ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/46579430

Reconciling the analysis of IBD and IBS in complex trait studies

Article *in* Nature Reviews Genetics · September 2010

DOI: 10.1038/nrg2865 · Source: PubMed

CITATIONS
164

READS
94

3 authors, including:

Joseph E Powell
The University of Queensland

167 PUBLICATIONS **7,981** CITATIONS

SEE PROFILE

All content following this page was uploaded by Joseph E Powell on 16 May 2014.

The user has requested enhancement of the downloaded file.

STUDY DESIGNS — OPINION

Reconciling the analysis of IBD and IBS in complex trait studies

Joseph E. Powell, Peter M. Visscher and Michael E. Goddard

Abstract | Identity by descent (IBD) is a fundamental concept in genetics and refers to alleles that are descended from a common ancestor in a base population. Identity by state (IBS) simply refers to alleles that are the same, irrespective of whether they are inherited from a recent ancestor. In modern applications, IBD relationships are estimated from genetic markers in individuals without any known relationship. This can lead to erroneous inference because a consistent base population is not used. We argue that the purpose of most IBD calculations is to predict IBS at unobserved loci. Recognizing this aim leads to better methods to estimating IBD with benefits in mapping genes, estimating genetic variance and predicting inbreeding depression.

The concept of identity by descent (IBD) was classically quantified[1] following Wright's work on the coefficients of relationship[2,3] and is used to indicate two homologous alleles that have descended from a common ancestor. This fundamental concept has many uses in genetics, including predicting genotype frequencies[4], mapping genes[5,6], estimating genetic variance[7] and predicting inbreeding depression[8]. The probability that two alleles are IBD has to be defined with respect to a base (reference) population; that is, the two alleles are descended from the same ancestral allele in the base population. Traditionally, the probability that two alleles are IBD was most often calculated from a known pedigree and so the individuals at the top of the pedigree (the founders) form a natural base population. However, it is becoming common to use data on genetic markers such as SNPs to estimate the probability of being IBD without reference to a known pedigree (for an example see REF. 9) and, in this case, there is no obvious base population. Moreover, the concept of IBD seems to conflict with the well-established coalescence theory[10–12] in which all alleles are descended from a common ancestor but at different times in the past. In practice, this conflict has been ignored by using IBD concepts for recent common ancestors and coalescence analysis for distant common ancestors; however, the two categories of ancestor merge, especially when using dense SNP or DNA sequence data.

The lack of a consistent definition of IBD probabilities based on genetic markers can lead to several practical problems. For instance, popular methods[6] to estimate IBD probabilities report that the relationship between many pairs of individuals is zero[9,13]. However, all individuals are related if traced back far enough and some pairs of individuals are more closely related than others despite all being called 'unrelated'. Consequently, this approach loses much of the information contained in the data and leads to incorrect conclusions with respect to the estimation of relatedness between pairs of individuals and the estimation of genome-wide inbreeding for individuals. Furthermore, using these SNP-derived estimates in combination with phenotypes will lead to imprecise and biased estimates of genetic variance and of inbreeding depression.

Most of the uses of IBD implicitly involve predicting the probability that alleles at an unobserved site are identical by state (IBS); that is, whether they 'look' the same. Consequently, we argue and demonstrate that methods of estimating the probability of IBD should be designed to estimate the probability that alleles at an unobserved locus are IBS. This provides a logical basis for deriving IBD probabilities and unifies this approach with that of coalescent analysis. The recognition that we need to estimate the probability that individuals carry alleles that are IBS at unobserved sites also leads to a new and unbiased method for estimating genetic variance and can provide an estimate of the genetic variance that is not captured by a panel of SNPs (termed the "missing heritability"[14,15]).

In this Opinion we first define the probability of IBD (F), and point out the equivalence of this probability to the correlation between the alleles carried by two different gametes. Then we show that by expressing this correlation relative to the current population instead of relative to some past population, we can overcome the practical problem of an undefined base population. With this definition of IBD, F becomes a convenient parameter for predicting the probability that two gametes carry IBS alleles at an unobserved site. Finally, we review methods for estimating F from SNP data and the use of IBD in gene mapping, estimation of genetic variation and prediction of inbreeding depression.

IBD theory

We first provide a clear definition of the probability of IBD and show how this measure is related to IBS at unseen loci.

Defining IBD, relationships and inbreeding at a single locus. We can use the symbol F to denote the probability that at a single site in the genome, homologous alleles in two different gametes are IBD with respect to a defined base population. This probability is also called the gametic relationship of the gametes. If the two alleles are in the same diploid individual then F is the inbreeding coefficient of the individual at this locus. If we consider two diploid individuals, each with two alleles at a locus, then there are four pairs of gametes (taking one from the first individual and one from the second). The co-ancestry of the two individuals (that is, the expected inbreeding

© 2010 Macmillan Publishers Limited. All rights reserved

of their offspring) is the average of these four F values and their numerator relationship (A)[16] is twice their co-ancestry[16,17]. The relationship of a gamete with itself is 1, so the numerator relationship of a diploid individual with itself is $1 + F$, in which F is the inbreeding coefficient of the individual.

Defining IBD at a chromosome segment. If we consider a segment of a chromosome instead of a single point in the genome, it is possible to define IBD slightly differently. We define two homologous chromosome segments as IBD if they descend from a common ancestor without either of them experiencing a recombination[18]. In this definition of 'chromosome segment IBD' there is no need for a base population.

IBS and IBD. We define alleles as being IBS if neither allele has experienced a mutation since their last common ancestor. Coalescent theory treats all alleles at a locus as being IBD and models the probability of mutation causing them not to be IBS[10–12,19]. This leads to tension between coalescent theory and both pedigree and marker IBD methods, which are based on the framework that all loci are independent in the base population. BOX 1 illustrates the concepts of IBD for single sites and for chromosome segments and a coalescent tree for the single site.

