Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals

H. Wang, I. Misztal, Andres Legarra

HAL Id: hal-02630962
https://hal.inrae.fr/hal-02630962

ORIGINAL ARTICLE

Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals

H. Wang[1], I. Misztal[2] & A. Legarra[3]

1 Genus Plc, Hendersonville, TN, USA
2 Department of Animal and Dairy Science, University of Georgia, Athens, GA, USA
3 INRA, UR 631 SAGA, Castanet Tolosan, France

Summary

This work studied differences between expected (calculated from pedigree) and realized (genomic, from markers) relationships in a real population, the influence of quality control on these differences, and their fit to current theory. Data included 4940 pure line chickens across five generations genotyped for 57 636 SNP. Pedigrees (5762 animals) were available for the five generations, pedigree starting on the first one. Three levels of quality control were used. With no quality control, mean difference between realized and expected relationships for different type of relationships was ≤ 0.04 with standard deviation ≤ 0.10. With strong quality control (call rate ≥ 0.9, parent-progeny conflicts, minor allele frequency and use of only autosomal chromosomes), these numbers reduced to ≤ 0.02 and ≤ 0.04, respectively. While the maximum difference was 1.02 with the complete data, it was only 0.18 with the latest three generations of genotypes (but including all pedigrees). Variation of expected minus realized relationships agreed with theoretical developments and suggests an effective number of loci of 70 for this population. When the pedigree is complete and as deep as the genotypes, the standard deviation of difference between the expected and realized relationships is around 0.04, all categories confounded. Standard deviation of differences larger than 0.10 suggests bad quality control, mistakes in pedigree recording or genotype labelling, or insufficient depth of pedigree.

Introduction

Pedigree-based relationships in the numerator relationship matrix ($\mathbf{A}$), are measures of expected relationships (Wright 1934), and unrelated founders are usually assumed. The value assigned to such relationships depends on the depth and completeness of pedigree (Cole & Franke 2002; Cassell et al. 2003). For animals that are not connected by pedigree, the relationship value is zero, although methods exist for the case of incomplete pedigrees (VanRaden 1992; Lutaaya et al. 1999).

Whereas pedigree-based relationships are expectations of genome sharing of two individuals, actual (or realized) relationships in nature differ from that expectation because genome size is finite and loci are linked (VanRaden 2008; Hayes et al. 2009). These deviations are in the core of increased accuracy of genomic predictions (Hayes et al. 2009). For instance, at a single locus, two full sibs may share none, one or both alleles. The extent of deviations of realized from expected pedigree-based relationships has been estimated either considering linkage and pairs of individuals (Hill & Weir 2011), or considering a general

pedigree and one locus (Garcia-Cortes *et al.* 2013). In general, deviations are small, but the less related two individuals are the greater the coefficient of variation for their actual relationship.

The genomic relationship matrix (**G**) is an estimator of the actual proportion of genome that is identical by descent across individuals. The **G** matrix is estimated based on SNP genotypes and does not depend on pedigree (VanRaden 2008; Hayes *et al.* 2009). However, it is dependent on the estimator and particularly on the assumed gene frequency for each SNP (Ritland 1996; Toro *et al.* 2011). Ideally, allele frequencies for the base population are needed, but they are difficult to estimate (Habier *et al.* 2010; Forni *et al.* 2011). For lengthy or incomplete pedigrees, what the genetic base should be is also unclear in practice. Also, the scale of **G** is somewhat arbitrary, because it depends on the assumed genetic base and variance. For instance, adding constants to **G** only shifts the estimated mean of the model (Stranden & Christensen, 2011), and multiplying **G** by a constant can be counterbalanced diminishing the genetic variance. To use the same genetic base and variance as regular genetic evaluations, algorithms to create **G** include mechanisms to adjust **G** for compatibility with **A** (VanRaden 2008; Powell *et al.* 2010; VanRaden *et al.* 2011; Vitezica *et al.* 2011) or adjust **A** to match **G** (Christensen 2012). Those adjustments generally correct for equal average breeding value and equal genetic variance in the two base populations (genomic-based and pedigree-based). Even after adjustment, animals unrelated in **A** can still appear related in **G**, and this makes sense: base generation animals are related (they share alleles at genes), in spite of the assumption of unrelatedness.

