Genetica
DOI 10.1007/s10709-008-9301-7

Whole genome approaches to quantitative genetics

Peter M. Visscher

Received: 15 July 2008 / Accepted: 16 July 2008

Abstract Apart from parent-offspring pairs and clones, relative pairs vary in the proportion of the genome that they share identical by descent. In the past, quantitative geneticists have used the *expected* value of sharing genes by descent to estimate genetic parameters and predict breeding values. With the possibility to genotype individuals for many markers across the genome it is now possible to empirically estimate the *actual* relationship between relatives. We review some of the theory underlying the variation in genetic identity, show applications to estimating genetic variance for height in humans and discuss other applications.

Keywords Quantitative genetics · Genome · Identity-by-descent · Estimation · Genetic variance · Actual relationship · Realised relationship

Introduction

Quantitative genetics theory has a solid foundation in mathematics and statistics and is well established following the ground breaking work by Fisher and Wright nearly a century ago. Estimation of genetic parameters, such as heritability and genetic correlation coefficients, and applications in human genetics, evolutionary biology and plant and animal breeding programmes, are based upon the specific theory of the resemblance between relatives due to genetic factors. The resemblance between relatives depends on the number of alleles that they share identical-by-descent (IBD) at loci influencing the trait of interest. Due to segregation and linkage, the actual number of alleles shared IBD is the outcome of a stochastic process. Until quite recently, IBD sharing between relatives could not be observed, and the theoretical expected value, based upon probabilities, is typically used in applications. If there are many (strictly infinite) independent loci influencing a phenotype, then there is no difference between the expected and actual proportion of alleles shared IBD between relatives. The infinitesimal model of quantitative genetics leads to appealing theoretical consequences (e.g., multivariate normality of additive genetic values) and its application has been highly successful, particularly in animal breeding selection programmes.

Nevertheless, genes reside on a finite, relatively small number of chromosomes, and recombination events during meiosis are relatively infrequent, typically 1–3 per 100 cM. This implies that large segments of chromosomes segregate from parents to progeny, which creates variation in the proportion of the genome shared between pairs of relatives around the expected value.

In this article, we will review the theory of the variation in the proportion of genomes shared IBD between relatives, and discuss applications that utilise this variation.

Theory

The covariance between relatives due to genetic factors is based upon the probability of identity by descent. For individuals X, Y with relationship a (=2× probability of identity by descent of random alleles) and probability of IBD at both alleles at a locus d, the genetic covariance is

$$\mathrm{cov}(X,Y) = \sum_{n}\sum_{m} a^{n} d^{m} V_{\mathrm{A}(n)\mathrm{D}(m)},$$

P. M. Visscher (✉)
Queensland Institute of Medical Research, Brisbane, Australia
e-mail: peter.visscher@qimr.edu.au

where $V_{A(n)D(m)}$ denotes the component with n A and m D terms, and A and D denote additive and dominance effects, respectively (Falconer and Mackay 1996; Lynch and Walsh 1998). The additive genetic covariance is simply aV_A. The genetic covariance therefore depends on allele sharing probabilities, which can be derived from the pedigree relationships using probability theory.

However, if the *actual* proportion of genes that account for quantitative trait variation in the genome varies between pairs of relatives with the same values of a and d, then the phenotypic covariance will vary accordingly. Hence, among all pairs of relatives with the same expected genetic identity, the pairs that share more alleles at trait loci IBD are expected to be phenotypically more similar. Genetic covariance would be the same for all pairs of relatives with the same value of a and d if there were a very large number of loci independently segregating in the pedigree. But the number of loci is finite and loci do not segregate independently due to genetic linkage, so we can expect variation in genetic identity.

Variation in genetic identity

The theory underlying the variation in the proportion of the genome shared IBD between relatives in crosses from inbred lines and in outbred populations was developed by several authors (Franklin 1977; Guo 1994, 1995, 1996; Hill 1993a, b; Risch and Lange 1979; Stam 1980; Stam and Zeven 1981), mostly for different reasons and with different applications in mind. I will give an example derivation for halfsibs (following Hill 1993a; Visscher et al. 2006).

