Proceedings, 10th World Congress of Genetics Applied to Livestock Production

The basis of genetic relationships in the era of genomic selection

T. Meuwissen[1], A. Sonesson[2], J. Ødegård[1,3]
[1]Norwegian University of Life Sciences, Ås, Norway; [2]NOFIMA, Ås, Norway; [3]AquaGen AS, Trondheim, Norway

ABSTRACT: In the literature an abundance of genomic relationship matrices have been described which mainly differ in the age of the relationships that they trace. Marker based relationship matrices (**G**) generally trace very old relationships, since the marker mutations occurred. Pedigree (**A**) and linkage analysis relationship matrices ($\mathbf{G_{LA}}$) trace relationships since pedigree recording started, i.e. since the founder population. Genomic selection (GS) is based on three sources of information: a) pedigree relationships (**A**); b) linkage analysis (LA) information and c)linkage disequilibrium (LD) information, where LD is defined between alleles in the founder population. LD due to cosegregation of alleles in the known pedigree is denoted LA information. The described relationship matrices follow the same pattern, i.e. **A**, $\mathbf{G_{LA}}$ and **G**, respectively. An ideal model for GS is proposed that fits all three sources simultaneously, i.e. the **A** and $\mathbf{G_{LA}}$ matrix and a Bayesian SNP selection model for the LD information.
Keywords: genomic selection; genomic relationships; genetic modeling; whole genome sequence data

Introduction

In genomic selection, the effects of many thousands of genetic markers, usually a genome-wide dense panel of SNPs, are simultaneously estimated in a training population, and used to predict genetic values in evaluation animals, usually selection candidates (Meuwissen et al., 2001). In theory, the linkage disequilibrium (LD) between QTL and the dense marker panel makes it possible to predict the former from the latter. It might have been expected that in this era of SNP effects, the concept of genetic relationship had become obsolete, but nothing is less true, since it has been demonstrated that the SNP based genomic selection model could be rewritten in an equivalent animal model based on a genomic relationship matrix, **G** (Habier et al, 2007; VanRaden 2008; Goddard 2009). Thus, when moving from the conventional BLUP animal model to Genomic-BLUP (GBLUP), we merely exchange the pedigree based relationship matrix **A** with the marker based relationship matrix **G**. The Bayesian models (BayesA/B/C/R; Meuwissen et al., 2001; Habier et al., 2010; Erbe et al., 2012)) attempt to differentially weight particular SNPs, i.e. particular regions in the genome, and thus are using a weighted genomic relationship matrix, with increased weights for important genomic regions.

In this perspective, it is perhaps not surprising that Habier et al. (2007) found that markers can capture and utilize genetic relationships amongst the genotyped animals. Habier et al. (2013) suggested to distinguish three types of information in genomic selection: (i) pedigree (A); (ii) linkage analysis (LA); and (iii) LD information, and that the markers could capture all three types. However, the only way how SNPs can predict QTL is through associations between SNPs and QTL and thus through LD, i.e. all genomic predictions are based on LD information. The latter problem is accommodated by restricting the definition of LD to the LD of the alleles in the founder population (Meuwissen et al., 2000), and any LD in later generations that is generated by the family structure, i.e. by co-segregation of alleles, is called linkage analysis information. Because the **G** resembles **A**, fitting a **G** matrix can capture **A** matrix relationships, and thus can even explain variation at unmarked chromosomes using information type (i). In QTL mapping, where a **G** matrix is fitted for a particular position, it well known that such **G** can capture variation at other unlinked positions through 'population/family structure'.

The pedigree relationship matrix, **A**, is defined relative to a founder population, whose individuals are assumed unrelated and non-inbred. The same holds for the linkage analysis relationship matrix $\mathbf{G_{LA}}$. The marker based relationship matrices are based on identical-by-state (IBS) relationships between the marker alleles and are thus not relative to a well-defined founder population. Two marker alleles are identical when no mutation has occurred on their inheritance paths since their most recent common ancestor. If mutations occur at a rate of ν per generation, the average time since a mutation on either of these inheritance paths is $1/(2\nu)$ generations. Thus, the marker based relationship matrix is relative to a base population that lived approximately $1/(2\nu)$ generations ago. Given that a mutation has occurred sometime at the SNP (otherwise there would not have been a SNP at this position), the per generation mutation rate is increased: with a total coalescence tree length of $\sim 10N_e$ generations, $\nu=1/(10N_e)$ and the founder population occurred $\sim 5N_e$ generations ago, where N_e is the effective population size. SNP markers on a SNP chip are often chosen for their high minor allele frequency (MAF) which implies that they are relatively old mutations, since they drifted to high MAF. In contrast, causative mutations will often be recent relative to neutral mutations, if the trait of interest is under (natural) selection. In this light, it seems that the marker based relationship matrix **G** is considering too old relationships relative to the relationships at the QTL, which are the relationships we want to model.

