Relatedness in the post-genomic era: is it still useful?

Doug Speed[1] *and David J. Balding*[1,2]

Abstract | Relatedness is a fundamental concept in genetics but is surprisingly hard to define in a rigorous yet useful way. Traditional relatedness coefficients specify expected genome sharing between individuals in pedigrees, but actual genome sharing can differ considerably from these expected values, which in any case vary according to the pedigree that happens to be available. Nowadays, we can measure genome sharing directly from genome-wide single-nucleotide polymorphism (SNP) data; however, there are many such measures in current use, and we lack good criteria for choosing among them. Here, we review SNP-based measures of relatedness and criteria for comparing them. We discuss how useful pedigree-based concepts remain today and highlight opportunities for further advances in quantitative genetics, with a focus on heritability estimation and phenotype prediction.

Relatedness
Two individuals are related if they have a recent common ancestor, where 'recent' can be variously defined as outlined under IBD (identity-by-descent).

Traditional measures of relatedness, which are based on probabilities of IBD (identity-by-descent) from common ancestors within a pedigree, depend on the choice of pedigree. However, in natural populations there is no complete pedigree or any optimal pedigree that could form the basis of a canonical definition of relatedness. Moreover, the random nature of recombination during meiosis means that expected genome sharing specified by IBD probabilities is an imprecise guide to actual genome sharing: human half-siblings are expected to share half of each chromosome that they received from their common parent, but the 95% credible interval for their actual amount shared ranges from 37% to 63% (see below). More recent approaches use genetic markers either to estimate IBD probabilities in an unobserved pedigree, of which founders are assumed to be unrelated, or to identify shared genomic regions that are unaffected by recombination since their most recent common ancestor (MRCA). These approaches also suffer from difficulties (see below). Although the problems in defining and measuring relatedness have been appreciated by some authors[1–4], no better approach has gained widespread acceptance, and the implications of different approaches for applications are rarely noted.

Genome-wide single-nucleotide polymorphism (SNP) data now allow us to measure realized genome sharing with great accuracy and without reference to pedigree-based concepts, but this advance brings new problems. First, two haploid human genomes are typically identical at >99.9% of sites owing to shared inheritance from common ancestors. However, sequence identity across SNPs is much lower. SNP-based measures of genome similarity will depend sensitively on the minor allele fractions (MAFs) of the SNP set, which reflect both choice of SNP genotyping technology and the quality control procedures used. Even when the SNP set is fixed, there remain many ways to measure the similarity of genomes, and we lack criteria to choose among them. The usual statistical criteria of bias and precision of an estimator are less useful for natural populations because of the lack of an interpretable relatedness parameter to be the target of estimation (see below).

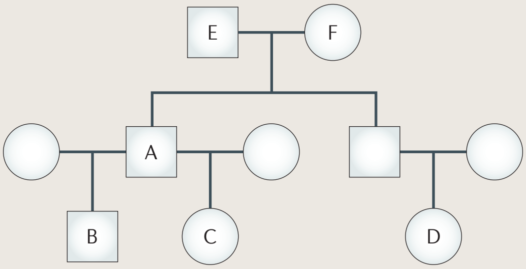
In this Review, we argue that IBD-based concepts of relatedness are now of limited value in genetics. Many previous uses can now be replaced by models and analyses that are based directly on actual genome similarity, although much work remains to be done to define concepts and to evaluate measures of genome similarity in the post-pedigree era. We begin with a recap and critique of pedigree-based relatedness, but our main focus is on SNP-based measures. Therefore, we complement a previous review that focused on pedigree-based relatedness[5]. We also discuss how relatedness can be defined in terms of genome-wide distributions of time since the MRCA (TMRCA). This seems to provide the most promising route to a

[1] *UCL Genetics Institute, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK.*
[2] *Present address: Department of Genetics and Department of Mathematics and Statistics, University of Melbourne, Parkville VIC 3010, Australia.*
Correspondence to D.J.B.
e-mail: david.balding@unimelb.edu.au

doi:10.1038/nrg3821
Published online 18 November 2014

Box 1 | Defining and measuring pedigree-based relatedness

Relatedness is popularly understood in terms of the shortest lineage path (or paths) linking two individuals; for example, cousins are linked by two lineage paths each of length four. However, pairs of individuals are linked by many lineage paths. Coefficients that incorporate all lineage paths within a specified pedigree have been developed since the 1940s[29,85,86]. The kinship coefficient (or coancestry) of individuals B and C is the probability that two homologous alleles, one drawn from each of two individuals, are IBD (identical-by-descent). It can be computed as follows.



$$\theta(B,C) = \sum_{A} \frac{1 + f_A}{2^{g_A+1}} \quad (13)$$

In this equation, the sum is over every most recent common ancestor (A) of B and C in the pedigree ('most recent' means that no descendant of A is also a common ancestor of B and C), g_A is the number of parent–child links in the lineage path linking B and C via A, and f_A is the inbreeding coefficient of A, which equals the coancestry of its parents.

In the pedigree shown (see the figure), assuming unrelated founders, half-siblings B and C have one common ancestor A with $g_A = 2$, so $\theta(B,C) = 1/2^3 = 1/8$. Cousins B and D have two common ancestors E and F, with $g_E = g_F = 4$, so $\theta(B,D) = 2 \times (1/2^5) = 1/16$. If in fact E and F were related with $\theta(E,F) = 1/20$, then $f_A = 1/20$ and $\theta(B,C) = (1+1/20)/2^3 = 21/160$.

$\theta(B,C)$ can be interpreted as the expected IBD fraction for two alleles at a locus, one each from B and C. There are 15 possible IBD states for the four alleles of B and C, which reduce to 9 if we regard each individual's two alleles as unordered[87], and reduce further to 3 (IBD = 0,1 or 2) if $f_B = f_C = 0$ (REF. 88). θ can then be expressed as follows, where $\phi = \text{P[IBD} = 1]$ and $\Delta = \text{P[IBD} = 2]$.

$$\theta = \frac{\text{E[IBD]}}{4} = \frac{\phi}{4} + \frac{\Delta}{2} \quad (14)$$

If $\Delta > 0$, then the relationship is bilineal and can help to assess the contribution of dominance to the genetic architecture of traits. Full siblings have $\Delta = 0.25$ and so they have matching genotypes owing to the shared parents at ~25% of the genome. The variance in the IBD fraction[89,90] can also be useful in distinguishing between relatives: in an outbred pedigree ($f_X = 0$ for all X), half-siblings, uncle–niece and grandparent–grandchild all have $\phi = 1/2$ and $\Delta = 0$, but uncle–niece have two common ancestors each with $g_A = 3$, which implies more but shorter regions shared IBD and therefore a lower variance than for the other two relationships (which have one common ancestor with $g_A = 2$). Early work on the distribution of lengths of IBD regions is referred to as the theory of junctions[91].

satisfactory conceptual definition of relatedness, but its practical usefulness has not yet been well explored. Finally, we review the use of relatedness in heritability estimation and phenotype prediction to capture the polygenic contribution to a complex trait, and discuss some implications of moving from pedigree-based to SNP-based measures of relatedness.

Throughout this Review, we refer to two sets of simulations that both use the Decode Genetic Map[6], which specifies male and female recombination rates over 2,667 Mb across the 22 autosomes. For simulations of Type A, we generate sequence data for pairs of individuals with one or two recent common ancestors, and examine the fraction of DNA that the two individuals share IBD from the ancestor (or ancestors). For Type B simulations, we generate sequence data from a Wright–Fisher population of 5,000 males and 5,000 females simulated over 50 generations from unrelated founders. The mating pattern is modified so that the probabilities for a female to have 0, 1, 2 or >2 children are 0.22, 0.20, 0.26 and 0.31, respectively, which is similar to Australian census data[7]; if two individuals have the same mother, then the probability that they have the same father is ~0.62 (see Supplementary information S1 (box)).

Pedigree-based relatedness

The classical theory of kinship coefficients based on lineage paths in pedigrees (BOX 1) provides a mathematically beautiful structure that has historically been useful, but its weaknesses are apparent. Pedigree founders are typically assumed to be unrelated, but this is only realistic in certain settings, such as some designed breeding programmes or an isolated population created by a specific founding event. All pairs of individuals with no common ancestor in the pedigree have coancestry (θ) of zero, but in practice they can have important differences in genome similarity. To overcome these problems, it may seem desirable to seek ever-larger pedigrees but, if we continue to add additional ancestors to an existing pedigree, then the co-ancestries of the original pedigree members will continue to increase and will eventually converge to one, which would be useless in practice. The lack of a complete or an ideal pedigree means that the choice of pedigree and hence any resulting kinship values are arbitrary to some extent. Similarly, pedigree-based inbreeding coefficients are of limited value: they also converge to one as the pedigree information increases and therefore make sense only with respect to a truncation of the pedigree, for example, at G generations before present[8,9]. Even then, interpretability remains a problem

IBD
(Identity-by-descent; also identical-by-descent). The phenomenon whereby two individuals share a genomic region as a result of inheritance from a recent common ancestor, where 'recent' can mean from an ancestor in a given pedigree, or with no intervening mutation event or with no intervening recombination event.

