The use of whole genome sequence data to estimate genetic relationships including rare alleles information

S.E. Eynard*†‡, J.J. Windig*§, G. Leroy†‡, E. Verrier†‡, S.J. Hiemstra*§, R. van Binsbergen*#, M.P.L. Calus
*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands, †AgroParisTech, Paris, France, ‡INRA UMR 1313 GABI, Jouy-en-Josas, France, §Centre for Genetic Resources the Netherlands, Wageningen UR, Wageningen, the Netherlands, #Biometris, Wageningen UR, Wageningen, the Netherlands

ABSTRACT: Whole genome sequencing technologies are rapidly developing. In some ways, the speed of this development has outstripped our capacity to use this type of data in selection strategies, especially in livestock diversity conservation. In this study, relationship matrices were computed for 118 Holstein bulls, key ancestors of the current population, from three different types of data: pedigree records, 50K SNP chips and whole genome sequences, considering three different scenarios (with, without or only using rare alleles). Estimates from different data were highly correlated. Rare alleles had a significant impact on relationship estimates, mostly when whole genome sequence data were used. Hence sequence data, and information from rare alleles, are potentially of use for improving relationship computation. Estimation of relationships made with this type of data may result in different individual optimal contributions and influence selection strategies and conservation decisions of livestock species.
Keywords: Whole genome sequence; Additive genomic relationship; Genetic diversity

Introduction

The use of sequence data in animal breeding is expected to increase rapidly (Stock and Reent (2013)) due to technology improvement and reduction of costs. Compared to SNP chip data, whole genome sequence harbours more complete information on each individual that can be used to optimize breeding and conservation decisions. Breeding decisions rely partly on measures of relatedness between individuals in a population to select parents. Whole genome sequence data is expected to give "true relationship" values as it also includes information on rare alleles that are not accessible when using pedigree or common SNP chips. The main objective of this study was first of all to compare relationship estimates computed from pedigree, SNP chips or whole genome sequence data. A secondary objective was to infer the impact of including rare alleles, defined as having a 1 to 5% frequency (Druet et al. (2014)), on such estimated relationships, as they are expected to be keystones for livestock diversity conservation.

Materials and methods

Data. This study was performed on data from 118 European and North American Holstein bulls, selected as being key ancestors of the current population. Pedigree information was recorded since the 1950s (19 generations maximum) and contained 4054 individuals. The 118 bulls, born between 1968 and 2004, had both parents recorded in the pedigree and included 43 parent-offspring pairs, two full-sib pairs and 48 half-sib pairs. Whole genome sequence data for the selected bulls, including SNPs and insertion-deletion variants, were accessible through the 1000 bull genomes project (Run 3.0), and were for each individual obtained as described by Daetwyler et al. (2013 (submitted)). SNPs that are included in the commonly used Illumina BovineSNP50 version2 BeadChip (Illumina Inc., San Diego, CA) were selected from the whole genome sequence.

Markers with a frequency lower <1%, meaning that fewer than three copies of the minor allele were observed in the whole data set, were excluded from the analysis as they may represent genotyping errors. Given the small sample size used in this study, 1% seemed to be a sensible threshold to distinguish errors and rare alleles. Larger sample sizes might enable using lower thresholds. Thus, across the 29 autosomes 15,871,933 SNPs among the initial 18,739,233 polymorphic markers in those 118 Holstein bulls were kept for the whole genome sequence data and 44,548 out of 45,729 SNPs were used for the 50K SNP chip.