Thus, with the advent of genome-wide dense SNP data, a plethora of relationship matrices have emerged with different properties and effects on the accuracy genomic selection. The aim is here to review these relationship matrices, their properties, and effects on the accuracy of genomic selection. The properties of the relationship

matrices are also relevant when they are used for the management of inbreeding (Sonesson et al., 2012).

The target Relationship Matrix, $\mathbf{G_T}$

The target or true relationship matrix is the relationship matrix at the QTL weighted by their effects squared as becomes clear from the following model (Goddard et al., 2011):

$$\mathbf{y} = \mathbf{g} + \mathbf{e} = \mathbf{Wu} + \mathbf{e} \qquad [1]$$

where **W** is a matrix of QTL genotypes with $W_{ij}=(X_{ij}-2q_j)$ and X_{ij} being the number of '1' alleles at the QTL (0, 1, or 2) and q_j being the frequency of the '1' alleles. On the one hand model [1] makes clear that the traditional animal model, which models **g** is equivalent to the genomic selection model which models **u** (Habier et al. 2007; VanRaden 2008, Goddard 2009). On the other hand, it becomes clear that our target relationship matrix is Var(**g**)=$\mathbf{G_T}$=**WDW**' with Var(**u**)=**D** is a diagonal matrix with diagonals equal to the variances of the effects of the QTL mutations. In case we assume equal variance for the QTL effects, σ_u^2: $\mathbf{G_T}$=**WW**'σ_u^2. In the case of sequence data, we know all mutations and $\mathbf{G_T}$=**WDW**' with $D_{ii}=0$ for all the SNPs without effects.

The expectation of the diagonal ii of **WW**' is $E(\Sigma_j(X_{ij}-2q_j)^2)=\Sigma_j Var(X_{ij})=\Sigma_j 2q_j(1-q_j)$ (in the absence of inbreeding). In the absence of inbreeding, the diagonals of a relationship matrix are 1, thus we define $\mathbf{G_T}$ as **WW**'/$\Sigma 2q_j(1-q_j)$ (following VanRaden, 2008, who used markers instead of QTL). The expectation of the offdiagonal $(WW')_{ik}$ is for position j (omitting subscripts j for brevity): $E[X_i-2q][X_k-2q]=E[(x_{i1}-q)+(x_{i2}-q)]\ [(x_{k1}-q)+(x_{k2}-q)]=Cov(x_{i1},x_{k1}) + Cov(x_{i1},x_{k2}) + Cov(x_{i2},x_{k1}) + Cov(x_{i2},x_{k2})$, where x_{i1} (x_{i2}) is the allele at the paternal (maternal) gamete of i ($x_{i1}=0$ or 1). Since $Cov(x_{i1},x_{k1}) = \phi_{i1,k1}q(1-q)$, where $\phi_{i1,k1}$ is the coancestry coefficient between gametes i1 and k1, $E[X_i-2q][X_k-2q]=(\phi_{i1,k1}+\phi_{i1,k2}+\phi_{i2,k1}+\phi_{i2,k2})q(1-q) = 2G_{ik}q(1-q)$. Again dividing by the heterozygosity 2q(1-q) shows that $E[X_i-2q][X_k-2q]$ and thus **WW**'/ $\Sigma_j 2q_j(1-q_j)$ indeed reflect the relationships (coancestries) between the animals.

In the above it was assumed that the frequencies q_j equals the current allele frequencies (following Powell et al., 2010), which implies that relationships are expressed relative to the average relationship in the current generation. WW'/ $\Sigma_j 2q_j(1-q_j)$ can contain negative relationships, which implies that the animals are less related than the average relationship in the population. If we use the allele frequencies of a historical generation G0, with frequency q_0, $Cov(x_{i1},x_{k1}) = \phi_{i1,k1}q(1-q) + (q-q_0)^2$, i.e. a term accounting for the drift in allele frequencies since G0 is added. Thus all relationships will be increased by $\Sigma(q_j-q_{0j})^2$ if not the current generation allele frequencies are used. Because $E((q-q_0)^2) = \bar{\phi}q_0(1-q_0)$, where $\bar{\phi}$ is the population average coancestry accumulated since G0, the use of q_0 implies that relationships are expressed relative to the base population G0.