If pedigree and genotypic information is correct, **G** and **A** are both close to actual relationships. Further, if the assumptions underlying the computation of both are right (**G** uses correctly marker information, and the base generation in **A** is drawn from an 'unrelated' large population mating at random), then $E(G) = A$ (VanRaden 2008; Hayes *et al.* 2009). Therefore, large differences between **G** and **A** would suggest pedigree or genotyping errors (Simeone *et al.* 2011). However, the extent and distribution of those differences are largely unknown, and therefore, a quality control threshold is difficult to establish.

One of the best populations to test differences between **A** and **G** is purebred commercial chicken because lines are closed. Differences in gene frequencies among lines are not a factor, and pedigrees are accurately recorded for many generations. Therefore, the base populations can be standardized to the same

point in time. The objectives of this study were to compare coefficients of **A** and **G** for purebred commercial chicken under various quality control levels, to investigate the distribution of such differences, to compare such distributions with theoretical predictions in Hill & Weir (2011) and Garcia-Cortes *et al.* (2013), and to suggest empirical thresholds for quality control checking.

Materials and methods

Genotypes from 4940 pure line chickens across five generations were obtained from Cobb-Vantress Inc. (Siloam Springs, AR, USA). A pedigree including these animals and their ancestors comprised 5762 individuals, always within the five generations. More generations were not included because individuals beyond these generations were *not* genotyped. Therefore, **A** and **G** can be compared on similar grounds. Genotyped individuals were roughly 150 males and 1100 females at the first two generations, and then 400 males and 400 females in generations 3–5.

Genotypes were determined using the Illumina 60K SNP BeadChip (Illumina Inc., San Diego, CA, USA) for chicken, which includes 57 636 single-nucleotide polymorphisms (SNPs) distributed over 34 chromosomes (Groenen *et al.* 2011). Three different quality control (QC) levels were applied to SNP data: none, weak and strong. The QC data edits (Wiggans *et al.* 2010) consisted in call rate (individual or marker-based; 0 for none, >0.7 for weak, >0.9 for strong), minor allele frequency (>0.05 for weak and strong, 12 261 markers removed), parent-progeny conflicts (<2% for individuals, <10% for markers, in both weak and strong), and exclusion of non-autosomal markers (in weak and strong QC, 4322 markers removed). A summary of removed data for each QC level is in Table 1. It can be observed that three animals showing parent-progeny conflict with weak QC were eliminated by call rate under strong QC. Parent-progeny conflicts were based on mendelian segregation, incompatible parent and progeny genotypes (i.e. alternative homozygotes) and were not based on divergences between numerator (**A**) and genomic (**G**) relationship matrices. Matrix **G** was computed for either all four generations, the last two generations, or the last generation. In this way, we can contemplate also the cases where pedigree goes a bit longer than genotypes.

The **A** and **G** for genotyped individuals were computed with algorithms by Aguilar *et al.* (2011). Initial G_b was computed as $G_b = ZZ'/2 \sum p_i q_i$, (VanRaden 2008) where **Z** contains centred genotypes. Allelic

QC level	QC edit for genotype	Chicken	SNP
None	–	0	0
Weak	Call rate of <0.7	5	1510
	Parent–progeny conflict rate of >2% for animal or >10% for SNP	219	1
	Minor allele frequency of <0.05		12 261
	Genotype from mitochondrial genome, unknown chromosomes or sex chromosome		4322
Strong	Call rate of <0.9	67	6095
	Parent–progeny conflict rate of >2% for animal or >10% for SNP	216	1
	Minor allele frequency of <0.05		12 261
	Genotype from mitochondrial genome, unknown chromosomes or sex chromosome		4322

