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

Genomic Heritability: What Is It?

Gustavo de los Campos[1], **Daniel Sorensen**[2] **and Daniel Gianola**[3]

[1]University of Alabama at Birmingham, US, [2]Aarhus University, Denmark. [3] University of Wisconsin, Madison, US.

ABSTRACT. Whole-genome regression models have become ubiquitous for analysis and prediction of complex traits. In human genetics, these methods are commonly used for inferences about genetic parameters. This is so despite the fact that some of the assumptions commonly adopted for data analysis are at odds with important quantitative genetic principles. In this article we develop theory that leads to a precise definition of parameters arising in regression models using genomic data. Our approach is framed within the classical quantitative genetics paradigm. We discuss how these parameters relate to statistical parameters, indicate potential inferential problems and provide a limited set of simulations where some statistical properties of likelihood-based estimates are assessed.

Keywords: Genomic heritability; G-BLUP; Quantitative genetics; Whole-genome regression; Missing heritability.

Introduction

Whole-genome regression (WGR) methods (Meuwissen, Hayes, and Goddard, 2001) are becoming increasingly used for analysis and prediction of complex traits. These methods were first used for prediction in animal and plant breeding. More recently, there has been interest in the use of WGR methods for inferences about "genomic heritability" (e.g., Yang et al. , 2010).

Prediction and inference are two different problems: a model that yields good (e.g., unbiased and precise) estimates may have poor prediction performance and vice versa. Unfortunately, little is known about the inferential properties of estimates derived from WGR models. For example, it is unclear whether the likelihood-based estimators commonly used estimate population parameters consistently (de los Campos and Sorensen 2013).

Before the introduction of dense genetic marker information, a common approach used to infer genetic variances and derived parameters was based on mixed linear models applied to family data (e.g., Henderson, 1975). With use of molecular markers it has become possible to assess kinship among nominally unrelated individuals. With this, it is now feasible to analyze data from nominally unrelated individuals using methods originally developed for family data. This has made evident the distinction between the data generating process and the model used for data analysis, or instrumental model: the marker genotype information embedded in the instrumental model is used in lieu of the "causal" genotypes that are part of the classical model of quantitative genetics. Once the distinction between the instrumental and the true model is made, it is no longer clear whether or not the parameters of the instrumental model (e.g., the genomic variance) can be equated to population parameters of interest (e.g., the genetic variance).

In human genetics, Yang et al. (2010) was the first study that used a WGR approach for estimation of 'genomic heritability'. Using a G-BLUP type model, these authors found that approximately half of the heritability of human height was captured by common SNPs as opposed to the 5-10% explained by GWAS-significant SNPs (Lango Allen et al. 2010). The proportion of unexplained genetic variance can be interpreted as 'missing heritability' and it has been attributed to imperfect LD between the markers used and the QTL affecting the trait. In the literature (e.g., Yang et al. 2010; Zaitlen and Kraft 2012; Speed et al. 2012) genetic parameters (e.g., heritability or genomic heritability) have been defined based on the statistical assumptions adopted in the instrumental model, some of which are at odds with principles of quantitative genetic theory. This results in a fuzzy connection between statistical parameters and the quantitative genetic parameters one wishes to infer.

The primary contribution of this paper is to develop theory leading to precise definitions of parameters arising in WGRs. Our approach is framed within the classical quantitative genetics paradigm. We discuss how these parameters relate to statistical parameters defined based on models commonly used for data analysis, consider potential estimation problems that may emerge, and provide a limited set of simulations where some properties of likelihood-based estimates are assessed.

Theory

In standard quantitative genetic theory (Falconer and Mackay 1996) additive genetic values are linear functions of allele content at one or more QTL. Concepts such as the allele substitution effect in a population or narrow sense heritability are defined with reference to this framework. However, in practice, the set of genes affecting a complex trait is typically unknown and empirical (instrumental) linear regression models are fitted using markers whose alleles are typically in imperfect LD with those at QTL. Concepts such as the "additive effect of a marker", or amount of variance explained by marker effects (the "genomic variance") have appeared in the literature (e.g., Goddard 2009). However, a precise definition of these parameters and a mathematical treatment that holds regardless of trait architecture or patterns of LD are lacking. In this section we attempt to fill this gap by presenting theory framed within a quantitative genetics perspective.

