods has enabled the performance of genome-wide association studies

A single nucleotide polymorphism (SNP) can be defined as a single base-pair difference in the DNA sequence of individual members of a species. They are the most abundant type of polymorphism and they show a remarkably uniform genomic distribution. About 1.48 million SNPs were discovered by using reduced representation shotgun sequencing in humans (Altshuler et al., 2000), while this number has increased up to 88 million due to the effort of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). A high number of SNPs has been also identified in domestic species such as cattle (26.7 million, Daetwyler et al., 2014), pigs (66.67 million, Groenen et al., 2012), and goats (10.70 million, Tosser-Klopp et al., 2014). High throughput genotyping systems have been implemented to characterize the variability of animal and plant populations. The identified SNP sites can be embedded into DNA microarrays that can be used to simultaneously type large numbers of SNPs (LaFramboise, 2009). Currently, there are two main companies which manufacture SNP chips, i.e., Affymetrix and Illumina, both of which essentially leverage the complementary mechanism of double stranded DNA (LaFramboise, 2009). Generally, a probe containing the SNP site and several surrounding nucleotides is fixed onto a microarray, which is subsequently hybridized to a target DNA (LaFramboise, 2009). The resulting changes of the fluorescent signal reflect the pattern of nucleotide hybridization and they can be used for identifying genotypes (LaFramboise, 2009). In 1998, the first commercial human SNP chip including 1,494 SNPs (i.e. HuSNP assay) became available (Wang et al., 1998). In contrast, the first commercial SNP array for cattle, the Illumina BovineSNP50 BeadChip (Matukumalli et al., 2009), appeared much later. Currently, there are about eleven commercial SNP arrays available for cattle, with SNPs numbers ranging from 6 K to 777 K. There are also four porcine SNP arrays, from 10 K to 68 K, and two SNP chips for sheep (Nicolazzi et al., 2015). For goats, Tosser-Klopp et al. (2014) designed the first caprine SNP chip, which

included 53,347 markers, under the framework of the International Goat Genome Consortium. Later on, Qiao et al. (2017) reported a 66 K caprine SNP chip which took advantage of a solution hybrid selection (SHS)-based target enrichment strategy. These SNP chips have been also widely used to characterize copy number variations (CNV), which can be defined as genomic duplications or deletions, with sizes between 50 bp and several Mb, that are polymorphic amongst individuals of a given species. Only a few genome-wide CNV scans have been performed in goats. Liu et al. (2019) used Goat SNP50 BeadChip data to detect 6,286 putative CNV in 1,023 samples from 50 goat breeds using PennCNV. These CNV were assembled into 978 CNV regions, covering ~262 Mb (~8.96%) of the goat genome. The segregation of CNV was concordant with the population history of goat breeds and several genes co-localizing with CNV had functions related with coat color, muscle development, metabolic processes, osteopetrosis, and embryonic development (Liu et al. 2019). Interestingly, genome-wide scans have also highlighted the existence of caprine CNV variants associated with growth and dairy traits (Kang et al. 2020, Liu et al. 2020) as well as with pigmentation patterns (Menzi et al. 2016, Henkel et al. 2019, see also section 1.5), indicating that variability in copy number could have important quantitative effects on simple and complex traits.

The development of high throughput SNP genotyping technologies also made it possible to implement genome-wide association studies (GWAS) for dissecting the genetic factors determining the phenotypic variation of observable traits. In a GWAS, a panel of markers with a genome-wide distribution are genotyped in a population with recorded phenotypes, then the magnitude and the significance of the associations between marker genotypes and phenotypes are evaluated with statistical methods (Bush and Moore, 2012, Tam et al., 2019). The success of the GWAS essentially depends on the amount of linkage disequilibrium between markers and causal mutations as well as population size, which has a great impact on statistical power (Bush and Moore, 2012, Sved and Hill, 2018). When populations are small, it becomes very difficult to identify

causal mutations with small effects on the trait or mutations, that despite having moderate or even large effects, are very rare (**Figure 4**). In reality, there are many factors determining the ability of GWAS to detect causal mutations including the heritability of the trait and the existence of cryptic population structure (Evangelou and Ioannidis, 2013, Schaid et al., 2018, Tam et al., 2019). These difficulties explain why the SNPs genotyped in the GWAS very often only explain a fraction of the phenotypic variance (h^2_{SNP}) that is lower than the heritability (h^2) of the trait (Manolio et al., 2009). This gap between genealogical heritability, which measures which fraction of the phenotypic variance is explained by additive factors, and h^2_{SNP} is often denominated as missing heritability. It should be taken into account, however, that genealogical heritability sometimes overestimates the magnitude of the additive component, so it should not be interpreted as an exact measurement of it. Last but not least, cryptic population structure and relatedness are important confounding factors that can result in spurious associations in GWAS. Therefore, adjusting population structure is critical in GWAS studies, and several tools have been developed to this end. For example, principal component analysis (PCA) data can be used to correct for population structure in GWAS (Price et al., 2006, Reich et al., 2008, Kang et al., 2010, Price et al., 2010, Bush and Moore, 2012, Pickrell and Pritchard, 2012). Other methods to infer population structure are multidimensional scaling (MDS), STRUCTURE, ADMIXTURE analysis and so on (Price et al., 2010, Alexander and Lange, 2011, Porras-Hurtado et al., 2013). In GWAS based on linear mixed models, individual relationships are accounted by the kinship matrix, so an additional correction based on PCA results would result in an overcorrection of the data (Kang et al., 2010, Price et al., 2010, Zhou and Stephens, 2012). In order to evaluate whether population stratification is well adjusted, Devlin and Roeder (1999) proposed the inflation factor λ, a genomic control parameter evaluating the impact of population stratification. If the λ value approximates 1, it means that population stratification does not exist; whilst anything below or above 1 means that population structure is not well corrected

(Devlin and Roeder, 1999, Price et al., 2010). Most recently, Wojcik et al. (2019) have found that the risk alleles responsible for a specific human disease differ from population to population, indicating that genetic heterogeneity is an important component of the inheritance of complex traits (Begum et al., 2012, Wojcik et al., 2019).

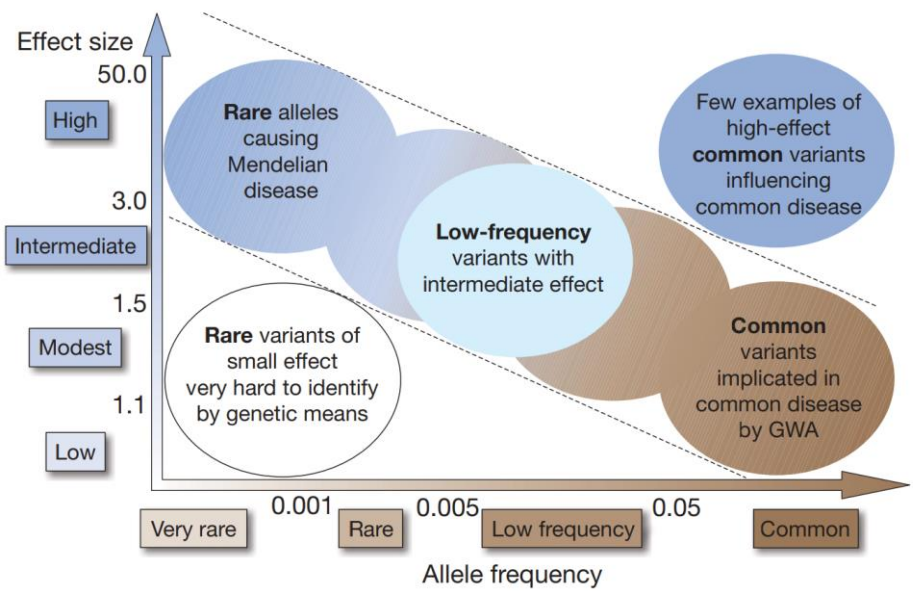

Figure 4. The ability of GWAS to detect causal mutations depends on their allele frequencies (*x*-axis) and effect sizes (*y*-axis) (Manolio et al., 2009). In general, genome-wide association (GWA) studies carried out in domestic species have a low statistical power due to constraints in sample size, making difficult the detection of causal alleles with small effects and/or very low frequencies.

Since approaches based on the linear mixed model (LMM) are able to correct for population structure and individual relatedness in GWAS analysis, the implementations based on LMM have become very popular. Examples of

software implementing this approach are GenABEL (Aulchenko et al., 2007), EMMAX (Kang et al., 2010), FaST-LMM (Lippert et al., 2011) and GEMMA (Zhou and Stephens, 2012). Specifically, GEMMA implements a genome-wide efficient mixed-model analysis that is summarized below (Zhou and Stephens, 2012):

$$y = W\alpha + x\beta + u + \varepsilon$$

Where y represents the vector of phenotypic values; W is a matrix with a column of 1 s and the fixed effects; α is a c-vector of the corresponding coefficients including the intercept; x is a n-vector of marker genotypes in each individual; β is the effect size of the marker (allele substitution effect); u is a n-vector of random effects with a n-dimensional multivariate normal distribution $(0, \lambda\tau^{-1}K)$, being τ^{-1} the variance of the residual error, λ the ratio between the two variance components and K is n × n relatedness matrix derived from marker genotypes (n is number of individuals); and ε is a vector of errors (Zhou and Stephens, 2012). The guidelines and protocols for GWAS analysis have been reviewed in several publications (Anderson et al., 2010, Ott et al., 2011, Barsh et al., 2012, Bush and Moore, 2012), so they will not be specified here. Moreover, it should be noted that due to the fact that thousands or millions of tests are carried out in a GWAS, it is absolutely necessary to correct for multiple testing with strategies such as the Bonferroni method, which can suffer from a high rate of false negatives (Haynes, 2013), or the false discovery rate (FDR) approach, that controls the expected proportion of falsely rejected hypotheses and it is less stringent (Benjamini and Hochberg, 1995). Obviously, GWAS data generated by different laboratories can be simultaneously analyzed by carrying out a meta-analysis, thus increasing very significantly population size and statistical power. Recently, a meta-analysis for stature, using 58,265 cattle from 17 populations with 25.4 million imputed whole-genome sequence variants, made it possible to identify 163 significantly associated genomic regions which explained at most 13.8% of the phenotypic variance (Bouwman et al., 2018).

At some instances, RNA-Seq and GWAS data are integrated to gain new insights into the genetic basis of complex traits. For instance, Deng et al. (2019) detected 1,420 differentially expressed (DE) genes across different lactation time points in buffaloes by using a RNA-Seq approach. Besides, they detected 976 genes displaying genome-wide associations with milk yield. By integrating both sources of information through a system biology approach, these authors identified 12 promising candidate genes with potential effects on milk (Deng et al., 2019). GWAS and RNA-Seq can also be combined to detect expression quantitative trait locus (eQTL), i.e. regions of the genome with quantitative effects on gene expression (Gilad et al., 2008, Westra and Franke, 2014). However, the positional concordance of QTL and eQTL mapping seems to be low (van den Berg et al., 2019), probably due to the high biological complexity of metric traits. Additionally, eQTL mapping is cost-intensive because measuring gene expression in a large sample of individuals is quite expensive.

1.3 Investigating the molecular basis of lactation through the analysis of gene expression

In order to nourish young babies, mammals developed lactation, which is a unique process conducted in the mammary gland which produces a fluid rich in proteins, lipids and calcium which can be used by the newborn as a rich source of nutrients (Akers, 2016, Hassiotou and Geddes, 2013). In goats, an empty udder can weight 6 kg, so it is attached to the body by a complex and strong set of suspensory ligaments in the rear, foreudder, lateral sides, as well as the medial suspensory ligament. The mammary gland contains connective and secretory tissue, which is constituted by alveoli, in which milk is synthesized by epithelial mammary cells, and the ductal system which brings milk to the gland cisterns

and from here to the teat cisterns (Akers, 2016, Ferreira et al., 2013, Hassiotou and Geddes, 2013). The general physiology of the mammary gland and lactation are well known in cattle, sheep and goats, three species that have been selected for millennia to increase milk production (Gross and Bruckmaier, 2019). A mechanistic lactation model demonstrated that dairy goats had greater milking potential compared to sheep, though both performed worse than cows (Dijkstra et al., 2010). The lactation curve of dairy goats (total milk yield level, lactation persistency, milk yield in mid-lactation) is greatly affected by diverse genetic and environmental factors such as parity, age of the doe, kidding season, herd effect, level of production, feeding practice as well as the breed itself (Gipson and Grossman, 1990, Arnal et al., 2018). In the study carried out by León et al. (2012), the lactation curve of MUG goats was best modeled by the quadratic spline function, resulting in a total lactation yield of 434 kg at 210 days, a daily milk yield of 1.93 kg and 2.42 kg in the beginning and peak (day 45) of lactation, respectively, followed by a gradual decrease in milk production (**Figure 5**). This study has also analyzed the influence of geographical regions, type of kidding, number of lactation and season of kidding on the lactation curve, and found that these factors could affect the shape of the lactation curve, especially the initial level of milk yield (**Figure 5**, León et al., 2012). Moreover, the secretion of milk is, to a large extent, synchronized by the endocrine system that produces hormones regulating the development of the mammary gland, the initialization of lactation and the maintenance of milking (Tucker, 1981, Akers, 2016, Ferreira et al., 2013, Rezaei et al., 2016). There are eight types of hormones with important roles in mammary gland development and milk synthesis (**Figure 6**), i.e. estrogen, progesterone, prolactin, growth hormone, placental lactogen, glucocorticoids, thyroid hormones and insulin (Tucker, 2000). At the initial step of milk synthesis, the mammary gland needs to grow and to develop under the stimulation of growth hormone and prolactin, adrenocortical steroids, oestrogens and progesterone (Svennersten-Sjaunja and Olsson, 2005, Brisken and O'Malley, 2010, Rezaei et al., 2016). Compared to cattle or even sheep, the physiology and

regulation of lactation of goats has been much less characterized. New and valuable knowledge could be obtained by using omics technologies (genomics, transcriptomics, metabolomics and proteomics) to understand the molecular events that make possible the production of milk in goats from a system biology perspective.

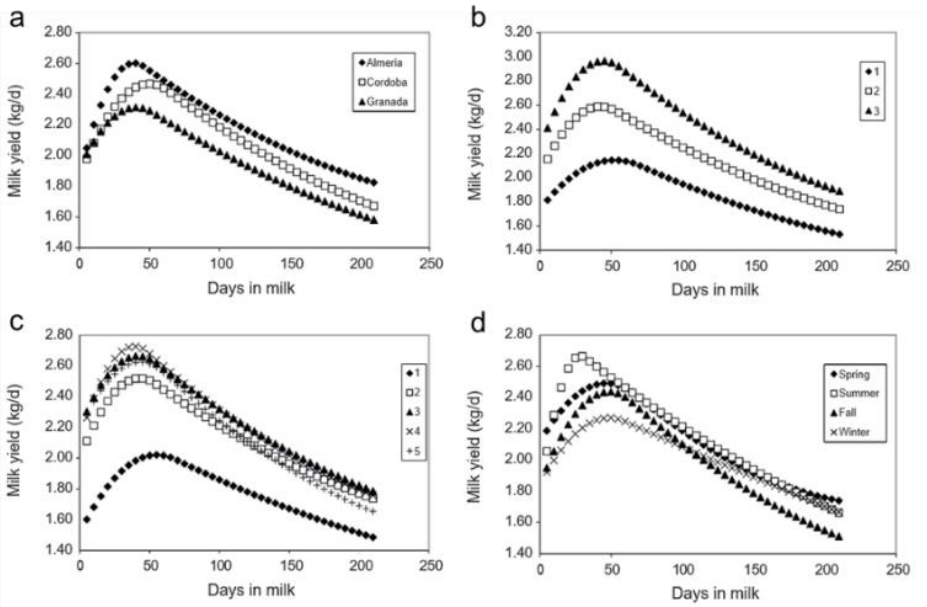

Figure 5. The lactation curve of Murciano-Granadina goats (León et al., 2012). The graphs show the impact of geographical regions (**a**), type of kidding (**b**), number of lactation (**c**) and season of kidding (**d**) on the shape of the lactation curve. Generally, milk yield of MUG goats reaches the production peak at ~50 days after parturition and then gradually decreases until the end of lactation. The analyzed factors affect generally the initial point of milking with the exception of "geographic region", which seems to have the same initial level of milking but reaches different peak levels at ~50 days.

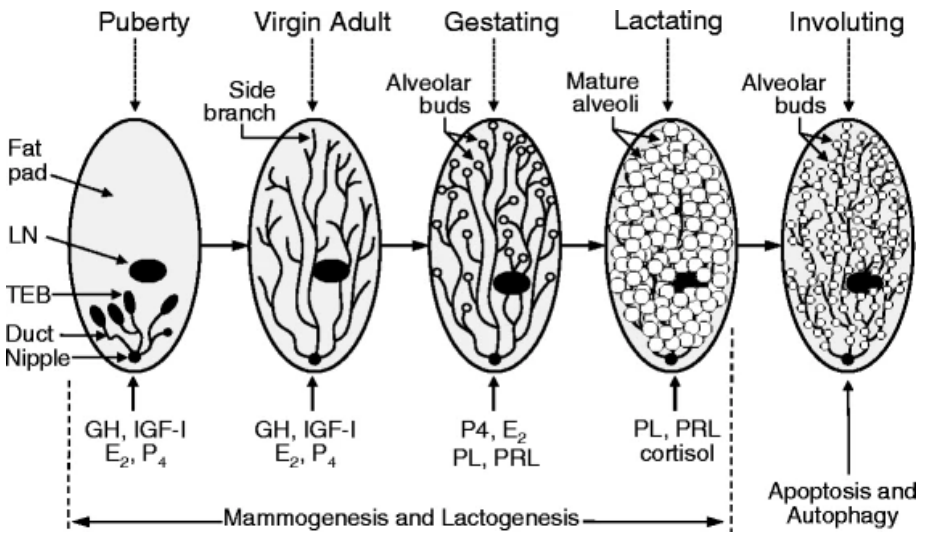

Figure 6. The developmental stages of the mammary gland are regulated by hormones (Rezaei et al., 2016). This process involves mammogenesis, lactogenesis, apoptosis and autophagy. Mammogenesis begins in puberty, and involves the formation of the terminal end buds (TEBs) at the tip of mammary duct and of the ductal branch system. This stage is regulated by growth hormone (GH), insulin-like growth factor (IGF-I), estrogen (E_2) and progesterone (P_4). Since the onset of pregnancy, the lobules and alveoli of the mammary gland further develop to form alveolar buds and such transformation is modulated by E_2, P_4, prolactin (PL) and prolactin receptor (PRL). After that, the initiation of lactation facilitates the formation of mature alveoli responsible for producing and secreting milk, a process influenced by PL, PRL and cortisol. Once breast-feeding ends, the involution of the mammary gland starts, involving apoptosis and autophagy. LN: lymph node.

1.3.1. Molecular analysis of lactation in cattle

Since the first study generating expressed sequence tags from pooled bovine mammary glands (Sonstegard et al., 2002), so far 186 studies have been conducted to understand the relationship between gene expression and lactation, and almost 76% of them focused on the dairy cow (**Figure 7**). Cánovas et al. (2010) found that as much as 19,175 genes (70% of the total number of bovine

mRNA genes) are expressed in the milk somatic cells from Holstein cows at two stages of lactation (day 15 and day 250) by using a RNA-Seq approach. However, this study did not investigate the potential function of expressed genes but just focused on the discovery of SNPs mapping to transcripts expressed in milk somatic cells (Cánovas et al., 2010). Since that, RNA-Seq has been widely used for investigating gene expression in the bovine lactating mammary gland in the framework of a system biology perspective. For instance, Wickramasinghe et al., (2012) sequenced the transcriptomes of milk somatic cells sampled at days 15 (transition), 90 (peak of lactation) and 250 (late lactation) of the lactation cycle of Holstein cows. They found that a total of 16,892 genes were expressed in transition lactation, 19,094 genes were expressed in peak lactation and 18,070 genes were expressed in late lactation, which means that ~69% of genes annotated in the bovine genome are expressed in milk somatic cells (Wickramasinghe et al., 2012). Moreover, genes involved in the synthesis of caseins, whey proteins and lactose displayed augmented mRNA levels in early lactation, while lipid genes were maximally expressed in transition and peak lactation (Wickramasinghe et al., 2012). Specifically, protein synthesis in the bovine mammary gland is centrally regulated by *ELF5* expression, and it is driven by the activities of glucose (e.g. *SLC2A1*, *SLC2A3*, *SLC2A8*) and amino acid transporters (e.g. *SLC1A1*, *SLC1A5*, *SLC36A1*, *SLC3A2*, *SLC7A1* and *SLC7A5*), and by the interaction of insulin signaling (e.g. *IRS1*) and mTOR signaling pathways (e.g. *FRAP1*, *EIF4E*, *EIF4EBP2*, *GSK3A* and *TSC1*). In contrast, milk fat synthesis and secretion are mediated by the coordinated action of the *SREBF1*, *SREBF2*, and *PPARG* genes (**Figure 8**, Bionaz and Loor, 2008, Bionaz and Loor, 2011). Mammary fat metabolism involves many different processes (**Figure 8**), such as fatty acid (FA) absorption from blood (*LPL* and *CD36*), intracellular FA trafficking (*FABP3*), activation of intracellular long-chain FA (*ACSL1*) and of short-chain FA (*ACSS2*), *de novo* FA synthesis (*ACACA*, *FASN*), FA desaturation (*SCD*, *FADS1*), triacylglycerol synthesis (*AGPAT6*, *GPAM*, *LPIN1*), lipid droplet formation (*BTN1A1*, *XDH*), ketone

body utilization (*BDH1*), and transcription regulation (*INSIG1*, *PPARG*, *PPARGC1A*). Moreover, Yang et al. (2015) profiled the gene expression of the milk fat globule at days 10 and 70 after calving via RNA-Seq and found 178 significantly DE genes. They also observed a functional enrichment of the mammary gland development, protein and lipid metabolism process, signal transduction, cellular process, differentiation and immune function (Yang et al., 2015). Besides, non-coding genomic elements are also important for bovine lactation, e.g. a total of 23,495 lncRNAs are expressed in the bovine mammary gland, and 3,746 show significant differences in expression between lactation and dry period (Yang et al., 2018). The comprehensive characterization of gene expression patterns in the bovine mammary gland provides an unprecedented systematic insight into the molecular dynamics of the lactation process.

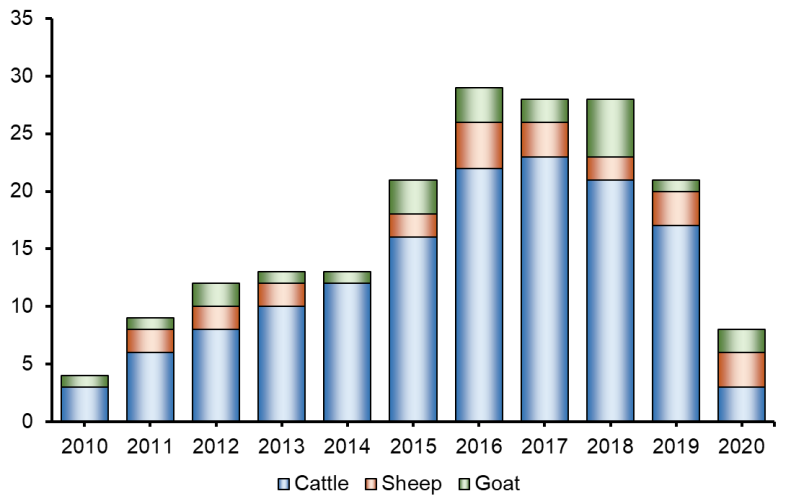

Figure 7. Number of scientific articles (*y*-axis) analyzing the lactation of dairy cattle, sheep and goats from a molecular perspective from 2010 to 2020 (*x*-axis). This graph is based on the results of a search in the PubMed database (https://pubmed.ncbi.nlm.nih.gov, the last accessed April 19, 2020) with terms "transcriptome + lactation" and "cattle/sheep/goats".

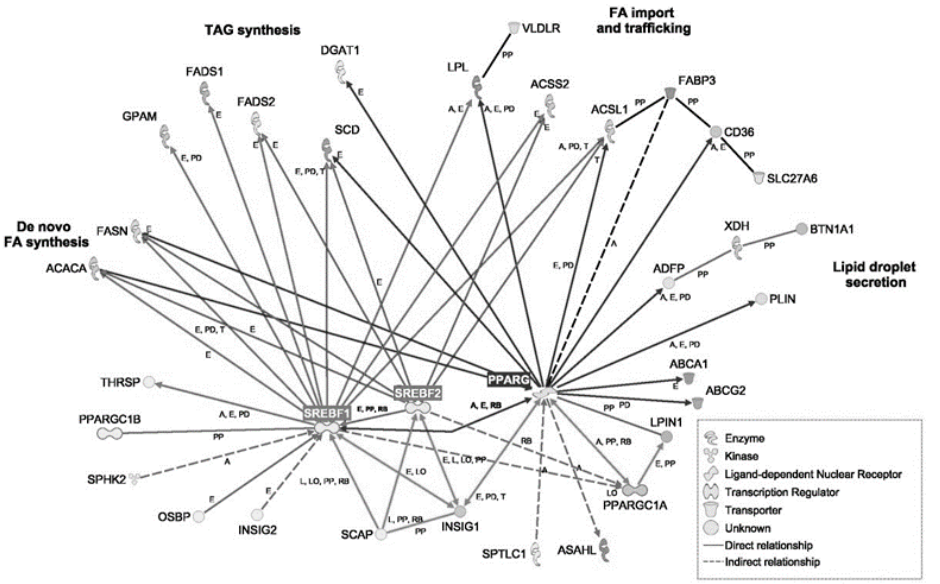

Figure 8. The proposed network of genes expressed in the bovine mammary gland that regulate and facilitate milk fat synthesis (Bionaz and Loor, 2008). This process involves *de novo* fatty acid (FA) synthesis, triacylglycerol (TAG) synthesis, FA import and trafficking, as well as lipid droplet secretion under the central coordination of the *SREBF1*, *SREBF2* and *PPARG* genes, which are also regulated by *SCAP*, *INSIG1*, *PPARGC1A* and *LPIN1* genes. More details can be found in the text.

1.3.2. Molecular analysis of lactation in sheep

In sheep, the first RNA-Seq study focused on the biology of late pregnancy and lactation was published in 2015, and revealed that about 13% of sheep genes were differentially expressed between the two studied time points (Paten et al., 2015). Genes related with cell proliferation, β-oxidation of fatty acids and translation were widely expressed in late pregnancy, while genes

mostly expressed during lactation were involved in the synthesis of fat and proteins, transportation, lipogenesis and remodeling (Paten et al., 2015). Another study carried out by Suárez-Vega et al. (2015) analyzed gene expression of milk somatic cells from four Assaf and four Churra ewes after lambing (days 10, 50, 120 and 150), resulting in about 67% of the annotated genes expressed in milk somatic cells and 573 DE genes across lactation points (Suárez-Vega et al., 2015). In this study, besides observing the differential expression of genes encoding casein α_{S1} (*CSN1S1*), casein α_{S2} (*CSN1S2*), casein β (*CSN2*), casein κ (*CSN3*), α-lactalbumin (*LALBA*) and progestagen-associated endometrial protein (*PAEP*), it was also found that the *GLYCAM1* and *B2M* genes are highly expressed in days 10, 50, 120 and 150 after lambing (Suárez-Vega et al., 2015). While the *GLYCAM1* gene encodes a protein that forms part of the fat globule; *B2M* is involved in the transfer of G immunoglobulins through the mammary epithelium (Suárez-Vega et al., 2015). Moreover, *GABRB3* (17.29), *COL4A2* (12.64), *CPXM2F* (12.58), *AM13C* (-11.78) and *IL20* (-11.75) were amongst the genes with the largest fold changes (log_2FC) in the late lactation stage (day 150) when compared to the beginning of lactation (day 50), possibly indicating their participation in mammary gland involution (Suárez-Vega et al., 2015). Very recently, the transcriptome profiling of the mammary gland has been performed in two Chinese ovine breeds, i.e. Small-Tailed Han and Gansu Alpine Merino sheep (Hao et al., 2019). This study has reported 407 and 373 breed-specifically expressed genes in the Small-Tailed Han and Gansu Alpine Merino breeds, respectively (Hao et al., 2019). Differentially expressed genes were enriched in functions related with catalytic activities, oxytocin signaling pathway and neuroactive ligand-receptor interaction (Hao et al., 2019).

1.3.3. Molecular analysis of lactation in goats

Early transcriptome studies in goats used microarrays from cattle to investigate gene expression (Faulconnier et al., 2011, Li et al., 2012). The first high-throughput sequencing study only profiled the expression of miRNAs in the mammary glands of goats in the dry period and peak lactation (Wang et al., 2017). Indeed, Wang et al. (2017) reported that miR-145 plays a critical role in fat metabolism by targeting insulin induced gene 1 (*INSIG1*), a key regulator of the expression of several genes linked to lipid synthesis during goat lactation. Likewise, Chen et al. (2018) identified another miRNA, chi-miR-3031, that by downregulating *IGFBP5* mRNA could promote the expression of β-casein, a major component of milk proteins. Another study analyzing the expression of genes in colostrum milk vs. milk collected at 120 days of lactation resulted in the identification of 207 upregulated and 122 downregulated genes (Crisà et al., 2016). As in dairy cow and sheep, genes with the highest expression during lactation are those encoding the main milk proteins, i.e. *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *PAEP* and *LALBA*. Moreover, Ji et al. (2019) profiled the gene expression of the caprine mammary gland across lactation and found DE genes related with biological regulation, cellular processes, metabolic processes, cells, organelles, binding, catalytic activity and transcriptional activity. Despite these efforts, the molecular basis of lactation in goats has not been characterized in depth yet.

1.4 Genetic analysis of dairy traits in goats

1.4.1 Heritability of dairy traits in goats

Narrow sense heritability (h^2) is the proportion of phenotypic variation explained by additive genetic variation segregating in the individuals of a given population (Visscher et al., 2008). Heritability is an important parameter in

animal breeding because it gives an idea of to what extent a trait can be modified by selection. Traits with high heritabilities will respond well to selection, while those with low heritabilities will be more difficult to improve because their inheritance, if any, is essentially non-additive. Generally, heritability values of dairy traits in goats are moderate. For instance, Gipson (2019) reported heritabilities for milk yield, fat and protein content of 0.26, 0.24 and 0.27, respectively. In MUG goats, heritabilities of 0.18, 0.16, and 0.25 for milk yield, fat content, and protein content have been reported, respectively (Analla et al., 1996). More recently, Miranda et al. (2019) measured heritabilities of parameters related with the lactation curve such as peak yield (0.13), yield (0.16) and persistency traits (0.08) in MUG goats (Miranda et al., 2019), obtaining low values. In contrast, much higher heritability values have been described in the Alpine and Saanen breeds, with registers of 0.30-0.34 for milk yield and 0.60-0.67 and 0.61-0.62 for protein and fat contents, respectively (Rupp et al., 2011). With regard to morphological traits, Rupp et al. (2011) have described moderate heritabilities for length (h^2 = 0.46-0.50), width (h^2 = 0.41-0.45), form (h^2 = 0.26-0.27), placement (h^2 = 0.30-0.38), angle (h^2 = 0.20-0.22) and orientation of the teats (h^2 = 0.32 -0.35), as well as udder floor position (h^2 = 0.34-0.37). With the availability of genome-wide genetic markers, the percentage of the phenotypic variance explained by SNPs (h^2_{SNP}) can be estimated (Yang et al., 2011, Carillier et al., 2014). For instance, Carillier et al. (2014), with 46,959 SNPs from the Illumina Goat SNP50 BeadChip, estimated the h^2_{SNP} values of traits indicated in **Table 2**, which ranged from 0.16 (somatic cell score) to 0.60 (protein content). In summary, there is a moderate degree of additive genetic variance for dairy and body conformation traits in goats, a feature that makes it possible to improve them by selection.

Table 2. Estimates of heritabilities of several dairy and morphological traits in Saanen and Alpine goats

Trait	Alpine goats	Saanen goats
Milk yield	0.31	0.26
Fat yield	0.28	0.25
Protein yield	0.31	0.25
Fat content	0.48	0.51
Protein content	0.60	0.56
Somatic cell score	0.20	0.16
Udder floor position	0.51	0.57
Udder shape	0.40	0.47
Rear udder attachment	0.47	0.52
Fore udder	0.44	0.42
Teat angle	0.42	0.45

Data presented in the current table are adapted from Carillier et al. (2014).

1.4.2 Early studies investigating the effects of casein and whey protein genotypes on milk yield and composition

The casein gene cluster is located in a 250 kb region of caprine chromosome 6. A total of 17 alleles have been identified in the *CSN1S1* gene (A, B1, B2, B3, B4, C, D, E, F, G, H, I, L, M, 0_1, 0_2 and N), 8 alleles in the *CSN2* gene (A, A1, O', O, B, C, D and E), 9 alleles in the *CSN1S2* gene (A, B, C, D, E, F, 0, Sub A and Sub B), and 16 alleles in the *CSN3* gene (A, B, B', B", C, C', D, E, F, G, H, I, J, K, L, M) (Marletta et al., 2007, Amills et al., 2002, Amills, 2014, Selvaggi et al., 2014). A molecular description of the casein alleles is outlined in **Figure 9**. By taking into account the effects of the *CSN1S1* alleles on the quantitative expression of the protein, they can be classified in four groups: strong alleles with ~3.5 g CSN1S1/L per allele (A, B1, B2, B3, B4, C, H, L and

M); medium alleles with ~1.1 g CSN1S1/L per allele (E and I); low alleles with ~0.45 g CSN1S1/L per allele (D, F and G); and null alleles (01, 02 and N) without CSN1S1 in milk (**Table 3**).

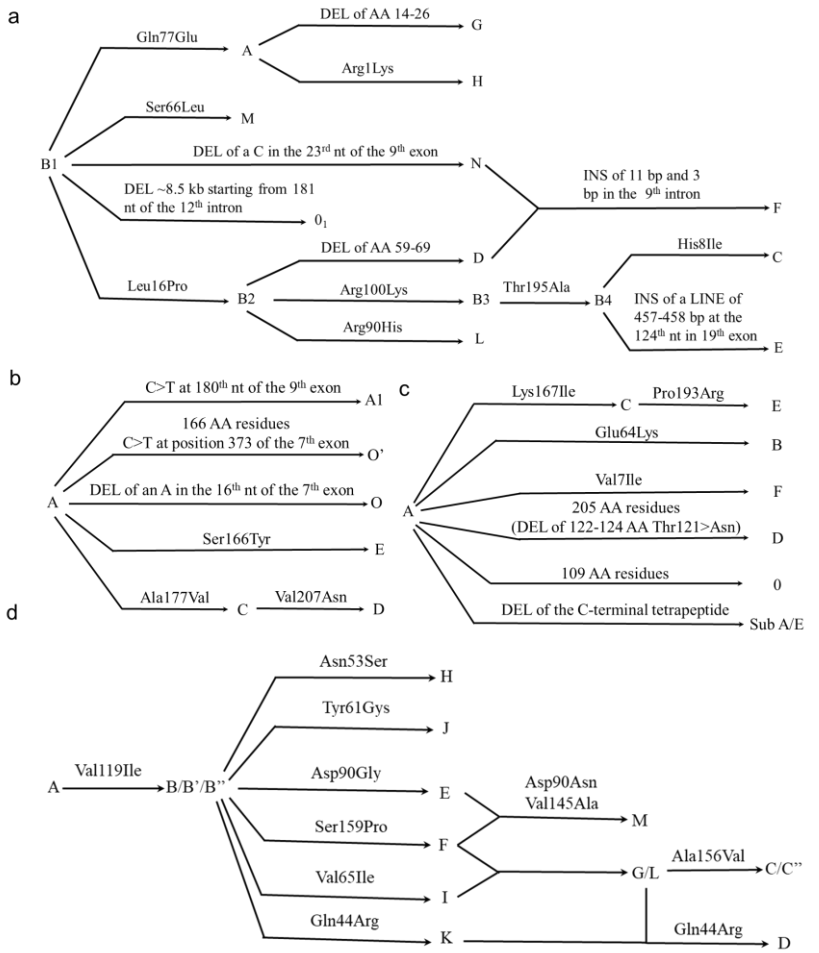

Figure 9. Diagram depicting the allelic variation of the *CSN1S1* (**a**), *CSN2* (**b**), *CSN1S2* (**c**) and *CSN3* (**d**) genes. This information is adapted from Marletta et al. (2007). Alleles that are differentiated just by silent mutations or that have not been well characterized are not included here. DEL: deletion; INS: insertion; AA: amino acid.

Table 3. The effect of *CSN1S1*, *CSN1S2* and *CSN2* alleles on the synthesis level of the corresponding proteins

Gene	Allele	Synthesis level (g casein /L/allele)
CSN1S1	A, B1, B2, B3, B4, C, H, L, M	3.5
	E, I	1.1
	D, F, G	0.45
	0_1, 0_2, N	0
CSN1S2	A, B, C, E, F	2.5
	D	~1.25
	0	0
CSN2	A, A1, B, C, D, E	5
	O, O'	0

Data presented in this table are adapted from Amills et al., (2002).

In Spanish MUG goats, the *CSN1S1*^{BB} genotype was associated with increased levels of the casein α_{S1} protein (Caravaca et al., 2008), but BB milk showed a lower curdling rate than EE milk (Caravaca et al., 2011). Notably, genotypes of the *CSN1S1* gene were not reported to influence protein, casein, and fat concentrations (Caravaca et al., 2009). In contrast, the AB and BB genotypes of the *CSN3* locus showed a tight association with increased levels of total casein and protein content in the MUG breed (Caravaca et al., 2009). Very recently, Pizarro Inostroza et al. (2019) genotyped 48 individual SNPs located in the casein loci of 159 MUG individuals and found associations between *CSN1S1* and *CSN3* polymorphisms and the milk fat and protein contents. In French Alpine and Saanen goats, *CSN1S1* genotypes showed highly significant effects on milk fat and protein contents and cheese yield (Carillier-Jacquin et al., 2016, Martin et al., 2002). Similar associations between *CSN1S1* haplotypes and

protein percentage and fat kilograms were observed in Norwegian goats (Hayes et al., 2006). Moreover, Hayes et al. (2006) found that *CSN3* haplotypes are associated with fat percentage and protein percentage (Hayes et al., 2006). Interestingly, they identified a unique deletion in the 12th exon of *CSN1S1* gene from Norwegian goats, which showed significant associations with fat yield instead of protein production (Hayes et al., 2006). In Sarda goats, the *CSN1S1* AB and BB genotypes showed strong associations with protein and fat percentages, while associations between *CSN2*^{AA} genotype and milk protein content and between *CSN1S2*^{AC} genotype and fat and protein production were also observed (Vacca et al., 2014). All in all, the findings mentioned above indicate that caprine casein genes show a high polymorphism and that part of this diversity is associated with milk yield and composition traits.

One of the major goat milk whey proteins is α-lactalbumin, which is encoded by the *LALBA* gene located on chromosome 5 (Selvaggi et al., 2014). For this gene, Cosenza et al. (2003) identified three variants in Italian goats, including a silent C>G substitution at the 5th position of the third exon (A^2 allele), an intronic T>C polymorphism in the 13th position of the 1st intron, and a C>G transversion in the 3' untranscribed region (Cosenza et al., 2003). In 2007, Lan and coworkers identified a missense mutation (p.Leu100Pro) in the 3rd exon (Lan et al., 2007). Afterwards, Zidi et al. (2014) reported 19 SNPs in the *LALBA* gene by analyzing goats from two Spanish breeds (MUG and Malagueña), but no significant association with lactose content was found (Zidi et al., 2014). Another important whey protein is β-lactoglobulin, currently named as progestagen-associated endometrial protein which is encoded by the *PAEP* gene located on caprine chromosome 11. Although many polymorphisms, including two missense mutations (p.Asp64Gly, p.Val118Ala) are found in the caprine *PAEP* gene (Yahyaoui et al., 2000, Ballester et al., 2005, Amills et al., 2002, Selvaggi et al., 2014), no association with milk phenotypes has been reported so far.

1.4.3 Using the genome-wide association study approach to elucidate the genomic architecture of dairy traits

A quantitative trait locus (QTL) is a genomic region containing one or several polymorphisms with quantitative effects on a phenotype. Dairy traits, including milk yield, fat yield and percentage, and protein yield and percentage are polygenic and have a complex inheritance (Hu et al., 2013). For instance, in dairy cattle 22,427 QTLs for milk fat content, 19,782 QTLs for milk protein content, 5,104 QTLs for milk yield, 3,933 QTLs for udder morphology and 2,382 QTLs for mastitis susceptibility have been detected so far (Cattle QTLdb, https://www.animalgenome.org/cgi-bin/QTLdb/BT/index). In dairy sheep, 234 QTLs for milk yield, 358 QTLs for milk fat yield and 134 QTLs for milk protein production have been reported (Sheep QTLdb, https://www.animalgenome.org/cgi-bin/QTLdb/OA/index). In contrast, no QTL for goat dairy traits have been deposited in the QTLdb database (Hu et al., 2013). Historically this lack of caprine QTL in public databases was due essentially to the absence of reports describing microsatellite markers uniformly distributed throughout the caprine genome (Amills, 2014). Before the release in 2014 of the Goat SNP50 BeadChip (Illumina Inc., San Diego, CA), that encompasses 53,347 SNPs (Tosser-Klopp et al., 2014), there was only one study mapping caprine QTL for milk production traits based on genotypic data provided by just 37 microsatellite markers (Roldán et al., 2008). By using the half-sib regression interval mapping approach implemented in the QTL Express software (Roldán et al., 2008, Seaton et al., 2002), this very preliminary study resulted in the identification of QTLs for milk yield, fat and protein percentages. The possibility of performing GWAS with the Goat SNP50 BeadChip has considerably improved the identification of QTLs associated with dairy traits in goats (**Table 4**). For instance, a total of 109 genomic regions have been associated with dairy traits in Saanen and Alpine goat populations, and the subsequent fine-mapping confirmed two missense polymorphisms in the *DGAT1* gene with causal effects

on fat content (Martin et al., 2017). Noteworthy, two missense mutations in this gene explained ~6% (p.Arg251Leu) and ~46% (p.Arg396Trp) of the genetic variance of milk fat content (Martin et al., 2017). Interestingly, Liu et al. (2018) detected a copy number variation overlapping the *DGAT1* gene which displayed a high frequency in a worldwide sample of goat breeds, but whether it plays role on milk fat production remains unknown. Another GWAS based on phenotypic and genotypic information provided by 2,381 goats made it possible to identify one genome-wide significant SNP on caprine chromosome 19 and four chromosome-wide significant SNPs for milk yield (Mucha et al., 2017). In the same study, four genome-wide significant QTLs, plus several chromosome-wide significant QTLs showed associations with conformation of udder attachment, udder depth, and front legs (Mucha et al., 2017). In a population of 810 Saanen and 1,185 Alpine goats, Martin et al. (2016a) also identified 17 genomic regions significantly associated with supernumerary teats, but only at the chromosome-wide level. In terms of susceptibility to mastitis, a genomic region (33-42 Mb) on chromosome 19 was significantly associated with somatic cell count (SCC) in the Saanen breed (Martin et al., 2018). Another QTL on this chromosome (19: 24.5-27 Mb) displayed negative pleiotropic effects on milk production (milk, fat yield, and protein yield) and udder traits (udder floor position and rear udder attachment) (Martin et al., 2018).

Table 4. Genomic regions showing associations with dairy and morphological traits in the Alpine (ALP) and Saanen (SAA) breeds and a composite population[a]

CHR[b]	QTL(Mb)	Phenotypes[c]	Breed[d]	Source
1	18.7	Rear udder attachment	ALP, SAA	Martin et al., 2018
1	20.5	Rear udder attachment	ALP	Martin et al., 2018
1	29.6	Rear udder	SAA	Martin et al., 2018
1	135.9-144.9	Teat form, fat content, teat length	ALP, SAA	Martin et al., 2017, Martin et al., 2018
2	88.6	Teat form	ALP	Martin et al., 2018
3	87.9	Foot orientation	ALP, SAA	Martin et al., 2018
4	71.2	Teat orientation	SAA	Martin et al., 2018
5	6.9	Udder profile	SAA	Martin et al., 2018
6	39.1-39.5	Chest depth	ALP, SAA	Martin et al., 2018
6	43.7	Teat length	SAA	Martin et al., 2018
6	74.3-87.2	Protein content, Fat content	ALP, SAA	Martin et al., 2017
8	38.83	Foot orientation	CMP	Mucha et al., 2017
8	40.1	Rear udder attachment	ALP, SAA	Martin et al., 2018
8	81.6	Chest depth	ALP	Martin et al., 2018
8	105	Rear udder	ALP, SAA	Martin et al., 2018
10	30.7	Chest depth	ALP	Martin et al., 2018
14	2.5-4.6	Fat content	ALP	Martin et al., 2017
14	9.2-16	Fat content	ALP, SAA	Martin et al., 2017
14	57.2-57.5	Teat length	ALP, SAA	Martin et al., 2018
14	79-79.9	Teat length, teat form	ALP	Martin et al., 2018
16	13.5	Rear udder	ALP, SAA	Martin et al., 2018
17	30.7-33.5	Fore udder	ALP, SAA	Martin et al., 2018
19	26.07-28.5	Udder attachment, udder depth, front legs, udder floor position, foot orientation	SAA, CMP	Mucha et al., 2017, Martin et al., 2018
19	33-42	Somatic cell count	SAA	Martin et al., 2018
19	58.2-58.3	Foot orientation	ALP, SAA	Martin et al., 2018
21	62.7-64.0	Protein yield	ALP, SAA	Martin et al., 2017
28	16	Rear udder attachment	SAA	Martin et al., 2018
28	33	Rear udder attachment	ALP, SAA	Martin et al., 2018
29	6.2	Teat length	SAA	Martin et al., 2018
29	40.5-41.9	Fore udder, teat angle	ALP, SAA	Martin et al., 2018

a: only those QTLs showing genome-wide significant associations and detected by GWAS are included here; b: Chromosome; c: milk traits are marked in bold; CMP: Composite population obtained by crossing Saanen, Toggenburg, and Alpine goats.

1.5. Genetic analysis of pigmentation in goats

According to classical genetic studies, color patterns in goats are determined by 12 alleles at the *Agouti* locus, 3 alleles at the *Brown* locus, 2 alleles at the *White Angora* locus and 1 allele at the *Extension* locus (Adalsteinsson et al., 1994, Sponenberg and LaMarsh, 1996, Sponenberg et al., 1998). The *Agouti* locus includes a dominant white or tan allele (A^{wt}), nine co-dominant alleles for black mask (A^{blm}), bezoar (A^{bz}), badgerface (A^{b}), grey (A^{g}), light belly (A^{lb}), swiss markings (A^{sm}), lateral stripes (A^{ls}), mahogany (A^{mh}), red cheek (A^{rc}), and a recessive nonagouti allele (A^{a}) (Adalsteinsson et al., 1994). Very recently, Henkel et al. (2019) reported a new peacock allele (A^{pc}) in this locus, which specifically exists in Peacock goats. At the *Brown* locus, the medium brown allele (B^{b}) is recessive to wild type (B^{+}) that is further recessive to a dark brown allele (B^{d}) (Sponenberg and LaMarsh, 1996). An experiment crossing Angora and non-Angora goats documented a dominant white (Wta^{D}) and a wild-type allele (Wta^{+}), indicating that non-white color patterns in goats are genetically caused by other loci (Sponenberg et al., 1998). Interestingly, a dominant black allele in the *Extension* locus displayed an epistatic effect on the *Agouti* locus (Sponenberg et al., 1998).

In mammals, the biosynthesis of melanin pigment takes place in melanocytes, which derive from the neural crest cells and migrate to hair follicles during embryonic development (Cichorek et al., 2013, Mort et al., 2015). The differentiation of melanocytes from the neural crests requires endothelin (END) signaling via homologous G-protein coupled, endothelin receptors (**Figure 10**, Saldana-Caboverde and Kos, 2010). Disruption of the process is expected to result in a diluted pigmentation due to the lack of mature melanocytes, e.g. mutated mice with $EDNRB^{sl/sl}$ genotype (*piebald lethal*) exhibited a white coat color (Saldana-Caboverde and Kos, 2010). Moreover, the maturation of

melanocytes is modulated by a complex array of regulatory factors including the melanocyte inducing transcription factor (MITF), KIT proto-oncogene, receptor tyrosine kinase (KIT), dopachrome tautomerase (DCT), the inactivation of which generally results in depigmentation (Cichorek et al., 2013, Mort et al., 2015). For instance, several polymorphisms in the *KIT* gene have causal effects on the color sidedness of cattle (Durkin et al., 2012), and on the patch and white coat colors of pigs (Rubin et al., 2012, Bickhart and Liu, 2014, Georges et al., 2018) and horses (Haase et al., 2007). In mature melanocytes, the synthesis of dark eumelanin and the red/yellow pheomelanin is controlled by a signaling pathway connected to the melanocortin 1 receptor (MC1R), which is located in-between the plasma membrane (Cichorek et al., 2013, Wolf Horrell et al., 2016). When MC1R binds to proopiomelanocortin (POMC), the levels of the second messenger cAMP are increased, thus activating enzymes that are essential for the synthesis of eumelanin such as tyrosinase (TYR), tyrosine related proteins 1 (TYRP1) and 2 (TYRP2), and MITF (**Figure 10**, Wolf Horrell et al., 2016). However, this pathway can be inhibited or blocked by the agouti signaling protein (ASIP). When ASIP, a negative agonist, binds to MC1R, it suppresses melanogenesis by decreasing melanocortin-induced cAMP production and further preventing expression of genes in downstream pathways including *TYR*, *TYRP1*, *TYRP2* and *MITF* (Aberdam et al., 1998). Many studies support the causal roles of MC1R and ASIP on animal coat colors (Kijas et al., 1998, Norris and Whan, 2008, Georges et al., 2018, Li et al., 2019), e.g. a 190-kb tandem duplication affecting the expression of the *ASIP* gene causes dominant white pigmentation in sheep breeds (Norris and Whan, 2008).

Several studies have investigated the causal factors determining pigmentation traits in goats by using a physiological candidate gene approach (Crepaldi and Nicoloso, 2007), but in general it has been difficult to associate genetic polymorphisms in pigmentation genes with color patterns (Feng et al., 2009, Fontanesi et al. 2009a, Fontanesi et al. 2009b, Badaoui et al., 2011), probably due to the existence of complex inheritance patterns and genetic

interactions (Badaoui et al., 2011). Fontanesi et al. (2009a) identified three missense mutations (p. Ala81Val, p. Val250Phe, and p.Cys267Trp) and one nonsense mutation (p.Glu225X) in the *MC1R* gene. These authors proposed a causal relationship between the p.Cys267Trp and the brown and black colorations of MUG goats (Fontanesi et al., 2009a). In another study performed by Wu et al. (2016), a relationship between a recessive allele in the *MC1R* gene and the red head and neck of Boer goats was established (Wu et al., 2016). The availability of high throughput technologies made it possible to investigate the genomic architecture of pigmentation traits with a much finer and comprehensive resolution than candidate gene studies. For instance, an undesirable coat "pink" color in the Saanen breed was associated with the genotype of the *ASIP* gene by performing a GWAS with 810 individuals (Martin et al., 2016b). Moreover, a 1 Mb copy number variant containing the *EDNRA* gene was tightly associated with a white coat color in Boer goats (Menzi et al., 2016). Additionally, a missense mutation, p.Gly496Asp in the *TYRP1* gene was identified in a GWAS as a causal factor for the brown (dominant) and black coat colors of Valais Blacknecked and Coppernecked goats (Becker et al., 2014, Dietrich et al., 2015).

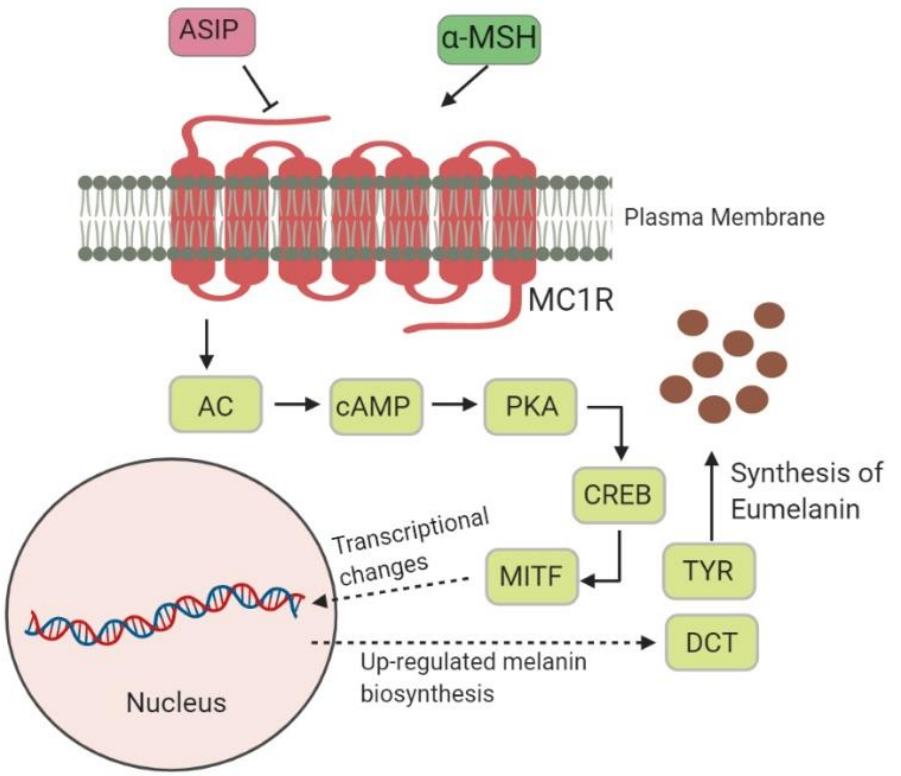

Figure 10. Melanin biosynthesis in the melanocyte (modified from Wolf Horrell et al., 2016). The binding of proopiomelanocortin (POMC, here indicated as α-MSH) to the melanocortin 1 receptor (MC1R) results in the activation of adenylyl cyclase (AC) and the accumulation of the second messenger cAMP. This would further activate the expression of downstream molecules including cAMP-dependent protein kinase (PKA), cAMP responsive binding element (CREB), melanocyte inducing transcription factor (MITF), tyrosinase (TYR), dopachrome tautomerase (DCT), as well as others (Wolf Horrell et al., 2016). In contrast, the binding of POMC to MC1R could be inhibited by agouti signaling protein (ASIP). If so, the ratio of eumelanin/pheomelanin would be changed affecting pigmentation. This image is created with BioRender.com.