Marker based genomic relationship matrices, G

Marker based estimates of genomic relationships are obtained by replacing the unknown QTL genotypes, W, by known marker genotypes, W^* (calculated in the same way as W), resulting in $G_1 = W^*W^{*\prime}/\Sigma_j 2p_j(1-p_j)$, where p_j is the frequency of the marker alleles (VanRaden, 2008). An estimate with smaller variance can be obtained by weighing the $[X_i-2q][X_k-2q]$ terms with the reciprocal of their prediction error variance. This results in the estimate $\mathbf{G_2}$=**ZZ**'/m, where $Z_{ij}=(X_{ij}-2p_j)/\sqrt{(2p_j(1-p_j)}$ and m is the number of SNPs (Yang et al., 2010; Powell et al., 2010). Yang et al. further suggested an improved calculation for the diagonals of G_2. There are philosophical differences in the sense that G_1 assumes that variance of a_j, i.e. the effect of the SNP is constant, and being independent of the frequency p_j. Whereas $\mathbf{G_2}$ assumes that the variance due to the SNP, $2p_j(1-p_j)a_j^2$, is constant, and thus it is assumed that rare SNPs have large effects to compensate for their low frequency. If the majority of the QTL have low MAF, $\mathbf{G_2}$ seems more appropriate than $\mathbf{G_1}$, since rare QTL alleles cannot be picked up by high MAF markers. In sequence data, the assumptions of $\mathbf{G_1}$ seem more logical, which are the same as for $\mathbf{G_T}$. However, in the case of selection, it seems unlikely that there are common causal mutations with large effects, i.e. some dependency between the size of the effect and the allele frequency is expected. Whether this dependency is as assumed by $\mathbf{G_2}$, where the variance due to causal mutations is independent of their frequency needs further investigations. A related difference between $\mathbf{G_1}$ and $\mathbf{G_2}$ is that $\mathbf{G_1}$ weighs common SNPs more heavily than $\mathbf{G_2}$. Since common SNPs have alleles that drifted to high frequencies, they are expected to be due to old mutations. Thus, $\mathbf{G_1}$ is expected to reflect more old relationships between animals than $\mathbf{G_2}$. For traits under selection, we expect the causal mutations to be relatively young, thus we would expect G_2 to better reflect $\mathbf{G_T}$. However, in practice very little difference in accuracy of genomic selection using $\mathbf{G_1}$ or $\mathbf{G_2}$ has been found. Since $\mathbf{G_1}$ and $\mathbf{G_2}$ are calculated from a quadratic form $\mathbf{W^*W^{*\prime}}$, they are both guaranteed to be semi-positive definite matrices. This implies that adding a small term to their diagonals, makes them positive definite and thus invertible.

Yang et al. (2010) suggested that the missing heritability for human height could be at least partly explained by too low LD between the SNPs and the causal mutations, due to causal mutations having lower allele frequencies than the SNPs on chip. Thus, they suggested that the SNP chip, using $\mathbf{G_2}$, is depicting too old relationships to fully reflect the true relationships ($\mathbf{G_T}$) for human height.

The situation where the SNPs are not in sufficient LD with the causal mutations, may also be due to insufficient SNP density, i.e. too few SNPs. In relationship matrix terms, this means that the relationship matrices are estimated with error, E, i.e.:

$$\mathbf{G} = \mathbf{G_T} + \mathbf{E}$$

In this case, $V(y) = G\sigma_g^2 + I\sigma_e^2 = G_T\sigma_g^2 + E\sigma_g^2 + I\sigma_e^2$, but this model is explaining erroneous (co)variances due to the $E\sigma_g^2$ term. When fitting the G matrix, the REML estimates will accommodate the $E\sigma_g^2$ term by reducing the estimate for σ_g^2. In such a case a better relationship matrix can be obtained by regressing the G elements back to A (VanRaden, 2008; Goddard et al. 2011), i.e.:

$$\hat{G} = bg + (1-b)A$$

The matrix $\hat{G}$ is unbiased in the BLUP sense namely $E(G_T|\hat{G}) = b(G-A)+A = \hat{G}$, where the former is a regression equation with coefficient $b = cov(G,G_T)/var(G)$, cov() and var() denoting (co)variance of the elements of the matrices here, and A being the overall expectation of G_T. The regressed relationship matrix $\hat{G}$ is thus needed if the LD between the SNPs and the QTL is incomplete, either due to incomplete density, holes in the marker map, and structural differences in frequencies between QTL and SNPs. The coefficient b may be estimated by fitting both G and A in a REML analysis (Goddard et al., 2011).

Haplotype based G matrix, $\mathbf{G_{hap}}$

Instead of using single markers to estimate **G**, it may be estimated using marker haplotypes. The haplotypes coalesce earlier, than the single SNPs setting the founder generations $1/(2\rho)$ generations back in time where ρ is the recombination rate within the haplotypes. The estimate is:

$$\mathbf{G_{hap}} = \mathbf{W^*W^{*\prime}} / \Sigma_j 2p_j(1 - p_j)$$

where summation is over all haplotype alleles (across all positions) and $W^*_{ij} = [X_{ij} - 2p_j]$ and X_{ij} is the number of copies of the haplotype allele j. In the literature some but not much increase in accuracy of GS using haplotypes in genomic selection has been found (Calus et al. 2008; Edriss et al. 2013; Meuwissen et al., 2014). Runs of homozygosity (Broman and Weber 1999) can also be used to calculate relationships based on identity of chromosome segments (assuming the marker data have been phased).