Conceptual QTL model

Assume that a trait of interest measured on individual i (y_i) is affected by alleles at q bi-allelic QTL. In quantitative genetics, the genetic value of an individual is defined as the expected phenotypic value given QTL genotypes, $z_i = \{z_{i1}, \dots, z_{iq}\}'$, $g_i = E(y_i|z_i)$. The conditional expectation function may not be linear on QTL; however, regardless of the genetic mechanism, one can always define a linear approximation of the form

$$y_i = \alpha' z_i + \delta_i. \qquad [1]$$

where α, a column vector of dimension q, represents the vector of effects of allele substitutions (Falconer and Mackay 1996), defined as the regression of g_i or y_i on z_i, that is

$$\alpha = Cov(z_i, z_i')^{-1} Cov(z_i, g_i) = \Sigma_z^{-1}\Sigma_{zg}. \qquad [2]$$

Above, Σ_z is a $q \times q$ matrix whose entries are the variances and covariances of allelic contents at the q QTL, and $\Sigma_{zg}$is a q-dimensional vector containing covariances between QTL genotypes and genetic value. In [1], the deviate δ_i is a random residual that includes environmental and genetic effects that cannot be captured by the linear regression on allele contents, e.g., dominance, epistasis and QTL-environment interactions. By construction δ_i is uncorrelated with z_i. The terms Σ_z and Σ_{zg}, and therefore α, are viewed as fixed population quantities and not as random variables. On the other hand, $\alpha' z_i$, is random because QTL genotypes vary between individuals.

Equation [1] leads to the decomposition of phenotypic variance, $Var(y_i) = Var(\alpha' z_i) + Var(\delta_i)$, or $\sigma_y^2 = \sigma_a^2 + \sigma_\delta^2$, where

$$\sigma_a^2 = \alpha' \Sigma_z \alpha = \sum_{j=1}^{q}\sum_{j'=1}^{q} Cov(z_{ij}, z_{ij'})\alpha_j \alpha_{j'} \qquad [3],$$

is the ***additive genetic variance***, stemming from the regression of phenotype on allelic contents at QTL. Randomness in [3] arises from variation and covariation of allelic contents at the QTL, as postulated in the standard quantitative genetic model, e.g., Falconer and Mackay (1996), and α is a fixed parameter. Expression [3] shows that the additive variance is not only a function of the variances of QTL genotypes (the diagonal elements of Σ_z) but also of the patterns of LD between QTL (the off-diagonal elements of Σ_z). For this reason, in general, the additive variance cannot be partitioned into locus-specific components (Daniel Gianola, Hospital, and Verrier 2013).

Narrow sense heritability is defined as the proportion of phenotypic variance explained by additive effects, that is

$$h^2 = \frac{\sigma_a^2}{\sigma_y^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\delta^2} = \frac{\alpha \sum_z \alpha}{\alpha \sum_z \alpha + \sigma_\delta^2}$$

Instrumental Model (regression on markers)

In practice, the QTL are unknown, thus their genotypes and effects. Empirically, the analysis is carried out using p markers with genotype codes in the vector $x_i = \{x_{i1}, \dots, x_{ip}\}'$; as before, we assume that marker genotypes have been centered. The marked additive genetic value can be defined as the regression of the true additive genetic value, $\alpha' z_i$, on allelic content at marker loci, that is

$$\alpha' z_i = \beta' x_i + \xi_i \,, \qquad [4]$$

where ξ_i is a model residual representing components of the true additive genetic values that cannot be explained by a regression on markers.