frequencies were computed from available genotyped animals and not from base animals; the reason for doing this is that this is the current practice in most genetic evaluations, because it results in **G** with better numerical properties (Stranden & Christensen 2011) and also in **G** with well defined properties (namely, an average of 1 in the diagonal and 0 off-diagonal). Matrix **G** was scaled by default to be compatible with **A**, using $G = a + bG_b$ so that average relationships and inbreeding agree across **G** and **A** (Christensen 2012). This scaling gives very similar results to use base allelic frequencies (Powell *et al.* 2010; Vitezica *et al.* 2011; Christensen 2012). For each QC level, several measures were investigated to determine the magnitude and distribution of difference between **G** and **A** (**G** − **A**): diagonal and off-diagonal elements, parent-offspring pairs, full-sib pairs and half-sib pairs.

Expected standard deviation of **G-A** was obtained from Hill & Weir (2011), Table 1) for specific pairs of relationships (0 for parent-progeny, 0.25 for half-sibs, 0.35 for full sibs), and compared with observed values across these kinds of relationships. For all possible relationships, the accuracy of the predictions of Garcia-Cortes *et al.* (2013) was verified graphically. We recall that, if all assumptions hold, $E(G_{ij}) = A_{ij}$, and therefore $Var(G_{ij} − A_{ij}) = Var(G_{ij})$. The *a priori* expected variation of realized relationship for two individuals i and j ($Var(G_{ij} - A_{ij})$) was calculated using their formula $Var(G_{ij} − A_{ij}) = 4(\phi_{ij,ij} − \phi_{ij}^2)$, where $\phi_{ij,ij}$ is Karigl (1981) double-pair co-ancestry coefficient computed from pedigree and $\phi_{ij} = \frac{A_{ij}}{2}$ is the coancestry between individuals i and j. We computed $Var(G_{ij} − A_{ij})$ from pedigree across all 12 204 270 distinct relationship coefficients (including self-relationships). Because of the difficulty of comparing a deviation

with its predicted variance, 44 equidistant bins were created according to expected variance of relationship. The within-bin variance of **G** − **A** was plotted against the expected variance of realized relationships from Garcia-Cortes *et al.* (2013).

Results and discussion

Ranges, means and standard deviations for **G** − **A** (including all 4940 animals across five generations) are in Table 2 for diagonals, off-diagonals and parent-progeny, full-sib and half-sib pairs under no, weak and strong QC. The most noticeable change with increased QC was for diagonal elements (Figure 1). From no QC to strong QC, the range for differences decreased over three times while the standard deviation almost halved because of poor genotyping of some individuals (e.g. homozygous for all loci). For simulated data, Simeone *et al.* (2011) found that diagonal elements of **G** were close to one; with observed data, they found that elements that were very different than one were the result of poor QC or animals from a different line. This may be the case for the extreme positive outliers that can be clearly seen in the Figure 1.

The next most noticeable change for **G** − **A** statistics (Table 2) was for parent-progeny pairs. Introducing weak QC decreased the range for differences by approximately 0.25 and the standard deviation by 0.05. Little additional change was observed with strong QC. Changes in differences for full- and half-sibs were minimal as QC increased. Large changes for parent-progeny pairs but not for full- or half-sibs could be the result of QC differences: parent-progeny conflicts were detected (and conflicting data removed) based on Mendelian inheritance, whereas full-sib and half-sib conflicts were not. The parent-progeny conflicts were caused mostly by sample (or pedigree) misidentification, which resulted in the wrong parent-progeny pairs. Detecting conflicts for relationships other than parent-progeny is more difficult, although it can be done with large data sets (VanRaden *et al.* 2013).