Twice the kinship coefficient for halfsibs, equal to the coefficient of additive relationship, is $a = 1/4$. Let π be the *actual* or *realised* coefficient of additive relationship for a pair of halfsibs. In a non-inbred population, this is equal to the actual proportion of the genome shared IBD. π is a random variable with $E(\pi) = a$ and we wish to quantify the variance of π. Let δ_i be an indicator variable for locus i, which is 1 if both halfsibs have inherited the same allele from the common parent and 0 otherwise. For halfsibs, the probability that $\delta_i = 1$ and $\delta_i = 0$ is 1/2 and 1/2, respectively, and $\pi_i = 1/2\delta_i$. Hence, $E(\pi_i) = 1/4$ and $\mathrm{var}(\pi_i) = 1/16$. These are the mean and variance of the coefficient of identity at a single locus for halfsibs. For two loci i and j and recombination fraction c, $E(\pi_i, \pi_j) = (1/16)\ [2(1 - c)^2 + 2c^2]$. Hence, the covariance of the indicator variables at two loci is,

$$\mathrm{cov}(\pi_i, \pi_j) = E(\pi_i, \pi_j) - E(\pi_i)E(\pi_j) = (1/16)(1 - 2c)^2.$$

Assuming the Haldane mapping function, i.e. that recombination events are continuously distributed with no interference, the covariance can be written as:

$$\mathrm{cov}(\pi_i, \pi_j) = (1/16)\exp(-4d_{ij}),$$

with d_{ij} the distance (in Morgan) between the loci. For n loci, the variance of chromosome-wide sharing between two halfsibs is:

$$\mathrm{var}(\pi) = (1/n^2)(1/16)\Sigma\Sigma\exp(-4d_{ij})$$

(following Hill 1993a; Stam and Zeven 1981). If n becomes very large this equation can be expressed as an integral (Hill 1993a; Stam and Zeven 1981),

$$\mathrm{var}(\pi) = \frac{1}{16}\int_0^l\int_0^l e^{-4|x_1-x_2|}dx_1dx_2 = \frac{1}{32l^2}(l - r_{2l}/2),$$

with l the length of the chromosome (in Morgan) and r_{2l} the recombination fraction for a segment of length $2l$ (=1/2(1 − exp(−4l)). Hence, the total variance in IBD sharing between two half siblings for a chromosome of length l is (Guo 1996; Hill 1993b):

$$\mathrm{var}(\pi) = (128l^2)^{-1}[4l - 1 + \exp(-4l)].$$

Finally, genome-wide π from k chromosomes is, $\pi_g = (1/L)\ \Sigma(l_i\ \pi_i)$, with $L = \Sigma(l_i)$, and

$$\mathrm{var}(\pi_a) = [1/(128L^2)][4L - k + \Sigma\exp(-4l_i)].$$

These results are the same as those of Hill (1993b) and Guo (1996). Derivation of the latter was based upon Markov chains. For human autosomes ($k = 22$ and $L = 35$ (Kong et al. 2004)), the variance of genome-wide IBD sharing of halfsibs is approximately $1/(32L) - 22/(128L^2) = 0.00089 - 0.00014 = 0.00075$, or a standard deviation of 0.027 about the mean of 0.25. To a first order approximation, the variance is determined by the total map length (L), and the number of chromosomes and the distribution of their lengths are less important. For example, the standard deviation of identity from the above prediction or that from taking the actual distribution of the 22 autosomes or that assuming 22 autosomes all with the same length ($l = 35/22$) all give the same value to the fourth decimal (0.0272).

The genome-wide variance can be compared to the variance at a single locus to calculate an equivalent number of independent segments (loci) that would give the same genome-wide variance (Gagnon et al. 2005; Visscher et al. 2006). For halfsibs in humans, this number is (1/16)/0.00075 = 83.3.

For fullsibs, the variance in the genome-wide additive coefficient of relationship is twice that of halfsibs because paternal and maternal chromosomes are inherited independently. In humans ($k = 22$) (Guo 1994; Visscher et al. 2006),

$$\mathrm{var}(\pi_a) \approx 1/(16L) - 22/(64L^2).$$

Dominance

The coefficient of dominance is a function of the probability that two relatives share both alleles IBD (= IBD2). In a non-inbred population, this probability is also called the coefficient of fraternity (Lynch and Walsh 1998). Here I consider the variation in the coefficient of dominance for fullsibs. The prior probability that full sibs share two alleles IBD is 1/4, and the mean and variance of an indicator variable that is one if both alleles are shared IBD and zero otherwise is 1/4 and 3/16, respectively. Hence the variance of the coefficient of dominance at a single locus is 1.5 times the variance in the additive coefficient of relationship (which is 1/8). The probability that the sibs share two alleles IBD at a linked locus, given that they are IBD2, is $(1-c)^4 + 2[(1-c)c]^2 + c^4 = [(1-c)^2 + c^2]^2$ (Visscher et al. 2006). Using the same methodology as before gives the genome-wide variance in the coefficient of fraternity as,

$$\mathrm{var}(\pi_d) = [1/(16L^2)][(5/4)L - 1/2\Sigma r_{2l} - (1/16)\Sigma r_{4l}]$$