Structural variations (SVs) also have an important role in the determination of the coat coloration of pigs (Rubin et al., 2012), cattle (Durkin et al., 2012) and sheep (Norris and Whan, 2008). In goats, Fontanesi and colleagues found a CNV co-localizing with the *ASIP* gene by using an array

comparative genome hybridization (aCGH) approach (Fontanesi et al., 2009b). This result was further supported by two whole genome sequencing studies that confirmed the existence of a CNV encompassing the caprine *ASIP* gene (Dong et al., 2015, Zhang et al., 2018). These authors consistently indicated the existence of a relationship between increased *ASIP* copy number and white coat color (Fontanesi et al., 2009b, Dong et al., 2015, Zhang et al., 2018), yet no functional experiment was performed. In a recent study carried out by Henkel et al. (2019), complex SVs in the caprine *ASIP* locus have been reported, including *ASIP*-SV1 (63.23-63.38 Mb), *ASIP*-SV2 (63.13-63.14 Mb), *ASIP*-SV3 (63.16-63.20 Mb) and *ASIP*-SV4 (63.13-63.25 Mb) (**Figure 11**). The *ASIP*-SV1 is a triplication spanning the *ASIP*, *AHCY* and part of *ITCH* loci (**Figure 11a**), which probably corresponds to a "white or tan" (A^{Wt}) allele segregating in the white Saanen and Appenzell breeds (**Figure 11b-11d**). The *ASIP*-SV2 is a tandem repeat with eight copies detected in Grisons Striped and Toggenburg goats (A^{sm}), while another tandem repeat (*ASIP*-SV3) was inferred to have five copies in the St. Gallen Booted goat (A^{b}). In Peacock goats, both *ASIP*-SV3 and *ASIP*-SV4 (a triplication) were simultaneously identified (Henkel et al., 2019). Moreover, these authors have identified two SVs downstream the *KIT* gene, including *KIT*-SV1 and *KIT*-SV2 (**Figure 12**). It is interesting to note that *KIT*-SV2 is a deletion but it is replaced by two copies of a triplication (89.21-89.23 Mb) located in the 5'-end of *RASSF6* gene (Henkel et al., 2019). Moreover, a CNV overlapping with the *ADAMTS20* gene was detected by Dong et al. (2015) and Liu et al. (2018). This CNV could be associated with goat pigmentation since the *ADAMTS20* gene plays a role in melanoblast survival (Silver et al., 2008) and co-localizes with a signature of selection detected when comparing white vs. colored goats (Bertolini et al., 2018). Despite these findings, the current knowledge about the genetic basis of color variation in goats, is still incomplete.

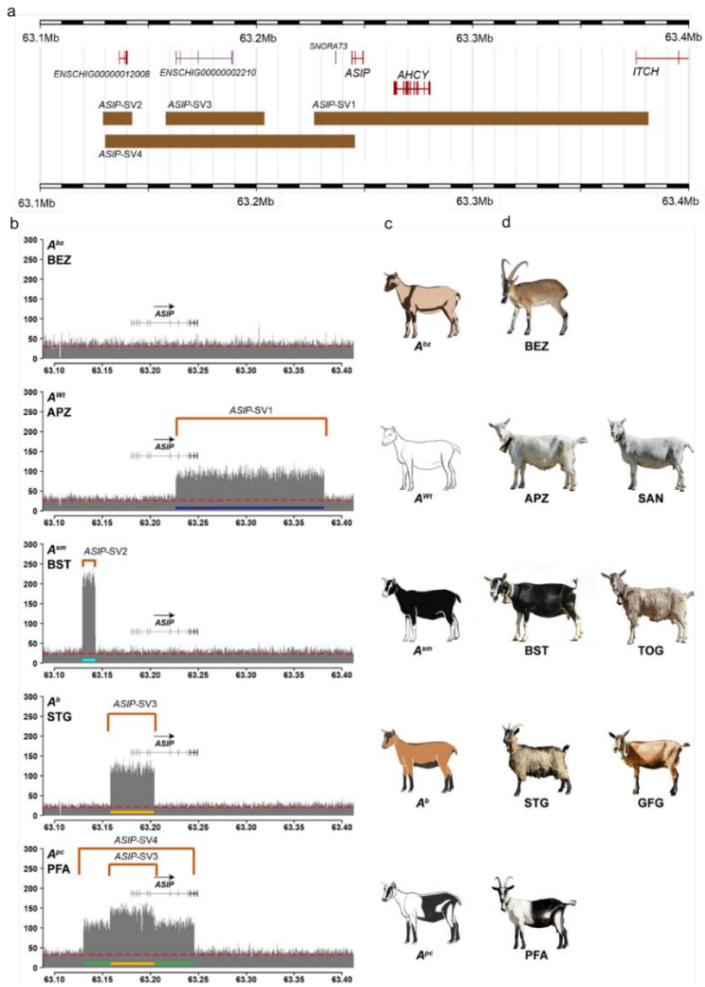

Figure 11. (**a**) The positional track of structural variations (SVs) in the caprine *ASIP* locus (Henkel et al. 2019). These authors reported four SVs in or near *ASIP* gene, i.e. *ASIP*-SV1 (63.23-63.38 Mb), *ASIP*-SV2 (63.13-63.14 Mb), *ASIP*-SV3 (63.16-63.20 Mb) and *ASIP*-SV4 (63.13-63.25 Mb). (**b**) Plotting of sequencing depth. The horizontal dashed line in red indicates the average sequencing depth. The positions of the *ASIP* gene and SV are shown in the upper part of the plotting. (**c**) and (**d**) Schematic pictures of coat coloration patterns associated with *ASIP* variation. The following abbreviations are used: BEZ, bezoar (*Capra aegagrus*, wild ancestor of domestic goat; APZ, Appenzell; SAN, Saanen; BST, Grisons Striped; TOG, Toggenburg; STG, St. Gallen Booted; GFG, Chamois Colored; PFA, Peacock.

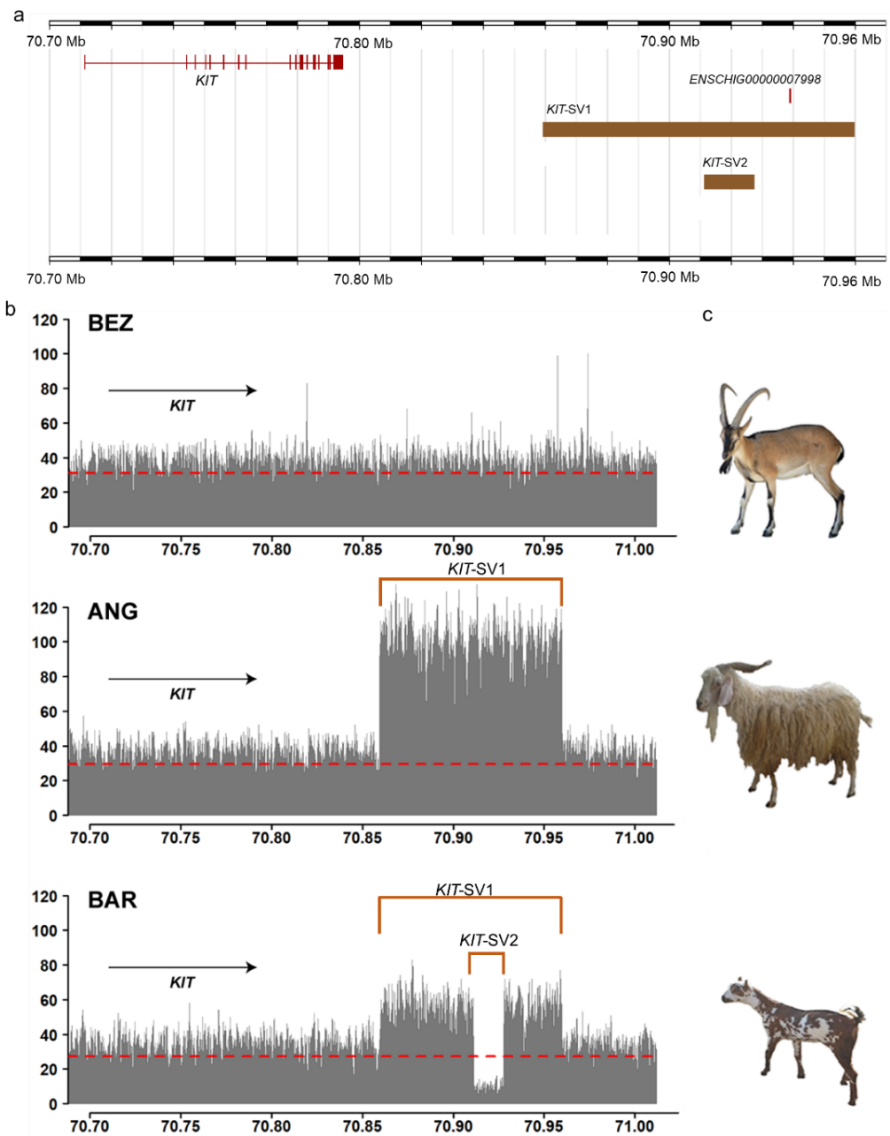

Figure 12. (**a**) The positional track of structural variations (SVs) near the caprine *KIT* locus (Henkel et al., 2019). There are two SVs mapping to ~63 kb downstream the *KIT* gene, i.e. *KIT*-SV1 (70.86-70.96 Mb), *KIT*-SV2 (70.91-70.92 Mb). (**b**) Plotting of sequencing depth. Compared to wild ancestor (bezoar, BEZ), SVs were detected in Pak Angora (ANG, *KIT*-SV1) and Barbari (BAR, both *KIT*-SV1 and *KIT*-SV2) goats. The authors identified a deletion (*KIT*-SV2) that was replaced by two copies of a triplication (89.21-89.23 Mb).

Chapter 2

Goals

This Ph.D. thesis was carried out under the framework of the project *Genomic analysis of the genetic determination of milk yield, composition and body condition and viability in Murciano-Granadina goats* (AGL2016-76108-R, 2017-2019), funded by the Spanish Ministry of Economy and Competitiveness. The general goals of this thesis can be divided in three main thematic blocks: (1) Understanding the molecular basis of lactation in goats and identifying the genetic determinants of milk yield and composition, (2) Dissecting the landscape of copy number variation in Murciano-Granadina goats, and (3) Exploring the genetic basis of coat color in the Murciano-Granadina breed. In this context, the specific goals of the current Thesis are:

- To investigate the molecular basis of lactation in Murciano-Granadina goats by using a RNA-Seq approach under the assumption that the elicitation and maintenance of lactation involves strong changes in the expression of genes involved in metabolism and other physiological processes (**Paper I**).

- Since, milk yield and composition traits have moderate heritabilities in goats, we wanted to identify the genetic determinants of these traits by performing a genome-wide association study comprising 822 Murciano-Granadina individuals with available phenotypes (**Paper I**).

- There is broad evidence that casein variation in goats has a relevant role in the genetic determinism of dairy traits, so we aimed to catalogue such diversity at a global scale and to elucidate whether it originated either before or after goat domestication by using a comprehensive data set of published whole-genome sequences from bezoars and European, Asian and African goats (**Paper II**).

- As a first step towards elucidating the potential role of structural

variation in goat phenotypes, one objective of the current Thesis was to identify copy number variations (CNV) segregating in the Murciano-Granadina breed by using SNP array data (**Paper III**).

- Copy number variation in the agouti signaling protein (*ASIP*) gene has been reported to influence coat color in goats. In the light of this, we were interested in quantifying the copy numbers of this gene in Murciano-Granadina goats as well as in other breeds with different pigmentation patterns to determine the relationship between *ASIP* CNV genotype and color (**Paper IV**).

- We also wanted to explore the molecular basis of the coat color of Murciano-Granadina goats by performing a genome-wide association study with 529 individuals with black (N=387) or brown (N=142) coat colors (**Paper V**). Our main interest was to dissect the genomic architecture of this trait (monogenic, oligogenic or polygenic) and to identify potential causal mutations.

Chapter 3

Papers and Studies

Journal of Animal Science and Biotechnology

Analyzing the genomic and transcriptomic architecture of milk traits in Murciano-Granadina goats

Dailu Guan[1], Vincenzo Landi[2], María Gracia Luigi-Sierra[1], Juan Vicente Delgado[2], Xavier Such[3], Anna Castelló[1,3], Betlem Cabrera[1,3], Emilio Mármol-Sánchez[1], Javier Fernández-Alvarez[4], José Luis Ruiz de la Torre Casañas[5], Amparo Martínez[2], Jordi Jordana[3], Marcel Amills[1,3]

[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain. [2]Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain. [3]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [4]Asociación Nacional de Criadores de Caprino de Raza Murciano-Granadina (CAPRIGRAN), 18340 Granada, Spain. [5]Servei de Granges i Camps Experimentals, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain.

Corresponding author: Marcel Amills (marcel.amills@uab.cat)

Journal of Animal Science and Biotechnology (2020). 11: 60.

https://doi.org/10.1186/s40104-020-00459-w

Abstract

Background

In this study, we aimed to investigate the molecular basis of lactation as well as to identify the genetic factors that influence milk yield and composition in goats. To achieve these two goals, we have analyzed how the mRNA profile of the mammary gland changes in seven Murciano-Granadina goats at each of three different time points, i.e. 78 d (T1, early lactation), 216 d (T2, late lactation) and 285 d (T3, dry period) after parturition. Moreover, we have performed a genome-wide association study (GWAS) for seven dairy traits recorded in the 1st lactation of 822 Murciano-Granadina goats.

Results

The expression profiles of the mammary gland in the early (T1) and late (T2) lactation were quite similar (42 differentially expressed genes), while strong transcriptomic differences (more than one thousand differentially expressed genes) were observed between the lactating (T1/T2) and non-lactating (T3) mammary glands. A large number of differentially expressed genes were involved in pathways related with the biosynthesis of amino acids, cholesterol, triglycerides and steroids as well as with glycerophospholipid metabolism, adipocytokine signaling, lipid binding, regulation of ion transmembrane transport, calcium ion binding, metalloendopeptidase activity and complement and coagulation cascades. With regard to the second goal of the study, the performance of the GWAS allowed us to detect 24 quantitative trait loci (QTLs), including three genome-wide significant associations: QTL1 (chromosome 2, 130.72-131.01 Mb) for lactose percentage, QTL6 (chromosome 6, 78.90-93.48 Mb) for protein percentage and QTL17 (chromosome 17, 11.20 Mb) for

both protein and dry matter percentages. Interestingly, QTL6 shows positional coincidence with the casein genes, which encode 80% of milk proteins.

Conclusions

The abrogation of lactation involves dramatic changes in the expression of genes participating in a broad array of physiological processes such as protein, lipid and carbohydrate metabolism, calcium homeostasis, cell death and tissue remodeling, as well as immunity. We also conclude that genetic variation at the casein genes has a major impact on the milk protein content of Murciano-Granadina goats.

Background

Understanding how genetic variation shapes the phenotypic diversity of milk traits not only implies the identification of such genetic determinants through genome-wide association studies (GWAS), but also a detailed knowledge about the genes playing a fundamental role in the progression of lactation. So far, very few GWAS have uncovered the genomic location and distribution of polymorphisms affecting milk yield and composition in goats. Martin et al. [1] genotyped, with the Goat SNP50 BeadChip, 2,209 Alpine and Saanen goats and performed association analyses with five dairy traits. Such work enabled the identification of 109 significant associations and further uncovered two polymorphisms in the *DGAT1* gene that have major effects on fat content by modifying the activity of this enzyme [1]. In another recent study, Mucha et al. [2] detected a single nucleotide polymorphism (SNP) on goat chromosome 19 displaying a genome-wide significant association with milk yield as well as a number of chromosome-wide significant associations with dairy traits on

chromosomes 4, 8, 14, and 29. Although these two studies represent a valuable step towards elucidating the genomic architecture of milk yield and composition traits in goats, analyzing a broader array of goat populations, as has been done in cattle [3], would provide a more comprehensive view of the genetic determinism of caprine dairy phenotypes.

The events promoting the initiation, maintenance and abrogation of lactation have been barely analyzed from a transcriptomic perspective in goats. Only one RNA-Seq study has investigated the changes experienced by the caprine mammary gland transcriptome across the production cycle (lactation vs. dry period) [4], while another one has compared the gene expression profile of goat milk somatic cells in colostrum and mature milk [5]. A third study investigated the transcriptomes of goat somatic cells, milk fat globules and blood cells via using microarrays [6]. This situation contrasts strongly with that of cattle, in which several studies outlining how the gene expression profile of the mammary gland changes in response to different experimental conditions have been published so far [7, 8, 9, 10]. Indeed, RNA-Seq studies performed in dairy cattle [11] and also in sheep [12] have revealed that hundreds of genes are differentially expressed (DE) in the mammary gland when lactating vs. non-lactating individuals are compared. Multiple lines of evidence indicate that many of these genes are related to mammary gland development, protein and lipid metabolism processes, signal transduction, differentiation and immune function, being very significant the downregulation of the protein and lipid biosynthetic machinery [11, 12].

The work presented here had two main objectives: 1) Elucidating the changes in the mammary transcriptome associated with the lactation stage by sequencing total RNA from mammary gland biopsies retrieved from seven Murciano-Granadina goats sampled at 78 d (early lactation), 216 d (late lactation) and 285 d (dry period) post-partum, and 2) Identifying the genetic determinants of milk yield and composition traits in Murciano-Granadina goats through a GWAS

approach comprising 822 individuals with records for 7 dairy traits registered during their 1st lactation.

Methods

Sequencing the mammary gland transcriptome along the lactation stage

Transcriptome sequencing

Mammary biopsies were retrieved from 7 Murciano-Granadina goats at each of the three time points, i.e. 78.25 ± 9.29 d (T1, early lactation), 216.25 ± 9.29 d (T2, late lactation) and 285.25 ± 9.29 d (T3, dry period) after parturition (**Additional file 1: Table S1**). The average age of the sampled goats was 5.88 ± 1.89 years and none of them was pregnant at T1, T2 or T3 (**Additional file 1: Table S1**). Mammary tissue was extracted with SPEEDYBELL 14G 150 mm semi-automatic biopsy needles (EVEREST Veterinary Technology, Barcelona, Spain) after applying local anesthesia to the region to be punctured. Samples were immediately submerged in RNAlater stabilization solution (Thermo Fisher Scientific, Barcelona, Spain) and shipped back to the laboratory for storage at -80 °C.

For isolating total RNA, a small piece of mammary gland tissue was submerged into liquid nitrogen and grinded to a fine powder with a mortar and a pestle. Subsequently, this powder was homogenized in 1 mL TRIzol reagent (Thermo Fisher Scientific, Barcelona, Spain) with a homogenizer device (IKA T10 basic ULTRA-TURRAX, Barcelona, Spain). The Ambion RiboPure kit (Thermo Fisher Scientific, Barcelona, Spain) was used to purify total RNA in accordance with the instructions of the manufacturer. The concentration and purity of RNA

preparations were evaluated with a Nanodrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Barcelona, Spain), while RNA integrity was checked in a Bioanalyzer-2100 (Agilent Technologies, Santa Clara, CA) by using an Agilent RNA 6000 Nano kit (Agilent Technologies, Inc., Santa Clara, CA). The RNA integrity number (RIN) ranged between 6.00-8.40, with an average of 7.43 ± 0.58.

Paired-end sequencing (2×76 bp) of the RNA was performed in the Centre Nacional de Anàlisi Genòmica (CNAG-CRG, http://www.cnag.crg.eu/). The RNA-Seq library was prepared with KAPA Stranded mRNA-Seq Illumina Platforms Kit (Roche). Briefly, 500 ng total RNA were used as the input material, the poly-A fraction was enriched with oligo-dT magnetic beads and the RNA was fragmented. The strand specificity was achieved during the second strand synthesis performed in the presence of dUTP. The blunt-ended double stranded cDNA was 3'-adenylated and Illumina platform compatible adaptors with unique dual indexes and unique molecular identifiers (Integrated DNA Technologies, Coralville, IA) were ligated. The ligation product was enriched by 15 cycles of PCR amplification and the quality of the final library was validated on an Agilent 2100 Bioanalyzer with the DNA 7500 assay (Agilent Technologies, Inc., Santa Clara, CA). The libraries were sequenced with a HiSeq 4000 instrument (Illumina, San Diego, CA) in a fraction of a HiSeq 4000 PE Cluster kit sequencing flow cell lane, following the manufacturer's protocol for dual indexing. Image analysis, base calling and quality scoring of the run were processed using the Real Time Analysis (RTA 2.7.7) tool and subsequently FASTQ sequence files were generated.

Bioinformatic analyses of gene expression

Sequencing quality was evaluated with the FastQC software v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptors were automatically detected and removed by using the TrimGalore 0.5.0 tool

(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), and we also trimmed reads shorter than 30 bp or those with more than 5 ambiguous bases (N). We excised 15 bp from both ends of each read because sequencing errors are more frequent in these regions [13, 14]. Clean reads were aligned to the goat reference genome ARS1 [15] with HISAT2 [16] by following the pipeline reported in [17]. The counts of unambiguously mapped reads of "protein-coding" features annotated in the general feature format (GFF) file were summarized by using the featureCounts tool [18]. Differential expression analyses were subsequently carried out by using DESeq2 software [19]. Correction for multiple testing was performed with the false discovery rate (FDR) procedure reported by Benjamini and Hochberg [20]. We considered that differential expression across two time points as relevant when two conditions were met: an absolute value of $\log_2$ fold change ($\log_2$FC) > 1.5 and a *q*-value ≤ 0.05. Moreover, we analyzed the functional enrichment of DE genes by employing the DAVID Bioinformatics Resources 6.8 database [21, 22]. This analysis was based on human and goat background gene sets, and statistical significance was set to a *q*-value ≤ 0.05.

Performance of a genome-wide association analysis for dairy traits

Phenotype recording

The population sampled in the current work comprised 1,023 Murciano-Granadina goats raised in 15 farms affiliated to the National Association of Murciano-Granadina Goat Breeders (CAPRIGRAN). All farms selected for this study were connected by artificial insemination. Raw records of phenotypic traits were routinely collected by CAPRIGRAN. Phenotypes under study included milk yield at 210 d (MY210), somatic cell count (SCC), fat percentage (FP), protein percentage (PP), lactose percentage (LP), dry matter percentage (DMP) and length of lactation (LOL). Phenotypes were normalized to a standard lactation of 210 d with the exception of LOL, which was not standardized. By

filtering out individuals without complete phenotypic records, 822 goats remained for GWAS analyses.

Genotyping with the goat SNP50 BeadChip

Blood samples were collected in EDTA K3 coated vacuum tubes and stored at −20 °C before processing. Genomic DNA was isolated by using a modified salting-out procedure [23]. Briefly, 3 mL of whole blood were centrifuged at a speed of 2,000 × *g* in the presence of 4 volumes of Red Cell Lysis Solution (Tris-HCl 10 mmol/L, pH = 6.5; EDTA 2 mmol/L; Tween 20 1%). The resulting white cell pellet was lysed with 3 mL lysis buffer (Tris-HCl 200 mmol/L, pH = 8, EDTA 30 mmol/L, SDS 1%; NaCl 250 mmol/L) and proteins were degraded by using 100 µL of proteinase K (20 mg/mL). After a 3-h incubation step at 55 °C, the lysate was chilled and 1 mL of ammonium acetate 10 mol/L was added to the lysate. After 10 min of centrifugation at 2,000 × *g*, the supernatant (~4 mL) was transferred to a new tube containing 3 mL of isopropanol 96%. Subsequently, samples were centrifuged at 2,000 × *g* for 3 min. Isopropanol was removed and the DNA pellet was washed with 3 mL of ethanol 70%. After a centrifugation step at 2,000 × *g* for 1 min, the DNA pellet was dried at room temperature and eluted with 1 mL of TE buffer (Tris-HCl 10 mmol/L, EDTA 1 mmol/L, pH = 8).

All goats were typed with the Goat SNP50 BeadChip (Illumina, USA) [24] according to the instructions of the manufacturer. Markers mapping to sex chromosomes, with calling rates < 90%, or with minor allele frequencies (MAF) < 0.01, or that deviated significantly from the Hardy-Weinberg expectation ($P \leq 1 \times 10^{-6}$) were filtered out. Individuals with calling rates < 90% were also excluded. By integrating available phenotypic records, 48,722 SNPs and 822 goats passed the filtering criteria.

Population structure and statistical analyses

We investigated population structure through the principal component analysis (PCA) approach implemented in the smartPCA program of the EIGENSOFT package (version 6.1.4) [25]. The proportion of the variance explained by each significant ($P < 0.05$) principal component was computed with the twstats program [26]. Association analyses were performed with the Genome-wide Efficient Mixed-Model Association (GEMMA, version 0.98) package [27] by fitting the following linear mixed model:

$$Y = W\alpha + x\beta + u + \varepsilon$$

where Y represents the vector of phenotypic values of the first lactation of 822 Murciano-Granadina goats; W is a matrix with a column of 1 s and the fixed effects, i.e. farm (15 levels), year of birth (10 levels) and litter size (5 levels); α is a c-vector of the corresponding coefficients including the intercept; x is a n-vector of marker genotypes in each individual; β is the effect size of the marker (allele substitution effect); u is a n-vector of random effects with a n-dimensional multivariate normal distribution $(0, \lambda\tau^{-1}K)$, being τ^{-1} the variance of the residual error, λ the ratio between the two variance components and K a $n \times n$ relatedness matrix derived from the 48,722 autosomal SNPs genotypes; and ε is a vector of errors. In this study, the GEMMA package performs likelihood ratio tests for each SNP by contrasting the alternative hypothesis (H_1: $\beta \neq 0$) against the null hypothesis (H_0: $\beta = 0$). Moreover, population structure is corrected by considering the relatedness matrix, which is built by taking into account all genome-wide SNPs as a random effect. After carrying out a correction for multiple testing based on a FDR approach [20], statistical significance was set to a q-value ≤ 0.05.

We retrieved a list of protein-coding genes that mapped within the genomic boundaries (± maximum distance of linkage disequilibrium decay, i.e. 988 kb) of leading SNPs (i.e. the SNP showing the most significant association with a

given trait) with the BEDTools v2.25.0 package [28]. The amount of linkage disequilibrium (LD) between adjacent SNPs was measured as the square of the correlation coefficient (r^2) by using the "--r^2" instruction implemented in PLINK v1.9 [29]. The objective of this analysis was to check whether protein-coding genes within or near quantitative traits loci (QTLs) are differentially expressed across lactation.

Results

An analysis of the mammary gene expression patterns across goat lactation

Differential expression analysis

We have individually sequenced 21 RNA samples representing three lactation time points (T1, T2 and T3, see **Methods**). This experiment generated approximately 120 gigabases of raw data, i.e. an average of 65 million reads were obtained for each sample. The overall alignment rate obtained with HISAT2 [16, 17] was above 92%. The uniquely mapped reads were summarized by using the featureCounts tool [18]. To reduce the influence of transcriptional noise, we removed the features with a number of raw counts below 10 in all samples. Principal component analysis (**Figure 1a**) based on the expression profiles of each one of the 21 samples showed a clear separation between T3 (dry period) and T1/T2 (lactation) samples. Indeed, the first component explained 73% of the total variance. The only exception was sample T3-22, which clustered with T1/T2 samples (**Figure 1a, Additional file 2: Figure S1**). Our interpretation is that this sample was retrieved from a goat that was not successfully dried off, so we decided to remove it from the data set. Although T1 and T2 samples

represented two different time points of lactation (**Figure 1a, Additional file 2: Figure S1**), they clustered tightly.

A total of 16,768 genes were found to be expressed in at least one of the 20 samples corresponding to the three lactation time points (T1, T2 and T3, see **Methods**). By establishing as a threshold of significance a *q*-value ≤ 0.05 and an absolute $\log_2$FC > 1.5, we found 42 (T1 vs. T2), 1377 (T1 vs. T3) and 1,039 (T2 vs. T3) DE genes (**Figures 1b-d, Additional file 3: Tables S2-S4**). The total set of 1,654 DE genes allowed us to differentiate T3 samples from the T1 and T2 samples (**Figure 2, Additional file 4: Figure S2**). Moreover, there was a comparable number of upregulated and downregulated genes in the pairwise T1 vs. T2 (22 upregulated and 20 downregulated) and T1 vs. T3 (649 upregulated and 728 downregulated) comparisons, while in T2 vs. T3 the number of downregulated genes (695) exceeded that of upregulated genes (344) (**Figures 1b-d, Additional file 3: Tables S2-S4**). In summary, our data evidenced that once lactation ceased, a large number of genes were downregulated (**Figures 1c-d, Additional file 3: Tables S3 and S4**). As expected, genes encoding the main milk protein constituents such as casein α_{S1} (*CSN1S1*), casein α_{S2} (*CSN1S2*), casein β (*CSN2*), casein κ (*CSN3*), lactalbumin α (*LALBA*) and progestagen-associated endometrial protein (*PAEP*) were strongly downregulated at T3 (**Table 1**). The insulin receptor 1 (*IRS1*) gene, a master regulator of carbohydrate, lipid and protein metabolism, also decreased in expression at T3 (**Table 1**).

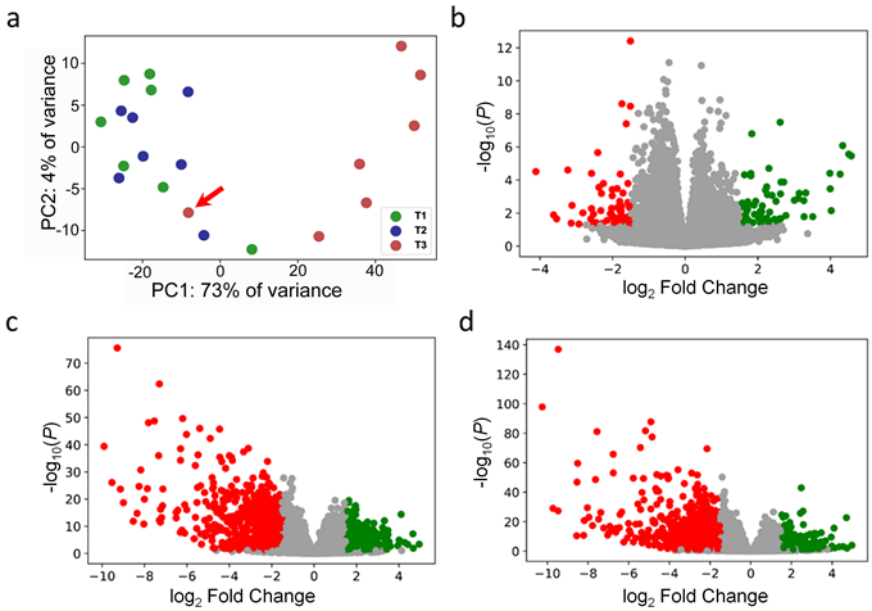

Figure 1. (**a**) Principal component analysis (PCA) of mammary samples on the basis of read counts of "protein-coding" features annotated in the general feature format (GFF) file. These samples were obtained 78 d (T1, early lactation), 216 d (late lactation, T2) and 285 d (T3, dry period) after parturition. The red arrow indicates the sample T3-22, which clusters with T1 and T2 samples probably due to an unsuccessful dry-off (**Additional file 2: Figure S1**). **b-d** Volcano plots displaying differentially expressed genes in the pairwise comparisons T1 vs. T2 (**b**), T1 vs. T3 (**c**) and T2 vs. T3 (**d**). The red and green dots denote significantly downregulated and upregulated genes, respectively

In T3, we also observed a marked downregulation of genes involved in lipid metabolic processes (**Table 1**), including: 1) Fatty acid synthesis, e.g. acetyl-CoA carboxylase α (*ACACA*) and fatty acid synthase (*FASN*); 2) Triglyceride synthesis, e.g. glycerol-3-phosphate acyltransferase, mitochondrial (*GPAM*), 1-acylglycerol-3-phosphate O-acyltransferase 1 (*AGPAT1*) and 4 (*AGPAT4*), glycerol-3-phosphate acyltransferase 2, mitochondrial (*GPAT2*) and 4 (*GPAT4*); 3) Cholesterol synthesis, e.g. 7-dehydrocholesterol reductase (*DHCR7*), 24-dehydrocholesterol reductase (*DHCR24*), lanosterol synthase (*LSS*), and

methylsterol monooxygenase 1 (*MSMO1*); 4) Sphingolipid synthesis, e.g. sphingolipid biosynthesis regulator 3 (*ORMDL3*), oxysterol binding protein like 10 (*OSBPL10*) and 1A (*OSBPL1A*), serine palmitoyltransferase long chain base subunit 2 (*SPTLC2*) and 3 (*SPTLC3*); 5) Acetate synthesis and fatty acid activation, e.g. acetyl-coenzyme A synthetase 2 (*ACSS2*) and acyl-CoA synthetase long chain family member 1 (*ACSL1*); 6) Fatty acid desaturation, e.g. stearoyl-CoA desaturase (*SCD*) and fatty acid desaturase 1 (*FADS1*); 7) Fatty acid absorption and transportation, e.g. CD36 molecule (*CD36*), low-density lipoprotein receptor (*LDLR*), fatty acid binding protein 3 (*FABP3*) and apolipoprotein A5 (*APOA5*); 8) Formation of milk fat globules, e.g. butyrophilin subfamily 1 member A1 (*BTN1A1*), perilipin 2 (*PLIN2*), RAB18, member RAS oncogene family (*RAB18*), and milk fat globule-EGF factor 8 protein (*MFGE8*); 9) Lipolysis, e.g. lipoprotein lipase (*LPL*), lipase G, endothelial type (*LIPG*), and pancreatic lipase related protein 2 (*PNLIPRP2*); 10) Transcriptional regulation of lipid metabolism, e.g. peroxisome proliferator activated receptor α (*PPARA*), estrogen receptor 2 (*ESR2*), leptin (*LEP*), and insulin-induced gene 1 (*INSIG1*).

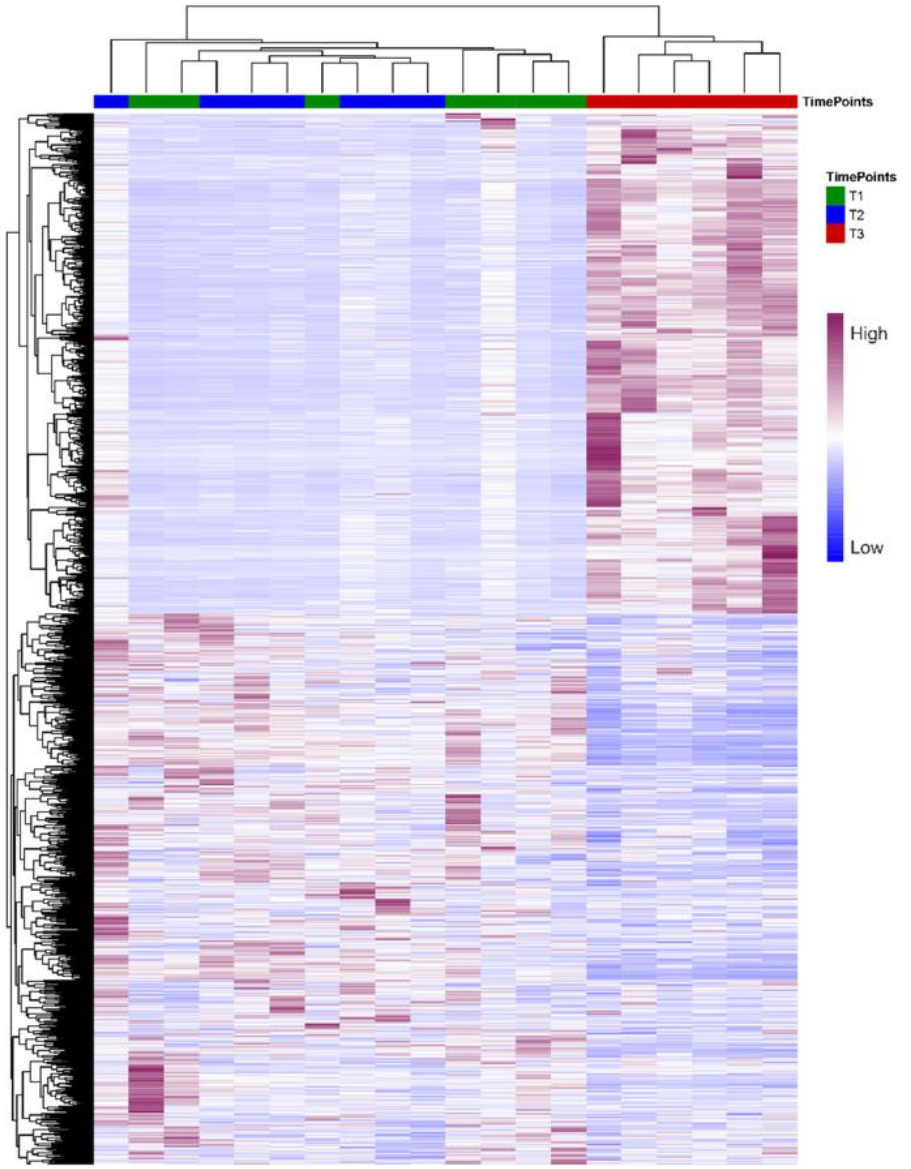

Figure 2. Heatmap of read counts of 1,654 differentially expressed genes identified in the three available comparisons (T1 vs. T2, T1 vs. T3 or T2 vs. T3). Samples were clustered by their read counts. The color scale varying from blue to purple depicts the number of read counts of differentially expressed genes which range from low to high, respectively.

The ceasing of lactation (T3) also involved an important decrease in the gene expression of solute carrier genes (**Table 1**) involved in the transportation of: 1) Carbohydrates, e.g. solute carrier family 2 member 1 (*SLC2A1*) and solute carrier family 35 member C1 (*SLC35C1*); 2) Amino acids, e.g. solute carrier family 1 member 1 (*SLC1A1*), solute carrier family 1 member 5 (*SLC1A5*), solute carrier family 7 member 14 (*SLC7A14*) and solute carrier family 36 member 2 (*SLC36A2*); and 3) Minerals, e.g. zinc (solute carrier family 30 member 4, *SLC30A4*), copper (solute carrier family 31 member 1, *SLC31A1*), divalent metals (solute carrier family 39 member 14, *SLC39A14*), to mention a few. With regard to the absorption of calcium, one of the main minerals present in milk, we observed a reduction in the expression of transient receptor potential cation channel subfamily V members 1 (*TRPV1*), while there was an upregulated expression of transient receptor potential cation channel subfamily V members 5 (*TRPV5*) and 6 (*TRPV6*). The gene expression of parathyroid hormone-like hormone (*PTHLH*) was reduced in the mammary gland at T3 but, at the same time, an increased expression of fibroblast growth factor 23 (*FGF23*) was also detected.

In general, genes involved in apoptosis displayed an upregulated expression in the mammary gland of goats at T3 (**Table 1**). Examples of these genes are the insulin-like growth factor binding protein 5 (*IGFBP5*), leukemia inhibitory factor (*LIF*), suppressor of cytokine signaling 3 (*SOCS3*), BCL2 like 14 (*BCL2L14*), oncostatin M (*OSM*), oncostatin M receptor (*OSMR*), Fos proto-oncogene, AP-1 transcription factor subunit (*FOS*) and JunB proto-oncogene, AP-1 transcription factor subunit (*JUNB*) as well as several genes belonging to the TNF superfamily such as tumor necrosis factor (*TNF*) and TNF receptor superfamily members 8 (*TNFSF8*), 13 (*TNFSF13*), 18 (*TNFRSF18*) and 6b (*TNFRSF6B*), and TNF-α induced protein 6 (*TNFAIP6*). In contrast, well known survival factors such as leukocyte receptor tyrosine kinase (*LTK*) and Wnt family member 5A (*WNT5A*) displayed a reduction in their expression at T3. Moreover, several genes belonging to the family of A disintegrin and metalloproteinase with

thrombospondin motifs (ADAMTS), such as *ADAMTS4*, *ADAMTS7*, *ADAMTS16* and *ADAMTS17*, which are involved in morphogenesis and tissue remodeling [30] increased in expression at T3.

Table 1. List of differentially expressed genes mentioned in the main text

Main Function	Gene symbol	T1 vs. T3		T2 vs. T3	
		log_2FC	***q*-value**	**log_2FC**	***q*-value**
Milk protein composition	*CSN1S1*	-8.17	1.31E-28	-8.03	8.28E-28
	CSN1S2	-8.00	1.83E-18	-8.18	1.50E-19
	CSN2	-7.24	5.06E-15	-7.23	9.68E-15
	CSN3	-5.16	2.76E-11	-5.12	4.67E-11
	LALBA	-9.53	3.72E-24	-9.47	8.10E-26
	PAEP	-4.67	8.16E-20	-4.41	1.61E-22
Regulator of carbohydrate, lipid and protein metabolism	*IRS1*	-1.85	1.60E-11	-	-
Fatty acid synthesis	*ACACA*	-3.10	3.26E-36	-2.69	1.31E-39
	FASN	-4.46	4.34E-43	-4.06	5.25E-47
Triglyceride synthesis	*GPAM*	-5.35	5.03E-23	-5.47	2.35E-17
	AGPAT1	-2.18	1.66E-23	-1.96	7.68E-27
	AGPAT4	-1.94	2.12E-02	-2.08	8.48E-03
	GPAT2	-	-	-2.80	3.49E-04
	GPAT4	-2.70	8.24E-19	-2.44	6.98E-35
Cholesterol synthesis	*DHCR7*	-2.53	1.57E-20	-2.91	1.40E-50
	DHCR24	-4.00	1.15E-33	-4.14	3.81E-49
	LSS	-1.82	3.88E-19	-2.30	1.60E-40
	MSMO1	-2.47	1.01E-22	-2.93	3.50E-31
Sphingolipid synthesis	*ORMDL3*	-2.24	3.76E-19	-2.16	3.44E-30
	OSBPL10	-1.63	6.25E-06	-	-
	OSBPL1A	-1.62	4.00E-13	-	-
	SPTLC2	-1.73	1.56E-18	-	-
	SPTLC3	-1.54	2.92E-06	-	-

Acetate synthesis and fatty acid activation	*ACSS2*	-2.26	5.04E-26	-2.53	1.57E-31
	ACSL1	-1.98	7.21E-09	-1.97	7.05E-13
Fatty acid desaturation	*SCD*	-6.30	4.75E-32	-6.43	2.17E-23
	FADS1	-2.23	5.87E-26	-2.41	4.70E-25
Fatty acid absorption and transportation	*CD36*	-2.64	5.12E-16	-2.59	6.81E-20
	LDLR	-1.72	7.96E-08	-2.37	3.19E-20
	FABP3	-5.00	1.93E-09	-5.37	5.60E-12
	APOA5	-5.12	3.17E-07	-5.08	8.00E-08
Milk fat globules	*BTN1A1*	-6.43	1.02E-14	-6.46	5.80E-15
	PLIN2	-2.00	6.30E-11	-1.96	7.69E-15
	RAB18	-2.65	1.05E-16	-2.32	4.00E-15
	MFGE8	-3.84	1.67E-20	-4.09	5.24E-49
Lipolysis	*LPL*	-5.60	3.33E-30	-5.78	2.29E-47
	LIPG	-4.10	7.56E-11	-4.25	1.14E-11
	PNLIPRP2	-	-	-2.46	1.12E-02
Transcriptional regulation of lipid metabolism	*PPARA*	-1.63	1.78E-18	-	-
	ESR2	-1.62	8.77E-08	-	-
	LEP	-4.01	2.11E-04	-5.39	2.39E-08
	INSIG1	-4.35	1.29E-31	-5.18	9.72E-79
Transportation of carbohydrates	*SLC2A1*	-2.18	2.76E-19	-2.00	2.34E-17
	SLC35C1	-2.94	1.59E-22	-2.85	8.11E-23
Transportation of amino acids	*SLC1A1*	-2.14	4.08E-07	-2.01	3.14E-06
	SLC1A5	-2.29	3.79E-09	-2.28	5.58E-14
	SLC7A14	-1.83	4.38E-03	-1.84	4.00E-03
	SLC36A2	-1.53	2.19E-03	-	-
Transportation of minerals	*SLC30A4*	-2.70	1.37E-11	-1.95	4.49E-12
	SLC31A1	-1.69	1.96E-10	-1.53	6.50E-09
	SLC39A14	-1.72	1.31E-09	-1.68	7.69E-11
Absorption of calcium	*TRPV1*	-3.43	4.50E-06	-2.83	2.64E-07
	TRPV5	4.66	7.54E-07	3.68	1.15E-06
	TRPV6	3.25	7.87E-06	3.25	8.06E-11

	PTHLH	-5.17	1.04E-20	-5.41	1.28E-26
	FGF23	3.48	1.21E-06	2.33	5.21E-03
Apoptosis	*IGFBP5*	-	-	-1.76	2.01E-05
	LIF	2.58	3.05E-05	1.86	4.80E-04
	SOCS3	2.65	2.64E-05	2.49	1.66E-06
	BCL2L14	1.53	1.55E-06	1.68	4.06E-12
	OSM	1.90	1.46E-04	-	-
	OSMR	1.64	6.38E-04	1.56	6.14E-09
	FOS	1.67	5.82E-03	1.72	2.61E-04
	JUNB	2.06	4.59E-06	1.80	4.63E-11
	TNF	1.78	8.90E-06	-	-
	TNFSF8	1.66	8.87E-12	-	-
	TNFSF13	-	-	-2.48	3.33E-17
	TNFRSF18	2.10	6.21E-05	1.59	1.04E-04
	TNFRSF6B	1.64	6.44E-04	-	-
	TNFAIP6	1.81	1.41E-05	-	-
	LTK	-2.45	2.59E-26	-2.31	2.79E-19
	WNT5A	-2.50	5.73E-20	-2.28	5.87E-26
Morphogenesis and tissue remodeling	*ADAMTS4*	2.46	5.84E-05	1.61	2.64E-03
	ADAMTS7	1.59	2.43E-06	-	-
	ADAMTS16	1.69	2.46E-03	-	-
	ADAMTS17	-1.54	1.18E-09	-1.64	2.02E-09
Immunity	*MUC1*	-3.11	2.59E-05	-2.78	1.21E-04
	MUC4	-2.10	7.80E-05	-1.81	1.03E-05
	MUC20	-3.01	4.49E-15	-2.78	1.09E-13
	ABCA3	-2.61	7.95E-21	-1.94	3.98E-20
	SFTPD	-5.47	6.59E-34	-5.35	6.57E-38
	BPIFA1	-4.41	2.67E-13	-3.86	5.54E-08
	BPIFA2	-8.53	5.70E-11	-6.61	1.56E-14
	BPIFA3	-5.11	1.34E-17	-4.08	6.36E-09
	BPIFB1	-4.32	3.14E-21	-4.28	3.16E-20

	BPIFB4	-3.93	8.81E-03	-	-
	CLDN6	3.09	1.68E-03	-	-
	CLDND2	1.68	1.57E-08	1.72	1.61E-22
Cytokines and/or their receptors	*IL5*	1.76	8.96E-03	-	-
	IL15RA	1.90	3.49E-07	-	-
	IL22RA2	1.99	9.41E-03	-	-
Defensins	*DEFB116*	2.75	2.34E-02	2.73	9.28E-03
	DEFB126	3.40	2.70E-03	3.55	2.21E-04
Chemokines	*CXCR4*	1.67	1.67E-05	-	-
Complement cascade	*C1QA*	1.60	1.77E-07	-	-
	C1S	1.69	8.27E-06	-	-
	C1R	1.76	4.57E-05	-	-
	C6	2.25	1.53E-07	1.83	2.36E-06
	C7	1.95	1.04E-03	-	-
	CTSL	1.78	8.36E-03	-	-

The dash symbol indicates the absence of a significant differential expression; log_2FC: log_2 of the fold-change in expression. A negative log_2FC value indicates that mRNA expression is downregulated in T3.

With regard to genes involved in immunity, the dynamics of their expression profiles was quite heterogeneous (**Table 1**). Genes with key roles in mucosal immunity, e.g. mucin 1 (*MUC1*), 4 (*MUC4*) and 20 (*MUC20*), ATP binding cassette subfamily A member 3 (*ABCA3*), and surfactant protein D (*SFTPD*), were downregulated at T3. In this time point, we also detected a decreased expression of several genes, e.g. the BPI fold containing family A member 1 (*BPIFA1*), member 2 (*BPIFA2*) and member 3 (*BPIFA3*), and the BPI fold containing family B member 1 (*BPIFB1*) and member 4 (*BPIFB4*), which have antimicrobial, surfactant and immunomodulatory properties, thus preventing the formation of bacterial biofilms [31]. In contrast, tight junction proteins claudin

6 (*CLDN6*) and D2 (*CLDND2*), which determine the permeability of the paracellular barrier [32], were highly upregulated at T3.

Finally, we detected an upregulation of a broad variety of immune response genes at T3 (**Table 1**), including: 1) Cytokines (and/or their receptors), e.g. interleukin 5 (*IL5*), interleukin 15 receptor subunit α (*IL15RA*) and interleukin 22 receptor subunit α2 (*IL22RA2*); 2) Defensins, e.g. defensin β116 (*DEFB116*) and β126 (*DEFB126*); 3) Chemokines, e.g. C-X-C motif chemokine receptor 4 (*CXCR4*); and 4) Genes participating in the complement cascade, e.g. complement C1q A chain (*C1QA*), complement C1s (*C1S*), complement C1r (*C1R*), complement 6 (*C6*), complement 7 (*C7*) and cathepsin L (*CTSL*).

Functional enrichment of differentially expressed genes

Due to the incomplete annotation of goat genes, the functional enrichment analysis of the 1,654 DE genes was based on both human and goat background gene sets retrieved from the DAVID database [21, 22]. As a result, we identified 10 pathways that were significantly enriched based on the human background gene set (q-value $\leq$ 0.05, **Additional file 5: Table S5**), and 11 significant pathways based on the goat background gene set (q-value $\leq$ 0.05, **Additional file 5: Table S6**). Six pathways were consistently detected in both analyses, i.e. PPAR signaling, metabolic pathways, steroid biosynthesis, complement and coagulation cascades, biosynthesis of antibiotics and adipocytokine signaling (**Table 2**). Moreover, the gene ontology (GO) analysis based on human background genes allowed us to detect 45 significant terms, while no term was identified when the goat background genes were used (**Additional file 5: Tables S5 and S6**).

Table 2. Enriched pathways in the set of 1,654 differentially expressed genes (T1-T2, T1-T3 and T2-T3)

Name	Human background gene set				Goat background gene set			
	Number	***P* value**	**Fold Enrichment**	***q*-value**	**Number**	***P* value**	**Fold Enrichment**	***q*-value**
PPAR signaling pathway	18	2.01E-06	3.86	2.65E-05	22	8.86E-08	3.84	1.16E-06
Steroid biosynthesis	9	3.06E-05	6.46	4.03E-04	10	2.58E-05	5.63	3.38E-04
Complement and coagulation cascades	16	5.85E-05	3.33	7.70E-04	18	2.78E-05	3.19	3.65E-04
Metabolic pathways	116	1.75E-04	1.37	2.30E-03	141	6.68E-06	1.41	8.78E-05
Biosynthesis of antibiotics	28	1.49E-03	1.90	1.95E-02	30	2.58E-03	1.78	3.34E-02
Adipocytokine signaling pathway	13	2.89E-03	2.67	3.73E-02	14	3.46E-03	2.48	4.45E-02

These are the pathways that were consistently detected in the analyses based on human and goat background gene sets

Identification of genomic regions associated with dairy traits

Descriptive statistics of seven dairy traits recorded in Murciano-Granadina goats are shown in **Additional file 6: Table S7**. The average values of milk fat percentage, protein percentage and milk yield normalized to 210 d were 5.20% ± 0.85%, 3.56% ± 0.41% and 387.65 ± 134.79 kg, respectively. Moreover, all traits showed a normal distribution with the exception of the somatic cell count (SCC), which was logarithmically transformed to achieve normality (**Additional file 7: Figure S3**). The analysis of the Murciano-Granadina individuals by PCA clustering based on the genotypes of the 48,722 available markers did not show any sign of population stratification (**Additional file 8: Figure S4**).

By performing association analyses between SNP genotypes and dairy traits recorded in 822 Murciano-Granadina goats, we identified 24 quantitative trait loci (QTLs) that reached the threshold of significance (q-value ≤ 0.05, **Table 3**) either at the genome-wide or chromosome-wide levels. Quantitative trait locus 6 (QTL6) on chromosome 6 was highly associated with protein percentage at the genome-wide level of significance (78.90-93.48 Mb, q-value = 1.54×10^{-06}, **Figure 3, Table 3**), and also with dry matter (84.67-86.86 Mb, q-value = 2.66×10^{-02}) and fat percentages (86.86 Mb, q-value = 1.36×10^{-02}) at the chromosome-wide level of significance. In addition, we found genome-wide significant associations for lactose percentage on chromosome 2 (QTL1, 130.72-131.01 Mb, q-value = 7.26×10^{-03}, **Figure 4a**), as well as for protein and dry matter percentages on chromosome 17 (QTL17, 11.20 Mb, **Figures 3a and 4b**). At the chromosome-wide level, we found 21 significant associations (**Table 3**) but only two of them were supported by more than 2 SNPs (QTL9 for somatic cell count and QTL24 for lactose percentage, **Table 3**).

According to data presented in **Additional file 9: Figure S5**, the maximum distance at which r^2 decays to its minimum value is 988 kb. Based on this, we retrieved 490 protein-coding genes mapping to ±988 kb of the leading SNP

corresponding to each QTL. This list of genes was compared with the list of genes DE across lactation time points. By doing so, we found 39 genes mapping to 14 QTLs that are also DE (**Table 4**). For instance, the QTL6 region, which shows significant associations with protein, fat and dry matter percentages, contains the casein genes, which are downregulated in T3 (**Tables 3** and **4**).