Pedigree
A set of individuals connected by parent–child relationships.

Most recent common ancestor
(MRCA). Although the ancestries of two alleles may both pass through the same individual, they pass through different alleles with probability 0.5, in which case that individual is not the MRCA of the alleles.

Time since the MRCA
(TMRCA; in generations). If the times back to a common ancestor differ between two individuals, then the average is used.

Heritability
The proportion of phenotypic variation that can be attributed to any genetic variation (broad-sense heritability) or to additive genetic variation (narrow-sense heritability (h^2)).

Lineage paths
Sequences of parent–child steps linking individuals with length equal to the number of steps.

Coancestry
(θ). A kinship coefficient defined as the probability that two homologous alleles, one drawn from each of two individuals, are IBD (identical-by-descent).

Inbreeding coefficients
The coancestries of the two parents of an individual.

because generations are typically not well defined (multiple lineage paths to a common ancestor can have different lengths) and G is arbitrary, yet shared DNA is treated very differently if the sharing originates in, for example, generation $(G-1)$ rather than generation $(G+1)$.

TABLE 1 shows the variability of θ' — the realized IBD fraction from the specified common ancestors in a Type A simulation — about its expected value θ. Indeed, differences in θ' can be exploited to estimate narrow-sense heritability (h^2) using pairs of individuals with the same θ, such as siblings[10,11] or even unrelated individuals ($\theta = 0$) (REF. 12). This contrasts with traditional h^2 estimates that require pairs of individuals with different θ, such as monozygotic and dizygotic twin pairs[13]. The table also reports $P[\theta' > 0]$, the probability of any IBD from the specified common ancestors. For example, two children with a common great-grandmother ($G = 3$; $A = 1$) will each have substantial genome-sharing IBD with her but could, in effect, not be related to each other. Although they are expected to share ~100 Mb IBD over ~7 regions, there is a probability of 0.005 that they share no DNA from her, despite the pedigree relationship. The values for $P[\theta' > 0]$ in this case are similar to those in a previous report[14] that assumed a sex-averaged human genetic map of 33 Morgans; we used 40.7 Morgans for women and 22.9 Morgans for men[15]. For a genome of length L Morgans and when $A = 1$, we have the approximation[9] $P[\theta' > 0] \approx 1 - \exp(-(2G-1)L/2^{2G-1})$. Supplementary information S2 (box) illustrates, using a simple simulation, the relationship between the number of pedigree ancestors and ancestors that actually transmitted DNA to the current generation.

FIGURE 1 demonstrates the potential impact of using θ (expected IBD) rather than θ' (realized IBD) for h^2 estimation (see below). Using a Type B simulation, we generated phenotypes with pairwise correlations between the genetic contributions to each phenotype equal to θ' over all 50 generations ($G = 50$) of the simulation, and using this information when estimating h^2 gives the best possible inferences (red). Inferences are less precise if we only have available θ' based on $G = 10$ (green) or $G = 5$ (blue), and worse again when instead using θ based on $G = 5$, which corresponds to a complete 5-generation pedigree (purple). Precision deteriorates as close relatives are progressively excluded from the analysis, particularly when using θ. However, reasonable estimation remains possible using θ' even when only distantly related individuals are considered.

Relatedness with unobserved pedigree

When no pedigree information is available, allelic correlations at genotyped markers have been used to estimate the pedigree-based coancestry. Many models of population genetics[3,16] incorporate the following expression for the probability that two homologous alleles are both of type a (p_{aa}), where p_a is the probability that one sampled allele is of type a.

$$p_{aa} = \theta p_a + (1-\theta)p_a^2 \qquad (1)$$

This equation is based on the idea that, with probability θ, the two alleles have the same source and so effectively reflect only one observed allele, which is a with probability p_a. Otherwise, the two alleles are independent and are both a with probability p_a^2. Defining $U_1 = 1$ if the first allele is a, otherwise $U_1 = 0$, and similarly U_2 for the second allele, we have $E[U_1] = E[U_2] = p_a$, $\mathrm{Var}[U_1] = \mathrm{Var}[U_2] = p_a(1-p_a)$ and $E[U_1U_2] = p_{aa}$.

Table 1 | **Properties of genomic regions shared IBD by two individuals from G generations in the past**

Relationship	G	A	$\theta = E[\theta']$	95% CI of θ'	$P[\theta' > 0]$	E[#SR]	μ_G (SD)
Sibling	1	2	$0.25 = (1/2)^2$	(0.204, 0.296)	1.000	85.9	31.1 (35.2)
Half-sibling	1	1	$0.125 = (1/2)^3$	(0.092, 0.158)	1.000	42.9	31.1 (35.2)*
First cousin	2	2	$0.062 = (1/2)^4$	(0.038, 0.089)	1.000	37.5	17.8 (21.5)
Half-cousin	2	1	$0.031 = (1/2)^5$	(0.012, 0.055)	1.000	18.8	17.8 (21.5)*
Second cousin	3	2	$0.016 = (1/2)^6$	(0.004, 0.031)	1.000	13.3	12.5 (15.4)
Half-second cousin	3	1	$0.008 = (1/2)^7$	(0.001, 0.020)	0.995	6.7	12.5 (15.4)*
Third cousin	4	2	$0.004 = (1/2)^8$	(0.000, 0.012)	0.970	4.3	9.6 (12.0)
Half-third cousin	4	1	$0.002 = (1/2)^9$	(0.000, 0.008)	0.834	2.2	9.6 (12.0)*
	5	1	$(1/2)^{11}$	(0.000, 0.004)	0.431	0.7	7.9 (9.9)
	6	1	$(1/2)^{13}$	(0.000, 0.001)	0.160	0.2	6.6 (8.4)
	8	1	$(1/2)^{17}$	(0.000, 0.000)	0.015	0.0	5.1 (6.5)
	10	1	$(1/2)^{21}$	(0.000, 0.000)	0.001	0.0	4.1 (5.3)

CI, credible interval; SR, shared region. We consider only IBD (identity-by-descent) sharing that results from the direct lineage path of length G from each ancestor to each individual. A denotes the number of common ancestors: if $A = 2$, then these ancestors are mates, and the two individuals descend from distinct offspring of this union. θ' is the realized IBD genomic fraction from the indicated common ancestors, for which we show the expected (E) value (which is equal to the coancestry (θ)), the equal-tailed 95% CI and $P[\theta' > 0]$, the probability that the two individuals share any genomic region IBD from those ancestors. Also shown are the average number of SRs and, conditional on SR > 0, the expected region length in megabase pairs (μ_G) and its standard deviation (SD). Estimates are based on 10^5 Type A simulations (see Supplementary information S1 (box)). *The value shown is the same as the one above by definition.