Relationship calculations. Pedigree (**A**) and genetic (**G**) relationship matrices were computed using the software calc_grm (Calus (2013)). The **G** matrix calculations were performed using the Yang et al. (2010) method: $\boldsymbol{G} = \frac{\boldsymbol{WW'}}{N}$, where N is the number of markers and $\boldsymbol{W}$ is the marker genotype matrix for all individuals, all loci. Each w_{ij} value was calculated as follows: $w_{ij} = \frac{x_{ij} - 2p_i}{\sqrt{2p_i * (1-p_i)}}$, with x_{ij} marker genotypes coded as 0, 1 and 2 for individual j locus i and p_i the allele frequency at marker i. Relationships were computed using SNPs and whole genome sequence data in three scenarios: (1) using all markers with Minor Allele Frequency (MAF) >1% (scenario ≥0.01); (2) all markers with MAF >5% only (scenario ≥0.05); and (3) using markers with MAF between 1 and 5% (scenario 0.01_0.05). After MAF selection 44,548, 41,225 and 3,323 SNPs were kept from 50K SNP chip data and 15,871,933, 11,953,905 and 3,918,028 from whole genome sequence data, in scenario ≥0.01, ≥0.05 and 0.01_0.05 respectively, for relationship computation.

Comparison of different relationship estimates. Estimated relationships using the three types of data were plotted against each other. Goodness of fit, measured as R^2, were estimated for diagonal and off-

diagonal elements. Differences between scenario ≥0.01 and ≥0.05 were tested, using Wilcoxon tests, and goodness of fit between scenarios using each type of data were estimated in order to infer the impact of rare alleles on estimated relationships.

Results

Relationship estimates. The relationship estimates from pedigree data (**A**) ranged from 0 to 1.16, the **G** elements from SNP data from -0.06 to 1.32 and from sequence data from -0.02 to 1.18, depending on the scenario. Mean values ranged from 0.1 to 1.02 for diagonal elements and were approximately 0.05 for off-diagonal elements (Table 1 and 2).

Table 1. Descriptive statistics of diagonal relationships, computed from pedigree (A), SNP (G_SNP) or sequence data (G_seq) from the three different scenarios: ≥0.01, ≥0.05 or 0.01_0.05.

	Minimum	Mean	Maximum	SD
A	1.000	1.027	1.163	0.031
G_SNP 0.01	0.850	0.994	1.325	0.086
G_SNP 0.05	0.812	0.924	1.110	0.059
G_SNP 0.01_0.05	0.069	0.124	0.270	0.033
G_seq 0.01	0.696	0.856	1.185	0.093
G_seq 0.05	0.585	0.658	0.772	0.034
G_seq 0.01_0.05	0.145	0.252	0.512	0.068

Table 2. Descriptive statistics of off-diagonal relationships, computed from pedigree (A), SNP (G_SNP) or sequence data (G_seq) from the three different scenarios: ≥0.01, ≥0.05 or 0.01_0.05.

	Minimum	Mean	Maximum	SD
A	0.000	0.063	0.663	0.067
G_SNP 0.01	-0.065	0.046	0.639	0.061
G_SNP 0.05	-0.064	0.047	0.583	0.059
G_SNP 0.01_0.05	0.045	0.053	0.128	0.004
G_seq 0.01	-0.022	0.050	0.501	0.045
G_seq 0.05	-0.017	0.050	0.398	0.038
G_seq 0.01_0.05	0.039	0.054	0.195	0.008

Correlation between types of estimated relationships. Diagonal elements contain information on within individual relationships only, and did not show significant non zero correlations between different types of data. Therefore, we focused further analysis on off-diagonal elements. Both plots of relationship (Figure 1) and goodness of fit estimates (Table 3) consistently show strong correlation between the three types of data. Pedigree relationships were more closely correlated to SNP than to whole genome sequence relationships, especially for the scenario 0.01_0.05. Estimated relationships in the **G** matrix had smaller (absolute) values than in the **A** matrix. Relationships for whole genome sequence and 50K SNP chip had high correlations for scenario ≥0.01 and ≥0.05 (R^2 0.969 and 0.981), and somewhat lower for scenario 0.01_0.05 (0.675). In the comparative analysis, for both types of data, scenario ≥0.01 gave higher relationship values than scenario ≥0.05 and estimates from sequence data were smaller than estimates from SNPs, except in scenario 0.01_0.05.