The linkage analysis relationship matrix, $\mathbf{G_{LA}}$.

Fernando and Grossman (1989) showed how marker data at a locus and pedigree data can be combined to set up a linkage analysis based G matrix for a locus j, $\mathbf{G_{LA_j}}$, and a fast algorithm for its inverse. $\mathbf{G_{LA_j}}$ is based on knowing the segregation probabilities S_{i1} and S_{i2}, where S_{i1} (S_{i2}) is the probability that animal i inherited its sires' (dam's) paternal allele. Thus, $1-S_{i1}$ is the probability that animal i inherited its sires' maternal allele. The LDMIP algorithm (Meuwissen and Goddard, 2010a) can calculate these S-probabilities for large and complex pedigrees, large numbers of linked loci, and complex patterns of missing genotype data. Working with a gametic model where both gametes of every animal form entries for the matrix, and working from young to old animals, the row of in the $\mathbf{G_{LAj}}$ matrix of gamete i is (assuming i is a paternally derived gamete):

$$G_{LAj}(i,1:i-1) = S_{i1} * G_{LAj}(s,1:i-1) + (1-S_{i1}) * G_{LAj}(d,1:i-1) \quad [2]$$

where s and d are the paternal and maternal gametes of the sire of i. The column of i follows from its row (since G_{LAj} is symmetric), the diagonal elements of G_{LAj} are all 1, and the founder animals are assumed unrelated and noninbred (as in **A**). These calculations are rather simple, but involve large matrices of 2N entries when the numbers of animals in the pedigree N is large. With many markers, the G_{LAj} matrices can be calculated in parallel but the (large) memory requirements increase linearly with the number of matrices calculated in parallel. With dense markers, the G_{LAj} will not change much from one position to the next, so some marker positions may be skipped, but how many needs further investigation. The overall $\mathbf{G_{LA}}$ matrix is obtained by averaging over all the G_{LAj} (Villaneuva et al. 2005; Luan et al., 2012):

$$\mathbf{G_{LA}} = \sum_{j=1}^{m} \mathbf{G_{LA_j}} / m$$

Because of the summation in this expression for $\mathbf{G_{LA}}$ it is not possible to calculate the inverse of $\mathbf{G_{LA}}$ in a rapid manner, although the inverses of the individual $\mathbf{G_{LAj}}$ matrices can be calculated rapidly (Fernando and Grossman, 1989).

The $\mathbf{G_{LA}}$ matrix has some desirable properties: (i) it is calculated for all animals in the pedigree, including the non-genotyped animals, and is as such a 'single-step method' (Legarra et al. (2009); Aguilar et al. (2010); Christensen & Lund (2010)), with a more natural integration of pedigree and marker data then the original single step method (see single-step section); (ii) it is unbiased in the BLUP sense just like **A** and does not need a correction like in $\hat{\mathbf{G}}$ (thus low SNP densities can be used without introducing erroneous covariances **E**); (iii) the founder population is well defined and identical to that of A; (iv) linkage analysis requires only a relatively sparse marker map; and (v) it reflects more recent relationships than the $\mathbf{G_1}$ and $\mathbf{G_2}$ matrices. The latter implies that it does not rely on old relationships to be correct, which may be problematic because old relationships are difficult to estimate accurately, they may not be relevant for relatively young mutations, and for most forms of complex inheritance (non-additive inheritance, epigenetic, etc.) where genetic relationships decay faster than for additive gene effects. Thus, complex inheritance patterns are partly captured by additive genetic relationships, if the relationships are short, but not if they represent old relationships. The lack of extra accuracy of genomic selection when including reference populations from other (related) breeds, demonstrates the difficulties in utilising long distance genetic relationships, even if the markers are able to depict them.

Property (v) has also undesirable effects, namely the $\mathbf{G_{LA}}$ matrix cannot be used in genomic selection to predict across large genetic distances, eg. for across breed predictions, and in situations where the information on close relatives is insufficient, $\mathbf{G_{LA}}$ based genomic selection will not be accurate. The latter occurs in family structures where there are few close relatives, and when heritability is

low, i.e. where many observations over long genetic distances need to be combined in order to achieve an accurate GEBV. The **A** matrix relationships are halved every generation and thus decay fast. In contrast, $\mathbf{G_{LA}}$ relationships decay at variable speed, depending on the actual recombination and inheritance of chromosome segments. Hence, the coefficient of variation for relationships increases as individuals become more distant (Hill and Weir, 2011).