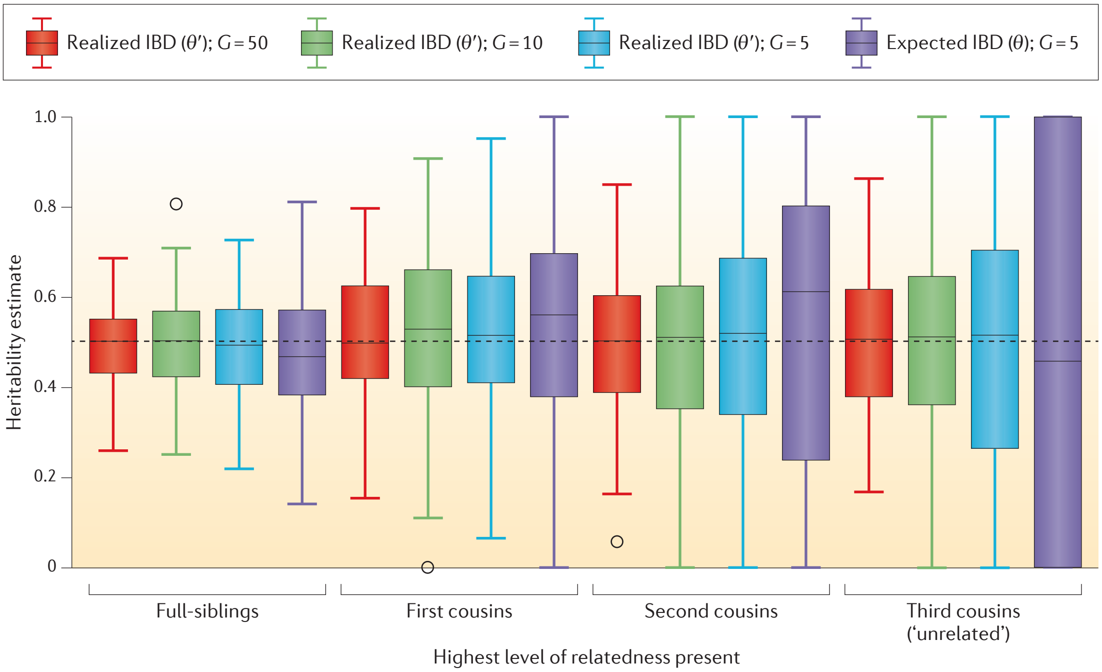


Figure 1 | **Estimation of narrow-sense heritability (h^2) using either expected IBD (θ) or realized IBD (θ') for varying levels of relatedness.** We used a Type B simulation (see Supplementary information S1 (box)). From the current generation, we drew 4 samples of 1,250 individuals, first with no filtering so that siblings were included, followed by filtering to exclude close relatives (the x axis labels indicate the closest relationship included). For each sample, we generated 100 phenotypes with narrow-sense heritability (h^2) of 0.5, where the correlation structure of the genetic contributions to each phenotype was specified by realized IBD (identity-by-descent; θ') based on $G = 50$ generations. To estimate h^2, it is necessary to specify a covariance matrix (K in equation 11). For each phenotype, we estimated h^2 using K constructed from θ' based on $G = 50$ (red boxes; the best-possible analysis in our model), θ' based on $G = 10$ (green), θ' based on $G = 5$ (blue) or θ based on $G = 5$ (purple; corresponding to a 5-generation pedigree). Boxes indicate the interquartile range for h^2 estimates, whiskers mark 1.5× this range, with values outside the 1.5× range individually plotted.

Therefore, equation 1 can be written in the form of a correlation coefficient as follows.

$$\theta = \frac{p_{aa} - p_a^2}{p_a(1-p_a)} = \frac{\mathrm{E}[(U_1 - p_a)(U_2 - p_a)]}{\sqrt{\mathrm{Var}[U_1]\mathrm{Var}[U_2]}} = \mathrm{Cor}[U_1, U_2] \qquad (2)$$

If the two alleles are sampled at random in a subpopulation, then θ equals the fixation index (F_{ST})[17]. However, if an allele is drawn from each of B and C — members of a finite pedigree with unrelated founders — then[18] θ in equation 1 is their coancestry ($\theta(B,C)$ in BOX 1), and the p_a are the founder allele probabilities in some reference population. Ignoring the expectation in equation 2 and using the observed values of U_1 and U_2, an unbiased estimator of θ is obtained if the p_a are known. This estimator is imprecise because it is based on only a single locus but, as θ is constant across loci, precision can be improved by averaging over loci (see equation 9 below).

The value of θ in equation 1 can be interpreted broadly, as representing any recent common origin of the two alleles; similarly, the concept of the reference population can be flexible. Interpreting θ in terms of IBD originating within the past G generations provides a coherent framework[9], but we have already outlined above some of the practical difficulties. Moreover, it is difficult to estimate the time depths of IBD genomic regions because their lengths are highly variable, with mean decreasing only linearly with time[14]. Alternatively, for two individuals sampled in a subpopulation, θ can be interpreted in terms of an intra-subpopulation pedigree of which founders are immigrants from a global population[16]. However, the assumptions that alleles drawn from the global population are independent and that the p_a are known or well estimated remain problematic[3], particularly in the presence of population structure[19].

There is a substantial amount of conservation genetics literature on estimators of θ based on equation 2, and these estimators are mainly designed for tens of multiallelic markers, such as short tandem repeat (STR) loci[20–22]. Maximum likelihood estimators have been proposed[23], but methods of moments estimators are often preferred despite their lower precision, because they are computationally efficient and can be unbiased if the p_a are known. In practice, not only are the p_a typically unknown, but the observed alleles from which they might be estimated are also not drawn from the reference population. Moreover, when θ is also estimated from the same data, the estimate $\hat{\theta}$ is biased downwards and is often negative, whereas $\theta \geq 0$. The limited number of markers available until recently meant that $\hat{\theta}$ was too

Maximum likelihood estimators
Estimates of unknown parameters obtained by maximizing the likelihood for the observed data given a statistical model.

Method of moments estimators
Estimates of unknown parameters obtained by equating theoretical moments (for example, mean, variance and skewness) under the assumed statistical model to empirical moments calculated from the observed data.

imprecise for the issues raised here to be of practical concern. However, high-density SNP data now permit precise estimation, forcing us to confront interpretation difficulties. Properties of estimators can be assessed in artificial, truncated pedigrees, but this may not be relevant in practical applications because real pedigrees are effectively infinite.

With 771 SNPs, a 2010 zebra finch study[24] found that direct estimation of θ was poor and recommended using SNPs to reconstruct the pedigree as a preliminary step. By contrast, the results from a 2013 study on pigs[25] indicated that 2,000 SNPs could give relatedness estimates that are superior to those from known pedigrees. However, although recognizing the potential for better measures of relatedness from markers rather than pedigrees, these authors still used pedigree-based measures as the 'gold standard' to assess marker panels. We suggest that use of equation 2 to estimate θ in natural populations should be avoided because of the problems of interpretation. Equation 2 may still give a useful summary of genome similarity, but it does not estimate a meaningful parameter except in artificial settings. Models and analyses can instead be formulated directly in terms of genome similarity, which raises the problem of how to compare measures of genome similarity (see below).

Coalescent theory

A different framework for describing relatedness in populations without pedigree information is provided by coalescent theory, in which alleles at a locus are connected through a coalescent tree[26]. In its simplest form, the standard coalescent describes the probability distribution of the TMRCA of a set of homologous alleles, assuming random mating in a constant-size population. In that case, the probability for two lineages to 'coalesce' at an ancestral allele more than G generations in the past is as follows, where N is the number of diploid individuals.

$$\mathrm{P}[\mathrm{TMRCA} > G] = \left(1 - \frac{1}{2N}\right)^{G} \approx e^{-G/2N} \qquad (3)$$

This model can be generalized to allow for variable population size and some forms of population structure and selection. The standard coalescent is based on assuming a Poisson number of offspring per individual and that each mating generates one offspring so full siblings are rare. However, it can be used to approximate the properties of some more-complex models, by replacing N in equation 3 with an effective population size (N_e). For example, our Type B simulation with $N = 10{,}000$ can be approximated by a coalescent model with $N_e = 8{,}450$, which is in close agreement with a theoretical formula[16] based on the variance of the number of offspring (see Supplementary information S3 (box)).

Under the coalescent model, the MRCA of two haploid human genomes at a given site is unlikely to be recent. In our Type B simulation model, the probability of an MRCA in generation G is ~6×10^{-5} for G up to several hundred, which supports the assumption that people are unrelated if nothing is known about their relatedness. However, even for $G = 78{,}000$ (which is the 99th percentile of the TMRCA distribution) and assuming a mutation rate of 1.2×10^{-8} per site per generation[27], the probability of a mutation in either lineage since the MRCA is still low (~0.002). Therefore, any two human genomes will be IBS (identical-by-state) for almost all genomic sites as a result of IBD. However, the situation is different for STR loci, which are often used in forensic identification and relatedness testing; STR mutation rates are ~10^{-3} per site per generation and so, under the coalescent model, the majority of IBS will not be a consequence of IBD.

Powell *et al.*[4] acknowledged a conflict between pedigree-based IBD theory and coalescent theory but, as they recognized, there is ultimately no conflict: IBD can be described in terms of the more general coalescent theory. Broadly speaking, the IBD versus non-IBD distinction is a simplification of the coalescent theory in which the TMRCA is classified into recent and non-recent. This can sometimes be a useful simplification, but it does not provide a satisfactory general notion of relatedness because none of the attempts to define 'recent' can solve the problem that a binary classification cannot capture the essentially continuous range of TMRCA values. A better approach is to define kinship coefficients in terms of genome-wide TMRCA distributions (BOX 2).

There is a fundamental connection between coalescent trees and pedigrees: a pedigree can be thought of as providing a 'scaffold' on which coalescent trees at different genomic loci are constructed. Pedigree members have maternal and paternal alleles at each locus, but each coalescent lineage passes through only one allele of each individual. Thus, we can consider the coalescent tree at a locus as a stochastic process on a fixed pedigree, making 'coin toss' decisions between the maternal and paternal chromosomes of each ancestor that is reached (see Supplementary information S4 (box)). Features of a more extensive pedigree, such as population structure, generate genome-wide influences on the coalescent distributions. This effect is evident in genome-wide association studies, in which population structure and cryptic relatedness (that is, pedigree effects) alter the genome-wide distribution of single-SNP association statistics[28]. Coalescent modelling usually ignores the pedigree because it is rarely observed and is difficult to infer. However, pedigree effects are not always negligible, and it may be useful in some settings to jointly model the pedigree and the coalescent trees embedded in it.

