Genetic relatedness analysis: modern data and new challenges

Bruce S. Weir, Amy D. Anderson* and Amanda B. Hepler‡*

Abstract | Individuals who belong to the same family or the same population are related because of their shared ancestry. Population and quantitative genetics theory is built with parameters that describe relatedness, and the estimation of these parameters from genetic markers enables progress in fields as disparate as plant breeding, human disease gene mapping and forensic science. The large number of multiallelic microsatellite loci and biallelic SNPs that are now available have markedly increased the precision with which relationships can be estimated, although they have also revealed unexpected levels of genomic heterogeneity of relationship measures.

Additive variance
The portion of the variance of a quantitative trait that is due to the single effects of alleles at the loci that influence the trait.

Dominance variance
The portion of the variance of a quantitative trait that is due to the interaction of the two alleles that an individual carries at the loci that influence the trait.

Affected-relative linkage studies
Studies that aim to estimate the degree of linkage between a disease and a marker locus on the basis of the marker genotypes of relatives who have the disease.

**Department of Biostatistics, University of Washington, BOX 357232, Seattle, Washington 98195-7232, USA.*
‡Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
Correspondence to B.S.W.
e-mail: bsweir@u.washington.edu
doi:10.1038/nrg1960

The concept of genetic relatedness is central to many aspects of life: marriage and inheritance laws are, at least in part, based on the degree of relationship among members of the same family. Similarly, forensic scientists need to know the degree of relatedness among members of the same population to estimate match probabilities for DNA profiles. In agriculture, measurements made on related individuals can be used to estimate the additive and dominance components of variance, which in turn are needed to predict the gain from breeding programmes for domesticated plant and animal species. In human genetics, a powerful approach to mapping disease genes is based on comparing the genetic marker profiles of affected relatives, and such affected-relative linkage studies require that family relationships be accurately known. In an ecological context, mating strategies in conservation programmes for endangered species, for example, require knowledge of the relatedness of potential mates. Relatedness reflects the shared history of members of the same family or the same population, and so it affects all characters that have a genetic component.

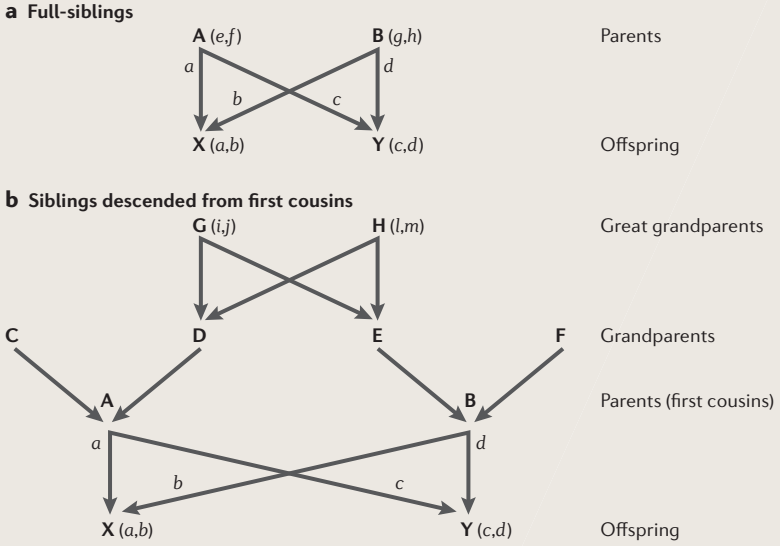
Studies of relationship are phrased in terms of probabilities that sets of genes have descended from a single ancestral gene — that is, the probability that they are identical-by-descent (IBD). There is a probability of one-in-four, for example, that an individual would receive identical copies of a gene from its parents if those parents were siblings. This is just the chance that both alleles have descended from the same one of the four alleles that are carried by the two grandparents. If that particular form of the gene were deleterious, the danger to the health of a child who receives two copies is sufficiently high that it probably accounts for the prohibition of marriage between siblings in all human societies.

Two individuals are said to be related if the allele or alleles of one are IBD to those of the other. This review begins by explaining some of the basic concepts in relatedness analysis. We then describe the statistical framework that is used to link observed genotypes to the probabilities of the IBD status of the constituent alleles, and therefore to the probabilities of particular relationships between the genotypes. These probabilities, which are derived on the basis of observed genotypes, can then be used to make statistical inferences about the degree of relatedness. For example, if two individuals are both heterozygous for different alleles at a microsatellite marker, their four alleles at that locus cannot be IBD. The observation would favour the hypothesis that they are unrelated over the hypothesis that they are half-siblings, although it might still be desirable to attach probabilities to the various possibilities for their actual relatedness.

We also discuss variation in relatedness across the genome. Two types of marker — multiallele microsatellites and the more numerous biallelic SNPs — are currently used for relatedness analysis. Both types of marker have revealed considerable variation in the degree of relatedness along a chromosome; this means that methods for mapping disease genes, devising breeding patterns or determining the probabilities of coincidental matches among forensic profiles might need to be tailored to specific genomic regions.

Current markers also allow relatedness to be studied when relatives are themselves inbred, that is, when their parents are related. Because of this added complexity,

Box 1 | **Measures of relatedness**



Relatedness measures are the probabilities of the identity-by-descent patterns that are possible among the four alleles of a gene that two individuals share. The figure shows some example pedigrees. Individuals are indicated by capital letters and allele labels are given in brackets. It is important to note that the genotypes are unknown: the notation (*a*,*b*) is simply a convenient way of referring to the constituent alleles of an individual.

Consider, first, one allele chosen at random from each individual: the coancestry coefficient between the two individuals is the probability that those two alleles are identical-by-descent (IBD). In panel **a** individuals X and Y are full-siblings whose parents are A and B. An allele (*a* or *b*) from individual X has a one-in-four chance of having descended from the same parental allele (*e*, *f*, *g* or *h*) as an allele (*c* or *d*) from its sibling Y. This common origin means that the two alleles from X and Y are IBD, so the coancestry coefficient of full-siblings is one-in-four. Also, there is a one-in-two chance that a random allele (*a*) from X descends from parent A and a further one-in-two probability that *a* has descended from one or the other random allele (*e* or *f*) from A. The coancestry coefficient for parent–child is therefore also one-in-four.

A more detailed description gives the number of IBD pairs of alleles that any two individuals share: 0, 1 or 2. For the full-siblings X and Y in panel **a**, each of the allele pairs *a*,*c* and *b*,*d* have a one-in-two probability of being IBD (copies of the same parental allele) independently of the other pair. The three events of neither, either or both pairs being IBD are therefore one-in-four, one-in-two and one-in-four, respectively ($k_2 = k_0 = 1/4$, $k_1 = 1/2$). For X, allele *a* must be IBD to one of the alleles in parent A, whereas allele *b* cannot be IBD to either of A's alleles. The event of one pair of IBD alleles for parent and child therefore has a probability of one ($k_2 = k_0 = 0$, $k_1 = 1$).

