Hill *Genetics Selection Evolution* 2013, **45**:32
http://www.gsejournal.org/content/45/1/32

GSE Genetics Selection Evolution

RESEARCH Open Access

On estimation of genetic variance within families using genome-wide identity-by-descent sharing

William G Hill

Abstract

Background: Traditionally, heritability and other genetic parameters are estimated from between-family variation. With the advent of dense genotyping, it is now possible to compute the proportion of the genome that is shared by pairs of sibs and thus undertake the estimation within families, thereby avoiding environmental covariances of family members. Formulae for the sampling variance of estimates have been derived previously for families with two sibs, which are relevant for humans, but sampling errors are large. In livestock and plants much larger families can be obtained, and simulation has shown sampling variances are then much smaller.

Methods: Based on the assumptions that realised relationship of sibs can be obtained from genomic data and that data are analyzed by restricted maximum likelihood, formulae were derived for the sampling variance of the estimates of genetic variance for arbitrary family sizes. The analysis used statistical differentiation, assuming the variance of relationships is small.

Results: The variance of the estimate of the additive genetic variance was approximately proportional to $1/(fn^2\sigma_R^2)$, for f families of size n and variance of relationships σ_R^2.

Conclusions: Because the standard error of the estimate of heritability decreased in proportion to family size, the use of within-family information becomes increasingly efficient as the family size increases. There are however, limitations, such as near complete confounding of additive and dominance variances in full sib families.

Background

Quantitative genetic parameters such as heritability have traditionally been estimated from the variation among full- or half-sib families, or from the parent-offspring covariance [1,2]. The covariance among sibs is assumed to be proportional to the pedigree relationship, but relatives may be further correlated because they share a common environment. This problem arises particularly in humans and, although sire families can be used in livestock to minimise the environmental covariance of sibs, these and weaker relationships come at the cost of higher sampling errors of heritability estimates because the correlation between sibs has to be multiplied by the inverse of the relationship to obtain an estimate of heritability. Estimates of heritability from non-pedigreed populations also rely heavily on getting good estimates of pedigree relationship [3], which is difficult unless relationships are very close, and environmental confounding can still a source of bias.

Correspondence: w.g.hill@ed.ac.uk
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

Although pairs of full-sibs, for example, share half their genome on average, individual pairs do not because of Mendelian sampling of large chromosome segments. Such a discrepancy at pairs of loci is the basis of QTL (quantitative trait locus) mapping using, for example, the method of Haseman and Elston [4], to associate the phenotypic divergence between sibs to differences in marker frequency. Dense marker genomes are now available, and Visscher et al. [5] proposed that the actual or realised relationships between sibs can be estimated from genomic data and the association between the actual relationship and phenotypic similarity used to estimate the genetic covariance within families, thereby eliminating correlations due to shared environment. Visscher and colleagues used data on human dizygotic twins and full-sibs, first from microsatellites [5] and subsequently from SNPs (single nucleotide polymorphisms) [6] to estimate the level of genome sharing and thus trait heritability. In a later paper, Visscher [7] discussed the theory further. However, the sampling error of the estimates of genetic variance was high because the variation in actual relationship was

small (typical standard deviation (SD) of 3.9% of the mean of 50% for human full-sibs, as expected from theory [5,7-10]). Since family sizes in humans are also very small, many are needed for precise estimation.

Ødegård and Meuwissen [11] pointed out that the method of Visscher et al. [5] could be used in very large families, such as for fish species, and for which it is not always practical to avoid rearing full-sibs together. They showed by simulation that sampling errors of the resulting estimates of heritability are substantially reduced as family size increases and are smaller with a few large families than with many small families. These results raise the following basic question: for a family of n sibs, is the information content, i.e. the inverse of the sampling variance of the estimate of heritability, approximately proportional to family size n (or e.g. to n -1) or to the number of pairs in the family, ½n(n -1)? The simulation results of Ødegård and Meuwissen [11] indicated the latter. Furthermore, PM Visscher (personal communication) showed that, using genomic relationships estimated from a sample of N individuals from the population, the sampling variance is a function of N^2. The difference between methods with sampling variances that depend on approximately squares of numbers rather than numbers of individuals is not trivial and clearly has an important impact on their design and potential utility.

The model used by Ødegård and Meuwissen [11] was based on a finite number (80) of genomic blocks that were individually marked, and with trait effects that were identically normally distributed for each block. In this note, we quantify these estimates and show how they depend on the design and variation in realised relationships. We adopt a model in which the realised relationship is continuous over the genome and with trait effects that are uniformly distributed across the genome. To calculate sampling errors, Visscher et al. [5] used regression of the squared phenotypic difference of sibs on the estimated actual relationship from tracking genome segments, whereas Ødegård and Meuwissen [11] used a REML (restricted maximum likelihood) analysis within and between families with estimated realised relationships for a finite number of genome segments. In the present analysis the data were assumed to be analysed by REML. Implications for design of experiments are discussed.

Analysis

Let us assume that the data are from matings of unrelated individuals and comprise f (≥ 1) families each of size n (≥ 2). The extension to variable n is straightforward and deferred meanwhile. The mean (i.e. pedigree) numerator relationship within families is A (e.g. 0.25 for half-sibs or 0.5 for full-sibs) and the within-family variance of actual relationships is σ^2_R. We also assume that all sibs share the same environment and, for simplicity, as in the work of Visscher et al. [5,6], that additive genetic variance is estimated using only within-family differences; in essence, family effects are regarded as fixed. Therefore information is accumulated independently across families and no bias or sampling error arises due to common environment, albeit at the cost of losing potential between-family genetic information.