$\approx 5/(64L) - 99/(256L^2) \approx 5/(64L) - 1/(3L^2)$ (Visscher et al. 2006). The variance of the dominance coefficient is larger (by about 30% if $L = 35$) than the variance of the genome-wide additive coefficient of relationship. The correlation between mean genome-wide allele sharing and mean genome-wide IBD2 sharing is the ratio of the SD because the regression of π_a on π_d is unity (Visscher et al. 2006),

$$r(\pi_a, \pi_d) = \sigma(\pi_a)/\sigma(\pi_d) \approx [1/(16L)/\{5/(64L)\}]^{0.5}.$$

For $L = 35$, this correlation is 0.89. Hence the genome-wide additive and dominance coefficients of relatedness are highly correlated.

X-chromosome

The previous calculations for genome-wide relationships were done ignoring the contribution of the X-chromosome to variation in genome-wide IBD sharing between relatives. For mammals, the X-chromosome constitutes 3–5% of the genome, so the error made in calculating the variance of identity by ignoring the X-chromosome is not large. For the X-chromosome, the sex of the pair of relatives determines the definition and properties of the coefficient of relationship (Lynch and Walsh 1998). Here I consider a pair of mammalian full siblings. Sister–sister (ss) pairs always share the allele inherited from the father and have a probability of 1/2 in sharing the allele inherited from the mother. Therefore, at a single locus, $E(\pi) = 3/4$ and $\mathrm{var}(\pi) = 1/16$. Although the mean coefficient of relationship is larger than for halfsibs, the variance of the coefficient of relationship is exactly the same as for halfsibs. Therefore,

$$\mathrm{var}(\pi_{\mathrm{ss}}) = (128 l_{\mathrm{X}}^2)^{-1}[4l_{\mathrm{X}} - 1 + \exp(-4l_{\mathrm{X}})].$$

For humans ($l_{\mathrm{X}} = 185$ cM), the SD of X-chromosome sharing between sister pairs is 0.121. Brother–brother (bb) pairs either share the allele on the X-chromosome IBD ($\pi = 1$) or they do not ($\pi = 0$), both with a probability of 1/2. Therefore, at a single locus, $E(\pi) = 1/2$ and $\mathrm{var}(\pi) = 1/4$. The segregation and recombination processes are again the same as for halfsibs but with a four-fold larger variance. Hence,

$$\mathrm{var}(\pi_{\mathrm{bb}}) = (32 l_{\mathrm{X}}^2)^{-1}[4l_{\mathrm{X}} - 1 + \exp(-4l_{\mathrm{X}})].$$

For humans, the SD of X-chromosome sharing between brother pairs is 0.242. Finally, for brother–sister (bs) pairs the allele on the X-chromosome in the male is either IBD with one of the two alleles in the female ($\pi = 1/2$) or it is IBD with neither allele in the female ($\pi = 0$), both with a probability of 1/2. It follows that the mean and variance of π are 1/4 and 1/16, respectively, and the variance $\mathrm{var}(\pi_{\mathrm{bs}})$ is the same as for sister–sister pairs.

Note that these coefficients of relationship are also the coefficient of additive genetic covariance for the brother–brother and sister–sister pairs, but not for the brother–sister pairs. The coefficient of additive relationship for brother–sister pairs is $(1/\sqrt{2})\pi_{\mathrm{bs}}$ (Bulmer 1985; Kent et al. 2005; Lynch and Walsh 1998). In addition, the above definition of coefficients of relationship does not take dosage compensation for expression of the phenotype into account.

Comparison of relative pairs

Using the theoretical derivations from (Guo 1996) and the map length of autosomes in humans (Kong et al. 2004), the SD of whole genome coefficients of relatedness were calculated for a number of relative pairs. Results are shown in Table 1. These results correspond closely to those in Table 2 of Guo (1996).