Marker effects are defined as the multivariate multiple regression of additive genetic values on markers, as

$$\beta = Var(x_i)^{-1} Cov(x_i, z_i'\alpha) = \Sigma_x^{-1}\Sigma_{xz}\alpha = B\alpha \qquad [5]$$

where Σ_x is the $p \times p$ covariance matrix among marker genotypes, Σ_{xz} is a $p \times q$ matrix of covariances between marker and QTL genotypes, and $B = \{b_{x_j z_k}\}$ is a $p \times q$ matrix of regression coefficients. Since Σ_x , Σ_{xz} and α are fixed population parameters, so is β.

Genomic values are then defined as $\beta' x_i = \alpha' \Sigma_{zx}\Sigma_x^{-1} x_i = \alpha' \hat{z}_i$, where $\hat{z}_i = \Sigma_{zx}\Sigma_x^{-1} x_i$ is the best linear predictor of allele content at QTL, given allele content of markers. The variance of genomic values, or ***genomic variance*** is

$$Var(\beta' x_i) = \beta' Cov(x_i, x_i')\beta = \alpha' \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz}\alpha \,. \qquad [7]$$

Its value depends on the QTL effects (α) and on the LD relationships among QTL and markers (via Σ_{xz}), and among markers (via Σ_x^{-1}). The variance of the regression residual is

$$Var(\xi_i) = Var(\alpha' z_i - x_i'\beta) = \alpha' \Sigma_z \alpha - \alpha' \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz}\alpha$$
$$= \alpha' (\Sigma_z - \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz})\alpha \,.$$

Because ξ_i is uncorrelated with x_i, the model in equation [4] yields the variance partition $Var(z_i'\alpha) = Var(x_i'\beta) + Var(\xi_i)$, leading to

$$\alpha' \Sigma_z \alpha = \alpha' \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz}\alpha \quad + \alpha'(\Sigma_z - \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz})\alpha \qquad [8]$$

or $\sigma_a^2 = \sigma_g^2 + \sigma_{\tilde{g}}^2$ where,

$$\sigma_g^2 = Var(x_i'\beta|\alpha) = \alpha' \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz}\alpha \,, \qquad [9]$$

the genomic variance, is interpretable as the amount of additive variance captured by the regression on markers. Likewise, $\sigma_{\tilde{g}}^2$ can be interpreted as the "missing" additive genetic variance, that is, the variability yet to be marked.

In the above definition we used standard quantitative genetic theory assumptions where genotypes are random and additive effects (α) are fixed. As stated, the genomic variance as defined in expression [9], depends on additive effects α at QTL and on the patterns of LD between markers and QTL (Σ_{xz}) and among markers (Σ_x).

The ratio σ_g^2/σ_a^2 represents the proportion of additive variance that is explained by a linear regression on available markers and the product of this ratio times the trait heritability is the proportion of variance of phenotypes explained by the regression on markers, or ***genomic heritability***:

$$h_g^2 = h^2 \frac{\sigma_g^2}{\sigma_a^2} = \frac{\sigma_a^2}{\sigma_y^2}\frac{\sigma_g^2}{\sigma_a^2} = \frac{\sigma_G^2}{\sigma_y^2} \qquad [10]$$

The proportion of ***missing heritability*** can be defined as a population parameter as

$$\frac{h^2 - h_g^2}{h^2} = \frac{\sigma_a^2 - \sigma_g^2}{\sigma_a^2} = \frac{\alpha'(\Sigma_z - \Sigma_{zx}\Sigma_x^{-1}\Sigma_{xz})\alpha}{\alpha' \Sigma_z \alpha}.$$

This parameter is defined with respect to a trait in a population and in reference to a technology (i.e., a set of markers).