Additive model

Initially, we assumed that gene effects were additive but subsequently extended the results to include dominance. The additive genetic variance is σ^2_A, the residual environmental variance is σ^2_E, and so the within-family variance is $\sigma^2_W = (1 - A)\sigma^2_A + \sigma^2_E$. The phenotypic variance is given by $\sigma^2_P = A\sigma^2_A + \sigma^2_C + \sigma^2_W$, where σ^2_C is the variance due to common environment. In the analysis, it is convenient to parameterise the actual relationship between family members i and j in terms of deviations from mean pedigree-based relationships: $r_{ij} = A_{ij} - A$. The $n \times n$ covariance matrix $\mathbf{V}$ of observations $\mathbf{y}$ within a family of n sibs is then $\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathbf{I}\sigma^2_W + \mathbf{R}\sigma^2_A$, where $\mathbf{I}$ is the identity matrix and elements of $\mathbf{R}$ are r_{ij}, $i \neq j$, and $r_{ii} = 0$.

The sampling variance of the parameter estimates can be approximated by using a Taylor series expansion in r_{ij} because these deviations are small, and then taking expectations so as to obtain Fisher's information matrix $\mathbf{S}$ (the inverse of the variance covariance matrix) for the REML estimates of variance components $\hat{\sigma}^2_A$ and $\hat{\sigma}^2_W$, respectively. The derivation is rather complicated, so details are given in Appendix 1. For a family of size n it is shown that:

$$\mathbf{S} = \frac{n-1}{2\sigma^4_W}\begin{pmatrix} m\sigma^2_R & -2m\sigma^2_R\sigma^2_A/\sigma^2_W \\ -2m\sigma^2_R\sigma^2_A/\sigma^2_W & 1+3m\sigma^2_R\sigma^4_A/\sigma^4_W \end{pmatrix} \quad (1)$$

where $m = n(1 - 2/n + 2/n^2)$. Since between-family relationships are not used, information $\mathbf{S}_k$ from family k is merely summed over families, with corresponding elements for family size n_k and m_k, $k = 1, \ldots, f$. The overall variance-covariance matrix of the estimates is:

$$\mathbf{C} = \begin{pmatrix} \mathrm{var}(\hat{\sigma}^2_A) & \mathrm{cov}(\hat{\sigma}^2_A, \hat{\sigma}^2_W) \\ \mathrm{cov}(\hat{\sigma}^2_A, \hat{\sigma}^2_W) & \mathrm{var}(\hat{\sigma}^2_W) \end{pmatrix} = \left(\sum\nolimits_k \mathbf{S}_k\right)^{-1}$$

With f families of equal size, from (1):

$$\mathbf{C} = \left(\frac{2\sigma^4_W}{f(n-1)m\sigma^2_R(1-m\sigma^2_R\sigma^4_A/\sigma^4_W)}\right) \times \begin{pmatrix} 1+3m\sigma^2_R\sigma^4_A/\sigma^4_W & 2m\sigma^2_R\sigma^2_A/\sigma^2_W \\ 2m\sigma^2_R\sigma^2_A/\sigma^2_W & m\sigma^2_R \end{pmatrix} \quad (2)$$

The estimate of the environmental variance is $\widehat{\sigma}^2_E = \widehat{\sigma}^2_W - \frac{1}{2}\widehat{\sigma}^2_A$ and hence $\mathrm{var}(\widehat{\sigma}^2_E) = c_{22} - c_{12} + \frac{1}{4}c_{11}$ and cov

$\left(\hat{\sigma}^2_A, \hat{\sigma}^2_E\right) = c_{12} - \frac{1}{2}c_{11}$, where c_{ij} are elements of **C**. Taking just σ^2_A and σ^2_E into account, $\hat{\sigma}^2_P = \hat{\sigma}^2_A + \hat{\sigma}^2_W$, and the sampling error of the corresponding heritability estimate, $\hat{h}^2 = \hat{\sigma}^2_A / \hat{\sigma}^2_P$, can be approximated using standard formulae for ratios (see e.g. page 818 in [2]). Between-family information, not included in the data used above, has to be incorporated to estimate the phenotypic variance and heritability if common family environment or allowance for non-additive effects is to be included.

If the quantity $m\sigma^2_R\sigma^4_A/\sigma^4_W$ is small, the determinant of **S** is dominated by its diagonal elements and var $(\hat{\sigma}^2_A)$ simplifies to:

$$\text{var}\left(\hat{\sigma}^2_A\right) \approx 1/s_{11} = 2\sigma^4_W / \left[f(n-1)m\sigma^2_R\right] \quad (3)$$

Hence for families of $n = 2$ individuals, $m = 1$ and var $\left(\hat{\sigma}^2_A\right) \approx 2\sigma^4_W / \left(f\sigma^2_R\right)$. This corresponds to the formula of Visscher et al. [5] for the sampling error of the heritability estimate: $2(1-t)^2 / \left(f\sigma^2_R\right)$, where t is the intra-class correlation of family members. As n increases, $m(n - 1) = n(n - 3 + 2/n - 2/n^2) \rightarrow n(n - 3) \rightarrow n^2$. If σ^2_R is small and n large, then var$(\hat{\sigma}^2_A) \sim 2\sigma^4_W / \left(fn^2\sigma^2_R\right)$.

The variation in relationships within a family depends on whether family members are full- or half-sibs, on the total map length (L) of the chromosomes and, to a limited extent, on their individual lengths [5,7,10]. To a good approximation, $\sigma^2_R \sim 1/(16L) - 1/(3L^2)$ for full-sibs and one-half of that for half-sibs [5,7]. For humans, the number of autosomes is 22 and the total map length is 35.9 M, so σ^2_R is approximately 0.00153 for full-sibs and 0.00077 for half-sibs (SD = 0.039 and 0.028). Therefore, for full-sib families of a species with a map length and chromosome number similar to humans, SE$(\hat{\sigma}^2_A) \sim 36\ \sigma^2_W / \left[\sqrt{fn(n-3)}\right.$, e.g. 0.28 σ^2_W for 50 families of size 20 and 0.17 σ^2_W for 20 families of size 50. Cattle, for example, have 29 autosomes and a map length of 32.5 M [12], so σ^2_R would be a little larger and the sampling variance of estimates of heritability correspondingly smaller.

