Agrociencia Uruguay 2026 | Volume 30 | Article e1771
DOI: 10.31285/AGRO.30.1771
ISSN 2730-5066

SNP Data Quality Control in the Uruguayan Sheep Breeding Program Database

Carracelas, B. [1]; Ciappesoni, G. [1 2], Navajas, E. A. [1], Aguilar, I. [1 3]

[1] *Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Ganadero Extensivo, Las Brujas, Uruguay*

[2] *Universidad de la República, Facultad de Agronomía, Departamento de Producción Animal y Pasturas, Montevideo, Uruguay*

[3] *Instituto Nacional de Investigación Agropecuaria (INIA), Sistema Lechero, Las Brujas, Uruguay*

Editor

Hugo Naya
Universidad de la República, Facultad de Agronomía, Montevideo, Uruguay

Received *6 Jun 2025*
Accepted *15 May 2026*
Published *27 May 2026*

Correspondence

Beatriz Carracelas
bcarracelas@inia.org.uy

Abstract

Genomic data provides enhanced accuracy to sheep genetic evaluations while speeding up genetic improvement and helping fix pedigree errors. Achieving these outcomes requires efficient data pipelines to automate quality control (QC) and optimize genotypic data analysis during routine genetic evaluations. Our pipeline includes three main steps: genotype QC, based on the per sample call rate; parentage verification against reported sires and dams, and animal QC, which detects duplicate entries and possible errors in sex and breed assignment. These QC procedures help detect sample mix-ups that occur because of laboratory or farm errors. This paper describes the design and implementation of the QC pipeline applied to the MGAdbSNP database, with the aim of supporting robust and accurate genomic evaluations in sheep.

Keywords: quality control, sheep, parentage verification, sex check, breed check

Carracelas, B., Ciappesoni, G., Navajas, E. A., & Aguilar, I. (2026). SNP data quality control in the Uruguayan sheep breeding program database. *Agrociencia Uruguay, 30*, Article e1771. https://doi.org/10.31285/AGRO.30.1771

Control de calidad de datos de genotipado en la base de datos del programa de mejoramiento ovino de Uruguay

Resumen

La información genómica puede mejorar la precisión de las evaluaciones genéticas en ovinos, lo que acelera la ganancia genética y ayuda a corregir los registros genealógicos. Para lograr estos resultados son necesarios procesos eficientes capaces de automatizar el control de calidad (QC) y optimizar el análisis de datos de genotipado en las evaluaciones de rutina. El flujo de trabajo incluye tres pasos principales: control de calidad de los genotipos, basado en el *call rate* por muestra; verificación de paternidad contra los padres declarados, y control de calidad de los animales, que detecta duplicados y posibles errores en la asignación de sexo y raza. Estos procedimientos de control de calidad ayudan a detectar muestras mal asignadas debido a errores de campo o de laboratorio. Este artículo describe el diseño y la implementación del flujo de control de calidad aplicado a la base de datos MGAdbSNP, con el objetivo de respaldar evaluaciones genómicas robustas y precisas en ovinos.

Palabras clave: control de calidad, ovinos, verificación de paternidad, chequeo de sexo, chequeo de raza

Controle de qualidade de dados de genotipagem no banco de dados do programa de melhoramento ovino do Uruguai

Resumo

A informação genômica pode melhorar a precisão das avaliações genéticas em ovinos, acelerando o ganho genético e ajudando a corrigir os registros genealógicos. Para atingir esses resultados são necessários processos eficientes capazes de automatizar o controle de qualidade (CQ) e otimizar a análise de dados de genotipagem nas avaliações de rotina. O fluxo de trabalho inclui três etapas principais: controle de qualidade dos genótipos, baseado na taxa de chamada por amostra; verificação de paternidade contra os pais declarados; e controle de qualidade animal, que detecta duplicatas e possíveis erros na atribuição de sexo e raça. Esses procedimentos de CQ ajudam a detectar amostras atribuídas incorretamente por causa de erros de campo ou de laboratório. Este artigo descreve o desenho e a implementação do pipeline de controle de qualidade aplicado ao banco de dados MGAdbSNP, com o objetivo de apoiar avaliações genômicas robustas e precisas em ovinos.

Palavras-chave: controle de qualidade, ovinos, verificação de paternidade, verificação de sexo, verificação de raça

1. Introduction

In Uruguay, genomic evaluations play a crucial role in livestock improvement and are currently conducted for several economically important breeds, including Hereford, Angus, Holstein, and Australian Merino. The integration of genomic information into breeding programs provides various advantages which improve genetic prediction accuracy, enable the integration of difficult-to-measure traits, and support pedigree records verification (Kaseja et al., 2022), a critical process since incomplete or inaccurate pedigrees can significantly reduce genetic gain (Israel & Weller, 2000). For instance, sire identification errors in Uruguay were estimated at 12.1% for Corriedale and 7.0% for Australian Merino breeds, while on dams the reported magnitude of errors was 8.7% and 5.7%, for the same breeds, respectively (Macedo et al., 2015). These errors can reduce the reliability of breeding programs, which emphasizes the need for advanced genomic tools to support accurate parentage assignment.

In 2013, the Instituto Nacional de Investigación Agropecuaria (National Institute of Agricultural Research; hereafter INIA) established the MGAdbSNP genotype database (I. Aguilar & S. Fernández, personal communication, February 12, 2024), which functions as a vital resource for Uruguay´s livestock research and breeding activities. This repository maintains genotypes with different single nucleotide polymorphism (SNP) densities which derive

from various sources. This database has expanded continuously and now contains about 51,000 cattle genotypes together with 18,500 sheep genotypes.

According to McClure et al. (2018), large-scale genomic datasets need extensive quality control (QC) procedures for both samples and SNPs. Effective QC checks are very important to ensure the accuracy and reliability of the genotypic data, not only to verify the quality of each SNP but also to ensure that the genotypes correspond to the correct animals. A QC pipeline is already in place for cattle genotypes at INIA, but sheep genotypes lack a similar system. To preserve data quality for sheep breeding programs and research on genomics, it is important to implement a QC pipeline for sheep as well.