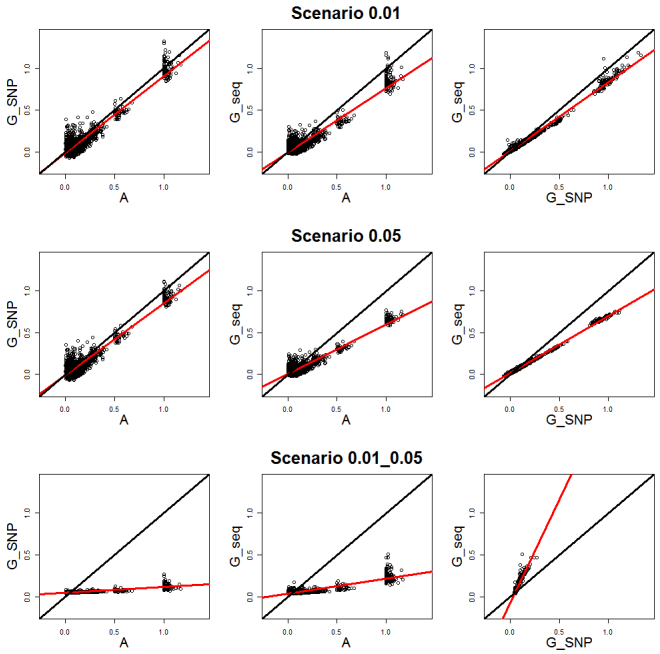

Figure 1. Comparative plots of linear regression of A relationship estimates, G relationship estimates for SNP and sequence data against each other in scenarios ≥0.01, ≥0.05 and 0.01_0.05. In black is the regression line for an exact linear model (intercept=0, slope=1), in red is the actual regression line for all elements.

Effect of MAF on relationship estimation. Comparative tests showed a significant difference between the scenarios ≥0.01 and ≥0.05. Wilcoxon test performed on relationship estimates from the 50K SNP chip or whole genome sequence in the scenario 0.01_0.05 showed significant differences from 0. The R^2 between same type of data for scenario ≥0.01 against ≥0.05 was 0.999 for SNP and 0.988 for sequence data. In Table 3, R^2 between SNP and sequence data was higher in the ≥0.05 scenario than in ≥0.01, so relationships computed from 50K SNPs explain less of the variation in relationships computed with whole genome sequence when rare alleles are included for computation (scenario ≥0.01). Slopes of the linear regression of scenario ≥0.01 on scenario ≥0.05 were higher than 1, indicating that estimates from scenario ≥0.01 have a higher variance than for scenario ≥0.05.

Discussion

Relationship estimates from SNPs and whole genome sequences differed from pedigree, as expected (Simeone et al. (2011)), even though a strong correlation between the SNP and pedigree relationships was observed, similar to the results shown in Rolf et al. (2010). From comparison of scenarios with different sets of loci we concluded that rare alleles have a significant impact on estimated relationships, predominantly when using whole genome sequence. Estimated genomic relationships were lower than those based on pedigree. Hypotheses to explain this difference are: (i) the subset of markers used in the different scenarios showed departure

Table 3. Numerical comparison between the three types of data in scenario ≥0.01, ≥0.05 and 0.01_0.05. R^2 (below diagonal) and linear regression slope (above diagonal), for diagonal and off-diagonal elements separately.