Predicting genotype probabilities using *F*

If there are two alleles (for example, G and T) with allele frequencies q and p in the base population, then the genotype probabilities for diploid individuals are as outlined in equation 1:

$$\begin{array}{ccc} \text{GG} & \text{GT} & \text{TT} \\ q^2 + pqF & 2pq\,(1 - F) & p^2 + pqF \end{array} \qquad (1)$$

In this equation F is the probability that the alleles are IBD with respect to the base population[16]. F can also be described as the correlation between the gametes (BOX 2). In fact, the choice of a base population is arbitrary and in BOX 2 we describe how the base population, to which IBD coefficients are expressed, can be changed. It is convenient to choose the current population as the base because we can easily estimate allele frequencies in the current population but it may be difficult to estimate them in a base population that has a more ancient ancestry.

If the base is the current population and if mating is at random then the mean F is zero and the genotypes show Hardy–Weinberg frequencies over the whole population. However, even in a randomly mating population some mates are more closely related than others and so F varies between individuals. If $F < 0$ this indicates an individual that is less homozygous than the average. Negative F values cannot be interpreted as a probability, but they can still be interpreted as a correlation and so equation 1 still applies. The frequencies of genotypes given by equation 1 can also be interpreted as the frequencies of pairs of gametes even if they are not in the same individual. Thus, equation 1 describes a model for predicting whether two gametes carry alleles that are IBS (both G or both T) or not (one G and one T).

We can use this model with data on observed genotypes to estimate F (as described in BOX 3) or we can use this model to predict the probability that two alleles are IBS when predicting genotypes at an unobserved locus. The assumption of the model is that the same F applies in both cases so that F can be estimated from SNP data and then used in applications such as gene mapping. Several applications are described in the following sections.

Estimating relatedness from SNPs

Estimating IBD coefficients from high-density marker data is standard practice for many population-based studies (for an example see REF. 9). Methods such as those of Milligan[20] and Purcell *et al.*[6] are useful for identifying cryptic relatedness[21] or recent IBD, but they are not suited to estimating ancient IBD among distantly related individuals because they do not have a well-defined base population of unrelated individuals. Equation 1 can be used as a model, and data on genetic markers such as SNPs can be used to estimate F. By defining the current population as the base we can easily estimate the allele frequency q needed in equation 1. BOX 3 describes the formula for a single locus. Although we have described the relationship among gametes, the method is easily extended to diploid individuals (the formulae are given in BOX 3). The estimates from single loci can be simply averaged over the whole genome or over a part of the genome[22]. We call this estimator of the genome-wide relationship between individuals the raw unified additive relationship (raw UAR or $\hat{A}$). This estimate contains sampling error due to the finite number of SNPs that are used to estimate it; a better estimate can be obtained by regressing this raw UAR towards the identity matrix as explained in Yang *et al.*[22] and in BOX 3. We call the regressed or shrunk estimate of the relationship matrix the adjusted UAR ($\tilde{A}$). The adjusted UAR is unbiased in the sense that $E(A|\tilde{A}) = \tilde{A}$.

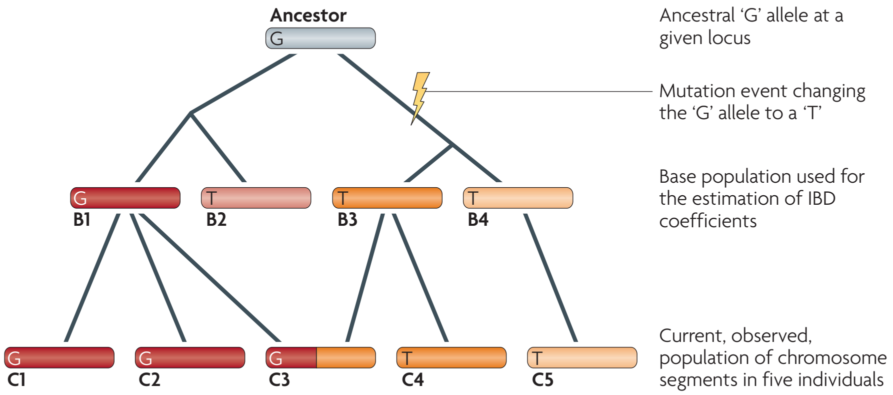
To illustrate this method we simulated 30 sets of sequence level data comprising 1,000 'unrelated' individuals in a randomly mating

Box 1 | **Illustration of the concepts of identity by descent and coalescence**

Here we illustrate the concepts of identity by descent (IBD), identity by state (IBS) and coalescence at a single nucleotide and at a chromosome segment. The figure depicts an ancestral allele at a locus, representing the point of coalescence for alleles in the current population (C1–C5). At the point of coalescence (the most recent common ancestor) this locus carries a copy of a G allele that is subject to a mutation event (G→T; lightning symbol) leading to a G/T polymorphism.

IBD at the polymorphic locus among individuals (C1–C5) can be defined with respect to a base population (B1–B4) in which individuals are assumed to be unrelated (shown by the differently coloured chromosome segments). Then the G alleles in C1, C2 and C3 are IBD to each other as all three descend from the G allele in B1. The T alleles in C4 and C5 are IBS but not IBD as they descend from different alleles in the base population.

The whole chromosome segments C1 and C2 are IBD because they descend from a common ancestor (B1) without recombination, but chromosome segment C3 is not IBD to C1 and C2.



Ancestral 'G' allele at a given locus

Mutation event changing the 'G' allele to a 'T'

Base population used for the estimation of IBD coefficients

Current, observed, population of chromosome segments in five individuals

© 2010 Macmillan Publishers Limited. All rights reserved

population[23,24]. Relatedness coefficients were estimated from a subset of polymorphisms, representing a high-density genotype panel, using four methods: standard relationship coefficients from ten generations of pedigree data[16]; standard IBD estimates using PLINK[6]; raw UAR; and adjusted UAR (BOX 3; see Supplementary information S1 (box)). The adjusted UAR provides an unbiased prediction of the numerator relationships (A) estimated from the pedigree as shown by the regression of A from pedigree on adjusted UAR being approximately 1.0. The raw UAR is slightly biased (regression coefficient = 0.97) owing to sampling error in the estimates of relationship. By contrast, the standard IBD function within PLINK (PLINK_IBD) estimates many of the relationship coefficients to be zero because it has no clearly defined base and therefore has difficulty estimating distant relationships. The estimates of relationship by UAR are more highly correlated with the pedigree relationships than are the estimates from PLINK.