The variance in genome-wide sharing is small relative to the mean, with coefficients of variation ranging from 8% (fullsibs) to 17% (first cousins). It was noted previously that although halfsibs and grandparent–grandchild have the same expectation, the variance in genome-wide sharing is larger for grandparent–grandchild pairs (Hill 1993b). For halfsibs, the probability of them sharing alleles IBD at two loci is $1/2(1-c)^2 + 1/2c^2$, whereas the probability that a grandchild inherits the same alleles from the grandparent is $1/2(1-c)$ (Hill 1993b). For a small value of c, these probabilities are $(1/2 - c)$ and $(1/2 – 1/2c)$, respectively. The latter is closer to 1/2 and has therefore larger variance.

Table 1 Predicted variance in genome-wide coefficients of additive relationship of human relative pairs

Relatives	Single locus		Genome-wide		Equivalent no. loci
	$E(\pi)$	var(π)	var(π)	SD(π)	
Fullsibs	1/2	1/8	0.00147	0.039	84
Halfsibs	1/4	1/16	0.00075	0.027	84
Grandparent–grandchild	1/4	1/16	0.00122	0.035	50
Uncle–niece	1/4	1/16	0.00063	0.025	99
First cousin	1/8	3/64	0.00044	0.021	102
Double first cousin	1/4	3/32	0.00090	0.030	102

For non-collateral (ancestor-descendant) relatives it can be seen that the more distant the relatives, the larger the CV. The expectation and variance of π at a single locus for relatives m generations apart (e.g., $m = 2$ for grandparent–grandchild pairs) are $(1/2)^m$ and $1/2(1/2)^m[1 - 2(1/2)^m]$, respectively, or a and $1/2a(1 - 2a)$, with a the additive coefficient of relatedness (twice the kinship coefficient). Hence $\mathrm{CV}(\pi) = \sqrt{[1/(2a) - 1]}$, so the smaller the coefficient of relationship, the larger the SD of sharing as a proportion of the mean. Similarly for genome-wide sharing, using results from (Hill 1993b), $\mathrm{CV}(\pi) = \frac{1}{L}\sqrt{\frac{1}{2}\sum_{j=1}^{m-1}\binom{m-1}{j}\frac{1}{j^2}\left(2jL - k + \sum_{i=1}^{k} e^{-2jl_i}\right)}$, which increases with increasing m.

Validation of theory

Previously we showed a comparison of the theory with empirical data from 4,401 pseudo-independent fullsib pairs from Australian families (Visscher et al. 2006). The siblings and their parents were genotyped for microsatellite markers and additive and dominance coefficients of relationships were estimated at each cM and genome-wide using exact multi-point probability calculations (Abecasis et al. 2002). Figures 1 and 2 show the empirical distribution of the additive and dominance genome-wide relationships. The mean and SD were 0.0498 and 0.036 for the additive coefficients and 0.248 and 0.040 for the dominance coefficients (Visscher et al. 2006). Both the mean and variance are close to expectation. The extreme values for the additive coefficients are ~0.37 and ~0.63, so that at the low range some fullsibs are in between average halfsibs and average fullsibs and at the high range the sibs are in between average fullsibs and monozygotic twins. For dominance coefficients, the range is from ~0.12 to ~0.41.

The empirical variances are approximately 85 and 83% of the theoretical values. The ratio of empirical to theoretical value can be seen as a measure of genome-wide multimarker information content, analogous to that ratio at a single location. A multimarker information content of ~0.8 is expected given that the estimates were based upon a ~5 cM genome-scan from microsatellite markers, but this concordance does not prove that the assumptions made in the theoretical prediction of the variance are correct.