Simulation check on approximations

In the analysis in Appendix 1, many simplifying assumptions were made in the Taylor series analysis. As a partial check, simulation was undertaken for a model of 22 chromosomes, each 1.632 M long, i.e. the mean length of human chromosomes, and relationships were simulated with the programme used previously to check formulae for variance in relationships [10]. (The distribution of relationships would be little affected if map lengths varied [10]). The information matrix **S** was then computed directly from equation (A1) and from the approximation in Equation (1). For simplicity, however, it was assumed that the contrast matrix **K** (see below equation (A1)) was invariant (see examples in Table 1). In general, there was good agreement between the observed and the approximate predicted estimates of sampling variance (Table 1), but this deteriorated as family size increased, with the approximation generally underestimating the sampling variance. This bias would be greater if σ^2_R were higher. Although, if only a single chromosome was fitted σ^2_R would be much greater, the additive variance contributed by it would be only a fraction of the total and, as the example in Table 1 shows, the approximation remains good. Table 1 also gives predictions based solely on Equation (3), showing a good fit with those obtained directly from Equation (2).

Dominance

In full-sib families, both additive and dominance variance can, in principle, be estimated. Derivation of the extended information matrix is given in Appendix 2. It depends on the variance σ^2_Q in dominance relationships (about its mean of ¼) and the covariance between dominance and additive relationships, cov_{RQ}. However, as Visscher et al. [5] pointed out, the additive and dominance relationships within families are very highly correlated, since the additive coefficient depends on the average number of paternal and maternal genes that are shared identical by descent at a locus and the dominance coefficient on whether both are shared. The regression of dominance on additive relationships (cov_{RQ} / σ^2_R) is equal to 1 and their correlation is approximately 0.9. This implies that, in practice, partitioning σ^2_A and σ^2_D using within-family information is probably not feasible and furthermore that if only an additive model is used, the estimate of σ^2_A is biased upwards by σ^2_D; indeed it essentially has expectation $\sigma^2_A + \sigma^2_D$.

Discussion and conclusions

The analysis shows that the sampling variances of estimates of heritability based on within-family realized relationships fall roughly in proportion to n^2 as family size n increases, i.e. based on the number of pairwise comparisons among individuals in the family, and in proportion to the number of families. Therefore, when undertaking such an analysis, it is more efficient to use few very large families, although one might be reluctant to use just one or very few families in case they are atypical [11]. Here, a model of a continuous genome was used, rather than a finite number of independent regions as by Ødegård and Meuwissen [11], and the calculations assumed a fairly even distribution of genetic variance along the genome. If there is much heterogeneity, e.g. a few QTL of large effect, the sampling errors of genetic variance estimates would increase. In the present analysis, we make the assumption that shared segments are identified accurately, for example using Merlin [13].

Table 1 Comparison of var($\hat{\sigma}^2_A$) predicted from the information matrix directly and from the Taylor series approximation*

Family	HS			FS			FS			FS			FS		
h^2	0.5			0.25			0.5			0.75			0.04		
chr	22			22			22			22			1		
n	5	15	25	5	15	25	5	15	25	5	15	25	5	15	25
Eq (A1)	174	12.2	4.15	94	6.67	2.26	82.6	5.94	2.04	71.4	5.20	1.81	4.80	0.354	0.127
Eq (1)	182	11.8	3.88	101	6.56	2.18	88.6	5.83	1.97	77.1	5.23	1.82	5.26	0.331	0.110
Eq (3)	182	11.7	3.80	101	6.53	2.16	88.1	5.69	1.88	76.0	4.90	1.62	5.26	0.330	0.110

*Predictions were obtained directly by inverting the realised information matrix (eq A1) obtained from sampling relationships, and from the Taylor series approximation eq. (1) using the variance of relationships directly; variances were computed by averaging information over samples of 100 families, but are expressed for a single family, so for *f* families var($\hat{\sigma}^2_A$) should be divided by *f*; predictions using the simplification eq. (3) are shown similarly; results are for half (HS) and full (FS) sib families; h^2 is the proportion of variance contributed by the fitted chromosomes; chr is the number of chromosomes; chr = 22 denotes the whole genome; chr = 1 denotes a single chromosome.

Ødegård and Meuwissen [11] investigated the effect of selectively genotyping only the individuals with high and low phenotypes within a family, when all phenotypes are included in the REML analysis. The efficiency of this approach was good in terms of sampling errors but estimates of heritability were biased downwards when sample sizes were small. This may reflect insufficient marker coverage of the genes of interest because of lack of linkage disequilibrium, in which case this bias may be hard to avoid, but possibly also bias caused by selection.

They also estimated actual relationships from a finite number of markers and, occasionally, obtained a singular matrix in their simulated replicates [11]. To check the causes, simulated relationships were sampled from a continuous chromosome model [10] and the exact allele sharing was computed. Pairs of individuals can inherit identical non-recombinant short chromosomes, thereby yielding a positive semi-definite relationship matrix (i.e. including zero but not negative eigenvalues). In the unlikely event that this occurs at all chromosomes, the data can still be analysed by REML. Negative eigenvalues were not obtained in our simulations and indeed seem infeasible, because the relationships were jointly sampled. Negative eigenvalues are a consequence of the estimation of weak relationships from marker data and might arise in practice.

A different approach to estimating the genetic variance free of common environment was suggested by Yang et al. [14]. They fitted by regression all the SNPs to data from individuals sampled from the population that are not known to be related and from which any pairs with a relationship above a low threshold have been removed, so as to minimise the chance of shared environment. Such an analysis is expected to give a lower estimate of heritability than the within-family analysis discussed here, however, because marker-associated effects in the population can be missed through incomplete linkage disequilibrium, especially when traits genes have low minor allele frequencies, as indeed seems to be the case [14].