Box 2 | Changing the base population for identity by descent coefficients

Sometimes the appropriate base population to use is obvious. For example, if we have a population consisting of full sib families but with no knowledge of the relationships between the parents of these families, then it is convenient to define the parents as the base population. However, for some purposes it may be desirable to change the base population, for instance, to the ancestors of these individuals who were alive in 1800. We can do this if we know the average relationship of gametes in the recent base population (that is, the parents) with respect to the older base population (that is, individuals alive in 1800). Let this probability be F_{st}, then the F between two gametes with respect to the old base population (F_{old}) is $F_{old} = F_{new} + (1 - F_{new})F_{st}$, in which F_{new} is the relationship with respect to the new base population. This equation can be solved to obtain a formula for F_{new} (equation 2) so that it is possible to transform between base populations in either direction.

$$F_{new} = \frac{F_{old} - F_{st}}{1 - F_{st}} \qquad (2)$$

There is another interpretation of F. For a biallelic locus, arbitrarily label one allele as 0 and the other allele as 1, then the allele in gamete i can be given the value x_i, in which this is either 0 or 1. Then F is simply the correlation between alleles taken from gametes with a relationship of F. Again this correlation is with respect to a base population. If the allele frequency in the base is p (which represents the mean of x in the base population), then $E(x_i - p)(x_j - p) = F_{ij}p(1 - p)$.

If the base population is old, genetic drift may mean that p is highly different from the allele frequency in the current population. It is possible to observe the allele frequency in the current population (u) but it may be extremely difficult to estimate p. However, we can define the 'new' base population as the current population with F_{new} as defined in equation 2. Then the correlation between alleles from two gametes with respect to the current population is F_{new} and $E(x_i - u)(x_j - u) = F_{new}u(1 - u)$. With this definition, the average value of F between alleles in the current population is per definition 0, and F between some pairs of gametes is negative. Consequently, this F cannot be interpreted as a probability; however, its interpretation as a correlation is straightforward.

Estimating genetic variance

Genetic variances and heritabilities for complex traits are traditionally estimated by using a relationship matrix that is calculated from a known pedigree. However, the relationship matrix can also be estimated from genome-wide genetic markers[25]. The genetic variance that is estimated is that assumed to have existed in the base population used to calculate the relationship matrix. Thus, if an ancient base is used, the genetic variance will be estimated in an ancient base. As the current population is assumed to be more inbred than an ancient base (BOX 2), the genetic variance in the ancient base will be estimated to be larger than the genetic variance in the current population. This makes the estimated genetic variance difficult to interpret. Therefore it is convenient to estimate the relationship matrix with the current population as the base and hence estimate the genetic variance in the current population.

Ideally, the genetic variance would be estimated using the relationship matrix at the causal loci controlling the trait. As these are typically unknown, we would like to use an unbiased estimate of this relationship; that is, the expected value of the relationship conditional on the marker data.

If the causal variants are unaffected by natural selection (that is, they are neutral), the relationship matrix estimated from pedigree is unbiased in this sense, as is the adjusted UAR, but the raw UAR estimated from equation 3 in BOX 3 is not because the variance of the raw UAR is inflated by sampling error due to the finite number of SNPs used in its calculation.

If causal variants are subject to selection, then relationships at these variants are systematically different to those at neutral markers. Although loci that harbour causal variants are typically unknown, we can investigate their properties by studying markers with different allele frequencies. For example, Yang *et al.*[22] showed how the relationship at markers with low minor allele frequencies differs from the relationship based on all markers. Consequently, if the causal variants had similar properties to low minor allele frequency SNPs the adjusted UAR underestimates the true heritability.

These principles are illustrated in BOX 4, in which simulated data are used to estimate heritability. When the causal variants are simulated to have similar properties to the SNPs and adjusted UAR is used, the genetic variance in the current population is estimated without bias. By contrast, using the relationships estimated by PLINK_IBD[6] or by the raw UAR gives biased estimates (BOX 4). The additive genetic value of individuals can be predicted by using the adjusted UAR in place of the pedigree-defined numerator relationship matrix. This is a form of "genomic selection"[26] used to predict breeding values in livestock[27,28]. The methods of calculating IBD probabilities described above can be used to calculate IBD probabilities over the whole genome, one chromosome or a small segment of chromosome and hence can be used to calculate the genetic variance due to a single chromosome or segment.

Estimating variance explained by SNPs

Estimating the genetic variance using the raw UAR as the relationship matrix (instead of the adjusted UAR) is equivalent to a model that fits the effects of all the SNPs[26–31]. Thus the genetic variance estimated when raw UAR is used is an estimate of the total genetic variance explained by the SNPs. This will tend to underestimate the full genetic variance, as shown in BOX 4, which is due to incomplete linkage disequilibrium (LD) between the SNPs and the causal variants. The statement that the relationship at the SNPs (raw UAR) is not a perfect estimate of the relationship at the causal variants is equivalent to the statement that the LD between the SNPs and the causal variant is incomplete. Thus the difference between the full genetic variance estimated using the pedigree-defined relationship

© 2010 Macmillan Publishers Limited. All rights reserved

matrix and the variance explained by the SNPs estimated using the raw UAR is the missing heritability[14,15]. For instance, we recently showed that genotyped SNPs on a commercial array collectively explain about 50% of the genetic variance for human height[22]. The SNPs have an apparent effect on the trait due to LD with causal variants. Therefore, if they only explain 50% of the variance, the LD between the causal variants and the SNPs is less than complete.

Gene mapping

Mapping methods. All gene mapping methods can be described in the following way[32,33]: for the possible positions of the causal locus, calculate the probabilities that individuals are IBS. Then choose the position at which the IBS probabilities most closely match the observed similarities in phenotype.

