Using genomic information to predict sex in dairy cattle

IL Zhang*, C Couldrey, RG Sherlock

LIC, Private Bag 3016, Hamilton.

**Corresponding author:* Email: izhang@lic.co.nz

Abstract

Sex misassignment of animals in breeding programmes can arise from sources such as misidentification and data-entry errors. The objective of this study was to develop a test for checking sex using Single Nucleotide Polymorphism (SNP) chip genotypes on chromosome X. All animals in this study were genotyped using the GeneSeek Genomic Profiler LD v3 (GGPv3) SNP chip. After filtering, 16478 sex-verified animals (15703 cows and 775 bulls) were available in the training set. The proportion of heterozygous SNPs on chromosome X was used to predict sex. The distribution of heterozygosity was bimodal in the training population with a clear gap between 0.01 and 0.03. Using the heterozygosity cutoff of 0.02 resulted in 100% and 99.6% prediction accuracy for males and females, respectively. Additional analysis determined that four incorrectly classified females were either XO or inbred. The test was validated on an independent set of 2212 animals, which resulted in seven animals for which the predicted sex differed to the recorded sex. Physical examination of these animals revealed that the predicted sex was correct in all cases. The method developed is a useful addition to quality control of genotypes obtained as part of a genomic selection programme.

Keywords: genotyping; sex prediction; dairy cattle; SNP arrays; sex chromosome aneuploidy

Introduction

Achieving high genetic gain in breeding programmes depends largely on the accuracy of animal data. However, data errors occur: for example, average misparentage of dairy calves is estimated at 23% (DairyNZ, 2012). A less-studied error type is misassignment of sex. Physical inspection is the traditional method for determining an animal's sex, however, selection decisions are increasingly being made remotely using DNA information. With the increasing availability of SNP genotypes for genomic selection, it is possible to detect and correct sex misassignments with available data for little additional cost.

Females possess two X chromosomes and can, therefore, have heterozygous genotypes for chromosome X SNPs. The pseudoautosomal region (PAR) is a region of homology between the X and Y chromosomes (Flaquer et al. 2008). Males may appear to be heterozygous for SNPs in the PAR but should be hemizygous for all other chromosome X SNPs because they only have one copy of chromosome X. Several tools exist to predict sex using genotyping data as part of quality control for genome-wide associations studies: PLINK (Purcell et al. 2007), GenABEL (Aulchenko et al. 2007), pyGenClean (Perreault et al. 2013), GWASTools (Gogarten et al. 2012), and SNPflow (Weissensteiner et al. 2013). However we were unable to find any published studies in which cattle sex was tested using these tools. The main objective of this study was to develop a test based on the proportion of heterozygous SNPs (heterozygosity) on chromosome X to predict sex in dairy cattle.

Materials and methods

A population of sex-verified animals was generated from all animals (n=37871) with valid (sample call rate of ≥ 0.95) GeneSeek Genomic Profiler LD v3 genotypes by only retaining females with lactation records and males with an LIC bull code (physically examined prior to purchase). Any animal that was unable to be parentage tested against at least one parent was excluded. For the remaining animals, those that failed parentage testing were excluded as their DNA sample may be potentially misidentified. In total, 15703 females and 775 males were "sex-verified".

The sex-verified animals were used to select a set of informative SNPs from the 1487 SNPs mapped to chromosome X. Firstly, SNPs with greater than 10% missing genotypes were identified as poor quality and excluded. We then identified and excluded SNPs for which more than 1% of males were heterozygous, because they were uninformative for predicting sex. The 1% threshold allows for genotyping errors which may misclassify a small percentage of homozygous SNPs as heterozygous (Hong et al. 2012; Nielsen et al. 2011).

We need to detect and exclude uninformative SNPs (displayed > 0.01 heterozygosity in the sex-verified males) as they are likely to be PAR SNPs or SNPs that are heterozygous due to mistakes in the genotyping or the alignment process. However, the lower the Minor Allele Frequency (MAF), the smaller the chance of observing a heterozygous SNP and this decreases the chance that such SNPs will be identified and excluded. We chose to exclude, based on MAF, any SNP with a less than 99% chance of being correctly identified as uninformative. Through simulation we determined that for our training population of 775 males we could achieve our goal by excluding SNPs with MAF < 0.015.