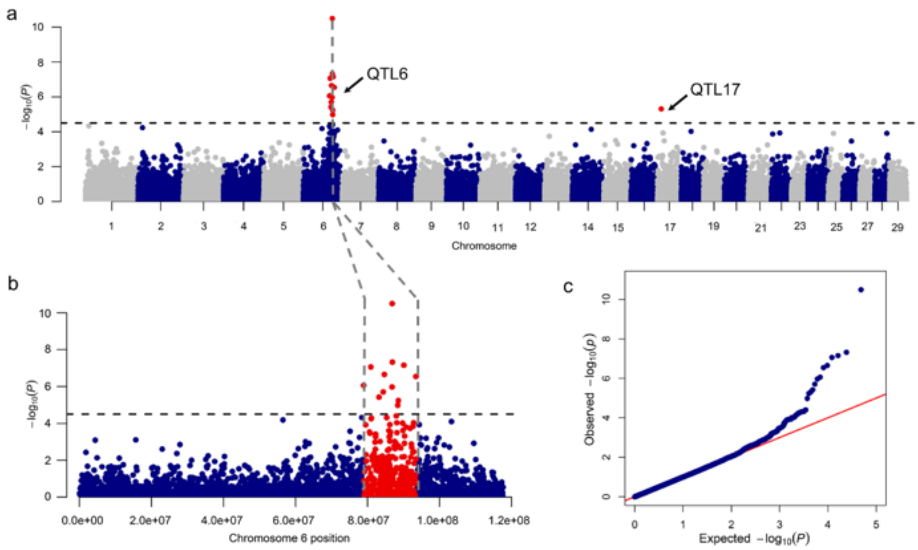

Figure 3. (**a**) Manhattan plot depicting the genome-wide association between milk protein percentage and a genomic region on chromosome 6 containing the casein genes (QTL6). Negative $\log_{10}P$ values of the associations between SNPs and phenotypes are plotted against the genomic location of each SNP marker. Markers on different chromosomes are denoted by different colors. The dashed line represents the genome-wide threshold of significance (q-value ≤ 0.05). (**b**) A detailed view of the chromosome 6 region associated with protein percentage. Significant SNPs within the QTL boundaries have been marked in red. (**c**) Quantile-quantile (QQ) plot of the data shown in the Manhattan plot.

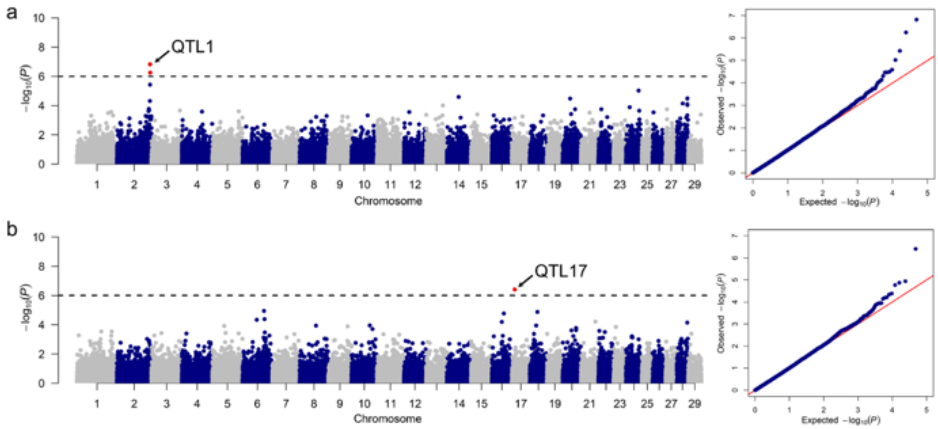

Figure 4. (**a**) Manhattan plot depicting the genome-wide significant associations between SNP markers and lactose percentage. The corresponding quantile-quantile (QQ) plot is shown at the right side of the Manhattan plot. (**b**) Manhattan plot depicting the genome-wide significant associations between SNP markers and dry matter percentage. The corresponding quantile-quantile (QQ) plot is shown at the right side of the Manhattan plot. Negative $\log_{10}P$ values of the associations between SNPs and phenotypes are plotted against the genomic location of each marker SNP. Markers on different chromosomes are denoted by different colors. The dashed lines represent the genome-wide threshold of significance (q-value ≤ 0.05).

Table 3. Quantitative trait loci (QTLs) associated with milk traits recorded in Murciano-Granadina goats

QTL	Chromosome	Leading SNP	Position (Mb)	#SNPs	MAF	Trait	β ± SE	*P* value	*q*-value
1	**2**	**rs268253425**	**130.72-131.01**	**2**	**0.20**	**LP**	**-0.09±0.02**	**1.50E-07**	**7.26E-03**
2	3	rs268258472	113.47	1	0.38	LOL	-14.20±3.21	1.04E-05	2.38E-02
3	6	rs268259784	4.40	1	0.35	PP	0.07±0.02	8.21E-04	4.09E-02
4	6	rs268251267	15.64	1	0.24	PP	0.08±0.03	7.85E-04	4.00E-02
5	6	rs268259390	56.51	1	0.02	DMP	1.08±0.26	4.56E-05	3.56E-02
						PP	0.29±0.07	6.52E-05	8.48E-03
6	**6**	**rs268290907**	**78.90-93.48**	**12**	**0.43**	**PP**	**-0.14±0.02**	**3.19E-11**	**1.54E-06**
	6	rs268268356	84.67-86.86	2	0.40	DMP	-0.35±0.08	1.14E-05	2.66E-02
	6	rs268290907	86.86	1	0.43	FP	-0.18±0.04	5.79E-06	1.36E-02
7	11	rs268250457	72.83	1	0.41	LOL	15.23±3.30	4.40E-06	9.13E-03
8	12	rs268256521	68.10	1	0.12	LOL	-20.35±4.76	2.07E-05	3.50E-02
9	13	rs268236131	53.62-54.38	2	0.32	SCC	0.22±0.05	1.92E-05	3.05E-02
11	14	rs268255959	46.10	1	0.09	LP	-0.10±0.02	2.56E-05	4.77E-02
10	14	rs268282962	56.92	1	0.01	FP	0.78±0.18	1.78E-05	3.32E-02
12	15	rs268235117	34.69	1	0.06	MY210	-61.98±13.53	4.95E-06	7.91E-03
13	15	rs268290053	35.51	1	0.07	FP	0.35±0.08	4.82E-06	7.69E-03
14	15	rs268266747	63.95	1	0.23	SCC	0.23±0.05	1.63E-05	2.60E-02

15	16	rs268236985	39.59	1	0.04	DMP	0.81±0.20	6.39E-05	4.93E-02
16	16	rs268253363	47.67	1	0.21	DMP	0.42±0.10	1.70E-05	2.62E-02
17	**17**	**rs268238952**	**11.20**	**1**	**0.06**	**PP**	**0.22±0.05**	**4.84E-06**	**2.13E-02**
						DMP	**0.88±0.17**	**3.89E-07**	**1.88E-02**
18	18	rs268278435	29.64	1	0.04	DMP	0.84±0.19	1.33E-05	1.70E-02
19	20	rs268277231	29.45	1	0.08	LP	-0.11±0.03	3.35E-05	4.84E-02
20	22	rs268253724	25.30	1	0.36	FP	-0.18±0.04	4.14E-05	4.68E-02
21	23	rs268243170	8.15	1	0.45	MY210	29.89±6.76	1.03E-05	1.01E-02
22	24	rs268240589	49.71	1	0.32	LP	0.07±0.02	9.51E-06	1.23E-02
23	28	rs268240830	23.02	1	0.42	LP	-0.06±0.01	7.16E-05	1.66E-02
24	28	rs268246445	41.15-41.42	4	0.51	LP	-0.06±0.01	3.21E-05	1.51E-02

QTL: Quantitative trait locus; Genome-wide significant associations are indicated in bold; leading SNP: a SNP showing the most significant association with a given trait; #SNPs; number of SNPs; MAF: minor allele frequency; PP: protein percentage, FP: fat percentage, LP: lactose percentage, DMP: dry matter percentage, SCC: somatic cell count, LOL: length of lactation, MY210: milk yield normalized to 210 days; β and SE denote the effect size of the marker (allele substitution effect) and its standard error, respectively.

Table 4. List of genes that are differentially expressed and that co-localize with dairy QTLs

QTL	Leading SNP	Chrom-osome	Start	End	Gene symbol	T1 vs. T2		T1 vs. T3		T2 vs. T3	
						log_2FC	*q*-value	log_2FC	*q*-value	log_2FC	*q*-value
1	rs268253425	2	130227819	130232923	*MSTN*	-	-	-2.17	3.30E-05	-1.66	1.14E-05
2	rs268258472	3	113070996	113098602	*SH2D1B*	-	-	1.53	3.42E-04	-	-
		3	113497530	113528836	*HSD17B7*	-	-	-2.18	1.87E-09	-2.08	5.91E-13
		3	113589998	113598576	*CCDC190*	-	-	2.15	3.31E-02	-	-
		3	113876670	113883784	*RGS4*	-	-	-3.45	1.10E-03	-3.25	1.73E-04
4	rs268251267	6	15922309	15965996	*CFI*	-	-	2.20	4.46E-04	2.33	1.08E-05
6	rs268268356, rs268290907	6	84458894	84563377	*LOC102185449*	-	-	-	-	-2.04	2.62E-02
		6	84667577	84721849	*LOC102186288*	-	-	-4.57	7.95E-08	-3.96	5.03E-10
		6	85137434	85152951	*LOC102172432*	-	-	-	-	-3.91	3.79E-03
		6	85878003	85901026	*LOC102169846*	-	-	-6.10	4.46E-07	-6.24	6.03E-06
		6	85978463	85995270	*CSN1S1*	-	-	-8.17	1.31E-28	-8.03	8.28E-28
		6	86006250	86015321	*CSN2*	-	-	-7.24	5.06E-15	-7.23	9.68E-15
		6	86076845	86093539	*CSN1S2*	-	-	-8.00	1.83E-18	-8.18	1.50E-19
		6	86093738	86115903	*LOC102178810*	-	-	-6.50	4.36E-12	-5.91	2.77E-11
		6	86197263	86211376	*CSN3*	-	-	-5.16	2.76E-11	-5.12	4.67E-11
		6	86427932	86443025	*AMTN*	1.94	2.72E-02	-2.59	4.31E-06	-4.56	4.61E-12

7	rs268250457	11	71885505	71909335	*GCKR*	-	-	-1.96	5.40E-09	-	-
		11	72204731	72208863	*TCF23*	-	-	2.04	3.56E-03	1.56	2.26E-02
		11	72818025	72863960	*DRC1*	-	-	2.55	2.54E-07	1.71	6.22E-04
8	rs268256521	12	67360683	67392416	*EBPL*	-	-	-2.05	6.24E-13	-2.32	7.71E-12
		12	68130994	68150636	*CYSLTR2*	-	-	1.57	1.43E-06	-	-
9	rs268236131	13	53376787	53378322	*TNFRSF6B*	-	-	1.64	6.44E-04	-	-
		13	53475687	53483660	*EEF1A2*	-	-	2.07	2.39E-02	-	-
		13	53771553	53785372	*SLC17A9*	-	-	-	-	-1.94	5.05E-10
		13	54495442	54974847	*CDH4*	-	-	1.55	1.19E-08	-	-
11	rs268255959	14	45768357	46228322	*KCNB2*	-	-	-	-	-1.63	1.45E-02
		14	46685533	46688514	*MSC*	-	-	1.58	4.90E-07	-	-
13	rs268290053	15	34071717	34073359	*LOC102175876*	4.50	3.70E-04	-4.06	9.75E-04	-8.56	7.66E-10
		15	34118282	34119849	*HBBC*	-	-	-	-	-3.72	6.30E-04
14	rs268266747	15	63018327	63025893	*C15H11orf87*	-	-	4.01	7.24E-05	-	-
15	rs268236985	16	39165138	39176197	*TNFSF18*	-	-	1.71	1.23E-02	-	-
		16	40049635	40096376	*TNFRSF8*	-	-	1.53	5.39E-04	-	-
16	rs268253363	16	47058035	47175045	*AJAP1*	-	-	2.42	2.31E-09	1.74	3.18E-05
		16	47851111	47857057	*SMIM1*	-	-	-1.53	3.69E-14	-1.52	2.29E-15
		16	47858258	47874280	*CCDC27*	-	-	-1.82	4.42E-03	-	-

		16	48000776	48022801	*LOC102183348*	-	-	1.70	1.01E-03	-	-
21	rs268243170	23	7662629	7687922	*CD83*	-	-	1.58	1.78E-06	-	-
22	rs268240589	24	49397175	49422345	*LIPG*	-	-	-4.10	7.56E-11	-4.25	1.14E-11
		24	50316631	50490273	*MAPK4*	-	-	-3.40	3.72E-15	-2.71	2.17E-12

These DE genes were retrieved from an interval of ± 988 kb around leading SNPs (see **Methods**); Leading SNP: a SNP displaying the most significant association with a given trait; The dash symbol indicates the absence of a significant differential expression; $\log_2$FC: $\log_2$ of the fold change in expression.

Discussion

The expression profiles of the goat mammary gland in early and late lactation are similar

The number of DE genes in T1 vs. T2 was quite low (only 42 genes were DE), implying that the physiological and metabolic state of the mammary gland in these two time points is not remarkably different. In sheep milk, an analysis of differential expression revealed 22 (d 10 vs. 50), 20 (d 50 vs. 120), 277 (d 10 vs. 120), 135 (d 50 vs. 150) and 578 (d 10 vs. 150) DE genes [12]. The comparison that more closely resembles ours (d 50 vs. 150, 135 DE genes) highlighted a higher number of DE genes than us. Many biological and technical factors might have produced this discrepancy. For instance, we have used mammary tissue while Suárez-Vega et al. [12] employed milk somatic cells as a source of RNA. Moreover, the shape and duration of the lactation curve is different in sheep and goats. Despite these differences, a steady increase was observed in the expression of the carboxypeptidase X, M14 family member 2 (*CPXM2*) gene by Suárez-Vega et al. [12] and us. This gene might have an important role in mammary gland development and involution [12]. Moreover, Suárez-Vega et al. [12] and us observed an upregulation of the gene encoding γ-aminobutyric acid receptor subunit β_3 (*GABRB3*) at T2, a change that has also been observed in rat lactation [33]. We have also detected an upregulation of the arylsulfatase family member I (*ARSI*), inhibin subunit β A (*INHBA*) and tenascin R (*TNR*) genes in T2, which might be indicative of the tissue remodeling and progressive involution that the mammary gland experiences through the progression of lactation [34,35,36]. In T2, the upregulated ST8 α-N-Acetyl-Neuraminide α-2,8-Sialyltransferase 6 (*ST8SIA6*) and polypeptide N-acetylgalactosaminyltransferase 14 (*GALNT14*) genes respectively catalyze the formation of milk sialoglycoconjugates [37] and the O-glycosylation of mucins [38]. Finally, two molecules, i.e. adiponectin

(*ADIPOQ*) and hexokinase domain containing 1 (*HKDC1*), showed an increased and reduced expression in T2, respectively. These two molecules increase glucose utilization, reflecting the complex metabolic changes that the mammary gland undergoes throughout lactation.

Remarkable differences in the mammary mRNA expression profiles of lactating and dried goats

The mRNA expression of genes involved in milk protein synthesis is reduced during the dry period

In contrast with the previous comparison, the gene expression profiles of the goat mammary gland are quite different when T1/T2 samples are compared to T3 samples. At T3, we have observed a 5-8 fold downregulation of the genes encoding caseins (the major protein components of milk), while there was also a 9.5-fold reduction in the gene expression of the milk whey LALBA protein, which is essential for the synthesis of lactose [39]. Likewise, the *PAEP* gene, which encodes the major whey protein β-lactoglobulin, was down-regulated 4.5-fold at T3. Similar results have also been obtained in sheep and cattle [10,11,12, 40]. The reduction in milk protein synthesis can be attributed to the fact that this is an energetically demanding process that is rapidly inhibited in the absence of proper hormonal and nutritional stimulation [41]. In rodents, milk protein synthesis appears to be under the control of the signal transducer and activator of transcription 5 (STAT5) factor [42], but in close similarity to what has been observed in cattle [43], we did not observe a change in the expression of the *STAT5A* or *STAT5B* genes. Conversely, there was a 2-fold reduction of the E74 like ETS transcription factor 5 (*ELF5*), which was also detected in cattle by Bionaz and Loor [43]. The *ELF5* gene is regulated by *STAT5* and induced by insulin, which might be a major player in the activation of protein synthesis in the bovine mammary gland. Furthermore, and as discussed by Bionaz and Loor

[43], one of the factors that probably contributes to the strongly lowered milk protein synthesis during the dry period (T3) is the mRNA downregulation of major amino acid transporters, such as *SLC1A1*, *SLC1A5*, *SLC7A14* and *SLC36A2* [44, 45, 46].

The expression of genes involved in carbohydrate and lipid metabolism is downregulated in the dry period

Carbohydrate metabolism is also affected by the ceasing of lactation and, as mentioned before, LALBA, an enzyme necessary for the synthesis of lactose [39], the major sugar in milk, was downregulated at T3. We also observed a decrease in the mRNA expression of the *IRS1* gene, which mediates the effects of insulin [47]. Besides being fundamental for the absorption and storage of glucose [48], insulin also has important effects on the synthesis of milk proteins [49]. While the abundance of *SLC2A1* mRNA, one of the main glucose transporters, decreased at T3, we did not observe the same trend for *SLC2A4*, which is another major insulin-responsive glucose transporter [50]. These results agree with data presented by Komatsu et al. [51] who showed that *SLC2A1* has a more predominant role than *SLC2A4* in the glucose metabolism of the mammary gland during lactation.

The metabolic downregulation of the mammary gland that takes place during dry period has also a major impact on lipid metabolism. At T3, important transcriptional regulators were downregulated, such as *PPARA*, which is expressed in tissues with a high rate of fatty acid catabolism [52]; *ESR2*, which can inhibit ligand-mediated PPARG-transcriptional activity [53]; *LEP*, encoding a hormone that stimulates fatty acid oxidation; and *INSIG1*, encoding a protein that inhibits the proteolytic activation of sterol regulatory element-binding proteins (SREBPs). As mentioned by Bionaz and Loor [54], the case of *INSIG1* is quite counterintuitive because the mRNA expression of this gene is upregulated during lactation despite its inhibitory action on SREBPs and

lipogenesis. Our interpretation is that the increased expression of *INSIG1* during lactation arises from the increased need to fine tune the activity of SREBPs. The pathway enrichment analysis also detected many biochemical routes related to lipid metabolism, including the PPAR signaling pathway. Indeed, *PPARG* is a master regulator of adipocyte differentiation and lipid and glucose homeostasis [55], and according to Bionaz and Loor [54], *PPARG*, *PPARGC1A*, and *INSIG1*, rather than *SREBP1*, have a pivotal role in milk fat synthesis in cattle.

Alterations in the expression of genes modulating calcium homeostasis

In mammals, maternal calcium homeostasis is often challenged by the high calcium demand associated with the lactation process [56]. In the epithelial mammary cell, calcium is stored in and around the Golgi apparatus, and it is secreted into milk in close association with caseins [56]. At T3, the mammary glands of Murciano-Granadina goats displayed reduced mRNA levels of *PTHLH*, a molecule that favors calcium mobilization through bone resorption during lactation [57], and in parallel, an increased mRNA expression of the *FGF23* gene, which inhibits the synthesis of parathyroid hormone [58]. We also observed an upregulation of the *TRPV5* and *TRPV6* mRNAs, which favor calcium uptake in a broad array of tissues with predominance of kidney [59] and of intestine [60], respectively. From our perspective, the increased expression of these two channels at T3 is quite paradoxical because the abrogation of lactation implies a strong reduction of the calcium demand. A possible explanation is that the increased expression of *TRPV5* and *TRPV6* genes might contribute to replenish the exhausted mammary calcium pool, but this hypothesis needs to be verified.

Increased mammary expression of genes related with cell death and tissue remodeling during the dry period

During the dry period (T3), there is an extensive involution, apoptosis and remodeling of the mammary gland that involves the death and replacement of

senescent alveolar cells [61], transforming the udder from a milk factory to a quiescent organ [62]. Probably, one of the main cues that triggers this process is milk stasis [63]. The *FOS* and *JUNB* genes are upregulated in the mammary gland of Murciano-Granadina goats at T3, a finding that is relevant because they form part of the activator protein 1 (AP-1) dimeric transcription factor. This dimeric transcription factor is probably involved in the initiation or execution of apoptosis after mammary gland stops to milk [64]. We have also detected an increased expression of *OSM* and its receptor (*OSMR*), *LIF*, *BCL2L14*, *IGFBP5* and *SOCS3* mRNAs, a set of molecules which are known to promote the death of mammary epithelial cells and to facilitate the involution of the mammary gland [65, 66, 67, 68]. Furthermore, metalloproteinases with aggrecanase (*ADAMTS4*) and cartilage oligomeric matrix protein-cleaving (*ADAMTS7*) activities [30] were also upregulated at T3, probably because of the extensive tissue remodeling takes place during mammary involution [69]. Indeed, metalloproteinases play a fundamental role not only in the remodeling of the epithelial ductal and vascular networks, but also in the correct synchronization of parenchymal, stromal and extracellular matrix homeostasis.

Complex changes in the expression of genes with immunological functions

Bacterial infections are seven times more prevalent during the early dry period than during lactation [70], thus increasing the risk to suffer mastitis in the subsequent lactation. The mammary gland can be considered as a temporal mucosal organ [71], and in this regard we have detected a downregulation, at T3, of several molecules that are involved in the synthesis of mucins (*MUC1*, *MUC4* and *MUC20*) or surfactant (*ABCA3* and *SFTPD*) substances. These are two major components of the chemical barrier that protects mucosal surfaces against bacterial infection and biofilm formation. Mucins are large O-linked glycoproteins that form part of the gel-like extracellular matrix known as mucus [72]. This is considered to be the first line of defense against pathogens because it can trap bacteria and slow down the diffusion of large viruses and, moreover,

it holds immunoglobulin A and antimicrobial peptides that facilitate the elimination of pathogenic microorganisms [72]. Surfactant, which is mainly constituted by proteins and lipids, can also stimulate the clearance of microorganisms by increasing the membrane permeability of bacteria and by enhancing phagocytosis featured by cells of the innate immune system [73]. At T3, we have also detected a lowered mammary expression of the *BPIFA1*, *BPIFA2*, *BPIFA3*, *BPIFB1* and *BPIFB4* mRNAs. These molecules also play an essential role in mucosal immunity, being particularly well known the BPIFA1 protein because of its abundance in respiratory secretions, its inhibitory effect on bacterial growth and biofilm formation and its immunomodulatory properties [31]. Our results might suggest that mucous and surfactant substances that protect the mammary epithelium from infectious agents are synthesized at lower levels during the dry period, but in the absence of protein data we cannot draw firm conclusions about this matter.

In parallel, we have detected an increased mRNA expression, at T3, of several complement factors that are an important component of mucosal immunity by favoring immune bacteriolysis, neutralization of viruses, immune adherence, immunoconglutination and phagocytosis [74]. Two β-defensins (*DEFB116* and *DEFB126*) were also upregulated at T3. Defensins are cationic antimicrobial peptides that bind the negatively charged outer membranes of bacteria and kill them through a variety of mechanisms including pore formation, interference with cell wall synthesis, and prokaryotic membrane depolarization [75]. Interleukin 5, *CXCR4* and specific subunits of interleukins 15 and 22 receptors also showed an increase in mRNA expression at T3. Interleukin 5 is a survival factor for B-cells and eosinophils [76], while the chemokine receptor CXCR4 is a major contributor to B-cell homeostasis and humoral immunity [77]. With regard to interleukin 15, it is a pleiotropic cytokine involved in the establishment of inflammatory and protective immune responses against invading pathogens by regulating the functions of cells belonging to both the innate and adaptive immune systems [78]. In contrast, interleukin 22 promotes the proliferation of

epithelial and stromal cells, thus contributing to tissue regeneration, and also to the modulation of host defense at barrier surfaces [79].

About the genetic determinism of dairy traits in Murciano-Granadina goats

The most significant association that we have detected in our study is that between the chromosome 6 region containing the casein genes (QTL6) and protein percentage. This result is relevant because caseins constitute ~80% of the total milk protein content [80]. Moreover, we have observed differential expression of the four casein genes when comparing T1/T2 vs. T3. By applying a physiological candidate gene approach, Caravaca et al. [81] found that the *CSN3* genotype is significantly associated with casein and protein contents in Murciano-Granadina goats, while the *CSN1S1* genotype did not show significant associations with protein, casein, and fat concentrations. In Norwegian goats, Hayes et al. [82] described significant associations between *CSN1S1* (protein percentage and fat kilograms) and *CSN3* genotypes (fat percentage and protein percentage) and the phenotypic variation of dairy traits. In 2016, Carillier-Jacquin and colleagues [83] reported that *CSN1S1* genotypes had a significant effect on milk yield and milk fat and protein contents in French goat breeds. Moreover, a GWAS for dairy traits in Alpine and Saanen goats detected highly significant associations between markers mapping to the casein cluster and milk protein and fat contents [1]. Indeed, we also detected a chromosome-wide significant association between QTL6 and fat percentage. The pleiotropic effects of the casein genotypes on milk protein and fat contents could be due to the fact that, in the mammary epithelial cell, the transport of proteins and lipids is coupled to a certain extent [84].

Another relevant genome-wide significant association was that between QTL1 on chromosome 2 (130.72-131.01 Mb) and lactose percentage. This region overlaps the NGFI-A binding protein 1 (*NAB1*) gene, also known as EGR1

binding protein 1 gene. This gene shows an increased expression during mouse lactation and encodes a molecule that binds to the proximal promoter of the galactokinase gene, which is involved in galactose catabolism [85]. We also identified a third genome-wide significant association between a chromosome 17 region (QTL17, 11.20 Mb) and protein and dry matter percentages. This region closely maps to the T-Box 3 (*TBX3*) gene, which is highly expressed in luminal cells during early mammary gland initiation by interacting with Wnt and fibroblast growth factor (Fgf) signaling [86, 87].

The comparison of the genome-wide and chromosome-wide significant associations detected by us vs. those reported by Martin et al. [1] revealed a low level of positional concordance, suggesting the existence of a remarkable level of genetic heterogeneity amongst caprine breeds with regard to the genetic determinism of milk traits. Indeed, in the GWAS carried out by Martin et al. [1] more than 50% of the associations were exclusively detected in one of the two breeds under analysis (Alpine and Saanen) despite their close genetic relatedness. This finding supports the proposal of using breed-specific reference genomes to increase the accuracy of genomic analyses [88]. Moreover, in humans a large amount of variants occurs at different frequencies in different populations, having variable effects on complex traits and producing a substantial level of genetic heterogeneity [89]. Technical and experimental factors related to population size and marker density may also influence statistical power to detect associations [90]. Many of the QTLs detected by us were represented by a single SNP, possibly due to the low LD between nearby markers [91, 92, 93, 94]. Finally, only a few genes located within or close to QTLs showed differential expression between T1/T2 and T3, suggesting that the set of DE genes in these two physiological states has a weak correspondence with the set of genes influencing the quantitative variation of milk traits.

Conclusions

The ceasing of lactation in Murciano-Granadina goats involves the downregulation of the mRNA expression of many genes related to the synthesis, uptake and transportation of proteins, lipids and carbohydrates as well as changes in the mRNA expression of genes involved in the maintenance of calcium homeostasis. We also observed an increased expression of genes modulating cell death and tissue remodeling that probably mediate the involution and regeneration of the mammary gland during the dry period. From an immunological perspective, genes that contribute to the formation of mucous and surfactant barriers are downregulated in the dry period, possibly increasing the risk of infection. However, we have also observed an increase in the mRNA expression of defensin, cytokine and complement genes which should ensure the elicitation of an effective immune response against pathogens. Finally, the results obtained in the GWAS allows us to conclude that the casein genes, which are strongly downregulated during the dry period, are major genetic determinants of the phenotypic variance of milk protein and fat composition traits recorded in Murciano-Granadina goats, thus supporting the use of casein genotypes as a source of information to improve these two phenotypes.

Supplementary Information

Additional file 1: Table S1. Information about the Murciano-Granadina goats sampled in the RNA-Seq experiment

Additional file 2: Figure S1. Similarity matrix of samples used for detecting differentially expressed genes. T1, T2 and T3 correspond to 78.25 ± 9.29 d (T1, early lactation), 216.25 ± 9.29 d (T2, late lactation) and 285.25 ± 9.29 d (T3, dry period) after parturition, respectively. The sample T3-22 (red arrow) clustered with T1/T2 samples probably because it was obtained from a goat that was not successfully dried-off at the time of sampling.

Additional file 3: Tables S2-S4. List of differentially expressed genes between the T1 and T2, T1 and T3, and T2 and T3 time points

Additional file 4: Figure S2. Venn diagram depicting the overlaps of differentially expressed genes between pair-wise T1 vs. T2, T1 vs. T3 and T2 vs. T3 comparisons. T1 and T2 represent early (78.25 ± 9.29 d after parturition) and late (216.25 ± 9.29 d) lactation, respectively, while T3 (285.25 ± 9.29 d) corresponds to the dry period.

Additional file 5: Table S5-S6. Pathways and gene ontology (GO) terms enriched in the set of 1,654 differentially expressed genes on the basis of human and goat background gene sets

Additional file 6: Table S7. Descriptive statistics of seven dairy traits recorded in the first lactation of 822 Murciano-Granadina goats

Additional file 7: Figure S3. Histograms of the phenotypic values of the percentage of protein (a), fat (b), lactose (c) and dry matter (d), milk yield normalized to 210 d (e), length of lactation (f), logarithmically transformed somatic cell count (g) recorded in the first lactation of Murciano-Granadina goats. The raw somatic cell count is ($\times 10^3$ cells/mL) shown in (h).

Additional file 8: Figure S4. Structure of the Murciano-Granadina population employed in the GWAS as assessed by principal component analysis (PCA) based on Goat SNP50 BeadChip genotypes. PC1 and PC2 indicate the principal components 1 and 2, respectively. Values in parentheses reflect the percentage of variance in the data explained by each principal component.

Additional file 9: Figure S5. Linkage disequilibrium (LD) decay in 822 Murciano-Granadina goats with available Goat SNP50 BeadChip genotypes. The scatter plot shows the decline of r^2 between single nucleotide polymorphisms (*y*-axis) with distance expressed in bp (*x*-axis). The fitting line is depicted in red.

Abbreviations

DE: Differentially expressed; DMP: Dry matter percentage; FDR: False discovery rate; FP: Fat percentage; GFF: General feature format; GO: Gene ontology; GWAS: Genome-wide association study; LD: Linkage disequilibrium; log_2FC: log_2 of the fold-change; LOL: Length of lactation; LP: Lactose percentage; MAF: Minor allele frequency; Mb: Megabase; MY210: Milk yield at 210 days; PCA: Principal component analysis; PP: Protein percentage; QTL: Quantitative trait locus; RNA-Seq: RNA sequencing; SCC: Somatic cell count; SNP: Single nucleotide polymorphism;

Acknowledgements

Many thanks to Manuel Delgado Vico, Emilio Martínez Hurtado, Miguel García García and Teresa Novo Díaz from CAPRIGRAN for carrying out phenotype recording and blood sample collection in Murciano-Granadina goats. We are also indebted to Ramón Costa and all the staff of the Farm and Experimental Field Service of the Universitat Autònoma de Barcelona for their collaboration in the collection of mammary biopsies from goats.

Author's contributions

MA, JJ, JVD and VL designed the study. JFA, JVD and AM coordinated all tasks involved in phenotype recording. VL did all DNA extractions. BC and AC assessed DNA quality and carried out the Goat SNP50 BeadChip genotyping tasks. XS, JLRTC, EMS, MA, MGL and DG collected mammary gland biopsies from goats. EMS, MGL and DG did the RNA extractions from the mammary gland samples. DG did the bioinformatic and statistical analyses, with the cooperation of EMS and MGL. MA and DG wrote the first draft of the paper. All authors read and approved the content of the paper.

Funding

This research was funded by the European Fund for Regional Development/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación/Project Reference AGL2016-76108-R. We acknowledge the financial support from the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa Programme for Centres of Excellence in R&D" 2016-2019 (SEV-2015-0533), and from the CERCA programme of the Generalitat de Catalunya. Dailu Guan was funded by a PhD fellowship from the China Scholarship Council (CSC). Maria Luigi-Sierra was funded with a PhD fellowship "Formación de Personal Investigador" (BES-C-2017-0024) awarded by the Spanish Ministry of Economy and Competitivity. Emilio Mármol-Sánchez was funded with a PhD fellowship (FPU15/01733) awarded by the Spanish Ministry of Education and Culture (MECD).

Availability of data and materials

Goat SNP50 BeadChip genotypes and mammary RNA-Seq data are accessible at Figshare (https://doi.org/10.6084/m9.figshare.11881617) and Sequence Read Archive (SRA) database (PRJNA607923), respectively.

Ethics approval and consent to participate

The protocol for collecting mammary biopsies was approved by the Ethics Committee on Animal and Human Experimentation of the Universitat Autònoma de Barcelona (procedure code: UAB 3859).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Martin P, Palhière I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, et al. A genome scan for milk production traits in dairy goats reveals two new mutations in Dgat1 reducing milk fat content. Sci Rep. 2017; 7:1872.
2. Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, Conington J. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. J Dairy Sci. 2018; 101:2213-25.
3. Sharma A, Lee JS, Dang CG, Sudrajad P, Kim HC, Yeon SH, et al. Stories and challenges of genome wide association studies in livestock - a review. Asian-Austral J Anim Sci. 2015; 28:1371-9.
4. Ji Z, Chao T, Zhang C, Liu Z, Hou L, Wang J, et al. Transcriptome analysis of dairy goat mammary gland tissues from different lactation stages. DNA Cell Biol. 2019; 38:129-43.
5. Crisà A, Ferrè F, Chillemi G, Moioli B. RNA-sequencing for profiling goat milk transcriptome in colostrum and mature milk. BMC Vet Res. 2016; 12:264.
6. Pławińska-Czarnak J, Zarzyńska J, Majewska A, Jank M, Kaba J, Bogdan J, et al. Selected tissues of two polish goat breeds do not differ on genomic level. Anim Sci Pap Rep. 2019; 37:53-64.
7. McCabe M, Waters S, Morris D, Kenny D, Lynn D, Creevey C. RNA-Seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance. BMC Genomics. 2012; 13:193.

8. Yang B, Jiao B, Ge W, Zhang X, Wang S, Zhao H, et al. Transcriptome sequencing to detect the potential role of long non-coding RNAs in bovine mammary gland during the dry and lactation period. BMC Genomics. 2018; 19:605.

9. Dai WT, Zou YX, White RR, Liu JX, Liu HY. Transcriptomic profiles of the bovine mammary gland during lactation and the dry period. Funct Integr Genomics. 2018; 18:125-40.

10. Fang L, Sahana G, Su G, Yu Y, Zhang S, Lund MS, et al. Integrating sequence-based GWAS and RNA-Seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle. Sci Rep. 2017; 7:45560.

11. Yang J, Jiang J, Liu X, Wang H, Guo G, Zhang Q, et al. Differential expression of genes in milk of dairy cattle during lactation. Anim Genet. 2016; 47:174-80.

12. Suárez-Vega A, Gutiérrez-Gil B, Klopp C, Robert-Granie C, Tosser-Klopp G, Arranz JJ. Characterization and comparative analysis of the milk transcriptome in two dairy sheep breeds using RNA sequencing. Sci Rep. 2015; 5:18399.

13. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010; 38:e131.

14. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17:13.

15. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017; 49:643-50.

16. Kim D, Landmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12:357-60.

17. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-Seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016; 11:1650-67.

18. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30:923-30.

19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 2014; 15:550.

20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B. 1995; 57:289-300.

21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008; 4:44-57.

22. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1-13.

23. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988; 16:1215.

24. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52K SNP Chip for goats. PLoS One. 2014; 9:e86227.

25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904-9.

26. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190.

27. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44:821-4.

28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841-2.

29. 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.

30. Kelwick R, Desanlis I, Wheeler GN, Edwards DR. The ADAMTS (a Disintegrin and metalloproteinase with Thrombospondin motifs) family. Genome Biol. 2015; 16:113.

31. Britto CJ, Cohn L. Bactericidal/permeability-increasing protein fold-containing family member A1 in airway host protection and respiratory disease. Am J Respir Cell Mol Biol. 2015; 52:525-34.

32. Kinugasa T, Sakaguchi T, Gu X, Reinecker HC. Claudins regulate the intestinal barrier in response to immune mediators. Gastroenterology. 2000; 118:1001-11.

33. Anantamongkol U, Charoenphandhu N, Wongdee K, Teerapornpuntakit J, Suthiphongchai T, Prapong S, et al. Transcriptome analysis of mammary tissues reveals complex patterns of transporter gene expression during pregnancy and lactation. Cell Biol Int. 2010; 34:67-74.

34. Robinson GW, Hennighausen L. Inhibins and activins regulate mammary epithelial cell differentiation through mesenchymal-epithelial interactions. Development. 1997; 124:2701-8.

35. Jones PL, Boudreau N, Myers CA, Erickson HP, Bissell MJ. Tenascin-C inhibits extracellular matrix-dependent gene expression in mammary

epithelial cells. Localization of active regions using recombinant tenascin fragments. J Cell Sci. 1995; 108:519-27.

36. van Hekken DL, Eigel WN. Distribution of the lysosomal enzyme aryl sulfatase in murine mammary tissue through pregnancy, lactation, and involution. J Dairy Sci. 1990; 73:2318-26.

37. Maksimovic J, Sharp JA, Nicholas KR, Cocks BG, Savin K. Conservation of the ST6Gal I gene and its expression in the mammary gland. Glycobiology. 2010; 21:467-81.

38. Bennett EP, Mandel U, Clausen H, Gerken TA, Fritz TA, Tabak LA. Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. Glycobiology. 2011; 22:736-56.

39. Osorio JS, Lohakare J, Bionaz M. Biosynthesis of milk fat, protein, and lactose: roles of transcriptional and posttranscriptional regulation. Physiol Genomics. 2016; 48:231-56.

40. Dado-Senn B, Skibiel AL, Fabris TF, Zhang Y, Dahl GE, Peñagaricano F, et al. RNA-Seq reveals novel genes and pathways involved in bovine mammary involution during the dry period and under environmental heat stress. Sci Rep. 2018; 8:11069.

41. Bionaz M, Hurley W, Loor J. Milk protein synthesis in the lactating mammary gland: insights from transcriptomics analyses. In: Hurley WL, editor. Milk Protein. London: IntechOpen; 2012. https://doi.org/10.5772/46054.

42. Wakao H, Gouilleux F, Groner B. Mammary-gland factor (Mgf) is a novel member of the cytokine regulated transcription factor gene family and confers the prolactin response. EMBO J. 1994; 13:2182-91.

43. Bionaz M, Loor JJ. Gene networks driving bovine mammary protein synthesis during the lactation cycle. Bioinform Biol Insights. 2011; 5:83-98.

44. Bailey CG, Ryan RM, Thoeng AD, Ng C, King K, Vanslambrouck JM, et al. Loss-of-function mutations in the glutamate transporter SLC1A1 cause human dicarboxylic aminoaciduria. J Clin Invest. 2011; 121:446-53.

45. Closs EI, Boissel JP, Habermeier A, Rotmann A. Structure and function of cationic amino acid transporters (CATs). J Membr Biol. 2006; 213:67-77.

46. Scalise M, Pochini L, Console L, Losso MA, Indiveri C. The human SLC1A5 (ASCT2) amino acid transporter: from function to structure and role in cell biology. Front Cell Dev Biol. 2018; 6:96.

47. Boucher J, Kleinridders A, Kahn CR. Insulin receptor signaling in normal and insulin-resistant states. Cold Spring Harb Perspect Biol. 2014; 6:a009191.

48. Bequette BJ, Kyle CE, Crompton LA, Buchan V, Hanigan MD. Insulin regulates milk production and mammary gland and hind-leg amino acid fluxes and blood flow in lactating goats. J Dairy Sci. 2001; 84:241-55.

49. Menzies KK, Lefèvre C, Macmillan KL, Nicholas KR. Insulin regulates milk protein synthesis at multiple levels in the bovine mammary gland. Funct Integr Genomics. 2009; 9:197-217.

50. Huang S, Czech MP. The GLUT4 glucose transporter. Cell Metab. 2007; 5:237-52.

51. Komatsu T, Itoh F, Kushibiki S, Hodate K. Changes in gene expression of glucose transporters in lactating and nonlactating cows. J Anim Sci. 2005; 83:557-64.

52. Yoon M. The role of PPARα in lipid metabolism and obesity: focusing on the effects of estrogen on PPARα actions. Pharm Res. 2009; 60:151-9.

53. Foryst-Ludwig A, Clemenz M, Hohmann S, Hartge M, Sprang C, Frost N, et al. Metabolic actions of estrogen receptor Beta (ERβ) are mediated by a negative cross-talk with PPARγ. PLoS Genet. 2008; 4:e1000108.

54. Bionaz M, Loor JJ. Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics. 2008; 9:366.

55. Ahmadian M, Suh JM, Hah N, Liddle C, Atkins AR, Downes M, et al. PPARγ signaling and metabolism: the good, the bad and the future. Nat Med. 2013; 19:557-66.

56. Horst RL, Goff JP, Reinhardt TA. Calcium and vitamin D metabolism during lactation. J Mammary Gland Biol Neoplasia. 1997; 2:253-63.

57. Wysolmerski JJ. Parathyroid hormone-related protein: an update. J Clin Endocrinol Metab. 2012; 97:2947-56.

58. Lanske B, Razzaque MS. Molecular interactions of FGF23 and PTH in phosphate regulation. Kidney Int. 2014; 86:1072-4.

59. Mensenkamp AR, Hoenderop JGJ, Bindels RJM. TRPV5, the gateway to Ca^{2+} homeostasis. In: Flockerzi V, Nilius B, editors. Transient Receptor Potential (TRP) channels. Berlin: Springer Berlin Heidelberg; 2007. p. 207-20.

60. Peng JB, Suzuki Y, Gyimesi G, Hediger MA. TRPV5 and TRPV6 calcium-selective channels. In: Kozak JA, Putney Jr JW, editors. Calcium entry channels in non-excitable cells. Boca Raton (FL): CRC Press/Taylor & Francis; 2018. https://doi.org/10.1201/9781315152592-13.

61. Yu TC, Chang CJ, Ho CH, Peh HC, Chen SE, Liu WB, et al. Modifications of the defense and remodeling functionalities of bovine

neutrophils inside the mammary gland of milk stasis cows received a commercial dry-cow treatment. Vet Immunol Immunopathol. 2011; 144:210-9.

62. Watson CJ. Involution: apoptosis and tissue remodelling that convert the mammary gland from milk factory to a quiescent organ. Breast Cancer Res. 2006; 8:203.

63. Stanford JC, Cook RS. Apoptosis and clearance of the secretory mammary epithelium. In: Rudner J, editor. Apoptosis. London: IntechOpen; 2013. https://doi.org/10.5772/52160.

64. Marti A, Lazar H, Ritter P, Jaggi R. Transcription factor activities and gene expression during mouse mammary gland involution. J Mammary Gland Biol Neoplasia. 1999; 4:145-52.

65. Le Provost F, Miyoshi K, Vilotte JL, Bierie B, Robinson GW, Hennighausen L. SOCS3 promotes apoptosis of mammary differentiated cells. Biochem Biophys Res Commun. 2005; 338:1696-701.

66. Schere-Levy C, Buggiano V, Quaglino A, Gattelli A, Cirio MC, Piazzon I, et al. Leukemia inhibitory factor induces apoptosis of the mammary epithelial cells and participates in mouse mammary gland involution. Exp Cell Res. 2003; 282:35-47.

67. Marshman E, Green KA, Flint DJ, White A, Streuli CH, Westwood M. Insulin-like growth factor binding protein 5 and apoptosis in mammary epithelial cells. J Cell Sci. 2003; 116:675-82.

68. Tiffen PG, Omidvar N, Marquez-Almuina N, Croston D, Watson CJ, Clarkson RWE. A dual role for oncostatin M signaling in the differentiation and death of mammary epithelial cells in vivo. Mol Endocrinol. 2008; 22:2677-88.

69. Khokha R, Werb Z. Mammary gland reprogramming: metalloproteinases couple form with function. Cold Spring Harb Perspect Biol. 2011; 3:a004333.

70. Leelahapongsathon K, Piroon T, Chaisri W, Suriyasathaporn W. Factors in dry period associated with intramammary infection and subsequent clinical mastitis in early postpartum cows. Asian-Austral J Anim Sci. 2016; 29:580-5.

71. Betts CB, Pennock ND, Caruso BP, Ruffell B, Borges VF, Schedin P. Mucosal immunity in the female murine mammary gland. J Immunol. 2018; 201:734-46.

72. Hasnain SZ, Gallagher AL, Grencis RK, Thornton DJ. A new role for mucins in immunity: insights from gastrointestinal nematode infection. Int J Biochem Cell Biol. 2013; 45:364-74.

73. Han S, Mallampalli RK. The role of surfactant in lung disease and host defense against pulmonary infections. Ann Am Thorac Soc. 2015; 12:765-74.

74. Rainard P. The complement in milk and defense of the bovine mammary gland against infections. Vet Res. 2003; 34:647-70.

75. Meade KG, O'Farrelly C. β-Defensins: farming the microbiome for homeostasis and health. Front Immunol. 2019; 9:3072.

76. Takatsu K. Interleukin-5 and IL-5 receptor in health and diseases. Proc Jpn Acad Ser B Phys Biol Sci. 2011; 87:463-85.

77. Nie Y, Waite J, Brewer F, Sunshine MJ, Littman DR, Zou YR. The role of CXCR4 in maintaining peripheral B cell compartments and humoral immunity. J Exp Med. 2004; 200:1145-56.

78. Perera P-Y, Lichy JH, Waldmann TA, Perera LP. The role of interleukin-15 in inflammation and immune responses to infection: implications for its therapeutic use. Microbes Infect. 2012; 14:247-61.

79. Dudakov JA, Hanash AM, van den Brink MRM. Interleukin-22: immunobiology and pathology. Annu Rev Immunol. 2015; 33:747-85.

80. Corredig M, Nair PK, Li Y, Eshpari H, Zhao Z. Invited review: understanding the behavior of caseins in milk concentrates. J Dairy Sci. 2019;102:4772-82.

81. Caravaca F, Carrizosa J, Urrutia B, Baena F, Jordana J, Amills M, et al. Short communication: effect of α_{S1}-casein (*CSN1S1*) and κ-casein (*CSN3*) genotypes on milk composition in Murciano-Granadina goats. J Dairy Sci. 2009; 92:2960-4.

82. Hayes B, Hagesæther N, Ådnøy T, Pellerud G, Berg PR, Lien S. Effects on production traits of haplotypes among casein genes in Norwegian goats and evidence for a site of preferential recombination. Genetics. 2006; 174:455-64.

83. Carillier-Jacquin C, Larroque H, Robert-Granié C. Including α_{s1} casein gene information in genomic evaluations of French dairy goats. Genet Sel Evol. 2016; 48:54.

84. Ollivier-Bousquet M. Milk lipid and protein traffic in mammary epithelial cells: joint and independent pathways. Reprod Nutr Dev. 2002; 42:149-62.

85. Yang F, Agulian T, Sudati JE, Rhoads DB, Levitsky LL. Developmental regulation of galactokinase in suckling mouse liver by the Egr-1 transcription factor. Pediatr Res. 2004; 55:822-9.

86. Platonova N, Scotti M, Babich P, Bertoli G, Mento E, Meneghini V, et al. *TBX3*, the gene mutated in ulnar-mammary syndrome, promotes growth of mammary epithelial cells via repression of p19ARF, independently of p53. Cell Tissue Res. 2007; 328:301-16.

87. Eblaghie MC, Song SJ, Kim JY, Akita K, Tickle C, Jung HS. Interactions between FGF and Wnt signals and *Tbx3* gene expression in mammary gland initiation in mouse embryos. J Anat. 2004; 205:1-13.

88. Czech B, Frąszczak M, Mielczarek M, Szyda J. Identification and annotation of breed-specific single nucleotide polymorphisms in Bos taurus genomes. PLoS One. 2018; 13:e0198419.

89. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature. 2019; 570:514-8.

90. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019; 20:467-84.

91. Lashmar SF, Visser C, Ev MK. SNP-based genetic diversity of South African commercial dairy and fiber goat breeds. Small Rumin Res. 2016; 136:65-71.

92. Michailidou S, Tsangaris GT, Tzora A, Skoufos I, Banos G, Argiriou A, et al. Analysis of genome-wide DNA arrays reveals the genomic population structure and diversity in autochthonous Greek goat breeds. PLoS One. 2019; 14:e0226179.

93. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017; 101:5-22.

94. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018; 19:491-504.

Genomic analysis of the origins of extant casein variation in goats

D. Guan,* E. Mármol-Sánchez,* T. F. Cardoso,*† X. Such,‡ V. Landi,§ N. R. Tawari,# and M. Amills*‡[1]

* Department of Animal Genetics, Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. † CAPES Foundation, Ministry of Education of Brazil, Brasilia D.F, 70.040-020 Brazil. ‡ Departament de Ciència animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. § Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain. # Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, Genome, #02-01, Singapore 138672.

[1]Corresponding author: Marcel Amills (marcel.amills@uab.cat)

Journal of Dairy Science 102:1–12

https://doi.org/10.3168/jds.2018-15281

ABSTRACT

The variation in the casein genes has a major impact on the milk composition of goats. Even though many casein polymorphisms have been identified so far, we do not know yet whether they are evolutionarily ancient (i.e., they existed before domestication) or young (i.e., they emerged after domestication). Herewith, we identified casein polymorphisms in a data set of 106 caprine whole-genome sequences corresponding to bezoars (*Capra aegagrus*, the ancestor of domestic goats) and 4 domestic goat (*Capra hircus*) populations from Europe, Africa, the Far East, and the Near East. Domestic and wild goat populations shared a substantial number of casein SNPs, from 36.1% (CSN2) to 55.1% (CSN1S2). The comparison of casein variation among bezoars and the 4 domestic goat populations demonstrated that more than 50% of the casein SNPs are shared by 2 or more populations, and 18 to 44% are shared by all populations. Moreover, the majority of casein alleles reported in domestic goats also segregate in the bezoar, including several alleles displaying significant associations with milk composition (e.g., the A/B alleles of the CSN1S1 and CSN3 genes, the A allele of the CSN2 gene). We conclude that much of the current diversity of the caprine casein genes comes from ancient standing variation segregating in the ancestor of modern domestic goats.

Key words

Domestication, standing variation, next-generation sequencing, single nucleotide polymorphism

INTRODUCTION

Caseins represent 80% of the protein content of milk and they have a major impact on dairy traits, as well as on cheese yield and texture (Remeuf et al., 1991). Goat caseins α_{S1}, α_{S2}, β, and κ are encoded by the *CSN1S1*, *CSN1S2*, *CSN2*, and *CSN3* genes, respectively, which map to a 250-kb region on chromosome 6 in the order CSN1S1-CSN2-CSN1S2-CSN3 (Rijnkels, 2002). Polymorphisms in these 4 genes have been implicated in the variation of milk yield and composition (protein and fat contents) as well as in the determination of milk rheological properties and the yield and organoleptic attributes of cheese (reviewed in Martin et al., 1999; Moioli et al., 2007; Amills et al., 2012).

The domestication of the bezoar (*Capra aegagrus*) in the Fertile Crescent 10,000 yr before present resulted in the domestic goat (*Capra hircus*), an important economic resource in developing countries (Zeder and Hesse 2000; Naderi et al., 2008). Even though many reports describing the variability of goat casein genes have been published (Martin et al., 2002; Moioli et al., 2007; Amills, 2014), we do not know yet whether casein polymorphisms are evolutionarily ancient (i.e., they existed before domestication) or young (i.e., they emerged after domestication). We addressed this question by identifying casein polymorphisms from 106 caprine whole-genome sequences and comparing the allelic variation of the 4 casein genes in (1) 2 populations: bezoars and domestic goats, (2) 5 populations: bezoars and 4 groups of domestic goats from Europe, Africa, the Far East, and Near East. We aimed to determine whether extant genetic variation in the goat casein genes was present before domestication (as standing variation segregating in the bezoar) or if it emerged in the context of the evolutionary processes that took place during and after domestication.

MATERIALS AND METHODS

Retrieval of Goat Whole-Genome Sequences

Whole-genome sequences from 110 wild and domestic goats (Becker et al., 2015; Benjelloun et al., 2015; Reber et al., 2015; Menzi et al., 2016; Wang et al., 2016; Li et al., 2017; Alberto et al., 2018) were retrieved from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra; **Supplemental Table S1**). Specifically, this data set included genome sequences from bezoars (n = 22) as well as 4 domestic goat populations from Europe (Alpine, n = 2; Chamois Colored, n = 2; Grisons Striped, n = 1; Saanen, n = 6; Coppernecked, n = 1; Tessin Grey, n = 1), Africa (local Moroccan population, n = 20), the Far East (Inner Mongolia Cashmere goat, n = 9, Liaoning Cashmere goat, n = 10; Tibetan goat, n = 16), and Near East (local Iranian breed, n = 20). We retrieved all goat genome sequences that were available at the time of initiating our experiment. All raw data in SRA format were converted into the fastq format by using the fastq-dump 2.8.2 tool available in the SRA-toolkits package (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/).

Discovery and Annotation of Genomic Variants

To obtain high-quality single nucleotide polymorphism (SNP) and insertion/deletion (INDEL), fastq files were filtered with the Trimommatic software (version 0.36, Bolger et al., 2014). Only paired-end reads were used in the alignment step. Sequences were aligned to the goat reference genome (ARS1, Bickhart et al., 2017) with the BWA MEM algorithm with default settings (Li, 2013). Files in sequence alignment map (SAM) format were sorted and

converted into binary format to remove PCR duplicates and to realign INDEL regions with the Picard tool (https://broadinstitute.github.io/picard/). The HaplotypeCaller function of the Genome Analysis Toolkit (GATK, version 3.8) was used to generate vcf (variant call format) files by considering default parameters (McKenna et al., 2010). Finally, a hard filtering step was performed by following the GATK best practices recommendations. The SNP data set was then imputed and phased by using the Beagle 4.1 software (Browning and Browning, 2016) to improve genotype calls based on genotype likelihoods.