Special Cases

With only a single marker-QTL pair, Σ_{zx} and Σ_x are scalars and expression [5] becomes $\beta = \Sigma_x^{-1}\Sigma_{xz}\alpha = b_{zx}\alpha$ where $b_{zx} = \frac{cov(x_i,z_i)}{var(x_i)}$ is the (population) linear regression of the QTL genotype on the marker genotype. The genomic value is then $x_i\beta = x_i b_{zx}\alpha$, and the expression for the genomic variance (equation [9]) reduces to $\sigma_G^2 = \frac{cov(x_i,z_i)^2}{var(x_i)}\alpha^2$. The proportion of additive variance explained by the regression on markers, $\frac{\sigma_G^2}{\sigma_a^2} = \frac{\left[\frac{cov(x_i,z_i)^2}{var(x_i)}\alpha^2\right]}{[var(z_i)\alpha^2]} = r^2$, is simply the squared correlation between genotypes at the marker locus and at the QTL. Therefore, the genomic heritability is $h_g^2 = r^2 h^2 \leq h^2$. If LD is perfect, $h_g^2 = h^2$; otherwise, it will get closer to 0 as LD becomes weaker.

Multi Marker-QTL pairs in LE. Goddard (2009) proposed a framework where QTL are in mutual LE and, for each QTL, there is a single associated marker. In this stylized setting the genome can be represented as independent QTL-markers pairs (z_{ij}, x_{ij}) . Under these conditions several simplifications occur. Marker effects are simply obtained by regressing QTL genotypes on markers within pairs, that is $\beta_j = b_{z_j x_j}\alpha_j$ where $b_{z_j x_j} = \frac{Cov(x_{ij},z_{ij})}{Var(x_{ij})}$ is the regression of the *j*th QTL on the *j*th marker. Also, the genomic variance can be uniquely decomposed as the sum of marker-specific terms: $\sigma_g^2 = \sum_j \frac{cov(x_{ij},z_{ij})^2}{var(x_{ij})}\alpha_j^2 = \sum_j var(z_{ij}) r_j^2 \alpha_j^2$ where r_j^2 is the squared correlation between the marker and the QTL genotype at the *j*th pair (Goddard, 2009). This decomposition does not hold if multiple markers are linked to the same QLT or if QTL are in LD.

Analysis with "Causal Variants". With sequence data all "causal variants" are expected to be included in the marker panel; therefore, it is reasonable to expect that there will be no missing heritability. The framework outlined in previous sections is consistent with this view. In fact, using results of inverses of partitioned matrices, it can be shown that when all "causal variants" are included in the marker panel, marker effects satisfy $\beta_j = \{\alpha_j \; if \; x_j \; is \; a \; QTL \; ; 0 \; otherwise\}$; therefore the genomic variance is equal to the additive variance and, as one would expect, there is no missing heritability.

On defining genomic variances based on statistical models

As stated, in classical quantitative genetics theory, genetic variance arises from variation and covariation of allelic contents at QTL, and both QTL and marker effects are fixed population parameters. On the other hand, in a typical Bayesian genomic model marker genotypes are fixed quantities and marker effects are regarded as random variables. For instance, in the family of models named the Bayesian Alphabet (e.g., Gianola et al. 2009, 2013) marker effects $b = \{b_j\}_{j=1}^p$ are assumed to be identically and independently distributed (IID) draws from a common prior distribution with null mean and variance $Var(b_j) = \sigma_b^2$. The regression model is built conditional on the observed marker genotypes, and the prior variance of the *i*th genomic value is $Var(x_i' b|x_i) = \sigma_b^2 \sum_j x_{ij}^2$. Under HW equilibrium, the expected value of this parameter is

$$\sigma_u^2 = E\{Var(x_i' b|x_i)\} = \sigma_b^2 \sum_j 2\pi_j(1-\pi_j) \quad [11]$$

where π_j represents the frequency of one of the alleles at marker *j*. Expression [11] is usually referred to as the "genomic variance" (VanRaden 2008). However, the link between [11] and the population parameter σ_g^2 defined in [7] is tenuous. First, from a classical quantitative genetics perspective, marker effects are fixed constants and not random variables. Importantly, expression [11] suggests that the genomic variance can be decomposed into locus-specific components. However as noted earlier, under general conditions this is not possible, because LD affects both the additive and the genomic variances, precluding a decomposition such as that implied by [11].