The heterozygosity of the informative SNP set was examined for the sex-verified female and male populations in the training set. Visual inspection of the histogram was used to determine a test threshold for separating males and females. Animals with a heterozygosity of less than the threshold were classified as male and those with a heterozygosity of greater than or equal to the threshold

were classified as female.

The test was used to predict sex for an independent validation set of 2212 animals that were genotyped using GGPv3. The sex of these animals had been routinely recorded on the LIC database by farmers. As with the training dataset, samples and SNPs with a call rate of less than 0.95 and 0.9, respectively, were excluded. For animals where the predicted sex differed from their recorded sex, two additional tests were undertaken. Firstly, called SNPs for the 19 Y chromosome SNPs available on GGPv3 were counted. If less than or equal to five Y SNPs were called, the animal was classified as female and if 17 or more SNPs were called, the animal was classified as male. Animals with between five and 17 SNPs called would be classified as inconclusive but none were observed. Secondly, LogR ratio values of the entire genotyping batch were read into Golden helix SVS (Golden Helix, Inc., Bozeman, MT) from Illumina final reports. Batches of genotyping were kept separate to avoid batch effects. Using a genetic marker map that contained only the informative SNP set, the following parameters were used to ensure a single segment was generated from chromosome X: univariate analysis, maximum segments per 10k markers=1, minimum number markers per segment=20, maximum pairwise segment p-value=0.005. The resulting segmentation file contained the average LogR ratio for the chromosome X segment generated in each animal. The average LogR ratio was examined for male and female animals separately to determine how many copies of chromosome X they possess. The average intensity of the Y chromosome was not examined in this study due to the low number of Y chromosome SNPs on GGPv3.

Results

Of the 1487 available genotyped SNPs on GGPv3 that map to chromosome X, 420 SNPs had a MAF of ≥

Figure 1 The distribution of the proportion of heterozygous animals for SNPs across chromosome X for female and male cattle. All 1487 SNPs before filtering for females (a) and males (b) ; Final informative set of 349 SNPs for females (c) and males (d).