Investigation of Population Structure

We used the autosomal SNP identified with GATK (McKenna et al., 2010) to investigate the population structure of bezoars and domestic goats. A thinned set of autosomal SNPs was selected with the command "--hwe 0.001 --maf 0.05 --geno 0.3 --indep-pairwise 50 5 0.2" of the PLINK v1.9 software (Purcell et al., 2007). Beforehand, individuals with pi-hat values, estimated based on an identity-by-descent (IBD) matrix, above 0.4 were removed from the data set to avoid biases produced by relatedness. By doing so, 4 individuals (2 EU and 2 FE goats) were excluded and the final data set was based on 106 caprine genomes. A neighbor-joining tree was constructed with the MEGA7 software (Kumar et al., 2016) based on an identity-by-state (IBS) distance matrix (Purcell et al., 2007). Principal components analyses (PCA) based on 11,226,125 SNPs with a whole-genome distribution and 1,221 SNPs mapping to the casein genes were performed with PLINK v1.9 software (Purcell et al., 2007) by using the flag "--pca" with default parameters. In addition, the Admixture software (version 1.3.0, Alexander et al., 2009) was used to estimate population structure with a block relaxation algorithm. The number of clusters (K-value) went from 2 to 5, and the K-value with the lowest cross-validation error was identified by using the

method of Alexander and Lange (2011). Moreover, we repeated the Admixture analysis considering just the data set of 1,221 SNPs mapping to the casein genes.

Annotating the Variation of Caprine Casein Genes

The genomic coordinates of the goat casein genes (*CSN1S1*, *CSN2*, *CSN1S2*, and *CSN3*) in the ARS1 reference genome (Bickhart et al., 2017) were used to retrieve polymorphisms mapping to these 4 loci with VCFtools 1.8 (Danecek et al., 2011; https://samtools.github.io/bcftools/bcftools.html). Casein polymorphisms were classified and their effects were predicted with the SnpEff 4.3 software (Cingolani et al., 2012). Moreover, the SIFT Annotator (Vaser et al., 2016) was used to predict deleteriousness of missense SNP. When the SIFT predicted score is < 0.05, an AA substitution is classified as deleterious (Vaser et al., 2016); otherwise, it is tolerated or neutral. By following the strategy outlined in **Supplemental Figure S1**, we were able to convert sequence data into casein alleles or groups of alleles. This classification, which was based on information provided by Marletta et al. (2007), took into account several missense mutations that are outlined in **Supplemental Figure S1**. We were unable to discriminate between the B4 and E alleles of the *CSN1S1* gene because we could not trace the presence of the LINE insertion characteristic of the E allele (repetitive elements are usually filtered out before the alignment step). The SnpEff 4.3 software did not detect any mutation introducing a premature stop codon, so we did not identify null alleles in the casein genes. It is difficult to know whether this was due to a biological reality (absence of null alleles in the analyzed populations) or to an annotation problem associated with SnpEff 4.3. In contrast, the O1 and O2 alleles of the *CSN1S1* gene are large copy number variants whose genomic coordinates have not been reported at a fine resolution. They might be detectable using software such as Cn.MOPS or CNVnator, but the main limitation of our experiment was that we had a very heterogeneous data

set composed of whole-genome sequences generated with different types of libraries, platforms, and, more importantly, coverages, so detecting copy number variants based on read depth would be inaccurate. We did not use INDEL information to classify alleles because we believe that INDEL calling from sequence data can be quite unreliable. Indeed, O'Rawe et al. (2013) compared the concordance rates among INDEL detected by the GATK Unified Genotyper (v1.5), SOAPindel (v1.0), and SAMtools (v0.1.18) and concluded that there was just 26.8% agreement across all 3 software programs. Hasan et al. (2015) compared the performance of 7 INDEL calling tools and reported that the number of common INDELs called by all 7 tools was very low. For this reason, we decided to report the B2, F, and D alleles as a group.

The nucleotide diversity (π value, average number of pairwise differences between all individuals in the population) of the casein loci (based on 1,221 SNPs mapping to casein genes) was calculated with the VCFtools software (Danecek et al. 2011) by using the "--site-pi" command. The same conditions were used to estimate nucleotide diversity at the whole-genome level. All results in this study were visualized under the R software environment (https://www.r-project.org/).

RESULTS

Genome-Wide Analysis of Population Structure

By using a data set of 106 whole-genome sequences (**Supplemental Table S1**) from domestic goats and bezoars, a total of 31 billion paired-end reads were mapped to the goat reference genome ARS1 (Bickhart et al., 2017). The average sequencing depth was 9.92× and the average mapping rate > 99%

(**Supplemental Table S1**). Analysis of the sequence data with the GATK package (McKenna et al., 2010) made it possible to identify 51 million SNPs. The majority of these SNPs were biallelic, and only 509,001 sites displayed 3 or more alleles. Moreover, 35.17% of SNPs had minor allele frequencies (MAF) > 0.05 (17.94 million), whereas rare (MAF between 0.01 and 0.05) and very rare (MAF < 0.01) SNPs displayed frequencies of 29.72 and 36.74%, respectively. The average ratio of transitions to transversions was 2.11 for the whole data set, a result consistent with previous reports (Guan et al., 2016; Li et al., 2017).

After filtering, 11,226,125 autosomal SNPs were used to assess the population structure of the 106 bezoars and domestic goats (4 highly related individuals with pi-hat values > 0.4 were removed). The PCA and the neighbor-joining tree (**Figures 1a** and **1b**) showed that individuals clustered according to their geographic origin. In the PCA, bezoars and domestic goats from the Near East occupied an intermediate position between Far East goats and those from Europe and Africa. Moreover, Far East domestic goats formed a tight cluster, whereas bezoars had a more scattered distribution (**Figure 1b**). At K = 2, the Admixture analysis showed the existence of 2 different backgrounds in domestic goats: Africa/Europe and Far East, whereas Near East goats displayed an intermediate or admixed background (**Figure 2**). At the K-value with the lowest cross-validation error (K = 3), bezoars formed a distinctive group clearly differentiated from domestic goats (**Figure 2**). At K = 4, we observed the existence of 2 genetically differentiated subgroups in Far East goats, whereas at K = 5, European and African goats displayed different genetic backgrounds (**Figure 2**). When we repeated the PCA and Admixture analysis by using a panel of 1,221 SNPs mapping to the casein genes, we observed a substantial weakening of population structure (**Supplemental Figure S2**). This result might be because this second analysis was based on a very reduced set of SNPs (1,221 SNPs versus 11,226,125 SNPs used in the first analysis).

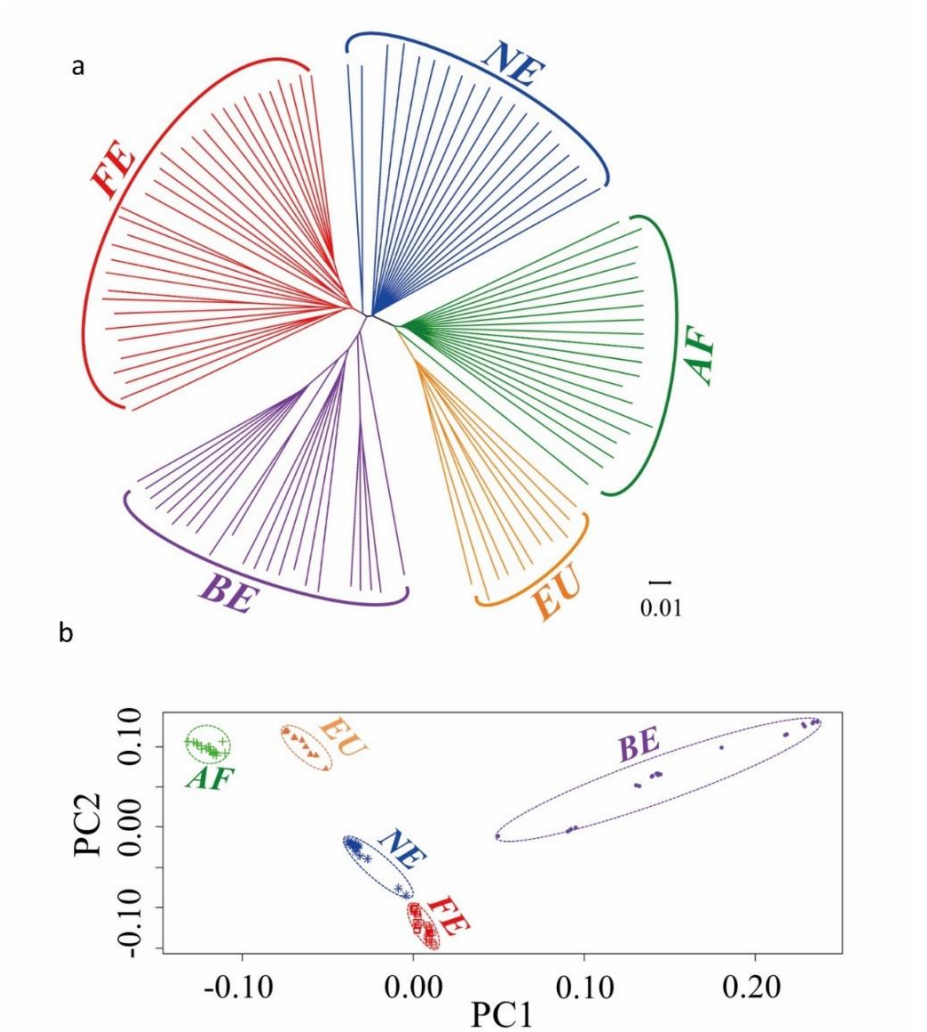

Figure 1. (**a**) Neighbor-joining tree, and (**b**) principal components analysis (PCA) of 106 bezoars (BE) and domestic goats from Europe (EU), Africa (AF), Near East (NE), and Far East (FE) based on a data set of 11,226,125 autosomal SNPs. The neighbor-joining tree was built according to an identity-by-state (IBS) distance matrix constructed with the PLINK software (Purcell et al., 2007) with default parameters. The PCA considered principal components (PC) 1 and 2, which explained 14.20% (6.20/eigenvalues) and 13.54% (5.91/eigenvalues) of the variance, respectively.

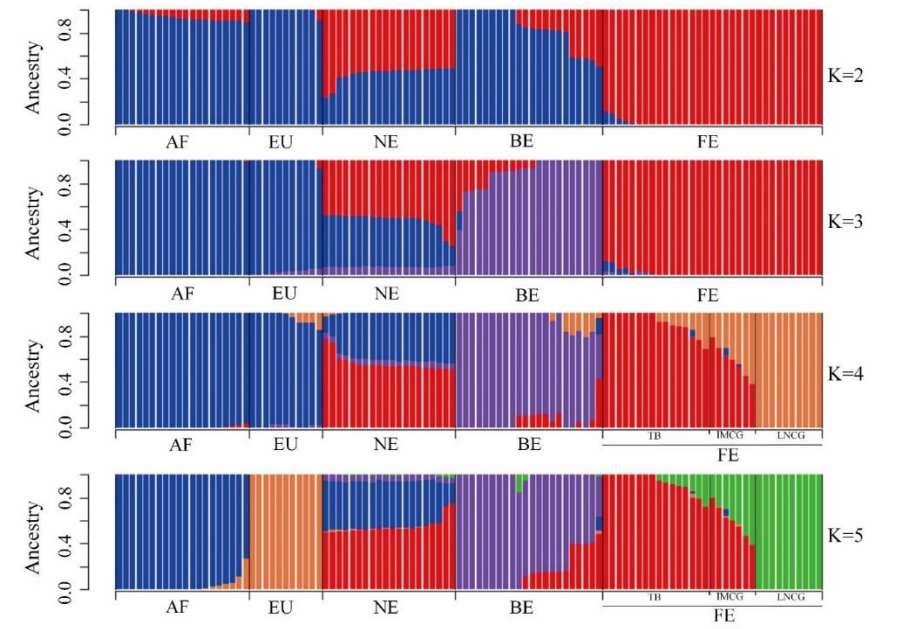

Figure 2. Analysis using Admixture software (version 1.3.0, Alexander et al., 2009) of 106 bezoars (BE) and domestic goats from Europe (EU), Africa (AF), Near East (NE), and Far East (FE) based on a data set of 11,226,125 autosomal SNPs. Each colored bar represents one individual and the length represents the proportion of the goat genome inherited from each ancestral population. In the Far East group, the following subpopulations are indicated: Tibetan (TB), Inner Mongolia Cashmere goats (IMCG), and Liaoning Cashmere goats (LNCG). The K-value defines the number of clusters.

Characterization of Variation in Goat Casein Genes

We identified hundreds of SNPs in the *CSN1S1* (455 SNPs, 6 missense), *CSN2* (194 SNPs, 5 missense), *CSN1S2* (292 SNPs, 11 missense), and *CSN3* (280 SNPs, 9 missense) genes (**Table 1**, **Supplemental Tables S2, S3,** and **S4**). In the 4 casein genes, most SNPs were intronic, and the second most abundant category was represented by SNPs located in upstream and downstream genic regions

(**Supplemental Table S4**). Annotation of SNPs with the SnpEff software (Cingolani et al., 2012) showed that the majority of casein polymorphisms were expected to have low or moderate effects. Indeed, only 1 SNP (g.85982647G > A), affecting a splice site in the *CSN1S1* gene (G allele), was predicted to have a high impact (**Supplemental Tables S2** to **S4**). The SIFT annotator (Vaser et al., 2016) captured additional missense SNPs predicted to be functionally relevant (**Supplemental Tables S2** to **S4**). Additionally, most casein SNPs identified in our investigation had MAF > 0.05 (47.1%), a result consistent with that obtained in the analysis of genome-wide diversity. With regard to INDEL, we found 81 in *CSN1S1*, 25 in CSN2, 59 in *CSN1S2*, and 49 in *CSN3* (**Supplemental Table S2**). However, we did not use INDELs in subsequent analyses because INDEL calling remains an error-prone process and the rate of false positives might be high because of alignment artifacts and to the fact that most of the INDEL calling tools lack accurate methods for checking sequencing errors before calling INDEL (Hasan et al. 2015). Indeed, concordance rates of INDEL calls between algorithms and sequencing platforms are reportedly low (Fang et al. 2014).

We analyzed casein SNP variation of bezoars and domestic goats (**Figure 3**). Domestic and wild populations shared a substantial number of casein SNPs, from 36.1% (*CSN2*) to 55.1% (*CSN1S2*). The comparison of casein variation among the 5 populations (bezoars and domestic goats from Europe, Africa, Near East, and Far East) also showed that more than 50% of casein SNPs are shared by 2 or more populations and 18% (*CSN3*) to 44% (*CSN1S1*) of SNPs are shared by all populations (**Figure 3**). Nucleotide diversity in the casein loci was similar in bezoar and domestic goat populations (**Figure 4a**), with the exception of Far East goats, which showed a reduced level of variation (t-test, $P < 0.05$). Moreover, the nucleotide diversity of the casein loci was higher (t-test, $P < 0.001$) than that observed along the autosomal genome (**Figure 4b**).

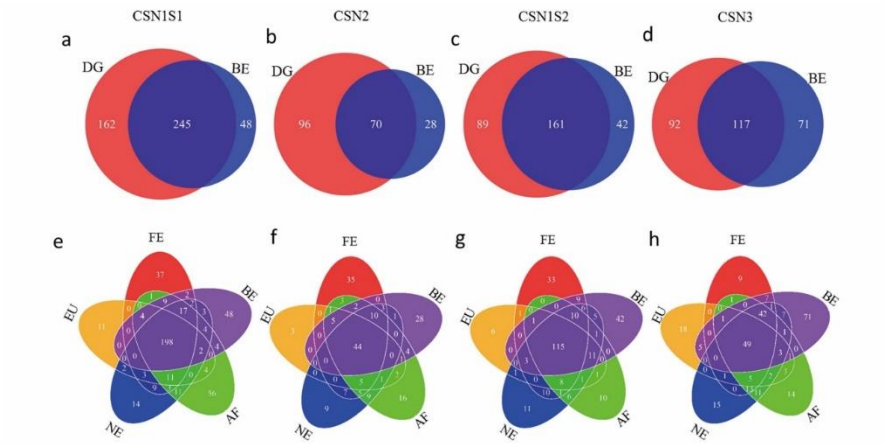

Figure 3. (**a**-**d**) Venn diagrams depicting the α_{S1}- (*CSN1S1*), α_{S2}- (*CSN1S2*), β- (*CSN2*), and κ- (*CSN3*) casein SNPs shared between bezoars (BE) and domestic goats (DG); (**e** to **h**) Venn diagrams depicting *CSN1S1*, *CSN1S2*, *CSN2*, and *CSN3* SNPs shared between bezoars (BE) and domestic goats from Europe (EU), Africa (AF), Near East (NE), and Far East (FE).

We used the pipeline reported in **Supplemental Figure S1** to detect casein alleles or groups of alleles based on sequence data (**Table 1**). In the *CSN1S1* gene, the A/I/N/O group of alleles was quite frequent in all populations, with an average frequency > 0.5. Moreover, we were able to detect combinations of SNPs that did not correspond to any of the *CSN1S1* alleles cataloged by Marletta et al. (2007); for example, $H_8P_{16}Q_{77}R_{100}A_{195}$, $H_8L_{16}Q_{77}R_{100}A_{195}$, $H_8P_{16}Q_{77}R_{100}T_{195}$, and $H_8L_{16}E_{77}K_{100}A_{195}$. These novel haplotypes were especially frequent in Far East goats. In the *CSN1S2*, *CSN2*, and *CSN3* genes, the most abundant alleles were A (average frequency = 0.54), C (average frequency = 0.67), and A/B (average frequency = 0.68), respectively. We also identified certain alleles in *CSN1S1* (B1, C and G), *CSN1S2* (E), and *CSN3* (I, K, and M) that are rare (average frequency < 0.05) or very rare (average frequency < 0.01). Five of these rare alleles were present in Far East goats at low frequencies, and 2 (K and I alleles of the *CSN3* gene) segregated exclusively in this population. We were unable to identify the D allele of the *CSN1S2* gene, and the null alleles plus the A1 allele of the *CSN2* gene remained undetected in our data set.

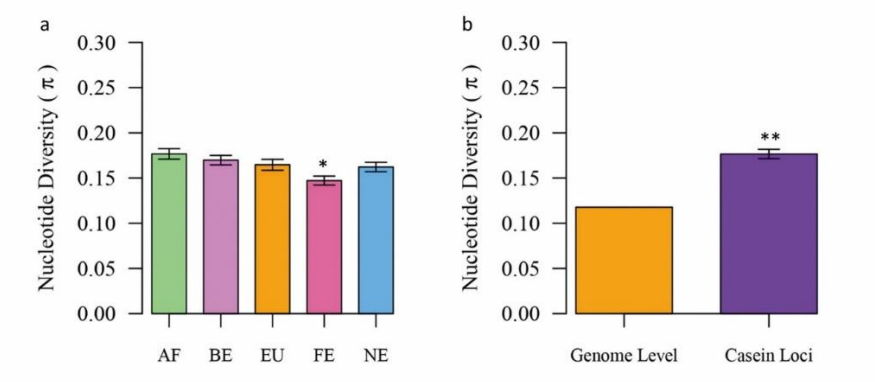

Figure 4. Nucleotide diversity of (**a**) the casein loci in bezoars (BE) and domestic goats from Africa (AF), Europe (EU), Near East (NE), and Far East (FE), and (**b**) the casein loci compared with the autosomal genome. Each bar represents the mean nucleotide diversity and its standard error. The standard error (2.08×10^{-5}) of the estimate of the nucleotide diversity corresponding to the autosomal genome is very small, so it is not depicted in the graph. *The nucleotide diversity of the Far East population was significantly lower ($P < 0.05$) than those of the other populations. **Nucleotide diversities of whole-genome and casein genes were significantly different ($P < 0.01$).

In certain cases, frequencies of casein alleles were quite divergent among populations. For instance, the H allele of the *CSN3* gene was relatively common in Far East goats but rare in goats from Africa, Europe, or the Near East (**Table 1**), and the D allele of the *CSN3* gene was quite frequent in goats from the Near and Far East but completely absent in the remaining caprine populations (**Table 1**). In the *CSN1S1* gene, the B3 allele segregated at moderate frequencies in African goats but did not segregate in the European population. Importantly, we found that the majority of casein alleles were present in the population of bezoars, indicating that their existence probably predates domestication. This finding was further supported by the segregation of the majority of casein alleles in two or more of the analyzed populations (**Table 1**), despite the fact that the populations are separated by considerable geographic distances.

Table 1. Frequencies of alleles or groups of alleles identified in the bezoars (BE) and domestic goats from Europe (EU), Africa (AF), Near East (NE) and Far East (FE) in current study

	Allele	**AF (N=20)**	**BE (N=22)**	**EU (N=11)**	**FE (N=33)**	**NE (N=20)**	**Total (N=106)**
CSN1S1	A-I-N-O1-O2	0.30	0.55	0.27	0.57	0.67	0.51
	B1	0	0.02	0	0.03	0	0.01
	B2-D-F	0.13	0.11	0.41	0.02	0	0.09
	B3	0.27	0.16	0	0.02	0.10	0.11
	B4-E	0.30	0.09	0.27	0.06	0.05	0.13
	C	0	0	0	0.03	0.08	0.02
	G	0	0	0.05	0	0	0.01
	Unreported	0	0.07	0	0.27	0.1	0.12
CSN1S2	A	0.45	0.62	0.64	0.59	0.43	0.54
	B	0	0.02	0	0.14	0	0.05
	C	0.43	0.10	0.09	0.18	0.50	0.26
	E	0	0.02	0	0.05	0	0.02
	F	0.12	0.24	0.27	0.04	0.07	0.13
CSN2	A	0.58	0.22	0.25	0.24	0.25	0.30
	C	0.39	0.75	0.60	0.76	0.75	0.67
	C1	0.03	0.03	0.15	0	0	0.03
CSN3	A	0.50	0.25	0.05	0.19	0.30	0.27
	B	0.45	0.52	0.95	0.23	0.25	0.41
	D	0	0	0	0.24	0.28	0.13
	G-L	0	0.14	0	0	0.10	0.05
	H	0	0.09	0	0.23	0	0.09
	I	0	0	0	0.08	0	0.02
	K	0	0	0	0.03	0	0.01
	M	0.05	0	0	0	0.07	0.02

DISCUSSION

A relevant evolutionary question that we aimed to answer in the current study was whether extant casein genetic variation segregating in domestic goats comes from standing variation (already present in bezoars before their domestication) or whether it emerged after goat domestication and dispersal (novel variation). Before analyzing casein variation, we investigated the genetic relationships and population structure of the 5 caprine populations under analysis (bezoars and goats from Europe, Africa, Near East, and Far East). Our results showed that individuals clustered according to their geographic origin (**Figure 1**). Goats were domesticated in a geographic area from the Central Zagros Mountains (Iran) to Eastern Anatolia (Zeder and Hesse, 2000; Naderi et al., 2008; reviewed in Pereira and Amorim, 2010) and subsequently spread into Europe, Africa, and Asia (Pereira and Amorim, 2010). The existence of genetic differentiation between the 5 populations analyzed in our study (**Figure 1**) is compatible with the hypothesis that different gene pools migrated through the Mediterranean and Danubian corridors in Europe, the Central Steppe and the Indus Valley in Asia, and North Africa (Pereira and Amorim, 2010). Indeed, the analysis of goat ancient genomes has shown that goats were domesticated at multiple locations (and time periods) in the Fertile Crescent (Daly et al., 2018). Genetic drift combined with the existence of differences in breed management, reproductive isolation, and selection goals probably contributed to the establishment of genetic differences between Asian, European, and African goats.

The analysis of population structure using Admixture analysis showed the existence of Western (Africa and Europe) and Eastern (Far East) genetic backgrounds, whereas the genetic background of Near Eastern domestic goats shared both components (K = 3, **Figure 2**). The third distinctive genetic background was represented by bezoars from the Near East (K = 3, **Figure 2**). Analysis of mitochondrial variation of Iranian wild boars revealed segregation

not only of Middle East haplotypes but also of haplotypes that are typically found in wild boars from the West (Europe and Africa) and Far East (Khalilzadeh et al., 2016). These results highlight that Iran has been an important contact zone between the East and the West, and also a key hotspot of genetic diversity (Khalilzadeh et al., 2016). Moreover, we also detected the existence of 2 different backgrounds in Far East goats, reflecting the existence of 2 different populations (K = 5, **Figure 2**); that is, Tibetan goats and 2 Cashmere breeds (Wang et al., 2016; Li et al., 2017). Inner Mongolia Cashmere goats displayed a genetic background intermediate between that of Tibetan and Liaoning Cashmere goats (**Figure 2**). This result points to the Mongolian Plateau being a critical hub for the dispersal of goats across East Asia (Pereira and Amorim, 2010), as reported for cattle (Ajmone-Marsan et al., 2010) and sheep (Zhao et al., 2017).

We annotated casein polymorphisms according to the genomic coordinates provided in the ARS1 assembly of the goat genome (Bickhart et al., 2017). Obviously, this annotation may differ from that used in previous publications. For instance, the missense *CSN3* Asn74Ser and Val86Ile polymorphisms identified by us (**Supplemental Table S2**) correspond to the Asn53Ser and Val65Ile substitutions reported by Marletta et al. (2007). These differences might be due, for instance, to the fact that AA residue numbering in a protein sequence may begin with the first AA of either the leader peptide or the mature protein sequence. Moreover, whole-genome sequencing with a modest coverage (average of 9.92 × in the current work) can yield thousands of false polymorphisms that are produced by sequencing errors (Robasky et al., 2014). However, these drawbacks should not have a major effect on the main conclusions of our study because we did not intend to build a curated catalog of casein variation in goats, which will eventually be reported in the Ensembl database (https://www.ensembl.org). Rather, we aimed to investigate the geographic distribution of caprine casein variation to make inferences about the

origins of such variation (i.e., to ascertain whether it arose from standing or novel variation).

The comparison of the nucleotide diversity of the casein loci in bezoars versus domestic goats showed that they have similar levels of variation (**Figure 4a**). However, goats from the Far East displayed lower levels of diversity, a feature that might be due to an ancient founder effect associated with goat dispersal after domestication. Moreover, the nucleotide diversity of the casein loci was higher than that observed in the autosomal genome (**Figure 4b**). Two preferential recombination sites have been reported in the casein cluster (Bevilacqua et al., 2002; Hayes et al., 2006), a circumstance that is known to promote the generation of diversity. Moreover, the casein genes are not essential to sustain life, so purifying selection is probably less intense than in other genomic regions that contain housekeeping genes.

In the 4 casein genes, a substantial number of SNPs (36-55%) were shared between the wild and domestic forms. Importantly, the number of analyzed bezoars was relatively low, so we cannot rule out the possibility that the percentage of shared variation between bezoars and domestic goats will increase if sample size is augmented. The variation shared between wild and domestic goats might have an ancestral origin, but such a pattern could also be produced by an introgression of the bezoar population with domestic goats. However, the analysis of **Figure 1** does not provide evidence of introgressed bezoars in our data set. Moreover, the comparison of casein polymorphisms across the 5 populations showed that more than 50% of the polymorphisms were shared between 2 or more populations, and that between 18 and 44% were shared by all populations. Casein diversity shared by all 5 populations probably has an ancestral origin (i.e., its existence probably predates the post-domestication dispersal of goats). These results indicate that a considerable proportion of casein variation might have been present in the bezoar before the domestication process.

Our findings agree well with other studies demonstrating that genetic variants of agricultural importance such as those related to tomato fruit size, maize plant

architecture (e.g., teosinte branched 1), seasonality controls, and seed size were already present as standing variation in the wild progenitors of domestic plant species (Larson et al., 2014). One exception to this general trend would be that of mutations that could be deleterious in the wild but not in a domestic context; however, in principle, mutations with functional consequences on the casein genes are not expected to have any effect on the biological efficacy of the individuals harboring them. These results contrast strongly with those obtained in dogs, where several mutations with large phenotypic effects are present in dogs but not in wolves, implying that these mutations emerged during or after domestication and reached detectable allelic frequencies because they were selected for (Boyko et al., 2010; Larson et al., 2014). We also observed that a relevant fraction of casein diversity is not shared across populations (**Figure 3**). This variation might be represented by mutations that emerged after the domestication and dispersal of goats or may be caused by insufficient sampling or to sequencing errors (because of limited genomic coverage).

The allelic frequencies of the casein genes reported in **Table 1** were consistent with those of previous studies, although it is important to emphasize that such frequencies can be very variable even when comparing breeds reared in the same geographic location; for example, the *CSN1S1* A and E alleles are the most frequent alleles in Italian Saanen and Alpine breeds (Frattini et al., 2014), whereas the most abundant *CSN1S1* allele in Sarda goats is B (Vacca et al., 2014). In the *CSN1S1* gene, we observed that the A-I-N-O1-O2, B2-D-F, and B4-E groups of alleles were well represented in the European and African populations (**Table 1**). Genotyping of *CSN1S1* in French (Grosclaude et al., 1994; Pepin, 1994; Carillier-Jacquin et al., 2016) and Italian (Sacchi et al., 2005; Caroli et al., 2006; Gigli et al., 2008; Mastrangelo et al., 2013; Frattini et al., 2014) goats showed that the A and F alleles are quite abundant, whereas in Spanish (Jordana et al., 1996; Caravaca et al., 2008) and African (Caroli et al., 2007) goats, the B/E and A/B pairs of alleles are the most frequent, respectively. In contrast, the C, G, N, and O1 alleles tend to have low frequencies in Italian (Sacchi et al.,

2005; Caroli et al., 2006; Gigli et al., 2008; Mastrangelo et al., 2013; Frattini et al., 2014) and African breeds (Caroli et al., 2007). According to our results (**Table 1**), the A-I-N-O1-O2 group of alleles is prevalent in Near East and Far East domestic goats as well as in bezoars. A previous study on Indian and Turkish goats (Chessa et al., 2007) reflected the same trend, with high frequencies of the A allele in the majority of the analyzed populations. With regard to the *CSN1S2* and *CSN2* genes, we found that the A and C alleles, respectively, were predominant, a finding consistent with what has been published in European, African, Turkish, and Indian goats (Sacchi et al., 2005; Caroli et al., 2006, 2007; Chessa et al., 2007; Gigli et al., 2008; Vacca et al., 2014; Tortorici et al., 2014; Kusza et al., 2016; Grobler et al., 2017). In the *CSN3* gene, the A and B alleles were predominant in most populations (**Table 1**), as previously published in a broad array of caprine breeds (Yahyaoui et al., 2003; Prinzenberg et al., 2005; Caroli et al., 2007; Chessa et al., 2007; Kiplagat et al., 2010; Di Gerlando et al., 2015). Interestingly, the D allele was frequent in Far East and Near East goats, and the G/L group of alleles was found in Near East goats and bezoars (**Table 1**). The genetic analysis of the *CSN3* locus in Turkish and Indian goats shows that the D allele is relatively frequent, and that the G allele also segregates in these 2 populations but at lower frequencies (Prinzenberg et al., 2005; Chessa et al., 2007). The *CSN3* D and G alleles, in contrast, are rare in African (Caroli et al., 2007) and most European breeds (Yahyaoui et al. 2003; Prinzenberg et al., 2005), with the exception of several Italian populations (Sacchi et al., 2005; Di Gerlando et al., 2015).

In summary, our main finding was that a significant number of casein alleles (or groups of alleles) are present in the bezoar, suggesting that they existed before domestication. Of note, several of the casein alleles detected in the bezoar have been associated with dairy traits in domestic goats. For instance, the A and B alleles of the *CSN1S1* gene determine a high content of α_{S1}-casein in milk and they increase milk protein, casein, and fat contents and improve cheese yield (reviewed in Martin et al., 1999; Moioli et al., 2007; Amills et al., 2012; Amills,

2014). The A allele of the *CSN2* gene and the B allele of the *CSN3* gene are also associated with a higher protein content (Caravaca et al., 2009; Vacca et al., 2014). We found that several *CSN3* polymorphisms that are very rare in European breeds were more frequent in goat populations from other continents, emphasizing the need to investigate their effects on dairy traits.

CONCLUSIONS

The main conclusion of this work is that a relevant fraction of the casein variation segregating in domestic goats probably emerged before the domestication process.

ACKNOWLEDGMENTS

This research was funded by grant AGL2016-76108-R awarded by the Spanish Ministry of Economy and Competitivity (Madrid, Spain). We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain). Dailu Guan was funded by a PhD fellowship from the China Scholarship Council (CSC). Emilio Mármol-Sánchez was funded by a FPU PhD grant from the Spanish Ministry of Education (FPU15/01733). Tainã Figueiredo Cardoso was funded by a fellowship from the CAPES Foundation-Coordination of Improvement of Higher Education, Ministry of Education of the Federal Government of Brazil. The authors thank the CERCA Programme of the Generalitat de Catalunya (Barcelona, Spain) for their support and those who provided publicly available data.

Supplementary Information

Supplementary Figure S1. The strategy used for inferring alleles of *CSN1S1* (a), *CSN2* (b), *CSN1S2* (c) and *CSN3* gene (d), which was based on information provided by Marletta et al. (2007).

Supplementary Figure S2. Principal components analysis (PCA, a) and ADMIXTURE analysis (b) based on 1,221 SNPs mapping to casein genes in bezoars (BE) and four domestic goat populations corresponding to Europe (EU), Africa (AF), Near East (NE) and Far East (FE).

Supplementary Table S1. List of the caprine whole-genome sequences used in the current study

Supplementary Table S2. Polymorphisms (SNPs in black, INDEL in red) detected in the goat casein genes

Supplementary Table S3. Distribution and numbers of SNPs mapping to casein genes identified in a data set of 106 genomes of domestic goats and bezoars

Supplementary Table S4. Classification of the casein SNPs identified in the current work

REFERENCES

Ajmone-Marsan, P., J. F. Garcia, and J. A. Lenstra. 2010. On the origin of cattle: How aurochs became cattle and colonized the world. Evol. Anthropol. 19:148-157. https://doi.org/10.1002/evan.20267.

Alberto, F. J., F. Boyer, P. Orozcoter Wengel, 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. https://doi.org/10.1038/s41467-018-03206-y.

Alexander, D. H., and K. Lange. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246. https://doi.org/10.1186/1471-2105-12-246.

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-1664. https://doi.org/10.1101/gr.094052.109.

Amills, M. 2014. The application of genomic technologies to investigate the inheritance of economically important traits in goats. Adv. Biol. 13. https://doi.org/10.1155/2014/904281.

Amills, M., J. Jordana, A. Zidi, and J. M. Serradilla. 2012. Genetic factors that regulate milk protein and lipid composition in goats. In Milk Production-Advanced Genetic Traits, Cellular Mechanism, Animal Management and Health. N. Chaiyabutr, ed. InTech, Rijeka, Croatia. https://doi.org/10.5772/51716.

Becker, D., M. Otto, P. Ammann, I. Keller, C. Drögemüller, and T. Leeb. 2015. The brown coat colour of Coppernecked goats is associated with a non-synonymous variant at the *TYRP1* locus on chromosome 8. Anim. Genet. 46:50-54. https://doi.org/10.1111/age.12240.

Benjelloun, B., F. J. Alberto, I. Streeter, F. Boyer, E. Coissac, S. Stucki, M. BenBati, M. Ibnelbachyr, M. Chentouf, A. Bechchari, K. Leempoel, A. Alberti, S. Engelen, A. Chikhi, L. Clarke, P. Flicek, S. Joost, P. Taberlet, and F. Pompanon, and NextGen Consortium. 2015. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data. Front. Genet. 6:107. https://doi.org/10.3389/fgene.2015.00107.

Bevilacqua, C., P. Ferranti, G. Garro, C. Veltri, R. Lagonigro, C. Leroux, E. Pietrolà, F. Addeo, F. Pilla, L. Chianese, and P. Martin. 2002. Interallelic recombination is probably responsible for the occurrence of a new α_{S1}-casein variant found in the goat species. Eur. J. Biochem. 269:1293-1303.

Bickhart, D. M., B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisa, F. A. Ponce de Leon, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat. Genet. 49:643-650. https://doi.org/10.1038/ng.3802.

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.

Boyko, A. R., P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt, K. E. Lohmueller, K. Y. Zhao, A. Brisbin, H. G. Parker, B. M. vonHoldt, M. Cargill, A. Auton, A. Reynolds, A. G. Elkahloun, M. Castelhano, D. S. Mosher, N. B. Sutter, G. S. Johnson, J. Novembre, M. J. Hubisz, A. Siepel, R. K. Wayne, C. D. Bustamante, and E. A. Ostrander. 2010. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol. 8:e1000451. https://doi.org/10.1371/journal.pbio.1000451.

Browning, B. L., and S. R. Browning. 2016. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98:116-126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Caravaca, F., M. Amills, J. Jordana, A. Angiolillo, P. Agüera, C. Aranda, A. Menéndez-Buxadera, A. Sánchez, J. Carrizosa, B. Urrutia, A. Sànchez, and J. M. Serradilla. 2008. Effect of α_{S1}-casein (*CSN1S1*) genotype on

milk CSN1S1 content in Malagueña and Murciano-Granadina goats. J. Dairy Res. 75:481-484. https://doi.org/10.1017/S0022029908003609.

Caravaca, F., J. Carrizosa, B. Urrutia, F. Baena, J. Jordana, M. Amills, B. Badaoui, A. Sánchez, A. Angiolillo, and J. M. Serradilla. 2009. Short communication: Effect of α_{S1}-casein (*CSN1S1*) and κ-casein (*CSN3*) genotypes on milk composition in Murciano-Granadina goats. J. Dairy Sci. 92:2960-2964. https://doi.org/10.3168/jds.2008-1510.

Carillier-Jacquin, C., H. Larroque, and C. Robert-Granié. 2016. Including α_{s1} casein gene information in genomic evaluations of French dairy goats. Genet. Sel. Evol. 48:54. https://doi.org/10 .1186/s12711-016-0233-x.

Caroli, A., F. Chiatti, S. Chessa, D. Rignanese, P. Bolla, and G. Pagnacco. 2006. Focusing on the goat casein complex. J. Dairy Sci. 89:3178-3187. https://doi.org/10.3168/jds.S0022-0302(06)72592-9.

Caroli, A., F. Chiatti, S. Chessa, D. Rignanese, E. M. Ibeagha-Awemu, and G. Erhardt. 2007. Characterization of the casein gene complex in West African goats and description of a new α_{S1}-casein polymorphism. J. Dairy Sci. 90:2989-2996. https://doi.org/10.3168/jds.2006-674.

Chessa, S., F. Chiatti, D. Rignanese, E. M. Ibeagha-Awemu, C. Özbeyaz, Y. A. Hassan, M. M. Baig, G. Erhardt, and A. Caroli. 2007. The casein genes in goat breeds from different continents: Analysis by polymerase chain reaction-single strand conformation polymorphism (PCR-SSCP). Ital. J. Anim. Sci. 6:73-75. https://doi .org/10 .4081/ijas.2007.1s.73.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92. https://doi.org/10.4161/fly.19695.

Daly, K. G., P. Maisano Delser, V. E. Mullin, A. Scheu, V. Mattiangeli, M. D. Teasdale, A. J. Hare, J. Burger, M. P. Verdugo, M. J. Collins, R. Kehati, C. M. Erek, G. BarOz, F. Pompanon, T. Cumer, C. Çakırlar, A. F. Mohaseb, D. Decruyenaere, H. Davoudi, Ö. Çevik, G. Rollefson, J.D. Vigne, R. Khazaeli, H. Fathi, S. B. Doost, R. Rahimi Sorkhani, A. A. Vahdati, E. W. Sauer, H. Azizi Kharanaghi, S. Maziar, B. Gasparian, R. Pinhasi, L. Martin, D. Orton, B. S. Arbuckle, N. Benecke, A. Manica, L. K. Horwitz, M. Mashkour, and D. G. Bradley. 2018. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. Science 361:85-88. https://doi.org/10.1126/science.aas9411.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin., and 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158. https://doi.org/10.1093/bioinformatics/btr330.

Di Gerlando, R., L. Tortorici, M. T. Sardina, G. Monteleone, S. Mastrangelo, and B. Portolano. 2015. Molecular characterisation of κ-casein gene in Girgentana dairy goat breed and identification of two new alleles. Ital. J. Anim. Sci. 14:3464. https://doi.org/10.4081/ijas.2015.3464.

Fang, H., Y. Y. Wu, G. Narzisi, J. A. O'Rawe, L. T. J. Barron, J. Rosenbaum, M. Ronemus, I. Iossifov, M. C. Schatz, and G. J. Lyon. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 6:89. https://doi.org/10.1186/s13073-014-0089-z.

Frattini, S., L. Nicoloso, B. Coizet, S. Chessa, L. Rapetti, G. Pagnacco, and P. Crepaldi. 2014. Short communication: The unusual genetic trend of α_{S1}-casein in Alpine and Saanen breeds. J. Dairy Sci. 97:7975-7979. https://doi.org/10.3168/jds.2014-7780.

Gigli, I., D. O. Maizon, V. Riggio, M. T. Sardina, and B. Portolano. 2008. Short communication: Casein haplotype variability in Sicilian dairy goat

breeds. J. Dairy Sci. 91:3687-3692. https://doi.org/10.3168/jds.2008-1067.

Grobler, R., C. Visser, S. Chessa, and E. van Marle-Köster. 2017. Genetic polymorphism of *CSN1S2* in South African dairy goat populations. S. Afr. J. Anim. Sci. 47:72-78. https://doi.org/10.4314/sajas.v47i1.11.

Grosclaude, F., G. Ricordeau, P. Martin, F. Remeuf, L. Vassal, and J. Bouillon. 1994. Du gène au fromage: Le polymorphisme de la caséine α_{S1} caprine, ses effets, son évolution. INRA Prod. Anim. 7:3-19.

Guan, D., N. Luo, X. Tan, Z. Zhao, Y. Huang, R. Na, J. Zhang, and Y. Zhao. 2016. Scanning of selection signature provides a glimpse into important economic traits in goats (*Capra hircus*). Sci. Rep. 6:36372. https://doi.org/10.1038/srep36372.

Hasan, M. S., X. Wu, and L. Zhang. 2015. Performance evaluation of indel calling tools using real short-read data. Hum. Genomics 9:20. https://doi.org/10.1186/s40246-015-0042-2.

Hayes, B., N. Hagesæther, T. Ådnøy, G. Pellerud, P. R. Berg, and S. Lien. 2006. Effects on production traits of haplotypes among casein genes in Norwegian goats and evidence for a site of preferential recombination. Genetics 174:455. https://doi.org/10.1534/genetics.10.058966.

Jordana, J., M. Amills, E. Díaz, C. Angulo, J. M. Serradilla, and A. Sànchez. 1996. Gene frequencies of caprine α_{s1}-casein polymorphism in Spanish goat breeds. Small Rumin. Res. 20:215-221. https://doi.org/10.1016/0921-4488(95)00813-6.

Khalilzadeh, P., H. R. Rezaei, D. Fadakar, M. Serati, M. Aliabadian, J. Haile, and H. Goshtasb. 2016. Contact zone of Asian and European wild boar at North West of Iran. PLoS One 11:e0159499. https://doi.org/10.1371/journal.pone.0159499.

Kiplagat, S. K., M. Agaba, I. S. Kosgey, A. M. Okeyo, D. Indetie, O. Hanotte, and M. K. Limo. 2010. Genetic polymorphism of kappa-casein gene in indigenous Eastern Africa goat populations. Int. J. Genet. Mol. Biol. 2:1-5.

Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870-1874. https://doi.org/10.1093/molbev/msw054.

Kusza, S., D. E. Ilie, M. Sauer, K. Nagy, I. Patras, and D. Gavojdian. 2016. Genetic polymorphism of *CSN2* gene in Banat White and Carpatina goats. Acta Biochim. Pol. 63:577-580. https://doi.org/10.18388/abp.2016_1266.

Larson, G., D. R. Piperno, R. G. Allaby, M. D. Purugganan, L. Andersson, M. Arroyo-Kalin, L. Barton, C. C. Vigueira, T. Denham, K. Dobney, A. N. Doust, P. Gepts, M. T. P. Gilbert, K. J. Gremillion, L. Lucas, L. Lukens, F. B. Marshall, K. M. Olsen, J. C. Pires, P. J. Richerson, R. R. de Casas, O. I. Sanjur, M. G. Thomas, and D. Q. Fuller. 2014. Current perspectives and the future of domestication studies. Proc. Natl. Acad. Sci. USA 111:6139-6146. https://doi.org/ 10 .1073/pnas.1323964111.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303 .3997v2 [q-bio.GN].

Li, X., R. Su, W. Wan, W. Zhang, H. Jiang, X. Qiao, Y. Fan, Y. Zhang, R. Wang, Z. Liu, Z. Wang, B. Liu, Y. Ma, H. Zhang, Q. Zhao, T. Zhong, R. Di, Y. Jiang, W. Chen, W. Wang, Y. Dong, and J. Li. 2017. Identification of selection signals by large-scale whole-genome resequencing of cashmere goats. Sci. Rep. 7:15142. https://doi.org/10.1038/s41598-017-15516-0.

Marletta, D., A. Criscione, S. Bordonaro, A. M. Guastella, and G. D'Urso. 2007. Casein polymorphism in goat's milk. Lait 87:491-504.

Martin, P., M. Ollivier-Bousquet, and F. Grosclaude. 1999. Genetic polymorphism of caseins: A tool to investigate casein micelle organization. Int. Dairy J. 9:163-171. https://doi .org/10.1016/S0958-6946(99)00055-2.

Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reprod. Nutr. Dev. 42:433-459. https://doi.org/10.1051/rnd:2002036.

Mastrangelo, S., M. T. Sardina, M. Tolone, and B. Portolano. 2013. Genetic polymorphism at the *CSN1S1* gene in Girgentana dairy goat breed. Anim. Prod. Sci. 53:403-406. https://doi.org/10.1071/AN12242.

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. https://doi.org/10 .1101/gr.107524.110.

Menzi, F., I. Keller, I. Reber, J. Beck, B. Brenig, E. Schutz, T. Leeb, and C. Drogemuller. 2016. Genomic amplification of the caprine *EDNRA* locus might lead to a dose dependent loss of pigmentation. Sci. Rep. 6:28438. https://doi .org/10 .1038/srep28438.

Moioli, B., M. D'Andrea, and F. Pilla. 2007. Candidate genes affecting sheep and goat milk quality. Small Rumin. Res. 68:179-192. https://doi .org/10 .1016/j.smallrumres.2006.09.008.

Naderi, S., H. R. Rezaei, F. Pompanon, M. G. Blum, R. Negrini, H. R. Naghash, O. Balkiz, M. Mashkour, O. E. Gaggiotti, P. Ajmone-Marsan, A. Kence, J. D. Vigne, and P. Taberlet. 2008. The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. Proc. Natl. Acad. Sci. USA 105:17659-17664. https://doi.org/10.1073/pnas.0804782105.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. 2013. Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med. 5:28.

Pepin, L. 1994. Recherche de polymorphisme genetique chez les caprins. Applications à l'étude de la diversité des populations, au contrôle de filiation et à la résistance génétique à la cowdriose. Doctoral Thesis. Universite de Paris XI Orsay, France.

Pereira, F., and A. Amorim. 2010. Origin and spread of goat pastoralism. In Encyclopedia of Life Sciences (eLS). John Wiley & Sons, Ltd., Chichester, UK. https://doi.org/10.1002/9780470015902.a0022864.

Prinzenberg, E. M., K. Gutscher, S. Chessa, A. Caroli, and G. Erhardt. 2005. Caprine κ-casein (*CSN3*) polymorphism: New developments in molecular knowledge. J. Dairy Sci. 88:1490-1498. https://doi.org/10.3168/jds.S0022-0302(05)72817-4.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based analyses. Am. J. Hum. Genet. 81:559-575. https://doi.org/10.1086/519795.

Reber, I., I. Keller, D. Becker, C. Flury, M. Welle, and C. Drogemuller. 2015. Wattles in goats are associated with the FMN1/GREM1 region on chromosome 10. Anim. Genet. 46:316-320. https://doi.org/10.1111/age.12279.

Remeuf, F., V. Cossin, C. Dervin, J. Lenoir, and R. Tomassone. 1991. Relationships between physicochemical characteristics of milks and their cheese-making properties. Lait 71:397-421.

Rijnkels, M. 2002. Multispecies comparison of the casein gene loci and evolution of casein gene family. J. Mammary Gland Biol. Neoplasia 7:327-345. https://doi.org/10.1023/A:1022808918013.

Robasky, K., N. E. Lewis, and G. M. Church. 2014. The role of replicates for error mitigation in next-generation sequencing. Nat. Rev. Genet. 15:56-62. https://doi.org/10.1038/nrg3655.

Sacchi, P., S. Chessa, E. Budelli, P. Bolla, G. Ceriotti, D. Soglia, R. Rasero, E. Cauvin, and A. Caroli. 2005. Casein haplotype structure in five Italian goat breeds. J. Dairy Sci. 88:1561-1568. https://doi.org/10.3168/jds.S0022-0302(05)72825-3.

Tortorici, L., R. Di Gerlando, S. Mastrangelo, M. T. Sardina, and B. Portolano. 2014. Genetic characterisation of *CSN2* gene in Girgentana goat breed. Ital. J. Anim. Sci. 13:3414. https://doi.org/10.4081/ijas.2014.3414.

Vacca, G. M., M. L. Dettori, G. Piras, F. Manca, P. Paschino, and M. Pazzola. 2014. Goat casein genotypes are associated with milk production traits in the Sarda breed. Anim. Genet. 45:723-731. https://doi.org/10.1111/age.12188.

Vaser, R., S. Adusumalli, S. N. Leng, M. Sikic, and P. C. Ng. 2016. SIFT missense predictions for genomes. Nat. Protoc. 11:1. https://doi.org/10.1038/nprot.2015.123.

Wang, X., J. Liu, G. Zhou, J. Guo, H. Yan, Y. Niu, Y. Li, C. Yuan, R. Geng, X. Lan, X. An, X. Tian, H. Zhou, J. Song, Y. Jiang, and Y. Chen. 2016. Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. Sci. Rep. 6:38932. https://doi.org/10.1038/srep38932.

Yahyaoui, M. H., A. Angiolillo, F. Pilla, A. Sànchez, and J. M. Folch. 2003. Characterization and genotyping of the caprine κ-casein variants. J Dairy Sci. 86:2715-2720. https://doi.org/10.3168/jds.S0022-0302(03)73867-3.

Zeder, M. A., and B. Hesse. 2000. The initial domestication of goats (*Capra hircus*) in the Zagros Mountains 10,000 years ago. Science 287:2254-2257. https://doi.org/ 10.1126/science.287.5461.2254.

Zhao, Y. X., J. Yang, F. Lv, X. Hu, X. Xie, M. Zhang, W. Li, M. Liu, Y. Wang, J. Li, Y. Liu, Y. Ren, F.Wang, E. Hehua, J. Kantanen, J. Arjen Lenstra, J. Han, and M. Li. 2017. Genomic reconstruction of the history of native sheep reveals the peopling patterns of nomads and the expansion of early pastoralism in East Asia. Mol. Biol. Evol. 34:2380-2395. https://doi .org/10 .1093/molbev/msx181.

GSE Genetics Selection Evolution

A genome-wide analysis of copy number variation in Murciano-Granadina goats

Dailu Guan[1], Amparo Martínez[2], Anna Castelló[1,3], Vincenzo Landi[2,4], María Gracia Luigi-Sierra[1], Javier Fernández Álvarez[5], Betlem Cabrera[1,3], Juan Vicente Delgado[2], Xavier Such[6], Jordi Jordana[3], Marcel Amills[1,3]*

[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain. [2]Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain. [3]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [4]Department of Veterinary Medicine, University of Bari "Aldo Moro", SP. 62 per Casamassima km. 3, 70010 Valenzano (BA), Italy. [5]Asociación Nacional de Criadores de Caprino de Raza Murciano-Granadina (CAPRIGRAN), 18340 Granada, Spain. [6]Group of Research in Ruminants (G2R), Department of Animal and Food Science, Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona, Spain

*Corresponding author: Marcel Amills (marcel.amills@uab.cat)

Genetics Selection Evolution 52: 44

https://doi.org/10.1186/s12711-020-00564-4

Abstract

Background: In this work, our aim was to generate a map of the copy number variations (CNV) segregating in a population of Murciano-Granadina goats, the most important dairy breed in Spain, and to ascertain the main biological functions of the genes that map to copy number variable regions.

Results: Using a dataset that comprised 1,036 Murciano-Granadina goats genotyped with the Goat SNP50 BeadChip, we were able to detect 4,617 and 7,750 autosomal CNV with the PennCNV and QuantiSNP software, respectively. By applying the EnsembleCNV algorithm, these CNV were assembled into 1,461 CNV regions (CNVR), of which 486 (33.3% of the total CNVR count) were consistently called by PennCNV and QuantiSNP and used in subsequent analyses. In this set of 486 CNVR, we identified 78 gain, 353 loss and 55 gain/loss events. The total length of all the CNVR (95.69 Mb) represented 3.9% of the goat autosomal genome (2,466.19 Mb), whereas their size ranged from 2.0 kb to 11.1 Mb, with an average size of 196.89 kb. Functional annotation of the genes that overlapped with the CNVR revealed an enrichment of pathways related with olfactory transduction (fold-enrichment = 2.33, q-value = 1.61×10^{-10}), ABC transporters (fold-enrichment = 5.27, q-value = 4.27×10^{-04}) and bile secretion (fold-enrichment = 3.90, q-value = 5.70×10^{-03}).