A reason for differences between **G** and **A** is limited pedigree depth. Animals are assumed to be unrelated in the base generation of **A**. However, they are actually related, and those relationships have been reflected in **G** (Christensen 2012). Table 3 shows statistics for **G** − **A** under strong QC when **A** was based on all five generations and **G** was based on either all five generations, the last three generations, or the last generation. For diagonal coefficients, when **G** was based on the last three generations, the maximum difference between **G** and **A** was 0.84. This decreased to

Table 2 Statistics for coefficient differences between genomic (**G**) and numerator (**A**) relationship matrices for genotyped chickens

Quality control level	**G** − **A** coefficient measure	Number of animal pairs	Minimum	Maximum	Mean	Standard deviation
None	Diagonals	4940	−0.53	3.11	0.000	0.083
	Off-diagonals	12 199 330	−0.58	1.15	0.000	0.036
	Parent-progeny pairs	6115	−0.58	0.16	−0.040	0.094
	Full-sib pairs	9970	−0.19	0.18	−0.016	0.050
	Half-sib pairs	69 154	−0.18	0.16	−0.014	0.041
Weak[1]	Diagonals	4727	−0.52	0.88	0.000	0.062
	Off-diagonals	11 169 901	−0.57	1.03	0.000	0.036
	Parent-progeny pairs	5377	−0.32	0.17	−0.016	0.044
	Full-sib pairs	9130	−0.19	0.18	−0.017	0.050
	Half-sib pairs	59,930	−0.18	0.16	−0.015	0.040
Strong[2]	Diagonals	4667	−0.18	0.84	0.000	0.048
	Off-diagonals	10 888 111	−0.57	1.02	0.000	0.037
	Parent-progeny pairs	5259	−0.16	0.17	−0.011	0.034
	Full-sib pairs	9126	−0.19	0.18	−0.017	0.050
	Half-sib pairs	59 870	−0.18	0.16	−0.015	0.040

[1]Call rate of ≥ 0.7.
[2]call rate of ≥ 0.9.
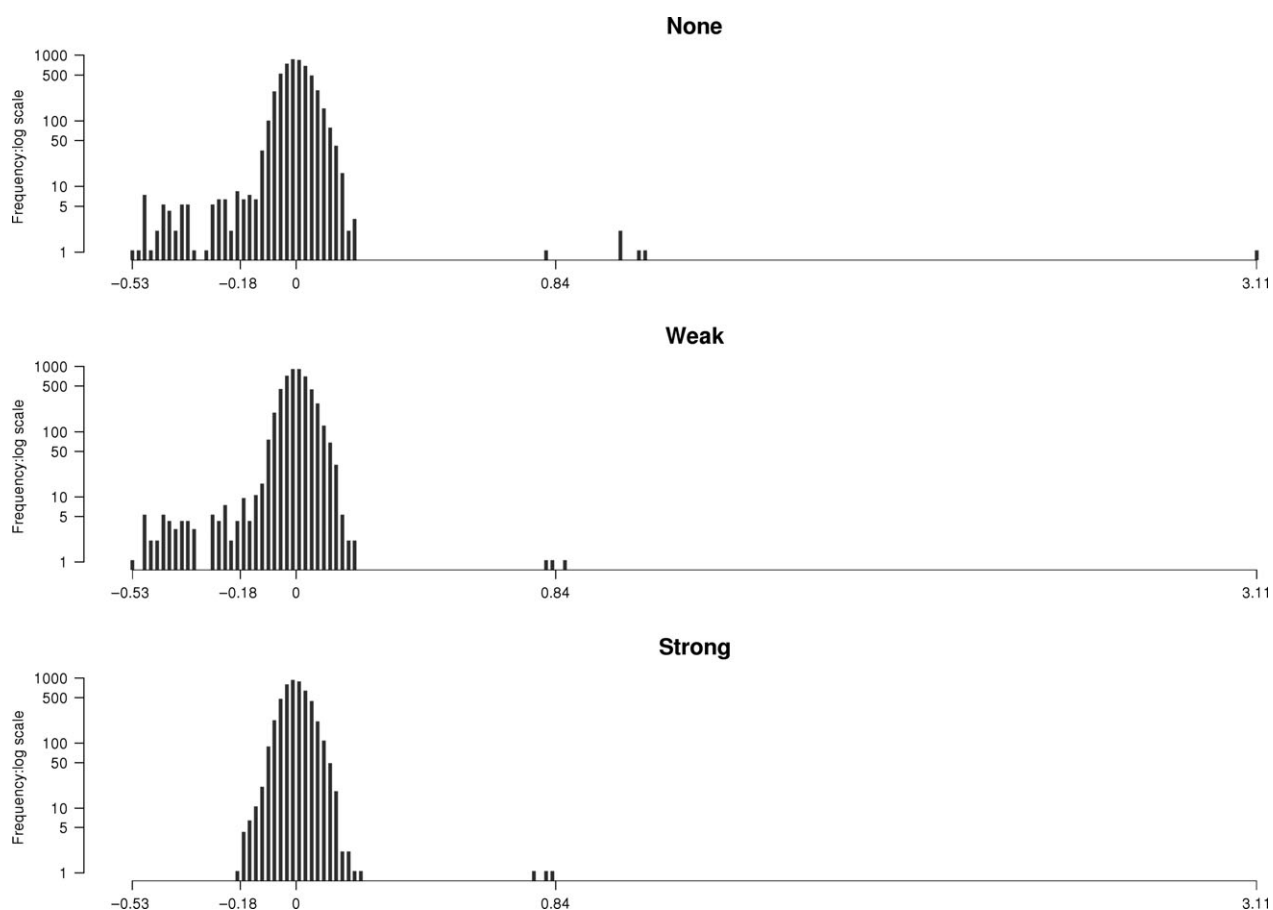
Figure 1 Distribution of diagonal differences between genomic (**G**) and numerator (**A**) relationship matrices for genotyped chickens. Differences (**G** − **A**) in diagonal coefficients were examined under no quality control (None) for genotypic data, weak quality control (Weak, call rate of ≥ 0.7), and strong quality control (Strong, call rate of ≥ 0.9). Frequencies (on *y*-axis) were in log scale to show bars with low counts.

