Impact of SNP calling quality on the detection of transmission ratio distortion in goats

María Gracia Luigi-Sierra[1], Joaquim Casellas[2]*, Amparo Martínez[3], Juan Vicente Delgado[3], Javier Fernández Álvarez[3], Francesc Xavier Such[4], Jordi Jordana[2], Marcel Amills[1,2]*

[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, Bellaterra, 08193, Spain. [2]Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [3]Departamento de Genética, Universidad de Córdoba, Córdoba 14071, Spain. [4]Group of Research in Ruminants (G2R), Department of Animal and Food Science, Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona, Spain.

*Corresponding authors: Marcel Amills: marcel.amills@uab.cat, Joaquim Casellas: joaquim.casellas@uab.cat.

SUMMARY

Transmission ratio distortion (TRD) is the preferential transmission of one specific allele to offspring at the expense of the other one. The existence of TRD is mostly explained by the segregation of genetic variants with deleterious effects on the developmental processes that go from the formation of gametes to fecundation and birth. A few years ago, a statistical methodology was implemented in order to detect TRD signals on a genome-wide scale as a first step to uncover the biological basis of TRD and reproductive success in domestic species. In the current work, we have analyzed the impact of SNP calling quality on the detection of TRD signals in a population of Murciano-Granadina goats. Seventeen bucks and their offspring (N=288) were typed with the Goat SNP50 BeadChip, while the genotypes of the dams were lacking. Performance of a genome-wide scan revealed the existence of 36 SNPs showing significant evidence of TRD. When we calculated GenTrain scores for each one of the SNPs, we observed that 25 SNPs showed scores below 0.8. The allele frequencies of these SNPs in the offspring were not correlated with the allele frequencies estimated in the dams with statistical methods, thus evidencing that flawed SNP calling quality might lead to the detection of spurious TRD signals. We conclude that, when performing TRD scans, the GenTrain scores of markers should be taken into account to discriminate SNPs that are truly under TRD from those yielding spurious signals due to technical problems.

KEYWORDS: Transmission ratio distortion, GenTrain score, genotyping, Mendelian segregation.

INTRODUCTION

In diploid organisms, allelic transmission from parents to offspring is expected to follow the Mendelian law of inheritance, implying that both paternal and maternal alleles are transmitted to the progeny following approximately a 1:1 ratio (Huang et al., 2013). A deviation from this ratio, the so-called transmission ratio distortion (TRD), is produced when either the paternal or the maternal allele is preferentially transmitted to the offspring (Fishman and Mcintosh, 2019). A few years ago, Casellas et al. (2014) implemented a new Bayesian methodology to scan TRD for biallelic markers in diploid organisms. This method was later refined by Vázquez-Gómez et al. (2020) to detect TRD even in pedigrees with incomplete trios. This can be achieved by inferring the probability that specific alleles are present in the parent with a missing genotype based on the allele frequencies of the SNP in the overall population. The accuracy of such inference might be substantially affected by the quality of SNP genotypes. In previous studies (Abdalla et al., 2020; Casellas et al., 2017, 2020 Gòdia et al., 2020; Lahoucine et al., 2020; Vázquez-Gómez et al., 2020), the filtering of SNPs used in TRD scans relied fundamentally on two parameters (genotype call rate and minimum allele frequency or MAF). Deviation from HWE has never been used for this purpose because SNPs showing TRD are expected to display significant departures from HWE. However, it should be noticed that SNPs displaying significant HWE deviations often do so because of genotyping errors (Hosking et al., 2004). In other words, HWE filtering eliminates many unreliable markers which might yield spurious TRD signals just because of technical problems.

The goal of the current work was to make a TRD scan with incomplete trios and then to assess the calling quality of the SNPs that are putatively under TRD by using the

GenTrain score implemented in the GenomeStudio software from Illumina (Zhao et al., 2018). The GenTrain score is a metric that fluctuates between 0 (very poor calling quality) and 1 (excellent calling quality) and indicates the reliability of SNP detection based on the distribution of genotypic classes (Pavy et al., 2008; Zhao et al., 2018). By doing so, we aimed to assess the usefulness of the GenTrain score as a complementary metric to be considered in TRD scans as well as to recommend a specific GenTrain score filtering threshold to researchers interested in the detection of TRD.

MATERIALS AND METHODS

Sampling and genotyping of Murciano-Granadina goats

As animal material, we have collected blood samples from 17 bucks and their offspring (N=288) in vacuum tubes with K_3EDTA. These samples have been subsequently stored at -20ºC. Since blood collection is a routine procedure performed by CAPRIGRAN, no approval by the Ethics Committee on Animal and Human Experimentation of the Universitat Autònoma de Barcelona was required to perform this experiment. Information about the number of offspring per sire is depicted in Table S1. Genomic DNA extractions were performed following the modified salting out procedure described by Guan et al. (2020). Animals were genotyped with the Illumina Goat SNP50 BeadChip (Illumina Inc., San Diego, CA), which contains 54,241 SNP, following the instructions of the manufacturer. Genotypic data were updated with PLINK 1.9 (Chang et al., 2015) based on the *Capra hircus* genome ARS1 assembly (Bickhart et al., 2017) and the annotation provided by the International Goat Genome Consortium (http://www.goatgenome.org/projects.html#50K_snp_chip). Genotypes

were pruned using PLINK 1.9 (Chang et al., 2015). We selected SNPs fulfilling the following criteria: (1) genotype call rate over 95%, (2) minor allele frequency above 0.05 and (3) no missing genotypes in any of the 17 sires. Besides, the percentage of sires heterozygous for each SNP was estimated from the output obtained with the --hwe command of PLINK 1.9 (Chang et al., 2015) in order to remove SNPs with less than 20% of heterozygosity in the sire population (only SNPs with heterozygous genotypes are informative). The reference allele in this subset of the population was set as the most common allele in all individuals.

Statistical methods

To estimate TRD, we used a frequentist modification (Vázquez-Gómez et al., 2020) of the Bayesian method implemented by Casellas et al. (2014). Assuming two alleles (A1 and A2) and the existence of genotyped heterozygous sires and of ungenotyped dams, this method allows to compute for every marker an α-value which ranges between −0.5 (the A1 allele is not transmitted) and 0.5 (the A2 allele is not transmitted) thus providing an estimate of the magnitude of TRD. Allele frequencies in the ungenotyped dams were inferred by calculating a π-parameter which varies from 0 to 1. The two α and π parameters were estimated by maximizing the likelihood function and the statistical significance of α was assessed by using a likelihood ratio test (Nelson, 2008). A correction for multiple testing was applied to the *P*-values obtained from the χ^2 distribution using the false discovery rate approach (FDR) reported by Benjamini & Hochberg (1995) to obtain the corresponding *q-values*. Markers with α-values above 0.15 or below -0.15 and *q-values* < 0.05 were considered to show significant TRD.

After quality control based on genotype call rate and MAF, the final genotypic data included 42,272 autosomal SNPs.

RESULTS AND DISCUSSION

The implementation of the TRD test allowed us to identify 2,944 SNPs that were deviating from the Mendelian ratio (α-value > 0.15 or < -0.15). The highest α-value detected in our population was 0.499, while the lowest was -0.428; but for the majority of the genotyped SNPs, α-values were comprised in the [-0.15 to 0.15] interval which indicates the absence of TRD (Fig. 1a). After applying a likelihood ratio test and FDR correction, 36 SNPs were selected as significant (*q-value* < 0.05), from which 15 SNPs had an α-value below -0.15 implying a major transmission of the alternative allele, while the remaining 21 SNPs showed an overtransmission (α-value over 0.15) of the reference allele (Fig. 1b).

In order to verify the accuracy of the genotyping of the 36 SNPs displaying significant TRD, we calculated their GenTrain scores with the GenomeStudio software (Illumina Inc., San Diego, CA). Clustering of SNPs with different GenTrain scores and distribution of the GenTrain scores of the genotyped SNPs are depicted in Fig. 1c and Fig. S1. In Table 1, it can be seen that 25 of these SNPs have GenTrain scores below 0.80, with values ranging from 0.16-0.63 and an average score of 0.51 ± 1.14 (**Group 1**). In contrast, eleven SNPs have GenTrain scores above such threshold, with an average value of 0.87 ± 0.04 (**Group 2**). For each of these two groups of SNPs, we have calculated the correlation between allele frequencies of the SNP in the offspring vs allele frequencies inferred for the ungenotyped dams with the methods reported by Vázquez-Gómez et al. (2020). In principle, allele frequencies of parents and their

6

offspring should be significantly and positively correlated. In the **Group 1** of SNPs such correlation was very weak and non-significant (r = -0.007, *P*-value = 0.9733). Even worse, when we retrieved from **Group 1** seventeen SNPs with GenTrain scores between 0.5-0.6, the correlation was -0.0932 (*P*-value 0.7217). In strong contrast, allele frequencies of mothers and offspring were highly correlated in the **Group 2** of SNPs (r = 0.8656, *P*-value = 0.0005). This result implies that the method implemented by Casellas et al. (2014) and subsequently modified by Vázquez-Gómez et al. (2020) works very well in reconstructing allele frequencies in parental individuals without genotypes when SNPs have high GenTrain scores (> 0.80 in our study), which are the vast majority (Fig. 1c).

Previous reports have indicated that SNPs with minimum GenTrain scores of 0.25 can be safely used in most analyses (Pavy et al., 2008). Indeed, Pavy et al. (2008) designed highly-multiplexed SNP arrays for the genotyping of black and white spruce and reported that SNPs with GenTrain scores of 0.25 are very reliable and have a low rate of missing data. According to Pavy et al. (2008), SNPs with GenTrain scores of 0.25 or more had average call rates above 99%, and for SNPs with GenTrain scores above 0.4 the rate of missing data became negligible and the average rate of missing data per successful SNP was very low, On the other hand, Guo et al. (2014) indicated that SNPs with GenTrain scores above 0.70 are correctly clustered in most cases, while the clustering of SNPs with GenTrain scores below 0.7 might be more problematic. Based on our results, we conclude that the performance of TRD scans, especially in the case in which full trios are not available, should rely on the establishment of a stringent threshold for SNP calling quality. In our study, family size was small (on average 17-18 offspring per sire) and the density of the Goat SNP50 BeadChip (Illumina) is modest. In these conditions, we advise to consider markers with GenTrain scores of 0.80 or higher.

In studies with a more optimal design (particularly regarding family size), such threshold could be determined empirically by analyzing the correlations between the allele frequencies estimated in the parental class without genotypes and those inferred experimentally in the offspring, which should be significant and positive. This simple approach should facilitate the elimination of spurious TRD signals produced by technical factors in order to concentrate efforts on those that have biological implications.

ACKNOWLEDGMENTS

AVAILABILITY OF DATA

Genotypes of the 305 Murciano-Granadina sires and offspring and pedigree information are available in 10.6084/m9.figshare.14686230.

REFERENCES

Abdalla E.A., Id-Lahoucine S., Cánovas A., Casellas J., Schenkel F.S., Wood B.J. & Baes C.F. (2020). Discovering lethal alleles across the turkey genome using a transmission ratio distortion approach. *Animal Genetics* **51**, 876–89. https://doi.org/10.1111/age.13003

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300. https://doi.org/10.2307/2346101

Bickhart D.M., Rosen B.D., Koren S., Sayre B.L., Hastie A.R., Chan S., Lee J., Lam E.T., Liachko I., Sullivan S.T., Burton J.N., Huson H.J., Nystrom J.C., Kelley C.M., Hutchison J.L., Zhou Y., Sun J., Crisà A., Ponce De León F.A., Schwartz J.C., Hammond J.A., Waldbieser G.C., Schroeder S.G., Liu G.E., Dunham M.J., Shendure J., Sonstegard T.S., Phillippy A.M., Van Tassell C.P. & Smith T.P.L. (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics* **49**, 643–50. https://doi.org/10.1038/ng.3802.

Casellas J., Manunza A., Mercade, A., Quintanilla R. & Amills, M. (2014) A flexible bayesian model for testing for transmission ratio distortion. *Genetics* **198**, 1357-67. https://doi.org/10.1534/genetics.114.169607

Casellas J., Cañas-Álvarez J.J., González-Rodríguez A., Puig-Oliveras A., Fina M., Piedrafita J., Molina A., Díaz C., Baró J.A. & Varona L. (2017) Bayesian analysis of parent-specific transmission ratio distortion in seven Spanish beef cattle breeds. *Animal Genetics* **48**, 93-96 . https://doi.org/10.1111/age.12509

Casellas J., Id-Lahoucine S. & Cánovas A. (2020) Discriminating between allele- and

9

genotype-specific transmission ratio distortion. *Animal Genetics* **51**, 847-54.
https://doi.org/10.1111/age.13007

Chang C.C., Chow C.C., Tellier L.C.A.M., Vattikuti S., Purcell S.M. & Lee J.J. (2015)
Second-generation PLINK: Rising to the challenge of larger and richer datasets.
Gigascience **4**, 7. https://doi.org/10.1186/s13742-015-0047-8

Fishman L. & Mcintosh M. (2019) Standard deviations: The biological bases of
transmission ratio distortion. *Annual Review of Genetics* **53**, 347–72.
https://doi.org/10.1146/annurev-genet-112618-043905

Gòdia M., Casellas J., Ruiz-Herrera A., Rodríguez-Gil J.E., Castelló A., Sánchez A. &
Clop A. (2020) Whole genome sequencing identifies allelic ratio distortion in
sperm involving genes related to spermatogenesis in a swine model. *DNA
Research* **27**, dsaa019. https://doi.org/10.1093/dnares/dsaa019

Guan D., Landi V., Luigi-Sierra M.G., Delgado J.V., Such X., Castelló A., Cabrera B.,
Mármol-Sánchez E., Fernández-Alvarez J., de la Torre Casañas J.L.R., Martínez
A., Jordana J. & Amills M. (2020) Analyzing the genomic and transcriptomic
architecture of milk traits in Murciano-Granadina goats. *Journal of Animal
Science and Biotechnology* **11**, 35. https://doi.org/10.1186/s40104-020-00435-4

Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y. & Long J.
(2014) Illumina human exome genotyping array clustering and quality control.
Nature Protocols **9**, 2643-62. https://doi.org/10.1038/nprot.2014.174

Hosking L., Lumsden S., Lewis K., Yeo A., McCarthy L., Bansal A., Riley J., Purvis I.
& Xu C.F. (2004) Detection of genotyping errors by Hardy-Weinberg
equilibrium testing. *European Journal of Human Genetics* **12**, 395-9.
https://doi.org/10.1038/sj.ejhg.5201164

Huang L.O., A. Labbe, & C. Infante-Rivard (2013) Transmission ratio distortion:

Review of concept and implications for genetic association studies. *Human Genetics* **132**, 245–63. https://doi:10.1007/s00439-012-1257-0.

Lahoucine S.I., Casellas J., Lu D., Sargolzaei M., Miller S.P. & Canovas A. (2020) Distortion of Mendelian segregation across Angus cattle genome reveal novel lethal haplotype affecting reproduction. *Journal of Animal Science* **98**, 5. https://doi.org/10.1093/jas/skaa278.009

Nelson W.A. (2008) Statistical Methods. In: *Encyclopedia of Ecology* (ed. by S.E.Jørgensen & B.D. Fath), pp. 3350–62. Academic Press, Oxford. https://doi.org/https://doi.org/10.1016/B978-008045405-4.00661-3

Pavy N., Pelgas B., Beauseigle S., Blais S., Gagnon F., Gosselin I., Lamothe M., Isabel N. & Bousquet J. (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: Application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics* **9**, 21. https://doi.org/10.1186/1471-2164-9-21

Vázquez-Gómez M., de Hijas-Villalba M.M., Varona L., Ibañez-Escriche N., Rosas J.P., Negro S., Noguera J.L. & Casellas, J. (2020) Maternal transmission ratio distortion in two iberian pig varieties. *Genes* **11**, 1050. https://doi.org/10.3390/genes11091050

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**, 765-75. https://doi.org/10.1093/bib/bbx012

Table 1. Single nucleotide polymorphisms displaying transmission ratio distortion (α-value above 0.15 or below -0.15, *q*-value < 0.05) in a

population composed by 17 families of Murciano-Granadina goats.

Chr[1]	POS[2]	rs[3]	A1[4]	A2[5]	GT[6]	FREQ[7]	FREQ_off[8]	α-value[9]	*P-value*[10]	*q*-value[11]
1	40908451	rs268237959	A	G	0.350	0.983	0.525	-0.428	1.62E-51	2.29E-47
1	41022348	rs268246840	C	A	0.315	0.096	0.543	0.491	3.13E-51	3.31E-47
1	151646075	rs268244954	G	A	0.570	0.352	0.608	0.364	7.66E-06	1.09E-02
2	108668880	rs268260553	A	G	0.827	0.669	0.605	-0.281	3.07E-05	3.90E-02
2	130031994	rs268247158	A	C	0.897	0.791	0.800	0.313	2.54E-05	3.36E-02
3	27454000	rs268275899	G	A	0.448	0.051	0.525	0.499	6.61E-68	2.79E-63
4	7588781	rs268237958	A	G	0.566	0.152	0.575	0.499	4.79E-41	3.38E-37
4	31832045	rs268233915	G	A	0.905	0.645	0.703	0.313	7.71E-06	1.09E-02
4	44425927	rs268259874	G	A	0.925	0.706	0.802	0.383	6.67E-06	1.01E-02
5	55791623	rs268265557	G	A	0.591	0.807	0.577	-0.262	6.09E-07	1.03E-03
5	73383101	rs268233427	C	A	0.498	0.139	0.546	0.455	3.75E-32	1.76E-28
5	101041288	rs268263725	G	A	0.626	0.272	0.636	0.499	6.06E-24	2.14E-20

7	65607106	rs268289107	A	G	0.823	0.378	0.579	0.313	3.88E-07	7.13E-04
8	12216924	rs268259462	G	A	0.827	0.493	0.589	0.289	4.11E-05	4.83E-02
10	78039665	rs268257968	G	A	0.572	0.259	0.608	0.451	1.99E-18	6.00E-15
12	47232738	rs268244023	A	G	0.556	0.893	0.590	-0.322	9.33E-16	2.47E-12
13	63273673	rs268286787	A	G	0.634	0.864	0.680	-0.234	5.68E-07	1.00E-03
13	63340917	rs268286785	A	C	0.543	0.824	0.612	-0.294	3.92E-09	7.90E-06
15	34972358	rs268249438	A	G	0.606	0.883	0.580	-0.415	6.34E-25	2.44E-21
16	25523264	rs268249901	G	A	0.474	0.314	0.649	0.47	1.67E-16	4.70E-13
17	56855465	rs268264837	A	G	0.568	0.869	0.510	-0.402	3.35E-23	1.09E-19
18	9175705	rs268237492	A	G	0.840	0.426	0.566	0.284	1.21E-05	1.65E-02
18	16733418	rs268267817	A	G	0.859	0.371	0.516	0.244	3.14E-05	3.90E-02
18	40369501	rs268242498	G	A	0.576	0.184	0.593	0.499	2.33E-34	1.23E-30
19	51486084	rs268292182	A	G	0.160	0.061	0.528	0.491	7.76E-59	1.64E-54
19	56025239	rs268255767	C	A	0.303	0.981	0.546	-0.386	7.48E-40	4.52E-36
20	44223928	rs268278017	G	A	0.885	0.493	0.540	0.316	4.24E-06	6.89E-03

Chr	POS	rs	A1	A2	GT	FREQ	FREQ_off	α-value	P-value	q-value
22	39194121	rs268248059	A	C	0.844	0.778	0.746	-0.236	3.56E-05	4.30E-02
23	11194999	rs268271701	A	G	0.904	0.558	0.705	0.299	6.63E-06	1.01E-02
23	20456303	rs268233718	G	A	0.623	0.826	0.520	-0.361	1.16E-13	2.72E-10
25	37037529	rs268259686	A	G	0.586	0.955	0.510	-0.428	6.57E-44	5.56E-40
26	35111415	rs268287592	G	A	0.173	0.931	0.518	-0.375	4.17E-29	1.76E-25
27	167943	rs268251856	A	G	0.584	0.797	0.686	-0.337	8.66E-09	1.66E-05
27	32499525	rs268276328	A	G	0.606	0.893	0.623	-0.322	2.55E-15	6.35E-12
29	36841752	rs268284775	A	G	0.622	0.466	0.732	0.499	2.28E-09	4.82E-06
29	46365476	rs268268604	G	A	0.526	0.191	0.525	0.313	3.50E-10	7.80E-07

[1]Chr, chromosome; [2]POS, chromosomal position in base pairs (bp); [3]rs, rs identifier of the SNP; [4]A1, major allele; [5]A2, minority allele; [6]GT, GenTrain scores of the SNPs; [7]FREQ, estimated frequency of A1 in the dam ungenotyped population; [8]FREQ_off, frequency of A1 in the offspring (N = 288); [9]α-value, see Methods for definition ; [10]P-value, nominal P-value; [11]q-value, P-value corrected for multiple testing with the false discovery rate approach.

LEGENDS TO FIGURES

Fig. 1a. Genome-wide detection of SNP markers that show evidence of transmission ratio distortion in a population comprising 17 sire-families of Murciano-Granadina goats. The α-value estimated for each SNP is plotted in the *y*-axis while the chromosomal locations of SNPs are indicated in the *x*-axis. **1b.** Manhattan plot indicating the statistical significance (y-axis), expressed as $-\log_{10}$ of the *q*-value, of the α- values calculated for 42,272 SNPs genotyped in a population comprising 17 sire-families of Murciano-Granadina goats. The chromosomal location of each SNP is indicated in the x-axis. The red line corresponds to the threshold of significance which corresponds to a *q*-value = 0.05 expressed in a $-\log_{10}$ scale. **1c.** GenTrain score distribution for 42,272 SNPs genotyped in 305 goats. It can be seen that the vast majority of SNPs have GenTrain scores above 0.70.