Conclusions: A previous study reported that the average number of CNVR per goat breed was ~20 (978 CNVR/50 breeds), which is much smaller than the number we found here (486 CNVR). We attribute this difference to the fact that the previous study included multiple caprine breeds that were represented by small to moderate numbers of individuals. Given the low frequencies of CNV (in our study, the average frequency of CNV is 1.44%), such a design would

probably underestimate the levels of the diversity of CNV at the within-breed level. We also observed that functions related with sensory perception, metabolism and embryo development are overrepresented in the set of genes that overlapped with CNV, and that these loci often belong to large multigene families with tens, hundreds or thousands of paralogous members, a feature that could favor the occurrence of duplications or deletions by non-allelic homologous recombination.

Background

Copy number variations (CNV) encompass genomic deletions or duplications, with sizes ranging from 50 base pairs (bp) to several megabases (Mb), and which display polymorphisms (in terms of copy number) among individuals of a particular species [1, 2, 3]. In livestock, a broad array of phenotypes related with, among others, morphology [4, 5], pigmentation [6, 7, 8, 9], sexual development [10] and susceptibility to disease [11] is caused by the segregation of CNV. Genome scans to detect structural variations in cattle have revealed that CNV regions (CNVR) are often enriched in genes that are involved in immunity [12, 13, 14, 15], metabolism [12, 13], embryo development [12, 15] and sensory perception [13, 14]. There is evidence that the d_N/d_S ratios of genes that map to taurine CNV are generally higher than those of genes that do not overlap with CNV, which indicates that CNV genes probably evolve under reduced selective constraint [13]. The analysis of gene networks has also shown that genes that co-localize with duplications tend to have fewer interactions with other genes than loci that do not overlap with CNV, reinforcing the idea that genes mapping to duplicated regions have fewer essential housekeeping functions than non-CNV genes, and also have reduced pleiotropy [13].

Although structural chromosomal variations can have strong effects on gene expression and phenotypic variability, technical limitations and the moderate quality of genome assemblies have hampered CNV mapping in livestock [1]. Until recently, this has been particularly true for goats. In 2010, Fontanesi et al. [16] published the first caprine CNV map by identifying, with the Bovine 385 k aCGH array, 127 CNVR including 86 and 41 copy loss and gain variants, respectively. Later on, resequencing the genome of individuals from several caprine breeds made it possible to identify CNV that overlap with 13 pigmentation genes and to detect an association between increased *ASIP* copy number and light pigmentation [17]. The first worldwide survey of copy number variation in goats was performed within the Goat ADAPTmap Project (http://www.goatadaptmap.org), and involved the genome-wide genotyping of 1,023 goats from 50 breeds [18]. This study resulted in the identification of 978 CNVR among which several overlapped with genes that are functionally related with local adaptation such as coat color, muscle development, metabolic processes, and embryonic development [18]. Moreover, the patterns of the diversity of CNV differed according to geographic origin, which indicates that they have been influenced by population history [18]. In another study on 433 individuals from 13 East African goat breeds, Nandolo et al. [19] detected 325 CNVR. More recently, Henkel et al. [8] demonstrated the existence of complex patterns of structural variation in the regions containing the caprine *ASIP* and *KIT* genes, with potential causal effects on pigmentation. In spite of these efforts, the description of structural chromosomal variation in goats is still lagging behind that of other domestic species. Most of the CNV surveys in goats have analyzed large populations that represent a mixture of different breeds each with a limited number of individuals [18, 19], thus making it difficult to assess the magnitude of the CNV diversity at the within-breed level. Our goal was to fill this gap by analyzing a population of 1,036 individuals from a single Spanish breed (Murciano-Granadina), and to investigate the functional roles of genes that

map to CNVR and compare these results with data obtained in composite goat populations.

Methods

Genomic DNA extraction and high-throughput genotyping

Blood samples from 1,036 Murciano-Granadina female goats from 15 farms that are connected through the use of artificial insemination were collected in EDTA K3 coated vacuum tubes and stored at -20 °C before processing. Genomic DNA was isolated by a modified salting-out procedure [20]. Four volumes of red cell lysis solution (Tris-HCl 10 mmol/L, pH = 6.5; EDTA 2 mmol/L; Tween 20 1%) were added to 3 mL of whole blood, and this mixture was centrifuged at 2000×g. Pelleted cells were resuspended in 3 mL lysis buffer (Tris-HCl 200 mmol/L, pH = 8, EDTA 30 mmol/L, SDS 1%; NaCl 250 mmol/L) plus 100 µL proteinase K (20 mg/mL). The resulting mixture was incubated at 55 °C for 3 h followed by centrifugation at 2000×g in the presence of 1 mL of ammonium acetate (10 mol/L). The supernatant (~4 mL) was mixed with 3 mL of isopropanol 96%, which was subsequently centrifuged at 2000×g for 3 min. The supernatant was removed and the DNA pellet was washed with 3 mL of ethanol 70%. After centrifuging at 2000×g for 1 min, the DNA precipitate was dried at room temperature and resuspended in 1 mL of TE buffer (10 mmol/L Tris, pH = 8.0; 1 mmol/L EDTA, pH = 8).

High-throughput genotyping of the 1,036 Murciano-Granadina DNA samples was carried out with the Goat SNP50 BeadChip [21] according to the manufacturer's instructions (Illumina). Signal intensity ratios i.e. log R Ratio or LRR (the total probe intensity of a SNP referred to a canonical set of normal

controls [22]), and B allele frequencies or BAF (relative quantity of one allele compared to the other one) [22], were exported for each single nucleotide polymorphism (SNP) with the GenomeStudio software 2.0.4 (Illumina, https://emea.illumina.com). Then, SNP coordinates were converted to the latest version of the goat reference genome (ARS1) [23]. After filtering out unmapped and non-autosomal SNPs and those with a call rate lower than 98%, a set of 50,551 SNPs remained for CNV mapping.

Copy number variant calling with PennCNV and QuantiSNP

Based on their excellent performance in comparative studies, we selected two software packages, PennCNV v1.0.5 [24] and QuantiSNP v2 [25], to call CNV in the Murciano-Granadina population [26, 27]. The PennCNV software [24] detects CNV by applying the default parameters of the Hidden-Markov model. Population frequencies of B alleles were compiled based on the BAF of each SNP in the population. We used the "--gcmodelfile" option to adjust "genomic waves" [28]. The number of goat chromosomes was set with the "--lastchr 29" instruction. The QuantiSNP analysis [25] assumes an objective Bayes hidden-Markov model to improve the accuracy of segmental aneuploidy identification and mapping. This CNV calling software was run under default parameters by modifying the "--chr 1:29" option. The CNV that were supported by less than three SNPs were removed from the filtered set used here.

Definition and functional annotation of copy number variant regions

We used the EnsembleCNV algorithm (beta version) [29] to assemble CNVR. All CNV called by PennCNV and/or QuantiSNP were combined to generate a set of initial CNVR by using the heuristic algorithm (threshold of minimum

overlap = 30%) described in [29]. Subsequently, CNVR boundaries were refined by considering the local correlation structure of the LRR values of the SNPs mapping to CNVR [29]. Then, we reassigned the CNV calls that were initially obtained with PennCNV and QuantiSNP to each refined CNVR, so that the final set of CNVR comprised only those that were simultaneously detected by both callers. The resulting CNVR were matched to gene features that are annotated in the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) by using BEDTools v2.25.0 [30]. In addition, we performed gene ontology (GO) enrichment and pathway analyses using the DAVID Bioinformatics Resources 6.8 [31, 32] based on human and goat background gene sets. The statistical significance was set to a q-value ≤ 0.05.

Confirmation of copy number variant regions by quantitative real-time PCR

In order to evaluate the rate of false positives in our experiment, we conducted quantitative real-time PCR (qPCR) experiments to obtain an independent estimate of the copy number of four putative CNVR (CNVR_371_chr5, CNVR_506_chr6, CNVR_160_chr2 and CNVR_1229_chr21). Primers were designed with the Primer Express software (Applied Biosystems) to amplify specific regions of the *ADAMTS20*, *BST1*, *NCKAP5* and *TNFAIP2* genes (see **Additional file 1: Table S1**). As reference genes, we used the melanocortin 1 receptor (*MC1R*) and glucagon (*GCG*) genes (see **Additional file 1: Table S1**) loci [18, 33, 34, 35]. Quantitative PCR reactions contained 7.5 ng genomic DNA, 7.5 µL 2 × SybrSelect Master mix (Applied Biosystems), 4.5 pmol of each forward and reverse primer, and ultrapure water to a maximum final volume of 15 µL. Each sample was analyzed in triplicate in order to obtain averaged copy number estimates. Reactions were loaded onto 384-well plates and run in a QuantStudio 12 K Flex Real-Time PCR System instrument (Applied

Biosystems). The specificity of the PCR reactions was evaluated with a melting curve analysis procedure, and the efficiency (96.2-105.4%) was assessed with standard curves. Thus, relative copy number was inferred with the qbase + software (Biogazelle, Ghent, Belgium) by using the $2^{-\Delta\Delta Ct}$ approach [36]. Copy number values were calibrated by taking as a reference, four samples which, according to Goat SNP50 BeadChip data, had two copies of the investigated genomic loci.

Results

Detection of copy number variation in Murciano-Granadina goats

The initial calling with PennCNV and QuantiSNP yielded 4,617 and 7,750 autosomal CNV, respectively. By using the EnsembleCNV tool [29], we assigned these CNV into 1,461 CNVR with refined boundaries, of which 486 (33.3% of the total CNVR count) were detected simultaneously by PennCNV and QuantiSNP. The resulting CNVR included 78 copy gain, 353 copy loss and 55 copy gain/loss variants (**Figure 1**, and **Table 1**) and (see **Additional file 2: Table S2**). The total length of the CNVR covered 95.69 Mb (3.9%) of the goat autosomal genome (2,466.19 Mb), whereas their individual size ranged from 2.0 kb to 11.1 Mb, with an average of 196.9 kb (**Figure 2a** and **Table 1**). Moreover, we found that 72.6% of the CNVR showed minimum allele frequencies lower than 0.01, with an average frequency of 1.44% (**Figure 2b**). In addition, 10 CNVR with frequencies higher than 10% were distributed over seven caprine chromosomes. With a frequency of 41%, CNVR_1229_chr21 was the CNVR with the highest frequency in the whole dataset (see **Additional file 2: Table S2**). By using the BEDTools v2.25.0 program [30], 212 of the CNVR that we detected overlapped with 191 unique CNVR published by Liu et al. [18] (**Figure**

1) and (see **Additional file 2: Table S2**). The CNVR that were detected in both studies are referred to as "shared CNVR", whereas those that were identified in our study only are referred to as "non-shared CNVR" (**Figure 1**). Six of the ten "shared CNVR" with frequencies higher than 0.1 show positional concordance with six CNVR detected by Liu et al. [18] (see **Additional file 2: Table S2**).

Table 1. Main features of copy number variation regions (CNVR) detected in 1,036 Murciano-Granadina goats

Summary statistics	Total	Gain	Loss	Gain/Loss
Total length (Mb)	95.69	26.52	61.17	8
Total number of CNVR	486	78	353	55
Number of CNVR (< 10 kb)	1	1	0	0
Number of CNVR (10-50 kb)	4	2	1	1
Number of CNVR (50-100 kb)	152	25	113	14
Number of CNVR (100-500 kb)	313	47	227	39
Number of CNVR (500 kb-1 Mb)	10	0	9	1
Number of CNVR (≥ 1Mb)	6	3	3	0
Average number of SNP per CNVR	5.59	9.01	5.03	4.35
Minimum size of CNVR (kb)	2.04	2.04	23.2	43.1
Maximum size of CNVR (kb)	11,124	11,124	1,629.39	534.16
Average CNVR size (kb)	196.89	339.99	173.28	145.49
Standard deviation of CNVR size (kb)	539.35	1299.49	156.89	91.51

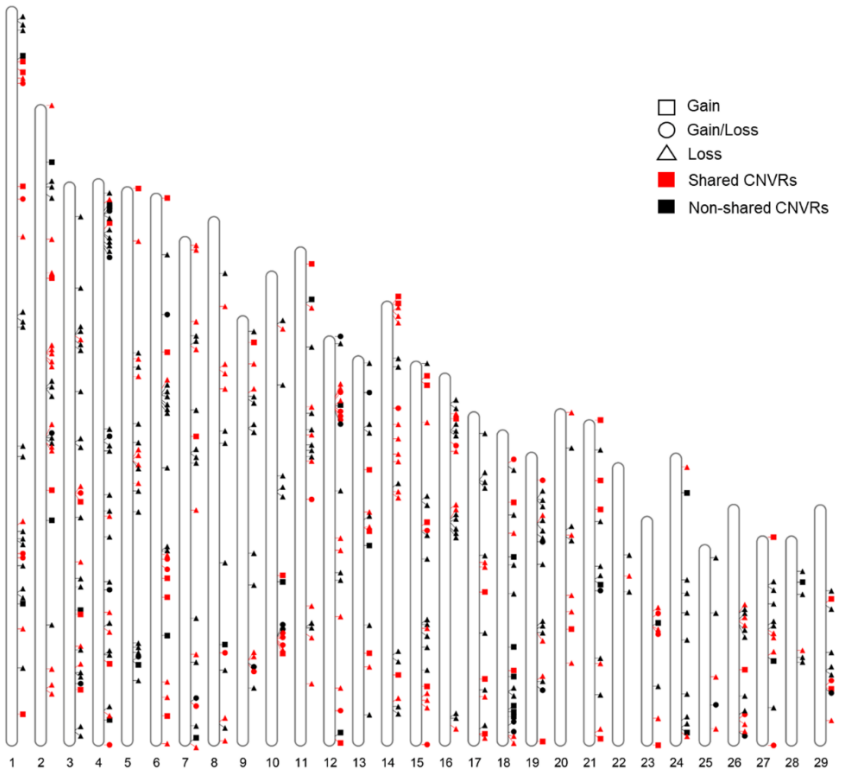

Figure 1. Genomic distribution of 486 CNVR detected with the PennCNV and QuantiSNP software on the 29 caprine autosomes. Squares, triangles and circles represent copy number gain, loss and gain/loss events, respectively. Red and black colors represent shared and non-shared CNVR, respectively. Shared CNVR are those detected both in our study and in Liu et al. [18], while non-shared CNVR are those identified only in our study.

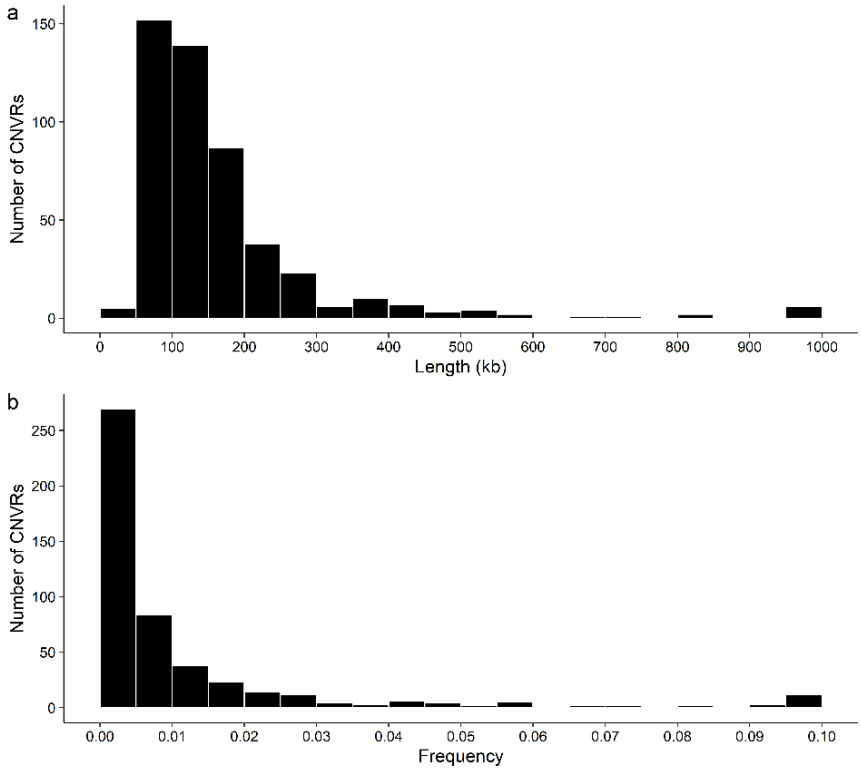

Figure 2. Histograms displaying the distribution of CNVR according to their size (**a**) and frequency (**b**). CNVR that were longer than 1000 kb were included in the 1000-kb bin, whereas those with frequencies above 0.1 were grouped in the 0.1 bin. The histograms were drawn by using the ggplot2 package (http://ggplot2.tidyverse.org/) implemented in R (https://www.r-project.org/).

Functional annotation of the genes that are located in copy number variable regions

Within the CNVR defined in our study, we detected 779 protein-coding genes according to the goat reference genome annotation (ARS1) [23] from the NCBI database (see **Additional file 2: Table S2** and **Additional file 3: Table S3**). In a survey of the diversity of CNV in goats with a worldwide distribution, Liu et al. [18] detected 1,437 copy number variable genes, of which 116 were also

identified in our study and are referred to as "shared copy number variable genes" (see **Additional file 3: Table S3**). Among the "shared copy number variable genes", the *ASIP* and *ADAMTS20* genes are particularly relevant: they are involved in pigmentation [6, 8, 17, 35, 37, 38, 39] and co-localize with selection signals detected in a worldwide sample of goats [40]. In addition, we found that about 11.4% (89) of the annotated genes that co-localize with CNVR are olfactory receptors or olfactory receptor-like genes (see **Additional file 3: Table S3**). Consistently, the most significantly enriched pathway was "Olfactory transduction" (q-value = 1.61×10^{-10}, **Table 2**), followed by "ABC transporter" (q-value = 4.27×10^{-4}, **Table 2**). A significant pathway related with immunity (i.e. Fc epsilon RI signaling, q-value = 0.02) was also identified based on a human background gene set (**Table 2**). Several overrepresented GO terms were related with embryonic skeletal system morphogenesis (q-value = 1.22×10^{-3}) and G-protein coupled purinergic nucleotide receptor activity (q-value = 6.22×10^{-3}, **Table 2**). Interestingly, the copy number variable genes were also enriched in pathways with metabolic significance, such as prolactin signaling and insulin signaling, as well as GO terms related with feeding behavior, but none of these pathways reached the significance threshold (q-value ≤ 0.05) after correction for multiple testing (see **Additional file 4: Table S4**). Several of the pathways outlined in **Additional file 4: Table S4** play important roles in immunity (e.g. chemokine signaling, B cell receptor signaling and T cell receptor signaling), cancer (e.g. endometrial cancer, proteoglycans in cancer, thyroid cancer), as well as in oncogenic signaling (e.g. Ras and ErbB signaling) (see **Additional file 4: Table S4**), but most of them are not significant after correction for multiple testing.

Table 2. Functional enrichment of genes co-localizing with CNVR detected in 1,036 Murciano-Granadina goats

Background gene set	Category	ID	Term	Number of genes	Fold Enrichment	*P* value	*q*-value
Goat	KEGG	chx04740	Olfactory transduction	69	2.33	1.26E-11	1.61E-10
Goat	KEGG	chx02010	ABC transporters	11	5.27	3.33E-05	4.27E-04
Goat	KEGG	chx04976	Bile secretion	11	3.90	4.46E-04	5.70E-03
Human	KEGG	hsa04664	Fc epsilon RI signaling pathway	8	4.71	1.40E-03	1.76E-02
Human	GO/BP	GO:0009952	Anterior/posterior pattern specification	12	5.56	9.36E-06	1.61E-04
Human	GO/BP	GO:0048704	Embryonic skeletal system morphogenesis	8	7.60	7.13E-05	1.22E-03
Human	GO/BP	GO:0035589	G-protein coupled purinergic nucleotide receptor signaling pathway	5	13.24	4.18E-04	7.16E-03
Human	GO/CC	GO:0016020	Membrane	81	1.40	1.45E-03	1.98E-02
Human	GO/MF	GO:0003677	DNA binding	67	1.48	1.10E-03	1.60E-02
Human	GO/MF	GO:0045028	G-protein coupled purinergic nucleotide receptor activity	5	13.19	4.24E-04	6.22E-03

KEGG: Kyoto encyclopedia of genes and genomes pathway; GO/MF: gene ontology (GO) term related with molecular function; GO/BP: GO term related with biological process; GO/CC: GO term related with cellular component.

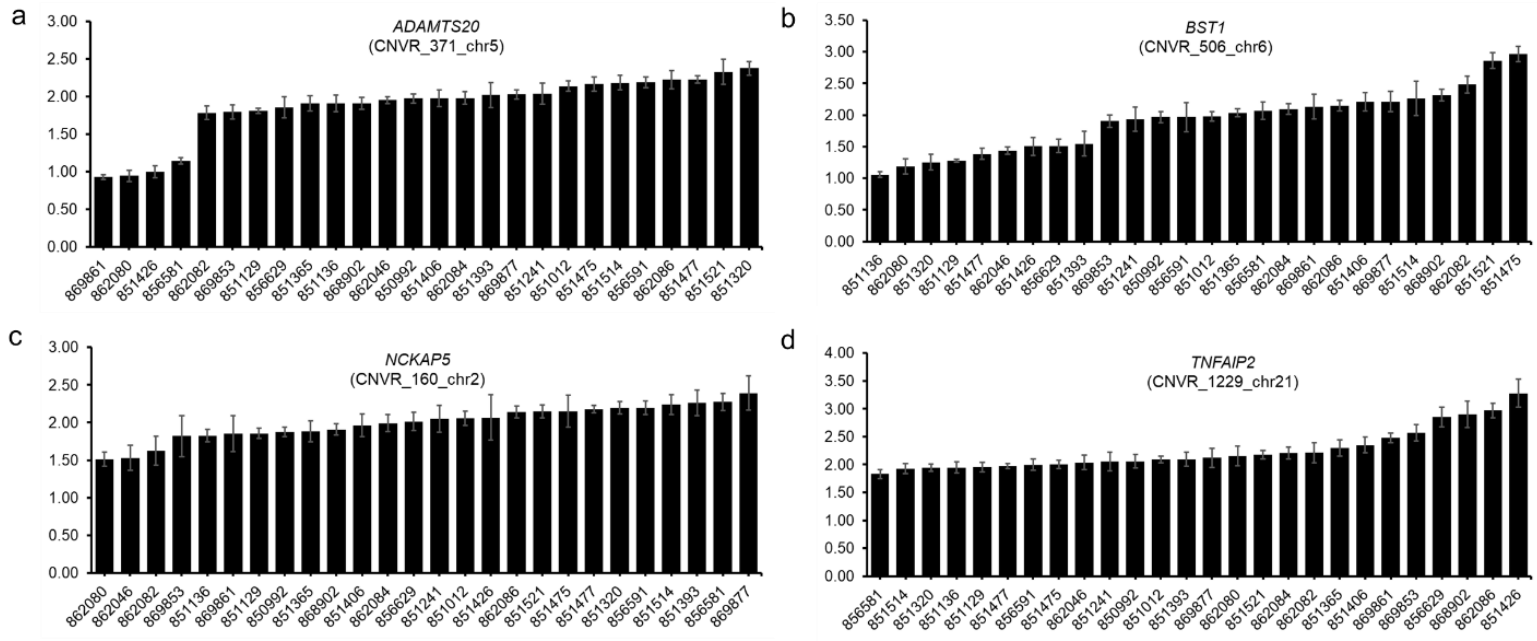

Figure 3. Relative quantification of four copy number variation regions by real-time quantitative polymerase chain reaction analysis: a CNVR_371_chr5 (*ADAMTS20*), b CNVR_506_chr6 (*BST1*), c CNVR_160_chr2 (*NCKAP5*), d CNVR_1229_chr21 (*TNFAIP2*). The *x* and *y* axes represent sample ID and relative quantification of CNVR (mean ± standard error, with each sample analyzed in triplicate), respectively. As calibrator, we used the average of four samples estimated to have two copies (diploid status) based on the Goat SNP50 BeadChip analysis.

Validation of four copy number variants by real-time quantitative polymerase chain reaction

In order to confirm our results, we selected four CNVR (i.e. CNVR_371_chr5, CNVR_506_chr6, CNVR_160_chr2 and CNVR_1229_chr21) that co-localized with the *ADAMTS20*, *BST1*, *NCKAP5* and *TNFAIP2* genes, respectively (the primers used to amplify these CNVR are listed in **Additional file 1: Table S1**). As shown in **Figure 3**, the estimated copy numbers obtained by qPCR analysis of Murciano-Granadina goat samples were: 0.93 to 2.38 copies relative to the calibrator (*ADAMTS20*), 1.06 to 2.96 copies (*BST1*), 1.51 to 2.39 copies (*NCKAP5*) and 1.83 to 3.28 copies (*TNFAIP2*). According to D'haene et al. [36], copy number estimates between 1.414 and 2.449 most likely correspond to a normal copy number of 2, whereas any number below or above these thresholds could represent a deletion or a duplication, respectively. Thus, based on these values, evidence of copy number variation was inferred for three of the four genes analyzed by qPCR.

Discussion

In this work, our aim was to characterize copy number variation in Murciano-Granadina goats, a native Spanish breed used for milk production. By genotyping 1,036 Murciano-Granadina goats with a SNP array, we were able to identify 486 CNVR covering 3.9% of the goat genome, whereas Liu et al. [18] identified CNVR that covered ~9% of the goat genome. The latter higher percentage reported by Liu et al. [18] can be explained by the fact that they analyzed 50 breeds with different geographical origins, i.e. a composite population that is probably much more diverse than that used in our work. Besides, the pipeline

that we used to identify CNVR is more stringent than that employed by Liu et al. [18], removing CNVR that were not consistently detected by PennCNV and QuantiSNP. In the literature, estimates of 4.8 to 9.5% for CNVR coverage in the human genome are reported [2]. Our results and those obtained by Liu et al. [18] are consistent with these values.

Indeed, when Liu et al. [18] calculated the CNVR length for each breed normalized by the goat genome size, their results agreed well with our estimate of 3.9%. For instance, this parameter reached values of 3.94% in goats from Southeastern Africa and 3.13% in goats from Northwestern Africa and Eastern Mediterranean, whereas it was lowest (0.70%) for individuals from West Asia [18]. The number of CNV detected at the within-breed level by Liu et al. [18] was on average 126 CNV per breed and ranged from 6 to 714, whereas the average number of CNVR was only ~20 per breed [18]. Since the number of detected CNVR is proportional to population size, for most of the breeds investigated in [18], the level of within-breed CNV variation is probably underestimated. In summary, one important conclusion from our study is that the magnitude of CNV diversity at the within-breed level is likely to be much larger than that previously reported in studies that analyzed multiple populations, each represented by a small or moderate number of individuals.

Most of the CNVR that we report here ranged in size from 50 to 500 kb, with a mean size of 196.89 kb. Similarly, the average CNVR size reported by Liu et al. [18] was 268 kb. Both estimates are quite large and reflect that medium-density SNP arrays are not well suited to detect small CNVR in spite of their high abundance. In cattle, the average sizes of CNVR detected with the Illumina BovineHD Genotyping BeadChip (777 K SNPs) [14], Illumina whole-genome sequencing and PacBio sequencing [41] were 66.15, 10 and 0.81 kb, respectively. Another consistent feature of CNVR is that, in general, their frequencies are low or very low. In our study, approximately 73% of the CNVR had frequencies lower than 1% and the average frequency was 1.44%. Liu et al. [18] reported lower CNVR frequencies ranging from 0.34% (Alpine and Northern European

goats) to 0.98% (Northwestern African goats). This decreased average CNVR frequency is not very significant and probably reflects differences in sampling size and the use of composite populations with multiple breeds, each one with its specific CNVR frequencies.

The CNVR detected in our study covered 779 protein-coding genes. Pathway analyses reflected a substantial enrichment of genes that are involved in olfactory perception, which is consistent with previous reports in cattle [13, 14]. In this regard, there is an important difference between our results and those by Liu et al. [18]. Whereas in the study of Liu et al. [18], the term "sensory perception" was underrepresented among the CNV genes (fold enrichment = 0.21), in our work the terms "olfactory transduction" (fold enrichment = 2.33) and "G-protein coupled purinergic nucleotide receptor activity" (fold enrichment = 13.19) were overrepresented, and many CNV genes were olfactory receptors. The two terms mentioned before are closely related because a broad array of purinergic receptors are differentially expressed in the olfactory receptor neurons that modulate odor responsiveness [42]. Moreover, purinergic nucleotides are important neuromodulators of peripheral auditory and visual sensory systems [42]. In cattle, Keel et al. [13] reported that "sensory perception of smell" and "G-protein coupled receptor signaling pathway" were significantly overrepresented in the protein-coding genes that overlapped with CNVR. Similarly, Upadhyay et al. [14] showed that "sensory perceptions of smell" and "chemical stimuli" are enriched in their set of CNV genes. A potential explanation for the underrepresentation of the "sensory perception" functional category among the genes overlapping CNV reported by Liu et al. [18] could be that in goats these genes are not well annotated yet, so the majority of them are identified with a LOC prefix and a number and, as a consequence of this, they are not correctly detected by PANTHER [43], thus biasing the results obtained in the gene ontology enrichment analysis.

Loci belonging to large multigene families might be more prone to co-localize with CNV because paralogous genes can act as templates in non-allelic

homologous recombination events, which promote increases or reductions in copy number [44]. It should be noted that olfactory receptor genes constitute the largest gene superfamily, and in humans more than 900 genes and pseudogenes have been identified [45]. In cattle, 1,071 olfactory receptor genes and pseudogenes are distributed in 49 clusters across 26 bovine chromosomes [46], and similar numbers have been reported for pigs [47]. Moreover, purifying selection against CNV is probably less intense in regions that contain olfactory-receptor genes than in genomic regions that contain genes with essential functions [48]. Interestingly, copy number changes in the olfactory receptor genes of wild and domestic mammals might have consequences on food foraging as well as on mate and predator recognition [49, 50].

In the set of genes that co-localize with CNVR, we also detected an enrichment of loci related with the multigene family of ATP binding cassette (ABC) transporters, a result that agrees well with previous findings in humans [51, 52, 53, 54] and cattle [14, 56]. In mammals, ABC transporters fulfill the mission of carrying a broad array of endogenous substrates, such as amino acids, peptides, sugars, anions and hydrophobic compounds and metabolites across lipid membranes. At least 49 ABC genes that belong to eight subfamilies have been identified in the human genome [52]. Copy number variation in the human *ABCC4* and *ABCC6* genes is associated with susceptibility to esophageal squamous cell carcinoma [51] and to the rare autosomal recessive disease pseudoxanthoma elasticum [54], respectively. Moreover, large-scale deletions of the human *ABCA1* gene are a causative factor for hypoalphalipoproteinemia [53], a disease that is characterized by the complete absence of the apolipoprotein AI and extremely low levels of plasma high-density lipoprotein (HDL) cholesterol. We also found a highly significant enrichment of pathways related with embryo development (anterior/posterior pattern specification, embryonic skeletal system morphogenesis), as previously reported [18]. These pathways are featured by genes that belong to the Hox multigene family of transcription factors, possibly reflecting the genomic instability of certain homeobox gene clusters as

evidenced by the existence of many synteny/paralogy breakpoints and assembly gaps as outlined in comparative studies [55].

Although not significant after correction for multiple testing, we detected an enrichment of pathways with metabolic significance, such as prolactin and insulin signaling, which could have an impact on milk production and growth [57, 58, 59]. Interestingly, the comparison of our work with that of Liu et al. [18] revealed 116 protein-coding genes that co-localize with the set of shared CNVR. One of the most relevant shared genes encodes ASIP, a protein that increases the ratio of pheomelanin to eumelanin by binding to the melanocortin 1 receptor and delivering an antagonist signal that blocks the downstream expression of eumelanogenic enzymes [60]. Mutations in the *ASIP* gene play critical roles in animal pigmentation [39]. For instance, the causal factor of the white color typical of many sheep breeds is the ubiquitous expression of a duplicated copy of the *ASIP* coding sequence, which is regulated by a duplicated promoter corresponding to the itchy E3 ubiquitin protein ligase gene [6, 39]. Although some studies proposed that the *ASIP* CNV might be associated with different pigmentation patterns in goats [8, 17, 37], no functional assay has verified an association of *ASIP* copy number with *ASIP* mRNA levels. Another interesting shared copy number variable gene is *ADAMTS20*, which was also identified in two previous CNV surveys [17, 18]. This gene encodes a metalloproteinase with an important role in melanoblast survival by mediating Kit signaling [38] and in palatogenesis [61]. Bertolini et al. [40] performed a selection scan in white vs. colored (black and red) goats and detected a selective sweep in the *ADAMTS20* gene. In the light of these results, the potential involvement of a structural variation in *ADAMTS20* in goat pigmentation should be explored further. Moreover, it is worthwhile to mention that several CNVR genes have functions related with production and reproduction traits. For instance, the *NCKAP5* gene, which co-localizes with CNVR_160_chr2 (frequency = 0.1), is associated with milk fat percentage in cattle [62]. Taking the above evidence into account, the implication of structural chromosomal variations in the genetic determinism of

traits of economic interest with a complex inheritance deserves further exploration by designing tools that allow inferring CNVR genotypes with high confidence.

Conclusions

With the PennCNV and QuantiSNP software, we detected 486 CNVR in the genome of the Murciano-Granadina breed. In a previous study [18] that used a less stringent pipeline (only PennCNV was used) and included multiple populations with small to moderate sample sizes, the average number of CNVR events per breed was ~20. One conclusion of our study is that CNV surveys, which are based on a broad array of breeds represented by only a few individuals, underestimate the true levels of the CNV diversity at the within-breed level. The main reason for this outcome is that since the majority of CNV have very low frequencies, they cannot be detected efficiently when sample size is small and, in consequence, much of the existing variation is missed. We have also found that genes that overlap with CNV are functionally related with olfactory transduction, embryo development, ABC transporters and G-protein coupled purinergic nucleotide receptor activity. Most of these genes belong to large multigene families encompassing tens, hundreds or thousands of paralogous genes that could act as substrates in non-allelic homologous recombination events, which is one of the main mechanisms generating duplications and deletions in humans and other species. Finally, we detected CNV that co-localize with the *ASIP* and *ADAMTS20* pigmentation genes, which according to previous studies have been subjected to positive selection for coat color in goats.

Supplementary Information

Additional file 1: Table S1. List of primers used in the real-time quantitative PCR experiment to validate four putative copy number variable genes

Additional file 2: Table S2. List of copy number variation regions (CNVR) consistently detected with PennCNV and QuantiSNP in 1,036 Murciano-Granadina goats

Additional file 3: Table S3. List of copy number variable genes detected in the current work and their concordance with those reported by Liu et al. [18]

Additional file 4: Table S4. Functional enrichment of genes co-localizing with copy number variation regions detected in 1,036 Murciano-Granadina goats

Declarations

Ethics approval and consent to participate

The collection of blood is a routine procedure carried out by trained veterinarians working for the CAPRIGRAN association, so it does not require a permission from the Committee on Ethics in Animal and Human Experimentation of the Universitat Autònoma de Barcelona.

Consent for publication

Not applicable

Availability of data and materials

The dataset supporting the conclusions of this article is accessible at Figshare (https://doi.org/10.6084/m9.figshare.12674357).

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the European Regional Development Fund (FEDER)/Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación/Project Reference grant: AGL2016-76108-R and by the CERCA Programme/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain). Dailu Guan was funded by a PhD fellowship from the China Scholarship Council (CSC).

Authors' contributions

MA, JJ, VL and JVD conceived and designed the experiment; JFA and XS contributed to the collection of materials and reagents, VL and AM performed DNA extractions, AC and BC carried out DNA quality assessment and genotyping tasks, DG made the bioinformatic analyses of the data, MGL contributed to the bioinformatic analyses of the data, AC and DG did the quantitative PCR analyses, MA and DG wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

The authors are indebted to the Asociación Nacional de Criadores de Caprino de Raza Murciano-Granadina (CAPRIGRAN), and specially to Miguel García García and Teresa Novo Díaz, who collected all blood samples. Moreover, we are also indebted to Dr. George Liu (USDA) for providing technical advice in CNV calling.

References

1. Clop A, Vidal O, Amills M. Copy number variation in the genomes of domestic animals. Anim Genet. 2012;43:503-17.
2. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16:172-83.
3. Bickhart DM, Liu GE. The challenges and importance of structural variation detection in livestock. Front Genet. 2014;5:37.
4. Wright D, Boije H, Meadows JRS, Bed'hom B, Gourichon D, Vieaud A, et al. Copy number variation in intron 1 of *SOX5* causes the pea-comb phenotype in chickens. PLoS Genet. 2009;5:e1000512.
5. Chen C, Liu C, Xiong X, Fang S, Yang H, Zhang Z, et al. Copy number variation in the *MSRB3* gene enlarges porcine ear size through a mechanism involving miR-584-5p. Genet Sel Evol. 2018;50:72.
6. Norris BJ, Whan VA. A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. Genome Res. 2008;18:1282-93.
7. Menzi F, Keller I, Reber I, Beck J, Brenig B, Schütz E, et al. Genomic amplification of the caprine *EDNRA* locus might lead to a dose dependent loss of pigmentation. Sci Rep. 2016;6:28438.
8. Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, et al. Selection signatures in goats reveal copy number variants

underlying breed-defining coat color phenotypes. PLoS Genet. 2019;15:e1008536.

9. Giuffra E, Törnsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, et al. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking *KIT*. Mamm Genome. 2002;13:569-77.
10. Pailhoux E, Vigier B, Chaffaux S, Servel N, Taourit S, Furet JP, et al. A 11.7-kb deletion triggers intersexuality and polledness in goats. Nat Genet. 2001;29:453-8.
11. Sundström E, Imsland F, Mikko S, Wade C, Sigurdsson S, Pielberg GR, et al. Copy number expansion of the *STX17* duplication in melanoma tissue from Grey horses. BMC Genomics. 2012;13:365.
12. Letaief R, Rebours E, Grohs C, Meersseman C, Fritz S, Trouilh L, et al. Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. Genet Sel Evol. 2017;49:77.
13. Keel BN, Lindholm-Perry AK, Snelling WM. Evolutionary and functional features of copy number variation in the cattle genome. Front Genet. 2016;7:207.
14. Upadhyay M, da Silva VH, Megens HJ, Visker MHPW, Ajmone-Marsan P, Bâlteanu VA, et al. Distribution and functionality of copy number variation across European cattle populations. Front Genet. 2017;8:108.
15. Hou Y, Bickhart DM, Hvinden ML, Li C, Song J, Boichard DA, et al. Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. BMC Genomics. 2012;13:376.
16. Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, et al. An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. BMC Genomics. 2010;11:639.
17. Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W, et al. Reference genome of wild goat (*capra aegagrus*) and sequencing of goat

breeds provide insight into genic basis of goat domestication. BMC Genomics. 2015;16:431.

18. Liu M, Zhou Y, Rosen BD, Van Tassell CP, Stella A, Tosser-Klopp G, et al. Diversity of copy number variation in the worldwide goat population. Heredity. 2018;122:636-46.
19. Nandolo W, Lamuno D, Banda L, Gondwe T, Mulindwa H, Nakimbugwe H, et al. Distribution of copy number variants in the genomes of East African goat breeds. In: Proceedings of the 11th world congress on genetics applied to livestock production, 11-16 Feb 2018, Auckland. 2018.
20. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16:1215.
21. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and characterization of a 52 K SNP Chip for goats. PLoS One. 2014;9:e86227.
22. Attiyeh EF, Diskin SJ, Attiyeh MA, Mossé YP, Hou C, Jackson EM, et al. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Res. 2009;19:276-83.
23. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643-50.
24. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17:1665-74.
25. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and

accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35:2013-25.

26. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomics. 2009;8:353-66.
27. Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol. 2011;29:512-20.
28. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 2008;36:e126.
29. Zhang Z, Cheng H, Hong X, Di Narzo AF, Franzen O, Peng S, et al. EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. Nucleic Acids Res. 2019;47:e39.
30. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2.
31. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44-57.
32. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1-13.
33. Ballester M, Castelló A, Ibáñez E, Sánchez A, Folch JM. Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals. Biotechniques. 2004;37:610-3.
34. Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, Souza CA, et al. Copy number variation in the porcine genome inferred from a 60 K SNP BeadChip. BMC Genomics. 2010;11:593.

35. Fontanesi L, Beretti F, Riggio V, Gómez González E, Dall'Olio S, Davoli R, et al. Copy number variation and missense mutations of the agouti signaling protein (*ASIP*) gene in goat breeds with different coat colors. Cytogenet Genome Res. 2009;126:333-47.
36. D'haene B, Vandesompele J, Hellemans J. Accurate and objective copy number profiling using real-time quantitative PCR. Methods. 2010;50:262-70.
37. Zhang B, Chang L, Lan X, Asif N, Guan F, Fu D, et al. Genome-wide definition of selective sweeps reveals molecular evidence of trait-driven domestication among elite goat (*Capra* species) breeds for the production of dairy, cashmere, and meat. GigaScience. 2018;7:giy105.
38. Silver DL, Hou L, Somerville R, Young ME, Apte SS, Pavan WJ. The secreted metalloprotease ADAMTS20 is required for melanoblast survival. PLoS Genet. 2008;4:e1000003.
39. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev Genet. 2019;20:135-56.
40. Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, et al. Signatures of selection and environmental adaptation across the goat genome post-domestication. Genet Sel Evol. 2018;50:57.
41. Couldrey C, Keehan M, Johnson T, Tiplady K, Winkelman A, Littlejohn MD, et al. Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. J Dairy Sci. 2017;100:5472-8.
42. Hegg CC, Greenwood D, Huang W, Han P, Lucero MT. Activation of purinergic receptor subtypes modulates odor sensitivity. J Neurosci. 2003;23:8291-301.
43. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14:703-21.

44. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10:551-64.
45. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome Res. 2001;11:685-702.
46. Lee K, Nguyen DT, Choi M, Cha SY, Kim JH, Dadi H, et al. Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. BMC Genomics. 2013;14:596.
47. Nguyen DT, Lee K, Choi H, Choi MK, Le MT, Song N, et al. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. BMC Genomics. 2012;13:584.
48. Young JM, Endicott RM, Parghi SS, Walker M, Kidd JM, Trask BJ. Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet. 2008;83:228.
49. Paudel Y, Madsen O, Megens HJ, Frantz LAF, Bosse M, Crooijmans RPMA, et al. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC Genomics. 2015;16:330.
50. Rinker DC, Specian NK, Zhao S, Gibbons JG. Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. Proc Natl Acad Sci USA. 2019;116:13446-51.
51. Sun Y, Shi N, Lu H, Zhang J, Ma Y, Qiao Y, et al. *ABCC4* copy number variation is associated with susceptibility to esophageal squamous cell carcinoma. Carcinogenesis. 2014;35:1941-50.
52. Vasiliou V, Vasiliou K, Nebert DW. Human ATP-binding cassette (ABC) transporter family. Hum Genomics. 2009;3:281-90.
53. Dron JS, Wang J, Berberich AJ, Iacocca MA, Cao H, Yang P, et al. Large-scale deletions of the *ABCA1* gene in patients with hypoalphalipoproteinemia. J Lipid Res. 2018;59:1529-35.
54. Kringen MK, Stormo C, Berg JP, Terry SF, Vocke CM, Rizvi S, et al. Copy number variation in the ATP-binding cassette transporter *ABCC6*

gene and *ABCC6* pseudogenes in patients with pseudoxanthoma elasticum. Mol Genet Genomic Med. 2015;3:233-7.

55. Wilming LG, Boychenko V, Harrow JL. Comprehensive comparative homeobox gene annotation in human and mouse. Database. 2015;2015:bav091.
56. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. Genome Res. 2010;20:693-703.
57. Fujita S, Rasmussen BB, Cadenas JG, Grady JJ, Volpi E. Effect of insulin on human skeletal muscle protein synthesis is modulated by insulin-induced changes in muscle blood flow and amino acid availability. Am J Physiol Endocrinol Metab. 2006;291:E745-54.
58. Bequette BJ, Kyle CE, Crompton LA, Buchan V, Hanigan MD. Insulin regulates milk production and mammary gland and hind-leg amino acid fluxes and blood flow in lactating goats. J Dairy Sci. 2001;84:241-55.
59. Freeman ME, Kanyicska B, Lerant A, Nagy G. Prolactin: structure, function, and regulation of secretion. Physiol Rev. 2000;80:1523-631.
60. Nasti TH, Timares L. MC1R, eumelanin and pheomelanin: their role in determining the susceptibility to skin cancer. Photochem Photobiol. 2015;91:188-200.
61. Wolf ZT, Brand HA, Shaffer JR, Leslie EJ, Arzi B, Willet CE, et al. Genome-wide association studies in dogs and humans identify *ADAMTS20* as a risk variant for cleft lip and palate. PLoS Genet. 2015;11:e1005059.
62. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. A large-scale genome-wide association study in U.S. Holstein cattle. Front Genet. 2019;10:420.

Estimating the copy number of the agouti signaling protein (*ASIP*) gene in goat breeds with different color patterns

Dailu Guan[1], Anna Castelló[1,3], María Gracia Luigi-Sierra[1], Vincenzo Landi[2], Juan Vicente Delgado[2], Amparo Martínez[2], Marcel Amills[1,3]

[1]Department of Animal Genetics, Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus de la Universitat Autònoma de Barcelona, Bellaterra, Spain; [2]Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain; [3]Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, Spain.

Corresponding author: Marcel Amills (marcel.amills@uab.cat)

(*Submitted to Livestock Science*)

Abstract

The agouti signaling protein (*ASIP*) gene has a crucial role in pigmentation by encoding a protein that binds the melanocortin 1 receptor and stimulates the synthesis of pheomelanin rather than eumelanin. Several studies have suggested that an increased copy number of the *ASIP* gene might explain the white pigmentation of certain goat breeds, as previously demonstrated in sheep. In the current work, we have identified the segregation of the *ASIP* CNV in Murciano-Granadina (black or brown coat), Malagueña (brown, blond or white coat) and Saanen (white coat) goats with available Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA) genotypes by using the PennCNV v1.0.5 and QuantiSNP v2 tools. This result shows that the *ASIP* CNV segregates in dark-colored breeds. To gain new insights into this issue, we have estimated the copy number of the *ASIP* gene in 83 goats from 8 breeds with different coloration patterns using a real-time quantitative PCR approach. Our results showed an increased *ASIP* copy number not only in Saanen (3.50 ± 0.23 copies relative to the calibrator) and white Malagueña (3.51 ± 0.51 copies) goats, but also in the Murciano-Granadina breed (3.33 ± 0.58 copies) as well as in blond/brown individuals from the Malagueña (3.58 ± 0.73 copies) breed. The number of *ASIP* copies was not significantly different in these four caprine populations (*P* value > 0.05). Moreover, we did not observe a trend towards increased *ASIP* copy number in breeds with predominantly white colors, such as Maltese (2.85 ± 0.28 copies), Jonica (2.82 ± 0.39 copies) and Blanca de Rasquera (2.37 ± 0.33 copies). Our results, combined with recent findings demonstrating the high structural complexity of the *ASIP* locus, indicate that additional functional and expression studies should be performed in order to fully understand the role of *ASIP* structural variation in goat pigmentation.

Keywords: copy number variation, pigmentation, real-time quantitative PCR

1. Introduction

Copy number variations (CNV) play a key role in the genetic determinism of several pigmentation phenotypes in domestic species (Freeman et al., 2006; Clop et al., 2012; Bickhart and Liu 2014; Zarrei et al., 2015). For instance, the loss of white pigmentation in the South African Boer breed is associated with a 1 Mb CNV mapping to the endothelin receptor type A (*EDNRA*) gene (Menzi et al., 2016). In sheep, the causal factor of the dominant white/tan (A^{Wt}) coat was mapped to a 190 kb tandem duplication encompassing the whole agouti signaling protein (*ASIP*) gene (Norris and Whan 2008). The transcription of the second copy of the ovine *ASIP* gene is controlled by the promoter of the itchy homolog E3 ubiquitin protein ligase (*ITCH*) gene and it shows a deregulated and ubiquitous pattern of expression associated with a white coloration (Norris and Whan 2008).

Classical genetic studies carried out in crossed goats revealed that the white color might be caused by the dominant A^{Wt} (white/tan) allele of the *ASIP* locus (Adalsteinsson et al., 1994). Fontanesi et al. (2009) identified one CNV in the caprine *ASIP* gene encompassing at least 100 kb, and they showed that in Saanen and Girgentana goats this CNV might have a correspondence with the A^{Wt} allele associated with the white pigmentation characteristic of these two breeds. Noteworthy, Fontanesi et al. (2009) detected an *ASIP* copy gain in one individual from the Murciano-Granadina breed, which can be black or brown but never white, and they also demonstrated that not all the analyzed Saanen goats carried 2 additional *ASIP* copies. They interpreted these findings in the light that there might be some degree of genetic heterogeneity in the *ASIP* locus and that the expression of the *ASIP* alleles could be modulated by epistatic interactions (Fontanesi et al. 2009). In a subsequent study based on whole-genome resequencing data, Dong et al. (2015) showed that black Yunnan Black goats and brown Australian Rangeland goats carry a single *ASIP* gene copy, while light colored Cashmere and Boer goats harbor multiple *ASIP* copies. However, the

reach of this experiment was limited by the low number of analyzed individuals. Zhang et al. (2018) also compared *ASIP* copy number in two white (Saanen and Liaoning) vs. two black (Leizhou goats and Dera Din Panah) goat breeds and inferred a lower *ASIP* copy number in the latter, suggesting that this CNV has causal effects on pigmentation. Indeed, a selection scan performed in white vs. colored (black and red) goats made it possible to detect a selective sweep co-localizing with the *ASIP* gene (Bertolini et al., 2018). More recently, Henkel et al. (2019) identified four different CNV, close to or encompassing the *ASIP* locus, that showed correspondence with the white or tan (A^{Wt}), Swiss markings (A^{sm}), badgerface (A^{b}), and peacock (A^{pc}) alleles of the *ASIP* locus. Moreover, transcriptomic analyses indicated that variation in copy number might involve changes in *ASIP* mRNA expression between eumelanistic and pheomelanistic body areas.

In a previous study (our unpublished data), we performed a CNV scan in a population of 1,036 Murciano-Granadina goats and found evidence of a CNV mapping to the *ASIP* locus. The goal of the current work is to characterize the segregation of the *ASIP* CNV in several goat breeds with different coat colors in order to find out whether dark-colored breeds have lower *ASIP* copy numbers than the ones with white coats.

2. Materials and Methods

2.1. CNV calling based on Goat SNP50 BeadChip genotyping data

In order to investigate the segregation of a CNV in the caprine *ASIP* locus and its potential association with coat color, we performed an initial experiment exclusively focused on three breeds for which Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA) data had been generated in previous experiments. Our initial data set comprised 1,036 Murciano-Granadina goats

which had been genotyped with the chip in the context of a genome-wide CNV scan (our unpublished data). From these, we selected 559 individuals with coloration records. We also obtained Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA) genotypes corresponding to 42 Saanen goats that were kindly provided by Dr. Gwenola Tosser-Klopp from INRA (Castanet-Tolosan). Finally, we generated Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA) genotypes for 54 Malagueña goats. We considered that this breed is of great interest because Malagueña goats can display white, light blond, dark blond and brown coat colors. Genomic DNA of Malagueña goats were genotyped by using the Illumina Goat SNP50 BeadChip (Tosser-Klopp et al., 2014) in accordance with the instructions of the manufacturer (Illumina Inc., San Diego, CA). Following Attiyeh et al. (2009), B allele frequencies (BAF) and signal intensity ratios (log R Ratio or LRR) were obtained with the GenomeStudio software 2.0.4 (Illumina, https://emea.illumina.com). Mapping of CNV was independently carried out for the datasets of 54 Malagueña goats and 42 Saanen goats. Specifically, we employed the EnsembleCNV pipeline (Zhang et al., 2019) to assemble initial calling data from PennCNV v1.0.5 (Wang et al., 2007; Diskin et al., 2008) and QuantiSNP v2 (Colella et al., 2007) into CNV regions (CNVR) with a heuristic algorithm (threshold of minimum overlap = 30%). The CNVR boundaries were subsequently refined by considering the local correlation structure of the LRR values of the SNPs mapping to CNVR (Zhang et al., 2019). Then, we reassigned CNV calls initially obtained with both PennCNV and QuantiSNP to each refined CNVR, so the final set of CNVR only comprised those simultaneously detected by both callers.

2.2. Estimating *ASIP* copy number by quantitative real time PCR

We performed a second experiment based on quantitative real-time PCR (qPCR) to estimate *ASIP* copy number in three breeds mentioned before plus five additional populations for which DNA was available. Goats under analysis

(N=83) belonged to the following breeds: Saanen (white, N=10), Jonica (white or rosy, sometimes with tawny spots in the head and neck, N=9), Carpathian (polychromatic, N=9), Maltese (white, with a raven-black area on the top and sides of the head, N=6), Blanca de Rasquera (white or white with black spots, N=9), Derivata di Siria (brown/blond, sometimes white pied, N=10), Malagueña (white, N=10; blond/brown, N=10) and Murciano-Granadina (black/brown, N=10) (**Figure S1**).

Primers were designed with the Primer Express Software (Applied Biosystems) to amplify specific regions of the caprine *ASIP* gene and two reference genes (**Table S1**): melanocortin 1 receptor (*MC1R*, Fontanesi et al., 2009; Liu et al., 2018) and glucagon (*GCG*, Ballester et al., 2004; Ramayo-Caldas et al., 2010). Specifically, polymerase chain reactions (PCR) were carried out in a final 15 µL volume containing 7.5 ng genomic DNA, 7.5 µL 2 × SybrSelect Master Mix (Applied Biosystems), 300 nM of each forward and reverse primer and ultrapure water. Each sample was analyzed in triplicate. Assays were loaded in 384-well plates and run in a QuantStudio 12K Flex Real-Time PCR System instrument (Applied Biosystems). The thermal cycling was 50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 seconds and 60°C for 1 min. The specificity of the PCR reactions was assessed with a melting curve analysis procedure based on the following thermal profile: 95°C for 15 seconds, 60°C for 15 seconds and a gradual increase in temperature with a ramp rate of 1% up to 95°C. By performing ten-fold serial dilutions of a goat DNA pool template, the generated standard curves showed a comparable amplification with efficiencies ranging from 107.2% to 108.8%. The relative copy number of the *ASIP* gene was inferred with the qbase+ software (Biogazelle, Ghent, Belgium) by using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen 2001). The average of the four samples with the lowest *ASIP* copy numbers was employed as calibrator for relative quantification.