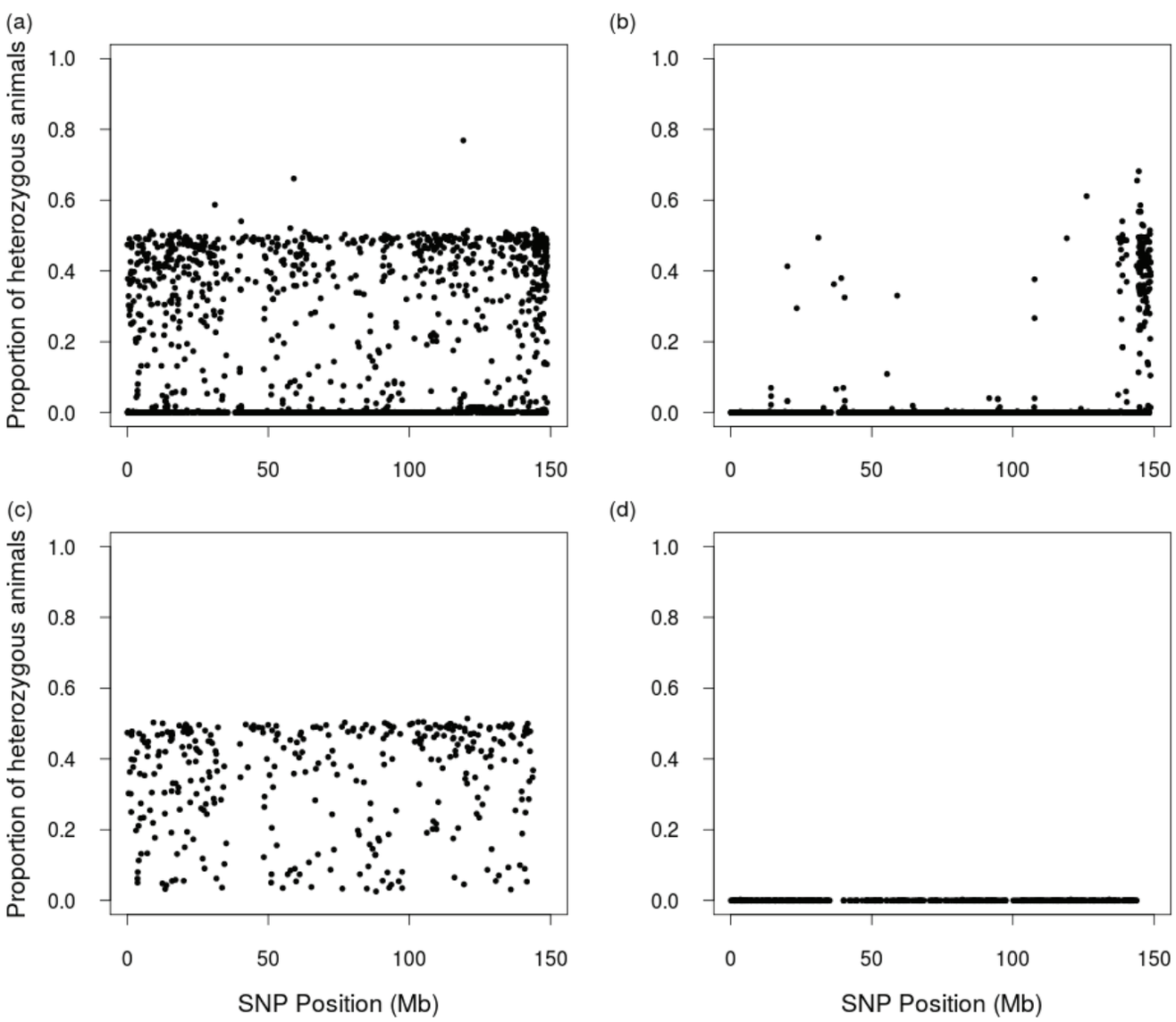


Figure 2 Histogram of the proportion of heterozygous SNPs for the informative SNP set for male cattle (white bars) and female cattle (grey bars).