Luan et al. (2012) attempted to distinguish the accuracy of genomic selection due to LD and due to LA by fitting G_{LA} and subsequently fitting G_2 in order to study the increase in accuracy. However, they found no increase in accuracy, and if anything, G_{LA} yielded a higher accuracy than G_2. Ilska et al. (2014) used a weighted average of G_2 and G_{LA} for genomic selection in chicken data and found that this weighted average yielded higher accuracy than either of the original matrices. In a pig genomic selection study, the G_{LA} matrix yielded a lower accuracy than G_2 (Meuwissen et al., 2014), which may be due to a too small number of genotyped generations for an accurate linkage analysis and due to a relatively recent admixture of European and Asian subspecies in pig breeding which likely created substantial LD and thus increased the importance of the use of LD information. Thus, at least in some circumstances, the use of the $\mathbf{G_{LA}}$ matrix can improve genomic predictions.

Combining A and G: the Single-Step matrix H

A problem with G_1 and G_2 is that it can only be calculated amongst genotyped animals, which resulted in multi-step EBV estimation methods, i.e. GEBV and traditional EBV were calculated and subsequently blended into an ultimate breeding value estimate. Hence, Legarra et al. (2009); Aguilar et al. (2010); Christensen & Lund (2010) proposed the single-step method, i.e. to set up a relationship matrix, **H**, that combines ungenotyped and genotyped animals, and uses this H matrix in a single-step GBLUP (SS-GBLUP) to predict GEBV. In this method, the genetic values of the ungenotyped ($\mathbf{g_1}$) animals are predicted from the genotyped ($\mathbf{g_2}$) animals by regression (Meuwissen et al., 2013):

$$\mathbf{g_1 = Bg_2 + \varepsilon}$$

where the regression coefficients **B** come from the **A** matrix, ie. $\mathbf{B=A_{12}A_{22}^{-1}}$. The variance of the residual, **ε**, is $\text{Var}(\boldsymbol{\varepsilon})=\mathbf{R}=\mathbf{A_{11}-BA_{22}B'}$. Since the marker data indicate $\text{Var}(\mathbf{g_2})=\mathbf{G}$, the above regression predicts $\text{Var}(\mathbf{g_1})=\mathbf{BGB'}+\mathbf{R}$, and $\text{Cov}(\mathbf{g_1},\mathbf{g_2})=\mathbf{BG}$. Thus, the combined relationship matrix of ($\mathbf{g_1}$ $\mathbf{g_2}$) is:

$$\mathbf{H} = \begin{bmatrix} \mathbf{BGB'} + \mathbf{R} & \mathbf{BG} \\ \mathbf{GB'} & \mathbf{G} \end{bmatrix}$$

The inverse of **H** is quite simple and fast to compute (Legarra et al. (2009); Christensen & Lund (2010)). In the context of SNP based genomic selection, regressions **B** are used to impute the missing marker genotypes, and $\text{Var}(\boldsymbol{\varepsilon})$ accounts for the imputation errors. The GEBV resulting from **H** have been found biased, which is probably because its derivation assumed that the regression coefficients **B** are not affected by the marker data (Meuwissen et al., 2011). **B** is not affected by marker data when the ungenotyped animals are descendants from the genotyped animals, but not if the ungenotyped animals are ancestors of the genotyped animals. For example in a small halfsib family, two half sibs may be related due to coinheritance of the same chromosome from their father, and their G matrix may reflect this, e.g. showing a relationship of 0.3 instead of 0.25. Linkage analysis will reveal such an inheritance pattern, but the **H** matrix will assume regression coefficients as in the **A** matrix and will translate the increased relationship between the halfsibs into quite strange relationships between the founder alleles (Ødegård and Meuwissen, 2013). Therefore, Meuwissen et al. (2011) suggested to replace **A** by $\mathbf{G_{LA}}$ in SS-GBLUP, because the linkage analysis matrix will account for changes in **B**.

When setting up **H**, care needs to be taken to express the relationships **A** and **G** relative to the same founder population. The founder population of **A** is defined by the depth of the pedigree (usually recent), whereas that of **G** by SNP mutation rates (old). Powell et al. (2010) and Meuwissen et al. (2011) used wrights F_{st}-coefficients to express **A** and **G** to the same founder population, where F_{st} is the average inbreeding of the new (pedigree) founder animals expressed to the old base. The old-base-inbreeding and new-base-inbreeding coefficients are related by:

$$F_{old_i} = F_{new_i} + (1 - F_{new_i})F_{st},$$

where F_{new} is inbreeding to the new, more recent base and F_{old} to the old base. This equation can be used to calculate F_{old_i} given F_{new_i} and vice versa assuming assuming known F_{st}. Meuwissen et al. (2011) distinguished diagonal and offdiagonal elements of the relationship matrix, but when the gametic relationship matrix is used this distinction is not needed (all diagonals are 1), and also the gametic relationships are identical to the coancestry coefficients, such that the above equation can be applied directly (no transformation between relationships and coancestries) .