Table 1. The number of individuals carrying copy number variation mapping to the caprine *ASIP* gene based on the analysis of Illumina Goat SNP50 BeadChip data (Illumina Inc., San Diego, CA)

Breed	Coat color	Number of individuals	Number of individuals carrying *ASIP* CNV
Saanen	White	42	28
Malagueña	White	16	8
	Light blond	17	4
	Dark blond	11	3
	Brown	10	2
Murciano-Granadina[2]	Brown	159	32
	Black	400	83

[2]only those with known coat colors were included here.

3. Results and discussion

By performing CNV mapping, we confirmed the existence of a CNVR co-localizing with the *ASIP* gene in white Saanen, black/brown Murciano-Granadina and white/blond/brown Malagueña goats (**Table 1**). This is a clear indication that, as pointed out by Fontanesi et al. (2009) there is increased copy number of the *ASIP* locus not only in goats from white breeds, such as Saanen, but also in Murciano-Granadina individuals that are black or brown (**Table 1**). To gain new insights into this issue, we decided to carry out a qPCR quantification of *ASIP* copy number in a panel of 83 goats from 8 breeds with diverse coat colors (**Figure S1**). The averaged estimates of *ASIP* relative copy number per breed and relative to the calibrator are shown in **Figure 1** and **Table 2**. It is important to emphasize that our data should not be interpreted in terms of absolute copy numbers. As previously said, all copy number estimates are calibrated with regard to the four individuals with lowest copy numbers. As expected, we observed high *ASIP* copy numbers in the Saanen breed (3.50 ± 0.23 copies, relative to the calibrator), a result that is consistent with data reported by

Fontanesi et al. (2009). Likewise, a similar high copy number was observed in pure white individuals from the Malagueña breed (3.51 ± 0.51 copies). However, we also detected high *ASIP* copy numbers in brown/blond Malagueña goats (3.58 ± 0.73 copies) and in black/brown Murciano-Granadina goats (3.33 ± 0.58 copies). Interestingly, the highest copy number was found in a light blond Malagueña goat (5.00 ± 0.18 copies). Performance of an ANOVA test with the "aov" function implemented in R (https://www.r-project.org/) revealed that *ASIP* copy numbers are not different amongst the four populations cited above ($P > 0.05$, **Table 2**). Moreover, several breeds, which exhibit a white or predominantly white coat (e.g. Jonica, Maltese and Blanca de Rasquera), showed lower *ASIP* copy numbers than brown/blond Malagueña and black/brown Murciano-Granadina goats (**Figure 1**). In a previous study, Zhang et al., (2018) reported that white Liaoning Cashmere goats have higher average *ASIP* copy numbers than black Leizhou goats, but several black Leizhou goats harbored higher *ASIP* copy numbers than their white Liaoning Cashmere counterparts. Similarly, Fontanesi et al. (2009) indicated that not all Saanen goats investigated in their experiment carried 2 additional copies of the *ASIP* gene and, even more, they also identified a Murciano-Granadina individual (MGb7) with an increased *ASIP* copy number similar to that estimated in Saanen goats. So, our data and results presented by other authors (Fontanesi et al. 2009, Zhang et al. 2018) do not evidence a perfect correlation between *ASIP* copy number and white coat color. As pointed out by Fontanesi et al. (2009), this could be due to a complex inheritance pattern involving epistasis and other genetic factors modulating pigmentation. Recently, Henkel et al., (2019) detected, with short-read resequencing data, at least four CNV located near or encompassing part of the caprine *ASIP* gene, which might be associated with different color patterns. These authors showed that in Grisons Striped goats (A^{sm}), Chamois Colored goats (A^{b}) and Peacock goats (A^{pc}), the eumelanistic skin displayed a weak *ASIP* mRNA expression, while the pheomelanistic skin regions in these three goats had at least 10-fold higher *ASIP* expression than the corresponding eumelanistic

samples. Moreover, the uniformly white Saanen goat (A^{Wt}) had the highest *ASIP* mRNA expression. In summary, these authors were able to correlate the eumelanistic/pheomelanistic pigmentation of skin with *ASIP* mRNA expression, but evidence correlating *ASIP* copy number estimates with *ASIP* mRNA expression in the skin were not provided. In the absence of evidence linking copy number with mRNA expression levels, it is difficult to assume that the CNV has a causal role on pigmentation because it is unknown whether increased copy number translates into an increased function. Indeed, the duplication of a gene does not necessarily involve a duplication of its expression due to compensatory mechanisms or to the loss of regulatory elements during the duplication process (Clop et al. 2012).

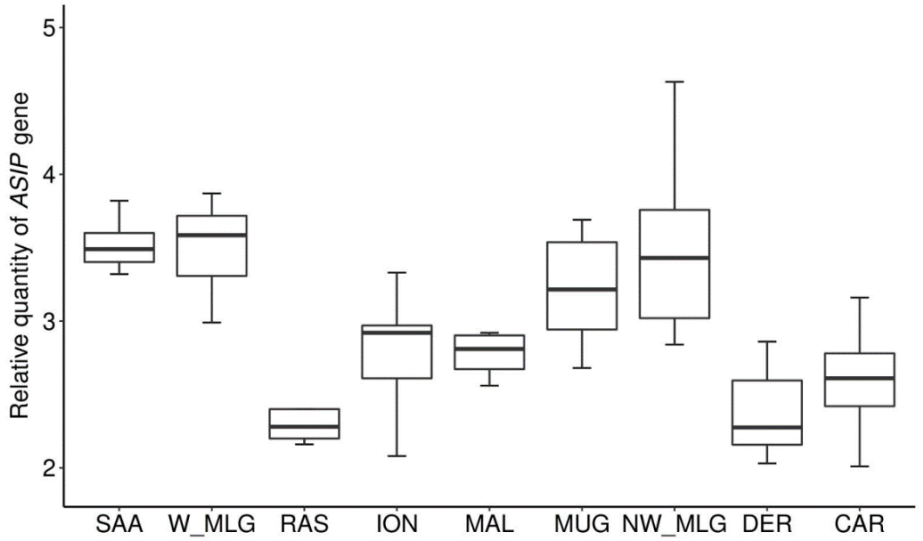

Figure 1. Boxplot depicting the relative copy number of the *ASIP* gene in eight goat breeds. The *y*-axis represents the median and the distribution of *ASIP* copy number in eight goat breeds (*x*-axis). The average of the four samples with the lowest *ASIP* copy numbers was used as calibrator. The following abbreviations have been used: SAA, Saanen (white, N=10); ION, Jonica (white or rosy, sometimes with tawny spots in the head and neck, N=9); CAR, Carpathian (polychromatic, N=9); RAS, Blanca de Rasquera (white or white with black spots, N=9); MAL, Maltese (white, with a raven-black area on the top and sides of the head, N=6); DER, Derivata di Siria (brown or blond, sometimes white pied, N=10); MUG, Murciano-Granadina (brown/black, N=10); Malagueña (white, W_MLG, N=10; blond or brown, NW_MLG, N=10). The pigmentation patterns of these populations are reported in **Figure S1**.

Table 2. Estimates of *ASIP* relative copy number in eight goat breeds

Coloration pattern	Population[1]	Code	Country	Number of individuals	CN range[2]	Mean ± SD[3]
Pure white	Saanen	SAA	Switzerland	10	3.04-3.82	3.5 ± 0.23^{a}
	White Malagueña	W_MLG	Spain	10	2.57-4.42	3.51 ± 0.51^{a}
Predominantly white	Blanca de Rasquera	RAS	Spain	9	1.89-3.04	2.37 ± 0.33^{c}
	Jonica	ION	Italy	9	2.08-3.33	2.82 ± 0.39^{bc}
	Maltese	MAL	Italy	6	2.56-3.36	2.85 ± 0.28^{abc}
Solid dark-colored	Murciano-Granadina	MUG	Spain	10	2.68-4.7	3.33 ± 0.58^{ab}
	Non-white Malagueña	NW_MLG	Spain	10	2.84-5	3.58 ± 0.73^{a}
	Derivata di Siria	DER	Italy	10	2.03-3.27	2.42 ± 0.40^{c}
Polychromatic	Carpathian	CAR	Romania	9	2.01-3.16	2.63 ± 0.33^{c}

[1]The specific colors of each breed can be found in **Figure S1**. Among them, Blanca de Rasquera goats are white or white with black spots; Jonica goats are white or rosy, sometimes with tawny spots in the head and neck; Maltese goats are white, with a raven-black area on the top and sides of the head; Murciano-Granadina goats are black or brown; Non-white Malagueña goats are blond or brown; Derivata di Siria goats have a light red coat, possibly white pied; Carpathian goats can be white, gray, reddish, black, or spotted.[2]For each individual, *ASIP* copy number (CN) was measured in triplicate and the average of the four samples with the lowest *ASIP* copy numbers was used as calibrator. [3]*ASIP* copy number averages with different letters are significantly different ($P < 0.05$) according to an ANOVA test. SD: standard deviation.

4. Conclusion

As a whole, our results and those published by other authors (Fontanesi et al., 2009; Dong et al., 2015; Zhang et al., 2018; Henkel et al., 2019) evidence that the potential role of structural variation in the *ASIP* locus on pigmentation has not been fully elucidated yet. In our study, we have not detected a consistent pattern by which light colored breeds display higher *ASIP* copy numbers than those observed in dark-colored breeds such as Murciano-Granadina. This could be due to the existence of genetic factors masking the effects of increased *ASIP* copy number or, alternatively, to the fact that increased *ASIP* copy number does not imply an increase in *ASIP* expression or function. For the four CNV mapping close or overlapping the *ASIP* gene, it would be crucial to investigate if copy number correlates with *ASIP* mRNA expression in the skin.

Supplementary Information

Table S1. List of primers used for the relative quantification of *ASIP* copy number by real-time qPCR.

Figure S1. Pictures of several goat breeds used in the qPCR experiment. The picture of the Saanen breed was retrieved from: https://commons.wikimedia.org/, and the remaining pictures were provided by Dr Jordi Jordana, Dr Juan Manuel Serradilla, Dr Baltasar Urrutia and Dr Juan carrizosa. Moreover, the pigmentation patterns of additional breeds can be found at the following links:
Carpathian: https://www.iga-goatworld.com/uploads/6/1/6/2/6162024/03_grosu_h_romania_ 09.04.2014.pdf .
Derivata di Siria: http:// www.agraria.org/caprini/derivatadisiria.htm.
Jonica: http://www. agraria.org/caprini/jonica.htm.
Maltese: http://eng.agraria.org/goat/maltese.htm.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

This study was funded by the European Regional Development Fund (FEDER)/Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación/Project Reference grant: AGL2016-76108-R and by the CERCA Programme/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain). Dailu Guan was funded by a PhD fellowship from the China Scholarship Council (CSC). María Gracia Luigi-Sierra was funded with an FPI Ph.D. grant (BES-C-2017-0024) awarded by the Spanish Ministry of Economy and Competitivity. Thanks to Dr Fabio Pilla, Dr Jordi Jordana, Dr Juan Manuel Serradilla Manrique and Dr Valentin Balteanu for providing goat samples from Italy, Spain and Romania. Thanks to Dr Gwenola Tosser-Klopp from INRA (Castanet-Tolosan) for providing Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA) data from 42 Saanen goats. We are also indebted to Dr Juan Manuel Serradilla Manrique, Dr. Baltasar Urrutia, Dr. Juan Carrizosa and Dr. Jordi Jordana for providing goat pictures.

Availability of data

Data can be accessed once the manuscript is accepted.

References

Adalsteinsson S., Sponenberg D.P., Alexieva S., Russel A.J., 1994. Inheritance of goat coat colors. J. Hered. 85, 267-72. https://doi.org/10.1093/oxfordjournals.jhered.a111454.

Attiyeh E.F., Diskin S.J., Attiyeh M.A., Mossé Y.P., Hou C., Jackson E.M., Kim C., Glessner J., Hakonarson H., Biegel J.A., Maris J.M., 2009. Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. Genome Res. 19, 276-83. https://doi.org/10.1101/gr.075671.107.

Ballester M., Castelló A., Ibáñez E., Sánchez A., Folch J.M., 2004. Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals. Biotechniques 37, 610-3. https://doi.org/10.2144/04374ST06.

Bertolini F., Servin B., Talenti A., Rochat E., Kim E.S., Oget C., Palhière I., Crisà A., Catillo G., Steri R., Amills M., Colli L., Marras G., Milanesi M., Nicolazzi E., Rosen B.D., Van Tassell C.P., Guldbrandtsen B., Sonstegard T.S., Tosser-Klopp G., Stella A., Rothschild M.F., Joost S., Crepaldi P., the AdaptMap consortium, 2018. Signatures of selection and environmental adaptation across the goat genome post-domestication. Genet. Sel. Evol. 50, 57. https://doi.org/10.1186/s12711-018-0421-y.

Bickhart D.M., Liu G.E., 2014. The challenges and importance of structural variation detection in livestock. Front. Genet. 5, 37. https://doi.org/10.3389/fgene.2014.00037.

Clop A, Vidal O, Amills M., 2012. Copy number variation in the genomes of domestic animals. Anim Genet. 43:503-17. https://doi: 10.1111/j.1365-2052.2012.02317.x

Colella S., Yau C., Taylor J.M., Mirza G., Butler H., Clouston P., Bassett A.S., Seller A., Holmes C.C., Ragoussis J., 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number

variation using SNP genotyping data. Nucleic Acids Res. 35, 2013-25. https://doi.org/10.1093/nar/gkm076.

Dong Y., Zhang X., Xie M., Arefnezhad B., Wang Z., Wang W., Feng S., Huang G., Guan R., Shen W., Bunch R., McCulloch R., Li Q., Li B., Zhang G., Xu X., Kijas J.W., Salekdeh G.H., Wang W., Jiang Y., 2015. Reference genome of wild goat (*capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication. BMC Genomics 16, 431. https://doi.org/10.1186/s12864-015-1606-1.

Fontanesi L., Beretti F., Riggio V., Gómez González E., Dall'Olio S., Davoli R., Russo V., Portolano B., 2009. Copy number variation and missense mutations of the agouti signaling protein (*ASIP*) gene in goat breeds with different coat colors. Cytogenet. Genome Res. 126, 333-47. https://doi.org/10.1159/000268089.

Freeman J.L., Perry G.H., Feuk L., Redon R., McCarroll S.A., Altshuler D.M., Aburatani H., Jones K.W., Tyler-Smith C., Hurles M.E., Carter N.P., Scherer S.W., Lee C., 2006. Copy number variation: new insights in genome diversity. Genome Res. 16, 949-61. https://doi.org/10.1101/gr.3677206.

Henkel J., Saif R., Jagannathan V., Schmocker C., Zeindler F., Bangerter E., Herren U., Posantzis D., Bulut Z., Ammann P., Drögemüller C., Flury C., Leeb T., 2019. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. PLoS Genet. 15, e1008536. https://doi.org/10.1371/journal.pgen.1008536.

Diskin S.J., Li M., Hou C., Yang S., Glessner J., Hakonarson H., Bucan M., Maris J.M., Wang K., 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 36, e126. https://doi.org/10.1093/nar/gkn556.

Livak K.J., Schmittgen T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods 25, 402-8. https://doi.org/10.1006/meth.2001.1262.

Liu M., Zhou Y., Rosen B.D., Van Tassell C.P., Stella A., Tosser-Klopp G., Rupp R., Palhière I., Colli L., Sayre B., Crepaldi P., Fang L., Mészáros G., Chen H., Liu G.E., the ADAPTmap Consortium, 2018. Diversity of copy number variation in the worldwide goat population. Heredity 122, 636-46. https://doi.org/10.1038/s41437-018-0150-6.

Menzi F., Keller I., Reber I., Beck J., Brenig B., Schütz E., Leeb T., Drögemüller C., 2016. Genomic amplification of the caprine *EDNRA* locus might lead to a dose dependent loss of pigmentation. Sci. Rep. 6, 28438. https://doi.org/10.1038/srep28438.

Norris B.J., Whan V.A., 2008. A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. Genome Res. 18, 1282-93. https://doi.org/10.1101/gr.072090.107.

Ramayo-Caldas Y., Castelló A., Pena R.N., Alves E., Mercadé A., Souza C.A., Fernández A.I., Perez-Enciso M., Folch J.M., 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC Genomics 11, 593. https://doi.org/10.1186/1471-2164-11-593.

Tosser-Klopp G., Bardou P., Bouchez O., Cabau C., Crooijmans R., Dong Y., Donnadieu-Tonon C., Eggen A., Heuven H.C.M., Jamli S., Jiken A.J., Klopp C., Lawley C.T., McEwan J., Martin P., Moreno C.R., Mulsant P., Nabihoudine I., Pailhoux E., Palhière I., Rupp R., Sarry J., Sayre B.L., Tircazes A., Wang J., Wang W., Zhang W., the International Goat Genome Consortium, 2014. Design and characterization of a 52K SNP Chip for goats. PLoS ONE 9, e86227. https://doi.org/10.1371/journal.pone.0086227.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F., Hakonarson H., Bucan M., 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17, 1665-74. https://doi.org/10.1101/gr.6861907.

Zarrei M., MacDonald J.R., Merico D., Scherer S.W., 2015. A copy number variation map of the human genome. Nat. Rev. Genet. 16, 172-83. https://doi.org/ 10.1038/nrg3871.

Zhang B., Chang L., Lan X., Asif N., Guan F., Fu D., Li B., Yan C., Zhang H., Zhang X., Huang Y., Chen H., Yu J., Li S., 2018. Genome-wide definition of selective sweeps reveals molecular evidence of trait-driven domestication among elite goat (*Capra species*) breeds for the production of dairy, cashmere, and meat. GigaScience 7, giy105. https://doi.org/10.1093/gigascience/giy105.

Zhang Z., Cheng H., Hong X., Di Narzo A.F., Franzen O., Peng S., Ruusalepp A., Kovacic J.C., Bjorkegren J.L.M., Wang X., Hao K., 2019. EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. Nucleic Acids Res. 47, e39. https://doi.org/10.1093/nar/gkz068.

Exploring the genomic architecture of coat color in Murciano-Granadina goats

Dailu Guan[1], Amparo Martínez[2], María Gracia Luigi-Sierra[1], Juan Vicente Delgado[2], Vincenzo Landi[2,3], Anna Castelló[1,4], Javier Fernández Álvarez[5], Xavier Such[6], Jordi Jordana[4], Marcel Amills[1,4]*

[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain. [2]Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain. [3]Department of Veterinary Medicine, University of Bari "Aldo Moro", SP. 62 per Casamassima km. 3, 70010 Valenzano (BA), Italy. [4]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [5]Asociación Nacional de Criadores de Caprino de Raza Murciano-Granadina (CAPRIGRAN), 18340 Granada, Spain. [6]Group of Research in Ruminants (G2R), Department of Animal and Food Science, Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona, Spain

*Corresponding author: Marcel Amills (marcel.amills@uab.cat)

(*In preparation*)

Abstract

The inheritance of pigmentation patterns in domestic animals is often determined by one or several genes with large effects that, in some instances, can interact through epistasis or other complex mechanisms. The genomic architecture of the black/brown color of Murciano-Granadina goats has not been investigated yet at the genome-wide level, although one candidate gene study reported the involvement of the *MC1R* gene. Herewith, we have carried out a genome-wide association study comprising 529 Murciano-Granadina goats with available coat color records and Goat SNP50 BeadChip genotypes. Statistical analysis of the data with the GEMMA software revealed a strong association between a chromosome 18 region containing the *MC1R* gene and coat color (q-value = 2.03×10^{-18}). This result is consistent with previous data and demonstrates that the inheritance of coat color in Murciano-Granadina goats is very simple, being determined by a single locus. Sequencing of the *MC1R* coding region and genotyping experiments evidenced that the c.801C>G (p.Cys267Trp) polymorphism tightly segregates with coat color, a result concordant with that generated in the aforementioned candidate gene study. In contrast with other pigmentation phenotypes, such as white spotting in cattle or pink coloring in goats, which are determined by several loci, our results clearly demonstrate that the inheritance of coat color in Murciano-Granadina goats is essentially monogenic.

Keywords: Coat color, Goat, GWAS, MC1R, Murciano-Granadina

1 Introduction

Coat color has important physiological functions related with camouflage from predators, protection from UV radiation and communication (Linderholm and Larson, 2013). More than 300 genes with effects of pigmentation have been reported, and many of them are dedicated to control the synthesis of eumelanin and pheomelanin (Montoliu et al., 2020). In this regard, the role of the melanocortin 1 receptor (*MC1R*) is essential. Binding of proopiomelanocortin (POMC) to MC1R activates a series of biochemical events, fundamentally regulated by the cAMP secondary messenger, which promote an increase in the activity of tyrosinase (TYR), the rate-limiting enzyme synthesizing melanin from tyrosine (Linderholm and Larson, 2013). This event, combined with higher levels of TYR and of tyrosinase-related proteins 1 (TYRP1) and 2 (TYRP2), enhances the synthesis of black/brown eumelanin (García-Borrón et al., 2014). In contrast, agouti signaling protein (*ASIP*) antagonizes the effect of MC1R and induces the production of red/yellow pheomelanin (Gracía-Borrón et al., 2014). Another key locus in animal pigmentation is the KIT proto-oncogene, receptor tyrosine kinase (*KIT*), which is crucial for melanoblast differentiation and proliferation and melanogenesis (Linderholm and Larson, 2013).

The majority of complex traits are controlled by a broad array of polymorphisms with very small quantitative effects, but coat color is considered to have a much simpler genetic basis, being usually controlled by a few loci with large effects (Hayes et al., 2010). In mice, at least ten genes have been reported to influence white spotting (Baxter et al., 2004). Moreover, a meta-analysis of genome-wide association studies (GWAS) for skin color in 17,262 Europeans highlighted nine significantly associated SNPs, making it possible to identify several candidate genes (Liu et al., 2015). Similarly, the inheritance of pigmentation patterns in domestic animals is often oligogenic. For instance, in cattle two highly significant quantitative trait loci (QTLs) on chromosomes 6 and 22, and co-localizing with the *KIT* and melanocyte inducing transcription factor (*MITF*)

genes, have been associated with white spotting (Jivanji et al., 2019). Moreover, a third signal co-localizing with the paired box 3 (*PAX3*) gene, which has a crucial role in melanogenesis, was also detected by the same authors. Indeed, *PAX3* has been reported as a potential causal factor for the splashed white coat phenotype in horses (Hauswirth et al., 2012). Similarly, in goats the genetic basis of the PINK and PINK NECK phenotypes, which are considered as a defect, was investigated through a GWAS approach (Martin et al., 2016). Four and five significant associations were detected for PINK and PINK NECK, respectively, and the *ASIP* gene was proposed as a causal factor for the PINK phenotype (Martin et al. 2016). Altogether, these findings indicate that the genomic architecture of coat color can be quite complex, involving the participation of several loci acting cooperatively to generate a specific pigmentation pattern. In the current work, we aimed to study the genomic architecture of coat color in Murciano-Granadina goats, which can be black or brown. A previous candidate gene study reported that *MC1R* genotype is associated with this phenotype (Fontanesi et al., 2009), but sample size was low and no genome-wide study was performed precluding the detection of additional genetic factors.

2 Materials and Methods

2.1 Ethics Statement

The collection of blood is a routine procedure carried out by trained veterinarians working for the CAPRIGRAN association, so it does not require a permission from the Committee on Ethics in Animal and Human Experimentation of the Universitat Autònoma de Barcelona.

2.2 Sample Collection and Genotyping

Blood samples from 529 Murciano-Granadina goats were collected in EDTA K3 coated vacuum tubes and stored at -20 ºC before processing. Genomic DNA was isolated using a previously reported salting-out procedure (Guan et al., 2020) and resuspended in 1 mL TE buffer (Tris-HCl 10 mmol/L, EDTA 1 mmol/L, pH = 8). These 529 goats were phenotyped for coat color by visual assessment, since the two pigmentation patterns (black and brown) are discernible to the naked eye (**Figures 1A** and **1B**). The Goat SNP50 BeadChip (Tosser-Klopp et al., 2014), which contains 53,347 single nucleotide polymorphisms (SNPs), was used to genotype the 529 goats in accordance with the instructions of the manufacturer (Illumina, San Diego, CA). Data normalization and genotype calling were performed with the GenomeStudio software 2.0.4 (Illumina, https://emea.illumina.com). The PLINK v1.9 software (Purcell et al., 2007) was used to filter out unmapped and non-autosomal SNPs, as well as those with a low call rate (< 90%) and low frequency (< 1%). After applying these filtering criteria, 43,240 SNPs were selected to perform subsequent analyses.

2.3 Genome-wide association study

The genome-wide association study (GWAS) was carried out with the GEMMA software (version 0.98.1), which implements the Genome-wide Efficient Mixed Model Association algorithm (Zhou and Stephens, 2012). The following statistical model was used:

$$Y = x\beta + u + \varepsilon$$

where Y is a vector of phenotypic values coded as 1 (black) or 2 (brown); x is a n-vector of marker genotypes harbored by each individual; β is the effect size of the marker (allele substitution effect); u is a n-vector of random effects with a n-dimensional multivariate normal distribution $(0, \lambda\tau^{-1}K)$, being τ^{-1} the variance of the residual error, λ the ratio between the two variance components and K a n × n

relatedness matrix derived from the 42,793 valid SNPs. Finally, ε is a vector of errors. In this model, we did not include the fixed factors defined in Guan et al. (2020) because they are not expected to have any effect on coat color. The GEMMA software contrasts the alternative hypothesis (H_1: $\beta \neq 0$) against the null one (H_0: $\beta = 0$) by carrying out likelihood ratio tests for each marker. Besides, the relatedness matrix, which is constructed by accounting all genome-wide SNPs as a random effect, is employed to correct for population structure. Multiple testing was implemented through a false discovery rate approach (Benjamini and Hochberg, 1995), and a *q*-value ≤ 0.05 was established as a threshold of significance in the GWAS.

2.4 Sanger sequencing

Two pairs of primers were designed with the Primer-BLAST tool (https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi) in order to amplify the coding region of the *MC1R* gene (**Supplementary Table S1**) in 9 black and 13 brown Murciano-Granadina individuals. The polymerase chain reaction (PCR) contained 50 ng of genomic DNA, 1× BIOTAQ PCR buffer (Bioline, Barcelona, Spain), 200 µmol/L of dNTPs, 0.2 µmol/L of each primer, 1.5 mmol/L $MgCl_2$, and 0.65 units of BIOTAQ DNA Polymerase (Bioline, Barcelona, Spain). Nuclease-free water was added to a final volume of 25 µL. The thermal cycle was as follows: a hot-start step at 95°C for 2 minutes, followed by 34 cycles of 95°C for 45 seconds (denaturation), 60°C for 45 seconds (annealing) and 72°C for 45 seconds (extension), plus a final extension at 72°C for 5 minutes. Five µL of the PCR were mixed with 1.5 µL of a mixture containing 1.13 µL PCR buffer 1× (composition for 1 mL: 100 µL PCR Gold Buffer 10× + 100 µL $MgCl_2$ 25 mmol/L + 800 µL H_2O), 0.12 µL Exonuclease I (20 units/uL, Thermo Fisher Scientific, Barcelona, Spain) and 0.25 µL FastAP Thermosensitive Alkaline Phosphatase (1 unit/µL, Thermo Fisher Scientific, Barcelona, Spain). This mixture was incubated at 37ºC for 15 minutes plus 85ºC during 15 minutes. Purified amplicons were sequenced with the BigDye Terminator Cycle

Sequencing Kit v1.1 (Applied Biosystems, Foster City, CA). Sequencing reactions were run on ABI 3730 DNA analyzer (Applied Biosystems, Foster City, CA). Finally, sequences were viewed and aligned with the Molecular Evolutionary Genetics Analysis software (MEGA X) (Kumar et al., 2018).

2.5 TaqMan genotyping experiment

In order to further confirm the causality of the c.801C>G mutation, we used a Custom Taqman SNP Genotyping Assay (Applied Biosystems) to genotype 49 black and 41 brown individuals. TaqMan probes are shown in the **Supplementary Table S2**. Five samples with GG (N=1), CC (N=2) and GC (N=2) genotypes ascertained by Sanger sequencing were used as positive controls. The genotyping reaction was carried out in a final volume of 15 μL containing 1× Taqman Universal PCR Master Mix (Applied Biosystems, Foster City, CA), 1× Taqman Custom Genotyping Assay designed for rs669694251 (Applied Biosystems, Foster City, CA) and 18 ng of genomic DNA. Real-time PCRs were performed in 96-well reaction plates and they were run in a 7900-HT Real Time PCR system (Applied Biosystems, Foster City, CA). The thermal profile was: 50°C for 2 min, 95°C for 10 min and 40 cycles of 95°C for 15 seconds and 60°C for 1 min. Genotypes were obtained using the Genotyping Analysis Module of the Applied Biosystems Analysis Software accessible in the Thermofisher Cloud (https://www.thermofisher.com/es/es/home/digital-science.html, accessed March 20, 2020).

3 Results and Discussion

Through GWAS analysis, we detected 25 SNPs on chromosome 18 (12.18-22.30 Mb) showing significant associations with coloration (**Figures 1C** and **1D**). The marker showing the highest significance was rs268287597 (q-value = $2.03 \times 10^{-}$

[18], **Figure 1D**). The analysis of the gene content of the aforementioned chromosome 18 region revealed the presence of a single peak coinciding with the position of the *MC1R* gene. In principle, this would indicate that the inheritance of the black/brown color in Murciano-Granadina goats is monogenic. This result contrasts that obtained in humans and domestic animals, in which color is determined by multiple loci. For instance, a GWAS focused on skin pigmentation pinpointed the existence of fifteen of major genes in Eurasians, but the analysis of the KhoeSan populations indigenous to southern Africa revealed that these genes explain only a minimal part of skin color and that there are many other loci yet to be discovered, evidencing that the genomic architecture of skin color in humans is much more complex than what was previously thought (Martin et al., 2017). Moreover, in humans more than 200 markers have been independently associated with a broad array of hair colors going from blond to black (Morgan et al., 2018). In cattle the *MC1R* genotype is the main determinant of the black, red and wild type color, but the *KIT* gene has been recently involved in the reddening of the black pigmentation indicating that it acts as a modifier gene (Hulsman Hanna et al., 2014). White spotting or proportion of black seem also to have an oligogenic inheritance in cattle, being the *KIT* and *MITF* genes two key players in the determination of these phenotypes (Jivanji et al., 2019; Hayes et al., 2010), and similar conclusions have been obtained when analyzing the PINK and PINK NECK phenotypes in goats (Martin et al., 2016).

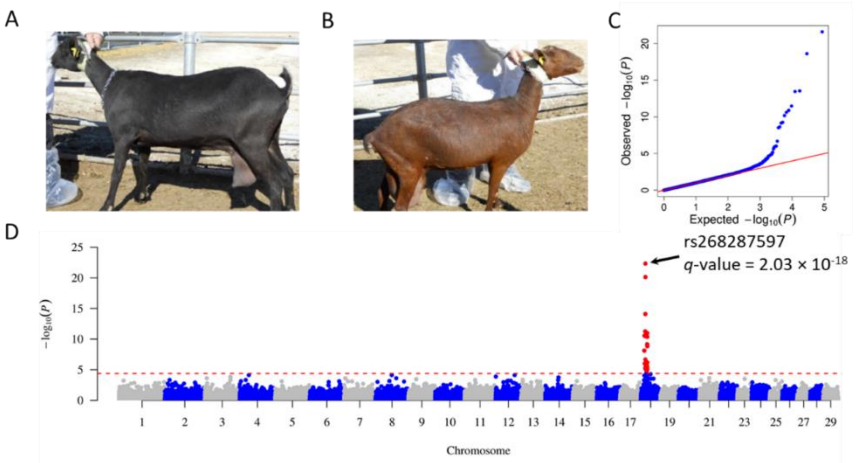

Figure 1. Murciano-Granadina goats with black (**A**) and brown (**B**) coat colors, which were respectively coded as 1 and 2 in the genome-wide association study (GWAS). (**C**) Quantile-quantile (QQ) plot of the expected versus observed *P* values in the GWAS analysis. (**D**) Manhattan plot depicting the associations between coat color and Goat SNP50 BeadChip genotypes in 387 black and 142 brown Murciano-Granadina goats. Negative $\log_{10}P$ values (*y*-axis) of the associations between SNPs and phenotypes are plotted against the genomic location of each SNP marker (*x*-axis). Markers on different chromosomes are denoted with different colors. The horizontal dashed line indicates statistical significance after correction for multiple testing by using the false discovery rate approach reported by Benjamini and Hochberg (1995). The arrow indicates the leading SNP that shows the highest association with phenotype (rs268287597, *q*-value = 2.03×10^{-18}).

The coincidence of the only genome-wide significant association with the *MC1R* locus agreed well with data reported in a candidate gene study for coat color in goats (Fontanesi et al., 2009), which indicated that *MC1R* genotype is associated with the pigmentation of Murciano-Granadina goats. Sanger sequencing revealed two missense mutations c.748G>T (p.Val250Phe) and c.801C>G (p.Cys267Trp) in the *MC1R* gene. These two mutations were also identified by Fontanesi et al. (2009) and they are registered in the Ensembl database (https://www.ensembl.org) with identifiers rs657434682 and rs669694251, respectively. Visual inspection of the sequences revealed a single G peak (GG

genotype) at position 748 in 9 black and 10 brown goats, while only 3 brown goats showed a double GT peak at this location and no goat displayed a single T peak (TT genotype). Altogether, these results rule out c.748G>T (p.Val250Phe) as a causal mutation for the black/brown coat color of Murciano-Granadina goats. With regard to c.801C>G, black goats showed either a single G peak (GG genotype, N=1) or a GC double peak (GC genotype, N=8), while all brown goats (N=13) displayed a single C peak (CC genotype, Figure 2A). By performing a TaqMan genotyping experiment in 49 black and 41 brown individuals, we found that all brown goats were CC and all black goats were either GG or GC (**Figure 2B**), indicating that the G-allele has a dominant inheritance. Moreover, a second GWAS was carried out by using a SNP data set comprising the 43,240 SNPs contained in the Goat SNP50 BeadChip plus the rs669694251 (c.801C>G) marker. The GEMMA software (Zhou and Stephens, 2012) was employed to investigate the association of these 43,241 markers and color phenotypes in 90 Murciano-Granadina goats as previously explained. In **Figure 2C**, it can be appreciated that marker rs669694251 displays the most significant association with coat color (c.801C>G, *q*-value = 2.91×10^{-25}). Fontanesi et al. (2009) investigated the association of the *MC1R* c.801C>G genotype with coat color in 28 Murciano-Granadina goats and found a fully consistent result. The G-allele would be a dominant gain-of-function mutation that is located in the extracellular loop 3 of the MC1R molecule and probably disrupts a disulfide bond between Cys267 and Cys275, thus altering the three-dimensional structure and activity of the protein towards an increased production of eumelanin. We have examined the genotype of 22 bezoars at genomic position 16,105,786 (which corresponds to c.801C>G) on chromosome 18 by using a data set of whole-genome sequences reported in Guan et al. (2019). This analysis revealed that all bezoars display CC genotypes at position 801 of the *MC1R* coding region, suggesting that the G-allele emerged during or after goat domestication.

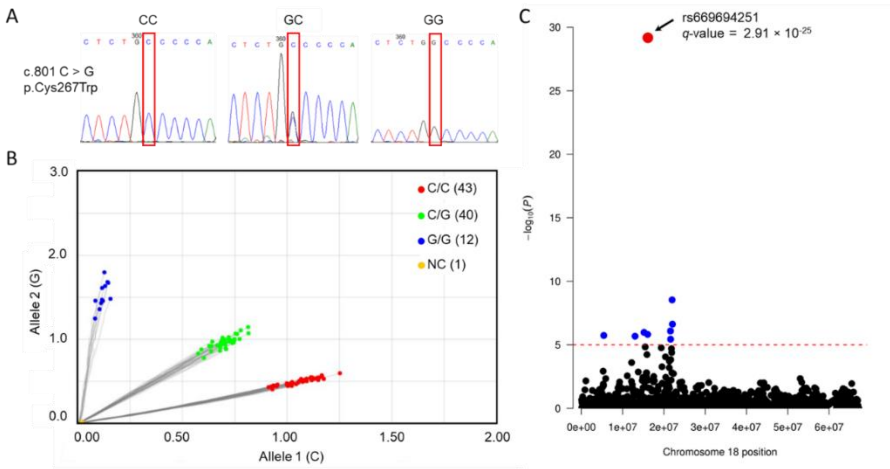

Figure 2. (**A**) Sequence electropherograms showing the region containing the c.801C>G, (p.Cys267Trp) SNP for individuals with CC, GC and GG genotypes. (**B**) Cluster plots of TaqMan genotyping results obtained with the Genotyping Analysis Module implemented in the ThermoFisher Cloud computing application (Applied Biosystems). The horizontal and vertical axes correspond to alleles C and G, respectively. The dots with red, green and blue colors represent CC, CG and GG genotypes, respectively. The negative control is indicated by an orange dot. (**C**) Manhattan plot depicting associations between coat color (41 brown and 49 black Murciano-Granadina goats) and the genotypes of marker rs669694251 (red dot) plus 1,134 additional Goat SNP50 BeadChip markers mapping to goat chromosome 18. The dashed line represents the negative $\log_{10}P$ value defining the threshold of significance (q-value ≤ 0.05) after correcting for multiple testing with a false discovery rate approach (Benjamini and Hochberg, 1995). Significant SNPs are indicated with blue dots.

4 Conclusion

Coloration seems to have a complex genetic basis in the majority of goat breeds, with patterns of association that are often difficult to interpret (Fontanesi et al., 2009). In this work, we have demonstrated that the genomic architecture of the black or brown coat color of the Murciano-Grnadina goat breed is very simple, being determined by a single locus: the *MC1R* gene. Likely, the Cys267Trp substitution alters substantially the three-dimensional structure of MC1R, but

instead of abolishing its function triggers an increase in its activity leading to an enhanced synthesis of eumelanin. In our GWAS we did not find any additional genome-wide significant association, confirming the monogenic inheritance of this trait. This result contrasts strongly with findings made in other domestic species in which pigmentation is determined by the concerted action of several genes.

5 Supplementary Information

Supplementary Table S1. Primers used in the amplification of the coding region of the melanocortin 1 receptor gene.

Supplementary Table S2. Custom TaqMan probes used to genotype the rs669694251 marker.

6 Conflict of Interest

The authors declare that they have no competing interests.

7 Author Contributions

MA, JJ, AM, JVD and XS conceived and designed the experiment; VL performed DNA extractions and DNA quality assessment; AM and JFA coordinated the collection of color phenotypes; DG and MGL performed the Sanger sequencing of the *MC1R* gene; AC and DG carried out MC1R genotyping tasks; DG did the bioinformatics and statistical analyses of the data; MA and DG wrote the manuscript. All authors read and approved the final version of the manuscript.

8 Funding

This research was funded by the European Regional Development Fund (FEDER)/Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación/Project Reference grant: AGL2016-76108-R and by the CERCA Programme/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016-2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain). Dailu Guan was funded by a PhD fellowship from the China Scholarship Council (CSC). María Gracia Luigi-Sierra was funded with a fellowship Formación de Personal Investigador (BES-C-2017-079709) awarded by the Spanish Ministry of Economy and Competitivity.

9 Acknowledgments

The authors are indebted to the Asociación Nacional de Criadores de Caprino de Raza Murciano-Granadina (CAPRIGRAN), and more specifically to Miguel García, Teresa Novoa and Manolo Delgado, for providing the blood samples and color phenotypes of goats analyzed in the current study. We also want to acknowledge Drs. Juan Manuel Serradilla (Universidad de Córdoba), Baltasar Urrutia (Instituto Murciano de Investigación y Desarrollo Agrario y Alimentario) and Juan Carrizosa (Instituto Murciano de Investigación y Desarrollo Agrario y Alimentario) for providing pictures of Murciano-Granadina goats.

10 Data Availability Statement

The dataset used to perform GWAS is accessible at Figshare (https://doi.org/10.6084/m9.figshare.11999823), and *MC1R* sequences have been deposited in the GenBank database (accession codes: MT186757-MT186778).

11 References

Baxter, L.L., Hou, L., Loftus, S.K., and Pavan, W.J. (2004). Spotlight on spotted mice: A review of white spotting mouse mutants and associated human pigmentation disorders. *Pigment Cell Res.* 17, 215-224. doi: 10.1111/j.1600-0749.2004.00147.x.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal. Stat. Soc. B.* 57, 289-300.

Fontanesi, L., Beretti, F., Riggio, V., Dall'Olio, S., González, E.G., Finocchiaro, R., et al. (2009). Missense and nonsense mutations in melanocortin 1 receptor (*MC1R*) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC Genet*. 10, 47. doi: 10.1186/1471-2156-10-47.

García-Borrón, J.C., Abdel-Malek, Z., and Jiménez-Cervantes, C. (2014). MC1R, the cAMP pathway, and the response to solar UV: extending the horizon beyond pigmentation. *Pigment Cell Melanoma Res.* 27, 699-720. doi: 10.1111/pcmr.12257.

Guan, D., Landi, V., Luigi-Sierra, M.G., Delgado, J.V., Such, X., Castelló A., et al. (2020). Analyzing the genomic and transcriptomic architecture of milk traits in Murciano-Granadina goats. *J. Anim. Sci. Biotechnol*. 11, 35. doi: 10.1186/s40104-020-00435-4.

Guan, D., Mármol-Sánchez, E., Cardoso, T.F., Such, X., Landi, V., Tawari, N.R., et al. (2019). Genomic analysis of the origins of extant casein variation in goats. *J. Dairy Sci.* 102, 5230-5241. doi: 10.3168/jds.2018-15281.

Hauswirth, R., Haase, B., Blatter, M., Brooks, S.A., Burger, D., Drögemüller, C., Gerber, V., Henke, D., Janda, J., Jude, R., Magdesian, K.G., Matthews, J.M.,

Poncet, P.A., Svansson, V., Tozaki, T., Wilkinson-White, L., Penedo, M.C.T., Rieder, S., and Leeb, T. (2012). Mutations in MITF and PAX3 cause "Splashed white" and other white spotting phenotypes in horses. *PLoS Genet.* 8, e1002653. doi: 10.1371/journal.pgen.1002653.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., and Goddard, M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genet.* 6, e1001139. doi: 10.1371/journal.pgen.1001139.

Hulsman Hanna, L.L., Sanders, J.O., Riley, D.G., Abbey, C.A., and Gill, C.A. (2014). Identification of a major locus interacting with MC1R and modifying black coat color in an F2 Nellore-Angus population. *Genet. Sel. Evol.* 46, 4. doi: 10.1186/1297-9686-46-4.

Jivanji, S., Worth, G., Lopdell, T.J., Yeates, A., Couldrey, C., Reynolds, E., Tiplady, K., Mcnaughton, L., Johnson, T.J.J., Davis, S.R., Harris, B., Spelman, R., Snell, R.G., Garrick, D., and Littlejohn, M.D. (2019). Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. *Genet. Sel. Evol.* 51, 62. doi: 10.1186/s12711-019-0506-2.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol. Biol. Evol*. 35, 1547-1549. doi: 10.1093/molbev/msy096.

Linderholm, A., and Larson, G. (2013). The role of humans in facilitating and sustaining coat colour variation in domestic animals. *Semin Cell Dev Biol*. 24, 587-593. doi: 10.1016/j.semcdb.2013.03.015.

Liu, F., Visser, M., Duffy, D.L., Hysi, P.G., Jacobs, L.C., Lao, O., Zhong, K., Walsh, S., Chaitanya, L., Wollstein, A., Zhu, G., Montgomery, G.W., Henders, A.K., Mangino, M., Glass, D., Bataille, V., Sturm, R.A., Rivadeneira, F., Hofman, A., Van Ijcken, W.F.J., Uitterlinden, A.G., Palstra, R.J.T.S., Spector, T.D., Martin, N.G., Nijsten, T.E.C., and Kayser, M. (2015). Genetics of skin

color variation in Europeans: genome-wide association studies with functional follow-up. *Hum. Genet.*134, 823-835. doi: 10.1007/s00439-015-1559-0.

Martin, P.M., Palhière, I., Ricard, A., Tosser-Klopp, G., and Rupp, R. (2016). Genome wide association study identifies new loci associated with undesired coat color phenotypes in Saanen goats. *PLoS One* 11, e0152426. doi: 10.1371/journal.pone.0152426.

Martin, A.R., Lin, M., Granka, J.M., Myrick, J.W., Liu, X., Sockell, A., Atkinson, E.G., Werely, C.J., Möller, M., Sandhu, M.S., Kingsley, D.M., Hoal, E.G., Liu, X., Daly, M.J., Feldman, M.W., Gignoux, C.R., Bustamante, C.D., and Henn, B.M. (2017). An unexpectedly complex architecture for skin pigmentation in Africans. *Cell* 171, 1340-1353.e1314. doi: 10.1016/j.cell.2017.11.015.

Montoliu L., Oetting W.S., Bennett D.C. (March, 2020) Color Genes. European Society for Pigment Cell Research. World Wide Web (http://www.espcr.org/micemut).

Morgan, M.D., Pairo-Castineira, E., Rawlik, K., Canela-Xandri, O., Rees, J., Sims, D., Tenesa, A., and Jackson, I.J. (2018). Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* 9, 5271. doi: 10.1038/s41467-018-07691-z.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81, 559-575. doi: 10.1086/519795.

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., et al. (2014). Design and characterization of a 52K SNP Chip for goats. *PLoS One*. 9, e86227. doi: 10.1371/journal.pone.0086227.

Zhou, X., Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 44, 821-824. doi: 10.1038/ng.2310.

Chapter 4

General Discussion

The main aim of this Thesis is to investigate the genomic architecture of dairy and pigmentation traits in Murciano-Granadina (MUG) goats through the utilization of high throughput technologies, such as Illumina Goat SNP50 BeadChip genotyping and next generation sequencing approaches. The results obtained in this research have been discussed in the five papers which constitute the core of this Thesis, while in the General Discussion section we will provide a global view about the main findings and conclusions generated in the present work. One of the main reasons for analyzing dairy and pigmentation phenotypes in the same Thesis is that, according to the literature, they are expected to have a very different genetic basis. While the determinism of pigmentation traits is usually monogenic or oligogenic and environmental influences are in general negligible, the inheritance of dairy traits tends to be highly polygenic, with many loci involved, and there is also an important environmental component. This feature gives us the opportunity to investigate and compare phenotypes with highly divergent genomic architectures, enriching our perspective about how goat phenotypes are genetically determined. We have also characterized copy number variation in MUG goats as a first step to understand their impact on phenotypic variation.

4.1 Molecular basis of lactation in Murciano-Granadina goats

4.1.1 Mammary gene expression in early and late lactation is similar but there are important changes in the mRNA levels of several genes related with proliferation and cell death

The performance of RNA-Seq analysis in paper I resulted in the identification of 42 differentially expressed (DE) genes between T1 vs. T2, a

significantly lower number than that obtained in the T1 vs. T3 (1,377), and T2 vs. T3 (1,039 DE genes) comparisons. In the study carried out by Bhat et al. (2019), a higher number of DE genes was observed in Kashmiri (455) and Jersey cattle (418 DE genes) when comparing early (90 days) and late (250 days) lactation (Bhat et al., 2019). One potential explanation for this discrepancy is that this study applied a less stringent threshold of significance than ours, only considering as a threshold of significance a *q*-value < 0.05 (instead of also taking into account the fold-change as we did). As mentioned in paper I, 135 DE genes were detected when comparing 50 days vs. 150 days of lactation in sheep (Suárez-Vega et al., 2015), but in this case the threshold of significance was the same as ours (an absolute $\log_2$FC >1.5 and a *q*-value < 0.05). In another study focusing on the lactation of buffaloes, Arora et al. (2019) compared early (30-54 d) vs. late lactation (250-273 d) and detected 125 upregulated and 32 downregulated DE genes, but they employed a threshold of significance that was more stringent than ours (absolute $\log_2$FC ≥ 2.0 and *q*-value ≤ 0.05). These findings indicate that the threshold of significance is probably not the main factor explaining why in early and late lactation we have detected less DE genes than what has been published in other studies. One potential explanation would be methodological, being especially important the use of different pipelines for differential expression analysis. For instance, Arora et al. (2019) carried out differential expression analysis using the CLC transcriptomics analysis tool implemented in the CLS Genomics Workbench 6.5.1 (CLC Bio, Aarhus, Denmark), while Bhat et al. (2019) used the Cuffdiff software. On the other hand, Suárez-Vega et al. (2015) performed differential expression analysis with two R packages: DESeq2 and edgeR. Both analyses yielded a variable number of DE genes, with concordance rates between these two tools fluctuating between 6% to 65%. Another reason explaining this discrepancy would be physiological, as it is well known that the lactation curve of goats tends to be flatter, with a less prominent peak and greater persistency than that of cattle (http://www.fao.org/dairy-production-products/production/dairy-animals/small-

ruminants/en/), thus suggesting that the molecular basis of lactation in these two species is probably similar but not identical. Moreover, the sample source used for isolating RNA is also different. The studies carried out in cattle, sheep and buffalo (Suárez-Vega et al., 2015, Arora et al., 2019, Bhat et al., 2019) used RNA extracted from milk somatic cells, while we isolated RNA from biopsies of mammary gland tissues. In healthy goats, neutrophils, macrophages and lymphocytes constitute 45-74%, 15-41% and 9-20% of the somatic cell fraction, while epithelial cells represent a small portion of somatic cells in goat milk (Paape et al., 2001). This is the reason why we used mammary biopsies instead of extracting RNA from milk somatic cells, which can have a very heterogeneous composition.

Amongst the 42 DE genes detected by us when comparing T1 vs. T2, we have discussed several genes in paper I that participate in metabolism (e.g. *ST8SIA6*, *GALNT14*, *ADIPOQ*, *HKDC1*) as well as in mammary gland development and involution (e.g. *GABRB3*, *ARSI*, *INHBA*, *TNR*). There are also several genes that have been also identified in lactation studies focusing on buffaloes, dairy sheep and cattle, such as *INHBA*, *CPXM2*, *CCDC152* and *NOV*, which are known to participate in cell proliferation and apoptosis (Seder et al., 2009, Suárez-Vega et al., 2015, Yao et al., 2015, Lin et al., 2017). Moreover, a substantial 4-fold downregulation was observed in the mRNA expression of the *ATP4A* gene encoding ATPase H+/K+ transporting subunit α, which modulates proton pumps and ion channels (Singh et al., 2013). In the end of pregnancy, and also at the beginning and middle lactation, the caprine mammary gland experiences a steady growth, with a significant augmentation in the number of secretory cells and lumen area which result in a marked increment of the parenchyma tissue and mammary gland volume (Lérias et al., 2014). In contrast, in late lactation there is an important reduction of secretory epithelial cells due to involution and apoptosis (Lérias et al., 2014). The genes mentioned above might play a fundamental role in this process. For instance, the *INHBA* gene ($\log_2$FC = 1.63, *q*-value = 2.57×10^{-3}) encodes inhibin subunit beta A that is a

subunit of both activin and inhibin, which regulate mammary epithelial cell differentiation through mesenchymal-epithelial interactions (Robinson and Hennighausen, 1997). The *NOV* gene ($log_2FC = 1.57$, q-value = 0.03) encodes cellular communication network factor 3 (CCN3), a matricellular protein of the CCN family involved in the lactogenic differentiation of mammary epithelial cells (Perbal, 2004, Morrison and Cutler, 2010). Moreover, the CCN3 protein has a role in calcium ion signaling (Li et al., 2002). Other genes with strong expression changes between T1 and T2 encode haemoglobin subunits, such as *LOC102168680* ($log_2FC = 4.57$, q-value = 4.24×10^{-4}), *LOC102175876* ($log_2FC = 4.5$, q-value = 3.7×10^{-4}), *LOC102168959* ($log_2FC = 4.34$, q-value = 1.63×10^{-4}), a finding consistent with previous data (Bhaskaran et al., 2005, Newton et al., 2006, Chang et al., 2010). Mammary remodeling involves changes in the microvasculature and the expression of endothelial cells, thus ensuring the continuous supply of nutrients and oxygen to mammary epithelial cells and the removal of waste products (Safayi et al., 2010).

4.1.2 Molecular mechanisms modulating lactation are similar across species

4.1.2.1 Metabolic changes associated with lactation

Protein metabolism

In paper I, we mentioned that genes encoding caseins (CSN1S1, CSN1S2, CSN2, CSN3) and whey proteins (PAEP, LALBA) are significantly upregulated in T1/T2 (lactation) vs. T3 (dry-off), a result consistent with what has been observed in dairy cattle, sheep and buffalo (Bionaz et al., 2012b, Suárez-Vega et al., 2015, Arora et al., 2019, Bhat et al., 2019). The synthesis of milk proteins is mainly regulated by insulin signaling and mTOR signaling-related genes in the bovine mammary gland (Bionaz and Loor, 2011, Bionaz et al., 2012a). When we compared our results with those obtained by Bionaz et al.

(2011), we found that both the *SLC1A5* and *SLC7A5* genes are upregulated in T1 and T2 compared to T3. The *SLC1A5* and *SLC7A5* genes mediate the uptake of the essential Leu and Ile amino acids. Interestingly, in the bovine lactating mammary gland these two genes also play a central role in activating the downstream mTOR signaling pathway (Bionaz and Loor, 2011), because amino acids are fundamental activators of the mammalian target of rapamycin (mTOR) kinase which modulates protein translation, cell growth, and autophagy (Nicklin et al., 2009).