If the pedigree is unknown, then standard linkage analysis cannot be used; another method must be found to estimate the probability that individuals carry the same variants at a specified position. For instance, genome-wide association studies estimate the apparent effect of the marker relying on the assumption that a marker will only have a strong association with the trait if it is in high LD with the causative polymorphism and that this will only happen if the marker and causative polymorphism are close together on the chromosome. In this case, the apparent effect of the marker underestimates the

Box 3 | Estimating relatedness coefficients

For any pair of gametes, genotyped at a biallelic locus, we can use equation 1 as a model and use it to estimate F. If x_i is the allele carried by gamete i coded as either a 0 or 1, with allele frequencies in the base population of q and p, then the gametic relationship F between gametes 1 and 2 can be estimated by equation 3:

$$\hat{F} = \frac{(x_1 - p)\,(x_2 - p)}{pq} \qquad (3)$$

This is consistent with the expectation of F given in BOX 2. This is the minimum variance, unbiased estimate with sampling variance = 1, if true $F \approx 0$. It is easy to calculate if we define the base as the current population because then p is the allele frequency in the current population. If genotype data exist on many loci for a pair of gametes, the $\hat{F}$ for that pair can simply be averaged over the loci yielding an estimate of the average F. In this way F can be estimated for the whole genome or a segment of the genome. Similar estimators have been used previously[30,31,40,41]. The sampling variance of $\hat{F}$ is then $1/m$, in which m is the number of markers used.

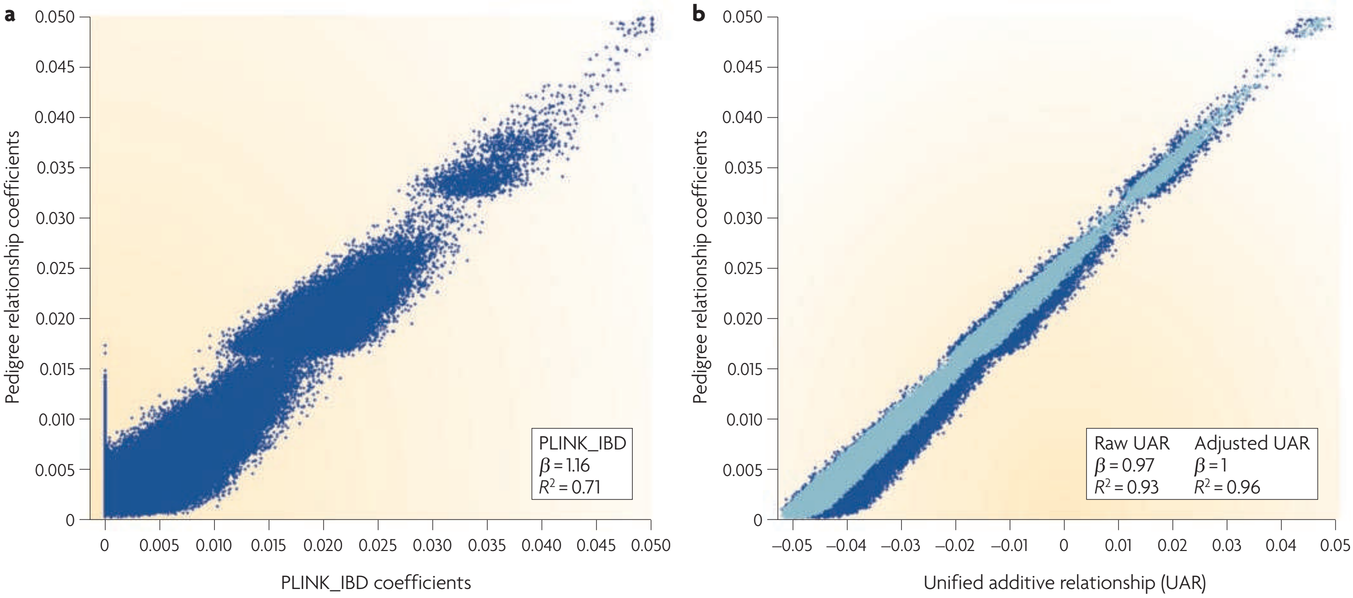
$\hat{F}$ is unbiased in the classical sense that $E(\hat{F}|F) = F$. That is, if we sample many pairs of gametes with the same true F then the average value of $\hat{F}$ will be F. These estimates will be distributed around F with a variance of $1/m$. However, unbiasedness in this sense is not the property that we want in an estimate of F[42]. We want to estimate F at unobserved loci and the best estimator of this is the expected value of F conditional on the markers data. That is, we want an estimate with the property $E(F|\tilde{F}) = \tilde{F}$. Yang *et al.*[22] showed how this can be empirically achieved by using the regression (β) of F on $\hat{F}$ to produce an estimate with the desired property $\tilde{F} = \beta\hat{F}$. That is, $\tilde{F}$ is a 'shrunk' or 'regressed' version of $\hat{F}$.

To estimate the relationship between diploid individuals it is not necessary to divide the genotypes into haploid alleles. This estimator of relationship between two distinct diploid individuals (i and j) can easily be calculated from genotype data as outlined in equation 4, in which x is the sum of the two alleles within an individual.

$$\hat{A} = \frac{(x_i - 2p)\,(x_j - 2p)}{2pq} \qquad (4)$$

For the relationship of an individual with itself, $\hat{A}$ should be estimated as $1 + \hat{F}$, in which $\hat{F}$ is the estimate of the relationship between the two gametes constituting the individual. This estimate of A should be regressed in the same way as $\hat{F}$. We call the estimators $\hat{A}$ and $\tilde{A}$ the raw unified additive relationship (UAR) and the adjusted UAR, respectively. This estimator of F or A can be applied to the whole genome or to a segment containing one or more markers.

We simulated data on a SNP panel (described in Supplementary information S1 (box)) and calculated four estimates of the relationship between each pair of 1,000 individuals: pedigree relationship coefficients[16]; PLINK_identity by descent (IBD)[6]; raw UAR[22]; and adjusted UAR[22]. Further details of methods used to estimate the relatedness coefficients are given in Supplementary information S1. The results are presented here in the figure. A large proportion (0.49) of IBD coefficient estimates from PLINK_IBD are zero (**a**), leading to high regression (β) and low correlation (R^2) coefficients. The adjusted UAR gives an unbiased estimate of pedigree relationship ($\beta = 1$) and the highest correlation with pedigree relationship. Correlations from both raw UAR (dark blue) and adjusted UAR (light blue) coefficient estimates are shown in (**b**).



