RUNNING HEAD: TRD and GenTrain scores

Markers with low GenTrain scores can generate spurious signals in genome-wide scans for transmission ratio distortion

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.

Transmission ratio distortion (TRD) takes place when either the paternal or the maternal allele is preferentially transmitted to the offspring[1]. Methods have been implemented to detect TRD even in pedigrees with incomplete trios[2,3] by computing two parameters for every SNP marker: the 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, and the π-value which reflects allele frequencies in the ungenotyped parents. The accuracy of such inferences might be substantially affected by the quality of SNP genotypes. In this work, we aimed to assess the impact of SNP quality on the identification of TRD signals.

After quality control based on genotype call rate and MAF, a genome-wide scan for TRD based on 42,272 autosomal SNPs was carried out in a population of Murciano-Granadina goats composed by 17 sires and 288 offspring (see **Supplementary Materials and Methods** and **Table S1**). The performance of the TRD scan[3] (**Figure 1a**) allowed us to identify 36 SNPs showing a significant deviation (*q-value* < 0.05) from the Mendelian 1:1 ratio (**Figure 1b**). We calculated the GenTrain scores with the GenomeStudio software (Illumina Inc., San Diego, CA). In **Table S2**, 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 previously reported methods[3]. In principle, allele frequencies of parents and their 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). 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 methods performed to carry out the TRD scan[2,3] work 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 (**Figure 1c**). 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 (i.e. GenTrain scores > 0.80). 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

Many thanks to CAPRIGRAN for carrying out phenotype recording and blood sample collection in Murciano-Granadina goats. This research was funded by the European Regional Development Fund (FEDER)/Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación/Project Reference grant: PID2019-105805RB-I00 and by the Spanish Ministry of Economy and Competitivity (CEX2019-000902-S) Maria Luigi-Sierra was funded with a PhD fellowship "Formación de Personal Investigador" (BES-C-2017-079709) awarded by the Spanish Ministry of Economy and Competitivity.

Conflict of interest: Authors declare that they have no conflict of interest

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

References

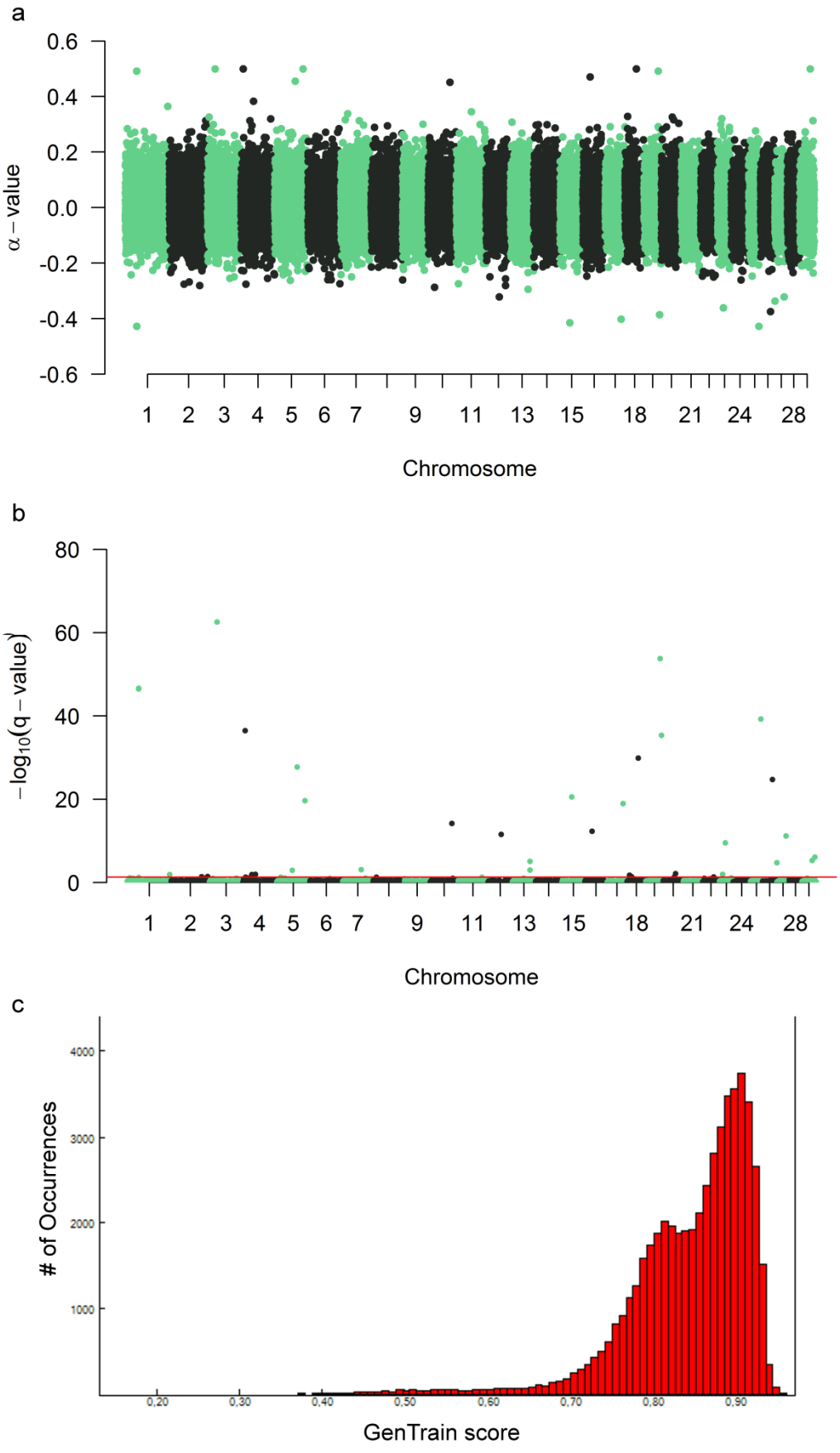
1 Fishman L. & Mcintosh M. (2019) *Annual Rev Genet* **53**, 347–72.

2 Casellas et al. (2014) *Genetics* **198**, 1357-67.

3 Vázquez-Gómez et al. (2020) *Genes* **11**, 1050.

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.


a
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
α – value
1 2 3 4 5 6 7 9 11 13 15 18 21 24 28
Chromosome
b
80
60
40
20
0
−log10(q – value)
1 2 3 4 5 6 7 9 11 13 15 18 21 24 28
Chromosome
c
4000
3000
2000
1000
# of Occurrences
0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90
GenTrain score

SUPPLEMENTARY MATERIALS

Supplementary Materials and Methods

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.

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 $\chi 2$ 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.

References

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

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

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

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

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

Table S1. Family size for each one of the 17 Murciano-Granadina sires.

ID sire	Number of daughters	Number of sons	Total Offspring
26137	7	3	10
47067	5	5	10
47227	5	7	12
49946	23	0	23
49952	21	0	21
56475	12	3	15
60401	13	11	24
148342	14	0	14
190384	24	0	24
190580	24	2	26
203008	11	0	11
203511	8	2	10
871651	10	0	10
871653	5	6	11
871674	27	3	30
871680	13	1	14
871681	11	12	23

Table S2. 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

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.