Estimation of Genomic Variance Using a G-BLUP model

Above we argued that the statistical parameter σ_u^2 has a tenuous connection with the population parameter σ_g^2. Another important consideration is to determine whether σ_u^2 can be correctly inferred. In this section we assess some statistical properties of maximum likelihood estimates using the Genomic Best Linear Unbiased Predictor (G-BLUP) model. In G-BLUP phenotypes are regressed on markers using the linear model $y_i = \sum_j x_{ij} b_j + \varepsilon_i$ where $b_j \sim^{iid} N(0, \sigma_b^2)$ and $\varepsilon_j \sim^{iid} N(0, \sigma_\varepsilon^2)$. The model implies the marginal distribution of phenotypes

$$y \sim N(0, G\sigma_u^2 + I\sigma_\varepsilon^2) \quad [12]$$

where G is a genomic-relationship matrix and σ_u^2 is a variance parameter. Maximizing the likelihood function associated with [12] yields maximum likelihood estimates of variance components and of the proportion of variance explained by the model $h_u^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$. Since G is computed based on markers that are in imperfect LD with QTL, the (co)variance patterns of additive effects and consequently the likelihood function, can be misspecified; this can lead to inconsistency of estimates.

Simulation Studies

Phenotypes were generated according to the additive QTL model of eq. [1]. QTL effects and error terms were sampled from IID normal distributions with variance parameters tuned to achieve a unit phenotypic variance and $h^2 = 0.5$. Two simulations were considered. In the first one, markers and QTL were generated according to stylized LD patterns; in a second simulation we used real human genotypes.

Simulation 1 (simplified LD patterns)

Here, allele frequencies were sampled from a beta distribution with shape parameters equal to 2 and 8. This gives an average allele frequency of 0.2 and a relatively large range of variability in MAF. Haplotypes were simulated with a Markov process where the extent of LD within blocks was controlled by transition probability parameters. Genotypes were formed by randomly mating haplotypes. Genomes consisted of 50,000 loci, 200 of which were QTL. We considered 4 simulation settings that differed in: (a) number of loci per block and number of blocks and (b) whether the patterns of LD were homogeneous or heterogeneous across blocks. In scenarios with short LD blocks (SB), 10,000 blocks were generated with 5 loci in LD in each block. In scenarios with long blocks (LB), 1,000 blocks were generated each with 50 loci. In scenarios with homogeneous LD patterns the transition probability was fixed. In scenarios with heterogeneous LD patterns, the transition probably at each LD block was sampled from a beta distribution with shape parameters 2 and 8. QTL positions were chosen at random: in the SB scenarios, 200 blocks were randomly selected out of the 10,000, and a QTL was assigned to a randomly chosen locus within the LD block. In LB scenarios the QTL positions were assigned at random within the 50K-loci genome.

A total of 3,000 Monte Carlo (MC) replicates, each with a population size of 10,000 were generated. All these 10,000 individuals were used to calculate population parameters such as genetic, phenotypic and genomic variances using formulae presented previously. From the 10K individuals, 1K were chosen at random, and data from these were used to estimate heritability and genomic heritability using a G-BLUP model. The G-BLUP model was fitted using genomic relationship matrices computed using only QTL; QTL and markers in the LD blocks containing QTL (QTL+MRK.LD); all loci (ALL); only MRK.LD; MRK.LD plus markers in LE with QTL (MRK.LD+MRK.LE), and only MRK.LE. According to the theory above-described, in the analysis setting where QTL are in the panel (QTL, QTL+MRK.LD, ALL loci) $h_g^2 = h^2 = 0.5$. In analysis using MRK.LD and MRK.LD+MRK.LE the proportion of variance explained is $h_G^2 \leq 0.5$. Finally, in the scenario including only MRK.LE $h_G^2 = 0$.