This review refers to recent work that provides estimates of a more detailed set of up to 15 probabilities[3] — these refer to all possible patterns of IBD among the four alleles for two individuals and are needed when the alleles within individuals are IBD — meaning the individuals are inbred. Consider any two individuals X and Y with alleles *a*,*b* and *c*,*d*. There might be no identity among the four alleles or any of the six pairs of alleles (*a*,*b*; *a*,*c*; *a*,*d*; *b*,*c*; *b*,*d*; *c*,*d*), or any of the four triples of alleles (*a*,*b*,*c*; *a*,*b*,*d*; *a*,*c*,*d*; *b*,*c*,*d*) might be IBD. There are also three possibilities that there are two IBD pairs (*a*,*b* and *c*,*d*; *a*,*c* and *b*,*d*; *a*,*d* and *b*,*c*) and, finally, all four alleles (*a*,*b*,*c*,*d*) might be IBD. For the siblings X and Y in panel **a**, it is only the two pairs *a*,*c* and *b*,*d* that might be IBD and the expanded set of measures is not needed. For siblings whose parents are first cousins, as in panel **b** (in which parents A and B of full-siblings X and Y are themselves first cousins with common grandparents G and H), all 15 IBD patterns are possible because each of the alleles *a*,*b*,*c*,*d* could have originated from any of the four alleles *i*,*j*,*l*,*m* that are carried by the two grandparents G and H who are shared by the cousins.

In most applications there is no need to distinguish between maternal and paternal alleles and the number of IBD classes reduces from 15 to 9: the event that alleles *a*,*b*,*c* are all IBD can be combined with the event that alleles *a*,*b*,*d* are all IBD, for example.

only the use of independent markers for pairs of individuals is covered in this review. Allowing for linkage or linkage disequilibrium between the markers, and the inclusion of multiple relatives, would greatly increase the number of relatedness parameters and would therefore complicate the analysis.

This review is intended to inform agricultural, ecological, forensic and medical geneticists about the ways to describe relatedness between pairs of individuals. Methods are described for using modern genetic marker data to estimate the degree of relatedness between individuals or to address suggested degrees of relatedness. The review also emphasizes that the actual relatedness for specific genes differs from the amount predicted by the genealogical history of the individuals.

Basic concepts in relatedness analysis

From genotypes to inferences about relatedness. Identity-by-descent is crucial to measuring relatedness; however, it is an unobservable quantity. What can often be observed instead are the allelic states (that is, the genotypes) at a locus, and so the challenge is to move from observation to inference about relatedness. For example, the observed allelic states of an unidentified body and the brother of a missing man can be used to address the question of whether the body is that of the missing man. Furthermore, the allelic states that were observed at several genetic markers for plants of the Cabernet Sauvignon, Cabernet Franc and Sauvignon Blanc grape varieties allowed a determination that the first variety was the offspring of the other two[1].

Alleles that seem to be the same are termed 'identical-in-state'. This could mean that they are both the same base type for a SNP or that they both have the same number of repeat units for a microsatellite. Identity-in-state does not generally equate to identity-by-descent, although it is sufficient to infer identity-by-descent in some cases. For example, if the allelic pairs for two parents are A_1A_2 and A_2A_2 then the A_1 alleles that are carried by their children, who both have genotypes A_1A_2, must be IBD, as they are copies of the same parental allele; by contrast, it is not known whether the childrens' A_2 alleles are copies of the same or different parental alleles.

In this review, we show that it is reasonably straightforward to find the probability of the genotypes of individuals when their relationship is known, but that it can be difficult to do the reverse and infer the probability of a relationship given the genotypes — as is required for most practical applications. So, if a man is observed to be homozygous A_1A_1 and it is known that the frequency of the allele A_1 in the population is 0.2, then the probability that his brother is also A_1A_1 is 0.36, as opposed to the value of 0.04 for unrelated people. However, it is difficult to determine the relationship of two men who are both observed to be A_1A_1, because even unrelated people can have the same genotype. We show that more reliable statements about the degree of relatedness are possible when more genetic markers are used.

Microsatellite
Also known as a short tandem repeat. A class of repetitive DNA that is made up of repeats that are 2–5 nucleotides in length. The number of these repeats is usually extremely variable in a population.

Linkage disequilibrium
The non-random association of alleles at different loci, whether or not the loci are linked.

Minisatellite
A region of DNA in which repeat units of 10–50 bp are tandemly arranged in arrays that are 0.5–30 kb in length.

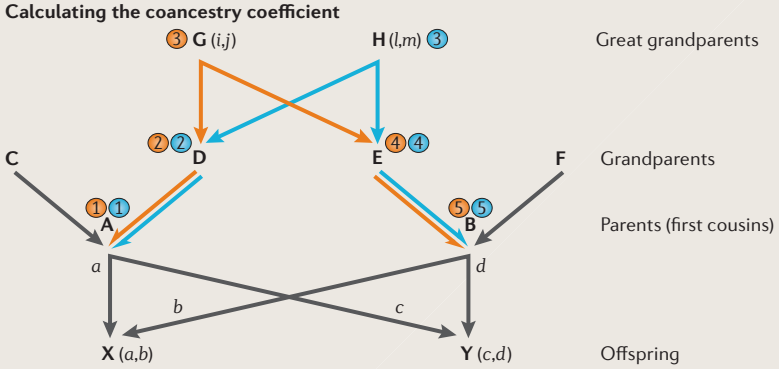
Association study
A study that aims to identify the joint occurrence of two genetically encoded characteristics in a population. Often, an association between a genetic marker and a phenotype (for example, a disease) is assessed.

Molecular marker types. The analysis of more markers increases the reliability of relationship inference and allows more detailed statements to be made, especially given the availability of microsatellite loci and large numbers of SNPs. Relatedness analysis in humans began with paternity testing in the 1920s, but the small number of marker loci (usually involving the ABO, Rhesus and MNS blood-group antigens), the small number of alleles at each locus and the dominance of some alleles over others made it impossible to be precise about estimating the relationship. In 1985 minisatellite markers were introduced by Jeffreys *et al.*[2], but even then the lack of direct correspondence between the observed bands (of the 'DNA fingerprint') on a gel and specific alleles at a locus made it difficult to quantify relatedness.

Minisatellites have now given way to microsatellites and, more recently, to SNPs. Microsatellites have been used widely in paternity testing and forensic science since the mid-1990s: these markers have the advantage of being multiallelic and co-dominant (so that there is no masking of one allele by another). The high degree of variability of microsatellite markers also makes them invaluable for human genetic linkage studies: all individuals within the CEPH (Centre d'Etude du Polymorphisme Humain) linkage panel have been typed at over 32,000 microsatellite markers. Association studies now almost exclusively make use of biallelic SNPs, which are present in far greater numbers; for example, the International HapMap Project data set has genotypes on almost six million SNPs. A comparison of the utility of microsatellite and SNP markers for relatedness estimation is given later.

Box 2 | **Calculating the coancestry coefficient**

Calculating the coancestry coefficient



To calculate the inbreeding coefficients in pedigrees, an individual's genealogy is traced back on both the maternal and paternal sides until an ancestor that is common to both lineages is found. The number *n* of individuals in the pathway that link the parents to the common ancestor, including the parents themselves, is used as a power of 0.5, and the 0.5^n terms are added over all pathways and common ancestors.

For first cousins, such as A and B, $n = 5$; as there is one path (shown in colour) to each of the two grandparents G and H whom they have in common, the inbreeding coefficient of their child X is $2(0.5)^5 = 1/16$. If a common ancestor (for example, G) is himself inbred, then his contribution to the inbreeding coefficient (F) of the descendant is $(1 + F_G)(0.5)^n$.