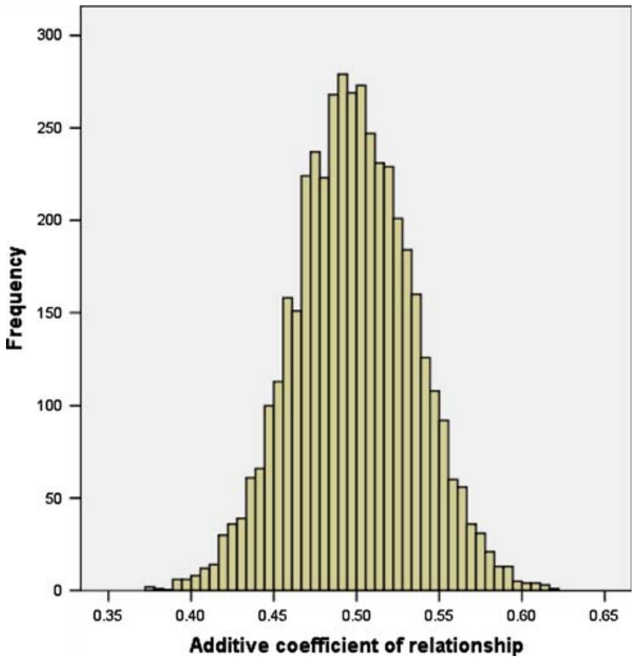

Fig. 1 Empirical distribution of genome-wide coefficients of additive relationships from 4,401 pairs of fullsibs (Visscher et al. 2006)

Quantification of the ratio of empirical to theoretically predicted variance is useful because the variance of locus or genome-wide coefficients of relationship is proportional to the power to detect QTL or genome-wide variance (Visscher and Hopper 2001; Visscher et al. 2006). Empirical and theoretical values can also be different because of the assumptions made in the theory, in particular the assumption of map function. For example, a localised distribution of chiasma positions gave an estimated standard deviation of identity of human full siblings of 0.055 (Suarez et al. 1979), whereas the use of Haldane's mapping function on the same data resulted in an estimate of 0.04 (Risch and Lange 1979). The increasing use of very

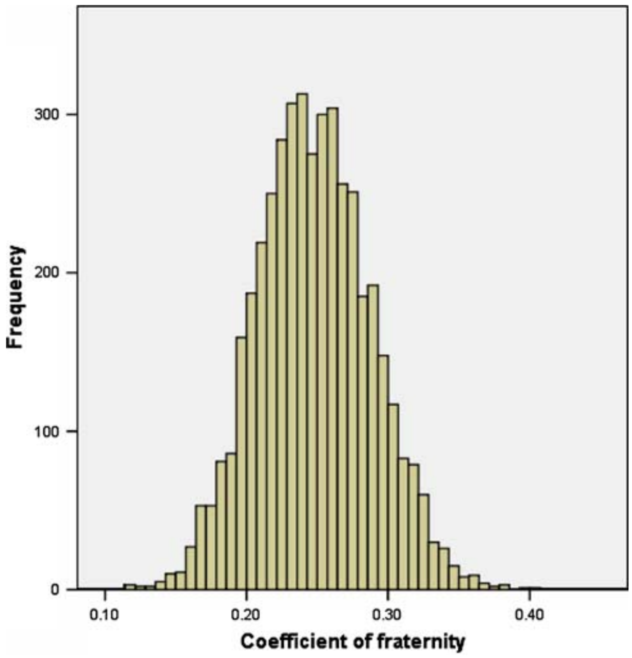

Fig. 2 Empirical distribution of genome-wide coefficients of dominance relationships from 4,401 pairs of fullsibs (Visscher et al. 2006)

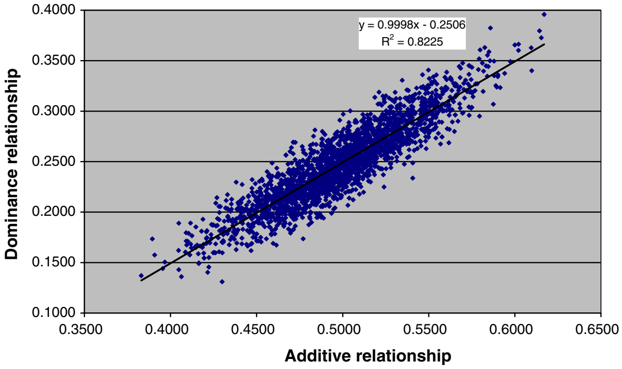

Fig. 3 Empirical relationship between genome-wide additive and dominance relationship from 4,401 pairs of fullsibs (Visscher et al. 2006)

dense SNP arrays will allow a more accurate estimation of variation in identity in the near future.

Figure 3 shows the relationship between the additive and dominance coefficients for the 4,401 pairs. Again, the correlation is close to expectation.

Applications

Estimation of genetic variance components

How can the variation in genetic identity be used in practice? One possibility that we explored recently in human populations (Visscher et al. 2007; Visscher et al. 2006) is to estimate additive genetic variance from the deviations of the average coefficients of relationship. In a simple design for which all pairs of relatives with a phenotype have the same expected coefficient of relationship (a), a model of the covariance of the relatives is, cov(y_i,y_j) $= V_C + \pi_{ij}V_A$, with $E(\pi_{ij}) = a$ and V_C the variance of between-pair effects not accounted for by the proportion of the genome shared IBD. This model is equivalent to cov(y_i,y_j) $= aV_C + \pi_{ij}V_A$, so clearly V_A is estimated from the deviation of the actual from the expected coefficient of relationship, independent of average between-family effects. Genetic variance can then be estimated from the covariation of the phenotypic resemblance of relatives and their actual genome-wide relationship. We used sibling pairs that had microsatellite marker genotype scans, typically 400–800 markers per individual, and the phenotype of interest was height. Maximum likelihood was used by fitting in addition to the usual A-matrix a matrix with realised genome-wide relationships. When the data structure is simple, e.g. pairs of fullsibs with no additional pedigree information and a small number of fixed effects in the model, the estimates of variance components from ML and REML are very similar. The reason for fitting the average relationship too was that we wished to estimate heritability free of assumptions regarding the between-family variance. The estimate of heritability from our method was 0.8 (Visscher et al. 2006) but with a large standard error (95% CI 0.5–0.9). The reason that the SE is so large is because the sampling variance of additive genetic variance is proportional to 1/var(π). For full siblings and a single locus, this is 1/8, but genome wide, as shown before, this value is $\sim 0.038^2$, about 80 times smaller. Therefore, large sample sizes are needed to estimate genetic variance from the deviations of actual relationships about their expected values, because the variation in actual relationships is small.

We subsequently increased the sample size to 11,214 sibling pairs with genome-wide marker data and a measurement on height, by combining samples from Australia, the USA and The Netherlands (Visscher et al. 2007). For each sibling pair we estimated realised additive relationship coefficients for each of the 22 autosomes and the X-chromosome and estimated the proportion of additive genetic variance associated with each chromosome using maximum likelihood. (Data were pre-adjusted for the fixed effects of sex and age, and with so few fixed effects the estimates from REML are almost identical to those from ML). We found that longer chromosomes contributed more additive genetic variance, that at least 6 chromosomes contributed to additive genetic variance for height, could not reject the hypothesis that additive variance was explained in proportion to the length of the chromosome and found no evidence for non-additive genetic variance (Visscher et al. 2007). If the distribution of genetic variance in the genome is approximately proportional to the length of the chromosome then this implies some kind of infinitesimal model and has implications for gene mapping by linkage and association.

Chromosome and whole-genome approaches and the number of trait loci

One possible disadvantage of a genome-wide approach to estimating genetic variance and breeding values is that the loci that explain genetic variance may not be uniformly spread across the genome and that their effect sizes varies (Xu 2006). Does the estimation of heritability from genome-wide sharing depend on a polygenic model? The derivation below for full siblings suggests that the estimate of genetic variance is not biased but inaccurate when there are not a large number of loci each with small effect.

Let n = number of loci (assumed unlinked); π_i = proportion of alleles shared IBD at locus i (for a given pair of sibs); π = genome-wide IBD = $(\Sigma\pi_i)/n$; $\sigma_i{}^2$ = proportion of (additive) genetic variance due to locus i; σ^2 = total (additive) genetic variance = $\Sigma\sigma_i^2$

π_i (and π) are random variables, and for all loci var(π_i) = var(π_j) {=1/8 for sibpairs}. Hence var(π) = var(π_i)/n. A maximum likelihood estimation procedure can be approximated by considering the (Haseman and Elston 1972) approach. Let $D^2 = (y_1 - y_2)^2$, the squared difference of the phenotypes of the sibs. Then,

$$E(D^2|\pi_i) = \Sigma\{2\sigma_i^2(1 - \pi_i)\}.$$

For analysis, the linear regression $D^2 = \alpha + \beta\pi + e$ is used.

$$\begin{aligned}\beta &= \mathrm{cov}(D^2, \pi)/\mathrm{var}(\pi) \propto \mathrm{cov}\left[\Sigma\pi_i\sigma_i^2, (\Sigma\pi_i)/n\right]/\left[\mathrm{var}(\pi_i)/n\right] \\ &= \mathrm{cov}\left[\Sigma\pi_i\sigma_i^2, \Sigma\pi_i\right]/\left[\mathrm{var}(\pi_i)\right] \\ &= \Sigma\left\{\mathrm{var}(\pi_i)\sigma_i^2\right\}/\left[\mathrm{var}(\pi_i)\right] \\ &= \Sigma\sigma_i^2,\end{aligned}$$

assuming no covariance between π_i and $\sigma_i{}^2$. Hence, the regression coefficient β is proportional to the total genetic variance, independent of the number of loci and their effects. This implies that there is no bias in the estimate when, for example, there are a few loci (or even a single one) of large effect. However, the sampling variance of the estimate of the regression coefficient would be greatly increased by the inefficient regression on the mean IBD of several loci, rather than directly on the IBD at the QTL. This reasoning is similar to saying that the estimate of QTL variance, when looking at the correct location, is not biased in the absence of complete IBD information. (Haseman and Elston 1972) derive the latter in a different way.

Prediction of breeding values from realised relationship

Previous applications were for a simple pedigree (sibling pairs) and to estimate and partition genetic variance. With sufficient marker data, realised relationships can be estimated between any pair of relatives in a complex pedigree and prediction of breeding values in an artificial selection programme can be made using the realised relationship matrix. The estimation of the realised relationship offers new opportunities in breeding programmes, in particular for species with large family sizes (e.g., livestock). For example, the breeding value of an individual without a phenotype can in principle be estimated with 100% accuracy on the basis of a large number of collateral relatives each with a phenotype (Meuwissen et al. 2001).

If genotyping is cheap relative to phenotyping then how much extra response to selection can we obtain from measuring the actual relationships between relatives? Here we consider a simple example of sib selection. All sibs are assumed to be genotyped but only one individual gets phenotyped (for example, because phenotyping is much more expensive than genotyping). For the derivation below we assume that the individual that gets phenotyped cannot be used for breeding (e.g., the phenotype is a carcass trait in a meat enterprise). Without genotype information, the choice of individual to be phenotyped is random and the choice of which sibling(s) are selected for breeding is random. With cheap genome-wide genotyping of n siblings, there are $n(n - 1)/2$ pairs of siblings for which we know the proportion (π) of their genome-wide IBD sharing. We suggest to pick out the pair with the largest genome-wide IBD sharing and to phenotype one of these two siblings. For full siblings, the mean and SD of the random variable π is 1/2 and, approximately, 0.04, respectively. We assume that π is normally distributed and that the genetic covariance between sibs is proportional to π. In the absence of genotypic information, the EBV of unphenotyped siblings are,

$$\mathrm{EBV} = (1/2)h^2P,$$

with P the phenotype of the sibling. The variance of the EBV is $1/4h^2V_\mathrm{A}$. With genotyping,

$$\mathrm{EBV} = \pi_\mathrm{s}h^2P,$$

with π_s the observed genome-wide IBD sharing between the phenotyped individual and its sib with which it shares the largest proportion of the genome.

$$\pi_\mathrm{s} = (1/2) + i\sigma_\pi$$

with i the selection intensity of pairs within a family. Response to selection is proportional to $E(\pi_\mathrm{s})$, so the gain in response by using genotyping is,

$$\text{Gain} = E(\pi_s)/E(\pi) = ((1/2) + i\sigma_\pi)/(1/2) = 1 + 2i\sigma_\pi.$$

Below are some examples of the extra gain, using the small sample values of *i* from (Falconer and Mackay 1996).

n Sibs	# Pairs	*i*	% Gain
2	1	0	0
3	3	0.85	7
4	6	1.27	10
8	28	2	16
10	45	2.2	18

Although this example is artificial and probably not practical, it illustrates that substantial gains in the accuracy of selection can be made. Note that even with 2 sibs in a family there is a tiny increase in response to selection if both sibs are genotyped because of the increase in accuracy; the variance in EBV is increased by 1 + 4var(π) but this increase is trivial.

Using genome-wide identity-by-state sharing

In principle, genome-wide approaches as applied to known pedigrees can be applied to 'unrelated' individuals by estimating the proportion of the genome identical-by-state (IBS) as a proxy for identical-by-descent. This approach has been used to estimate genetic parameters in natural populations, usually in a two-stage procedure in which relationships are first inferred and genetic parameters are estimated subsequently (Lynch and Walsh 1998; Thomas 2005; Thomas et al. 2002; Thomas et al. 2000). There are two potential problems related to statistical power with such approaches. Firstly, the more distant the actual relationship between a pair of individuals, the larger the number of markers required to accurately estimate the relationship. Distant relatives have a low probability of sharing alleles IBD and look like random pairs drawn from the population, so many markers are needed to estimate allele sharing in excess of what would be expected by chance. Secondly, given an accurate estimation of genome-wide IBS sharing, the precision with which quantitative genetic parameters are estimated is inversely proportional to the variation in estimated relationships. In a population of large effective size this variation is likely to be very small (say, $<0.01^2$) so large sample sizes are needed (>100,000s). Recent estimates of co-ancestry of supposedly unrelated individuals from four human populations from millions of SNPs showed that there were pairs of relatives with surprisingly large coefficients of relatedness (>0.01) (Frazer et al. 2007).

Discussion and conclusions

We have given a brief overview of the theory of variation in genetic identity, have given a number of applications and have hinted at possible future applications. No doubt that there are many more opportunities to exploit the variation in coefficients of relationship. For example, if the sample size is large enough it may become feasible to estimate dominance and other non-additive variance accurately and unbiased, finally getting round the curse of confounding between dominance variance and non-genetic sources of family resemblance (Visscher et al. 2006).

The main caveat of the suggested applications is that very large sample sizes are needed, both in terms of the number of genetic markers per individual and the number of individuals with a phenotype. Such sample sizes are now emerging in human genetics and should be achievable in livestock populations. In natural populations, the absence of pedigree information should not be an obstacle to applying genome-wide approaches, provided that there is sufficient variation in relatedness that can be sampled.

The merging of quantitative and population genetics, driven by data generated by large-scale high-throughput genomics platforms, offers new approaches to classical problems in quantitative genetics.

Acknowledgements We acknowledge support from the Australian National Health and Medical Research Council (grants 389892 and 442915) and the Australian Research Council (grant DP0770096), and thank Bill Hill, Mike Goddard, Nick Martin and Naomi Wray for many stimulating discussions and two referees for helpful comments.

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

Bulmer MG (1985) The mathematical theory of quantitative genetics. Clarendon Press, Oxford

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow

Franklin IR (1977) The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor Popul Biol 11:60–80

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Gagnon A, Beise J, Vaupel JW (2005) Genome-wide identity-by-descent sharing among CEPH siblings. Genet Epidemiol 29:215–224

Guo SW (1994) Computation of identity-by-descent proportions shared by two siblings. Am J Hum Genet 54:1104–1109

Guo SW (1995) Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am J Hum Genet 56:1468–1476

Guo SW (1996) Variation in genetic identity among relatives. Hum Hered 46:61–70

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Hill WG (1993a) Variation in genetic composition in backcrossing programs. J Hered 84:212–213

Hill WG (1993b) Variation in genetic identity within kinships. Heredity 71:652–653

Kent JW Jr, Dyer TD, Blangero J (2005) Estimating the additive genetic effect of the X chromosome. Genet Epidemiol 29:377–388

Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–1148

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Risch N, Lange K (1979) Application of a recombination model in calculating the variance of sib pair genetic identity. Ann Hum Genet 43:177–186

Stam P (1980) The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet Res 35:131–155

Stam P, Zeven AC (1981) The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by back-crossing. Euphytica 30:227–238

Suarez BK, Reich T, Fishman PM (1979) Variability in sib pair genetic identity. Hum Hered 29:37–41

Thomas SC (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. Philos Trans R Soc Lond B Biol Sci 360:1457–1467

Thomas SC, Pemberton JM, Hill WG (2000) Estimating variance components in natural populations using inferred relationships. Heredity 84(Pt 4):427–436

Thomas SC, Coltman DW, Pemberton JM (2002) The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. J Evol Biol 15:92–99

Visscher PM, Hopper JL (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. Ann Hum Genet 65:583–601

Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2:e41

Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, Medland S, Hill WG, Hottenga JJ, Willemsen G, Boomsma DI, Liu Y-Z, Deng HW, Montgomery GW, Martin NG (2007) Genome partitioning of genetic variation for height from 11, 214 sibling pairs. Am J Hum Genet 81:1104–1110

Xu S (2006) Population genetics: separating nurture from nature in estimating heritability. Heredity 97:256–257