	A	G_SNP0.01	G_SNP0.05	G_SNP0.01_0.05	G_seq0.01	G_seq0.05	G_seq0.01_0.05
Diagonal							
A	-	0.039	0.303	-0.264	-0.264	0.241	-0.504
G_SNP0.01	0.000	-	**1.398**	**2.291**	**0.918**		
G_SNP0.05	0.024	0.929	-	**1.291**		**0.535**	
G_SNP0.01_0.05	0.060	0.770	0.515	-			**1.685**
G_seq0.01	0.007	0.720			-	**2.322**	**1.326**
G_seq0.05	0.048		0.894		0.701	-	**0.326**
G_seq0.01_0.05	0.052			0.675	0.926	0.432	-
Off-diagonal							
A	-	**0.721**	**0.689**	**0.047**	**0.524**	**0.446**	**0.078**
G_SNP0.01	0.616	-	**1.043**	**12.080**	**0.720**		
G_SNP0.05	0.612	0.999	-	**11.261**		**0.646**	
G_SNP0.01_0.05	0.613	0.634	0.600	-			**1.655**
G_seq0.01	0.608	0.969			-	**1.163**	**4.626**
G_seq0.05	0.604		0.981		0.987	-	**3.628**
G_seq0.01_0.05	0.413			0.675	0.704	0.594	-

Values significantly different from 0 and 1 for slope are in bold.

from Hardy-Weinberg proportions, (ii) differences in the base population used in **A** and **G** matrix computation (Li et al. (2011)) which results in scaling differences in inbreeding estimates for individuals (Forni et al. (2011)) or, (iii) the use of IBS instead of IBD (Engelsma et al. (2012), Makgahlela et al. (2013)) to compute relationships. Departure from Hardy-Weinberg proportion is the most plausible explanation, especially for the extremely low relationships of scenario 0.01_0.05, and will be further investigated. This study focused on the use of different sets of MAF loci to compute relationship estimates. The Yang et al. (2010) method uses frequency for each locus independently, and is expected to give a more appropriate weighing of the markers compared to the Van Raden (2008) method for our purpose.

Differences in relationship estimates due to inclusion of rare alleles, as observed in this study, can lead to selection of different individuals and change optimal contribution values of potentially selected bulls for breeding and genetic diversity conservation purposes. Information on rare alleles is of major interest for long-term genetic improvement and management of livestock genetic diversity. Including rare alleles for optimal contribution decisions in livestock species is expected to help safeguard available genetic diversity for future genetic improvement.

Conclusion

Ignoring rare alleles, when using pedigree or SNP chip data instead of whole genome sequence, had a significant impact on relationship estimates. Although strongly correlated, pedigree, SNP and whole genome sequence relationship estimates showed differences when rare alleles were taken into consideration. Whole genome sequence, as it is the only type of data providing information on rare alleles, will be valuable for improved relationship computation for long term genetic diversity conservation purposes. Differences between relationship estimates computed including or not rare alleles is expected to affect selection decisions, long term selection strategy and conservation of livestock species.

Acknowledgement

SE benefited from a grant from the European Commission, within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG", co-funded by the Dutch Ministry of economic Affairs (KB-12-005-03-001). The authors thank the 1000 Bull genomes consortium for providing the sequence data.

Literature cited

Calus M. P. L. 2013. ABGC (ed.). Wageningen UR Livestock Research.

Daetwyler H. D., Capitan A., Pausch H., et al. 2013 Nat genetics (submitted).

Druet T., Macleod I. M. & Hayes B. J. 2014. Heredity, 112**,** 39-47.

Engelsma K. A., Veerkamp R. F., Calus M. P. L., et al. 2012. J. Animal Breeding and Genet, 129**,** 195-205.

Forni S., Aguilar I. & Misztal I. 2011. Genet Sel Evol, 43.

Li M. H., Stranden I., Tiirikka T., et al. 2011. Plos One, 6.

Makgahlela M. L., Stranden I., Nielsen U. S., et al. 2013. J Dairy Sci, 96**,** 5364-5375.

Rolf M. M., Taylor J. F., Schnabel R. D., et al. 2010. BMC Genet, 11.

Simeone R., Misztal I., Aguilar I., et al 2011. J Animal Breeding and Genet, 128**,** 386-393.

Stock K. F. & Reents R. 2013. Reprod in Domestic Animals, 48**,** 2-10.

VanRaden P. M. 2008. J Dairy Sci, 91**,** 4414-4423.

Yang J. A., Benyamin B., McEvoy B. P., et al. 2010. Nat Genetics, 42**,** 565-U131.