Combining LA and LD information: $\mathbf{G_{LDLA}}$

It seems both LA and LD provide useful information for GS, so our aim is here to combine both information sources. A simple approach is to weigh the G_{LA} and G matrix in to an overall matrix G_{LDLA1} (Meuwissen et al., 2011; Ilska et al., 2014). $\mathbf{G_{LDLA1}}$ is similar to the $\hat{\mathbf{G}}$ matrix, except that **A** is replaced by $\mathbf{G_{LA}}$. The weighting factor could be obtained from fitting both variance matrices in a REML analysis and estimate the variance components/weights. However, Ilska et al. did not always find the highest accuracy for the maximum likelihood estimate of the weighting factor. $\mathbf{G_{LDLA1}}$ can only be setup for genotyped animals.

A second alternative for $\mathbf{G_{LDLA}}$ is obtained by defining LD as associations between the loci in the founder population, and LA explains than any associations generated by cosegregation of alleles within the known pedigree, thereby

separating LD and LA information (Meuwissen et al., 2000). This removes the assumption of the $\mathbf{G_{LA}}$ matrix that founder animals are unrelated, i.e. have an identity relationship matrix, since the founder alleles are assumed here related through the LD between the SNPs. In the case, where all animals are genotyped, the relationships between the founder animals could be obtained from the markers using $\mathbf{G_1}$ or $\mathbf{G_2}$, say $\mathbf{G_{founders}}$. And later generation relationships could be obtained by linkage analysis, and applying equation [2] for all later gametes, resulting in $\mathbf{G_{LDLA2}}$. In the case, where all animals are genotyped and the inheritance of the gametes could be traced at all positions, this would result in the entire **G** matrix, as also could be calculated directly from the markers, assuming that the marker density is sufficient to accurately estimate **G**. If the founders are genotyped, but (some) descendants are not, this approach would still result in the best possible $\mathbf{G_{LDLA2}}$ matrix, showing the importance to densely genotyping founders in genomic selection schemes. The latter is however not always possible, and in which case the genotype probabilities of the founder animals, as obtained from the linkage analysis (LDMIP), can be used to estimate $\mathbf{G_{founders}}$. These genotype probabilities are not as extreme as actual genotypes, and thus show less variance, which implies that the diagonals of the $\mathbf{G_{founders}}$ will be severely underestimated. In a gametic model, the diagonals are 1 by definition (relationship of a gamete with itself), and thus they should be set to 1 in $\mathbf{G_{LDLA2}}$.

A third alternative to set up $\mathbf{G_{LDLA3}}$ is to use the same approach as for $\mathbf{G_1}$ or $\mathbf{G_2}$, but substituting the genotypes, when they are missing, by genotype probabilities from linkage analysis (LDMIP). When this algorithm is used for a gametic **G** matrix (2N entries),

$$\mathbf{G_{LDLA(a)}} = \mathbf{W^*W^{*\prime}} / \Sigma_j p_j(1 - p_j)$$

where W^*_{ij}= [prob("1"allele)-p_j]. The diagonals of $\mathbf{G_{LDLA(a)}}$ may deviate from 1 (despite $\mathbf{G_{LDLA(a)}}$ being a gametic relationship matrix), which is remedied by scaling the diagonals >1, back towards 1, i.e. $\mathbf{G_{LDLA(b)}} = \mathbf{\Delta_1 G_{LDLA(a)} \Delta_1}$, where $\mathbf{\Delta_1}$ is a diagonal matrix with elements equal $1/\sqrt{diag(\mathbf{G_{LDLA(a)}})}$ if diag($\mathbf{G_{LDLA(a)}}$)>1 or equal 1 otherwise. The too small diagonal elements are remedied by adding some scaled parts of the **A** matrix, i.e. $\mathbf{G_{LDLA3}} = \mathbf{G_{LDLA(b)}} + \mathbf{\Delta_2 A \Delta_2}$, where $\mathbf{\Delta_2}$ is a diagonal matrix with elements equal $1/\sqrt{1 - diag(\mathbf{G_{LDLA(a)}})}$ if diag($\mathbf{G_{LDLA(a)}}$)<1 or equal 1 otherwise. The $\mathbf{\Delta_2 A \Delta_2}$ term has similarities with the residual matrix **R** in SS-BLUP in that both account for the relationship that is not predicted by the marker data (Fernando et al., 2013). We are not using **R** here because its use does not guarantee that the diagonals to become 1. As $\mathbf{G_1}$ and $\mathbf{G_2}$, $\mathbf{G_{LDLA(a)}}$ is guaranteed to be semi-positive definite and the transitions to $\mathbf{G_{LDLA(b)}}$ and $\mathbf{G_{LDLA3}}$ preserve this property. If the founder animals are not genotyped, $\mathbf{G_{LDLA2}}$ relies on genotype probabilities to introduce the LD information, whereas $\mathbf{G_{LDLA3}}$ uses the genotyped animals directly to infer LD information, i.e. there is a more direct use of LD information.