© 2010 Macmillan Publishers Limited. All rights reserved

effect of the causal variant if the LD is not complete. An equivalent way to describe an association study would be to say that individuals carrying identical marker alleles are likely to carry identical causal alleles and therefore be similar in phenotype.

Alternatively, homozygosity mapping uses haplotypes that are IBS to imply that unobserved sites within the haplotype are likely to be IBS. An individual carrying two IBS haplotypes is therefore likely to be homozygous at additional sites within the haplotype that might contain a recessive allele for the trait under study[34].

The benefits of a unified approach. Linkage analysis, single-marker association methods and haplotype methods all use the concept of IBS at unobserved putative gene positions but often do not attempt to estimate this probability formally. A unified approach to all gene mapping methods can be provided by formally calculating the probabilities that individuals carry IBS alleles at a putative site. Then the likelihood of the phenotypic data given this set of probabilities can be calculated and the site with the highest likelihood chosen[33].

Predicting inbreeding depression

Many traits, especially those related to fitness, show a decline in the mean in inbred individuals, which is known as inbreeding depression[8]. This is thought to be largely due to increased homozygosity at loci in which a deleterious allele is at least partially recessive. In this case, the mean trait value declines in proportion to the inbreeding coefficient. The inbreeding coefficient can be calculated from the pedigree if it is known. In the absence of a known pedigree, the inbreeding coefficient can be calculated from genetic markers[35,36]. The inbreeding coefficient is the gametic relationship between the two gametes forming the individual, so the methods for estimating it are the same as for estimating the relationships between individuals. As before, we can use a method that estimates inbreeding for each SNP, averages these estimates over all SNPs and regresses the estimate to allow for sampling error.

Box 4 | Estimating heritability from marker-derived relationship coefficients

A relationship matrix, calculated from genetic markers, can be used to estimate the genetic variance and hence heritability of a complex trait. To obtain an unbiased estimate of genetic variance, the estimate of the relationship matrix should be unbiased in the sense $E(A|\tilde{A}) = \tilde{A}$[22] as explained in BOX 1 and illustrated in BOX 3.

	Causal variant scenario	
	MAF > 0.05	MAF < 0.05
Raw UAR	0.435 (0.012)	0.412 (0.014)
Adjusted UAR	0.495 (0.011)	0.427 (0.012)
PLINK_IBD	0.362 (0.027)	0.315 (0.029)

For each of the simulated data sets described in BOX 3 we generated a quantitative phenotype based on genotypic data with heritability (h^2) = 0.5. Random samples of 500 SNPs were chosen to represent causal variants from the set of polymorphisms not included in the generated genotype panels of the simulated data. Causal variants were selected from two minor allele frequency (MAF) scenarios. The first scenario is randomly sampled causal variants from polymorphisms with MAF > 0.05. This leads to a causal variant MAF distribution similar to that of the SNPs in the genotype panel. The second scenario is randomly sampled causal variants from the SNPs with MAF < 0.05, which represents rare variants in a standard population genetics framework. Details of the methods used to generate phenotypic values are given in Supplementary information S3 (box).

Using the simulated phenotypes, we estimated h^2 (details of estimating h^2 are given in Supplementary information S3) using the relatedness coefficients described in BOX 3. Results are summarized here in the table. In the first scenario, using the raw unified additive relationship (raw UAR or $\hat{A}$) coefficients, we underestimate heritability owing to incomplete linkage disequilibrium (LD) between SNPs and causal variants (mean standard errors are shown in brackets). However, by correcting for sampling error caused by estimating genetic variance from a finite number of SNPs[18] (see Supplementary information S2 (box)) we obtain an unbiased estimate of the genetic variance. In the second scenario, causal variants have lower MAFs than SNPs on average and less LD than between markers and causal variants, thereby leading to underestimation of h^2 even when adjusted UAR (or $\tilde{A}$) is used. The use of $\tilde{A}$ instead of $\hat{A}$ corrects for the sampling error in $\hat{A}$, but it does not correct for systematic differences between SNPs and causal variants. Estimates of h^2 using PLINK_IBD coefficients lead to a lower estimation of h^2, with larger standard errors than those obtained from our unified method (UAR).

Homozygosity across the genome. However, it is homozygosity at the causal variants that show dominance that is important rather than the homozygosity at SNPs. Selection operating at these causal variants may affect the evolution of allele frequency at such loci so that the relationship at such sites is systematically different to that at common SNPs. For instance, if the mutant allele at such a locus is selected against, it is likely to be eliminated from the population after some generations. Consequently, mutants that are segregating in the population are likely to be evolutionarily young. This means that relationships based on an ancient common ancestor may be poor predictors of inbreeding depression because the mutation creating the causal variant will have occurred since the common ancestor. Chromosome segments that share a long common haplotype are likely to derive from a recent common ancestor and so be a better predictor of homozygosity at the relevant causal variants than relationships based on similarity at single SNPs[36,37].There has been little research to discover how best to predict inbreeding depression, but the argument above shows that it involves the same concepts as the estimation of the additive genetic relationship.

Homozygosity at specific sites. Inbreeding depression usually refers to the effect of homozygosity over the whole genome, but the same phenomenon applies to specific sites in the genome in which a recessive mutation is segregating[38]. Logically these sites would be mapped using a method that estimated the probability that an unseen locus was homozygous, but often informal methods such as runs of homozygosity (ROH) among genetic markers are used. ROH SNPs in individuals are used to estimate regions of the genome that are inherited from a common ancestor and subsequently to see whether these runs are correlated with a phenotype[37,39]. Such approaches are based on the idea that a long ROH implies recent inbreeding and harmful recessives may be contained within the segment[8]. These analyses implicitly assume that runs of SNPs that are IBS are also IBD[10,39]. The appropriate question to ask, however, is whether the intervening, unobserved chromosome stretches (for example, causal variants) within the ROH are also identical. To illustrate this principle we show the probability that unobserved SNPs within a ROH are homozygous (see Supplementary information S2 (box)). The illustration shows that the longer the ROH the higher the probability that unobserved variants within the segment are also homozygous.