In rodents, the STAT5 protein, encoded by the *STAT5A* and *STAT5B* genes, is a critical player in the Jak2-Stat5 pathway regulating CSN3 synthesis (Wakao et al., 1994, Barash, 2006). In mice, STAT5 phosphorylation is high during late pregnancy, but very low in immature and non-lactating tissues, which is directly linked to the transcription of milk protein genes (Liu et al., 1996). In our study, however, these two genes did not reach the threshold of significance in any of the pairwise comparisons for differential expression analysis. It is important to emphasize, however, that in cattle the Jak2-Stat5 signaling pathway seems to play a minor role in milk protein synthesis (Bionaz and Loor, 2011). In contrast, we observed a downregulation of the *ELF5* mRNA at T3. This gene encodes a transcription factor with an essential role in the control of milk protein synthesis in cattle and mice (Zhou et al., 2005, Bionaz and Loor, 2011). Bionaz and Loor (2011) found that the *ELF5* mRNA is highly upregulated during lactation, being insulin the main modulator of its expression. In mouse mammary alveolar epithelial cells, *ELF5* mRNA expression appears to be regulated by prolactin, thus triggering the initiation of lactogenesis (Zhou et al., 2005). Our data combined with previous reports suggest that the ELF5 transcription factor is a major player regulating the metabolic changes associated with lactation in a broad array of species, while the exact role of STAT5 in ruminants needs to be further clarified. In this respect, it would be worth to look into the occurrence of post-translational modifications (mainly phosphorylation) associated with the lactation status.

Lipid metabolism

Many genes related with lipid metabolism showed an enhanced expression during lactation, being particularly important *FASN* which catalyzes the synthesis of fatty acids. We have also detected increased mRNA levels of genes related with fatty acid synthesis, triglyceride synthesis, cholesterol synthesis, sphingolipid synthesis, acetate synthesis and fatty acid activation, fatty acid desaturation, fatty acid absorption and transportation, formation of milk fat globules, lipolysis, and transcriptional regulation of lipid metabolism, a result consistent with those of Bionaz et al. (2012b). Pathway enrichment analysis also detected many biochemical routes related to lipid metabolism, including the PPAR signaling pathway. The PPARG transcription factor is an essential regulator of lipid metabolism in the bovine and mouse lactating mammary glands (Bionaz and Loor, 2008, Osorio et al., 2016). In bovine mammary cells, PPARG controls the expression of key lipid-related genes such as *ACACA*, *FASN*, *LPIN1*, *AGPAT6*, *DGAT1*, *SREBF1*, *SREBF2* and *INSIG1* (Bionaz and Loor, 2008, Kadegowda et al., 2009). Moreover, Shi et al. (2013) proposed that in lactating goat mammary epithelial cells, PPARG regulates the expression of the lipogenic *SCD* gene and contributes to monounsaturated fatty acid synthesis. According to our data, the expression of the *PPARG* gene did not reach the threshold of significance in the three pairwise comparisons, but as previously said PPARG signaling was detected in the pathway analysis. One important limitation of our study is that we just analyze mRNA expression, so changes in protein expression or post-translational modifications go unnoticed. Interestingly, we also found a significant downregulation of *PPARA* in T1 vs. T3 (log_2FC = -1.63, *q*-value = 1.78×10^{-18}). This gene encodes another isotype of peroxisome proliferator-activated receptors (PPAR) (Bionaz et al., 2013). Indeed, Tian et al. (2020) demonstrated that PPARA was able to promote the synthesis of monounsaturated fatty acids in primary goat mammary epithelial cells by stimulating the

expression of genes related to fatty acid synthesis, oxidation, transport, and triacylglycerol synthesis.

In the set of enriched pathways, we did not find any pathway regulated by the SREBP transcription factors 1 (SREBP1) and 2 (SREBP2), although they play central roles in regulating fatty acid import and trafficking, triacylglycerol and *de novo* fatty acid synthesis both in dairy cattle and sheep (Bionaz and Loor, 2008, Carcangiu et al., 2013, Osorio et al., 2016). Bionaz and Loor (2008) observed that the change of *SREBF1* mRNA expression during bovine lactation was ≤ 2-fold in magnitude, and suggested that PPARG, PPARGC1A, and INSIG1 might have a more pivotal role in milk fat synthesis. Noteworthy, the knockdown of caprine *PPARG* involves a 50% downregulation of key lipid genes, such as *SCD*, *DGAT1*, *AGPAT6*, *SREBF1*, *ACACA*, *FASN*, *FABP3*, *ATGL* (Shi et al., 2013), but Xu et al. (2016) also demonstrated that the overexpression of *SREBP1* in the goat mammary gland resulted in the increased mRNA expression of genes participating in the *de novo* synthesis of fatty acids, e.g. *ACSS2*, *ACLY*, *IDH1*, *ACACA*, *FASN*, *ELOVL6*; long-chain fatty acid activation, e.g. *ACSL1*; fatty acid transportation, e.g. *FABP3*; desaturation, e.g. *SCD*; lipid droplet formation, e.g. *LPIN1*, triglyceride synthesis, e.g. *DGAT1*, and transcriptional regulation, e.g. *NR1H3*, *PPARG*, *INSIG1*, *SCAP*. Therefore, in the absence of complementary proteomic data it is hard to figure out the relative importance of *PPARG* vs. *SREBF* in the control of lipid metabolism in the lactating mammary gland of goats.

Another key regulatory factor of lipid metabolism in goat mammary epithelial cells is the insulin-induced gene 1 (INSIG1). In our study, the *INSIG1* gene displayed a 4 to 5-fold upregulation during lactation (T1 and T2), a result consistent with that presented by Li et al. (2019). This finding is very consistent because Bionaz and Loor (2008) detected a 12-fold upregulation of the expression of this gene close to the peak of lactation. As we mentioned in paper I, the upregulated expression of the *INSIG1* gene during lactation is

counterintuitive because it would result in the decreased expression of genes participating in triacylglycerol, cholesterol synthesis, and lipid droplet accumulation (Li et al., 2019). Binding of SREBF Chaperone (SCAP) to INSIG proteins inhibits the delivery of SCAP/SREBP complex to the Golgi and downregulates the expression of SREBP target genes, leading to a reduction in cholesterol synthesis and uptake (Dong and Tang, 2010). Moreover, binding of HMG-CoA reductase to INSIG proteins leads to the degradation of the former (Dong and Tang, 2010). On the other hand, hepatic over-expression of INSIG1 in transgenic mice inhibits SREBP processing and reduces insulin-stimulated lipogenesis (Dong and Tang, 2010). So, INSIG1 has antilipogenic effects and paradoxically its expression is augmented during lactation, not only in goats but also in cattle (Bionaz and Loor, 2008). Our interpretation is that INSIG1 plays a key role in fine tuning lipid homeostasis in ruminants, thus ensuring that lipids are synthesized at adequate levels. Indeed, mutant hamster cells that are deficient in INSIG1, but not INSIG2, show partial defects in the regulation of reductase degradation and SREBF processing (Dong and Tang, 2010).

The expression of the *GLYCAM1* gene is also highly upregulated during lactation (T1 vs. T3: $\log_2$FC = -6.5, *q*-value = 5.72×10^{-14}; T2 vs. T3: $\log_2$FC = -6.93, *q*-value = 8.76×10^{-16}). This gene encodes glycosylation-dependent cell adhesion molecule-1, which was also highly expressed in the mammary gland of dairy cattle (Ibeagha-Awemu et al., 2016, Bu et al., 2017) and sheep (Suárez-Vega et al., 2015). The GLYCAM1 protein belongs to the glycoprotein mucin family and it is expressed in the ruminant mammary gland in a lactogenic-dependent manner under the influence of hormones such as prolactin, insulin, and steroids (Dowbenko et al., 1993, Le Provost et al., 2003). This molecule is a key component of the milk fat globule membrane (MFGM) (Sørensen et al., 1997, Lu et al., 2016a). Moreover, Lu et al. (2016a) found a high concentration of the GLYCAM1 and PP3 complex, which is able to inhibit lipolysis, in the large milk fat globule fraction. Other highly expressed genes during lactation are *XDH*, which encodes xanthine dehydrogenase; and *BTN1A1*, which encodes

butyrophilin subfamily 1 member A1. The XDH molecule could interact with the BTN1A1 protein and further contribute to lipid droplet secretion according to data presented by Bionaz and Loor (2008). Such interaction would also include the ADFP and MFGE8 proteins, which are major components of the MFGM in mammary gland epithelial cells. Due to the fact that XDH represents the most abundant protein in goat MFGM (~25%), and the ratio of expressed XDH/BTN1A1 in MFGM is higher than in other species, it could be concluded that XDH had a key role in the secretion of milk lipids from the goat mammary gland (Zamora et al., 2009, Lu et al., 2016b).

Carbohydrate metabolism and insulin signaling

The major sugar in milk is lactose, which is synthesized by using glucose and galactose as precursors. Lactose is an important energy-carrier and because of its osmotic properties it largely determines the amount of milk produced (Mardones and Villagrán, 2020). The synthesis of lactose is catalyzed by lactose synthase that is formed by α-lactalbumin and β-galactosyltransferase (Kuhn et al., 1980, Shendurse and Khedkar, 2016). Consistent with this fact, in paper I we observed a significant downregulation of the *LALBA* gene at T3. The expression of many genes encoding galactosyltransferases (*B3GALT1*, *B4GALT1*, *B3GALT5* and *B4GALT6)* was also downregulated during the dry period, reflecting the strong reduction in lactose synthesis that takes place during the dry-off period. Genes with key roles in glucose transport, such as *SLC2A1*, *SLC2A9*, also showed a decreased mRNA expression at T3. For instance, the *SLC2A1* gene encodes the insulin-dependent glucose transporter 1 (GLUT1) that mediates the transportation of glucose across the plasma membrane in order to synthesize lactose (Kuhn et al., 1980, Shendurse and Khedkar, 2016). Given that the role of glucose transporters is mostly tissue-specific, GLUT1 would be a predominant glucose transporter in the ruminant mammary gland (Nielsen et al., 2001, Zhao and Keating, 2007), and SLC2A9 might fulfill a similar function.

This probably explains why we did not observe a significant differential expression for other glucose transporters (e.g. GLUT3, GLUT4, GLUT5, GLUT8, GLUT12).

Insulin signaling through its receptor (INSR), insulin receptor substrate-1 (IRS1) as well as mTOR signaling pathway are critical for the coordinated metabolism of proteins, lipids and carbohydrates in the mammary gland. In this regard, we observed significant differences in the expression of the insulin receptor substrate 1 (*IRS1*) mRNA, which transfers insulin signals to insulin-like growth factor-1 receptors. More specifically, *IRS1* mRNA was downregulated in the comparison of T1 vs. T3 (log_2FC = -1.85, *q*-value = 1.6×10^{-11}) but not of T2 vs. T3. This latter result could be attributed to the stringent threshold used in our study because the downregulation of *IRS1* mRNA expression almost reaches significance in T2 vs. T3 (log_2FC = -1.45, *q*-value = 3.62×10^{-7}). Coincident results were presented by Bionaz and Loor (2011), who showed that *IRS1* was 2-fold upregulated during cattle lactation while insulin receptor followed a similar but less pronounced trend. Our findings and those of Bionaz and Loor (2011) are consistent with the fundamental role of insulin in lactogenesis. In this way, the mammary gland of knockout mice for the insulin receptor gene had 50% fewer alveoli at mid-pregnancy; and casein and lipid droplets were reduced by 60 and 75%, respectively, evidencing a role for insulin signaling both in alveolar development and differentiation (Neville et al., 2013). From a carbohydrate metabolism perspective, the increased expression of insulin-related genes during lactation would imply an increased uptake of glucose in the mammary gland through the augmented expression of the GLUT1 transporter (Bionaz and Loor 2011), thus enhancing lactose synthesis.

Mineral homeostasis

One of the essential components of milk is calcium, that is secreted by mammary epithelial cells and prevents the softening and the weakening of the bones of the newborn by ensuring an adequate mineralization. The transfer of calcium from blood to milk requires an optimal balance between the transport of calcium across the mammary epithelium and the supply of calcium (VanHouten, 2005). In cattle, there is substantial evidence that dietary calcium cannot maintain maternal calcium concentrations during lactation, so normocalcemia fundamentally relies on bone resorption (Hernandez, 2017). In our study, we have detected changes in the expression of several genes related with calcium homeostasis. Indeed, the increased expression of the *PTHLH* gene in the caprine lactating mammary gland might be interpreted as evidence of the mobilization of maternal skeletal calcium through the mechanism of increasing bone turnover (VanHouten, 2005). In cattle, serotonin enhances PTHLH function during bovine lactation to maintain maternal calcium homeostasis through bone resorption (Hernandez, 2017). We hypothesize that a similar mechanism operates in goats, but such assertion needs to be demonstrated yet. We have also detected a significantly upregulated (e.g. *TRPV5*, *TRPV6*) or downregulated (e.g. *TRPV1*) mRNA expression of members of the transient receptor potential channel family during the dry period. The decreased expression of *TRPV1* would be consistent with the abolishment of lactation and the consequent reduction of calcium uptake by the mammary gland, while the increased mRNA levels of *TRPV5* and *TRPV6* are less obvious to interpret from a biological point of view. Indeed, very little information (if any) exists about the role of transient receptor potential channel family members in lactation and mammary gland physiology.

Another fundamental component of milk is phosphorus, which, as calcium, is necessary for the mineralization of the skeleton of the neonate. Indeed, the major constituent of bones is hydroxyapatite, which is a mineral complex between calcium and phosphorus. In T3 (dry-off), we detected a 2 to 3-fold upregulation of the *FGF23* mRNA, that encodes fibroblast growth factor 23. In mice, increased levels of the *FGF23* mRNA result in phosphaturia,

hypophosphatemia, low serum 1,25-dihydroxyvitamin D levels, and rickets and osteomalacia as well as hyperparathyroidism (Martin et al., 2012). So, the lower levels of *FGF23* mRNA in the mammary gland of the lactating goat would be consistent with the physiological need of ensuring the retention of as much as phosphorus as possible. There are indications that insulin promotes the downregulation of FGF23 (Bär et al., 2018), thus reinforcing the notion that this hormone is crucial in all aspects of lactation.

4.1.2.2 Tissue remodeling, cell death and involution

The cessation of milking triggers mammary gland involution, a process involving cell death and extensive tissue remodeling in order to bring the mammary gland back to a non-lactating state (Zhao et al., 2019). The prerequisite of a successful mammary gland involution is to break down milk components and remove them from the mammary gland (Zhao et al., 2019). In this regard, we found an upregulated mRNA expression of the 5-hydroxytryptamine receptor 1D (*HTR1D*) in the comparison of T1 vs. T3 ($log_2FC = 2.85$, q-value $= 3.77 \times 10^{-3}$), and of the plasminogen activat*or* (*PLAT*) in the comparisons of T1 vs. T3 ($log_2FC = 1.93$, q-value $= 5.58 \times 10^{-16}$) and T2 vs. T3 ($log_2FC = 1.6$, q-value $= 2.70 \times 10^{-6}$). The *HTR1D* gene encodes a serotonin receptor that is able to inhibit milk synthesis (Hernandez et al., 2008), while the PLAT molecule is an activator in the plasmin-plasminogen-plasminogen system, which promotes the breakdown of casein subtypes (Zhao et al., 2019). Consistently, the expression of plasmin, plasminogen, and plasminogen activator also increases in response to mammary gland involution at drying-off in dairy cattle (Aslam and Hurley, 1997, Athie et al., 1997).

It is reported that the activator protein 1 (AP-1) plays key roles in initiating and executing apoptosis to ensure an adequate mammary gland involution (Jaggi et al., 1996, Marti et al., 1999). In this regard, we observed the

upregulated mRNA expression of the *FOS* and *JUNB* genes at T3, both encoding subunits of the AP-1 dimeric transcription factor (Marti et al., 1999). The early stage of mammary gland involution is an apoptosis-only process, which would be promoted by the Jak/Stat pathway. Two important transcription factors in this pathway, STAT5 and STAT3, showed downregulated and upregulated mRNA expression during the dry period in cattle, respectively. The STAT5 protein plays a key role in regulating the synthesis of milk components during lactation as mentioned above, while STAT3 generally induces mammary involution and apoptosis at drying-off (Stein et al., 2007, Zhao et al., 2019). However, our study did not detect the existence of differential expression for these two transcription factors. It is quite possible that the activity of STAT3 and STAT5 is mostly mediated by post-translational changes rather than by modifications in their mRNA levels (Johnston et al., 1995). During mice mammary involution, the expression of STAT3 is activated by the leukemia inhibitory factor (LIF), suggesting that LIF is a key upstream regulator of STAT3 and a mediator of mammary epithelial cell death (Hughes and Watson, 2018). In paper I, we have observed that the *LIF* gene is significantly upregulated in the mammary gland of MUG goats at T3, an observation consistent with such role. Moreover, we found a downregulation of the *IGFBP5* gene (T2 vs. T3: $\log_2$FC = -1.76, *q*-value = 2.01×10^{-5}), which encodes insulin-like growth factor-binding protein 5, a target of the STAT3 molecule. Indeed, the IGF-IGFBP system is another essential regulator of mammary gland involution, and IGFBP5 seems to be a fundamental mediator that regulates IGF1 hormonal signaling and stimulates cell proliferation (Allan et al., 2004).

The progression of mammary gland involution induces morphological changes in epithelial cells and mammary tissue, which includes degradation of the extracellular matrix and basement membrane (Zhao et al., 2019). This is a key step in which the size and number of ducts in the mammary gland decrease markedly, and epithelial cell numbers also become drastically reduced, finally resulting in the loss of approximately 50% of epithelial cells (Zhao et al., 2019).

During the dry period, we observed increased mRNA levels of genes encoding matrix metalloproteinases (MMP), such as *MMP1*, *MMP2*, *MMP3*, *MMP7*, *MMP12*, *MMP13*, *MMP14*, *MMP17*, and *MMP19*, a result also reported in dairy cattle, sheep and buffaloes (Rabot et al., 2007, Suárez-Vega et al., 2015, Arora et al., 2019, Bhat et al., 2019). Indeed, matrix metalloproteinases (MMP) play a major role in degrading extracellular matrix and basement membranes, as well as in remodeling mammary gland architecture. For instance, MMP3 is able to degrade the basement membrane, thus causing a substantial enhancement of apoptosis (Uria and Werb, 1998, Zhao et al., 2019). Similarly, both MMP2 and MMP9 possess gelatinolytic activity and lead to the degradation of the basement membrane (Uria and Werb, 1998). However, the activation of these MMP can be inhibited by the tissue inhibitor of metalloproteinases (TIMP) (Zhao et al., 2019). Indeed, the MMP:TIMP ratio decides the fate of the basement membrane and the extracellular matrix as well as the remodeling process of the mammary gland after abrogation of lactation (Zhao et al., 2019). In this way, an increased MMP:TIMP ratio would enhance the process of mammary gland degradation and remodeling (Zhao et al., 2019). Murphy (2011) demonstrated that the TIMP molecule can inhibit the action of disintegrin metalloproteinases (ADAM and ADAMTS). Although we did not find differential expression for *TIMP* mRNA in our study, we detected an upregulated expression of the *ADAMTS4*, *ADAMTS7*, *ADAMTS16* genes during the dry period. This suggests that disintegrin metalloproteinases might play a key role in the apoptosis of mammary epithelial cells (Murphy, 2011).

The proteolytic degradation of the extracellular matrix by MMP would result in reduced communication between epithelial cells and the extracellular matrix, which induces and accelerates the impairment of tight junction integrity (Zhao et al., 2019). The loss of communication between mammary epithelial cells and the extracellular matrix is generally promoted by the downregulation of integrins (Zhao et al., 2019). In paper I an upregulation of the *ITGB6* gene ($\log_2$FC = 2.18, *q*-value = 4.27×10^{-9}) was found at T3. This gene encodes β6-

integrin that generally forms complexes with integrin alpha-V inducing active proliferation and impairing apoptosis (Liang et al., 2017). However, its upregulation is often coupled with the downregulation of αvβ5 integrin complex (Janes and Watt, 2004). The final step of mammary involution involves the removal of casein micelles, lipid droplets, and cellular debris. In paper I, we observed a 3 to 4-fold downregulation of the glycoprotein milk fat globule epidermal growth factor 8 (*MFGE8*) mRNA at T3, which has antiapoptotic effects (Gao et al., 2018) but also regulates the clearance of apoptotic epithelial cells (Atabai et al., 2005). It should be emphasized that the clearance of apoptotic cells largely relies on phagocytosis, a process mainly driven by cells of the immune system, thus we will discuss it in the next section ("Mammary immunity").

4.1.2.3 Mammary immunity

After the abrogation of lactation, mammary gland enters into the dry period, during which the mammary gland is particularly vulnerable to pathogens and easily suffers from new infections, especially in the beginning of the drying-off (Jain, 1979, Sordillo and Streicher, 2002). The infection of the mammary gland results in the development of an inflammatory response (mastitis), which can have adverse consequences on both health and production. It is reported that the main factor increasing the rate of mastitis is the invasion of bacterial pathogens into the mammary gland (Jain, 1979, Katsafadou et al., 2019).

To protect the mammary gland against infection, both innate and specific immunity work together in a synergistic way. As previously mentioned, the abrogation of lactation initiates mammary gland involution, during which casein micelles, lipid droplets, and cellular debris are removed, and the mammary epithelium is renewed (Zhao et al., 2019). After the migration and recruitment of neutrophils into the mammary gland, proinflammatory cytokines (IL1β, IL6,

and IL17) promote the initiation of inflammatory responses (Alnakip et al., 2014). In this regard, we found an upregulation of the mRNA expression of the *CCL20* chemokine in the non-lactating mammary gland of MUG goats, possibly supporting a role of the immune system in the clearance of apoptotic epithelial cells (Zhao et al., 2019). In paper I, the upregulation of the *TNF* mRNA expression at T3 (T1 vs. T3: $\log_2$FC = 1.78, *q*-value = 8.9×10^{-6}) evidences its key role in mammary involution. In mice, TNF is deeply involved in the apoptosis of mammary cells after weaning (Kojima et al., 1996), but this proinflammatory cytokine is also essential in the host defense against pathogens (Pfeffer, 2003) and it has also been reported to promote the growth and development of the mammary gland (Varela and Ip, 1996), thus illustrating its multifunctional role in mammary physiology.

The cessation of lactation is usually followed by the enhancement of humoral defenses, which leads to an increased expression of genes related with the complement cascade (e.g. *C6*, *C7*, *C1R*, *C1QA*, *C1S*, *CTSL*), cytokines (e.g. *IL5*, *IL15RA*, *IL22RA2*), lysozymes (e.g. *LYG2*), and chemokines (e.g. *CXCR4*). It is reported that the concentration of complement is lowest in healthy milk during lactation, but at drying-off higher concentrations are reached (Alnakip et al., 2014). Complement has versatile effects on the immune response, including the recruitment of phagocytes, opsonization of bacteria, initiation of inflammation, and the digestion and killing of microorganisms (Wellnitz and Bruckmaier, 2012, Alnakip et al., 2014, Katsafadou et al., 2019). Moreover, we observed a 9 and 3-fold downregulation of lactoperoxidase (*LPO*) and myeloperoxidase (*MPO*) mRNAs at T3, respectively. Lactoperoxidase accounts for 0.5% of the total whey proteins and it has antibacterial effects on gram-negative and gram-positive bacteria (Alnakip et al., 2014). Myeloperoxidase is a peroxidase enzyme which is mainly located in neutrophil granulocytes and it has a strong antimicrobial activity mediated through the generation of hypohalous acids (Alnakip et al., 2014). Since these two enzymes are secreted into milk, it is

logic that the expression of these two genes decreases markedly in the dry-off period.

In paper I, we also observed the downregulation of genes related with mucins (e.g. *MUC1*, *MUC4*, *MUC20*) and surfactant molecules (e.g. *ABCA3*, *SFTPD*) at T3. Surfactant is a lipoprotein complex that plays a critical role in immunity (Pastva et al., 2007). Surfactant-D (SFTPD) is related to the inhibition of inflammation and enhancement of pathogen clearance (Pastva et al., 2007). On the other hand, mucins are a family of highly glycosylated glycoproteins expressed in the epithelial surface of the mammary gland (Patton et al., 1995). Though mucins constitute in many tissues (e.g. intestine) a chemical barrier to protect mucosal surfaces against bacterial infection, they are also able to inhibit bacterial invasion of epithelial cells (Liu et al., 2012). In this respect, mucins act as adhesion decoys thus exerting an antimicrobial activity, both MUC1 and lactadherin maintain their integrity in the stomachs of infants to prevent infection (Peterson et al., 1998, Liu et al., 2012). This may be relevant because in our study we observed a consistent downregulation of the *MUC1* and lactadherin (i.e. *MFGE8*) mRNAs. Possibly, low levels of mucins are associated with the cessation of lactation because mucins are part of the milk fat globule membrane (Schroten, 2001). On the other hand, aberrantly secreted mucins might impair the efficacy of the antimicrobial response, thus suggesting a potential role of mucins in mammary immunity (Linden et al., 2008).

4.2 Genetic determinism of milk traits in Murciano-Granadina goats

4.2.1 The casein gene cluster is strongly associated with milk protein content

In our GWAS study, the most significant quantitative trait locus (QTL) was associated with protein percentage and mapped to a caprine chromosome 6

region containing the casein gene cluster (Martin et al., 2017). A study carried out in 89 MUG individuals with 316 phenotypic records revealed that *CSN1S1* genotype was not associated with total casein and protein contents, while *CSN3* genotypes (AB and BB) had significant effects on total casein and protein contents (Caravaca et al., 2009). Caravaca et al. (2011) reported that the BB, EE and EF genotypes of the *CSN1S1* locus display similar associations with cheese yield, but the $CSN1S1^{EE}$ genotype was associated with a higher milk curdling rate when compared to the $CSN1S1^{BB}$ genotype. These findings suggested that the *CSN1S1* locus has a weak effect on milk traits in MUG goats, a result that is not concordant with what has been observed in French goats (Mahé et al., 1994, Manfredi et al., 1995). In the French Saanen and Alpine breeds, genetic polymorphisms in the *CSN1S1* locus were reported to have significant effects on protein, casein, and fat production, as well as on milk rheology and organoleptic properties of cheese (Mahé et al., 1994, Manfredi et al., 1995, Caravaca et al., 2009). It should be noticed that sample sizes used by Caravaca et al. (2009, 2011) were modest, so their results might not be completely conclusive. As a matter of fact, Caravaca et al., (2008) demonstrated that the BB genotype in the *CSN1S1* locus significantly increased CSN1S1 levels by performing an association study comprising 138 MUG goats with 460 phenotypic records. By carrying out a non-parametric association analysis between 48 single nucleotide polymorphisms (SNPs) mapping to the four casein genes (*CSN1S1*, *CSN1S2*, *CSN2* and *CSN3*) and dairy phenotypes recorded in 159 MUG goats, Pizarro Inostroza et al. (2019) were able to conclude that SNPs in the *CSN1S1* and *CSN3* genes have major effects on protein percentage. Our GWAS does not allow us to identify exactly which polymorphisms in the casein cluster have causal effects on protein percentage, but an experiment to finely map these causal mutations is currently under way. Moreover, Pizarro Inostroza et al. (2019) identified an effect of *CSN1S1* genotype on fat percentage, a finding concordant with ours. The effects of casein loci on fat content in caprine milk was also observed in Saanen goats (Martin et al., 2017), suggesting that the synthesis of milk protein and lipid

components is, to some extent, coupled in the mammary gland (Bionaz and Loor, 2008, 2011). As inferred by systems biology analysis, a cross regulation of milk fat, protein and carbohydrate synthesis exists, being a central hub the mTOR signaling pathway (Osorio et al., 2016).

4.2.2 Other loci related with milk traits

At the genome-wide level, we obtained two additional significant associations, i.e. QTL1 on chromosome 2 (130.72-131.01 Mb) for lactose percentage and QTL17 on chromosome 17 (11.20 Mb) for both protein and dry matter percentages. The QTL1 region contains the NGFI-A binding protein 1 (*NAB1*) gene that shows an increased expression during mouse lactation (Yang et al., 2004) and it is involved in the metabolism of lactose, a disaccharide formed by glucose and galactose. Interestingly, NAB1 is a corepressor of the EGR-1 transcription factor (Swirnoff et al., 1998), and in mouse the promoter of the galactokinase gene, which forms part of the Leloir pathway converting galactose into glucose, has binding sites for EGR-1 (Yang et al., 2004). The QTL17 region is represented by a single SNP (rs268238952), that is located in the downstream part of the T-Box 3 (*TBX3*) gene. This gene is highly expressed in luminal cells during early mammary gland initiation and its inactivation is associated with the ulnar mammary syndrome characterized by hypoplastic or aplastic breasts and supernumerary, inverted, or absent nipples with an inability to lactate (Eblaghie et al., 2004, Platonova et al., 2007, Douglas and Papaioannou, 2013). However, no direct link with milk protein metabolism has been reported and, moreover, QTLs represented by a single SNP are sometimes statistical artifacts, so this result should be interpreted with caution.

At the chromosome-wide level, we identified QTLs related with lactose percentage, including QTL11 on chromosome 14 (46.1 Mb); QTL19 on chromosome 20 (29.45 Mb), which co-localizes with the *HCN1* gene; QTL22 on

chromosome 24 (49.71 Mb), which is supported by a single SNP (rs268240589) located in the *MYO5B* gene; and QTL23 on chromosome 28 (23.03 Mb), which physically overlaps the *CTNNA3* gene. Among them, the *MYO5B* gene encodes myosin VB, which regulates glucose metabolism genes and thus affects blood glucose levels (Cartón-García et al., 2015, Tomić et al., 2020). In bovine mammary epithelial cells, glucose upregulates the expression of genes such as *GLUT1*, *SLC35A2*, and *SLC35B1*, thus enhancing its own uptake and favoring lactose synthesis (Lin et al., 2016). In addition, we found two genes with important roles in fatty acid metabolism in the upstream region of the QTL22, i.e. *LIPG* (~400 kb) and *ACAA2* (~45 kb). The latter gene encodes an acyl-CoA transferase, which is able to catalyze fatty acid β-oxidation (Dunning et al., 2014). Increased fatty acid β-oxidation provides substrates for the gluconeogenesis pathway that yields glucose, one of the components necessary for lactose synthesis in the mammary gland (Orford et al., 2012). In dairy sheep, the polymorphism of this gene has been associated with milk yield in the Chios breed (Orford et al., 2012).

Regarding to somatic cell count, we would like to highlight QTL14 (rs268266747, chromosome 15: 63.95 Mb), which maps to the DEAD-box helicase 10 (*DDX10*) gene that plays a key role in innate immunity (Morerio et al., 2006) and in viral resistance (Lee et al., 2015); and QTL9 which mapped to chromosome 13 (53.62-54.38 Mb). This latter genomic region contains the ADP ribosylation factor GTPase activating protein 1 (*ARFGAP1*) gene that plays a critical role in regulating rearrangements of the actin cytoskeleton to impede the entry of mycobacteria into epithelial cells (Song et al., 2018). With regard to the length of lactation, we detected QTL2 on chromosome 3 (113.47 Mb), QTL7 on chromosome 11 (72.83 Mb) and QTL8 on chromosome 12 (68.1 Mb). The QTL7 overlaps the dynein regulatory complex subunit 1 (*DRC1*) gene, encoding an essential cell cycle-regulated molecule required for DNA replication (Wang and Elledge, 1999). This gene showed a significant differential expression in early lactation (T1) vs. drying-off (T3) ($\log_2$FC = 2.55, *q*-value = 2.54×10^{-7}), as well

as in T2 (end of lactation) vs. T3 ($log_2FC = 1.71$, q-value $= 6.22 \times 10^{-4}$) (paper I). In addition, DRC1 forms part of the dynein regulatory complex which plays a major regulatory role in the motility of cilia and flagella (Wirschell et al., 2013). It has been proposed that primary non-motile cilia may intervene in the involution of the bovine mammary gland by transducing chemosensory and mechanosensory signals produced by milk accumulation, and the consequent distension of the udder (Biet et al., 2016).

4.2.3 Genetic heterogeneity of milk traits in Murciano-Granadina goats

The concordance rate of the 24 QTLs associated with milk traits in MUG goats and those identified in French goats is quite low (Martin et al., 2017, Mucha et al., 2017, Martin et al., 2018). Consistently, the positional overlapping of QTLs identified in Alpine and Saanen breeds by Martin et al. (2017) was lower than 50% despite the fact that these two breeds are closely related. Through a functional approach, Martin et al. (2017) demonstrated that missense mutations p.Arg396Trp and p.Arg251Leu in the *DGAT1* gene affect activity the DGAT1 enzyme which is fundamental for triglyceride synthesis, and this is why these two missense polymorphisms are highly associated with milk fat content in Saanen and Alpine goats. However, we did not find any association with milk fat content at or near the *DGAT1* locus, thus suggesting that these two polymorphisms do not segregate (or segregate at very low frequencies) in MUG goats. One of the reasons for the heterogeneity in the genetic determinism of dairy traits in different goat breeds could be genetic differentiation amongst breeds due to drift and demographic factors. By analyzing Y-chromosomal markers, Vidal et al. (2017) found that Spanish goats displayed quite different haplotypes (mainly Y2) than those harbored by goats from Central and East Europe (Y1A, Y1B1, Y1B2 and Y1C). Based on autosomal markers genotyped with the Illumina Goat SNP50 BeadChip, Colli et al. (2018) showed that French and Spanish goats are clustered into two different clades, and a different genomic

composition was observed in the ADMIXTURE analysis. Another reason for low positional concordance between GWAS signals across breeds and populations could be limited sample size. In other words, by increasing sample size GWAS hits could be more concordant when comparing the results obtained in different populations. Wojcik et al. (2019) made a GWAS for 26 clinical and behavioral phenotypes in 49,839 non-European individuals and 1,444 were concordant with those obtained in Europeans, while only 27 novel loci and 38 secondary signals at known loci were identified (Wojcik et al., 2019).

4.2.4 Modest positional concordance between differentially expressed genes and GWAS signals

We found that the positional concordance between protein-coding genes identified as DE in the RNA-Seq analysis and GWAS signals is low. Many factors could contribute to this result. For instance, we used a medium density Goat SNP50 BeadChip and sample size was also modest, thus limiting our ability to fully discover all dairy QTLs. Indeed, 20 out of 24 QTLs are only supported by a single SNP, meaning that further studies will be needed to confirm their existence. Limited number of samples in the RNA-Seq experiment and the high stringency of the differential expression analysis may also cause this low concordance. Another explanation is that the set of genes that triggers and maintains lactation and the set of genes containing polymorphisms with causal effects on dairy traits is, to some extent, different. For instance, a SNP or an indel could create or suppress a phosphorylation site changing drastically the activity of a protein, but this would not be detected in a differential expression analysis. On the other hand, it is clear to us that understanding the effects of polymorphisms on phenotypes requires the generation of massive amounts of biological information, either through differential expression analysis or by using many other techniques and integrating different sources of information.

4.3 A substantial fraction of the variation of the goat casein genes was generated before domestication

We have investigated extant casein diversity in goats and bezoars (paper II) by using 106 whole genome sequences and 51 million SNPs. By exclusively focusing on casein loci, we found that a substantial part of polymorphisms was shared between goats and bezoars. This extensive sharing of variations between bezoars and domestic goats could be due to introgression events, which are common in *Capra* species. For instance, a recent study carried out by Zheng et al. (2020) demonstrated an ancient gene flow from a West Caucasian tur species to bezoars. These authors identified an introgressed locus encompassing the *MUC6* gene which affects resistance to gastrointestinal nematodes (Zheng et al., 2020). Likewise, Alpine ibex (*Capra ibex ibex*) was also introgressed by domestic goats (Grossen et al., 2014). In our study (paper II) we did not detect ancestral gene flow between goats and bezoars, but it should be noticed that the ADMIXTURE tool is designed to detect recent admixture events (Alexander et al., 2009, Alexander and Lange, 2011, Lawson et al., 2018). Nevertheless, the most plausible reason for the extensive sharing of polymorphisms between goats and bezoars is recent divergence. Indeed, goats diverged from their wild ancestor only 10,000 YBP, which is a very short time on an evolutionary scale (Amills et al., 2017). Certainly, many studies have demonstrated that standing genetic variation substantially contributes to adaptation to new environments and to animal domestication (Larson et al., 2014, Ramos-Onsins et al., 2014, Lai et al., 2019).

By reconstructing allelic variants of the casein genes, we found differences in the frequencies of casein alleles in goats with different geographic origins. Especially, we found several unreported casein alleles in Far Eastern

goats. Although the low number of samples (~20 samples per geographic group) could be one of the reasons for the observed differences in allele frequencies, it should be noted that goat domestication took place in several different locations in the Near East (i.e. there were multiple centers of domestication) and different gene pools were dispersed into Africa, Europe and Asia (Daly et al., 2018). This is reflected in the ADMIXTURE analysis of modern goat breeds based on 46,654 autosomal SNPs (Colli et al., 2018) as well as in the principal component analysis (PCA) and neighbor-joining (NJ) tree presented in paper II. The dispersal of different gene pools to Africa, Europe and Asia would partly explain the differences in allelic frequencies that we have detected amongst goats with different continental origins. Moreover, differences in allelic frequencies could be also due to selection for different production purposes. Chinese breeds are bred for cashmere (northern) and meat (southern) production, while many European breeds are devoted to the production of milk to manufacture cheese. While the associations of casein alleles with dairy traits have been extensively studied in European breeds, there is a large gap in our knowledge about the associations of casein polymorphisms and milk yield and composition traits in African and Asian breeds. Filling this gap could be a future avenue of research that could have relevant practical applications in the selection of genotypes with beneficial effects on dairy traits.

4.4 A first assessment of copy number variation in the Murciano-Granadina breed

4.4.1 Experimental factors influencing the discovery of copy number variations

Although we have used a more stringent pipeline than that reported by Liu et al. (2018), who described approximately 20 copy number variations regions (CNVR) per breed (Liu et al., 2018), we have been able to identify 486 CNVR in the MUG breed. This difference could be explained by the lower sample size (N=4-53) used by Liu et al. (2018) to characterize each one of the breeds included in their study, while in our survey we have investigated 1,036 MUG individuals. We conclude that Liu et al. (2018) underestimated the true levels of CNV diversity in goat breeds, but our measurement is also probably an underestimate. This is because SNP arrays have a very low power to detect small CNV, which are the most abundant ones, and, at the same time, the rate of false positives can be also high (Yau et al., 2009, Pinto et al., 2011). Based on pooled whole genome sequences, Zhang et al. (2019) identified 2,056 and 2,153 CNV respectively in low (N = 20 with a single offspring) and high fertility goats (N = 14 with litter size of 3 or 4) from the Laoshan dairy breed. Moreover, Dong et al. (2015), on the basis of individual whole genome sequences of six domestic goats and two wild counterparts, reported 13,347 CNV. Likewise, a genomic resequencing study detected a total of 2,317 CNV in three individuals from Saanen, Liaoning cashmere, Leizhou goat, and two Sindh ibex (*Capra aegagrus blythi*) and Markhor (*Capra falconeri*) individuals (Zhang et al., 2018). These studies detected a higher number of CNV than those reported by us despite the fact that they just included a few individuals (Dong et al., 2015, Zhang et al., 2018, Zhang et al., 2019). This implies that experimental resolution is a critical factor determining the discovery of CNV. In this respect, we could expect that long-read sequencing technologies would substantially contribute to the discovery of structural variations (SVs) in the near future (Ho et al., 2019, Mahmoud et al., 2019). In paper III, we also found that the majority of the CNVR (72.6%) showed allele frequencies lower than 0.01, an estimate that exceeds values obtained in cattle (Upadhyay et al., 2017) but that is similar to that described in sheep (Yang et al., 2018) and goats (Liu et al., 2018). This difference could be due to the fact that studies by Yang et al. (2018), Liu et al. (2018) and

us included 2,254, 1,023 and 1,036 samples, respectively, while Upadhyay et al., (2017) only employed 196 animals from 38 different cattle breeds (so they were unable to detect CNV with low or very low frequencies). This outcome evidences that low-frequency CNV events, which are the most abundant ones, cannot be discovered unless large populations are used.

4.4.2 Landscape of copy number variation in Murciano-Granadina goats

Copy loss (353) is more prevalent than both copy gain (78) and copy gain/loss (55) events in MUG goats, a finding concordant with results obtained in sheep and cattle (Komura et al., 2006, Liu et al., 2010, Zarrei et al., 2015, Yang et al., 2018). This is probably caused by the existence of a mutational bias toward deletions, as shown in the analysis of *Drosophila melanogaster* genomes (Leushkin et al., 2013). One relevant observation drawn in paper III is that the majority of copy number variable genes belong to large multigene families, such as olfactory receptors and ABC transporters. This may be caused by the fact that genomic regions with highly similar paralogous sequences (probably segmental duplications) would facilitate the emergence of CNV by non-allelic homologous recombination (NAHR). This process makes it possible to transfer a copy from a chromosome to its sister chromosome and simultaneously remove the original copy (Bickhart and Liu, 2014). It is reported that paralogs derived from gene duplications can have important roles in the adaptation to new environments by generating new functions once they diverge from the original copy (Kondrashov et al., 2002, Sudmant et al., 2015). Moreover, fork stalling and template switching (FOSTES), mobile element insertion (MEI), and non-homologous end-joining (NHEJ) could also mediate the emergence of CNV (Hastings et al., 2009, Bickhart and Liu, 2014). Noteworthy, CNV tend to occur in complex genomic regions (i.e. hotspots), rather than having a uniform distribution in the genome (Hastings et al., 2009). One of the most interesting copy number variable

genes is *ASIP*, which encodes the agouti signaling protein. This molecule has an essential role in the synthesis of yellow/red pheomelanin by binding to the melanocortin 1 receptor (Wolf Horrell et al., 2016, Caro and Mallarino, 2020). The segregation of an *ASIP* CNVR in MUG goats is not fully consistent with the hypothesis that increased *ASIP* copy number is associated with a white pigmentation in goat breeds (Fontanesi et al., 2009b, Dong et al., 2015, Zhang et al., 2018), as previously demonstrated in sheep (Norris and Whan, 2008). Our qPCR experiment (paper IV) demonstrated the existence of increased *ASIP* copy number in black/brown MUG goats and blond/brown Malagueña goats, indicating that the involvement of *ASIP* SV in goat pigmentation has not been fully clarified. A potential interpretation is that the extra copies of the caprine *ASIP* locus could have combinatorial effects with other mutations. A similar case was reported for the porcine *KIT* locus, in which at least four types of SVs and a splice mutation were detected, and the solid white coat would appear only when two types of duplications and the splice mutation are present concurrently (Rubin et al., 2012, Wu et al., 2019). Indeed, there are at least four types of SVs reported in the caprine *ASIP* locus, possibly associated with different pigmentation patterns (Henkel et al., 2019). Due to this high structural complexity, the expression pattern of the caprine *ASIP* gene becomes very complex, with at least nine different transcripts detected by using a RNA-Seq approach (Henkel et al., 2019). Our data, in summary, indicate that a simple and linear relationship between *ASIP* copy number and white pigmentation in goats does not exist, and that further studies need to be carried out in order to understand whether *ASIP* SV affects the *ASIP* mRNA levels and melanin synthesis.

4.5 Coat color in the Murciano-Granadina breed is explained by *MC1R* genotype

In addition to the discovery of the CNV co-localizing with the *ASIP* gene, we detected a CNV overlapping with the *ADAMTS20* gene, which was also found by Dong et al. (2015) and Liu et al. (2018). The *ADAMTS20* gene is essential for melanoblast survival (Silver et al., 2008) and co-localizes with a signature of selection when comparing white vs. colored goats (Bertolini et al., 2018). These evidences indicate that both *ASIP* and *ADAMTS20* could have a key role in goat pigmentation. Fontanesi et al. (2009a) reported that the *MC1R* genotype (c.801C>G, p.Cys267Trp) is tightly associated with the color of MUG goats, but they performed a candidate gene study with a low number of MUG goats making it impossible to discern the existence of additional genetic factors influencing the pigmentation pattern of this breed. Therefore, we conducted a GWAS comprising 529 MUG goats with brown and black colors, and by doing so we obtained a very clear GWAS signal encompassing the *MC1R* gene on chromosome 18 (paper V). Indeed, the *MC1R* gene is quite polymorphic and plays an essential role in regulating the synthesis of eumelanin/pheomelanin pigments (Wolf Horrell et al., 2016, Andersson, 2020, Orteu and Jiggins, 2020). By carrying out a sequencing experiment, we identified a missense mutation (c.801C>G, p.Cys267Trp) fully explaining the pigmentation patterns of MUG goats (paper V). Although the nature of our results is fundamentally confirmatory of previous data published by Fontanesi et al. (2009a), we demonstrate that the inheritance of coat color in MUG goats is fundamentally monogenic (Fontanesi et al., 2009a). Generally, animal pigmentation patterns are explained by one or few loci rather than having a polygenic basis (Georges et al., 2018). The inheritance of pigmentation can be quite complex and it is often oligogenic or, more rarely, polygenic. For instance, Menzi et al. (2016) revealed that the white pigmentation pattern in Boer goats could be explained by a CNV mapping to the *EDNRA* gene that regulates the levels of this molecule in a dose dependent manner. In contrast, the white spotting of cattle was associated to three QTLs in chromosomes 2, 6 and 22, respectively co-localizing with the *PAX3*, *KIT* and *MITF* genes (Jivanji et al., 2019). In humans, blond hair

pigmentation was explained by at least 200 loci (Morgan et al., 2018), illustrating that in certain cases the inheritance of pigmentation can be very complex. A next step for our study would be to demonstrate the causality of the Cys267Trp missense substitution by carrying out a functional test measuring the activity of MC1R in cultured cells with alternative genotypes for this mutation and the consequences of differential activity, if any, on melanin synthesis.

Chapter 5

Conclusions

The main conclusions of the present Ph.D. Thesis are as follows:

1. The abrogation of lactation in goats involves an important reduction in the mRNA expression of genes related with protein, lipid and carbohydrate biosynthesis and transportation. We have also detected an increased mRNA expression of genes involved in apoptosis and tissue remodeling, while the mRNA levels of several known survival factors were reduced, an observation consistent with the physiological changes associated with mammary involution and the extensive loss of epithelial cells associated with it. With regard to mammary immunity, the cessation of lactation involved changes in the mRNA expression of complement, cytokine, mucin and surfactant genes.

2. The performance of genome-wide association studies resulted in the identification of 3 genome-wide significant QTLs on chromosomes 1 (lactose percentage), 6 (protein percentage) and 17 (protein and dry matter percentages). The QTL6 region contains the casein genes, the polymorphism of which has been associated with the variation of milk traits in French and Spanish breeds. We have also identified 21 QTLs that were significant at the chromosome-wide level. Comparison of our QTL data with previous results obtained in French goats revealed a low level of positional concordance, a feature that could be due to technical factors (mainly limited sample size) and/or to the existence of a substantial degree of genetic heterogeneity.

3. The substantial sharing of single nucleotide polymorphisms (SNPs) and haplotypes mapping to casein loci between bezoars and domestic goats as well as between goat populations with different continental origins indicates that an important fraction of the extant diversity in the casein genes of domestic goats is derived from standing variation segregating in bezoars well before domestication.

4. The assessment of copy number variations (CNV) in a population of 1,036 Murciano-Granadina goats revealed a substantial degree of diversity. The set of genes mapping to CNV regions was functionally related with sensory perception, metabolism and embryo development, and in general these genes belonged to large multigene families with tens, hundreds or thousands of paralogs, a feature that might favor the emergence of CNV by nonallelic homologous recombination.

5. Measuring the copy number of the agouti signaling protein (*ASIP*) gene in eight goat breeds with different pigmentation patterns revealed the segregation of an *ASIP* CNV not only in white breeds, such as Saanen, but also in the black/brown Murciano-Granadina and the blond/brown Malagueña breeds. These results evidence the lack of a simple linear relationship between increased *ASIP* copy number and goat white pigmentation.

6. Performance of a genome-wide association study for coat color in Murciano-Granadina goats demonstrated that the inheritance of this trait is fundamentally monogenic, and confirmed a tight association between the missense mutation c.801C>G, p.Cys267Trp in the melanocortin 1 receptor gene and the black (GG or GC) or brown (CC) colorations.

Chapter 6

General References

Aberdam E, Bertolotto C, Sviderskaya EV, de Thillot V, Hemesath TJ, Fisher DE, Bennett DC, Ortonne JP, and Ballotti R. 1998. Involvement of microphthalmia in the inhibition of melanocyte lineage differentiation and of melanogenesis by agouti signal protein. *The Journal of Biological Chemistry*. 273:19560-5.

Adalsteinsson S, Sponenberg DP, Alexieva S, and Russel AJ. 1994. Inheritance of goat coat colors. *Journal of Heredity*. 85:267-72.

Akers RM. 2016. Lactation and the mammary gland. Iowa State Press, Ames, Iowa, USA. http://dx.doi.org/10.1002/9781119264880.

Alexander DH and Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 12:246.

Alexander DH, Novembre J, and Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*. 19:1655-64.

Allan GJ, Beattie J, and Flint DJ. 2004. The role of IGFBP-5 in mammary gland development and involution. *Domestic Animal Endocrinology*. 27:257-66.

Alnakip M, Quintela Baluja M, Böhme K, Fernández-No I, Caamaño-Antelo S, Calo-Mata P, and Barros-Velázquez J. 2014. The immunology of mammary gland of dairy ruminants between healthy and inflammatory conditions. *Journal of Veterinary Science.* 2014:1-32.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, and Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*. 407:513-6.

Amills M, Capote J, and Tosser-Klopp G. 2017. Goat domestication and breeding: A jigsaw of historical, biological and molecular data with missing pieces. *Animal Genetics*. 48:631-44.

Amills M. 2014. The application of genomic technologies to investigate the inheritance of economically important traits in goats. *Advances in Biology*. 2014:1-13.

Amills M, Jordana J, Zidi A, and Serradilla JM. 2002. Genetic factors that regulate milk protein and lipid composition in goats. In: Chaiyabutr N. (eds) Milk production-advanced genetic traits, cellular mechanism, animal management and health. IntechOpen: London, UK. http://dx.doi.org/10.5772/51716.

Analla M, Jiménez-Gamero I, Muñoz-Serrano A, Serradilla JM, and Falagán A. 1996. Estimation of genetic parameters for milk yield and fat and protein contents of milk from Murciano-Granadina goats. *Journal of Dairy Science*. 79:1895-8.

Anders S, Pyl PT, and Huber W. 2015. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31:166-9.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, and Zondervan KT. 2010. Data quality control in genetic case-control association studies. *Nature Protocols*. 5:1564-73.

Andersson L. 2020. Mutations in domestic animals disrupting or creating pigmentation patterns. *Frontiers in Ecology and Evolution.* 8:116.

Arnal M, Robert-Granié C, and Larroque H. 2018. Diversity of dairy goat lactation curves in France. *Journal of Dairy Science*. 101:11040-51.

Arora R, Sharma A, Sharma U, Girdhar Y, Kaur M, Kapoor P, Ahlawat S, and Vijh RK. 2019. Buffalo milk transcriptome: A comparative analysis of early, mid and late lactation. *Scientific Reports*. 9:5993.

Aslam M and Hurley WL. 1997. Proteolysis of milk proteins during involution of the bovine mammary gland. *Journal of Dairy Science*. 80:2004-10.

Atabai K, Fernandez R, Huang X, Ueki I, Kline A, Li Y, Sadatmansoori S, Smith-Steinhart C, Zhu W, Pytela R, et al. 2005. Mfge8 is critical for mammary gland remodeling during involution. *Molecular Biology of the Cell*. 16:5528-37.

Atabai K, Sheppard D, and Werb Z. 2007. Roles of the innate immune system in mammary gland remodeling during involution. *Journal of Mammary Gland Biology and Neoplasia*. 12:37-45.

Athie F, Bachman KC, Head HH, Hayen MJ, and Wilcox CJ. 1997. Milk plasmin during bovine mammary involution that has been accelerated by estrogen. *Journal of Dairy Science*. 80:1561-8.

Aulchenko YS, Ripke S, Isaacs A, and van Duijn CM. 2007. GenABEL: An R library for genome-wide association analysis. *Bioinformatics*. 23:1294-6.

Badaoui B, D'Andrea M, Pilla F, Capote J, Zidi A, Jordana J, Ferrando A, Delgado JV, Martínez A, Vidal O, et al. 2011. Polymorphism of the goat agouti signaling protein gene and its relationship with coat color in Italian and Spanish breeds. *Biochemical Genetics*. 49:523-32.

Ballester M, Sánchez A, and Folch JM. 2005. Polymorphisms in the goat β-lactoglobulin gene. *Journal of Dairy Research*. 72:379-84.

Bär L, Feger M, Fajol A, Klotz LO, Zeng S, Lang F, and Föller M. 2018. Insulin suppresses the production of fibroblast growth factor 23 (FGF23). *Proceedings of the National Academy of Sciences of the United States of America*. 115:5804-9.

Barash I. 2006. Stat5 in the mammary gland: Controlling normal development and cancer. *Journal of Cellular Physiology*. 209:305-13.

Barsh GS, Copenhaver GP, Gibson G, and Williams SM. 2012. Guidelines for genome-wide association studies. *PLoS Genetics*. 8:e1002812.

Becker D, Otto M, Ammann P, Keller I, Drögemüller C, and Leeb T. 2014. The brown coat colour of Coppernecked goats is associated with a non-synonymous variant at the *TYRP1* locus on chromosome 8. *Animal Genetics*. 46:50-4.

Begum F, Ghosh D, Tseng GC, and Feingold E. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research*. 40:3777-84.

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. 57:289-300.

Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, Palhière I, Crisà A, Catillo G, Steri R, et al. 2018. Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genetics Selection Evolution*. 50:57.

Bhaskaran M, Chen H, Chen Z, and Liu L. 2005. Hemoglobin is expressed in alveolar epithelial type II cells. *Biochemical and Biophysical Research Communications*. 333:1348-52.

Bhat SA, Ahmad SM, Ibeagha-Awemu EM, Bhat BA, Dar MA, Mumtaz PT, Shah RA, and Ganai NA. 2019. Comparative transcriptome analysis of mammary epithelial cells at different stages of lactation reveals wide differences in gene expression and pathways regulating milk synthesis between Jersey and Kashmiri cattle. *PLoS One*. 14:e0211773.

Bickhart D and Liu G. 2014. The challenges and importance of structural variation detection in livestock. *Frontiers in Genetics*. 5:37.

Biet J, Poole CA, Stelwagen K, Margerison JK, and Singh K. 2016. Primary cilia distribution and orientation during involution of the bovine mammary gland. *Journal of Dairy Science*. 99:3966-78.