Discussion

In the above an abundance of **G** matrices have been described and the main question is when to use which **G**? To answer this question we need to assume a true genetic model. We will assume a broad genetic model here consisting of several forms of genetic effects: (a) some genetic effects are truly polygenic; allelic effects are small and nearly neutral and their relationship matrix is not structurally different from that of randomly picked SNPs, i.e. $\mathbf{G_1}$ is appropriate ($\mathbf{G_2}$ if deviations from neutrality have equalised the SNPs genetic variances across the allele frequency spectrum or SNPs on the chip are selected for high MAF); (b) some genetic effects are due to major haplotypes, who carry several causal mutations, whose individual effects are small, but due to interactions their combined and summed effect is substantial; (c) some major genes may occur, best modelled by BayesB/C/R; (d) transgenerational epigenetic inheritance systems acting on the chromosomes (e.g., due to DNA methylation) may be best modelled with $\mathbf{G_{LA}}$, or **A** (depending on the "half-life" for such effects). Because recombinations break up the interaction effects of the major haplotypes (b), they are short lived and best explained by short genetic relationships such as $\mathbf{G_{LA}}$ or $\mathbf{G_{hap}}$. The more short lived relationships may also capture dominance effects better than older additive relationships, and recent mutations (which often explain substantial amounts of variation in disease traits).

The ideal model of analysis would thus look like (ignoring fixed and other environmental effects):

$$\mathbf{y} = \mathbf{a} + \mathbf{h} + \mathbf{g} + \Sigma_j \mathbf{W_j} * b_j + \mathbf{e} \qquad [3]$$

where **a**, **h** and **g** are all animal effect with variances $V(\mathbf{a})=\mathbf{A}\sigma_a^2$, $V(\mathbf{h})=\mathbf{G_{LA}}\sigma_h^2$ (or $\mathbf{G_{hap}}\sigma_h^2$), $V(\mathbf{g})=\mathbf{G_1}\sigma_g^2$ (or $\mathbf{G_2}\sigma_g^2$), and the $\Sigma_j \mathbf{W_j} * b_j$ term represents the BayesB/C/R model fitting the occasional SNP with large effects. The **a**-effects can be dropped from the model, when epigenetic effects are absent or show half-lifes similar to $\mathbf{G_{LA}}$ relationships, and all chromosomes are covered with dense markers. The terms in the above model have all been tried individually but have not been fitted simultaneously. Simultaneous estimation of the variance components σ_a^2, σ_h^2, σ_g^2, and the parameters of the Bayesian model will be a challenge, especially since the **G** matrices will be highly correlated. GBLUP and the Bayesian models have often been compared, favouring one or the other, but they assume very different genetic models, which might imply that when fitted to together they could explain more of the genetic effects. The weights given to the matrices may also be optimised by crossvalidation, but this will yield an upward bias for the crossvalidation accuracies.

If not all animals are genotyped, the $\mathbf{G_{1/2}}$ relationships for **g** in [3] are not available for all animals. In this case, Var(**g**) may be set equal to $\mathbf{G_{LDLA3}}$ or **H**, where the advantage of $\mathbf{G_{LDLA3}}$ is that imputation of missing genotypes is by linkage analysis instead of by **A**-matrix based regression coefficients **B**. For the BayesB/C/R model, LA based

genotype probabilities can be used for the missing W_{ij}^*. This resembles Haley-Knott regression in QTL mapping (Haley and Knott, 1992),where also QTL probabilities were used instead of actual genotype, and biases due to this were small. Whether such biases will also be small here needs further investigation.

Some of the described relationship matrices automatically combine two or more information sources. E.g., $\mathbf{G_{LA}}$ combines pedigree and the LA part of the marker information. Such automatic combination has the advantage that only variance component needs to be estimated, but has disadvantages when the theory underlying the combination of the matrices does not hold. E.g. in the case of $\mathbf{G_{LA}}$, the pedigree relationship matrix should obtain extra weight when there are holes in the marker map. Similarly, the $\mathbf{G_{LDLA}}$ matrices combine LD and LA information. But for instance in $\mathbf{G_{LDLA2}}$ the genotype information on the founder animals may be too poor to obtain an accurate estimate of their LD based G matrices and the errors will propagate through the later generations through the linkage analysis. A separate weighing of $\mathbf{G_{LA}}$ and $\mathbf{G_{1/2}}$ into a blended G matrix may yield better accuracies of genomic selection (Meuwissen et al. (2011); Ilska et al. (2014)):

$$\mathbf{G_{blend}} = \mathbf{G_{LA}}\lambda + (1-\lambda)\,\mathbf{G_2}.$$

and different λ coefficients may be appropriate for different situations. In $\mathbf{G_{blend}}$, λ can also account for erroneous covariances, **E**, found in the estimate $\mathbf{G_2}$ due to finite number of SNPs being used in its estimation.