Several QC pipelines have been described for the management of large-scale genotype datasets in livestock breeding programs. In dairy cattle, Wiggans et al. (2009) described SNP selection and QC procedures used to build the genomic evaluation panel adopted in the United States and Canada. In a later study, McClure et al. (2018) reported a comprehensive QC pipeline for the Irish national beef and dairy cattle system, based on a curated 800 SNP panel for parentage verification. General strategies for processing and QC of Illumina genotyping arrays have also been described (Zhao et al., 2018). In sheep, however, integrated QC pipelines designed for the routine processing of breeding-program databases are less commonly reported. The pipeline described here adapts these principles to the particular characteristics of sheep genotypes, in particular the absence of Y-chromosome markers in commercial chips and the diversity of SNP densities used over time within the same database.

This paper focuses on both the design and implementation of a complete QC pipeline that focuses on sheep SNP genotypes. The development of this pipeline aims to improve genomic data accuracy through effective QC procedures supporting the advancement of Uruguay's sheep industry.

2. Materials and Methods

2.1 MGAdbSNP Database

The MGAdbSNP database stores SNP genotypes from two sources: INIA's Animal Genomic DNA Laboratory (Carracelas et al., 2022) and external institutions such as the Asociación Rural del Uruguay (Rural Association of Uruguay; ARU) and several breed associations. The database contains comprehensive information about each genotype including tag numbers, species, breed and sex, along with pedigree records and phenotypic data. The sheep genomic data contains mainly purebred breeds that have been genotyped with different SNP chips (Table 1). For the analyses described in this paper, two filters were applied: genotypes from very low-density chips (170 and 507 SNPs) were excluded because their marker content was insufficient for the QC procedures described below, and crossbred animals were excluded so that the analyses were restricted to purebred animals from the seven breeds with the largest number of samples in the database. After applying these filters, 8,532 genotypes remained (Table 1).

The chips used in the database vary widely in marker density and manufacturer (Table 1), ranging from low-density panels (5,000 to 15,000 SNPs) to medium-density panels (40,000 to 60,000 SNPs) and a single high-density panel (Illumina OvineHD 600K, 606,006 SNPs). Because the chips share only a fraction of their content, the number of SNPs common across chips is markedly smaller than the density of any individual chip: 5,882 SNPs were common to all chips used across the seven breeds, whereas 34,591 SNPs were common to the subset of chips used to genotype Corriedale, Creole, and Merilin.

2.2 Genotype QC

The QC pipeline applied to the MGAdbSNP database, summarized in Figure 1, applies baseline QC checks at sample and animal levels. Genotypes received by the MGAdbSNP database are not raw signal data but already processed genotype calls delivered by the genotyping laboratories, which apply their own QC procedures (clustering quality, per SNP call rate) before releasing the data. The QC pipeline described in this paper applies baseline QC checks at sample and animal levels, providing curated genotypes that can subsequently be used for any downstream analysis. Additional SNP level filters (minor allele frequency or Hardy-Weinberg equilibrium thresholds) and stricter sample filters are not applied at the database level, since their definition depends on the specific downstream analysis.

The initial stage of the QC pipeline involves assessing genotype quality using the per-sample missing call rate (Laurie et al., 2010), computed over all SNPs available on each animal's chip, defined as the proportion of SNPs without a reliable call (Berry & Spangler, 2023). Per sample call rates are provided by the genotyping laboratory, and an in-house script applies a fixed threshold to the database. Samples with a call rate ≥ 85% are retained and made available for downstream analyses, while samples below this threshold are flagged in the MGAdbSNP database as unreliable and cannot be exported, although the records are preserved. The 85% threshold follows the recommendation by Purfield et al. (2016).

2.3 Parentage Verification

Parentage verification was performed with seekparentf90 v1.55 (Aguilar, 2014) by comparing each animal's genotype to its reported parents and counting the proportion of mismatches at homozygous SNPs (Wiggans et al., 2009). We used 856 highly informative SNPs for parentage assignment, compiled from those reported by ISGC (Kijas et al., 2012), CSIRO (Bell et al., 2013), AgResearch (Clarke et al., 2014), USDA (Heaton et al., 2014), INRA (Tortereau et al., 2017), and INIA (Macedo et al., 2014). Some SNPs were reported by more than one institution, so the final non-redundant set used in this study contained 856 SNPs. Details on the contributing groups, target breeds, and source publications are provided in Supplementary Table S1.

Match, doubtful, and exclusion statuses were assigned according to the ICAR cattle guidelines (ICAR, 2022) for parentage verification, applied with the *--icar* option of seekparentf90. ICAR defines these thresholds as absolute SNP mismatch counts for panels of 100 and 200 SNPs, and as percentages for larger panels. Since our set contains 856 SNPs, the percentage-based thresholds were applied: in single parent testing, the genotype of an animal is compared against the genotype of one reported parent (sire or dam), the combination is classified as match when the proportion of mismatches was ≤1.0%, doubtful for 1.0-3.0%, and excluded above 3%. In trio testing, the genotype of an animal is compared simultaneously against both reported parents (sire and dam) when both have been genotyped, and consistency is checked across the trio: the thresholds were ≤1.5% (match), 1.5-4.0% (doubtful), and >4% (exclusion). Parentage verification was also tested using the set of SNPs common to the chips used within each breed instead of the 856 SNP set; however, the parentage SNP set yielded fewer doubtful cases than the breed-common SNP set, supporting the rationale of using a small set of high-quality SNPs validated specifically for parentage assignment, in line with ICAR guidelines (ICAR, 2022).

2.4 Animal QC

The QC pipeline distinguishes two types of outcomes: some checks exclude problematic genotypes from downstream analyses, while others flag suspicious cases for individual review without excluding them automatically. Automatic exclusions are applied by the call rate filter and by the detection of duplicate genotypes with matching sample IDs. The remaining checks (duplicate sample IDs with differing genotypes, duplicate genotypes with different sample IDs, parentage verification, sex check, and breed check) produce flags that require manual verification of pedigree, farm, or laboratory records before any decision can be made.

2.4.1 Duplicates Check

The next step in the QC pipeline is checking for duplicate genotypes, which usually indicate farm or lab errors, as identical genotypes occur only in monozygotic twins, split embryos, or somatic cell nuclear transfer (SCNT) clones (McClure et al., 2018; VanRaden et al., 2023; Vera et al., 2021). Screening for duplicated genotypes was performed with seekparentf90 v1.55 (Aguilar, 2014) separately within each breed, using the SNPs common to all chips used to genotype animals of that breed, after passing the call rate check. The program flags pairs of samples as duplicates based on a modified Hamming distance, using the default similarity threshold of 0.99 (samples sharing at least 99% of their genotypes are considered duplicates). Three outcomes are identified:

Duplicate genotypes with matching sample IDs: Keep the genotype with the higher SNP density or call rate.