Results. Table 1 shows the average (over MC replicates) estimates of heritability and genomic heritability. Columns 3-4 of Table 1 provide the population parameters and columns 5-10 provide the average MC estimates and the corresponding SEs, by scenario and model. Heritability was 0.5, and genomic heritability ranged from values close to 0.3 in SB scenarios and slightly higher 0.327-0.328 in LB scenarios. When only QTL genotypes were used to compute the G-matrix (column 5 in Table 1) the average estimated genomic heritability was 0.5 and the estimates were rather precise. This suggests that ML estimation with this sample size yields estimates with no detectable bias, if the model holds. When QTL+MRK.LD genotypes were used to compute G (column 6 of Table 1) the ML estimate of h_G^2 was close to the true population parameter in the SB scenario, and had a small upward bias in the LB scenarios. When MRK in LE were included the SEs increased relative to those in the analysis based on QTL only. When genotypes at all loci were used to compute G a substantial upward bias in estimates of heritability was observed, and the bias was largest in the SB scenarios. This is likely due to the fact that in this scenario there are many more markers in LE with QTL than in the LB scenarios.

Columns 8-9 of Table 1 show results obtained using MRK.LD and MRK.LD+MRK.LE. Here there is missing heritability (compare columns 3 and 4 of Table 1). In the SB scenario with homogeneous LD patterns (fixed transition probability) using MRK.LD genotypes only, the genomic heritability was estimated almost without bias. However, in all other simulation scenarios there was an important upward bias in estimates of genomic heritability. This was accentuated when MRK.LE were added to MRK.LD to compute G. These results are in line with what we observed in the analysis including ALL loci (column 7 of Table 1) and suggest that adding large numbers of markers in LE with QTL increases the sampling variance of estimates and may induce bias.

Table 1. Mean (SD) of estimates of genomic heritability by simulation scenario (rows) and information used for analysis (Columns 5-1).

LD Block					.505 (.072)

Simulation 2 (with real human genotypes)

This simulation made use of real human genotypes; these genotypes reflect LD patterns that are more realistic than those considered in the previous section. On the other hand, with real genotypes, the population parameters Σ_z, Σ_{zx} and Σ_x, cannot be reliably estimated using a sample of 5,000 individuals; therefore the true genomic heritability remains unknown. Nevertheless, the analyses were performed incorporating only loci assigned as QTL, including only loci assigned as markers (MRK), and combining markers and QTL (MRK+QTL). According to theory, in analyses using QTL and MRK+QTL there is no missing heritability. When only MRK information is used there may be missing heritability, but the actual extent is unknown because Σ_z, Σ_{zx} and Σ_x are unknown.

The genotypes used in the simulation were obtained from the type-2 diabetes case-control data set from the Nurses' Health Study and the Health Professionals Follow-up; both are part of the Gene-Environment Association Studies consortia (GENEVA, https://www.genevastudy.org/). We used only genotypes of

nominally unrelated individuals of Caucasian origin and with less than 5% missing genotypes. This left 5,000 individuals for the analysis.

The simulation setting was similar to that described in (de los Campos et al. 2013): from a set of 400K (K=1,000) SNPs, 300K loci were randomly chosen and designated as markers. From the remaining 100K SNPs, 5,000 were chosen and designated as QTL using a sampling method that over-sampled markers with low minor-allele frequency. We generated 1,000 MC replicates and in each MC replicate 2,500 individuals were randomly sampled and used for estimation of variance parameters.

For each MC replicate a G-BLUP model was fitted to the 2,500 records, using a G matrix computed from: QTL genotypes, MRK, and QTL+MRK. In the analysis using only QTL information the average estimated genomic heritability (.498) was very close to the population heritability (.5); the estimated 90% confidence interval ranged from .438 to .558. The analysis with markers only showed an average estimated genomic heritability of .328, suggesting an extent of missing heritability of 34%, similar to that reported by de los Campos et al. (2013) who analyzed data simulated with a similar but not identical scheme. The sampling variance of the estimator was very large, with a 90% CI for the estimated genomic heritability ranging from .136 to .517. This indicates that adding markers that are in imperfect LD with QTL not only induces missing heritability but also adds considerable uncertainty to estimates of variance components. Finally, the distribution of the estimated genomic heritability obtained with markers and QTL was similar to that obtained only with markers, but shifted to the right, with a mean equal to .411. In this scenario, as stated, there is no missing heritability; however the result indicates that likelihood estimates of genomic heritability based on the G-BLUP model can be biased.