Conclusions

IBD is a widely used concept in genetics but it requires the definition of a base population. Traditionally the probability that alleles were

© 2010 Macmillan Publishers Limited. All rights reserved

IBD was predicted from pedigrees and the base population consisted of the founders of the pedigree. However, using dense SNP data it is possible to estimate IBD probabilities without knowledge of pedigrees. In many cases the most convenient base to use is the current population. The use of the current generation as the base causes some of the relationships to be negative and so they cannot be interpreted as probabilities but they can be interpreted as the correlation of homologous alleles in different gametes.

The purposes to which IBD relationships are used rely on their ability to predict the probability that alleles at unobserved loci, such as those harbouring causal variants, are IBS. Thus methods to estimate IBD relationships should in fact aim to estimate the probabilities that gametes are IBS at unobserved loci. This Perspective eliminates the conflict between IBD methods and coalescent methods: coalescent methods are examples of methods that can be used to estimate the probability that gametes are IBS. The same methods for estimating IBD relationships can be used for gene mapping, for estimating genetic variance and for predicting inbreeding depression, thereby leading to a more consistent approach and more accurate results.

Glossary

Coalescence theory

A population genetics model of inheritance relationships among alleles at a given locus. The coalescence of two alleles is the most recent point (going back in time) at which they shared a common ancestor.

Cryptic relatedness

The presence of close relatives in a sample of ostensibly unrelated individuals. It is characterized by a recent common ancestry that can be revealed from marker-based relatedness coefficients.

Genome-wide association study

Analysis of the entire genome using association models to identify regions of the genome that contribute to genetic variation in a phenotype. These studies typically analyse data from high-density SNP arrays.

Heritability

The proportion of phenotypic variation in a population that is attributable to genetic variation among individuals. Statistical methods are used to estimate the relative contributions of differences in genetic and non-genetic factors to the total phenotypic variation in a population.

Identity by descent

(IBD). Two or more alleles are IBD if they are identical copies of the same ancestral allele in a base population. IBD can be estimated for alleles at single loci in a diploid individual or between individuals.

Identity by state

(IBS). Refers to two or more alleles that 'look' the same. For example, if two individuals both carry a 'G' allele at a specific locus.

Pedigree

A population of individuals in which the mating records for multiple generations are known. Pedigree information is typically available for livestock populations, in which controlled breeding has been implemented to maximize the response to genetic selection.

Joseph E. Powell and Peter M. Visscher are at the Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia.

Michael Goddard is at the Department of Agriculture and Food Systems, University of Melbourne, Royal Parade, Parkville, Victoria 3010, Australia, and at the Department of Primary Industries, 1 Park Drive, Bundoora, Victoria 3083, Australia.

Correspondence to J.E.P. or P.M.V.
e-mails: Peter.Visscher@qimr.edu.au; Joseph.Powell@qimr.edu.au

All authors contributed equally to this work.

doi:10.1038/nrg2865
Published online 28 September 2010

1. Malécot, G. *Les Mathématiques de l'Hérédité*. (Masson, Paris, 1948).
2. Wright, S. Coefficients of inbreeding and relationship. *Am. Nat.* **51**, 636–639 (1917).
3. Wright, S. Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6**, 111–123 (1921).
4. Kong, A. *et al*. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
5. Albrechtsen, A. *et al*. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* **33**, 266–274 (2009).
6. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
7. Visscher, P. M. *et al*. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, 316–325 (2006).
8. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature Rev. Genet.* **10**, 783–796 (2009).
9. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–862 (2007).
10. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer Associates, Massachusetts, 1997).
11. Rosenberg, N. A. & Nordborg, M. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390 (2002).
12. Nordborg, M. in *Handbook of Statistical Genetics* (eds Balding, D. J., Bishop, M. & Cannings, C.) 179–212 (Wiley, Chichester, 2001).
13. Browning, S. R. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* **178**, 2123–2132 (2008).
14. Manolio, T. A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
15. Maher, B. The case of the missing heritability. *Nature* **456**, 18–21 (2008).
16. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Massachusetts, 1998).
17. Jacquard, A. *The Genetic Structure of Populations* (Springer, New York, 1974).
18. Hayes, B. J., Visscher, P. M., McPartlan, H. C. & Goddard, M. E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**, 635–643 (2003).
19. Slatkin, M. Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175 (1991).
20. Milligan, B. G. Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153–1167 (2003).
21. Astle, W. & Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
22. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human. *Nature Genet.* **42**, 565–571 (2010).
23. Chadeau-Hyam, M. *et al*. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics* **9**, 364–375 (2008).
24. Hoggart, C. J. *et al*. Sequence-level population simulations over large genomic regions. *Genetics* **177**, 1725–1731 (2007).
25. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
26. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
27. Goddard, M. E. & Hayes, B. J. Genomic selection. *J. Anim. Breed. Genet.* **124**, 323–330 (2007).
28. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**, 47–60 (2009).
29. Goddard, M. E. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257 (2009).
30. Habier, D., Fernando, R. L. & Dekkers, J. C. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397 (2008).
31. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
32. Goddard, M. E. & Meuwissen, T. H. E. The use of linkage disequilibrium to map quantitative trait loci. *Aust. J. Exp. Agric.* **45**, 837–845 (2005).
33. Meuwissen, T. H. E. & Goddard, M. E. Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* **176**, 2551–2560 (2007).
34. Lander, E. S. & Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
35. Lynch, M. & Ritland, K. Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766 (1999).
36. Carothers, A. D. *et al*. Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. *Ann. Hum. Genet.* **70**, 666–676 (2006).
37. McQuillan, R. *et al*. Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
38. Charlesworth, B. & Charlesworth, D. The genetic basis of inbreeding depression. *Genet. Res.* **74**, 571–576 (1999).
39. Broman, K. W. & Weber, J. L. Long homozygous chromosome segments in reference families from the centre d'étude du polymorphisme humain. *Am. J. Hum. Genet.* **65**, 1493–1500 (1999).
40. Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K. & Goddard, M. E. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* **41**, 51 (2009).
41. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
42. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* **24**, 517–529 (2009).