Duplicate sample IDs with differing genotypes: Likely a farm error; additional checks (parentage, breed, sex) help assign the correct genotype.

Duplicate genotypes with different sample IDs: Likely a laboratory mix-up; QC steps determine the correct assignment.

In both cases of error, the genotype ID mismatch cannot be resolved automatically. The other QC steps of the pipeline (parentage, sex, and breed checks) provide additional evidence that helps identify the most likely correct association, but the final assignment typically requires manual verification of pedigree, farm, or laboratory records. Until the case is resolved, the genotype involved in the conflict is flagged as unusable to prevent its use in downstream analyses.

2.4.2 Sex Check

Sex prediction through genotyping serves as a key QC method to verify that DNA samples correspond to the animal's reported sex (McClure et al., 2018). It detects when genetic males are misidentified as females or when genetic females are misidentified as males. Although the ideal approach to predict sex involves using SNPs located in the non-pseudoautosomal (nPAR) region of the X and Y chromosomes, the commercial chips used in our database report SNPs only on the X chromosome.

A previous study (Carracelas, Navajas, et al., 2025) identified the PAR region on the sheep X chromosome (0 to 7.24 Mb) using the ARS-UI_Ramb_v2.0 reference genome and described a sex prediction method relying only on X chromosome SNPs. Following that approach, the X chromosome SNPs in the MGAdbSNP database were split into the PAR and nPAR regions according to the boundary reported in Carracelas, Navajas, et al. (2025), and recoded as two separate chromosomes: chromosome 27 for the PAR region and chromosome 28 for the nPAR region. Sex prediction was then performed with seekparentf90 v1.55 (Aguilar, 2014) using the --chr_x 28 option, so that per-animal heterozygosity was computed only over the nPAR SNPs, where males are expected to be hemizygous and therefore show very low heterozygosity. The Irish Cattle Breeding Federation (ICBF) sex prediction rules (McClure et al., 2018) were then applied to the resulting nPAR heterozygosity values: <5% indicates male, >15% female, and 5-15% is ambiguous.

Different chip versions were combined into groups (e.g., Illumina Ovine SNP50 variants and GGP Ovine 50K versions). The AgResearch Sheep Genomic 8K, 18K, and 60K panels were also grouped due to their exclusive use in the Highlander breed.

Accuracy was computed as:

$$\frac{\boldsymbol{True\ Positives\ (TP)}}{\boldsymbol{True\ Positives\ (TP) + False\ Negatives\ (FN) + Ambiguous\ (A)}}$$

2.4.3 PCA for Breed Check

To confirm each animal's declared breed, principal component analysis (PCA) was performed with PLINK v1.9 (Purcell et al., 2007). PCA results were plotted using R v4.5.1 with the ggplot2 package v4.0.2 (Wickham, 2016). No MAF or LD-based pruning filters were applied, because variants discriminating among breeds may have low MAF in the combined sample, and removing them could mask breed structure. Two analyses were carried out: a multi-breed PCA including all seven breeds, based on the 5,882 SNPs common to every chip used; and a PCA restricted to Corriedale, Creole, and Merilin, based on the 34,591 SNPs common to the chips used in those breeds. The first two principal components (PC1 and PC2) were inspected to visualize breed structure and to identify outliers that suggest incorrect breed assignment (Zhao et al., 2018).

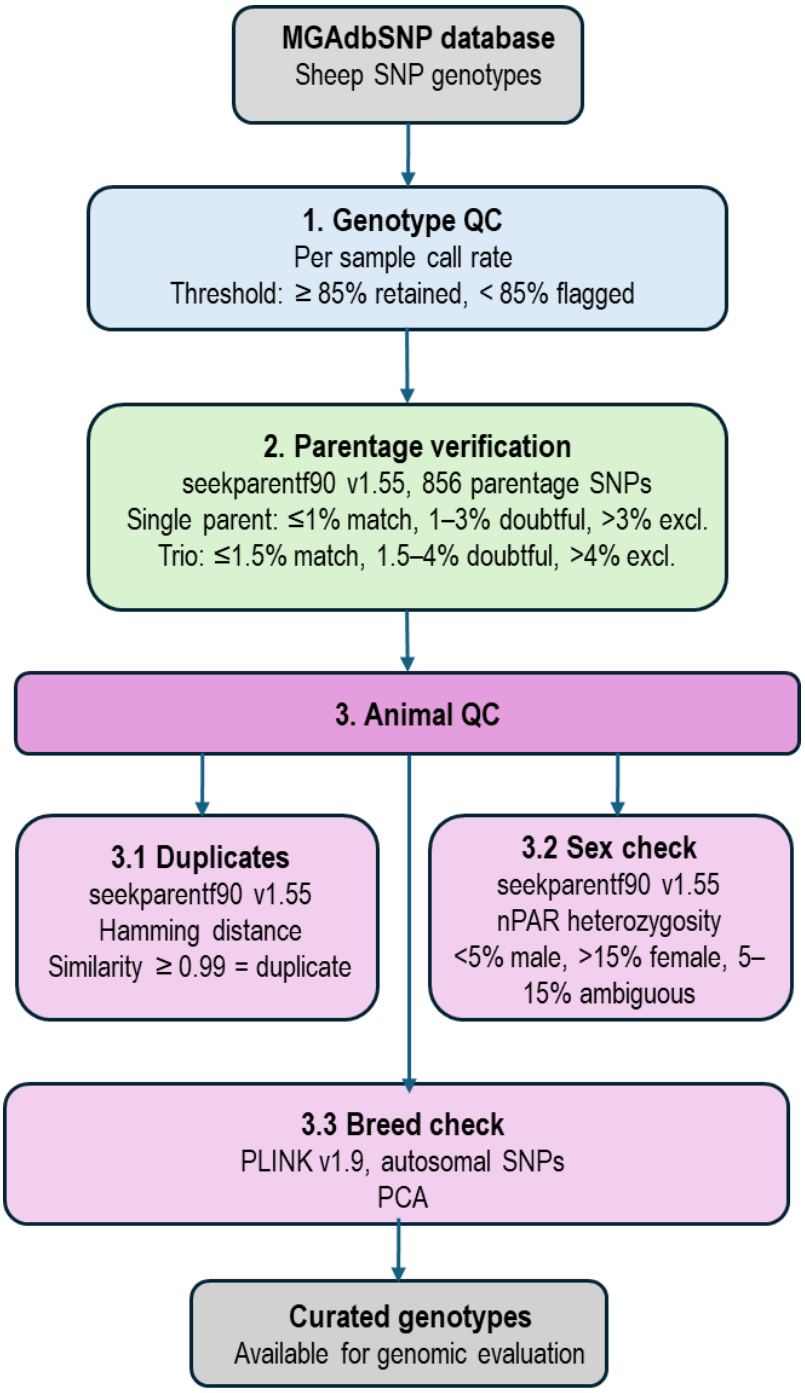

Figure 1. Workflow of the QC pipeline applied to the MGAdbSNP sheep database, showing the three main steps (genotype QC, parentage verification, and animal QC, the latter including duplicate, sex, and breed checks), and the software, parameters, and thresholds applied at each step