Table 3 Statistics for generation-restricted coefficient differences between genomic (**G**) and numerator (**A**) relationship matrices for genotyped chickens

Generations included in G	**G** − **A** coefficient measure[1]	Number of animal pairs	Minimum	Maximum	Range	Standard deviation
All five	Diagonals	4667	−0.18	0.84	1.02	0.048
	Off-diagonals	10 888 111	−0.57	1.02	1.59	0.037
Last three	Diagonals	2304	−0.18	0.13	0.31	0.041
	Off-diagonals	2 653 056	−0.26	0.20	0.46	0.032
Last	Diagonals	772	−0.18	0.13	0.31	0.040
	Off-diagonals	297 606	−0.18	0.17	0.35	0.033

[1]Strong quality control (call rate ≥ 0.9).

0.18 when **G** was based on the last generation. Such change was probably the result of eliminating an error in the first two generations, an error that could not be corrected by QC. For instance, this could be a misla-belled animal with no ancestor genotyped; in this case, parent-offspring checking do not detect the error. For off-diagonal coefficients, the largest difference decreased from -0.57 when **G** included all generations to −0.26 for **G** based on the last three generations, and to -0.18 based on the last generation. When pedigrees start at the same time than genotypes, related animals (e.g. full sibs) in the base populations that are unaccounted for in **A** would cause a difference in **G** − **A** of approximately 0.5. This cannot be corrected for using base allele frequencies, because those frequencies would be computed as if full sibs were unrelated. With one generation of pedigrees beyond genotypes, all full sibs are accounted for in **A**, but cousins are not, which would cause differences in **G** − **A** of approximately 0.125. Each additional generation of pedigrees causes **A** to capture an additional level of relationships. However, unequal pedigree length, which is usual in ruminants (but not in chicken) increases discrepancies between **A** and **G** (Christensen 2012; Misztal *et al.* 2013).

Hill & Weir (2011) and Garcia-Cortes *et al.* (2013) provided formulas, shown above, for the standard deviation of realized relationships. These formulas assume unrelated founders. For a single locus, that would be 0 for parent-progeny pairs theoretically if no inbreeding existed. A possible explanation for the standard deviation of 0.034 for parent-progeny pairs under strong QC (Table 2) is that founders are rarely from an extremely large population for chicken lines (and most domestic species). Therefore, they cannot be considered as unrelated. All parent-progeny relationships will include some inbreeding because the dam will always be related to the sire in real closed populations. For example, if the dam and the sire are first-degree cousins, the standard deviation of the actual parent-progeny relationship is 0.16 instead of

zero, using expressions in Garcia-Cortes *et al.* (2013) that allow for complex pedigrees. Another possibly complementary explanation is noise from not using base allelic frequencies.