Whole-genome sequence (WGS) data

When WGS data are available (imputed) on all animals, Meuwissen and Goddard (2010b) concluded that all causative mutations are in the data and a SNP model selection algorithm (e.g. BayesB) should find the best model, i.e. there is no need to fit relationship matrices. However, if we assume the earlier described, broad genetic model of inheritance, fitting the truly polygenic effects (a) may be computationally more efficient by fitting $\mathbf{G_{1/2}}$ than by fitting >20 millions of SNPs each with tiny effects. In WGS data all SNPs are available, such that selection biases of the SNPs can be avoided. Effects of major haplotypes (b), which cause short-lived genetic relationships would be poorly accommodated by the SNP selection model. Also Copy Number Variants (CNVs) with their higher mutation rates would cause an increased importance of recent relationships. The major genes (c) would be well accommodated by BayesB, but their effects are so big that all models will capture them reasonable well. Any epigenetics or holes in the WGS data would obviously not be captured by WGS based BayesB. Thus, in the light of the aforementioned broad model of inheritance, relationship matrices are very relevant even in the presence of WGS data.

Literature Cited

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J. (2010) J. Dairy Sci. 93:743-752.

Broman, K.W., Weber, JL. (1999) Am.J.Hum.Gen. 65:1493

Calus M., T. Meuwissen , A. de Roos., R. Veerkamp (2008) Genetics 178: 553 - 561

Christensen, O. and M. Lund. (2010). Gen. Sel. Evol. 42, 2.

Edriss V., R. Fernando, G. Su, M. Lund, B. Guldbrandtsen (2013) Gen. Sel. Evol. 45:5.

Erbe M. et al. (2012) J Dairy Sci 95:4114-29

Fernando R.L., M. Grossman (1989) Gen. Sel. Evol. 21: 467-477

Fernando R.L., D. Garrick, J. Dekkers (2014) 64th EAAP.

Goddard, M.E. (2009) Genetica 136:245-57.

Goddard, M.E., B.J. Hayes, T. Meuwissen (2011) J. Anim. Breed. Genet. 128:409-21.

Habier, D., R.L. Fernando, J.C. Dekkers (2007). Genetics 177:2389-97.

Habier D, R.L. Fernando, K.Kizilkaya (2010) 9th WCGALP

Habier, D., R. L. Fernando, and D. J. Garrick. (2013). Genetics 194, 597-607.

Haley, C.S., S.A. Knott (1992) Heredity 69:315-324

Ilska J.J., T. Meuwissen, A. Kranis, J. Woolliams (2014) Gen. Sel. Evol. Submitted.

Legarra, A., I. Aguilar, and I. Misztal. (2009). J. Dairy Sci. 92, 4656-4663.

Luan, T., J. A. Wooliams, J. Ødegård, et al. (2012). Gen. Sel. Evol. 44, 28.

Meuwissen, T.H.E., M.E. Goddard (2000) Genetics155:421-430

Meuwissen, T.H.E., B.J. Hayes, M.E. Goddard (2001) Genetics 157: 1819-1829

Meuwissen, T.H.E. and M.E. Goddard. (2010a). Genetics 185, 1441-1449.

Meuwissen, T.H.E. and M.E. Goddard. (2010b). Genetics 185:623-31

Meuwissen, T., T. Luan, J. Woolliams (2011). J. Anim. Breed. Genet. 128:429-439.

Meuwissen, T., B. Hayes, M. Goddard (2013). Ann. Rev. Anim. Biosci. 1: 221-237

Meuwissen T., J. Ødegård, I. Ranberg, E. Grindflek (2014) Gen. Sel. Evol. Submitted.

Powell J.E., P.M. Visscher, M.E. Goddard (2010) Nature Reviews 11:800-805

Sonesson A., J.A. Woolliams, T. Meuwissen (2012) Gen. Sel. Evol. 44:27.

VanRaden, R. M. (2008). J. Dairy Sci. 91:4414-23.

Villanueva B, Pong-Wong R, Fernández J, Toro MA (2005) J. Anim. Sci. 83:1747-1752

Yang J. et al. (2010) Nature Genet. 42:565-571

Ødegård, J., T. Meuwissen. (2013). 64th EAAP, Nantes.