3. Results and Discussion

3.1 MGAdbSNP Genotype Database

After applying the two automatic filtering steps of the pipeline (call rate filter and removal of duplicate genotypes with matching sample IDs), the MGAdbSNP database includes 8,532 genotypes from seven breeds with marked differences in sample size (Table 1). Sample distribution across chips is also uneven, reflecting the historical adoption of different SNP arrays at INIA over the years. A small number of animals appear in the database with more than one genotype across different chips. In those exceptional cases, only one genotype per animal is retained for downstream analyses, prioritizing the chip with higher SNP density and higher call rate.

Table 1. Number of samples genotyped with different SNP chips per breed in the MGAdbSNP database

Chip	Total SNPs	Australian Merino	Highlander	Corriedale	Texel	Creole	Dohne Merino	Merilin
ISGC SheepLD2015 15K	15,000	326						
GGP Ovine 50K	45,201	1,543						
GGP Ovine 50Kv2	51,863	378						
Illumina Ovine SNP50	54,177	186		102				
Illumina Ovine SNP50v2	53,453			24				
Illumina Ovine SNP50 (Weatherbys)	51,076			66	218			
Axiom Ovine Genotyping Array (50K)	52,339	1,382		848	685	196	160	43
Axiom Bovine-Ovine-Caprine Genotyping Array	42,855	34						
Illumina OvineHD 600K	606,006			24		170		
AgResearch Sheep Genomic 8K	5,405		391					
AgResearch Sheep Genomic 18K	11,476		1,137					
AgResearch Sheep Genomic 60K	57,760		619					
Total		**3,849**	**2,147**	**1,064**	**903**	**366**	**160**	**43**

*The nominal number of markers per chip is reported as provided by the manufacturer. This table excludes genotypes from low density arrays (170 and 507 SNPs).

3.2 Genotype QC

The genotype QC process flagged 1,142 genotypes (11.7% of the genotypes initially available before this filter) as unusable because their call rates were below the 85% threshold. This filtering step is comparable to the one applied in the Irish national cattle system, where genotypes with call rates below 90% are also excluded from downstream analyses (McClure et al., 2018). The proportion of genotypes excluded in our database is higher than those reported in studies of dairy and beef cattle, where mean call rates are typically above 99% and very few samples fall below the 90% threshold (Purfield et al., 2016). This difference likely reflects three particular features of the MGAdbSNP database: the accumulation of genotypes over more than ten years, including samples generated with older arrays; the diversity of biological sample types and processing flows, since blood samples are processed at the INIA Animal Genomic DNA Laboratory under standardized conditions, while hair samples submitted by external institutions are sent directly to genotyping laboratories where DNA is extracted on-site, and hair-derived DNA is generally of lower quantity and quality than blood-derived DNA, particularly when storage and transport conditions vary (Berry & Spangler, 2023); and the heterogeneity of SNP arrays used, ranging from low to high-density chips of different manufacturers (Table 1). Routine application of this QC step removes unreliable genotypes from genomic evaluations while preserving the original genomic information in the database.

3.3 Parentage Verification

Parentage analysis (Table 2) showed high consistency in dam-progeny and trio testing, while sire-progeny testing yielded a much higher exclusion rate concentrated in Australian Merino (133 of 222 exclusions), Corriedale (57), and Texel (32). Assuming these genotypes correspond to the correct animals, the exclusions may suggest potential errors in sire assignment.

The 13.9% sire exclusion rate observed in our population is consistent with the 12.1% previously estimated for Corriedale in Uruguay using microsatellite markers (Macedo et al., 2015), and falls within the range of values reported for SNP-based parentage testing in cattle: in the Mexican Holstein population, where parentage could not be validated for 17% of the sires of cows and 12% of the sires of bulls (García-Ruiz et al., 2019). The relatively low number of doubtful results across all tests (0.2% in sire-progeny, 0.0% in dam-progeny, and 1.7% in trio testing) supports the discriminative capacity of the ICAR percentage-based thresholds when applied to the 856 SNP set. The complete absence of exclusions in trio testing further confirms that using both dam and sire information provides the most conclusive parentage verification.

Table 2. Parentage verification results for sire, dam, and trio testing

	Sire-progeny tested	Dam-progeny tested	Trio tested
Match	1375 (85.9%)	320 (95.2%)	2613 (98.3%)
Doubtful	3 (0.2%)	0 (0.0%)	44 (1.7%)
Exclusion	222 (13.9%)	16 (4.8%)	0 (0.0%)
Total	1600	336	2657

3.4 Animal QC

3.4.1 Duplicates Check

Table 3. Duplicate IDs (Dupl_ID_all), duplicate genotypes with matching sample IDs (Dupl_ID_geno), duplicate IDs with differing genotypes (Dupl_ID_diff_geno), and duplicate genotypes with different IDs (Dupl_geno_diff_ID) per breed

Breed	Dupl_ID_all	Dupl_ID_geno	Dupl_ID_diff_geno	Dupl_geno_diff_ID
Australian Merino	153	82	71	7
Highlander	0	0	0	31
Corriedale	40	36	4	0
Texel	12	12	0	4
Creole	0	0	0	10
Dohne Merino	0	0	0	0
Merilin	0	0	0	0

Table 3 shows that the 153 duplicated sample IDs for Australian Merino resulted in 82 true duplicates, while the remaining samples had different genotypes. For Corriedale, 36 out of 40 duplicate sample IDs were true duplicates, with only four mismatched genotypes. In the case of the Texel breed, all 12 duplicate sample IDs were confirmed as true duplicates. Additionally, the analysis detected 52 genotypes from Australian Merino, Texel, Highlander and Creole breeds that were identified as duplicate genotypes but linked to different sample IDs. Duplicate detection has been highlighted as a critical QC issue in large national breeding programs such as the Irish (ICBF) (McClure et al., 2018) and the US/Canadian dairy genomic evaluation (Wiggans et al., 2009), where comparing each new genotype against millions of records is computationally demanding. The 52 cases of duplicate genotypes with different sample IDs detected here are particularly relevant, since without a genotype-based comparison they would silently appear as independent animals in the genomic evaluation.

Of these duplicate cases, the 130 duplicate genotypes with matching sample IDs (sum of Dupl_ID_geno across breeds) were resolved automatically by retaining one genotype per pair, prioritizing the chip with higher SNP density and higher call rate. The remaining cases, 75 duplicate sample IDs with differing genotypes (sum of Dupl_ID_diff_geno) and 52 duplicate genotypes with different sample IDs (sum of Dupl_geno_diff_ID), were flagged in the database for individual review, since their resolution requires verification of farm or laboratory records.

3.4.2 Sex Check

The number of X chromosome nPAR SNPs available for sex prediction varied substantially across chips, ranging from 169 to 19,687 (Table 4).

Sex prediction accuracy varied substantially across chips (Table 4; Supplementary Table S2). Most chips achieved accuracies above 97% for both sexes, with the ISGC SheepLD2015 15K, GGP Ovine 50K/50Kv2, and AgResearch series performing the best (≥99% in males and 98% in females). The Illumina OvineHD 600K showed the lowest female accuracy (84.5%) and the highest proportion of ambiguous female calls (16%), in contrast with its high male accuracy (98.7%). The Axiom Ovine Genotyping Array (50K) showed the opposite pattern, with female accuracy higher than male accuracy. The Axiom Bovine-Ovine-Caprine Genotyping Array, tested only on males, achieved 94% accuracy.

Table 4. Sex prediction accuracy for males and females by chip

Chip	Total SNPs[1]	nPAR SNPs[2]	Accuracy males	Accuracy females
Illumina Ovine SNP50				
Illumina Ovine SNP50 Weatherbys	41,625	1,103	97.8	88.6
Illumina Ovine SNP50v2				
Axiom Ovine Genotyping Array (50K)	32,182	707	94.4	98.7
ISGC SheepLD2015 15K	12,899	219	100.0	100.0
Axiom Bovine-Ovine-Caprine Genotyping Array	35,698	742	94.1	
AgResearch Sheep Genomic 8K				
AgResearch Sheep Genomic 18K	4,610	169	99.1	98.6
AgResearch Sheep Genomic 60K				
Illumina OvineHD 600K	485,615	19,687	98.7	84.5
GGP Ovine 50K	18,360	545	99.2	99.5
GGP Ovine 50Kv2				

[1] Total SNPs: number of SNPs common to all chips within the group, used for the analysis. [2] nPAR SNPs: number of SNPs from the non-pseudoautosomal region of the X chromosome (chromosome 28 in seekparentf90 coding) used for sex prediction.

The number of nPAR SNPs available on each chip is one factor that may explain the differences in sex-prediction accuracy across chips, but the breed composition of the samples genotyped on each chip is equally important. Chips with very few but well-positioned nPAR SNPs (e.g., AgResearch with 169 SNPs) achieved high accuracy in both sexes (99.1% males, 98.6% females), confirming that a small but informative set of nPAR markers is sufficient for reliable sex calls. In contrast, the Illumina OvineHD 600K, which carries the largest number of nPAR SNPs in the database (19,687), showed the highest proportion of ambiguous female calls (16%) and the lowest female accuracy (84.5%). In our database, this chip was used almost exclusively for Creole sheep genotyping (Table 1), a breed that was not represented among the breeds used for SNP discovery during the design of the OvineHD 600K array. This is consistent with ascertainment bias (Ajmone-Marsan et al., 2023): SNPs selected as polymorphic in the breeds used for chip development tend to be less polymorphic, and therefore less informative, when applied to breeds that were not included in the discovery panel, reducing the heterozygosity observed

on the X chromosome and increasing the chance of female calls falling into the ambiguous range. Thus, beyond chip density, the choice of nPAR SNPs and the match between the breeds genotyped and those used in chip design are both key determinants of sex-prediction performance.

3.4.3 PCA for Breed Check

For the multi-breed PCA, 391 Highlander samples genotyped with the AgResearch Sheep Genomic 8K chip were excluded to retain a larger set of SNPs common across all chips, leaving 8,141 animals. The resulting PCA revealed distinct clusters for two of the seven breeds analyzed: Texel and Highlander (Figure 2). The Australian Merino cluster stretched into a distinctive shape while it partially overlapped with the Dohne Merino cluster on one side. This overlap is expected, as the Dohne Merino was developed in South Africa by crossing the Peppin-type Merino with the German Mutton breed (https://dohne.com.au).

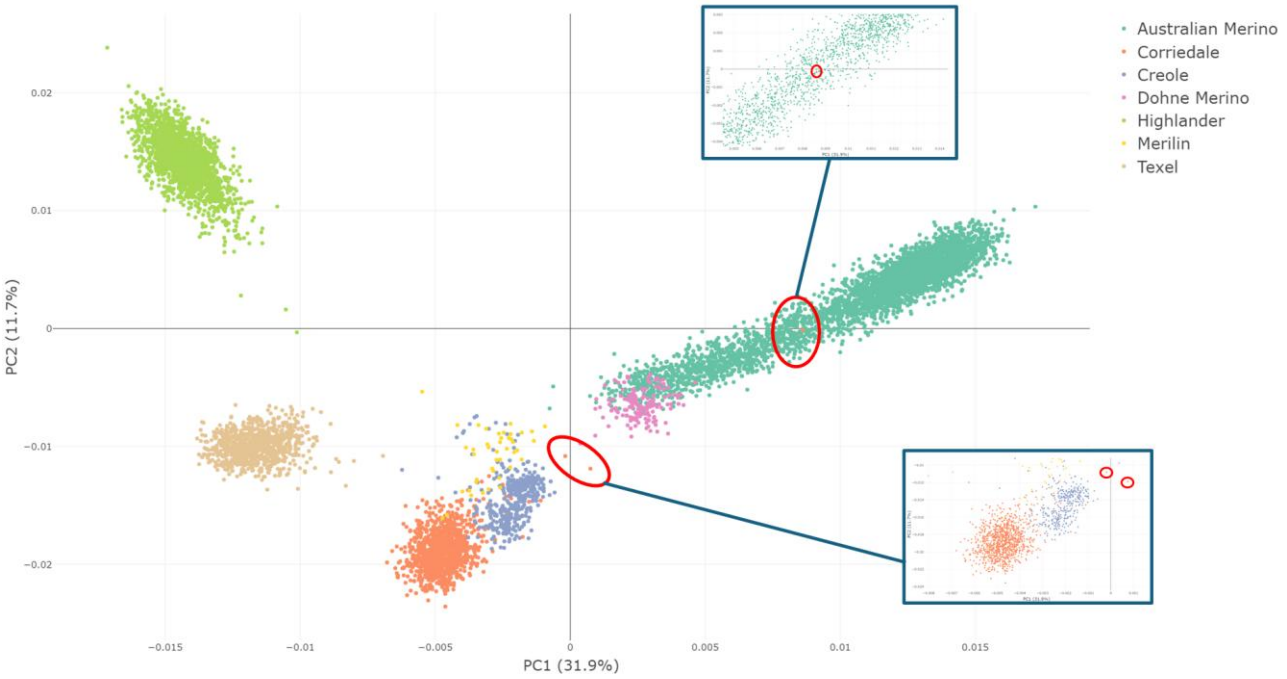

Figure 2. Principal components (PC1 and PC2) plot from a PCA analysis of 8,141 animals across 7 breeds, based on 5,882 SNPs common to all chips used

PC1 and PC2 explain 31.9% and 11.7% of the total genotypic variance, respectively. Highlander samples genotyped with the AgResearch Sheep Genomic 8K chip were excluded to preserve a larger set of common SNPs.

Creole and Corriedale overlapped in the PC1-PC2 plane, with additional overlap observed with Merilin. The proximity between Corriedale and Merilin is consistent with their shared ancestry through Lincoln and Merino crossbreeding, while the proximity between Creole and Corriedale can be explained because Creole was the genetic foundation when the Corriedale breed was introduced to Uruguay. These breeds have since been managed together on farms, which has enabled gene exchange between them (Carracelas, Peraza, et al., 2025).

Some samples are positioned outside their expected breed clusters, which may suggest possible sample mix-ups or the presence of non-purebred animals. These cases need further evaluation. For instance, a Corriedale sample within the Australian Merino cluster (upper zoom box in Figure 2) had a sire exclusion when parentage verification was performed. This discrepancy suggests that the genotype may not correspond to this animal, indicating an error at the farm or laboratory level.

Additionally, several Corriedale genotypes appeared within the Creole cluster (Figure 2). One possible explanation is that, in the multi-breed PCA, the smaller set of SNPs shared across all chips reduces resolution and makes Corriedale and Creole appear genetically closer than they actually are. When the PCA was restricted to Corriedale, Creole and Merilin, the larger set of shared SNPs allowed the Creole genotypes to separate into three distinct clusters based on their origin: Creoles from San Miguel National Park (Creole_SM), Creoles from INIA Las Brujas (Creole_LB), and Creoles from commercial farms (Creole_COM), which group together with Creoles from Catamarca, Argentina (Creole_AR). These clusters are clearly distinct from the Corriedale and Merilin clusters, which appear closely related (Figure 3).

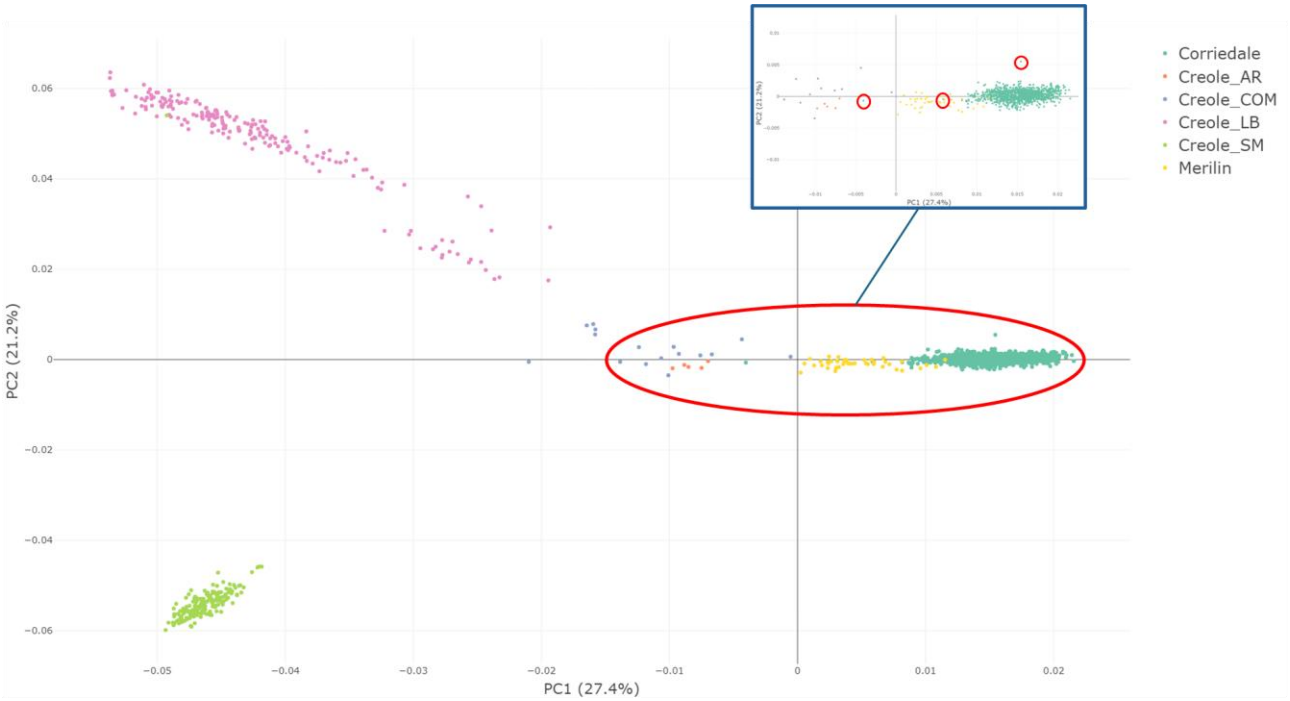

Figure 3. Plot of principal components (PC1 and PC2) from a PCA analysis including only Corriedale, Creole, and Merilin breeds, based on 34,591 SNPs common to the chips used in those breeds

PC1 and PC2 explain 27.4% and 21.2% of the total genotypic variance, respectively. Creole samples are differentiated by origin: San Miguel National Park (Creole_SM), INIA Las Brujas (Creole_LB), commercial farms (Creole_COM), and Catamarca, Argentina (Creole_AR).

Another example is observed with two Corriedale genotypes that appear distant from their breed cluster and position themselves to the right of the Creole cluster (bottom zoom box in Figure 2). These two samples display proximity to the Dohne Merino cluster. The first animal confirmed parentage with both its sire and progeny but the second animal showed sire exclusion. Both genotypes, along with the Corriedale genotype that grouped with Australian Merino in Figure 2, are positioned outside the Corriedale cluster in Figure 3 (marked with red circles in the zoom box). The first genotype passed parentage verification but the other two may not correspond to their expected breed, leading to potential misidentification.

PCA-based breed verification is also used in other large-scale livestock genotyping pipelines (ICBF, with 22,610 reference animals from 14 breeds) (McClure et al., 2018). In our case, combining PCA outliers with parentage and sex check results was particularly informative: the Corriedale sample positioned within the Australian Merino cluster (Figure 2) also failed sire verification, providing stronger evidence of sample mislabeling than either QC step alone. PCA is used here as an exploratory and diagnostic step rather than as a formal criterion for excluding animals: when PCA results remain inconclusive, suspicious samples are flagged for individual review together with the other QC outputs (parentage, sex, duplicates). Formal model-based approaches such as Admixture

(Alexander et al., 2009) or supervised classification methods could be incorporated in future versions of the pipeline to provide quantitative probabilities of breed assignment, particularly for closely related breeds with overlapping clusters or for individuals suspected of admixture.

4. Conclusions

Multiple factors which impact genotype quality were identified through the QC pipeline applied to the MGAdbSNP database. The detection of misassigned samples relies on genomic QC checks which serve as critical tools for identifying errors that occur at the farm or in the laboratory.

Routine application of the pipeline has direct implications for sheep genomic evaluations and breeding programs in Uruguay. By removing unreliable genotypes and flagging cases that require manual verification (parentage conflicts, sex discrepancies, and breed inconsistencies), the pipeline contributes to more accurate genomic predictions, more reliable estimation of genetic parameters, and ultimately more efficient genetic improvement decisions. The modular design of the pipeline, combining a baseline filtering step with multiple flagging steps, allows it to be adapted to other sheep breeding programs operating with multi-chip, multi-source genotype databases. Future extensions may incorporate model-based approaches for breed assignment like Admixture (Alexander et al., 2009) and additional statistical filters at the analysis stage, depending on the specific downstream applications.

Acknowledgements

This work has received funding from the National Institute of Agricultural Research (INIA SGE_06).

This research was conducted as part of a PhD thesis at the School of Agronomy, Universidad de la República, Uruguay.

Transparency of Data

Available data: Part of the data used in this study, corresponding to Australian Merino, Corriedale, Creole, and Texel populations, were uploaded by the SMARTER Project and are accessible via the GigaDB repository.

Cozzi, P., Manunza, A., Ramirez-Diaz, J., Tsartsianidou, V., Gkagkavouzis, K., Peraza, P., Johansson, A. M., Arranz, J. J., Freire, F., Kusza, S., Biscarini, F., Peters, L., Tosser-Klopp, G., Ciappesoni, G., Triantafyllidis, A., Rupp, R., Servin, B., & Stella, A. (2024). SMARTER-database: A tool to integrate SNP array datasets for sheep and goat breeds. *Gigabyte*. https://doi.org/10.46471/gigabyte.139

SMARTER: Small ruminants breeding for efficiency and resilence. (n.d). https://webserver.ibba.cnr.it/smarter/about

The remaining data can be obtained from the authors upon reasonable request.

Author Contribution Statement

	B Carracelas	G Ciappesoni	EA Navajas	I Aguilar
Conceptualization	■			■
Data curation	■			
Formal analysis	■			
Methodology	■			■
Resources		■	■	
Supervision		■	■	■
Visualization	■			■
Writing – original draft	■			
Writing – review and editing		■	■	■

References

Aguilar, I. (2014). *SeekParentf90* (Version 1.55) [Software]. University of Georgia. http://nce.ads.uga.edu/wiki/doku.php?id=readme.seekparentf90

Ajmone-Marsan, P., Boettcher, P., Colli, L., Ginja, C., Kantanen, J., & Lenstra, J. A. (2023). *Genomic characterization of animal genetic resources: Practical guide*. FAO. https://doi.org/10.4060/cc3079en

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*(9), 1655-1664. https://doi.org/10.1101/gr.094052.109

Bell, A., Henshall, J., Gill, S., Gore, K., & Kijas, J. (2013). Success rates of commercial SNP based parentage assignment in sheep. In Association for the Advancement of Animal Breeding and Genetics (Ed.), *Proceedings of the Twentieth Conference: Translating Science into Action* (pp. 278-281). https://www.aaabg.org/aaabghome/AAABG20papers/bell20278.pdf

Berry, D., & Spangler, M. (2023). Animal board invited review: Practical applications of genomic information in livestock. *Animal, 17*(11), Article 100996. https://doi.org/10.1016/j.animal.2023.100996

Carracelas, B., Navajas, E., Ciappesoni, G., & Aguilar, I. (2025). Identification of the pseudoautosomal region of the X chromosome in sheep and sex prediction using the ARS-UI_Ramb_v2.0 genome assembly. *Agrociencia Uruguay, 29*, Article e1587. https://doi.org/10.31285/AGRO.29.1587

Carracelas, B., Peraza, P., Vera, B., & Ciappesoni, G. (2025). Genetic diversity and population structure of a Creole sheep flock from Uruguay. *Czech Journal of Animal Science, 70*(5), 173-182. https://doi.org/10.17221/93/2024-CJAS

Carracelas, B., Peraza, P., Vergara, A., Ciappesoni, G., Ravagnolo, O., Aguilar, I., Lema, O. M., & Navajas, E. A. (2022). Banco de ADN genómico animal: Plataforma de evaluación genómica. *Revista INIA,* (71), 38-42.

Clarke, S. M., Henry, H. M., Dodds, K. G., Jowett, T. W. D., Manley, T. R., Anderson, R. M., & McEwan, J. C. (2014). A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep. *PLoS ONE, 9*(4), Article e93392. https://doi.org/10.1371/journal.pone.0093392

García-Ruiz, A., Wiggans, G. R., & Ruiz-López, F. J. (2019). Pedigree verification and parentage assignment using genomic information in the Mexican Holstein population. *Journal of Dairy Science, 102*(2), 1806-1810. https://doi.org/10.3168/jds.2018-15076

Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., Maddox, J. F., Basnayake, V., Petrik, D. T., Simpson, B., Smith, T. P. L., & Chitko-McKown, C. G. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS ONE, 9*(4), Article e94851. https://doi.org/10.1371/journal.pone.0094851

ICAR. (2022). Section 4: Guidelines for DNA technologies. In *The global standard for livestock data.* https://www.icar.org/Guidelines/04-DNA-Technology.pdf

Israel, C., & Weller, J. I. (2000). Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. *Journal of Dairy Science, 83*(1), 181-187. https://doi.org/10.3168/jds.S0022-0302(00)74869-7

Kaseja, K., Mucha, S., Yates, J., Smith, E., Banos, G., & Conington, J. (2022). Discovery of hidden pedigree errors combining genomic information with the genomic relationship matrix in Texel sheep. *Animal, 16*(3), Article 100468. https://doi.org/10.1016/j.animal.2022.100468

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K., Paiva, S., Barendse, W., Ciani, E., Raadsma, H., McEwan, J., & Dalrymple, B. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology, 10*(2), Article e1001258. https://doi.org/10.1371/journal.pbio.1001258

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., … Weir, B. S. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology, 34*(6), 591-602. https://doi.org/10.1002/gepi.20516

Macedo, F., Navajas, E. A., Aguilar, I., Grasso, A. N., Pieruccioni, F., & Ciappesoni, G. (2014). New parentage testing SNP panel for commercial breeds will be a useful tool for conservation of Creole sheep. In *Proceedings of the 10th World Congress on Genetics Applied to Livestock Production*. https://ainfo.inia.uy/digital/bitstream/item/4411/1/Ciappesoni.-441-paper-9067-manuscript-402-0.pdf

Macedo, F., Pieruccioni, F., Ciappesoni, G., & Navajas, E. A. (2015). Algunas aplicaciones de la genómica en poblaciones con y sin genealogía conocida. In *IX Jornada de Agrobiotecnología INIA: Apostando a la innovación para un futuro innovador* (pp. 25-29). INIA.

McClure, M. C., McCarthy, J., Flynn, P., McClure, J. C., Dair, E., O'Connell, D. K., & Kearney, J. F. (2018). SNP data quality control in a national beef and dairy cattle system and highly accurate SNP based parentage verification and identification. *Frontiers in Genetics, 9*, Article 84. https://doi.org/10.3389/fgene.2018.00084

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics, 81*(3), 559-575. https://doi.org/10.1086/519795

Purfield, D. C., McClure, M., & Berry, D. P. (2016). Justification for setting the individual animal genotype call rate threshold at eighty-five percent. *Journal of Animal Science, 94*(11), 4558-4569. https://doi.org/10.2527/jas.2016-0802

Tortereau, F., Moreno, C. R., Tosser-Klopp, G., Servin, B., & Raoul, J. (2017). Development of a SNP panel dedicated to parentage assignment in French sheep populations. *BMC Genetics, 18*(1), Article 50. https://doi.org/10.1186/s12863-017-0518-2

VanRaden, P. M., Fok, G., Toghiani, S., & Nicolazzi, E. (2023). Modeling identical twins and clones in genetic evaluations. *Interbull Bulletin, 59*, 63-68.

Vera, B., De Barbieri, I., Ferreira, G., Navajas, E. A., Carracelas, B., & Ciappesoni, G. (2021). Asignación de parentesco y detección de superfecundación heteropaternal en ovinos Merino Australiano mediante paneles de SNP. *Archivos Latinoamericanos de Producción Animal, 29*(Suppl. 1), 85-87. https://revista.alpaenlinea.org/index.php/alpa/article/view/2950

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Wiggans, G. R., Sonstegard, T. S., VanRaden, P. M., Matukumalli, L. K., Schnabel, R. D., Taylor, J. F., Schenkel, F. S., & Van Tassell, C. P. (2009). Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science, 92*(7), 3431-3436. https://doi.org/10.3168/jds.2008-1758

Zhao, S., Jing, W., Samuels, D. C., Sheng, Q., Shyr, Y., & Guo, Y. (2018). Strategies for processing and quality control of Illumina genotyping arrays. *Briefings in Bioinformatics, 19*(5), 765-775. https://doi.org/10.1093/bib/bbx012

Supplementary Material

Table S1. Origin of the 856 SNPs used for parentage verification

Contributing group	Target breeds	SNPs reported	Reference
ISGC	Various	88	Kijas et al. (2012)
CSIRO	Australian Merino, Dohne Merino	383	Bell et al. (2013)
AgResearch	Texel, Romney, Perendale, Booroola, Merino × Romney	84	Clarke et al. (2014)
USDA	74 breed groups	163	Heaton et al. (2014)
INIA Uruguay	Australian Merino, Corriedale, Texel	258	Macedo et al. (2014)
INRA	30 French breeds	192	Tortereau et al. (2017)
Final non-redundant set used in this study		**856**	

*The sum of SNPs reported by the contributing groups exceeds 856 because some SNPs were reported by more than one institution; duplicates were removed to obtain the final non-redundant set.

Table S2. Detailed sex prediction results by SNP chip

Chip	Total SNPs	N	nPAR SNPs	Males				Females			
				Reported	TP	FN	A	Reported	TP	FN	A
Illumina Ovine SNP50 / SNP50 (Weatherbys) / SNP50v2	41.625	517	1.103	403	394	9	0	114	101	13	0
Axiom Ovine Genotyping Array (50K)	32.182	3.607	707	1.831	1.728	25	78	1.776	1.753	15	8
ISGC SheepLD2015 15K	12.899	326	219	279	279	0	0	47	47	0	0
Axiom Bovine-Ovine-Caprine Genotyping Array	35.698	34	742	34	32	1	1	–	–	–	–
AgResearch Sheep Genomic 8K / 18K / 60K	4.610	2.142	169	972	963	9	0	1.170	1.154	12	4
Illumina OvineHD 600K	485.615	193	19.687	77	76	1	0	116	98	2	16
GGP Ovine 50K / 50Kv2	18.360	1.639	545	602	597	1	4	1.037	1.032	1	4

N: total animals genotyped per chip; nPAR SNPs: SNPs in the non-pseudoautosomal region of the X chromosome used for sex prediction; TP: True Positives (predicted sex matches the recorded sex); FN: False Negatives (predicted sex does not match the recorded sex); A: Ambiguous predictions.