For the two siblings, X and Y, whose parents, A and B, are first cousins, there are four common ancestors: A and B and the two great-grandparents G and H. There are three people in the paths XAY and XBY, and seven people in each of the paths XADGEBY, XBEGDAY, XADHEBY and XBEHDAY (the paths are not shown). The coancestry coefficient for X and Y is therefore $2(0.5)^3 + 4(0.5)^7 = 9/32$.

Background relatedness. Any two individuals in a finite population are related in the sense that they must have a common ancestor at some point in the past. This means that any relatedness between individuals occurs against a background level of relatedness in the population. Background relatedness is low for human populations, but statistical tools that allow its effects to be quantified are needed for both human and non-human populations. Its effects can be felt in human linkage studies: if background relatedness is neglected, the predicted level of allele sharing between affected relatives will be less than it should be and the increased difference between the predicted and observed sharing might lead to false declarations of linkage. In conservation biology, the relatedness between potential mates could be underestimated if population effects are ignored, and this can lead to increased homozygosity and reduced fitness among the resulting offspring.

This review describes how to measure the relatedness of any two individuals who might themselves be inbred, either as a consequence of inbreeding within the family or by belonging to the same population.

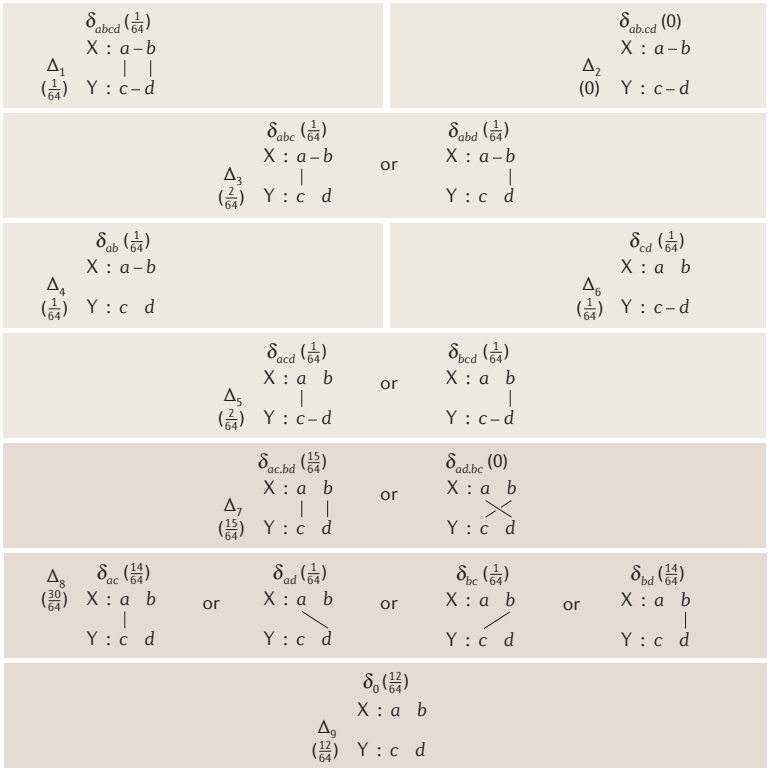
Statistical methods: general principles

Here we describe the general principles of the statistical methods that underlie relatedness analysis. The possible patterns of identity-by-descent among the alleles that are carried by two individuals are described, and how the probabilities of these patterns can allow one to measure the degree of relationship. For example, the allele that a child receives from its parent is IBD to one of that parent's alleles; we can therefore say that parent and child share exactly one pair of IBD alleles, and quantify the relationship by saying that the probability of one pair of IBD alleles is one. We can conclude that two people who are observed to have no alleles in common cannot be related as parent and child.

Inferential studies fall into two broad categories: the observed genotypes can be used to distinguish between a set of possible alternative degrees of relationship, or they can be used to estimate an unknown degree of relationship. For example, it would be possible to use the observed genotypes of individuals X and Y, who are offspring of parents A and B (BOX 1a), to conclude that they are more related than cousins. Alternatively, the actual degree of relatedness between X and Y can be estimated.

This section begins with a definition of the probabilities of identity-by-descent and discusses the evaluation of these quantities. We then show how these probabilities allow genotype probabilities for related individuals to be expressed in terms of allele frequencies. The section concludes with a discussion of distinguishing between alternative relationships or estimating the degree of relationship for two individuals.

Measures of relatedness. Characterizing the relatedness between individuals rests on the probabilities that their alleles are IBD. Before undertaking any relatedness analysis it is important to establish the level of detail needed. In theory, it would be possible to use a single number



$$\theta = \frac{1}{4}(\delta_{abcd} + \delta_{abc} + \delta_{acd} + \delta_{ac.bd} + \delta_{ac}) + \frac{1}{4}(\delta_{abcd} + \delta_{abd} + \delta_{acd} + \delta_{ad.bc} + \delta_{ad})$$
$$\frac{1}{4}(\delta_{abcd} + \delta_{abc} + \delta_{bcd} + \delta_{ad.bc} + \delta_{bc}) + \frac{1}{4}(\delta_{abcd} + \delta_{abd} + \delta_{bcd} + \delta_{ac.bd} + \delta_{bd})$$
$$= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8(\frac{9}{32})$$
$$= k_1/2 + k_1/4$$

Figure 1 | **Complete set of identity-by-descent measures.** Suppose individuals X and Y have alleles labelled *a*,*b* and *c*,*d* at some locus. The 15 patterns of identity-by-descent among the four alleles are shown here and the corresponding probabilities (δ) are given using the notation of Cockerham[33]. Alleles that are identical-by-descent (IBD) are shown within shaded boxes. It is generally neither possible nor necessary to distinguish between maternal and paternal alleles, and this leads to the reduction of the 15 δ probabilities to 9 IBD arrangements, for which the probabilities are shown as Jacquard's coefficients[34] (Δ_{1-9}) in the figure (shaded boxes). For non-inbred individuals, the 15 probabilities are collapsed to a set of three *k* coefficients (k_0, k_1, k_2), which are indicated by the darker shaded boxes. Numerical values shown in parentheses are the δ and Δ probabilities of the IBD pattern for the pedigree in BOX 1b. Each set can be summarized by the coancestry coefficient (θ) for those individuals. This coefficient is the probability that a random allele (*a* or *b*) from one of them is IBD to a random allele (*c* or *d*) from the other. The calculation of θ is provided below the figure using δ, Δ or *k*.

Inbreeding coefficient
The probability that an individual carries two identical-by-descent alleles at a locus.

Coancestry coefficient
The probability that two alleles at a locus, one taken at random from two individuals, are identical-by-descent. It is also called the coefficient of parentage or coefficient of consanguinity.

— the probability that a random allele from one individual is IBD to a random allele from the other — but then it would not be possible to distinguish between full-sibling and parent–offspring relationships, for example. A more useful description uses three probabilities: those for the individuals having zero, one or two pairs of IBD alleles (k_0, k_1 and k_2).

As an introduction, consider the identity of the two alleles *a*,*b*, which are carried by an individual (nothing is implied about identity-in-state by this notation). Then, the probability that *a* and *b* are IBD is defined as *F*, the inbreeding coefficient of the individual. This and all other probabilities of identity among alleles are defined relative to some reference point in the past (the point at which all ancestors are assumed to be unrelated). For a child of a first-cousin marriage, this reference point might be the grandparents of the cousins, in which case the inbreeding coefficient is one-sixteenth: there is a one-in-four chance that the two alleles of the child both come from the parents' common grandparents, and then a one-in-four chance that they are from the same grandparental allele. For the case of full-siblings, in which the inbreeding coefficient of their child is one-quarter (BOX 1a), the reference population is the parents of the siblings. Inbreeding coefficients in pedigrees can be calculated by a simple counting rule, which is described in BOX 2.