A 'back of the envelope' calculation allows a simple comparison of the sampling errors of estimates of additive genetic variance from within families utilising variation in relationship, $\hat{\sigma}^2_{\mathrm{Aw}}$, and from between families using ANOVA, $\hat{\sigma}^2_{\mathrm{Ab}}$ (Appendix 3). Provided the families are not small, $\mathrm{var}(\hat{\sigma}^2_{\mathrm{Aw}})/\mathrm{var}(\hat{\sigma}^2_{Ab}) \approx (A^2/\sigma^2_{\mathrm{R}})/[1 + nA\sigma^2_{\mathrm{A}}/\sigma^2_{\mathrm{W}}]^2$. With use of half-sib families ($A = 1/4$) to eliminate maternal effects in the between-family estimate, for a genome of 'human' length, $(A^2/\sigma^2_R) = (0.25/0.028)^2 \sim 80$. Assuming the heritability is 1/3, such that $A\sigma^2_{\mathrm{A}} = \frac{1}{5}\sigma^2_{\mathrm{W}}$, the ratio of variances is approximately $80/(1 + n/5)^2$, equalling 1.0 when $n \sim 40$. This implies that, with half-sib families of size 40, a similar amount of information would be obtained from within- and between-family data. With fewer larger families, the estimate from within-family information would have the lower standard error. Furthermore, because the within- and between-family estimates use the data in a different way they are, presumably, uncorrelated and so they can be simply combined. However, estimates from both sources may be biased to different extents by common environment, dominance, epistasis, etc., so specific applications require specific consideration.

There are other aspects that could be examined. For example, additive and within-family genetic covariances and correlations among traits can be estimated from a multi-trait analysis with the same data structure. Clearly the magnitude of their sampling errors is structured similarly to those of the corresponding variances of the individual traits. Estimation of variation due to any individual autosome can be achieved by fitting just the relationship on this chromosome, and similarly for the sex chromosome [6]. The variance of the corresponding relationships is then much higher and depends on the length of the chromosome, decreasing roughly in proportion to its length. Although $\mathrm{var}(\hat{\sigma}^2_{\mathrm{A}})$ per chromosome is then much smaller, the coefficient of variation of its estimate may be similar to that for the whole genome under the simplest assumption that the contribution by any chromosome to σ^2_{A} is roughly proportional to its length.

A problem specific to the within-family approach is the high degree of confounding between additive and

dominance effects in full-sib families (albeit there is also complete confounding in a between full-sib family analysis). This is not resolved by estimating σ^2_A separately from maternal and paternal sharing, since the dominance coefficient is the correlated intersection of these. The point is that, while maternal genomic similarity appears to include only the additive component because only one sire is involved, interactions between sire and dam effects, i.e. dominance, are included. Half-sib families with multiple dams per sire or a cross classified structure are needed, similar to when between-family correlations are used for estimation.

If, for example, a number of males and females are put together for mating in a single environment, then the pedigree can be obtained from genetic markers. Hence, paternal half-sibs, maternal half-sibs and full-sibs can be distinguished and the between-family covariance can be used. Additional information from within-family segregation could be identified via the markers, but this would likely contribute little. For example, in a pen comprising such a diallel structure, the variation in pedigree relationships ($A = 0$, ¼ or ½) is likely to be much larger than the variation in realised relationships among pairs with the same pedigree relationship.

Epistatic variance provides other associated difficulties of potential confounding and estimation. On a whole-genome basis, the relevant coefficient for the additive × additive variance component is the square of the relationship, which is highly correlated with the additive coefficient. Thus, similar to the analyses between families, obtaining a satisfactory partition between additive and additive × additive or higher order components is probably not feasible. A further problem is potential bias due to epistatic effects in the estimation of additive (e.g. from additive × additive effects) and dominance variance. Although the expected probability that sibs share alleles at pairs of genomic sites is small for the genome as a whole, it is much higher for nearby sites. Thus, if epistatic effects are substantial and predominately cis-acting, this bias could be important. To partially address this, Visscher et al. [6] fitted the mean relationship for each chromosome in a multiple regression model for human height. The variance removed by fitting variation in relationships for each chromosome was essentially the same whether chromosomes were fitted independently or in a joint analysis, indicating little or no interaction between regions on different chromosomes. Extending this more generally needs genomic regions to be defined such that joint identity by descent can be computed.

Within-family analysis, particularly when families are large, has attractive features because. it avoids bias due to common environment effects, but it introduces other potential confounding effects, as noted above. It also requires much genotyping and associated costs. Although in a breeding context this type of information may be available when collecting data to implement genomic prediction and subsequent selection, estimates of the variance components may not in themselves have value beyond what is obtained from the marker trait associations. But this is something to think about.

Appendix 1: Derivation of the sampling variance for the additive model

For the REML analysis, the information matrix **S**, which in turn yields the sampling variances based on $\mathbf{S}^{-1}$ for the estimates of σ^2_A and σ^2_W for each family, is defined by Lynch and Walsh (see page 791 in [2]):

$$\mathbf{S} = \frac{1}{2}\begin{pmatrix} \mathrm{tr}(\mathbf{PRPR}) & \mathrm{tr}(\mathbf{PRP} \\ \mathrm{tr}(\mathbf{PRP}) & \mathrm{tr}(\mathbf{PP}) \end{pmatrix}, \tag{A1}$$

where tr denotes the trace operator. Matrix $\mathbf{P} = \mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}$ and $\mathbf{K}_{(n-1)\times n}$ defines contrasts such that $\mathbf{KX} = \mathbf{0}$, where **X** is the design matrix and, since family members are contemporaneous in the same environment, **X** is a unit vector. The Helmert contrasts are suitable for **K**: for $i = 1, \ldots, n-1$: $k_{ij} = [(i(i+1)]^{-1/2}$, $j \le i$; $k_{i,i+1} = -[(i/(i+1)]^{1/2}$ and $k_{ij} = 0$, $j > i+1$. Note that $\mathbf{KK}' = \mathbf{I}_{(n-1)\times(n-1)}$ and $\mathbf{K}'\mathbf{K} = \mathbf{I}_{n\times n} - \frac{1}{n}\mathbf{J}_{n\times n}$, where all elements of **J** equal 1, and $(\mathbf{K}'\mathbf{K})^2 = \mathbf{K}'\mathbf{K}$.