Discussion

In the literature on genomic analysis of complex traits, genetic parameters have been commonly defined taking the model used for data analysis as starting point (Yang et al. 2010; Zaitlen and Kraft 2012; Speed et al. 2012). This approach has two potential problems. First, some of the assumptions of the statistical models used for data analysis are at odds with those in standard quantitative genetic theory (Falconer and Mackay 1996) making the connection between quantitative genetic parameters and statistical parameters tenuous. Secondly, because the patterns of allele sharing vary across the genome and because markers are typically in imperfect LD with QTL, marker-based models may largely misrepresent the underlying data generating process, leading, potentially, to important inferential problems (e.g., inconsistency).

A first contribution of this article is to provide **theory**, framed within the principles of quantitative genetics, that leads to precise definitions of parameters of marker-regressions at the population level. A few important results emerge from the definitions and derivations presented in this article.

Marker and QTL effects are fixed population quantities. Genetic and genomic variance stem from variation of allele content at QTL and at markers, respectively, and not due to uncertainty about QTL or marker effects. This is of course at odds with definitions of genomic variance based on Bayesian models where marker genotypes are treated as fixed and marker effects as random variables. From a Bayesian perspective it makes perfect sense to implement regressions conditioning on markers and with marker effects treated as random variables; however, **we question the use of** these **Bayesian models for definition of genetic parameters**.

Marker effects are linear combination of QTL effects. With high marker density multiple markers are likely to track variance from the same QTL. This **questions the treatment of marker effects as independent random variables**. For example, if, from a Bayesian perspective, QTL effects are treated as IID draws from a normal density, then it follows from expression [5] that marker effects are MVN distributed with null mean and covariance matrix $Cov(\beta) \propto \Sigma_x^{-1}\Sigma_{xz}\Sigma_{zx}\Sigma_x^{-1}$. Assuming that all covariances are null ignores the fact that multiple markers can track variance from the same QTL. Determining the correct covariance function is not possible because QTL positions are typically unknown. However attempts can be made to incorporate LD information in the prior density assigned to marker effects (e.g., Yang and Tempelman, 2012).

The recognition that marker effects are linear combinations of QTL effects has a second important consequence: **LD between markers plays a central role in the determination of genomic variances**. It is only under very idealized (and unrealistic) conditions that the total genomic variance can be decomposed into marker-specific components. On the other hand, in most of the parametric models used for data analysis the assumptions lead to a decomposition of the genomic variance that does not involve LD (Zaitlen and Kraft 2012). Depending on the patterns of LD between markers, ignoring LD may lead to under or over estimation of the genomic variance.

Estimation difficulties. Under regularity conditions, likelihood estimates are asymptotically unbiased (Lehmann and Casella 1998). However consistency requires that the likelihood is correctly specified. The proportion of allele sharing at any given set of loci can be viewed as a random variable with expected value given by twice the kinship coefficient between individuals (derived from a full pedigree) and random variation due to Mendelian sampling (Hill and Weir 2011). Because of imperfect LD between markers and QTL the proportion of allele sharing at markers and at QTL can be very different. This is particularly important for distantly related individuals (de los Campos et al. 2013). It follows that assessment of variance and covariance based on markers can misrepresent the true patterns of variance and covariance for a given trait. Under these conditions the likelihood can be misspecified leading to potential inconsistency of estimates.

There are two cases where the likelihood function will not be in error. The first one is when patterns of allele sharing at markers and at QTL are very similar. This will occur if markers are in tight LD with QTL, or with family

data because in this case markers and QTL co-segregate. Here, the instrumental model provides inferences about the true trait heritability. A second case will occur if the component of genetic values that cannot be explained by markers ($\xi_i$in expression [4]) are IID, and therefore $h_G^2 < h^2$ will be inferred consistently. However, there is no good reason to believe that the $\xi_i's$ are IID, and therefore it is not exactly clear if h_G^2 can be consistently estimated using G-BLUP.