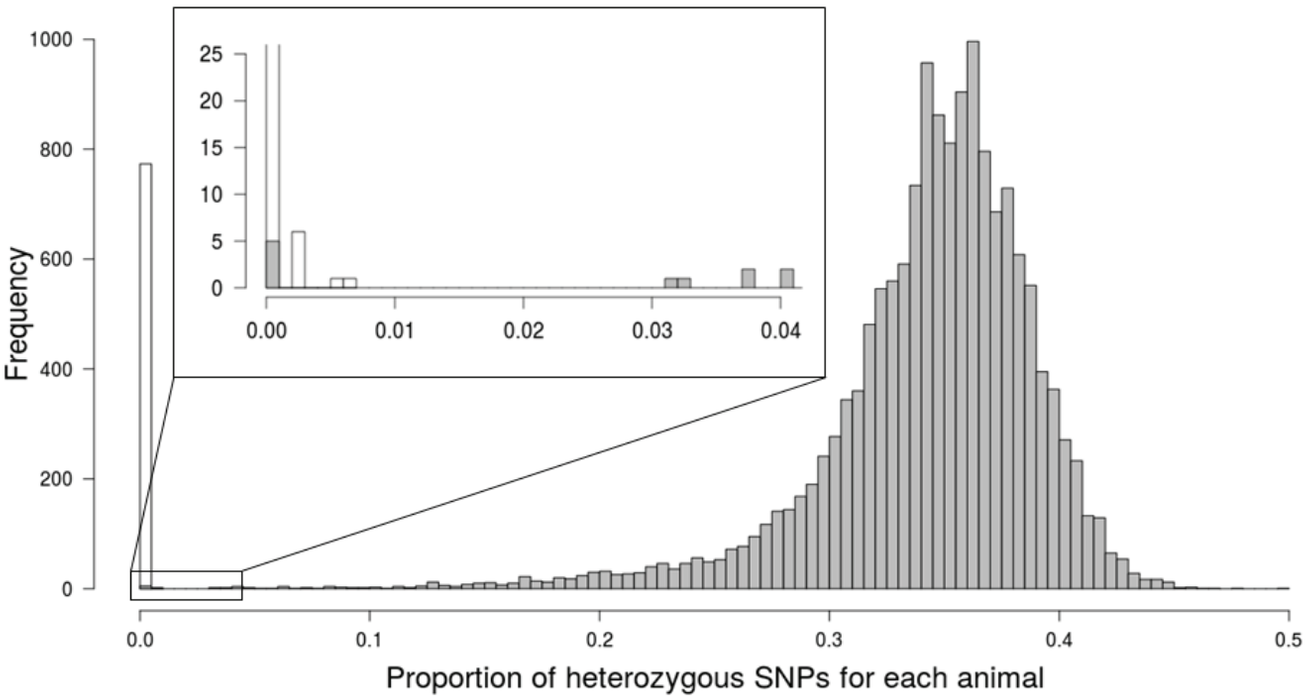


Table 1 Details for the four cattle in the training dataset whose sex was incorrectly predicted.

Animal	Verified sex	Predicted sex	Y chromosome	Number of X chromosomes	Pedigree	Number of offspring
A	Female	Male	No	No data[1]	Recent common ancestor	1
B	Female	Male	No	No data[1]	Recent common ancestor	1
C	Female	Male	No	XX	Dam unrecorded[2]	4
D	Female	Male	No	X	NA	1

[1]LogR ratio unavailable

[2]No dam was recorded for this animal on database

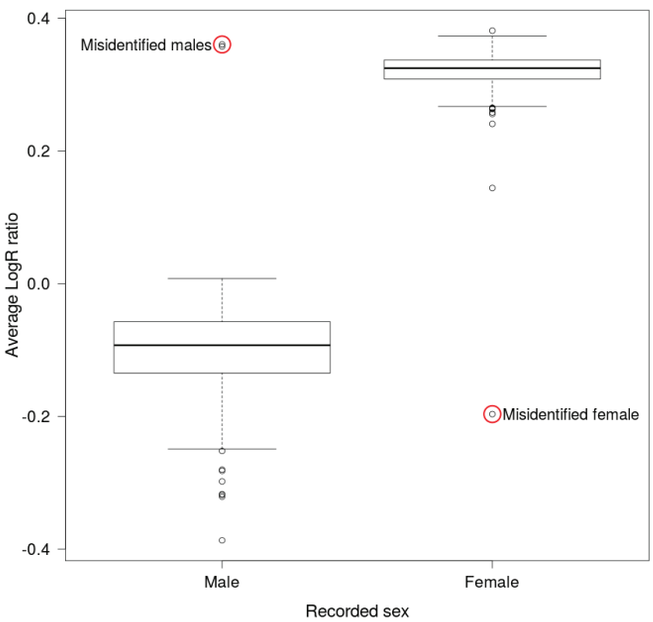
0.015, 1094 SNPs had ≤0.01 heterozygosity, and 349 SNPs satisfied both criteria and became the informative SNP set. Figure 1 shows the proportion of heterozygous female and male animals for SNPs along chromosome X: all chromosome X SNPs (a and b) and informative SNPs (c and d).

Figure 2 shows that all sex-verified males had a proportion of heterozygous SNPs below 0.01, and the majority of sex-verified females had a proportion of heterozygous SNPs above 0.03. Using a threshold of 0.02, the test predicted females in the training dataset with 99.96% accuracy and males with 100% accuracy. Four females were incorrectly predicted to be male: all these animals had a heterozygosity of 0 (i.e., all SNPs in the informative set were homozygous). Further information about each animal is presented in Table 1.

The pedigrees for animals A and B revealed that in both cases their dam and paternal grand dam shared the same sire. The informative SNPs on chromosome X for each animal, her sire, and her maternal grand sire were genotypically concordant, suggesting that the homozygosity on chromosome X was due to two copies of the same ancestral X chromosome passed on by the common ancestor. LogR ratio data indicated that animal C has two copies of chromosome X however she had no dam recorded on database, therefore, we could not perform the same pedigree genotypic concordance check. Animal D had one X chromosome and did not have a Y chromosome and was therefore suspected to have an XO karyotype.