Recombination-sense IBD

In the absence of an explicit pedigree, IBD was initially defined in terms of mutation[29]: pairs of alleles at a locus are mutation-sense IBD if there has been no mutation since their MRCA. IBD is now more commonly defined as a property of a genomic region[30,31]: two haploid genomes are recombination-sense IBD if there has been no recombination within the region since their MRCA, ignoring mutations. There is no reference population in this approach, but the problem now is how to identify IBD regions, which are often

Coalescent tree
Each leaf of the tree corresponds to an observed allele, and the root represents the most recent common ancestor (MRCA) of all observed alleles. The internal nodes (branching points) represent the MRCA of the alleles at the leaves connected to that node (without passing the root). Distances along branches represent time, measured in generations.

IBS
(Identical-by-state; also identity-by-state). When two homologous alleles have matching type. Some definitions of IBS exclude IBD (identity-by-descent).

Box 2 | Relatedness in terms of times since common ancestors

The coancestry $\theta(B,C)$ is the expected value of a function equal to 1 when homologous alleles from individuals B and C have a most recent common ancestor (MRCA) that is 'recent' (within a specified pedigree, within G generations or within a subpopulation), and $\theta(B,C) = 0$ otherwise. A better approach is to use a genome-wide average of a more informative function of the time since the MRCA (TMRCA). Although the TMRCA is typically unknown, it can be estimated from dense markers or sequence data under a demographic model. The estimate is imprecise at any one locus, but genome-wide estimates can be highly informative and have been used to draw inferences about historical demographic parameters[92].

If bilineal relationships are of interest, then it is necessary to consider two genome-wide TMRCA distributions: the minimum TMRCA over the four pairs of alleles and the TMRCA of the remaining allele pair. Distributions are often summarized by their expectation, but for the TMRCA the probability assigned to recent times is of most importance, which suggests summarizing the TMRCA distribution by the expectation of a function that upweighs recent times, such as exp(–TMRCA/c) for some constant c.

Slatkin[93] proposed defining fixation indices in terms of ratios of coalescence times, and Rousset[3] developed this idea to propose definitions of co-ancestries based on the excess TMRCA probability in recent generations. The excess is relative to a random pair of individuals, assuming that the TMRCA probabilities for (B,C) and the random pair are proportional for large TMRCA. This definition reduces to equation 1 in simple settings but does not require any pedigree or founder population to be specified. No estimator from marker data was proposed, but this may provide a promising approach to develop new statistics that summarize relatedness without the requirement of pedigree information or a reference population.

short, and recombination events may not be detectable. Recombination-sense IBD does not lead naturally to a useful measure of genome-wide relatedness because all pairs of haploid genomes are entirely IBD; the question is where the breakpoints are between the IBD regions, and these can be hard to infer.

The largest consumer genetic ancestry companies have databases with $>10^5$ individuals, predominantly of European ancestry, and each genotyped at $>10^5$ SNPs. A focus for such companies is to identify pairs of individuals connected by short lineage paths. We discuss below the difficulties in using inferred IBD regions to achieve this, but here we question why discovering a poorly inferred, distant pedigree relationship based on sharing perhaps only one genomic region is preferred over seeking the highest level of genome-wide similarity. Part of the answer may be that measures of genome similarity have not been adequately developed.

Another motive for inferring IBD regions is to understand population structure and demographic history[9,31,32]. For example, detecting long regions shared IBD between individuals in different parts of the world can point to a recent migration event. The observed data may first be summarized in terms of IBD regions, and demographic inferences can then be based on these inferred regions. However, this two-step process will disregard any demographic information that is not captured by the IBD inference, suggesting that more direct and statistically efficient demographic inferences may be possible.

The statistics of IBD regions

IBD regions can be measured in genetic distances (Morgans) or physical distances (base pairs). Genetic maps specify the relationship between them, which varies across the genome; they also differ across human populations, substantially at a fine scale[33,34].

We report below region lengths in megabase pairs. The simulation-based estimates reported in TABLE 1 for E[#SR] — the expected number of IBD regions for a pair of individuals — and for μ_G — the expected length of these regions — agree closely with theoretical (sex-averaged) values for 22 autosomes spanning 2,667 Mb with map lengths of 40.7 Morgans for females and 22.9 Morgans for males (see Supplementary information S5 (box)).

$$\mathrm{E}[\#\mathrm{SR}] = A \times \frac{22 + (40.7 + 22.9) \times G}{2^{2G-1}} \tag{4}$$

and

$$\mu_G = \frac{2667}{22 + (40.7 + 22.9) \times G} \tag{5}$$

This implies that the mean length of IBD regions is just more than 2 Mb when $G = 20$ and just more than 1 Mb when $G = 40$. Assuming again a mutation rate of 1.2×10^{-8} per site per generation, the probability that an average-length genomic region shared (recombination-sense) IBD from an ancestor G generations back is also mutation-sense IBD is approximately constant over G at ~0.37 (see Supplementary information S6 (box)). Therefore, average-length pairs of IBD regions are unlikely to be identical at the sequence level, even if they are from a recent common ancestor.

FIGURE 2A shows the distribution of IBD region lengths for $G = 1$ and $G = 10$ based on the Type A simulations underlying TABLE 1. For $G = 1$ there is a peak in lengths close to 30 Mb because about one-third of sibling pairs have complete IBD for at least one of chromosomes 21 and 22. A gamma distribution generally gives a good fit, except when G is very small owing to the difference between male and female recombination rates. We estimate the gamma shape parameter to be approximately constant over G at ~0.76, implying that the standard deviation (SD) is ~$\mu_G/\sqrt{0.76}$. This is a higher SD than that for the exponential distribution (gamma distribution with shape parameter 1), which would apply to IBD region lengths if the recombination rate were uniform across the genome and sexes. Even when an IBD region arises from a shared parent, there is a substantial probability for its length to be short, whereas the shared region could still be large even for an ancestor >20 generations in the past. FIGURE 2B shows the inverse distribution based on the Type B simulations used for FIG. 1, and indicates how well the time depth of the common ancestor can be inferred from the region length. Very long shared regions (>80 Mb) are highly likely to descend from a recent ancestor ($G \leq 5$), but G has a wide range for regions <40 Mb. Up to 10 Mb in length, the majority of shared regions descend from an ancestor >20 generations back. Our estimates are based on a very simple simulation model but are broadly consistent with an estimated age range of 32–52 generations for a 10-centimorgan region shared by a pair of UK residents[35]. There are some excellent blog discussions of issues around the statistics of genome sharing among relatives at gcbias, Genetic Inference and On Genetics.

SNP-based measures of relatedness

We distinguish genome-wide averages of single-SNP statistics from haplotype-based methods (which are sometimes referred to, respectively, as IBS and IBD methods, or as methods that do not and do take account of linkage[32]). Matrices of SNP-based kinship coefficients have been called genetic relatedness matrices in recent literature[36,37] but, to avoid any implication that these matrices estimate pedigree-based relatedness, we prefer to call them genetic similarity matrices (GSMs).

Single-SNP averages. We use S_{Bj} to denote the genotype of B at the j^{th} diallelic SNP, coded as 0, 1 or 2. Analogous with the definition of coancestry, a natural way to score the similarity of two individuals at each SNP is as the probability of a match between alleles drawn at random from each of them. In that case matching homozygotes (0,0) or (2,2) score 1; discordant homozygotes (0,2) score 0; while (0,1), (1,1) and (1,2) all score 0.5. Averaged over m SNPs, this gives an allele-sharing coefficient[29,38] as follows, where X_B is a (row) vector with j^{th} entry $S_{Bj} - 1$.

$$K_{as}(B,C) = \frac{1}{2} + \frac{1}{2m}\sum_{j=1}^{m}(S_{Bj}-1)(S_{Cj}-1)$$
$$= \frac{1}{2} + \frac{1}{2m}X_B X_C^T \qquad (6)$$

The corresponding GSM is then $(1 + XX^T/m)/2$, where X is a genotype matrix with row B equal to X_B. The range of K_{as} depends on the MAF spectrum of the SNPs, and $K_{as}(B,B) = (1 + h_B)/2$, where h_B denotes the homozygosity of B. Recall that $\theta(B,B) = (1 + f_B)/2$, where f_B denotes the inbreeding coefficient of B. K_{as} can be interpreted as average mutation-sense IBD under the assumption that IBS implies IBD, which is reasonable for SNPs given their low mutation rate.