Bionaz M, Chen S, Khan MJ, and Loor JJ. 2013. Functional role of PPARs in ruminants: Potential targets for fine-tuning metabolism during growth and lactation. *PPAR research*. 2013:684159.

Bionaz M, Hurley W, and Loor J. 2012a. Milk protein synthesis in the lactating mammary gland: Insights from transcriptomics analyses. In: Walter H. (eds) Milk Protein. IntechOpen: London, UK. http://dx.doi.org/10.5772/46054.

Bionaz M and Loor JJ. 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*. 9:366.

Bionaz M and Loor JJ. 2011. Gene networks driving bovine mammary protein synthesis during the lactation cycle. *Bioinformatics and Biology Insights*. 5:83-98.

Bionaz M, Periasamy K, Rodriguez-Zas SL, Everts RE, Lewin HA, Hurley WL, and Loor JJ. 2012b. Old and new stories: Revelations from functional analysis of the bovine mammary transcriptome during the lactation cycle. *PLoS One*. 7:e33268.

Boldman K, Kriese LA, Van Vleck L, Tassell CP, and Kachman SD. 1993. A manual for use of MTDFREML: A set of programs to obtain estimates of variances and covariances (Draft). United States Department of Agriculture. Agricultural Research Service. Clay Center, Nebraska.

Bolger AM, Lohse M, and Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 30:2114-20.

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, Sahana G, Govignon-Gion A, Boitard S, Dolezal M, et al. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature Genetics*. 50:362-7.

Bray NL, Pimentel H, Melsted P, and Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*. 34:525-7.

Brisken C and O'Malley B. 2010. Hormone action in the mammary gland. *Cold Spring Harbor Perspectives in Biology*. 2:a003178.

Bu D, Bionaz M, Wang M, Nan X, Ma L, and Wang J. 2017. Transcriptome difference and potential crosstalk between liver and mammary tissue in mid-lactation primiparous dairy cows. *PLoS One*. 12:e0173082.

Burrows M and Wheeler DJ. 1994. A block-sorting lossless data compression algorithm. Digital, Systems Research Center., Palo Alto, CA.

Bush WS and Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*. 8:e1002822.

Cánovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, and Medrano JF. 2010. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. *Mammalian Genome*. 21:592-8.

Caravaca F, Amills M, Jordana J, Angiolillo A, Agüera P, Aranda C, Menéndez-Buxadera A, Sánchez A, Carrizosa J, Urrutia B, et al. 2008. Effect of α_{s1}-casein (*CSN1S1*) genotype on milk CSN1S1 content in Malagueña and Murciano-Granadina goats. *Journal of Dairy Research*. 75:481-4.

Caravaca F, Ares JL, Carrizosa J, Urrutia B, Baena F, Jordana J, Badaoui B, Sànchez A, Angiolillo A, Amills M, et al. 2011. Effects of α_{s1}-casein (*CSN1S1*) and κ-casein (*CSN3*) genotypes on milk coagulation properties in Murciano-Granadina goats. *Journal of Dairy Research*. 78:32-7.

Caravaca F, Carrizosa J, Urrutia B, Baena F, Jordana J, Amills M, Badaoui B, Sánchez A, Angiolillo A, and Serradilla JM. 2009. Short communication: Effect of α_{S1}-casein (*CSN1S1*) and κ-casein (*CSN3*) genotypes on milk composition in Murciano-Granadina goats. *Journal of Dairy Science*. 92:2960-4.

Carcangiu V, Mura MC, Daga C, Luridiana S, Bodano S, Sanna GA, Diaz ML, and Cosso G. 2013. Association between *SREBP-1* gene expression in mammary gland and milk fat yield in Sarda breed sheep. *Meta Gene*. 1:43-49.

Carillier-Jacquin C, Larroque H, and Robert-Granié C. 2016. Including *α_{s1} casein* gene information in genomic evaluations of French dairy goats. *Genetics Selection Evolution*. 48:54.

Carillier C, Larroque H, and Robert-Granié C. 2014. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. *Genetics Selection Evolution*. 46:67.

Caro T and Mallarino R. 2020. Coloration in mammals. *Trends in Ecology & Evolution*. 35:357-66.

Cartón-García F, Overeem AW, Nieto R, Bazzocco S, Dopeso H, Macaya I, Bilic J, Landolfi S, Hernandez-Losa J, Schwartz S, JR, et al. 2015. *Myo5b* knockout mice as a model of microvillus inclusion disease. *Scientific reports*. 5:12312.

Chang SC, Chen HF, Chou MH, Wang HC, Su HY, and Wong ML. 2010. Haemoglobin in normal and neoplastic canine mammary glands. *Veterinary and Comparative Oncology*. 8:302-9.

Chen K, Hou J, Song Y, Zhang X, Liu Y, Zhang G, Wen K, Ma H, Li G, Cao B, et al. 2018. Chi-miR-3031 regulates beta-casein via the PI3K/AKT-mTOR signaling pathway in goat mammary epithelial cells (GMECs). *BMC Veterinary Research*. 14:369.

Cichorek M, Wachulska M, Stasiewicz A, and Tymińska A. 2013. Skin melanocytes: Biology and development. *Advances in Dermatology and Allergology/Postępy Dermatologii i Alergologii*. 30:30-41.

Colli L, Milanesi M, Talenti A, Bertolini F, Chen M, Crisà A, Daly KG, Del Corvo M, Guldbrandtsen B, Lenstra JA, et al. 2018. Genome-wide SNP profiling of worldwide goat populations reveals strong partitioning of diversity and highlights post-domestication migration routes. *Genetics Selection Evolution*. 50:58.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*. 17:13.

Cosenza G, Gallo D, Illario R, Di Gregorio P, Senese C, Ferrara L, and Ramunno L. 2003. A Mval PCR-RFLP detecting a silent allele at the goat alpha-lactalbumin locus. *Journal of Dairy Research*. 70:355-7.

Crepaldi P and Nicoloso L. 2007. SNPs in coat colour genes in goats. *Italian Journal of Animal Science*. 6:91-3.

Crisà A, Ferrè F, Chillemi G, and Moioli B. 2016. RNA-Sequencing for profiling goat milk transcriptome in colostrum and mature milk. *BMC Veterinary Research*. 12:264.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*. 46:858-65.

Daly KG, Maisano Delser P, Mullin VE, Scheu A, Mattiangeli V, Teasdale MD, Hare AJ, Burger J, Verdugo MP, Collins MJ, et al. 2018. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science*. 361:85-8.

Delgado JV, Landi V, Barba CJ, Fernández J, Gómez MM, Camacho ME, Martínez MA, Navas FJ, and León JM. 2017. Murciano-Granadina goat: A Spanish local breed ready for the challenges of the twenty-first century. In: Simões J, Gutiérrez C. (eds) Sustainable goat production in adverse environments: Volume II. Springer, Cham. https://doi.org/10.1007/978-3-319-71294-9_15.

Deng T, Liang A, Liang S, Ma X, Lu X, Duan A, Pang C, Hua G, Liu S, Campanile G, et al. 2019. Integrative analysis of transcriptome and GWAS data to identify the hub genes associated with milk yield trait in buffalo. *Frontiers in Genetics*. 10:36.

Devlin B and Roeder K. 1999. Genomic control for association studies. *Biometrics*. 55:997-1004.

Dietrich J, Menzi F, Ammann P, Drogemuller C, and Leeb T. 2015. A breeding experiment confirms the dominant mode of inheritance of the brown coat colour associated with the ^{496}Asp *TYRP1* allele in goats. *Animal Genetics*. 46:587-8.

Dijkstra J, Lopez S, Bannink A, Dhanoa MS, Kebreab E, Odongo NE, Fathi Nasri MH, Behera UK, Hernandez-Ferrer D, and France J. 2010. Evaluation of a mechanistic lactation model using cow, goat and sheep data. *The Journal of Agricultural Science*. 148:249-62.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. 2012. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 29:15-21.

Dobin A and Gingeras TR. 2015. Mapping RNA-seq reads with STAR. *Current Protocols in Bioinformatics*. 51:11.14.11-11.14.19.

Dong XY and Tang SQ. 2010. Insulin-induced gene: A new regulator in lipid metabolism. *Peptides*. 31:2145-50.

Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W, Feng S, Huang G, Guan R, Shen W, et al. 2015. Reference genome of wild goat (*capra aegagrus*) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genomics*. 16:431.

Douglas NC and Papaioannou VE. 2013. The T-box transcription factors TBX2 and TBX3 in mammary gland development and breast cancer. *Journal of Mammary Gland Biology and Neoplasia*. 18:143-7.

Dowbenko D, Kikuta A, Fennie C, Gillett N, and Lasky L. 1993. Glycosylation-dependent cell adhesion molecule 1 (GlyCAM 1) mucin is expressed by lactating mammary gland epithelial cells and is present in milk. *The Journal of clinical investigation*. 92:952-60.

Dunning KR, Anastasi MR, Zhang VJ, Russell DL, and Robker RL. 2014. Regulation of fatty acid oxidation in mouse cumulus-oocyte complexes during maturation and modulation by PPAR agonists. *PLoS One*. 9:e87327.

Durkin K, Coppieters W, Drögemüller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L, et al. 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature*. 482:81-4.

Eblaghie MC, Song SJ, Kim JY, Akita K, Tickle C, and Jung HS. 2004. Interactions between FGF and Wnt signals and *Tbx3* gene expression in mammary gland initiation in mouse embryos. *Journal of Anatomy*. 205:1-13.

Emrich SJ, Barbazuk WB, Li L, and Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*. 17:69-73.

Evangelou E and Ioannidis JP. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*. 14: 379-9.

Faulconnier Y, Chilliard Y, Torbati MBM, and Leroux C. 2011. The transcriptomic profiles of adipose tissues are modified by feed deprivation in lactating goats. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*. 6:139-49.

Feng FJ, Li XL, Zhou RY, Zheng GR, Li LH, and Li DF. 2009. Characterization and SNP identification of part of the goat melanophilin gene. *Biochemical Genetics*. 47:198-206.

Ferragina P and Manzini G. 2000. Opportunistic data structures with applications. Proceedings 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, USA. https://doi.org/10.1109/SFCS.2000.892127.

Ferreira AM, Bislev SL, Bendixen E, and Almeida AM. 2013. The mammary gland in domestic ruminants: A systems biology perspective. *Journal of Proteomics*. 94:110-23.

Fonseca NA, Rung J, Brazma A, and Marioni JC. 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics*. 28:3169-77.

Fontanesi L, Beretti F, Riggio V, Dall'Olio S, González EG, Finocchiaro R, Davoli R, Russo V, and Portolano B. 2009a. Missense and nonsense mutations in melanocortin 1 receptor (*MC1R*) gene of different goat breeds: Association with red and black coat colour phenotypes but with unexpected evidences. *BMC Genetics*. 10:47.

Fontanesi L, Beretti F, Riggio V, Gomez Gonzalez E, Dall'Olio S, Davoli R, Russo V, and Portolano B. 2009b. Copy number variation and missense mutations of the agouti signaling protein (*ASIP*) gene in goat breeds with different coat colors. *Cytogenetic and Genome Research*. 126:333-47.

Frattini S, Nicoloso L, Coizet B, et al. 2014. Short communication: The unusual genetic trend of α_{S1}-casein in Alpine and Saanen breeds. *Journal of Dairy Science*. 97:7975-9.

Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, and Leek JT. 2015. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nature Biotechnology*. 33:243-6.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, et al. 2009. The challenges of sequencing by synthesis. *Nature Biotechnology*. 27:1013-23.

Gao YY, Zhang ZH, Zhuang Z, Lu Y, Wu LY, Ye Zn, Zhang XS, Chen CL, Li W, and Hang CH. 2018. Recombinant milk fat globule-EGF factor-8 reduces apoptosis via integrin β3/FAK/PI3K/AKT signaling pathway in rats after traumatic brain injury. *Cell Death & Disease*. 9:845.

Georges M, Charlier C, and Hayes B. 2018. Harnessing genomic information for livestock improvement. *Nature Reviews Genetics*. 20:135-56.

Gilad Y, Rifkin SA, and Pritchard JK. 2008. Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends in Genetics*. 24:408-15.

Gipson TA. 2019. Recent advances in breeding and genetics for dairy goats. *Asian-Australasian Journal of Animal Sciences*. 32:1275-83.

Gipson TA and Grossman M. 1990. Lactation curves in dairy goats: A review. *Small Ruminant Research*. 3:383-96.

Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, and Caldas C. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*. 16:991-1006.

Goodwin S, McPherson JD, and McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. 17:333-51.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012.

Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 491:393-98.

Gross JJ and Bruckmaier RM. 2019. Invited review: Metabolic challenges and adaptation during different functional stages of the mammary gland in dairy cows: Perspectives for sustainable milk production. *Journal of Dairy Science*. 102:2828-43.

Grossen C, Keller L, Biebach I, International Goat Genome Consortium, and Croll D. 2014. Introgression from domestic goat generated variation at the major histocompatibility complex of Alpine Ibex. *PLoS Genetics*. 10:e1004438.

Haase B, Brooks SA, Schlumbaum A, Azor PJ, Bailey E, Alaeddine F, Mevissen M, Burger D, Poncet PA, Rieder S, et al. 2007. Allelic heterogeneity at the equine *KIT* locus in dominant white (*W*) horses. *PLoS Genetics*. 3:e195.

Hammerle-Fickinger A, Riedmaier I, Becker C, Meyer HHD, Pfaffl MW, and Ulbrich SE. 2009. Validation of extraction methods for total RNA and miRNA from bovine blood prior to quantitative gene expression analyses. *Biotechnology Letters*. 32:35-44.

Han Y, Gao S, Muegge K, Zhang W, and Zhou B. 2015. Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*. 9:29-46.

Hao Z, Zhou H, Hickford JGH, Gong H, Wang J, Hu J, Liu X, Li S, Zhao M, and Luo Y. 2019. Transcriptome profile analysis of mammary gland tissue from two breeds of lactating sheep. *Genes*. 10:781.

Hassiotou F and Geddes D. 2013. Anatomy of the human mammary gland: Current status of knowledge. *Clinical Anatomy*. 26:29-48.

Hastings PJ, Lupski JR, Rosenberg SM, and Ira G. 2009. Mechanisms of change in gene copy number. *Nature Reviews Genetics*. 10:551-64.

Hayes B, Bowman PJ, Chamberlain AJ, and Goddard ME. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*. 92:433-43.

Hayes B, Hagesæther N, Ådnøy T, Pellerud G, Berg PR, and Lien S. 2006. Effects on production traits of haplotypes among casein genes in Norwegian goats and evidence for a site of preferential recombination. *Genetics*. 174:455-64.

Hayes B, Lewin HA, and Goddard ME. 2013. The future of livestock breeding: Genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*. 29:206-14.

Haynes W. 2013. Bonferroni correction. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H. (eds) Encyclopedia of systems biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, and Ordoukhanian P. 2014. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*. 56:61-4.

Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, Herren U, Posantzis D, Bulut Z, Ammann P, et al. 2019. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLoS Genetics*. 15:e1008536.

Hernandez LL. 2017. TRIENNIAL LACTATION SYMPOSIUM/BOLFA: Serotonin and the regulation of calcium transport in dairy cows. *Journal of Animal Science*. 95:5711-9.

Hernandez LL, Stiening CM, Wheelock JB, Baumgard LH, Parkhurst AM, and Collier RJ. 2008. Evaluation of serotonin as a feedback inhibitor of lactation in the bovine. *Journal of Dairy Science*. 91:1834-44.

Ho SS, Urban AE, and Mills RE. 2019. Structural variation in the sequencing era. *Nature Reviews Genetics*. 21:171-89

Houdebine LM, Djiane J, Dusanter-Fourt I, Martel P, Kelly PA, Devinoy E, and Servely JL. 1985. Hormonal action controlling mammary activity. *Journal of Dairy Science*. 68:489-500.

Hrdlickova R, Toloue M, and Tian B. 2017. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*. 8:10.1002/wrna.1364.

Hu ZL, Park CA, Wu XL, and Reecy JM. 2013. Animal QTLdb: An improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Research*. 41:D871-9.

Hughes K and Watson CJ. 2018. The multifaceted role of STAT3 in mammary gland involution and breast cancer. *International Journal of Molecular Sciences.* 19:1695.

Ibeagha-Awemu EM, Li R, Ammah AA, Dudemaine PL, Bissonnette N, Benchaar C, and Zhao X. 2016. Transcriptome adaptation of the bovine mammary gland to diets rich in unsaturated fatty acids shows greater impact of linseed oil over safflower oil on gene expression and metabolic pathways. *BMC Genomics*. 17:104.

Iguchi T. 1996. Involvement of the TNF-α system and the Fas system in the induction of apoptosis of mouse mammary glands after weaning. *Apoptosis*. 1:201-8.

Jaggi R, Marti A, Guo K, Feng Z, and Friis RR. 1996. Regulation of a physiological apoptosis: Mouse mammary involution. *Journal of Dairy Science*. 79:1074-84.

Jain NC. 1979. Common mammary pathogens and factors in infection and mastitis. *Journal of Dairy Science*. 62:128-34.

Janes SM and Watt FM. 2004. Switch from alphavbeta5 to alphavbeta6 integrin expression protects squamous cell carcinomas from anoikis. *Journal of Cell Biology*. 166:419-31.

Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, Tiplady K, McNaughton L, Johnson TJJ, Davis SR, et al. 2019. Genome-wide

association analysis reveals QTL and candidate mutations involved in white spotting in cattle. *Genetics Selection Evolution*. 51:62.

Johnston JA, Bacon CM, Finbloom DS, Rees RC, Kaplan D, Shibuya K, Ortaldo JR, Gupta S, Chen YQ, Giri JD, et al. 1995. Tyrosine phosphorylation and activation of STAT5, STAT3, and Janus kinases by interleukins 2 and 15. *Proceedings of the National Academy of Sciences of the United States of America*. 92:8705-9.

Kadegowda AKG, Bionaz M, Piperova LS, Erdman RA, and Loor JJ. 2009. Peroxisome proliferator-activated receptor-gamma activation and long-chain fatty acids alter lipogenic gene networks in bovine mammary epithelial cells to various extents. *Journal of Dairy Science*. 92:4276-89.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, and Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*. 42:348-54.

Kang X, Li M, Liu M, Liu S, Pan MG, Wiggans GR, Rosen BD, and Liu GE. 2020. Copy number variation analysis reveals variants associated with milk production traits in dairy goats. *Genomics*. 112:4934-7.

Katsafadou AI, Politis AP, Mavrogianni VS, Barbagianni MS, Vasileiou NGC, Fthenakis GC, and Fragkou IA. 2019. Mammary defences and immunity against mastitis in Sheep. *Animals*. 9:726.

Kijas JMH, Wales R, Törnsten A, Chardon P, Moller M, and Andersson L. 1998. Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics*. 150:1177-85.

Kim D, Landmead B, and Salzberg SL. 2015. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*. 12:357-60.

Kim D, Paggi JM, Park C, Bennett C, and Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*. 37:907-15.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 14:R36.

Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, et al. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research*. 16:1575-84.

Kondrashov FA, Rogozin IB, Wolf YI, and Koonin EV. 2002. Selection in the evolution of gene duplications. *Genome biology*. 3:RESEARCH0008.

Kuhn NJ, Carrick DT, and Wilde CJ. 1980. Lactose Synthesis: The possibilities of regulation. *Journal of Dairy Science*. 63:328-36.

Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, Maloof JN, and Sinha NR. 2012. A high-throughput method for Illumina RNA-Seq library preparation. *Frontiers in Plant Science*. 3:202.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*. 37:4181-93.

Lai YT, Yeung CKL, Omland KE, Pang EL, Hao Y, Liao BY, Cao HF, Zhang BW, Yeh CF, Hung CM, et al. 2019. Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the National Academy of Sciences of the United States of America*. 116:2152-7.

Lan Y, Pan C, Chen H, Zhang C, Zhang A, Zhang L, Li J, and Lei C. 2007. An MspI PCR-RFLP detecting a single nucleotide polymorphism at alpha-lactalbumin gene in goat. *Czech Journal of Animal Science*. 52:138-42.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*. 11:Unit 11.7.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357-9.

Langmead B, Trapnell C, Pop M, and Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 10:R25.

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Vigueira CC, Denham T, Dobney K, et al. 2014. Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences of the United States of America*. 111:6139-46.

Lawson DJ, van Dorp L, and Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*. 9:3258.

Le Provost F, Cassy S, Hayes H, and Martin P. 2003. Structure and expression of goat *GLYCAM1* gene: Lactogenic-dependent expression in ruminant mammary gland and interspecies conservation of the proximal promoter. *Gene*. 313:83-9.

Lee BY, Lee KN, Lee T, Park JH, Kim SM, Lee HS, Chung DS, Shim HS, Lee HK, and Kim H. 2015. Bovine genome-wide association study for genetic elements to resist the infection of foot-and-mouth disease in the field. *Asian-Australasian Journal of Animal Sciences*. 28:166-70.

Lérias JR, Hernández-Castellano LE, Suárez-Trujillo A, Castro N, Pourlis A, and Almeida AM. 2014. The mammary gland in small ruminants: Major morphological and functional events underlying milk production - a review. *Journal of Dairy Research*. 81:304-18.

León JM, Macciotta NPP, Gama LT, Barba C, and Delgado JV. 2012. Characterization of the lactation curve in Murciano-Granadina dairy goats. *Small Ruminant Research*. 107:76-84.

Leushkin EV, Bazykin GA, and Kondrashov AS. 2013. Strong mutational bias toward deletions in the *Drosophila Melanogaster* genome is compensated by selection. *Genome Biology and Evolution*. 5:514-24.

Li C, Wang M, Zhang T, He Q, Shi H, Luo J, and Loor JJ. 2019. Insulin-induced gene 1 and 2 isoforms synergistically regulate triacylglycerol accumulation, lipid droplet formation, and lipogenic gene expression in goat mammary epithelial cells. *Journal of Dairy Science*. 102:1736-46.

Li CL, Martinez V, He B, Lombet A, and Perbal B. 2002. A role for CCN3 (NOV) in calcium signalling. *Molecular Pathology*. 55:250-61.

Li J, Bed'hom B, Marthey S, Valade M, Dureux A, Moroldo M, Péchoux C, Coville JL, Gourichon D, Vieaud A, et al. 2019. A missense mutation in *TYRP1* causes the chocolate plumage color in chicken and alters melanosome structure. *Pigment Cell & Melanoma Research*. 32:381-90.

Li Z, Lan X, Guo W, Sun J, Huang Y, Wang J, Huang T, Lei C, Fang X, and Chen H. 2012. Comparative transcriptome profiling of dairy goat MicroRNAs from dry period and peak lactation mammary gland tissues. *PLoS One*. 7:e52388.

Liang D, Xu W, Zhang Q, and Tao BB. 2017. Study on the effect of Integrin αVβ6 on proliferation and apoptosis of cervical cancer cells. *European review for medical and pharmacological sciences.* 21:2811-5.

Liao Y, Smyth GK, and Shi W. 2014. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 30:923-30.

Lin S, Luo W, Jiang M, Luo W, Abdalla BA, Nie Q, Zhang L, and Zhang X. 2017. Chicken *CCDC152* shares an NFYB-regulated bidirectional promoter with a growth hormone receptor antisense transcript and inhibits cells proliferation and migration. *Oncotarget*. 8:84039-53.

Lin Y, Sun X, Hou X, Qu B, Gao X, and Li Q. 2016. Effects of glucose on lactose synthesis in mammary epithelial cells from dairy cow. *BMC Veterinary Research*. 12:81.

Linden SK, Sutton P, Karlsson NG, Korolik V, and McGuckin MA. 2008. Mucins in the mucosal barrier to infection. *Mucosal Immunology*. 1:183-97.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, and Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. *Nature Methods*. 8:833-5.

Liu B, Yu Z, Chen C, Kling DE, and Newburg DS. 2012. Human milk mucin 1 and mucin 4 inhibit Salmonella enterica serovar Typhimurium invasion of human intestinal epithelial cells in vitro. *Journal of Nutrition*. 142:1504-9.

Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME, et al. 2010. Analysis of copy number variations among diverse cattle breeds. *Genome Research*. 20:693-703.

Liu M, Woodward-Greene J, Kang X, Pan MG, Rosen B, Van Tassell CP, Chen H, and Liu GE. 2019. Genome-wide CNV analysis revealed variants associated with growth traits in African indigenous goats. *Genomics*. 112:1477-80.

Liu M, Zhou Y, Rosen BD, Van Tassell CP, Stella A, Tosser-Klopp G, Rupp R, Palhière I, Colli L, Sayre B, et al. 2018. Diversity of copy number variation in the worldwide goat population. *Heredity*. 122:636-46.

Liu X, Robinson GW, and Hennighausen L. 1996. Activation of Stat5a and Stat5b by tyrosine phosphorylation is tightly linked to mammary gland differentiation. *Molecular Endocrinology*. 10:1496-506.

Love MI, Huber W, and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 15:550.

Lu J, Argov-Argaman N, Anggrek J, Boeren S, van Hooijdonk T, Vervoort J, and Hettinga KA. 2016a. The protein and lipid composition of the membrane of milk fat globules depends on their size. *Journal of Dairy Science*. 99:4726-38.

Lu J, Wang X, Zhang W, Liu L, Pang X, Zhang S, and Lv J. 2016b. Comparative proteomics of milk fat globule membrane in different species reveals variations in lactation and nutrition. *Food Chemistry*. 196:665-72.

Mahé MF, Manfredi E, Ricordeau G, Piacère A, and Grosclaude F. 1994. Effets du polymorphisme de la caséine α_{s1} caprine sur les performances laitières: analyse intradescendance de boucs de race Alpine. *Genetics Selection Evolution*. 26:151-7.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, and Sedlazeck FJ. 2019. Structural variant calling: The long and the short of it. *Genome Biology*. 20:246.

Manfredi E, Ricordeau G, Barbieri ME, Amigues Y, and Bibé B. 1995. Génotype caséine α_{s1} et sélection des boucs sur descendance dans les races Alpine et Saanen. *Genetics Selection Evolution*. 27:451-8.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature*. 461:747-53.

Mardones L and Villagrán M. 2020. Lactose Synthesis. In: Gutiérrez-Méndez N. (eds) Lactose. IntechOpen: London, UK. http://dx.doi.org/10.5772/intechopen.91399.

Marletta D, Criscione A, Bordonaro S, Guastella AM, and D'Urso G. 2007. Casein polymorphism in goat's milk. *Lait*. 87:491-504.

Marti A, Lazar H, Ritter P, and Jaggi R. 1999. Transcription factor activities and gene expression during mouse mammary gland involution. *Journal of Mammary Gland Biology and Neoplasia*. 4:145-52.

Martin A, David V, and Quarles LD. 2012. Regulation and function of the FGF23/klotho endocrine pathways. *Physiological Reviews*. 92:131-55.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17:3.

Martin P, Palhiere I, Maroteau C, Bardou P, Canale-Tabet K, Sarry J, Woloszyn F, Bertrand-Michel J, Racke I, Besir H, et al. 2017. A genome scan for milk production traits in dairy goats reveals two new mutations in *Dgat1* reducing milk fat content. *Scientific Reports*. 7:1872.

Martin P, Palhière I, Maroteau C, Clément V, David I, Klopp GT, and Rupp R. 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. *Journal of Dairy Science*. 101:5214-26.

Martin P, Palhière I, Tosser-Klopp G, and Rupp R. 2016a. Heritability and genome-wide association mapping for supernumerary teats in French Alpine and Saanen dairy goats. *Journal of Dairy Science*. 99:8891-900.

Martin P, Szymanowska M, Zwierzchowski L, and Leroux C. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reproduction Nutrition Development*. 42:433-59.

Martin PM, Palhière I, Ricard A, Tosser-Klopp G, and Rupp R. 2016b. Genome wide association study identifies new loci associated with undesired coat color phenotypes in Saanen goats. *PLoS One*. 11:e0152426.

Martinez A, Vega-Pla JL, León JM, Camacho M, Delgado JV, and Ribeiro M. 2010. Is the Murciano-Granadina a single goat breed? A molecular genetics approach. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia*. 62:1191-8.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, et al. 2009. Development and characterization of a high density SNP genotyping assay for Cattle. *PLoS One*. 4:e5350.

Menzi F, Keller I, Reber I, Beck J, Brenig B, Schutz E, Leeb T, and Drogemuller C. 2016. Genomic amplification of the caprine *EDNRA* locus might lead to a dose dependent loss of pigmentation. *Scientific Reports*. 6:28438.

Miller BA and Lu CD. 2019. Current status of global dairy goat production: An overview. *Asian-Australasian Journal of Animal Sciences*. 32:1219-32.

Miranda JC, León JM, Pieramati C, Gómez MM, Valdés J, and Barba C. 2019. Estimation of genetic parameters for peak yield, yield and persistency

traits in Murciano-Granadina goats using multi-traits models. *Animals*. 9:411.

Morerio C, Acquila M, Rapella A, Tassano E, Rosanda C, and Panarello C. 2006. Inversion (11)(p15q22) with NUP98-DDX10 fusion gene in pediatric acute myeloid leukemia. *Cancer Genetics and Cytogenetics*. 171:122-5.

Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, Tenesa A, and Jackson IJ. 2018. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nature Communications*. 9:5271.

Morrison B and Cutler ML. 2010. The contribution of adhesion signaling to lactogenesis. *Journal of Cell Communication and Signaling*. 4:131-9.

Mort RL, Jackson IJ, and Patton EE. 2015. The melanocyte lineage in development and disease. *Development*. 142:620-32.

Mucha S, Mrode R, Coffey M, Kizilaslan M, Desire S, and Conington J. 2017. Genome-wide association study of conformation and milk yield in mixed-breed dairy goats. *Journal of Dairy Science*. 101:2213-25.

Murphy G. 2011. Tissue inhibitors of metalloproteinases. *Genome biology*. 12:233.

Neville MC, Webb P, Ramanathan P, Mannino MP, Pecorini C, Monks J, Anderson SM, and MacLean P. 2013. The insulin receptor plays an important role in secretory differentiation in the mammary gland. *American Journal of Physiology-Endocrinology and Metabolism*. 305:E1103-14.

Newton DA, Rao KMK, Dluhy RA, and Baatz JE. 2006. Hemoglobin is expressed by alveolar epithelial cells. *Journal of Biological Chemistry*. 281:5668-76.

Nicklin P, Bergman P, Zhang B, Triantafellow E, Wang H, Nyfeler B, Yang H, Hild M, Kung C, Wilson C, et al. 2009. Bidirectional transport of amino acids regulates mTOR and autophagy. *Cell*. 136:521-34.

Nicolazzi EL, Biffani S, Biscarini F, Orozco ter Wengel P, Caprera A, Nazzicari N, and Stella A. 2015. Software solutions for the livestock genomics SNP array revolution. *Animal Genetics*. 46:343-53.

Nielsen MO, Madsen TG, and Hedeboe AM. 2001. Regulation of mammary glucose uptake in goats: Role of mammary gland supply, insulin, IGF-1 and synthetic capacity. *Journal of Dairy Research*. 68:337-49.

Norris BJ and Whan VA. 2008. A gene duplication affecting expression of the ovine *ASIP* gene is responsible for white and black sheep. *Genome Research*. 18: 1282-93.

Orford M, Hadjipavlou G, Tzamaloukas O, Chatziplis D, Koumas A, Mavrogenis A, Papachristoforou C, and Miltiadou D. 2012. A single nucleotide polymorphism in the acetyl-coenzyme A acyltransferase 2 (*ACAA2*) gene is associated with milk yield in Chios sheep. *Journal of Dairy Science*. 95:3419-27.

Orteu A and Jiggins CD. 2020. The genomics of coloration provides insights into adaptive evolution. *Nature Reviews Genetics*. 21:461-75

Oshlack A, Robinson MD, and Young MD. 2010. From RNA-seq reads to differential expression results. *Genome Biology*. 11:220.

Osorio JS, Lohakare J, and Bionaz M. 2016. Biosynthesis of milk fat, protein, and lactose: Roles of transcriptional and posttranscriptional regulation. *Physiological Genomics*. 48:231-56.

Ott J, Kamatani Y, and Lathrop M. 2011. Family-based designs for genome-wide association studies. *Nature Reviews Genetics*. 12:465-74.

Ozsolak F and Milos PM. 2011. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdisciplinary Reviews: RNA*. 2:565-70.

Paape MJ, Poutrel B, Contreras A, Marco JC, and Capuco AV. 2001. Milk somatic cells and lactation in small ruminants. *Journal of Dairy Science*. 84:E237-44.

Pastva AM, Wright JR, and Williams KL. 2007. Immunomodulatory roles of surfactant proteins A and D: Implications in lung disease. *Proceedings of the American Thoracic Society*. 4:252-7.

Paten AM, Duncan EJ, Pain SJ, Peterson SW, Kenyon PR, Blair HT, and Dearden PK. 2015. Functional development of the adult ovine mammary gland-insights from gene expression profiling. *BMC Genomics*. 16:748.

Patton S, Gendler SJ, and Spicer AP. 1995. The epithelial mucin, MUC1, of milk, mammary gland and other tissues. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes*. 1241:407-23.

Perbal B. 2004. CCN proteins: Multifunctional signalling regulators. *The Lancet*. 363:62-4.

Pertea M, Kim D, Pertea GM, Leek JT, and Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*. 11:1650-67.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, and Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*. 33:290-5.

Peterson JA, Hamosh M, Scallan CD, Ceriani RL, Henderson TR, Mehta NR, Armand M, and Hamosh P. 1998. Milk fat globule glycoproteins in human milk and in gastric aspirates of mother's milk-fed preterm infants. *Pediatric Research*. 44:499-506.

Pfeffer K. 2003. Biological functions of tumor necrosis factor cytokines and their receptors. *Cytokine & Growth Factor Reviews*. 14:185-91.

Pickrell JK and Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*. 8:e1002967.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, MacDonald JR, Mills R, et al. 2011. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*. 29:512-20.

Pizarro Inostroza MG, Landi V, Navas González FJ, León Jurado JM, Martínez Martínez MdA, Fernández Álvarez J, and Delgado Bermejo JV. 2019. Non-parametric association analysis of additive and dominance effects of casein complex SNPs on milk content and quality in Murciano-Granadina goats. *Journal of Animal Breeding and Genetics*. 00:1-16.

Platonova N, Scotti M, Babich P, Bertoli G, Mento E, Meneghini V, Egeo A, Zucchi I, and Merlo GR. 2007. *TBX3*, the gene mutated in ulnar-mammary syndrome, promotes growth of mammary epithelial cells via repression of p19ARF, independently of p53. *Cell and Tissue Research*. 328:301-16.

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, and Lareu M. 2013. An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Frontiers in Genetics*. 4:98.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 38:904-9.

Price AL, Zaitlen NA, Reich D, and Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*. 11:459-63.

Pulina G, Milán MJ, Lavín MP, Theodoridis A, Morin E, Capote J, Thomas DL, Francesconi AHD, and Caja G. 2018. Invited review: Current production trends, farm structures, and economics of the dairy sheep and goat sectors. *Journal of Dairy Science*. 101:6715-29.

Qiao X, Su R, Wang Y, Wang R, Yang T, Li X, Chen W, He S, Jiang Y, Xu Q, et al. 2017. Genome-wide target enrichment-aided chip design: A 66 K SNP chip for cashmere goat. *Scientific Reports*. 7:8621.

Rabot A, Sinowatz F, Berisha B, Meyer HHD, and Schams D. 2007. Expression and localization of extracellular matrix-degrading proteinases and their inhibitors in the bovine mammary gland during development, function, and involution. *Journal of Dairy Science*. 90:740-8.

Ramos-Onsins SE, Burgos-Paz W, Manunza A, and Amills M. 2014. Mining the pig genome to investigate the domestication process. *Heredity*. 113:471-84

Reich D, Price AL, and Patterson N. 2008. Principal component analysis of genetic data. *Nature Genetics*. 40:491-2.

Rezaei R, Wu Z, Hou Y, Bazer FW, and Wu G. 2016. Amino acids and mammary gland development: Nutritional implications for milk production and neonatal growth. *Journal of Animal Science and Biotechnology*. 7:20.

Robinson GW and Hennighausen L. 1997. Inhibins and activins regulate mammary epithelial cell differentiation through mesenchymal-epithelial interactions. *Development*. 124:2701-8.

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 26:139-40.

Roldán DL, Rabasa AE, Saldaño S, Holgado F, Poli MA, and Cantet RJC. 2008. QTL detection for milk production traits in goats using a longitudinal model. *Journal of Animal Breeding and Genetics*. 125:187-93.

Rubin CJ, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg O, Jern P, Jorgensen CB, et al. 2012. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*. 109:19529-36.

Rupp R, Clément V, Piacere A, Robert-Granié C, and Manfredi E. 2011. Genetic parameters for milk somatic cell score and relationship with production and udder type traits in dairy Alpine and Saanen primiparous goats. *Journal of Dairy Science*. 94:3629-34.

Rupp R, Mucha S, Larroque H, McEwan J, and Conington J. 2016. Genomic application in sheep and goat breeding. *Animal Frontiers*. 6:39-44.

Safayi S, Theil PK, Hou L, Engbæk M, Nørgaard JV, Sejrsen K, and Nielsen MO. 2010. Continuous lactation effects on mammary remodeling during late gestation and lactation in dairy goats. *Journal of Dairy Science*. 93:203-17.

Saldana-Caboverde A and Kos L. 2010. Roles of endothelin signaling in melanocyte development and melanoma. *Pigment Cell & Melanoma Research*. 23:160-70.

Schaid DJ, Chen W, and Larson NB. 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*. 19:491-504.

Schroten H. 2001. Chemistry of milk mucins and their anti-microbial action. In: Woodward B and Draper HH. (eds) Advances in nutritional research: Immunological properties of milk. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4615-0661-4_11.

Seaton G, Haley CS, Knott SA, Kearsey M, Visscher PM. 2002. QTL Express: Mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics*. 18:339-40.

Seder CW, Hartojo W, Lin L, Silvers AL, Wang Z, Thomas DG, Giordano TJ, Chen G, Chang AC, Orringer MB, et al. 2009. *INHBA* overexpression promotes cell proliferation and may be epigenetically regulated in esophageal adenocarcinoma. *Journal of Thoracic Oncology*. 4:455-62.

Selvaggi M, Laudadio V, Dario C, and Tufarelli V. 2014. Major proteins in goat milk: An updated overview on genetic variability. *Molecular Biology Reports*. 41:1035-48.

Sepe L and Argüello A. 2019. Recent advances in dairy goat products. *Asian-Australasian Journal of Animal Sciences*. 32:1306-20.

Seyednasrollah F, Laiho A, and Elo LL. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*. 16:59-70.

Shendurse AM and Khedkar CD. 2016. Lactose. In: Caballero B, Finglas PM, and Toldrá F. (eds) Encyclopedia of food and health. Academic Press, Oxford.

Shendure J and Ji H. 2008. Next-generation DNA sequencing. *Nature Biotechnology*. 26:1135-45.

Shi HB, Luo J, Yao DW, Zhu JJ, Xu HF, Shi HP, and Loor JJ. 2013. Peroxisome proliferator-activated receptor-γ stimulates the synthesis of monounsaturated fatty acids in dairy goat mammary epithelial cells via the control of stearoyl-coenzyme A desaturase. *Journal of Dairy Science*. 96:7844-53.

Silver DL, Hou L, Somerville R, Young ME, Apte SS, and Pavan WJ. 2008. The secreted metalloprotease *ADAMTS20* is required for melanoblast survival. *PLoS Genetics*. 4:e1000003.

Singh V, Mani I, and Chaudhary DK. 2013. *ATP4A* gene regulatory network for fine-tuning of proton pump and ion channels. *Systems and Synthetic Biology*. 7:23-32.

Song OR, Queval CJ, Iantomasi R, Delorme V, Marion S, Veyron-Churlet R, Werkmeister E, Popoff M, Ricard I, Jouny S, et al. 2018. ArfGAP1 restricts mycobacterium tuberculosis entry by controlling the actin cytoskeleton. *EMBO reports*. 19:29-42.

Sonstegard T, Capuco AV, White J, Van Tassell CP, Connor EE, Cho J, Sultana R, Shade L, Wray JE, Wells KD, et al. 2002. Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. *Mammalian Genome*. 13:373-9.

Sordillo LM and Streicher KL. 2002. Mammary gland immunity and mastitis susceptibility. *Journal of Mammary Gland Biology and Neoplasia*. 7:135-46.

Sørensen ES, Rasmussen LK, Møller L, and Petersen TE. 1997. The localization and multimeric nature of component PP3 in bovine milk: Purification

and characterization of PP3 from caprine and ovine Milks. *Journal of Dairy Science*. 80:3176-81.

Sponenberg DP, Alexieva S, and Adalsteinsson S. 1998. Inheritance of color in Angora goats. *Genetics Selection Evolution*. 30:385-95.

Sponenberg DP and LaMarsh C. 1996. Dominant and recessive brown in goats. *Genetics Selection Evolution*. 28:117-120.

Stark R, Grzelak M, and Hadfield J. 2019. RNA sequencing: The teenage years. *Nature Reviews Genetics*. 20:631-56

Stein T, Salomonis N, and Gusterson BA. 2007. Mammary gland involution as a multi-step process. *Journal of Mammary Gland Biology and Neoplasia*. 12:25-35.

Suárez-Vega A, Gutiérrez-Gil B, Klopp C, Robert-Granie C, Tosser-Klopp G, and Arranz JJ. 2015. Characterization and comparative analysis of the milk transcriptome in two dairy sheep breeds using RNA sequencing. *Scientific Reports*. 5:18399.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science*. 349:aab3761.

Sved JA and Hill WG. 2018. One hundred years of linkage disequilibrium. *Genetics*. 209:629-36.

Svennersten-Sjaunja K and Olsson K. 2005. Endocrinology of milk production. *Domestic Animal Endocrinology*. 29:241-58.

Swirnoff AH, Apel ED, Svaren J, Sevetson BR, Zimonjic DB, Popescu NC, and Milbrandt J. 1998. Nab1, a corepressor of NGFI-A (Egr-1), contains an active transcriptional repression domain. *Molecular and cellular biology*. 18:512-24.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, and Meyre D. 2019. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*. 20:467-84

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*. 526:68-74.

Tian H, Luo J, Shi H, Chen X, Wu J, Liang Y, Li C, and Loor JJ. 2020. Role of peroxisome proliferator-activated receptor-α on the synthesis of monounsaturated fatty acids in goat mammary epithelial cells. *Journal of Animal Science*. 98:skaa062.

Tomić TT, Olausson J, Rehammar A, Deland L, Muth A, Ejeskär K, Nilsson S, Kristiansson E, Wassén ON, and Abel F. 2020. *MYO5B* mutations in pheochromocytoma/paraganglioma promote cancer progression. *PLoS Genetics*. 16:e1008803.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HCM, Jamli S, et al. 2014. Design and characterization of a 52K SNP chip for goats. *PLoS One*. 9:e86227.

Trapnell C, Pachter L, and Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. 25:1105-11.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*. 7:562-78.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 28:511-5.

Tucker HA. 1981. Physiological control of mammary growth, lactogenesis, and lactation. *Journal of Dairy Science*. 64:1403-21.

Tucker HA. 2000. Hormones, mammary growth, and lactation: A 41-year perspective. *Journal of Dairy Science*. 83:874-84.

Upadhyay M, da Silva VH, Megens HJ, Visker MHPW, Ajmone-Marsan P, Bâlteanu VA, Dunner S, Garcia JF, Ginja C, Kantanen J, et al. 2017.

Distribution and functionality of copy number variation across European cattle populations. *Frontiers in Genetics*. 8:108.

Uria JA and Werb Z. 1998. Matrix metalloproteinases and their expression in mammary gland. *Cell Research*. 8:187-94.

Vacca GM, Dettori ML, Piras G, Manca F, Paschino P, and Pazzola M. 2014. Goat casein genotypes are associated with milk production traits in the Sarda breed. *Animal Genetics*. 45:723-31.

VanHouten JN. 2005. Calcium sensing by the mammary gland. *Journal of Mammary Gland Biology and Neoplasia*. 10:129-39.

van den Berg I, Hayes BJ, Chamberlain AJ, and Goddard ME. 2019. Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. *BMC Genomics*. 20:291.

Varela LM and Ip MM. 1996. Tumor necrosis factor-alpha: A multifunctional regulator of mammary gland development. *Endocrinology*. 137:4915-24.

Vidal O, Drögemüller C, Obexer-Ruff G, Reber I, Jordana J, Martínez A, Bâlteanu VA, Delgado JV, Eghbalsaied S, Landi V, et al. 2017. Differential distribution of Y-chromosome haplotypes in Swiss and Southern European goat breeds. *Scientific Reports*. 7:16161.

Visscher PM, Hill WG, and Wray NR. 2008. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*. 9:255-66.

Wakao H, Gouilleux F, and Groner B. 1994. Mammary-gland factor (Mgf) is a novel member of the cytokine regulated transcription factor gene family and confers the prolactin response. *The EMBO Journal*. 13:2182-91.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 280:1077-82.

Wang H and Elledge SJ. 1999. DRC1, DNA replication and checkpoint protein 1, functions with DPB11 to control DNA replication and the S-phase

checkpoint in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*. 96:3824-9.

Wang H, Shi H, Luo J, Yi Y, Yao D, Zhang X, Ma G, and Loor JJ. 2017. MiR-145 regulates lipogenesis in goat mammary cells via targeting *INSIG1* and epigenetic regulation of lipid-related genes. *Journal of Cellular Physiology*. 232:1030-40.

Wang Z, Gerstein M, and Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 10:57-63.

Wellnitz O and Bruckmaier RM. 2012. The innate immune response of the bovine mammary gland to bacterial infection. *The Veterinary Journal*. 192:148-52.

Westra HJ and Franke L. 2014. From genome to function by studying eQTLs. *Biochimica et Biophysica Acta*. 1842:1896-902.

Wickramasinghe S, Rincon G, Islas-Trejo A, and Medrano JF. 2012. Transcriptional profiling of bovine milk using RNA sequencing. *BMC Genomics*. 13:45.

Wirschell M, Olbrich H, Werner C, Tritschler D, Bower R, Sale WS, Loges NT, Pennekamp P, Lindberg S, Stenram U, et al. 2013. The nexin-dynein regulatory complex subunit DRC1 is essential for motile cilia function in algae and humans. *Nature genetics*. 45:262-8.

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, et al. 2019. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 570:514-8

Wolf Horrell EM, Boulanger MC, and D'Orazio JA. 2016. Melanocortin 1 receptor: Structure, function, and regulation. *Frontiers in Genetics*. 7:95.

Wu Z, Deng Z, Huang M, Hou Y, Zhang H, Chen H, and Ren J. 2019. Whole-genome resequencing identifies *KIT* new alleles that affect coat color phenotypes in pigs. *Frontiers in Genetics*. 10:218.

Wu ZL, Li XL, Liu YQ, Gong YF, Liu ZZ, Wang XJ, Xin TR, and Ji Q. 2006. The red head and neck of Boer goats may be controlled by the recessive allele of the *MC1R* gene. *Animal Research*. 55:313-22.

Yahyaoui MH, Pena RN, Sánchez A, and Folch JM. 2000. Rapid communication: Polymorphism in the goat β-lactoglobulin proximal promoter region. *Journal of Animal Science*. 78:1100-1.

Yang B, Jiao B, Ge W, Zhang X, Wang S, Zhao H, and Wang X. 2018. Transcriptome sequencing to detect the potential role of long non-coding RNAs in bovine mammary gland during the dry and lactation period. *BMC Genomics*. 19:605.

Yang F, Agulian T, Sudati JE, Rhoads DB, and Levitsky LL. 2004. Developmental regulation of galactokinase in suckling mouse liver by the Egr-1 transcription factor. *Pediatric Research*. 55:822-9.

Yang J, Jiang J, Liu X, Wang H, Guo G, Zhang Q, and Jiang L. 2015. Differential expression of genes in milk of dairy cattle during lactation. *Animal Genetics*. 47:174-80.

Yang J, Lee SH, Goddard ME, and Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*. 88:76-82.

Yang L, Xu L, Zhou Y, Liu M, Wang L, Kijas JW, Zhang H, Li L, and Liu GE. 2018. Diversity of copy number variation in a worldwide population of sheep. *Genomics*. 110:143-8.

Yao J, Weng YG, Yan SJ, Hou MY, Shi Q, and Zuo GW. 2015. Effects of nephroblastoma overexpressed gene on proliferation, apoptosis and migration of human osteosarcoma 143B cells. *Tumor*. 35:119-28.

Yau C, Ragoussis J, and Winchester L. 2009. Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics*. 8:353-66.

Yorukoglu D, Yu YW, Peng J, and Berger B. 2016. Compressive mapping for next-generation sequencing. *Nature Biotechnology*. 34:374-6.

Zamora A, Guamis B, and Trujillo AJ. 2009. Protein composition of caprine milk fat globule membrane. *Small Ruminant Research*. 82:122-9.

Zarrei M, MacDonald JR, Merico D, and Scherer SW. 2015. A copy number variation map of the human genome. *Nature Reviews Genetics*. 16: 172-83.

Zhang B, Chang L, Lan X, Asif N, Guan F, Fu D, Li B, Yan C, Zhang H, Zhang X, et al. 2018. Genome-wide definition of selective sweeps reveals molecular evidence of trait-driven domestication among elite goat (*Capra species*) breeds for the production of dairy, cashmere, and meat. *GigaScience*. 7:giy105.

Zhang RQ, Wang JJ, Zhang T, Zhai HL, and Shen W. 2019. Copy-number variation in goat genome sequence: A comparative analysis of the different litter size trait groups. *Gene*. 696:40-6.

Zhao FQ and Keating AF. 2007. Expression and regulation of glucose transporters in the bovine mammary gland. *Journal of Dairy Science*. 90:E76-86.

Zhao X, Ponchon B, Lanctôt S, and Lacasse P. 2019. Invited review: Accelerating mammary gland involution after drying-off in dairy cattle. *Journal of Dairy Science*. 102:6701-17.

Zheng Z, Wang X, Li M, Li Y, Yang Z, Wang X, Pan X, Gong M, Zhang Y, Guo Y, et al. 2020. The origin of domestication genes in goats. *Science Advances*. 6:eaaz5216.

Zhou J, Chehab R, Tkalcevic J, Naylor MJ, Harris J, Wilson TJ, Tsao S, Tellis I, Zavarsek S, Xu D, et al. 2005. Elf5 is essential for early embryogenesis and mammary gland development during pregnancy and lactation. *The EMBO journal*. 24:635-44.

Zhou X and Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. 44:821-4.

Zidi A, Casas E, Amills M, Jordana J, Carrizosa J, Urrutia B, and Serradilla JM. 2014. Genetic variation at the caprine lactalbumin, alpha (*LALBA*) gene

and its association with milk lactose concentration. *Animal Genetics*. 45: 612-3.

Chapter 7

Annexes

All Supplementary Tables, Figures and related Documents for the published (paper I, II and III) and unpublished papers (paper IV and V) are respectively available at their corresponding online versions and at Figshare database:

Paper I

https://jasbsci.biomedcentral.com/articles/10.1186/s40104-020-00435-4

Paper II

https://www.journalofdairyscience.org/article/S0022-0302(19)30294-2/fulltext

Paper III

https://gsejournal.biomedcentral.com/articles/10.1186/s12711-020-00564-4

Paper IV

https://doi.org/10.6084/m9.figshare.12933881

Paper V

https://doi.org/10.6084/m9.figshare.12933848

Additionally, all Supplementary Tables, Figures and related Documents are publicly available and can be downloaded from the following link: https://figshare.com/projects/Supplementary_Documents_included_in_the_PhD_Thesis_authored_by_Dailu_Guan_August_2020_/88346

Acknowledgements

The three-year Ph.D. study is very fast, and now it is closing to the completion of my Thesis. I would love to appreciate many people who provide direct and indirect help and support during these three years. First of all, I give my thanks to my supervisor Marcel Amills. Your supervision and knowledge allowed me to form critical thinking and scientific vision. When a new project was launched, you always explained experimental design and hypothesis very carefully to me. Thanks also for your patience in correcting manuscripts, which led me understand how to prepare a scientific publication. As well, I thank for your help at every stage of this thesis. Without you, it's impossible to finish all of the Ph.D. works.

I also thank to the rest of the researchers from the Animal Genomics group at CRAG, Alex Clop, Miguel Pérez, Sebastián Ramos, Armand Sanchez and Josep Maria Folch. All of you gave me wise advice and invaluable encouragement. Many thanks to Anna Castelló and Betlem Cabrera, who organized and managed the lab, importantly provided invaluable help in my lab works.

Many thanks, of course, to our team members, Tainã Figueiredo Cardoso, Emilio Mármol Sánchez and Maria Gracia Luigi. I remember Emilio took me to visit CRAG and taught me bioinformatics when I first came to CRAG, and the first lab work that I have done at CRAG was under the direction of Tainã. Needless to say, Maria helped me a lot, including the translation of thesis summary into Spanish. All of these things seem to have happened yesterday. Your help is really invaluable.

I would also give thanks to my colleagues at CRAG, Marta Godia, Lourdes Criado, Lino César Ramírez, Laura Zingaretti, Daniel Crespo, Jordi Leno, Yron Joseph Yabut Manaig, Alice Iob, Magi Passols, Jesús Valdés. With you I had a happy three-year in Spain. What I was very impressed is the celebration of the "La Gran Festa del Calçot" happened every year. Thanks, guys!

Many thanks are also given to my Chinese friends. Bin Liu, thank you for your advices and encouragement in these three years. Baoyi Zhang, thanks for your kindness and cares about us. I give my best wishes to your babies, Youheng Liu (Guan Guan) and Youxuan Liu (Xiao Yu), and wish them health and happiness every day. Thanks also for Yu Lian, Xiaoqing Shi, Yaxing Wang, Jinqiang Yan, Rong Zhang, Tingting Qiu, Yi Xiao, Liu Duan, to mention a few here. Without all of you, it would be a tough three-year.

A special thanks to my wife, Lang Xiong. No matter what happens, bad or good, you always stand with me. No words in this world can express my gratitude to you. Love you forever!

“儿行千里母担忧”。最后感谢父母的养育之恩，祝你们平安快乐，健康长寿！