The expected information using the Taylor series expansion has terms of the following form:

$$\mathrm{E}(\mathbf{PRPR}) = \mathbf{PRPR}_{|\mathbf{R}=\mathbf{0}} + \sum_{i\le j}\partial(\mathbf{PRPR})/\partial r_{ij|\mathbf{R}=\mathbf{0}}\mathrm{E}(r_{ij}) + \frac{1}{2}\sum_{i\le j}\sum_{k\le l}\partial^2(\mathbf{PRPR})/\partial r_{ij}\partial r_{kl|\mathbf{R}=\mathbf{0}}\mathrm{E}(r_{ij}r_{kl}) + \ldots$$

We note that $\mathrm{E}(r_{ij}) = 0$ and, assuming independent Mendelian segregation to each offspring, $\mathrm{E}(r_{ij}r_{kl}) = 0$, $i \ne k$ and/or $j \ne l$ and $\mathrm{E}(r_{ij})^2 = \sigma^2_R$, where σ^2_R is the variance in relationship. Differentiating

$$\frac{\partial(\mathbf{PRPR})}{\partial r_{ij}} = \frac{\partial\mathbf{P}}{\partial r_{ij}}\mathbf{RPR} + \mathbf{P}\frac{\partial\mathbf{R}}{\partial r_{ij}}\mathbf{PR} + \mathbf{PR}\frac{\partial\mathbf{P}}{\partial r_{ij}}\mathbf{R} + \mathbf{RPR}\frac{\partial\mathbf{R}}{\partial r_{ij}}, \tag{A2}$$

and when evaluated as $\mathbf{R} \to \mathbf{0}$, all terms in (A2) become zero. Furthermore, differentiating (A2) to obtain the second derivative, all remaining terms in **R** are also zero; and as **R** is linear in r_{ij}, $\partial^2\mathbf{R}/\partial r_{ij}\partial r_{kl} = \mathbf{0}$. Finally, as $\mathrm{E}(r_{ij}r_{kl}) = 0$ unless $i = k$ and $j = l$, $\mathrm{E}(\mathbf{PRPR})$ reduces to

$$\mathrm{E}(\mathbf{PRPR}) \approx \frac{1}{2}\sum_{i<j}\left(\mathbf{P}\frac{\partial\mathbf{R}}{\partial r_{ij}}\mathbf{P}\frac{\partial\mathbf{R}}{\partial r_{ij}} + \frac{\partial\mathbf{R}}{\partial r_{ij}}\mathbf{P}\frac{\partial\mathbf{R}}{\partial r_{ij}}\mathbf{P}\right)\sigma^2_R.$$

Let $\partial\mathbf{R}/r_{ij} = \mathbf{X}_{ij}$, with elements $x_{ij} = x_{ji} = 1$ and 0 otherwise; so taking $\mathbf{R} \to \mathbf{0}$,

$$\mathrm{E}(\mathbf{PRPR}) \approx \frac{1}{2}\sum_{i<j}\left(\mathbf{PX}_{ij}\mathbf{PX}_{ij} + \mathbf{X}_{ij}\mathbf{PX}_{ij}\mathbf{P}\right)\sigma^2_R \tag{A3}$$

As $\mathbf{R} \to \mathbf{0}, \mathbf{V} \to \mathbf{P} = \mathbf{K}'(\mathbf{KVK}')^{-1}\,\mathbf{K} \to (\mathbf{I} - \frac{1}{n}\mathbf{J})/\sigma^2_{\mathrm{W}}$. Defining further matrices, $\mathbf{Y}_{ij}$ where $y_{ii} = y_{jj} = 1$ and 0 otherwise, and $\mathbf{W}_{ij}$ where $w_{ik} = w_{jk} = 1$, $k = 1, \ldots, n$, and 0 otherwise, we have $\mathbf{X}_{ij}\ \mathbf{X}_{ij} = \mathbf{Y}_{ij}$, $\mathbf{JX}_{ij} = \mathbf{JY}_{ij} = \mathbf{W}_{ij}$, $\mathbf{W}_{ij}\mathbf{W}_{ij} = 2\mathbf{W}_{ij}$, and $\mathrm{tr}(\mathbf{X}_{ij}) = 0$, $\mathrm{tr}(\mathbf{Y}_{ij}) = \mathrm{tr}(\mathbf{W}_{ij}) = 2$. As the trace operator is commutative, it follows that by summing over the $n(n-1)/2$ off diagonal elements in (A3), all having the same expectation,

$$\begin{aligned}\mathrm{E}[\mathrm{tr}(\mathbf{PRPR})] &\approx \tfrac{1}{2}n(n-1)\mathrm{tr}\left[(\mathbf{I}-\mathbf{J}/n)\mathbf{X}_{ij}(\mathbf{I}-\mathbf{J}/n)\mathbf{X}_{ij}\right]\sigma^2_{\mathrm{R}}/\sigma^4_{\mathrm{W}}\\ &\approx \frac{1}{2}n(n-1)\mathrm{tr}\left(\mathbf{Y}_{ij}-2\mathbf{W}_{ij}/n+2\mathbf{W}_{ij}/n^2\right)\sigma^2_{\mathrm{R}}/\sigma^4_{\mathrm{W}}\\ &\approx n(n-1)\left(1-2/n+2/n^2\right)\sigma^2_{\mathrm{R}}/\sigma^4_{\mathrm{W}} = (n-1)m\sigma^2_{\mathrm{R}}/\sigma^4_{\mathrm{W}}\end{aligned} \tag{A4}$$

where $m = n(1 - 2/n + 2/n^2)$.

We give less detail for other terms in the information matrix.

$$\frac{\partial(\mathbf{PRP})}{\partial r_{ij}} = \frac{\partial \mathbf{P}}{\partial r_{ij}}\mathbf{RP} + \mathbf{P}\frac{\partial \mathbf{R}}{\partial r_{ij}}\mathbf{P} + \mathbf{PR}\frac{\partial \mathbf{P}}{\partial r_{ij}}.$$

Non-zero second derivatives must involve differentiation once of **P** and once of **R**. Hence

$$\begin{aligned}&\mathrm{E}(\mathbf{PRP}) \approx \frac{1}{2}\sum_{i<j}\left(2\frac{\partial \mathbf{P}}{\partial r_{ij}}\frac{\partial \mathbf{R}}{\partial r_{ij}}\mathbf{P} + 2\mathbf{P}\frac{\partial \mathbf{R}}{\partial r_{ij}}\frac{\partial \mathbf{P}}{\partial r_{ij}}\right)\sigma^2_R\\ &\frac{\partial \mathbf{P}}{\partial r_{ij}} = -\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K} \text{ and, as } \mathbf{R}\to\mathbf{0},\\ &\frac{\partial \mathbf{P}}{\partial r_{ij}} \to -\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{X}_{ij}\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\sigma^2_{\mathrm{A}}/\sigma^6_{\mathrm{W}}, \text{ so}\\ &\mathrm{E}(\mathbf{PRP}) \approx \frac{1}{2}\sum_{i<j} -4\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{X}_{ij}\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{X}_{ij}\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\sigma^2_{\mathrm{R}}\sigma^2_{\mathrm{A}}/\sigma^6_{\mathrm{W}}\end{aligned} \tag{A5}$$

As the trace is commutative and $\mathbf{I} - \frac{1}{n}\mathbf{J}$ is idempotent, putting the last such matrix in (A5) first, we see that:

$$\begin{aligned}\mathrm{E}[\mathrm{tr}(\mathbf{PRP})] &\approx -2n(n-1)\left(1-2/n+2/n^2\right)\sigma^2_{\mathrm{A}}\sigma^2_{\mathrm{R}}/\sigma^6_{\mathrm{W}}\\ &= -2(n-1)m\ \ \sigma^2_{\mathrm{A}}\sigma^2_{\mathrm{R}}/\sigma^6_{\mathrm{W}}.\end{aligned}$$

When $\mathbf{R} = \mathbf{0}$, $\mathbf{P} = (1 - 1/n)/\sigma^2_{\mathrm{W}}$ and $\mathrm{tr}(\mathbf{PP}) = (n-1)/\sigma^4_{\mathrm{W}}$. Now considering the terms in r_{ij},

$$\frac{\partial^2(\mathbf{PP})}{\partial r^2_{ij}} = \frac{\partial^2 \mathbf{P}}{\partial r^2_{ij}}\mathbf{P} + 2\frac{\partial \mathbf{P}}{\partial r_{ij}}\frac{\partial \mathbf{P}}{\partial r_{ij}} + \mathbf{P}\frac{\partial^2 \mathbf{P}}{\partial r^2_{ij}} \tag{A6}$$

with additional terms that become 0 as $\mathbf{R} \to \mathbf{0}$.

In (A6) $\frac{\partial \mathbf{P}}{\partial r_{ij}} = -\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}$

$$\frac{\partial^2 \mathbf{P}}{\partial r^2_{ij}} = 2\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}$$

And hence, using the commutative property,

$$\begin{aligned}\mathrm{tr}\left(\frac{\partial^2(\mathbf{PP})}{\partial r^2_{ij}}\right) &= 6\mathrm{tr}\left(\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial r_{ij}}\mathbf{K}'(\mathbf{KVK}')^{-1}\mathbf{K}\right)\\ &= 6\mathrm{tr}\left((\mathbf{I}-\tfrac{1}{n}\mathbf{J})\mathbf{X}_{ij}(\mathbf{I}-\tfrac{1}{n}\mathbf{J})\mathbf{X}_{ij}(\mathbf{I}-\tfrac{1}{n}\mathbf{J})\right)\sigma^4_{\mathrm{A}}/\sigma^8_{\mathrm{W}}.\end{aligned}$$

Therefore, using previous results,

$$\mathrm{E}[\mathrm{tr}(\mathbf{PP})] \approx (n-1)/\sigma^4_{\mathrm{W}} + 3nm\ \ \sigma^2_{\mathrm{R}}\sigma^4_{\mathrm{A}}/\sigma^8_{\mathrm{W}}$$

thus completing the derivation of the information matrix in Equation (1) of the main text.

Appendix 2: Fitting additive and dominance variances

Let $\mathbf{V} = \mathbf{I}\sigma^2_{\mathrm{w}} + \mathbf{R}\sigma^2_{\mathrm{A}} + \mathbf{Q}\sigma^2_{\mathrm{D}}$ of dimension $n \times n$, where, for full sib families, $\sigma^2_{\mathrm{W}} = \sigma^2_{\mathrm{E}} + \frac{1}{2}\sigma^2_{\mathrm{A}} + \frac{3}{4}\sigma^2_{\mathrm{D}}$. Additive and dominance effects of the loci are assumed to be uncorrelated. Let **Q** with elements q_{ij} define the departure of the realised dominance correlation of full sibs from the expected ¼, and let σ^2_{Q} denote var(q_{ij}) and similarly cov_{RQ} denote cov(r_{ij}, q_{ij}). The information matrix is now [2]:

$$S = \frac{1}{2}\begin{pmatrix}\mathrm{tr}(\mathbf{PRPR}) & \mathrm{tr}(\mathbf{PRPQ}) & \mathrm{tr}(\mathbf{PRP}\\ \mathrm{tr}(\mathbf{PRPQ}) & \mathrm{tr}(\mathbf{PQPQ}) & \mathrm{tr}(\mathbf{PQP})\\ \mathrm{tr}(\mathbf{PRP}) & \mathrm{tr}(\mathbf{PQP}) & \mathrm{tr}(\mathbf{PP})\end{pmatrix}.$$

The term $\mathrm{E}[\mathrm{tr}(\mathbf{PRPR})] \approx (n\text{ -}1)m\,\sigma^2_{\mathrm{R}}/\sigma^4_{\mathrm{W}}$ is unchanged from the additive case and, by symmetry,

$$\mathrm{E}[\mathrm{tr}(\mathbf{PQPQ})] \approx (n-1)m\sigma^2_{\mathrm{Q}}/\sigma^4_{\mathrm{W}} \text{ and}$$
$$\mathrm{E}[\mathrm{tr}(\mathbf{PRPQ})] \approx (n-1)mcov_{\mathrm{RQ}}/\sigma^4_{\mathrm{W}}.$$

The derivative of the term **PRP** with respect to r_{ij} remains

$$\frac{\partial(\mathbf{PRP})}{\partial r_{ij}} = \frac{\partial \mathbf{P}}{\partial r_{ij}}\mathbf{RP} + \mathbf{P}\frac{\partial \mathbf{R}}{\partial r_{ij}}\mathbf{P} + \mathbf{PR}\frac{\partial \mathbf{P}}{\partial r_{ij}},$$

and the expectation of its second derivative with respect to r_{ij} is unchanged. However, now taking the second derivative with respect to q_{ij}, we obtain additional terms with non zero expectation,

$$\frac{\partial^2(\mathbf{PRP})}{\partial r_{ij}\partial q_{ij}} = \frac{\partial \mathbf{P}}{\partial q_{ij}}\frac{\partial \mathbf{R}}{\partial r_{ij}}\mathbf{P} + \mathbf{P}\frac{\partial \mathbf{R}}{\partial r_{ij}}\frac{\partial \mathbf{P}}{\partial q_{ij}}.$$

Hence $\mathrm{E}[\mathrm{tr}(\mathbf{PRP})] \approx -2(n-1)m\left(\sigma^2_{\mathrm{R}}\sigma^2_{\mathrm{A}} + cov_{\mathrm{RQ}}\sigma^2_{\mathrm{D}}\right)/\sigma^4_{\mathrm{W}}$,

and similarly

$\mathrm{E}[\mathrm{tr}(\mathbf{PQP})] \approx -2(n-1)m\left(cov_{\mathrm{RQ}}\sigma^2_{\mathrm{A}} + \sigma^2_{\mathrm{Q}}\sigma^2_{\mathrm{D}}\right)/\sigma^4_{\mathrm{W}}$. The term $\mathrm{E}[\mathrm{tr}(\mathbf{PP})]$ is non-zero when differentiated twice

with respect to r_{ij} and to q_{ij} and once each with both variables. Hence

$$\mathrm{E}[\mathrm{tr}(\mathbf{PP})]\approx(n-1)/\sigma_{\mathrm{W}}^4 + 3(n-1)m\left(\sigma_{\mathrm{R}}^2\sigma_{\mathrm{A}}^4 + 2cov_{\mathrm{RQ}}\sigma_{\mathrm{A}}^2\sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{Q}}^2\sigma_{\mathrm{D}}^4\right)/\sigma_{\mathrm{W}}^8.$$

The information matrix for a single family is therefore

$$\mathbf{S} = \left(\frac{n-1}{2\sigma_{\mathrm{W}}^4}\right) \times \begin{pmatrix} m\sigma_{\mathrm{R}}^2 & m\,\mathrm{cov}_{\mathrm{RQ}} & -2m\left(\sigma_{\mathrm{R}}^2\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{D}}^2 cov_{\mathrm{RQ}}\right)/\sigma_{\mathrm{W}}^2 \\ & m\sigma_{\mathrm{Q}}^2 & -2m\left(cov_{\mathrm{RQ}}\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{Q}}^2\sigma_{\mathrm{D}}^2\right)/\sigma_{\mathrm{W}}^2 \\ \mathrm{symm} & & 1+3m\left(\sigma_{\mathrm{R}}^2\sigma_{\mathrm{A}}^4 + 2cov_{\mathrm{RQ}}\sigma_{\mathrm{A}}^2\sigma_{\mathrm{D}}^2 + \sigma_{\mathrm{Q}}^2\sigma_{\mathrm{D}}^4\right)/\sigma_{\mathrm{W}}^4 \end{pmatrix}.$$

These equations apply to estimates of $\widehat{\sigma}_{\mathrm{A}}^2$, $\hat{\sigma}_{\mathrm{D}}^2$ and $\hat{\sigma}_{\mathrm{W}}^2$. For full sib families, the estimate of the error variance would be $\widehat{\sigma}_{\mathrm{E}}^2 = \widehat{\sigma}_{\mathrm{W}}^2 - \frac{1}{2}\widehat{\sigma}_{\mathrm{A}}^2 - \frac{3}{4}\widehat{\sigma}_{\mathrm{D}}^2$, and its sampling error computed accordingly from $\mathbf{S}^{-1}$.

As noted in the main text, $cov_{\mathrm{RQ}} = \sigma_{\mathrm{R}}^2$, so $\mathbf{S}$ simplifies to

$$\mathbf{S} = \frac{n-1}{2\sigma_{\mathrm{W}}^4}\begin{pmatrix} m\sigma_{\mathrm{R}}^2 & m\sigma_{\mathrm{R}}^2 & -2m\sigma_{\mathrm{R}}^2\left(\sigma_{\mathrm{A}}^2+\sigma_{\mathrm{D}}^2\right)/\sigma_{\mathrm{W}}^2 \\ & m\sigma_{\mathrm{Q}}^2 & -2m\left(\sigma_{\mathrm{R}}^2\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{Q}}^2\sigma_{\mathrm{D}}^2\right)/\sigma_{\mathrm{W}}^2 \\ symm & & 1+3m\left[\sigma_{\mathrm{R}}^2\left(\sigma_{\mathrm{A}}^2+2\sigma_{\mathrm{D}}^2\right)\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{Q}}^2\sigma_{\mathrm{D}}^4\right]/\sigma_{\mathrm{W}}^4 \end{pmatrix}.$$

However, as σ_{R}^2 and σ_{Q}^2 have similar magnitude, $\mathbf{S}$ is almost singular and thus the genotypic variance cannot be partitioned into additive and dominance components unless the dataset is very large.

Appendix 3: Comparison of between and within family estimators

Let us assume a balanced one-way ANOVA (which is also REML if there are no unbalanced fixed effects) is used to estimate σ_{A}^2, i.e. $\hat{\sigma}_{\mathrm{Ab}}^2 = (MSB - MSW)/(nA)$ where MSB and MSW are the mean squares and A is the pedigree relationship (½ or ¼). It is assumed that there is no environmental correlation among sibs. Hence, with f families each of size n, $\mathrm{var}(MSB) = 2\left[\sigma_{\mathrm{W}}^2 + (n-1)A\sigma_{\mathrm{A}}^2\right]^2/(f-1)$, $\mathrm{var}(MSW) = 2\sigma_{\mathrm{W}}^4/[f(n-1)]$ and, as these are uncorrelated,

$$\mathrm{var}\left(\hat{\sigma}_{\mathrm{Ab}}^2\right) = \frac{2\sigma_{\mathrm{W}}^4}{(nA)^2}\left(\frac{\left[1+(n-1)\left(A\sigma_{\mathrm{A}}^2/\sigma_{\mathrm{W}}^2\right)\right]^2}{f-1} + \frac{1}{f(n-1)}\right).$$

For the within-family estimates, $\mathrm{var}\left(\hat{\sigma}_{\mathrm{Aw}}^2\right)$ is given by (3). Further simplification requires making some assumptions about numbers and size of families. As a first approximation, assume neither is small, so

$$\mathrm{var}\left(\hat{\sigma}_{\mathrm{Ab}}^2\right)\approx\frac{2\sigma_{\mathrm{W}}^4\left[1+nA\sigma_{\mathrm{A}}^2/\sigma_{\mathrm{W}}^2\right)\right]^2}{fn^2A^2}, \quad \mathrm{var}\left(\hat{\sigma}_{\mathrm{Aw}}^2\right)\approx\frac{2\sigma_{\mathrm{W}}^4}{fn^2\sigma_{\mathrm{R}}^2}$$

$$\text{and} \quad \frac{\mathrm{var}\left(\hat{\sigma}_{\mathrm{Aw}}^2\right)}{\mathrm{var}\left(\hat{\sigma}_{\mathrm{Ab}}^2\right)}\approx\frac{A^2/\sigma_{\mathrm{R}}^2}{\left[1+nA\sigma_{\mathrm{A}}^2/\sigma_{\mathrm{W}}^2\right]^2}.$$

Competing interests
The author declares no competing interests.

Author's contributions
WGH proposed, executed and reported the study.

Acknowledgements
I wish to thank Ian White, Peter Visscher, reviewers and editors for helpful comments.

Received: 15 March 2013 Accepted: 23 July 2013
Published: 3 September 2013

References

1. Falconer DS, Mackay TFC: *Introduction to quantitative genetics.* Essex: Longman Group Ltd; 1996.
2. Lynch M, Walsh JB: *Genetics and analysis of quantitative traits.* Sunderland, MA: Sinauer Associates; 1998.
3. Ritland K: **Marker-based method for inferences about quantitative inheritance in natural populations.** *Evolution* 1996, **50:**1062–1073.
4. Haseman JK, Elston RC: **The investigation of linkage between a quantitative trait and a marker locus.** *Behav Genet* 1972, **2:**3–19.
5. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG: **Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings.** *PLoS Genet* 2006, **2:**e41.
6. Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S, Hill WG, Hottenga JJ, Willemsen G, Boomsma DI, Liu YZ, Deng HW, Montgomery GW, Martin NG: **Genome partitioning of genetic variation for height from 11,214 sibling pairs.** *Am J Hum Genet* 2007, **81:**1104–1110.
7. Visscher PM: **Whole genome approaches to quantitative genetics.** *Genetica* 2009, **136:**351–358.
8. Hill WG: **Variation in genetic identity within kinships.** *Heredity* 1993, **71:**652–653.
9. Guo SW: **Proportion of genome shared identical by descent by relatives: concept, computation, and applications.** *Am J Hum Genet* 1995, **56:**1468–1476.
10. Hill WG, Weir BS: **Variation in actual relationship as a consequence of Mendelian sampling and linkage.** *Genet Res (Camb)* 2011, **93:**47–64.
11. Ødegård J, Meuwissen THE: **Estimation of heritability from limited family data using genome-wide identity-by-descent sharing.** *Genet Sel Evol* 2012, **44:**16.
12. Arias JA, Keehan M, Fisher P, Coppieters W, Spelman R: **A high density linkage map of the bovine genome.** *BMC Genet* 2009, **10:**18.
13. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin-rapid analysis of dense genetic maps using sparse gene flow trees.** *Nat Genet* 2002, **30:**97–101.
14. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42:**565–569.

doi:10.1186/1297-9686-45-32
Cite this article as: Hill: On estimation of genetic variance within families using genome-wide identity-by-descent sharing. *Genetics Selection Evolution* 2013 **45**:32.