The case (1,1) of two heterozygotes can correspond to either IBD = 2 or IBD = 0 depending on phase, which is often unknown. Many authors[39] prefer to score matching heterozygotes as 1 rather than 0.5, resulting in the following allele-sharing similarity, which has $K'_{as}(B,B) = 1$.

$$K'_{as}(B,C) = 1 - \frac{1}{2m}\sum_{j=1}^{m}|S_{Bj} - S_{Cj}| \qquad (7)$$

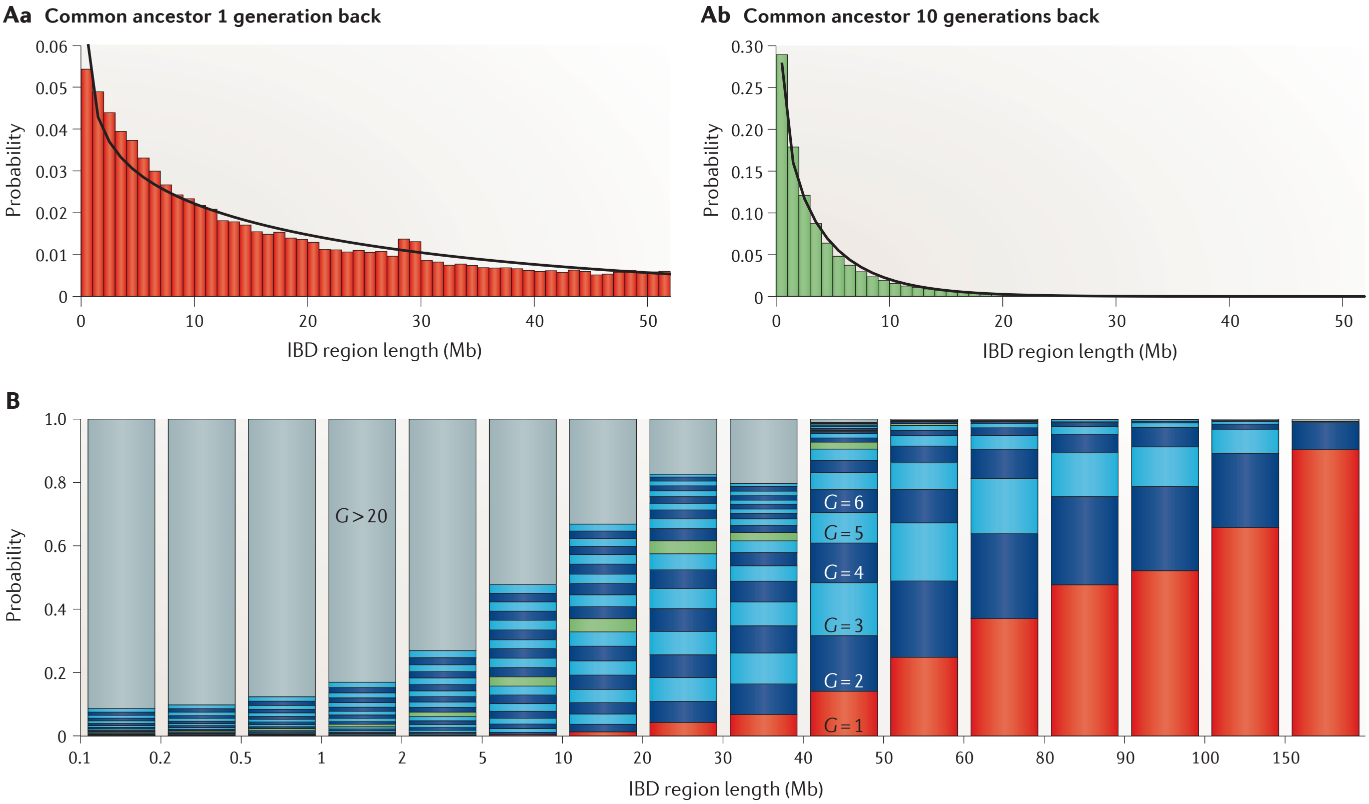


Figure 2 | **Statistics of IBD genomic regions.** Distributions of lengths of genomic regions shared IBD (identity-by-descent) from a common ancestor of 1 generation (part **Aa**) and 10 generations (part **Ab**) in the past are shown. The solid black curves show an approximating gamma distribution, with mean given by equation 4 and a shape parameter of 0.76. For all IBD regions arising from a common ancestor within the last 50 generations, the bars show how the distribution of the generation of the common ancestor depends on the length of the region (part **B**). From bottom to top in the graph, the tranches correspond to $G = 1$ (red), $G = 2\ldots9$ (alternating dark and light blue), $G = 10$ (green), $G = 11\ldots20$ (alternating dark and light blue) and $G > 20$ (grey). Plots in part **A** are based on Type A simulations and part **B** on a Type B simulation, details of which are provided in Supplementary information S1 (box).

K_{as} and K'_{as} give all SNPs equal weight, irrespective of MAFs, which has the advantage of avoiding estimation of population MAF values. However, the lower the MAF the greater the evidence for a recent common ancestor, which suggests giving more weight to rare shared alleles. One step in this direction is to centre the genotype scores. If p_j is the population MAF at the j^{th} SNP, and now defining X_B as having entries $S_{Bj}-2p_j$, then we can define the centred coefficient as follows.

$$\begin{aligned} K_{c0}(B,C) &= \frac{1}{m}\sum_{j=1}^{m}(S_{Bj}-2p_j)(S_{Cj}-2p_j)^T \\ &= \frac{1}{m}X_B X_C^T \end{aligned} \qquad (8)$$

The GSM is XX^T/m. Some authors[19,40,41] use $2\Sigma_{j=1}^{m}p_j(1-p_j)$ in place of m in the denominator of equation 7, so that $K_{c0}(B,B)\approx 1$ if B is outbred. Unlike K_{as} and K'_{as}, the value of K_{c0} can be negative, and this must frequently occur if the p_j are estimated from the same data, as then K_{c0} measures the excess or deficit of allele sharing from that expected under a random assignment of alleles. Some authors replace negative values with zero[42], which may be motivated by the belief that a non-negative θ is being estimated. However, the differences in genome similarity among unrelated individuals can be informative (FIG. 1), and setting negative values to zero discards much of this information.

In addition to centring, it is common to standardize, which assumes each SNP to be equally informative, leading to the following definition[43,44].

$$K_{c-1}(B,C) = \frac{1}{m}X_B X_C^T \quad \text{where} \quad X_{Bj} = \frac{S_{Bj}-2p_j}{\sqrt{2p_j(1-p_j)}} \qquad (9)$$

$K_{c-1}(B,C)$ is an average over SNPs of an estimator of an allelic correlation coefficient (see equation 2). It is the basis of principal component methods to adjust for population structure in association studies[45]. A regression adjustment to K_{c-1} has been proposed, imposing shrinkage towards zero in accordance with sample size[12], which was shown in simulations[4] to offset errors when estimating θ or h^2. With high SNP density, genetic variation tagged by multiple SNPs in high linkage disequilibrium (LD) will be over-represented when calculating genomic similarities, which can have adverse implications for heritability analyses. Adjusting for this by reweighting SNPs in the calculation of K_{c-1} gives improved h^2 estimates[46].

K_{c-1} is an unbiased and efficient estimator of θ in equation 2, provided that the modelling assumptions hold but, as discussed below equation 2, those assumptions rarely hold in practice (similar comments apply to K_{c0}). Although K_{as}, K'_{as}, K_{c0} and K_{c-1} all provide widely used SNP-based measures of genome similarity, there are no good grounds to regard any of them as a canonical SNP-based measure of relatedness. This frees us to search for measures that perform well in specific applications. In practice, genome-wide genome similarity is useful because it gives a guide to allele sharing at causal loci for the trait of interest[4], and so the best genome similarity coefficient will reflect the genomic architecture of the trait. This contrasts fundamentally with kinship coefficients, which are trait-independent, but the new flexibility can be highly profitable in terms of understanding genetic mechanisms.

One place to start the search for better GSMs is the one-parameter family $K_{c\alpha}$, defined as in equation 9 except that X_{Bj} is now defined as follows.

$$X_{Bj} = (S_{Bj}-2p_j)\times\left(\sqrt{2p_j(1-p_j)}\right)^{\alpha} \qquad (10)$$

The special cases $\alpha=-1$ (equation 9) and $\alpha=0$ (equation 8) are widely used in practice. We compare below the performance of GSMs based on $K_{c\alpha}$ for $\alpha=-2$, -1, 0 and 1. These values are chosen for illustration, and a thorough study should consider additional values for α.