Simulation studies using real human genotypes, including the one presented here, indicate that estimates of genomic heritability (h_G^2) based on a G-BLUP model incorporating both markers and QTL genotypes may be biased (e.g., Speed et al. 2012). Our simulation study indicates that problems are more serious when the patterns of LD vary strongly along the genome. Speed et al. (2012) argued that a main reason for inconsistency of G-BLUP estimates of heritability is that LD is ignored in the computation of the G matrix. These authors suggested an alternative method for computing G that resulted in estimates of genomic heritability closer to the simulated heritability of the trait. However, a follow up discussion (Lee et al. 2013; Speed et al. 2013) presented alternative simulations scenarios where the method of Speed et al. (2012) yielded biased estimates. All in all, this suggests that the appropriate choice of method for computing G depends on the genetic architecture of the trait. From our perspective, the main problem does not reside in the way G is computed, but rather in the use of massive numbers of markers that are in imperfect LD, and some in complete LE, with QTL.

On Genomic Analysis Using Whole-Genome Regressions

Complex traits are affected by large numbers of small-effect QTL. The analysis of such traits requires fitting large number of variants concurrently using the WGR approach originally proposed by (Meuwissen, Hayes, and Goddard 2001). Close relatives share long chromosome segments and, under these circumstances, the patterns of allele sharing at markers and at QTL are very similar. This leads to high prediction accuracy and very small bias in heritability estimates. On the other hand, with distantly related individuals, the addition of large numbers of markers that are in LE with QTL can lead to serious problems in the specification of genomic relationships. This can result in potential inconsistencies of estimates of genomic heritability. Importantly, this does not invalidate the use of WGR as a prediction machine. Rather, we warn about problems arising when these methods are used for inferences. We believe that this problem has been overlooked and that further research is needed to understand if and under what circumstances WGRs such as a G-BLUP model can be used to assess the true proportion of variance that can be explained by regression on markers in the population.

Acknowledgments. The authors wish to thank the participants of the GENEVA study. GDLC acknowledges financial support from NIH grants GM099992 and GM101219. DS acknowledges financial support from the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research. DG was supported by the Wisconsin Agricultural Experiment Station.

Literature Cited

de los Campos, G., and Sorensen, D.A. (2013). Nature Reviews Genetics 14: 894–894.

de los Campos, G., Vazquez, A.I., Fernando, R.L., et al. (2013). PLoS Genetics 9: e1003608.

Falconer, D.S. and Mackay, T.F.C. (1996). Introduction to Quantitative Genetics (4th Edition). Benjamin Cummings.

Gianola, D., de Los Campos G., Hill, W.G., et al. (2009). Genetics 183: 347.

Gianola, D., Hospital, F., and Verrier, E. (2013).Theoretical and Applied Genetics, 126: 1457–72.

Gianola, D. (2013). Genetics, 194:573-596.

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

Hill, W.G., and B.S. Weir. (2011). Genetics Research 93: 47–64.

Lango Allen, H., Estrada, K., Lettre, G., et al. (2010). Nature 467: 832–38.

Lee, S.H., Yang, J., Chen, G., et al. (2013). American Journal of Human Genetics 93: 1151–55.

Lehmann, E.L. and Casella, G. (1998). Theory of Point Estimation. Vol. 31. Springer.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.W. (2001). Genetics 157: 1819–29.

Speed, D., Hemani, G., Johnson, M.R. et al. (2012). The American Journal of Human Genetics 91:1011–21.

Speed, D., Hemani, G., Johnson, M.R. et al. (2013). The American Journal of Human Genetics 93:1155–57.

Yang, W. and Tempelman, R. (2012). Genetics 190:1491-150.

VanRaden, PM. (2008). J. Dairy Sci. 91: 4414–23.

Yang, J., B. Benyamin, B. P., McEvoy, S., et al. (2010). Nature Genetics 42: 565–69.

Zaitlen, N., and Kraft, P. (2012). Human Genetics 131: 1655–64.