The half- and full-sib standard deviations of **G** − **A** (0.05 and 0.04, respectively) appear to fit theoretical results (0.35 and 0.25) by a scale factor of roughly 7, which indicates around 49 independent chromosomal segments (Goddard 2009). Half- and full-sib pairs appear to be less sensitive than parent-progeny pairs to unknown relationships of founders: the theoretical standard deviation of actual full-sib relationships considering the full pedigree (Garcia-Cortes *et al.* 2013) is 0.40 (closer to 0.35).

Figure 2 reflects the distribution of within-bin variance of **G** − **A** to its expected value. Each point represents the *observed variance* (on the *y*-axis) for the set of points that are in a bin centred on a certain *expected variance* (on the *x*-axis). This graph is a generalization of Table 2 for complex pedigrees, where there are all
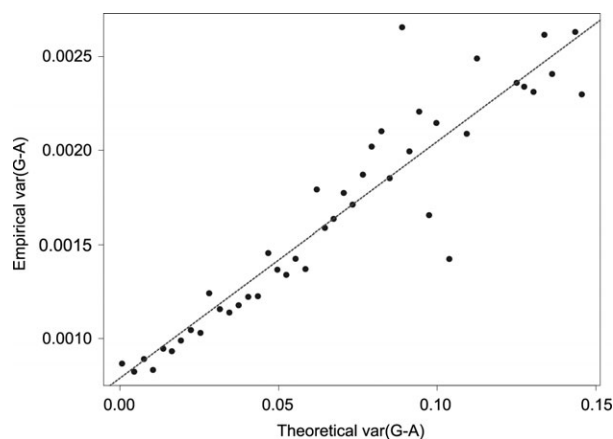
Figure 2 Plot of variance (on *y*-axis) of differences between genomic (**G**) and numerator (**A**) relationships for genotyped chicken versus theoretical values (on *x*-axis). The theoretical variance of **G** − **A** for a single locus was computed and divided in 44 bins. The empirical variance of **G** − **A** was computed within each bin. The correlation between theoretical and empirical var (**G** − **A**) was 0.93.

kind of relationships. For instance, there are many pairs of relationships where theoretically a variance of 0.05 is expected. The variance of computed **G–A** across these points is approximately 0.014. The figure shows good agreement with theory: the observed (empirical) variance of $\mathbf{G} - \mathbf{A}$ is directly proportional to the theoretical prediction. For instance, an increase value of approximately 0.10 in the theoretical variance of $\mathbf{G} - \mathbf{A}$ corresponds to an increase of approximately 0.014 in the observed variance of $\mathbf{G} - \mathbf{A}$. This relationship is constant (a straight line). Deviations from the line at high values were the result of data scarcity in extreme bins. A straight line fits the plot well (correlation of 0.93). The slope of the fitted line was 0.0143, which implies that the complexity of the genome is equal to that of 70 (i.e. 1/0.0143) equivalent independent loci. This value of 70 agrees well with the value of 49 estimated according to results in the previous paragraph, which looked at half-sibs and full-sibs pairs.

The results of this study reflect several layers of incertitude in the computation of $\mathbf{G} - \mathbf{A}$: first associated with genotyping mistakes, misidentifications, and random drift of allelic frequencies, and second associated with pedigree errors or incompletion. The issue of (not) using base allelic frequencies is of utmost importance for unbiased genetic evaluations, yet in most cases they are unavailable – except for dairy cattle (VanRaden 2008).