Methods based on detecting shared haplotypes. Averages of single-SNP coefficients do not take into account the lengths of genomic regions shared between two individuals. On average, the longer the shared region (or regions), the more recent the ancestor (or ancestors), which we have noted is relevant to some population genetics applications[32,35]. With unphased genotypes, a simple approach is to seek genomic regions for which two individuals share at least one allele at every SNP[15]. More sophisticated approaches usually require genotype data to be phased. Methods for phasing[47–51] typically use a hidden Markov model with the aim of constructing haplotypes (the hidden states) that are consistent with the observed genotypes, allowing for mutation and recombination[52]. Given the haploid data, it is then straightforward to identify IBS regions. However, a long region shared IBS by two individuals may have resulted from two or more lineage paths and is therefore not (recombination-sense) IBD. Conversely, an IBD region may consist of several IBS regions that are interrupted by occasional data errors or mutations since the MRCA. Inferring such regions requires a model for recombinations and mutations, which in turn implies a model for the demographic history of the population[53]. Methods for identifying IBD regions[15,54–59] differ in the size of data sets they can handle, which depends on the type of deterministic or stochastic search algorithm used. These algorithms can be sensitive to parameter choices, yet there is often no obvious way to tune these based on the data[32]. For the largest data sets, simultaneous phasing of all individuals is computationally infeasible. A recently developed method avoids explicit phasing but 'penalizes' proposed IBD intervals according to the estimated number of implied phasing switches[60].

Current approaches typically neglect shorter IBD regions because they are harder to infer. This means that distantly related pairs of individuals will have little or no IBD detected, and so differences in genome-sharing among them will be poorly recorded. FIGURE 2 shows that inferring the TMRCA based on region length is difficult but, for many applications, exploitation of the information in short IBD regions would be advantageous.

Linkage disequilibrium

(LD). A population correlation of allele pairs drawn at different genomic loci in the same gamete (that is, in a haploid genome).

Chromopainter[61] offers a different approach that is based directly on the haplotype copying model[52]. Every chromosome is regarded as a mosaic of fragments copied from other sampled chromosomes, possibly with some mutations, and the coancestry of two individuals is measured by the number of distinct copying events between them. Although copying is intended to reflect IBD, every part of every chromosome is copied from another chromosome, and so an individual that is not closely related with anyone else in the sample will have the closest genome matches recorded even if these are remote.

Heritability and phenotype prediction

We focus below on the use of GSMs to estimate h^2 and to predict phenotypes. Both applications traditionally used a matrix of θ values, which is now usually replaced by a GSM, to model phenotypic correlations among individuals, with the intuition being that the higher the genome similarity of two individuals, the more correlated their phenotypes that are under genetic control.

The linear mixed model. Underlying both types of inference is the following regression model, in which a matrix K specifies the covariance structure of a vector of observed phenotypes Y, where N represents the normal (or Gaussian) distribution.

$$Y \sim \mathrm{N}(Z\beta_0, K\sigma_g^2 + I\sigma_e^2) \qquad (11)$$

In this equation, β_0 represents fixed effects corresponding to covariates in Z, I denotes the identity matrix, and σ_g^2 and σ_e^2 are the genetic and environmental variances, respectively. Given Y, Z and K, we typically estimate σ_g^2 and σ_e^2 using restricted maximum likelihood (REML)[62], which seeks values that maximize the restricted model likelihood (see Supplementary information S7 (box)). For h^2 estimation, we are interested in the ratio of variance terms; when K is standardized to have a mean of zero and a mean diagonal value of one, $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. A key technique for phenotype prediction in plant and animal breeding is best linear unbiased prediction (BLUP)[63,64], which predicts the phenotypes of new individuals from estimates of $K\sigma_g^2$.

SNP-based analyses that were pioneered in wild populations[65] have been extensively applied in animal and plant breeding[66–69] and, more recently, in humans[12,70,71]. A feature of SNP-based $\hat{h}^2$ in humans is the use of unrelated individuals. This is counterintuitive because more relatedness generates more precise inferences. However, the problem is that inferences vary according to the levels of relatedness among the sampled individuals. In addition, most readily available data are from population samples that include little relatedness. By excluding any close relatives, sampled individuals only share the short genomic regions from remote ancestors that generate LD, which is reasonably stable across population samples. Furthermore, although high levels of relatedness would generate long-range tagging of causal variants, which can therefore all contribute to $\hat{h}^2$, the ability to attribute $\hat{h}^2$ to specific genomic regions would be greatly reduced. Despite the information loss from reduced relatedness, reasonable precision (SD < 0.05) can be achieved with, for example, 5,000 unrelated individuals[72], and the estimates are more consistent across studies. SNP-based $\hat{h}^2$ values using unrelated individuals have been interpreted in terms of common causal variants because these are better tagged by SNPs than rare variants. However, poorly tagged rare variants will contribute to $\hat{h}^2$ (REF. 73), which hinders interpretation. Even so, it is possible to make useful comparisons across genomic regions[74] and across phenotypes.

Prediction accuracy can be measured by the correlation (r) between observed and predicted phenotypes across test individuals. The squared correlation (r^2) is bounded above by h^2, but in practice $r^2 << h^2$. Relatedness between training and test individuals enhances predictive accuracy in the test set, but this may give an over-optimistic assessment of performance if the model is applied to new, less-related individuals. In humans, prediction of complex traits is typically poor in the general population but can be useful in high-risk groups[75–78].

Table 2 | **Model log likelihood, heritability estimates ($\hat{h}^2$) and predictive accuracy (r^2) for different SNP-based GSMs**

Trait	Log likelihood				Heritability ($\hat{h}^2$)				Prediction accuracy (r^2)			
	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}	K_{c-2}	K_{c-1}	K_{c0}	K_{c1}
Bipolar disorder	−97	0*	−12	−32	1.00*	0.98	0.92	0.81	0.040	0.074*	0.073	0.069
Coronary artery disease	−24	−3	0*	−1	0.33	0.41*	0.17	0.06	0.000	0.017	0.020*	0.02
Crohn's disease	−178	−5	0*	−3	1.00	1.00	1.00	1.00	0.057	0.096	0.098*	0.095
Hypertension	−32	−3	0*	−1	0.57*	0.48	0.21	0.08	0.005	0.024	0.026*	0.026
Rheumatoid arthritis	−125	0*	−15	−72	0.77	1.00*	0.99	0.17	0.016	0.043	0.042	0.043*
Type 1 diabetes	−65	0*	−7	−16	0.85*	0.82	0.41	0.16	0.031	0.060	0.060*	0.056
Type 2 diabetes	−28	0*	0	−3	0.64*	0.52	0.22	0.08	0.009	0.026*	0.025	0.024
Average	−78	−2*	−5	−18	0.74	0.74*	0.56	0.34	0.022	0.048	0.049*	0.047

GSM, genetic similarity matrix; SNP, single-nucleotide polymorphism. Data for seven disease traits are from the Wellcome Trust Case Control Consortium. The GSMs considered are $K_{c\alpha}$ for $\alpha \in \{-2, -1, 0, 1\}$. Log likelihoods, computed under the mixed model (equation 11), are reported relative to the maximum observed across GSMs. $\hat{h}^2$ values correspond to the observed scale. *The GSMs marked by asterisks indicate those that maximize the model likelihood, $\hat{h}^2$ and r^2.

***How to choose* K?** The success of both h^2 estimation and phenotype prediction using BLUP depends on the choice of K in equation 11. Pedigree-based θ values vary with the choice of pedigree, although in practice there is often little choice. The range of options for SNP-based K is much greater. In human genetics, K_{c-1} is often chosen, which implies an assumption that all SNPs explain the same fraction of phenotypic variance, and so effect sizes tend to increase as MAF decreases. By contrast, K_{c0} is often preferred in animal and plant breeding, and implies the same effect size distribution for all SNPs. These are both special cases of $K_{c\alpha}$, and some other value of α may provide better results, depending on the true relationship between MAFs and effect sizes.

With many GSMs available, we can now choose K based on performance in applications. In simulations (see Supplementary information S8 (box)) we found near-perfect h^2 estimation and phenotype prediction when the true K (that was used for simulating the phenotypes) was used in the analysis. If instead we use $K_{c\alpha}$ for various α, we find that maximizing $\hat{h}^2$ does not reliably recover the true value of α, but that maximizing the model likelihood and the predictive accuracy both give a good guide to α, suggesting that these provide useful criteria for choosing K. These results also confirm that a GSM chosen to suit the architecture of the trait under study will perform better than a phenotype-independent choice.