Moving on to the relationship between individuals, the coancestry coefficient of two individuals is the same as the inbreeding coefficient of any child they might have. The coefficient can be calculated for pedigrees by applying the counting rule to the path that links the individuals to their common ancestor(s) (BOX 2).

The most detailed description of relatedness is needed to accommodate the additional inbreeding that results from moving the reference point further back in time. In the case of cousins, grandparents might be regarded as being taken from a population in which all members are considered to be related because of the evolutionary history of the population. The magnitude of this background relatedness depends on the loci under consideration, as it is influenced by mutation.

In this framework, it is recognized that alleles within, as well as between, individuals might be IBD; this consideration leads to 15 possible IBD patterns[3] for the four alleles that are carried by two individuals (BOX 1b; FIG. 1). It is generally neither possible nor necessary to distinguish maternal and paternal alleles, and this leads to a reduction to nine IBD arrangements (FIG. 1). When the two individuals are inbred to the same degree, the number of IBD states reduces to seven. There is a final reduction when neither individual is inbred relative to the reference population, as then neither *a*,*b* nor *c*,*d* are IBD: in this case there are only three IBD states. The values of the three probabilities that are needed for some common non-inbred relationships are given in TABLE 1.

Note that throughout this review it is assumed that mutation destroys identity-by-descent. Specific mutation regimes can be postulated if IBD measures are to be predicted, but the estimation procedures to be described do not need specification of mutation or any other evolutionary process.

Joint genotypic probabilities. In the previous section we described the pattern of IBD status in a pedigree, but it is genotype rather than IBD status that can be observed, and the first step in making inferences about IBD probabilities is to express the genotype probabilities for pairs of individuals (or 'joint probabilities') as functions of the allele probabilities. The probability that two unrelated and non-inbred people are both homozygous A_1A_1 is P_1^4, whereas the probability that two full-siblings are homozygous A_1A_1 is $P_1^2(1 + P_1)^2/4$.

Table 1 | **Identity-by-descent probabilities for common, non-inbred relatives**

Relationship	k_2	k_1	k_0	$\theta = k_1/4 + k_2/2$
Identical twins	1	0	0	1/2
Full-siblings	1/4	1/2	1/4	1/4
Parent–child	0	1	0	1/4
Double first cousins	1/16	3/8	9/16	1/8
Half-siblings*	0	1/2	1/2	1/8
First cousins	0	1/4	3/4	1/16
Unrelated	0	0	1	0

*Also grandparent–grandchild and avuncular (for example, uncle–niece). The table shows the three identity-by-descent probabilities (k_{0-2}) and the coancestry coefficients (θ) for common relationships. Note that the coancestry coefficient for full-siblings and parent–child is the same (1/4), but that the pattern of allele sharing is different in each case (that is, there is a different set of k values). k_i, the probability of sharing i number of identical-by-descent alleles (where i = 0–2; see also BOX 1; FIG. 1; θ, the coancestry coefficient of two individuals (equivalent to the inbreeding coefficient of their offspring).

Unordered genotypes
The probability of unordered genotypes does not require specifying which genotype belongs to which individual (for example, which is for the parent and which is for the child). By contrast, the probability of ordered genotypes requires this information.

For a single individual, the two alleles at a locus are either IBD or not IBD with probabilities F and $(1 - F)$, respectively. In the first situation, the IBD alleles must be the same type, so the chance that they are both of type A_i is the same as the chance that either of them is of that type; this is the population frequency P_i of that allele. If two alleles at a locus are not IBD then they are independent and each has its own chance P_i of being of type A_i. The probability (Pr) of a homozygote A_iA_i is therefore $\Pr(A_iA_i) = FP_i + (1 - F)P_i^2$, and the corresponding result for a heterozygote A_iA_j, i ≠ j is $\Pr(A_iA_j) = 2(1 - F)P_iP_j$. The factor of 2 allows for each allele to be either maternal or paternal. The same logic leads to the joint probabilities of all seven possible pairs of unordered genotypes, which are shown in TABLE 2.

Distinguishing between relationships. In paternity testing, it is necessary to decide whether an individual is the father of a child or unrelated to the child. For remains identification, it is necessary to decide whether the remains are from a person with a specified relationship to a family member of a missing person. Although an absolute determination of relationship cannot be made, it is possible to find which of the competing putative relationships makes the observed genotypes most probable by using likelihood ratios, which compare the probabilities of the observed genotypes under alternative hypotheses about relationships. For non-inbred relatives, when only the three relationship coefficients are needed, and in the case in which the alternative is that the individuals are unrelated, the likelihood ratio has a simple form[4] (Supplementary information S1 (box)).

Approaches based on likelihood ratios have been used since the earliest days of paternity testing. Here, the putative relationships are that the alleged father is indeed the father of a child or that he is unrelated to the child, and the likelihood ratio is called the paternity index. In a forensic setting, the relationship alternatives might be 'self' or 'unrelated': the suspect in a crime is either the source of a biological stain or is unrelated to the source of that stain.

More recently, a likelihood ratio expression was used[4] to identify remains from the World Trade Center; genotypes from tissue found at the site and from a family member of a missing person were examined for possible full-sibling or parent–offspring relationships. This approach considerably reduced the number of calculations that would have been necessary if all the possible relationships between a tissue sample and everyone who had lost a relative were considered. In practice it can be difficult to distinguish between full- and half-siblings, because loci with the same genotype are more common in full-siblings whereas loci with different genotypes are more common in half-siblings[5]. Nevertheless, provided the two degrees of relationship that are being

Table 2 | **Joint genotypic probabilities**

	Genotypes	Genotypic state	Number of shared alleles	General	Non-inbred
1	A_iA_i, A_iA_i	Hom/hom	2	$\Delta_1P_i + (\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7)P_i^2 + (\Delta_4 + \Delta_6 + \Delta_8)P_i^3 + \Delta_9P_i^4$	$k_2P_i^2 + k_1P_i^3 + k_0P_i^4$
2	A_iA_i, A_jA_j	Hom/hom	0	$\Delta_2P_iP_j + \Delta_4P_iP_j^2 + \Delta_6P_i^2P_j + \Delta_9P_i^2P_j^2$	$k_0P_i^2P_j^2$
3	A_iA_i, A_iA_j	Hom/het	1	$\Delta_3P_iP_j + (2\Delta_4 + \Delta_8)P_i^2P_j + 2\Delta_9P_i^3P_j$	$k_1P_i^2P_j + 2k_0P_i^3P_j$
4	A_iA_i, A_jA_m	Hom/het	0	$2\Delta_4P_iP_jP_m + 2\Delta_9P_i^2P_jP_m$	$2k_0P_i^2P_jP_m$
5	A_iA_j, A_iA_j	Het/het	2	$2\Delta_7P_iP_j + \Delta_8P_iP_j(P_i + P_j) + 4\Delta_9P_i^2P_j^2$	$2k_2P_iP_j + k_1P_iP_j(P_i + P_j) + 4k_0P_i^2P_j^2$
6	A_iA_j, A_iA_m	Het/het	1	$\Delta_8P_iP_jP_m + 4\Delta_9P_i^2P_jP_m$	$k_1P_iP_jP_m + 4k_0P_i^2P_jP_m$
7	A_iA_j, A_mA_l	Het/het	0	$4\Delta_9P_iP_jP_mP_l$	$4k_0P_iP_jP_mP_l$

The table shows seven distinct patterns of genotypes that are possible for two unordered individuals, and the probabilities of these pairs of genotypes in general, or assuming no inbreeding. Two genotypes could be homozygous (hom) for the same or different alleles (rows 1 and 2), one could be homozygous and the other heterozygous (het) with one or zero shared alleles with the homozygote (rows 3 and 4), or both individuals could be heterozygous with two, one or zero shared alleles (rows 5–7). There are nine pairs of genotypes if the ordering of individuals is important (not shown), as the genotypes in rows 3 and 4 (one homozygote and one heterozygote) each have two orders. k_i, the probability of sharing i number of alleles that are identical-by-descent (where i = 0–2; see also FIG. 1); P, allele frequency; Δ_{1-9}, Jacquard coefficients, which are measures of identity-by-descent status (BOX 1; FIG. 1).

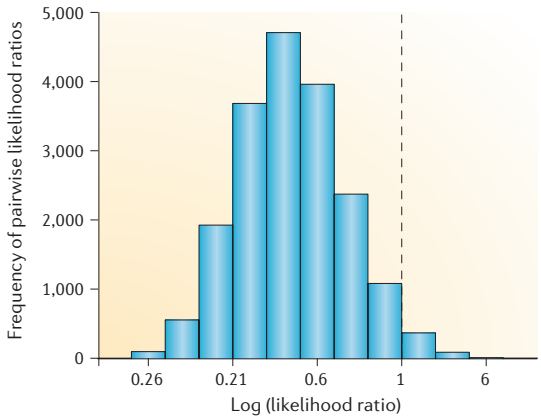


Figure 2 | **Likelihood ratios for putative full-siblings.** Owing to the probabilistic nature of relationship inference, distinguishing between relationships can be difficult, and evidence for a relationship might be found when none is justified. This point is exemplified by a study of 195 Caucasians[35] who were drawn randomly from a population and are presumed to be unrelated. All pairs of individuals in the sample were typed for all of 13 microsatellite loci (the CODIS forensic set[35]). For each pair of individuals, the likelihood ratio for the hypotheses of full-siblings versus unrelated was calculated for each locus and the results were multiplied over loci. A histogram of the 18,195 log-likelihood ratio values is shown: just under 3% of the values have a likelihood ratio greater than 1 (those to the right of the vertical dashed line), which would favour the hypothesis of sibship.

Likelihood ratio
The ratio of two probabilities for the same observations, calculated under alternative hypotheses. In the context of relatedness analysis, the likelihood ratio is formed by dividing the probability of the observed pair of genotypes using the identical-by-descent probabilities for one possible relationship by the probability of the genotypes using identical-by-descent probabilities for the other possible relationship. The likelihood ratio is a continuous variable that can take any non-negative value, and values greater than one support the relationship used for the numerator.

CODIS forensic set
A set of 13 highly polymorphic and essentially unlinked microsatellite markers that were developed by the US Federal Bureau of Investigations for human identification purposes.

Bayesian (framework)
An inference framework in which the posterior probability of a parameter depends explicitly on its prior probability, reflecting some previous belief about this parameter.

Maximum likelihood (method)
The process of estimating parameters by choosing their values to maximize the probability of some observed data.

Bayes theorem
The means of going from a probability of one event, given another, to the probability of the second event, given the first. It is often used to express the (posterior) probability of a hypothesis, given some data, as being proportional to the probability of the data, given the hypothesis, multiplied by the (prior) probability of the hypothesis.

Prior probability
The probability of an event or hypothesis before consideration of some data that will alter the probability of that event or hypothesis.

compared have different relatedness coefficients, they can be distinguished by using a sufficient number of markers (a small number of markers might not allow a distinction[6]).

Reid *et al.*[7] used likelihood ratios to distinguish between 50 pairs of known full-siblings and 50 pairs of known unrelated individuals using a panel of 15 microsatellite loci. The likelihood ratios for comparing full-sibling with unrelated relationships ranged from 4.6 to over 10^9 for the true siblings, and from 4.5×10^{-8} to 0.12 for the unrelated individuals. In this study, classifying a pair as full-siblings when the likelihood ratio is greater than 1 would have given the correct conclusion in all 100 pairs. This is not generally the case, however, as shown in FIG. 2.

These examples demonstrate the probabilistic nature of relationship inference. Even if two individuals are unrelated, it is possible to obtain genetic information that supports the hypothesis that they are related. Similarly, the observed marker genotypes might suggest an incorrect relationship over the correct one for related individuals. This uncertainty is inevitable given the random nature of the choice of which of its two alleles an individual transmits to its offspring, but the use of likelihood ratios allows the most information about the relationship to be extracted from the observed genotypes.

A recent application of likelihood ratios is described on the web site for The a-China DNA Project, which was set up to assist parents who wish to determine whether a Chinese child is a sibling of a child they have already adopted from China. Determining sibling relationships for Chinese family reunions was also an issue following the thawing of political hostility across the Taiwan Strait[8]. In forensics, Bieber *et al.*[9] described a technique known as 'familial searching', in which a genetic profile of interest in a crime is compared to every profile in a database of known offenders with the goal of identifying either a person who has that profile or some close relative of that person.

Estimating relationships: Bayesian approaches. Instead of distinguishing between alternative relationships, it is possible to estimate the actual degree of relationship, or at least to estimate the various relationship parameters (for example, the coancestry coefficient). Two approaches will be considered: Bayesian (this section) and maximum likelihood (next section).

The likelihood ratios that were described in the previous section compare the probabilities of the observed genotypes that are conditional on the assumed relationship. What is needed in practical applications, however, is the probability of a relationship that is conditional on the genotypes. The probability that two individuals are homozygous *AA* given that they are parent and child is P_A^3, but the probability that they are parent and child given that they are both homozygous *AA* cannot be found without additional information. The process of converting the conditional probability of relationship given the genotype to the probability of an observed genotype given a relationship is accomplished with Bayes theorem, and requires the specification of a prior probability of relationship.

If there was prior probability (π_0) for the relationship, compared with being unrelated, then the likelihood ratio (L) for this situation of two alternatives will give a posterior probability (π) from the expression $\pi = [L\pi_0]/[1 + (L - 1)\pi_0]$. This expression is most likely to be useful in a situation in which a relatively small number of remains must be identified, as is the case following an airplane disaster[10,11]: if a genetic profile is available from the parent of one of the 100 victims, then it might be reasonable to assign a prior probability of 1/100 to a parent–child relationship for each of the 100 remains.

Estimating the degree of relationship: maximum likelihood approach. Instead of starting with the prior probabilities of a relationship, an alternative is to estimate the IBD probabilities that characterize relatedness. The best estimation procedure is that of maximum likelihood, whereby the IBD probabilities are chosen to maximize the probability of an observed pair of genotypes (FIG. 3).

Estimating IBD probabilities is useful in situations in which the degree of relationship does not fall into one of the simple standard cases such as full- or half-siblings — for example, when individuals are inbred. Observations from individuals such as X and Y in BOX 1b would support a full-sibling relationship, but simply comparing likelihood ratios for full-siblings versus other standard relationships would fail to detect the increased relatedness that arises from their parents

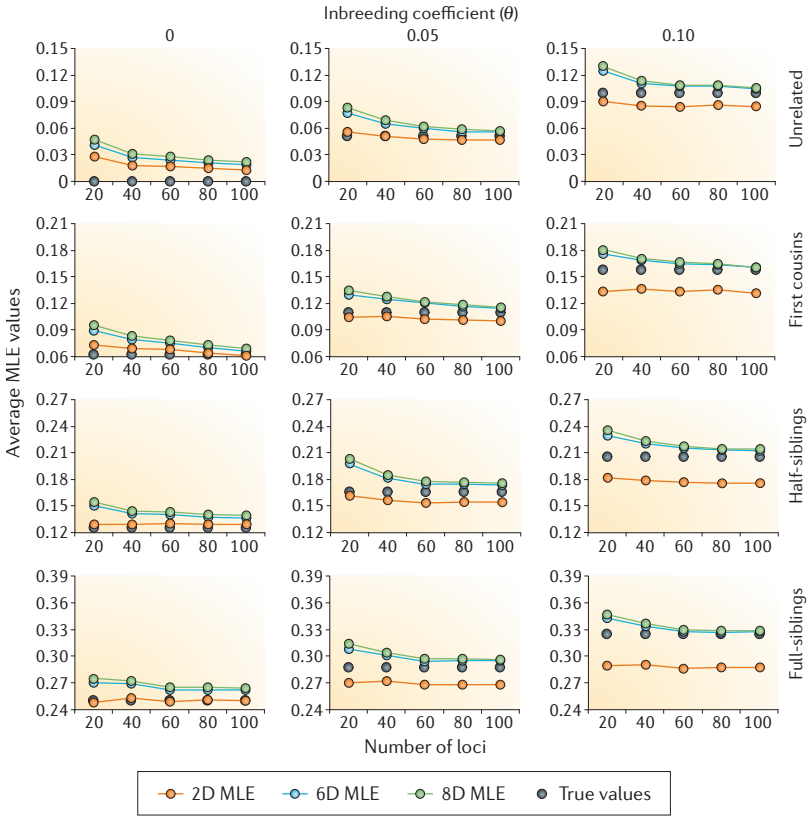


Figure 3 | **Effect of background relatedness on coancestry estimates.** Mean values over 500 replicates of maximum likelihood estimates (MLE) of coancestry coefficients for pairs of individuals with known relationships (unrelated, first cousins, half-siblings and full-siblings) in populations with known background coancestries ($\theta = 0, 0.05, 0.1$). Estimates are based on three, seven or nine identical-by-descent (IBD) coefficients. Because the coefficients sum to 1, only two, six or eight need to be estimated, and the estimates are labelled 2D, 6D or 8D, respectively. When values of θ are low (including $\theta = 0$), the cases that involve a larger number of parameters (6D and 8D) each produced estimates with some bias and sampling error compared with the 2D case (the red line in the $\theta = 0$ column is closer to the black line than either the blue and green lines), leading to less accurate estimates of θ. As θ increases, the 2D parameters become less able to capture the complete pattern of IBD among the four alleles, whereas the coancestry estimates involving 6D and 8D parameters become more accurate.

being cousins. Thompson[12] described the maximum likelihood method in the three-parameter non-inbred case, and Milligan[13] gave the nine-parameter likelihood that allows for inbreeding. Details of maximum likelihood estimation are shown in the Supplementary information S2 (box).

There is an immediate application of maximum likelihood to affected-relative linkage studies, which require knowledge of the relationships among the individuals under study. If the marker-based relationship estimates differ from those inferred from the stated relationship, this indicates that either the relationship is not as stated or the marker genotyping contains an error. In plant or animal breeding there is the additional complication that previous generations of artificial selection could have changed the actual relationships from what would be predicted from known pedigrees, and it is the actual relationships that are needed to predict gains under further selection regimes[14].

Posterior probability
The probability of an event or hypothesis after consideration of some data that have altered the probability of that event or hypothesis.

Population substructure
The existence of groups of individuals within a population that have some degree of reproductive isolation from the rest of the population, and for which the allele frequencies are likely to be different from the population as a whole.

Whether two or more possible degrees of relationship are to be compared for remains identification, or whether the coefficients of relationship are to be estimated for conservation genetics or plant and animal breeding, it is first necessary to express the probabilities of the observed genotypes as functions of IBD measures. In the first case the IBD measures are specified and in the second case they are estimated. In either situation, the availability of rich marker sets allows for more detailed sets of IBD measures to be used than the usual set (k_0, k_1 and k_2; BOX 1). This means that there are several markers in a short region of the genome, and it might be reasonable to assume equality of the IBD status at each marker. The effects of linkage or linkage disequilibrium among such close markers on estimation of the detailed measures remains to be investigated (linkage is discussed in Supplementary information S3 (box)).

New challenges

The substantial amount of genetic marker information that can now be generated with ease and low cost has allowed new questions to be asked. For example, what effect does inbreeding have on estimates of relatedness? Although inbreeding requires a set of nine IBD measures instead of three, how large are these more detailed measures and can small amounts of inbreeding be ignored? The magnitude of the nine IBD measures for individuals in natural or domesticated populations is an open question, but it is under investigation in our laboratories. There is also the new question of which type of marker to use: highly informative microsatellites with multiple alleles, or the much more numerous and inexpensive biallelic SNPs? Dense SNP maps have shown considerable heterogeneity in actual relatedness along the human genome, which might reflect the effects of natural selection in previous generations. Each of these questions — the magnitude of the nine IBD measures and the choice of markers — will now be considered.

Background relatedness. In some situations it is necessary to know the degree of relationship between individuals who do not fit into one of the simple categories in TABLE 1. For example, if affected siblings who are to be used in an affected sib-pair test for linkage have an additional relationship because their parents are cousins (BOX 1B), the challenge is to determine their actual relationship in order to determine the extent of marker allele sharing they would have if the disease locus were unlinked to the marker[15]. Population substructure also poses a challenge because the allele frequencies that are needed for estimating the relationship can vary among subpopulations[16,17]. The use of population-wide allele frequencies for individuals within a subpopulation is one approach to solving this problem, as outlined in Supplementary information S4 (box).

Choice of marker and marker number. Almost all current population genetic studies that use genetic markers use microsatellites or SNPs. The use of microsatellites is now well entrenched in forensic science, whereas SNPs have become the standard for the association mapping

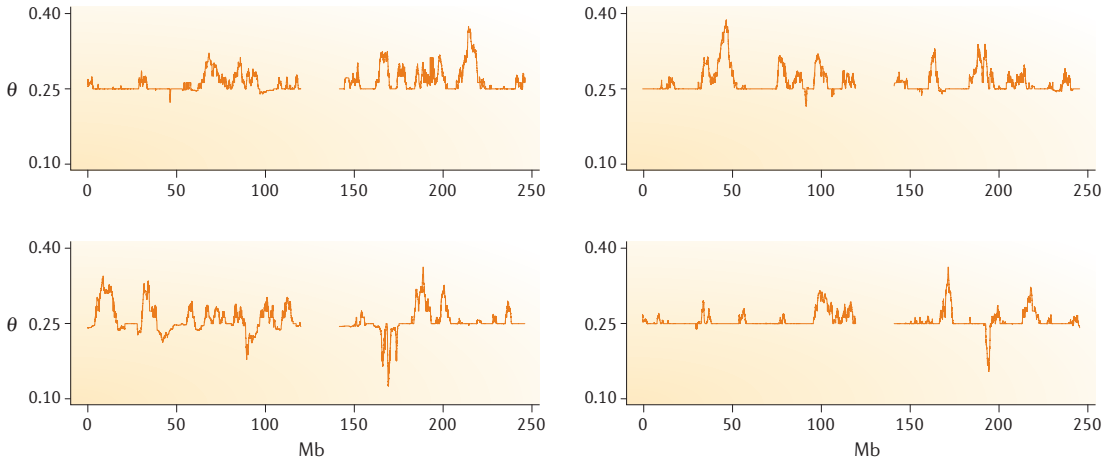


Figure 4 | **Variation in estimated coancestries along a chromosome.** Estimated coancestry coefficients (θ) for four separate parent–child pairs, using successive sets of 100 SNP markers on human chromosome 1. For each pair, the expected coancestry is 0.25, but the estimated actual coancestry varies along the chromosome. The patterns of variation for the four parent–child pairs shown are different, indicating that the variation is not the result of some evolutionary process such as natural selection, but probably represents sampling variation and different genealogies for different chromosomal regions. Because the detailed genealogies of two individuals can vary along the genome, the concept of 'actual' as opposed to 'expected' identity-by-descent becomes important. The variation in actual identity–by-descent is a consequence of evolutionary history and cannot be eliminated, whereas the variation in estimates of identity-by-descent parameters can be reduced (although not eliminated) by using more loci. The smoothing of estimates over loci comes at the expense of masking any variation that might be of real interest.

of human disease genes. The challenge is to weigh the greater discriminatory power of microsatellites against the lower cost of SNPs.

Recent publications[18,19] have described studies in which over one million SNPs are typed per individual. With such a density, is there still a place for the less dense microsatellite marker sets? On an individual marker basis, there is a clear advantage to using multiallelic microsatellites as they allow for all of the seven classes of genotype pairs that are shown in TABLE 2, as opposed to only four classes with biallelic SNPs; the use of multiple alleles is also less likely to allow two individuals to share two pairs of alleles by chance. As many as eight SNPs would be needed to provide the forensic discriminating power of one microsatellite locus[20]. However, despite the advantages of microsatellite markers, the rapid development of commercial SNP-typing technology[21] makes it likely that SNPs will be used in practice because of the cost considerations. Note, also, that although more SNPs than microsatellite loci are available, the greater number of SNPs is partly illusory, as an increased marker density implies increased dependencies due to genetic linkage.

What is the most informative number of SNPs to use? This issue has been addressed empirically by forensic scientists. Gill[22] suggested that 50 SNP loci are needed for forensic identification, in which the relationships that are being compared are 'self' and 'unrelated': relatedness questions involving less than complete genetic identity would require more loci. Amorim and Pereira[23] found that 50 SNP-marker panels were not as informative as 15 or 16 microsatellite panels in assessing parent–child relationships in paternity testing. Hepler[24] compared the use of 50 SNP and 50 microsatellite loci on 15 unrelated and 15 parent–child pairs from the CEPH data and found that one of the unrelated pairs had SNP-based estimates that were close to the values found for half-siblings. It seems that 50 SNPs are insufficient and that 200 SNPs or more will be needed to characterize relatedness.

Genomic heterogeneity of relationship. Estimating IBD probabilities requires data from more than one locus, and this, in turn, assumes equal probabilities at those loci. Recombination between even close loci on the same chromosome, and independent segregation of loci on different chromosomes, means that there is inherent variation in actual IBD values along the genome, along with any variation that might be caused by evolutionary forces such as natural selection. This concept of 'actual' as opposed to 'expected' IBD is important, but little attention has been paid to this variation at the individual level.

The predicted or average IBD status is used in linkage or conservation studies because the actual status is unknown. If the actual level of identity in the region of a disease susceptibility locus, for example, is much greater than that expected for full-siblings, then an affected sib-pair test for linkage might give a false-positive result. Even when identity is averaged over several nearby loci, the standard deviation of the actual IBD can be an appreciable fraction of its expected value[25]; this suggests that there is a need for caution when estimating relatedness from limited regions of the genome, and when inferring the presence of selection if relatedness parameter estimates vary. Some appreciation for the magnitude of genomic variation is given in FIG. 4, which shows the estimated coancestry coefficients for four parent–offspring pairs from HapMap CEU (Caucasians of European origin) trios.

Kin selection
William D. Hamilton's theory to explain the evolution of the hallmark of social life: altruistic cooperation (carrying out functions that are costly to the individual but that benefit others). By helping a relative, an individual increases its fitness by increasing the number of copies of its genes in the population.

Discussion

The study of relatedness between individuals has a rich history, and previous results form the basis of the substantial parentage-testing industry and more recent efforts to identify remains after war or other mass disasters[26]. These successful applications have rested on relatively few genetic markers, and it is generally accepted that some degrees of relationship could not be distinguished by this approach. Now that blood-protein markers have been replaced by microsatellites and overshadowed by SNPs, it has become possible to revisit some old problems. For example, it is no longer necessary to assume that individuals are not inbred or to ignore the accumulation of relatedness by evolutionary processes.

The increased richness of the data, however, has brought the genomic heterogeneity of relatedness to the fore. Descriptions of the degree to which two individuals are related might need to be qualified by stating that conventional values are genome-wide averages and that considerable variation exists. Some of this variation is due to the sampling that is inherent in the evolutionary process (FIG. 4). However, some of the variation might be due to evolutionary forces, such as selection, that affect all members of a population: this is exemplified by elevated coancestry coefficient values in the Caucasian population in the region of the *LCT* (lactase) gene[25] following natural selection for lactose tolerance.

DeWoody[27] has reviewed the applications of relatedness estimation to wildlife populations. Evolutionary hypotheses about mate choice, kin selection and inbreeding can be tested only when there is accurate information on the relatedness of the animals under study. There are also practical implications: reintroduction efforts that rely on translocating wild animals (such as elk[28]) from a donor population should attempt to capture unrelated individuals to reduce the risk to future offspring.

Future work on non-human populations will depend on the development of cost-efficient marker systems. The technology for marker detection (particularly for SNPs) in humans has been mainly driven by the efforts to locate human disease genes by case–control association tests[29]. The efficiency of these studies can be increased when the cases include affected relatives, as then the frequency of high-risk alleles is increased for cases, whereas it is unchanged for unrelated controls[30]. The relationships among people that are used as related cases must be taken into account, but existing analyses have used only the three-parameter IBD measures for the non-inbred situation[30]. It could be valuable to extend these analyses to the more detailed set of IBD measures to accommodate the background relatedness among people that are affected by diseases that have a genetic basis[31].

There is also scope to expand the work covered in this review to quantitative traits. The genetic theory of plant and animal breeding rests on the partitioning of genetic variance into additive, dominance and epistatic components, and these can be estimated from the observed co-variances of quantitative traits between individuals of known relatedness. Little attention has been paid to the magnitude of the complete set of IBD measures that are needed for inbred populations. Some attention has been given to the use of quantitative trait data, as opposed to discrete-marker data, for the estimation of the coancestry coefficient in natural populations, although these efforts have been limited to traits without dominance[32].

The growing amount of genetic marker data for humans and other species means that inferences about relatedness among individuals will be made with increasing precision, and with allowance for inbreeding and evolutionary relatedness. Future progress can be anticipated to allow for the use of multiple linked markers that can be applied to sets of more than two inbred individuals.

1. Bowers, J. E. & Meredith, C. E. The parentage of a classic wine grape: Cabernet Sauvignon. *Nature Genet.* **16**, 84–87 (1997).
2. Jeffreys, A. J., Wilson, V. & Thein, S. L. Individual-specific fingerprints of human DNA. *Nature* **316**, 76–79 (1985).
3. Harris, D. L. Genotypic covariances between inbred relatives. *Genetics* **50**, 1319–1348 (1964).
An early paper in which co-variances in trait values between individuals are expressed as functions of IBD probabilities.
4. Brenner, C. H. & Weir, B. S. Issues and strategies in the DNA identification of World Trade Center victims. *Theor. Popul. Biol.* **63**, 173–178 (2003).
Describes the procedures that are used to identify victims from mass disasters using DNA from known relatives, with a focus on statistical, combinatorial and population genetic issues.
5. Wenk, R. E. & Chiafari, F. A. Distinguishing full siblings from half-siblings in limited pedigrees. *Transfusion* **40**, 44–47 (2000).
6. Gaytmenn, R., Hildebrand, D. P., Sweet, D. & Pretty, I. A. Determination of the sensitivity and specificity of sibship calculations using AmpF/STR Profiler Plus. *Int. J. Legal Med.* **116**, 161–164 (2002).
7. Reid, T. M. *et al.* Specificity of sibship determination using the ABI identifiler multiplex system. *J. Forensic Sci.* **49**, 1262–1264 (2004).
8. Tzeng, C. H. *et al*. Determination of sibship by PCR-amplified short tandem repeat analysis in Taiwan. *Transfusion* **40**, 840–845 (2000).
9. Bieber, F. R., Brenner, C. H. & Lazer, D. Finding criminals through DNA of their relatives. *Science* **312**, 1315–1316 (2006).
10. Olaisen, B., Stenersen, M. & Mevag, B. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genet.* **15**, 402–405 (1997).
11. Leclair, B., Fegeau, C. J., Bowen, K. L. & Fourney, R. M. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair Flight 111 disaster. *J. Forensic Sci.* **49**, 939–953 (2004).
12. Thompson, E. A. Estimation of pairwise relationships. *Ann. Hum. Genet.* **39**, 173–188 (1975).
The classical treatment of maximum likelihood estimation of the three-parameter set of relatedness measures for non-inbred relatives.
13. Milligan, B. G. Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153–1167 (2003).
An important demonstration of the superiority of maximum likelihood methods. Contains details on the implementation and performance of maximum likelihood relatedness estimation when the individuals who are being compared might be inbred.
14. Yu, J. *et al.* A unified mixed-model method for association mapping accounting for multiple levels of relatedness. *Nature Genet.* **38**, 203–208 (2006).
15. Liu, W. & Weir, B. S. Affected sib-pair tests in inbred populations. *Ann. Hum. Genet.* **68**, 606–619 (2004).
The authors develop an analogue of a standard affected sib-pair test for linkage for use in inbred populations.
16. Balding, D. J. & Nichols, R. A. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* **64**, 125–140 (1994).
One of the early papers that deals with the calculation of genotype probabilities for individuals who are allowed to come from a subpopulation of the population from which the allele frequencies have been calculated.
17. Ewens, W. J. *Mathematical Population Genetics. 1. Theoretical Introduction* 2nd edn (Springer, New York, 2004).
A useful reference that contains a treatment of the theory that underlies the calculation of genotypic probabilities for structured populations.
18. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
19. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
20. Ayres, K. The expected performance of single nucleotide polymorphism loci in paternity testing. *Forensic Sci. Int.* **154**, 167–172 (2005).
21. Sobrino, B., Brion, M. & Carracedo, A. SNPs in forensic genetics: a review of SNP typing methodologies. *Forensic Sci. Int.* **154**, 181–194 (2005).
Shows that when IBD is measured with respect to distant ancestry, IBD sharing between two individuals varies appreciably across the genome.
22. Gill, P. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int. J. Legal Med.* **114**, 204–210 (2001).
23. Amorim, A. & Pereira, L. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci. Int.* **150**, 17–21 (2005).

24. Hepler, A. B. *Improving Forensic Identification using Bayesian Networks and Relatedness Estimation*. Ph.D. Thesis, North Carolina State Univ., Raleigh (2005).
25. Weir, B. S., Cardon, L., Anderson, A. D., Nielsen, D. M. & Hill, W. G. Heterogeneity of measures of population structure along the human genome. *Genome Res.* **15**, 1468–1476 (2005).
26. Ballantyne, J. Mass disaster genetics. *Nature Genet.* **15**, 329–331 (1997).
27. DeWoody, J. A. Molecular approaches to the study of parentage, relatedness, and fitness: practical applications for wild animals. *J. Wildl. Manage.* **69**, 1400–1418 (2005).
28. Williams, C. L., Serfass, T. L., Cogan, R. & Rhodes, O. E. Microsatellite variation in the reintroduced Pennsylvania elk herd. *Mol. Ecol.* **11**, 1299–1310 (2002).
29. Slager, S. L. & Schaid, D. J. Evaluation of candidate genes in case–control studies: a statistical method to account for related subjects. *Am. J. Hum. Genet.* **68**, 1457–1462 (2001).
30. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case–control association studies. *PLoS Genet.* **1**, 302–311 (2005).
Demonstrates that unknown relatedness between supposedly unrelated cases or controls can lead to an increased false-positive rate in genetic association studies.
31. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
32. Merila, J. & Crnokrak, P. Comparison of genetic differentiation at marker loci and quantitative traits. *J. Evol. Biol.* **14**, 892–903 (2001).
33. Cockerham, C. C. Higher order probability functions of identity of alleles by descent. *Genetics* **69**, 235–246 (1971).
Cockerham considers the sharing of IBD alleles between two individuals in terms of 15 IBD parameters, develops procedures for calculating the values of these parameters from pedigree data and examines their properties under various mating schemes.
34. Jacquard, A. *Structures Génétiques des Populations* (Masson & Cie, Paris, 1970); English translation available in Charlesworth, D. & Chalesworth, B. *Genetics of Human Populations* (Springer, New York, 1974).
Considers relatedness in terms of nine IBD coefficients: these are now the most commonly used parameters for describing relatedness between two (possibly inbred) individuals.
35. Budowle, B. & Moretti, T. R. Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. *US Department of Justice Forensic Science Communications* [online], < http://www.fbi.gov/hq/lab/fsc/backissu/july1999/budowle.htm > (1999).

Acknowledgements
This work was supported in part by grants from the National Institutes of Health, the National Institute of Justice and the National Science Foundation. We are grateful to W.G. Hill and the reviewers for helpful comments.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION

Centre d'Etude du Polymorphisme Humain: http://www.cephb.fr
Homepage for the Department of Biostatistics, University of Washington: http://www.biostat.washington.edu
International HapMap Project: http://www.hapmap.org
***Nature Reviews Genetics* audio supplement:** http://www.nature.com/nrg/focus/stats/audio
The a-China DNA Project: http://www.a-chinadnaproject.org

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box) | S3 (box) | S4 (box)
Access to this links box is available online.