For the independent validation set of animals, seven animals out of 2212 had a recorded sex that differed from their predicted sex. All seven animals were sex-verified by physical examination which showed that the sex predicted by the test was correct for all cases. Three of the animals were incorrectly recorded as female, were completely homozygous for the informative SNP set, and tested positive for the Y chromosome. Figure 3 shows an example of a misidentified female having an average LogR ratio consistent with having one X chromosome. Four animals were incorrectly recorded as males, had high heterozygosity for the informative SNP set, and tested negative for the Y chromosome. These four animals had average LogR ratios consistent with having two X chromosomes (i.e., similar to females); two of these are shown in Figure 3.

Figure 3 Box plot of the average LogR ratio for the informative SNP set for validation male and female cattle from one genotyping batch. Two animals were misidentified as males and another animal was misidentified as female



Discussion

We have developed a test to predict sex in NZ dairy cattle from genomic information. The test correctly identified seven animals that had been misassigned sex records in the independent validation set (2122 animals), indicating that this is an accurate test for the current dataset, in which all animals are genotyped on GGPv3.

A literature search identified a number of tools that use genotypes to predict sex on a variety of platforms. The majority checked sex based on proportion of heterozygous SNPs on chromosome X in combination with a check for the presence of a Y chromosome (Gogarten et al. 2012; Perreault et al. 2013; Purcell et al. 2007). PLINK requires the user to supply a set of SNPs which excludes PAR SNPs as the input and has an in-built tool for achieving this by discarding SNPs beyond specified boundaries. However, this method does not take into account the mistakes caused by incorrect probe binding or inaccuracies in alignments to the chromosome X reference genome. In contrast, our method selects SNPs that are likely to be informative by only including SNPs that displayed ≤ 0.01 heterozygosity in males regardless of position on the X chromosome. Our method and others (for example GenABEL (Aulchenko et al. 2007)) which use customized SNPs for the test population are likely to yield more accurate sex prediction than simply using chromosome X SNPs with SNPs excluded beyond certain cutoff positions.

Sometimes, checking the proportion of heterozygous X chromosome SNPs alone could lead to an incorrect prediction of sex. Example cases include animals with sex chromosome aneuploidies and females where two X chromosomes are identical by descent (IBD). Manually checking for the presence of a Y chromosome and the number of X chromosomes adds another level of confidence to the test.

We identified two animals where the same bull appeared on both sides of the pedigree. By tracing the X chromosome from the ancestral bull through the animal's sire and by checking the concordance between the animals with their respective ancestral bull, we were able to determine that the similarity of the two X chromosomes was due to receiving the same X chromosome from both sides of the pedigree. Therefore, the two X chromosomes are likely to be IBD and this explains the low proportion of heterozygous SNPs on chromosome X for these females. In our study just two females were detected to have IBD X chromosomes (another was suspected but unable to be verified) out of 15703 sex-verified females. The vast majority of genotyped females were sourced from progeny-test herds chosen to represent the NZ dairy population, indicating that the prevalence of these females in the national herd is also likely to be very low. This is expected as the industry has preventative measures such as DataMATE™ to limit inbreeding in the national herd (LIC, 2016).

In the training population, one animal possessed one copy of the X chromosome and no Y chromosome. It is very unlikely that we would be able to detect differences in LogR ratio as the result of mosaicism or chimeraism, therefore, we suspect that the animal in question has an XO karyotype. In general, domestic XO karyotype animals seem to be sterile (Chandley et al. 1975; Di Meo et al. 2008; Iannuzzi et al. 2000; Prakash et al. 1995; Romano et al. 2015; Smith et al. 1989; Zartman, Hinesley & Gnatkowski, 1981). Although the potential XO karyotype female has one recorded lactation, we were unable to rule out the possibility of a sample misidentification as she had already died and could not be resampled. Therefore, we could not confirm her fertility status.

The test offers high accuracy while using existing genotype information and, therefore, offers quality control for little additional cost. Database accuracy is an important factor in genomic selection. In addition, having the correct recorded sex is important to ensure all male calves from contract matings are considered as potential service bulls and some are not missed simply because they were misassigned as females. Sex chromosome aneuploidies are not often considered when selecting animals for genetic improvement, possibly due to the large amount of time and resources needed for karyotyping. Our current test could be extended to routinely examine the number of copies of X chromosomes and the presence of a Y chromosome, providing a fast and low-cost way to detect sex chromosome aneuploidies. This could help improve our understanding of the physical manifestations associated with sex chromosome aneuploidies and how they affect the fertility of New Zealand dairy cattle.

Conclusion

Our study demonstrates the feasibility of applying a sex checking test based on genotyping information in dairy cattle. This will improve database accuracy and may enhance genetic gain in the New Zealand dairy herd.

Acknowledgements

The authors would like to acknowledge Hannah Lee for communicating with farmers, Thomas Lopdell for assistance with figure generation, and Catherine Neeley for data extractions.

References

Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM 2007. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294–1296.

Chandley AC, Fletcher J, Rossdale PD, Peace CK, Ricketts SW, McEnery RJ, Thorne JP, Short RV, Allen WR 1975. Chromosome abnormalities as a cause of infertility in mares. Journal of Reproduction and Fertility. Supplement 23: 377–383.

DairyNZ 2012. Reproof bias correction. http://www.dairynz.co.nz/animal/animal-evaluation/interpreting-the-info/reproof-bias-correction/ [accessed 9 February 2016].

Di Meo GP, Perucatti A, Di Palo R, Iannuzzi A, Ciotola F, Peretti V, Neglia G, Campanile G, Zicarelli L, Iannuzzi L 2008. Sex chromosome abnormalities and sterility in river buffalo. Cytogenetic and Genome Research 120: 127–131.

Flaquer A, Rappold GA, Wienker TF, Fischer C 2008. The human pseudoautosomal regions: a review for genetic epidemiologists. European Journal of Human Genetics 16: 771–779.

Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, Nelson SC 2012. GWASTools: an R/Bioconductor package for quality control and analysis of Genome-Wide Association Studies. Bioinformatics 28: 3329–3331.

Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B, Perkins R, Ge W, Miclaus K, Zhang L, Park K 2012. Technical reproducibility of genotyping SNP arrays used in genome-wide association studies. PLoS One 7: e44483.

Iannuzzi L, Di Meo GP, Perucatti A, Zicarelli L 2000. Sex chromosome monosomy (2n= 49, X) in a river buffalo (Bubalus bubalis). Veterinary Record 147: 690–691.

LIC 2016. DataMATE™. http://www.lic.co.nz/lic_premier_sires.cfm?lid=28 [accessed 2 May 2016].

Nielsen R, Paul JS, Albrechtsen A, Song YS 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443–451.

Perreault LPL, Provost S, Legault MA, Barhdadi A, Dubé MP 2013. pyGenClean: efficient tool for genetic data clean up before association testing. Bioinformatics 29: 1704-1705.

Prakash B, Balain DS, Lathwal SS, Malik RK 1995. Infertility associated with monosomy-X in a crossbred cattle heifer. Veterinary Record 137: 436–437.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575.

Romano JE, Raussdepp T, Mulon PY, Villadóniga GB 2015. Non-mosaic monosomy 59, X in cattle: A case report. Animal Reproduction Science 156: 83–90.

Smith FWK, Buoen LC, Weber AF, Johnston SD, Randolph JF, Waters DJ 1989. X-chromosomal monosomy (77, XO) in a Doberman Pinscher with gonadal dysgenesis. Journal of Veterinary Internal Medicine 3: 90–95.

Weissensteiner H, Haun M, Schönherr S, Neuner M, Forer L, Specht G, Kloss-Brandstätter A, Kronenberg F, Coassin S 2013. SNPflow: a lightweight application for the processing, storing and automatic quality checking of genotyping assays. PloS one 8: e59508.

Zartman DL, Hinesley LL, Gnatkowski MW 1981. A 53, X female sheep (Ovis aries). Cytogenetic and Genome Research 30: 54–58.