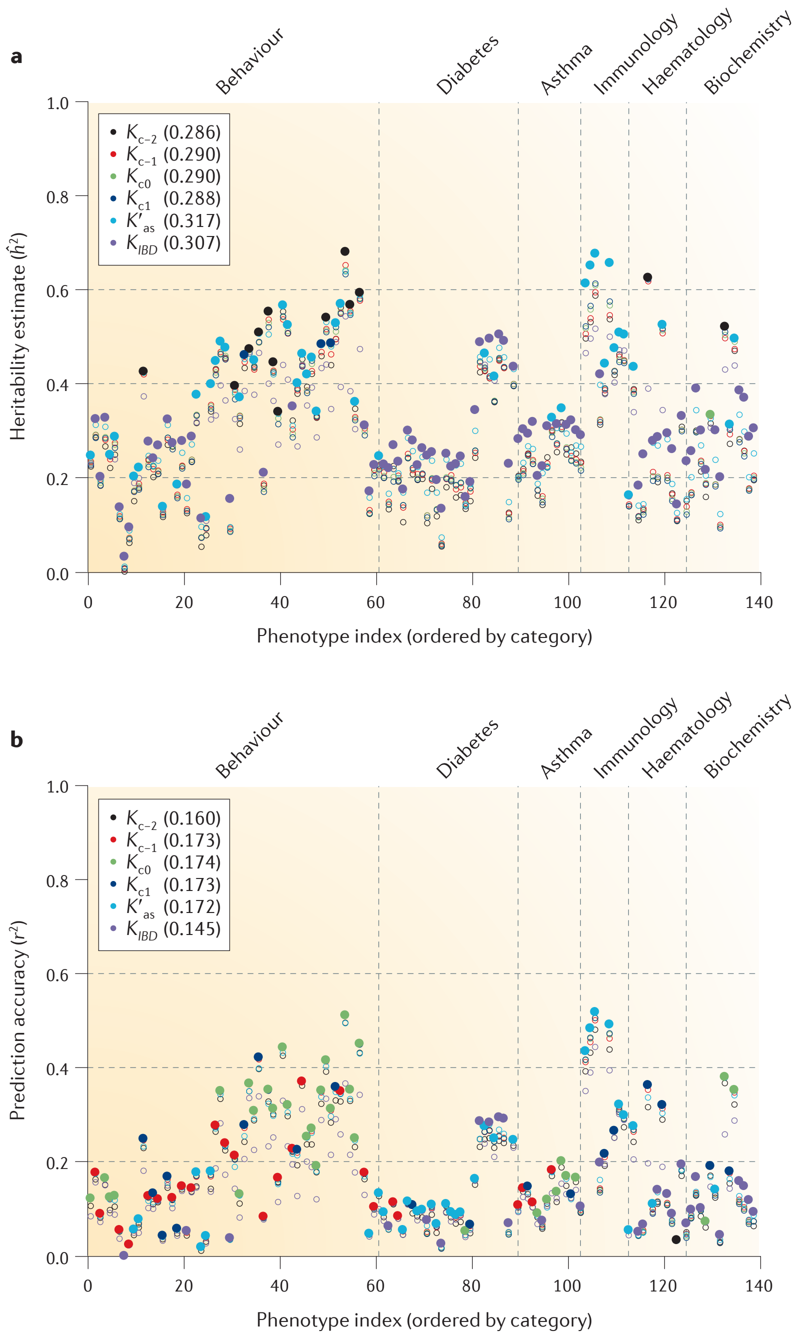


Figure 3 | **Heritability estimates ($\hat{h}^2$) and predictive accuracy (r^2) for 139 mouse phenotypes using different SNP-based GSMs.** The genetic similarity matrices (GSMs) considered are $K_{c\alpha}$ for $\alpha \in \{-2, -1, 0, 1\}$, K'_{as} and K_{IBD} (which is a matrix recording inferred fractions of IBD (identity-by-descent)). The vertical dashed lines separate different categories of phenotype: behaviour, diabetes, asthma, immunology, haematology and biochemistry. Solid points indicate the GSM providing highest heritability estimates ($\hat{h}^2$; part **a**) or predictive accuracy (r^2; part **b**) for each phenotype. Average $\hat{h}^2$ and r^2 values for each GSM are provided in parentheses. Data are from the Wellcome Trust Heterogeneous Stock Mouse Collection. SNP, single-nucleotide polymorphism.

TABLE 2 shows results from $\hat{h}^2$ estimation and phenotype prediction for seven human disease traits from the Wellcome Trust Case Control Consortium[79], for which we use $K_{c\alpha}$ in equation 11 with $\alpha = -2, -1, 0$ and 1. When dealing with binary outcomes, it is preferable to consider $\hat{h}^2$ on the liability scale[80], but for the purpose of making comparisons the observed scale is adequate[37]. On average, the model likelihoods are maximized when $\alpha = -1$, whereas $\alpha = 0$ is slightly superior for prediction. The latter result may reflect that, owing to the difficulty in estimating effect sizes for rare variants, prediction can perform well with α above its true value, as this increases the weight given to common variants.

FIGURE 3 shows $\hat{h}^2$ and r^2 for highly related mice using data from the Wellcome Trust Heterogeneous Stock Mouse Collection[81]. We consider 139 phenotypes spanning behavioural, haematological, biochemical and disease-related phenotypes. In addition to $K_{c\alpha}$ for $\alpha = -2, -1, 0$ and 1, we also consider K'_{as} and K_{IBD} (a matrix recording pairwise IBD fractions inferred using FastIBD[57]). The presence of high relatedness enables accurate estimation and prediction, and the different GSMs perform similarly overall. K_{IBD} is the worst overall for prediction, which suggests that its high $\hat{h}^2$ value may be inflated. However, K_{IBD} gives the best prediction for two of the seven phenotype categories, reflecting that the best GSM depends on the genetic architecture of the trait. Specifically, K_{IBD} performs well when the causal variants are not well tagged by the SNPs, which in turn suggests a lesser role for common causal variants.

Random regression models. For any K of the form XX^T, which includes $K_{c\alpha}$ but neither K_{as} nor K'_{as}, the mixed regression model (equation 11) can be reformulated as a random regression model in which the phenotype of B is expressed as follows[66], where $\beta_j \sim N(0, c\sigma_g^2)$ and c is a constant.

$$Y_B = \beta_0 + \sum_{j=1}^{m} \beta_j X_{Bj} \qquad (12)$$

Simulations indicate[46] that this model is effective for estimating h^2. However, for prediction, the assumption that all effect sizes follow a Gaussian distribution with constant variance is a severe limitation, so recently there have been many attempts to find more suitable models[82–84].

When equation 12 is available, it provides a more interpretable model in which K has no role: h^2 is just variance explained in a regression model with SNP predictors. Historically, kinship has played a central part in h^2 analysis, so it seems unthinkable to estimate h^2 without this concept or a SNP-based proxy, but we believe that benefits accrue from the freedom to develop statistical structures that are tailored to particular traits, which differ fundamentally from traditional kinship coefficients.

Conclusion

Quantitative genetics has undergone exciting developments in recent years, and this is affecting even the most fundamental aspects of the discipline, including our understanding of relatedness. The relatively simple IBD theory based on pedigrees remains useful in some areas, but particularly in natural populations we require more flexible and interpretable concepts to fully take advantage of the high numerical precision afforded by genome-wide SNP and sequence data. On the one hand, it seems clear that a satisfactory general definition of relatedness must be based on concepts from coalescent theory, and particularly on genome-wide TMRCAs. On the other hand, the choice of numerical measures of relatedness can be driven by optimizing criteria that are relevant to applications, such as model likelihood and predictive accuracy. There is much progress to be made in both directions.