Conclusions

In a closed population with long and complete pedigrees, clean genotypes, and proper scaling, the difference between coefficients in **G** and **A** follows a symmetric distribution with a mean of 0 and standard deviation of < 0.05. Large differences (beyond, say, 0.10) between relationships in **G** and **A** indicate mistakes in pedigree, incorrect genotypes, insufficient pedigree depth or knowledge, improper scaling or admixed populations, and implies that a stricter QC must be applied or that either the pedigree or the genotypes are incorrect. The figures that we have obtained are for a chicken population but, given the similar sizes of typical livestock genomes, they are possibly adequate for other species. Otherwise, the data correctly fit recent findings on variation of realized relationships around expected pedigree-based values.

Acknowledgements

References

Aguilar I., Misztal I., Legarra A., Tsuruta S. (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.*, **128**, 422–428.

Cassell B.G., Adamec V., Pearson R.E. (2003) Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. *J. Dairy Sci.*, **86**, 2967–2976.

Christensen O.F. (2012) Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet. Sel. Evol.*, **44**, 37.

Cole J.B., Franke D.E. (2002) Pedigree analysis using the Python programming language. *J. Dairy Sci.*, **85**(Suppl 1), 323.

Forni S., Aguilar I., Misztal I. (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.*, **43**, 1.

Garcia-Cortes L.A., Legarra A., Chevalet C., Toro M.A. (2013) Variance and covariance of actual relationships between relatives at one locus. *PLoS ONE*, **8**, e57003.

Goddard M. (2009) Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, **136**, 245–257.

Groenen M.A., Megens H.J., Zare Y., Warren W.C., Hillier L.W., Crooijmans R.P., Vereijken A., Okimoto R., Muir W.M., Cheng H.H. (2011) The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*, **12**, 274.

Habier D., Tetens J., Seefried F.R., Lichtner P., Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.*, **42**, 5.

Hayes B.J., Visscher P.M., Goddard M.E. (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb.)*, **91**, 47–60.

Hill W.G., Weir B.S. (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res. (Camb.)*, **93**, 47–64.

Karigl G. (1981) A recursive algorithm for the calculation of identity coefficients. *Ann. Hum. Genet.*, **45**, 299–305.

Lutaaya B.E., Misztal I., Bertrand J.K., Mabry J.W. (1999) Inbreeding in populations with incomplete pedigrees. *J. Anim. Breed. Genet.*, **116**, 475–480.

Misztal I., Vitezica Z.G., Legarra A., Aguilar I., Swan A.A. (2013) Unknown-parent groups in single-step genomic evaluation. *J. Anim. Breed. Genet.*, **130**, 252–258.

Powell J.E., Visscher P.M., Goddard M.E. (2010) Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.*, **11**, 800–805.

Ritland K. (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res. (Camb.)*, **67**, 175–185.

Simeone R., Misztal I., Aguilar I., Legarra A. (2011) Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. *J. Anim. Breed. Genet.*, **128**, 386–393.

Stranden I., Christensen O.F. (2011) Allele coding in genomic evaluation. *Genet. Sel. Evol.*, **43**, 25.

Toro M.A., Garcia-Cortes L.A., Legarra A. (2011) A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet. Sel. Evol.*, **43**, 1–10.

VanRaden P.M. (1992) Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J. Dairy Sci.*, **75**, 3136–3144.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

VanRaden P.M., Olson K.M., Wiggans G.R., Cole J.B., Tooker M.E. (2011) Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.*, **94**, 5673–5682.

VanRaden P.M., Cooper T.A., Wiggans G.R., O'Connell J.R., Bacheller L.R. (2013) Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. *J. Dairy Sci.*, **96**, 1874–1879.

Vitezica Z.G., Aguilar I., Misztal I., Legarra A. (2011) Bias in genomic predictions for populations under selection. *Genet. Res. (Camb.)*, **93**, 357–366.

Wiggans G.R., VanRaden P.M., Bacheller L.R., Tooker M.E., Hutchison J.L., Cooper T.A., Sonstegard T.S. (2010) Selection and management of DNA markers for use in genomic evaluation. *J. Dairy Sci.*, **93**, 2287–2292.

Wright S. (1934) The method of path coefficients. *Ann. Math. Stat.*, **5**, 161–215.