Acknowledgements

We acknowledge funding from the Australian National Health and Medical Research Council (grants 389892, 613672 and 613601) and the Australian Research Council (grants DP0770096 and DP1093900). We thank M. Keller for discussions and advice on simulation.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Genetic Epidemiology, Molecular Epidemiology and Queensland Statistical Genetics Laboratories Brisbane, Australia: http://genepi.qimr.edu.au
Nature Reviews Genetics series on Study Designs: http://www.nature.com/nrg/series/studydesigns/index.html

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box) | S3 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2010 Macmillan Publishers Limited. All rights reserved

Supplementary information S1 | Simulation of genotypic data and estimation of relatedness coefficients

Simulation of genotypic information

The forward in time simulation program Fregene[1, 2] was used to generate samples of sequence level data from a Wright-Fisher panmictic, non-selfing, diploid population of constant size. The founder generation consisted of individuals with identical chromosome segments of 50Mb, with polymorphisms arising through mutation events over 100,000 generations of random mating. Alleles had an equal probability of mutation and each mutation event was recorded uniquely, giving an infinite alleles model. A per-site mutation probability (μ) = $2.3e^{-08}$ was based on values estimated from human populations[3, 4]. Pedigree information was recorded for the final 10 generation of simulation.

From the sequence genotype data markers were chosen to represent typed SNPs in a 'genotype panel'. The SNP panel was generated by randomly sampling SNPs from MAF bins (9 bins of equal range 0.05 – 0.1 ..., 0.45 – 0.5), generating an approximately uniform distribution of SNP minor allele frequencies (MAF). Markers with MAF < 0.05 were omitted from the genotype panel. Genotype panels comprised of a total of 5000 SNPs, representing a density approximately equal to a 300K chip in humans (on average 1 SNP every 10kb). The remaining SNPs in the simulated data, not included in the genotype panel, represent hidden variation on chromosome stretches between markers. Two plink data objects were made, one comprising of SNPs in the "SNP panel" and the other comprising of all SNPs in a dataset. An index file was also generated to list which SNPs were included in the genotype panel.

Estimating relationship matrices

PLINK_IBD – The PLINK method of estimating IBD applies a hidden Markov model to estimate IBD from IBS assuming independence among markers[5]. We applied the standard method suggested in the PLINK documentation using the default setting to estimate the IBD coefficients. SNPs were pruned with a window size 100 (adjacent SNPs) and a step size of 25 (SNPs) with an r^2 threshold of 0.2. The resulting number of SNPs remaining is approximately 2000 in each dataset.

Pedigree relationship coefficients – A-matrix coefficients were formed by calculating twice the kinship coefficient between individuals i and j using pedigree information from the final 10 generations of the simulation[6]. Relationship coefficients were calculated for all individuals in the pedigree (10 generations) although only coefficients from the final generation were used for comparisons against other relationship coefficients.

© 2010 Macmillan Publishers Limited. All rights reserved.

Raw UAR – Pairwise genetic relationships were estimated using all markers in the genotype panel (MAF > 0.05). For individuals i and j the weighted average across SNPs is calculated by; $A_{ij} = \frac{1}{m}\sum_k \frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$, where x is an indicator for SNP k in individuals i and j, p_k is the allele frequency (frequency of either allele) of SNP k, and m is the number of SNPs in the sample. The genetic relationship for an individual with its self is given by; $A_{ii} = 1 + \frac{1}{m}\sum_k \frac{x_{ik}^2-(1+2p_k)x_{ik}+2p_k^2}{2p_k(1-p_k)}$. Individuals in the sample (final generation) were used as a base population such that the mean of the off-diagonal elements of the relationship matrix is zero and the mean of the diagonal elements is 1[7]. Further details are given in Yang *et al.*[7].

Adjusted UAR – Adjusted estimates of genetic relatedness were corrected for the sampling error of the raw UAR (sampling variance = $1/m$) by the regression of true A on $\hat{A}(1/m)$. The regression coefficient is $var(A)/(var(A)+1/m)$ and the $var(\hat{A}) = var(A) + 1/m$, so the adjusted UAR is

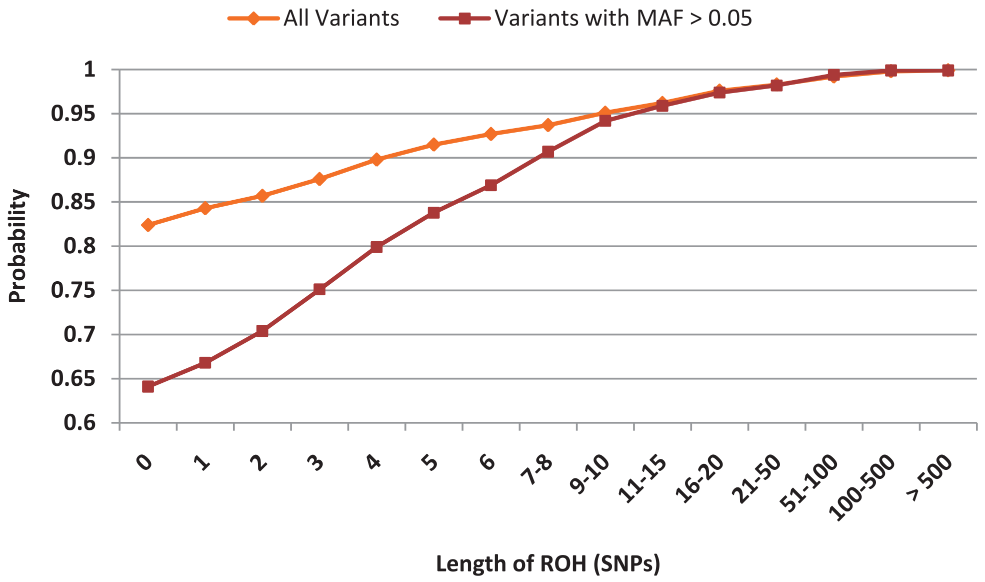
$$A_{ij}^{*} = \begin{cases} \left(1-\frac{\frac{1}{m}}{var(\hat{A})}\right)\hat{A}\ , i \neq j \\ 1+\left(1-\frac{\frac{1}{m}}{var(\hat{A})}\right)(\hat{A}-1)\,, i = k \end{cases}$$

References

1. Hoggart, C. J. *et al*. Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725-1731 (2007).
2. Chadeau-Hyam, M. *et al*. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinfo.* **9**, 364-375 (2008).
3. Nachman, M. W. & Crowell, S. J. Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297-304 (2000).
4. Reich, D. E. *et al*. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet*. **32**, 135-142 (2002).
5. Purcell, S. *et al*. PLINK: A tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
6. Lynch, M. & Walsh, B. Genetics and analysis of quantitative traits. *Sinauer Associates*, Massachusetts (1998).
7. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human. *Nat. Genet.* **42**, 565-571 (2010).

© 2010 Macmillan Publishers Limited. All rights reserved.

Supporting information S2 | Probability that a hidden SNP within a ROH is homozygous



Simulation of genotype data

Here genotypic data were simulated using similar principles to the generation of genotype information described in **S1**, with the addition that population demographics and recombination probability were included to reflect patterns estimated within human populations[1, 2]. As is described in **S1** 30 independent datasets of sequence level data were simulated using the forward in time simulation program Fregene[3, 4]. These datasets were simulated with the following population parameters;

1. Chromosome length = 50Mb
2. Per-site mutation probability (μ) = 2.3e^{-08}
3. Per-site recombination probability (r) = 1.1e^{-08}
4. Proportion of recombination occurring within hotspots = 80%
5. Hotspot length = 2.0 kb
6. Mean distance between hotspots = 9 kb
7. An initial population size of 10,000 remained constant for 70,000 generations. This was followed by a bottleneck event in generation 70,001 to 2000 individuals, and subsequently an exponential population expansion for a further 2000 generations.

From the simulated data the same procedure as described in **S1** was used to generate the genotype panel. Within the genotype panel of each dataset runs of homozygosity (ROH) were identified via the Runs of Homozygosity option implemented in PLINK[5] and divided into groupings based on their

© 2010 Macmillan Publishers Limited. All rights reserved.

maximum length in number of SNPs. ROH of length = 1 represents homozygous hidden SNPs located between homozygous and heterozygous SNPs in the genotype panel, whilst, ROH of length = 0 are homozygous hidden markers between two heterozygous genotyped SNPs. Within a ROH the probability that hidden SNPs, not included in the genotype panel, are homozygous was calculated from the proportions that were homozygous, averaged across the 30 datasets. This probability was calculated for all hidden SNPs (all variants), and a subset of hidden SNPs with MAF > 0.05. Mean homozygosity (across all datasets) was 0.11 (all variants), and 0.27 (variants with MAF > 0.05).

References

1. Nachman, M. W. & Crowell, S. J. Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297-304 (2000).
2. Reich, D. E. *et al*. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet*. **32**, 135-142 (2002).
3. Hoggart, C. J. *et al*. Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725-1731 (2007).
4. Chadeau-Hyam, M. *et al*. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinfo.* **9**, 364-375 (2008).
5. Purcell, S. *et al*. PLINK: A tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

© 2010 Macmillan Publishers Limited. All rights reserved.

Supplementary information S3 | Estimating heritability using relatedness coefficients

Simulating a phenotype

Quantitative trait phenotypes with a heritability (h^2) of 0.5 were simulated for each of the 30 datasets. Random samples of 500 causal variants were selected from two MAF scenarios: I) randomly sample causal variants from polymorphisms with MAF > 0.05. By sampling only from polymorphisms with MAF > 0.05 their allele frequency distribution will similar to that of the SNPs in the genotype panel; II) randomly sample causal variants from the SNPs with MAF < 0.05, representing rare variants in a standard population genetics framework. Whilst this represents a high number of causal variants for a region representing a chromosome, it means the total effect attached to each casual variant will remain small. The phenotypic value for each individual was determined as follows;

1) The effect (α) of each causal variant was drawn from a standard normal distribution
2) The overall genetic effect of an individual (g_i) was calculated by; $g_i = \sum_{ik} x_{ik}\alpha_{ik}$ where x_{ik} is an indicator for the number of 'A' alleles carried at causal variant k by individual i.
3) Residual effects (e_i) for each individual were drawn from a normal distribution (mean = 0) and variance $\left(\frac{1}{h^2}\ \sigma_g^2\right) - \sigma_g^2$ where σ_g^2 is the variance of g in the dataset.
4) The phenotypic value for each individual was calculated by; $y_i = g_i + e_i$.

Estimating h^2 from markers

Each of the 30 datasets consists of 1000 'unrelated' individuals simulated under a Wright-Fisher panmictic, non-selfing, diploid population of constant size, as described in **S1**. Pairwise genetic relationships were estimated from markers within the SNP panel for individuals within each dataset using raw UAR, adjusted UAR and PLINK_IBD methods. Each of these methods produces a 1000 by 1000 relationship matrix (A) based on the SNPs in the simulated genotype panel (**S1**). Each pair of individuals i and k has a relationship coefficient $A_{i,k}$. We fitted a linear model to the simulated phenotype and used restricted maximum likelihood[1, 2] (REML) analyses to estimate the variance explained by the markers within the genotype panel. This process was repeated for each of the methods used to estimate $A_{i,k}$.

Estimates of h^2 can be less than the true h^2 due to;

1) Sampling error associated with estimating UAR from m SNPs.
2) Incomplete linkage disequilibrium (LD) between causal variants and SNPs. This is exacerbated when MAF distributions of markers and causal variants differ.

© 2010 Macmillan Publishers Limited. All rights reserved.

Note: Yang *et al.*[3] established an empirical linear relationship between β and *N*; $\beta = 1 - \frac{(c+\frac{1}{N})}{var(A_{ij})}$ where *c* is a constant depending on the causal variant MAF threshold.

References

1. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554 (1971).
2. Gilmour, A. R. *et al*. ASReml User Guide Release 2.0. VSN International, Hemel Hempstead, UK (2006).
3. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human. *Nat. Genet.* **42**, 565-571 (2010).

© 2010 Macmillan Publishers Limited. All rights reserved.

View publication stats