1. Grafen, A. A geometric view of relatedness. *Oxford Surv. Evol. Biol.* **2**, 28–90 (1985).
2. Maynard Smith, J. *Evolutionary Genetics* (Oxford Univ. Press, 1998).
3. Rousset, F. Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**, 371–380 (2002).
This paper gives a critical examination of kinship coefficients and proposes a new approach to measure kinship based on a cumulative excess of recent coalescences.
4. Powell, J., Visscher, P. & Goddard, M. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Rev. Genet.* **11**, 800–805 (2010).
This is a review on IBS and IBD concepts, with a focus on choice of reference population; it also discusses SNP-based computation of relatedness coefficients and their use in heritability estimation.
5. Weir, B., Anderson, A. & Hepler, A. Genetic relatedness analysis: modern data and new challenges. *Nature Rev. Genet.* **7**, 771–780 (2006).
6. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
7. Corr, P. & Kippen, R. The case for parity and birth-order statistics. *Aust. N. Z. J. Stat.* **48**, 171–200 (2006).
8. Calboli, F., Sampson, J., Fretwell, N. & Balding, D. Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics* **179**, 593–601 (2008).
9. Thompson, E. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326 (2013).
This is an extensive review on the IBD concept that covers many applications and citations to early literature. We disagree with the conceptual framework, but there is much that is valuable in this review.
10. Visscher, P. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
This paper introduces a clever innovation for heritability estimation and is the first to exploit differences in realized IBD among pairs of individuals with the same pedigree-based relatedness.
11. Hill, W. G. On estimation of genetic variance within families using genome-wide identity-by-descent sharing. *Genet. Sel. Evol.* **45**, 32 (2013).
12. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
13. Falconer, D. & Mackay, T. *Introduction to Quantitative Genetics* 4th edn (Longman, 1996).
14. Donnelly, K. The probability that related individuals share some section of genome identical by descent. *Theor. Popul. Biol.* **23**, 34–63 (1983).
15. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
16. Crow, J. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper and Row, 1970).
17. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 159–171 (1951).
18. Wright, S. Coefficients of inbreeding and relationship. *Amer. Nat.* **61**, 330–338 (1922).
19. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
20. Csillery, K. *et al.* Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics* **173**, 2091–2101 (2006).
21. Oliehoek, P., Windig, J., van Arendonk, J. & Bijma, P. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**, 483–496 (2006).
22. Beaumont, M. in *Handbook of Statistical Genetics* (eds Balding, D., Bishop, M. & Cannings, C.) Ch. 30 (Wiley, 2007).
23. Thompson, E. The estimation of pairwise relationships. *Ann. Hum. Genet.* **39**, 173–188 (1975).
24. Santure, A. *et al.* On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Mol. Ecol.* **19**, 1439–1451 (2010).
25. Lopes, M. *et al.* Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet.* **14**, 92 (2013).
26. Nordborg, M. in *Handbook of Statistical Genetics* (eds Balding, D., Bishop, M. & Cannings, C.) Ch. 25 (Wiley, 2007).
27. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
28. Astle, W. & Balding, D. Population structure and cryptic relatedness in genetic association studies. *Statist. Sci.* **24**, 451–471 (2009).
29. Malécot, G. *The Mathematics of Heredity* (Freeman, 1969).
30. Sved, J. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Popul. Biol.* **2**, 125–141 (1971).
31. Hayes, B., Visscher, P., McPartlan, H. & Goddard, M. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res.* **13**, 635–643 (2003).
32. Lawson, D. & Falush, D. Population identification using genetic data. *Annu. Rev. Genet.* **13**, 337–361 (2012).
This is a review on available GSMs, both that do and do not take account of linkage, from the perspective of classifying individuals into populations.
33. Graffelman, J., Balding, D., Gonzalez-Neira, A. & Bertranpetit, J. Variation in estimated recombination rates across human populations. *Hum. Genet.* **122**, 301–310 (2007).
34. Wegmann, D. *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genet.* **43**, 847–853 (2011).
35. Ralph, P. & Coop, G. The geography of recent genetic ancestry across europe. *PLoS Biol.* **11**, e1001555 (2013).
This paper investigates IBD genome sharing across Europe and how this reflects population size and migrations over recent millennia.
36. Forni, S., Aguilar, I. & Misztal, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* **43**, 1–7 (2011).
37. Lee, J. Y. S., Goddard, M. & Visscher, P. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
38. Toro, M. *et al.* Estimation of coancestry in Iberian pigs using molecular markers. *Conserv. Genet.* **3**, 309–320 (2002).
39. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
40. Habier, D., Fernando, R. & Dekkers, J. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397 (2007).
41. VanRaden, P. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
42. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38**, 203–208 (2006).
43. Loiselle, B., Sork, V., Nason, J. & Graham, C. Spatial genetic structure of a tropical understory shrub *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* **82**, 1420–1425 (1995).
44. Amin, N., van Duijn, C. & Aulchenko, Y. A genomic background based method for association analysis in related individuals. *PLoS ONE* **2**, e1274 (2007).
45. Price, A. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
46. Speed, D., Hemani, G., Johnson, M. & Balding, D. Improved heritability estimation from genome-wide SNP data. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
47. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
48. Li, Y., Willer, C., Ding, J., Scheet, P. & Abecasis, G. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
49. Browning, B. & Browning, S. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
50. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457–470 (2011).
51. Delaneau, O., Zagury, J. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2013).

52. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using singlenucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
53. Thompson, E. The IBD process along four chromosomes. *Theor. Popul. Biol.* **73**, 369–373 (2008).
54. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
55. Bercovici, S., Meek, C., Wexler, Y. & Geiger, D. Estimating genome-wide IBD sharing from SNP data via an efficient hidden Markov model of LD with application to gene mapping. *Bioinformatics* **26**, i175–i182 (2010).
56. Moltke, I., Albrechtsen, A., Hansen, T., Nielsen, F. & Nielsen, R. A method for detecting IBD regions simultaneously in multiple individuals — with applications to disease genetics. *Genome Res.* **121**, 1168–1180 (2011).
57. Browning, B. & Browning, S. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* **88**, 173–182 (2011).
58. Browning, B. & Browning, S. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
59. Li, H. *et al.* Relationship estimation from whole-genome sequence data. *PLoS Genet.* **10**, e1004144 (2014).
60. Durand, E., Eriksson, N. & McLean, C. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. *Mol. Biol. Evol.* **31**, 2212–2222 (2014).
61. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4**, e1000078 (2008).
62. Corbeil, R. & Searle, S. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).
63. Henderson, C. Estimation of genetic parameters. *Ann. Math. Stat.* **21**, 309–310 (1950).
64. Henderson, C., Kempthorne, O., Searle, S. & von Krosigk, C. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218 (1959).
65. Mousseau, T., Ritland, K. & Heath, D. A novel method for estimating heritability using molecular markers. *Heredity* **80**, 218–224 (1998).
66. Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M. Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**, 433–443 (2009).
67. Goddard, M. & Hayes, B. Genomic selection. *J. Anim. Breed. Genet.* **124**, 323–330 (2007).
68. Goddard, M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **139**, 245–257 (2009).
69. Scutari, M., Mackay, I. & Balding, D. Improving the efficiency of genomic selection. *Stat. Appl. Genet. Mol.* **12**, 517–527 (2013).
70. Makowsky, R. *et al.* Beyond missing heritability: prediction of complex traits. *PLoS Genet.* **7**, e1002051 (2011).
71. de los Campos, G., Hickey, J., Pong-Wong, R. & Daetwyler, H. Whole genome regression and prediction methods applied to plan and animal breeding. *Genetics* **193**, 327–345 (2013).
72. Visscher, P. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
73. Dickson, S., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
74. Yang, J. *et al.* Genomic partitioning of genetic variation for complex traits using common SNPs. *Nature Genet.* **43**, 519–525 (2011).
75. de los Campos, G., Gianola, D. & Allison, D. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Rev. Genet.* **11**, 880–886 (2010).
76. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e03348 (2013).
77. Wray, N. *et al.* Pitfalls of predicting complex traits from SNPs. *Nature Rev. Genet.* **14**, 507–515 (2013).
78. Speed, D. *et al.* Describing the genetic architecture of epilepsy through heritability analysis. *Brain* **137**, 2680–2689 (2014).
79. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14, 000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
80. Dempster, E. & Lerner, I. Heritability of threshold characters. *Genetics* **35**, 212–236 (1950).
81. Valdar, W. *et al.* Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genet.* **38**, 879–887 (2006).
82. Zhou, X., Carbonetto, P. & Stephens, M. Polygeneic modeling with Bayesian sparse linear mixed models. *PLoS. Genet.* **9**, e1003264 (2013).
83. Speed, D. & Balding, D. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).
84. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep.* **3**, 1815 (2013).
85. Cockerham, C. Higher order probability functions of identity of alleles by descent. *Genetics* **69**, 235–246 (1971).
86. Cannings, C. & Thomas, A. in *Handbook of Statistical Genetics* (eds Balding, D., Bishop, M. & Cannings, C.) Ch. 23 (Wiley, 2007).
87. Jacquard, A. *The Genetic Structure of Populations* (Springer, 1974).
88. Cotterman, C. A. in *Genetics and Social Structure* (ed. Ballonoff, P. A.) 157–272 (Dowden, Hutchinson & Ross, 1974).
89. Guo, S. Variation in genetic identity among relatives. *Hum. Hered.* **46**, 61–70 (1996).
90. Hill, W. G. & Weir, B. S. Variation in actual relationship among descendants of inbred individuals. *Genet. Res.* **94**, 267–274 (2012).
91. Fisher, R. *The Theory of Inbreeding* (Oliver and Boyd, 1949).
92. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
93. Slatkin, M. Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175 (1991).

Acknowledgements

The authors thank G. Hellenthal and D. Kennett (both University College London), and M. Beaumont (University of Bristol) for discussion. This work is funded by the UK Medical Research Council under grant G0901388, with support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Access to Wellcome Trust Case Control Consortium data was authorized as work related to the project "Genome-wide association study of susceptibility and clinical phenotypes in epilepsy".

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

gcbias: http://gcbias.org/
Genetic inference: http://www.genetic-inference.co.uk/blog/2009/11/how-many-ancestors-share-our-dna/
On Genetics: http://ongenetics.blogspot.co.uk/2011/02/genetic-genealogy-and-single-segment.html
Wellcome Trust Heterogeneous Stock Mouse Collection: mus.well.ox.ac.uk/mouse

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box) | S3 (box) | S4 (box) | S5 (box) | S6 (box) | S7 (box) | S8 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF