Title: Ambiguous inference of identity coefficients from independent biallelic loci

Author: Miklós Csűrös

Author's affiliation: University of Montreal, Department of Computer Science and Operations Research, Montréal, Québec, Canada H3C 3J7

Short running title: Ambiguous identity coefficients

Key words: genetic relatedness, identity by descent, non-identifiability

Corresponding author: Miklós Csűrös

Mailing address:

Université de Montréal

Pavillon André-Aisenstadt

Département d'informatique et de recherche opérationnelle

Faculté des Arts et des Sciences

C.P. 6128, succ. Centre-Ville

Montréal, Qué. H3C 3J7

Canada

Courrier address:

Université de Montréal

Département d'informatique et de recherche opérationnelle

Local 2145, Pavillon André-Aisenstadt

2920 chemin de la Tour

Montréal, Qué. H3T 1J4

Canada

Phone: ++1 (514) 343-6111 extension 1655

E-mail: csuros@iro.umontreal.ca

Abstract

Shared genealogies introduce allele dependencies in diploid genotypes, as alleles within an individual or between different individuals will likely match when they originate from a recent common ancestor. The genotype distribution is thus a mixture of distributions proper to different modes corresponding to combinatorially distinct patterns of identity by descent (IBD). In the absence of known pedigrees, the genetic relatedness between two individuals is described by the nine-element distribution comprising the probabilities for different identity modes, called (Jacquard's) identity coefficients. At a locus with two possible alleles, identity coefficients are not identifiable from the joint genotypes because different coefficients can generate the same genotype distribution.

We analyze precisely how different identity modes combine into identical genotype distributions at diallelic loci. The analysis yields an exhaustive characterization of statistical measures over joint genotype distributions that are stable; i.e., that stay the same for equivalent identity coefficients. Importantly, we show that stable relatedness statistics include the kinship coefficient (probability that a random pair of alleles are identical by descent between individuals) and a number of inbreeding-related measures, which can thus be estimated from genotype distributions at independent biallelic loci despite the non-identifiability of the IBD distribution. We provide simple moment-based estimators for this purpose, and analyze their behavior on various data sets, including horses from various breeds, and human population samples from the 1000 Genomes project.

Non-random mating histories, selection, finite population sizes, and many other causes create dependencies between alleles in a diploid population. Because of joint genealogies, alleles may match within or across genotypes for being unmodified copies of a common ancestral state. Such alleles are said to be *identical by descent* (Malécot, 1969; Jacquard, 1974). Combinatorially distinct partitionings of identical-by-descent (IBD) alleles are called *identity modes* (Jacquard, 1974). Two diploid individuals' joint pedigree defines the possible inheritance histories for four alleles, which combine into a nine-element distribution over the identity modes. Every identity mode generates its own probability distribution over the joint genotypes at a locus, and the observable genotypic distribution is the mixture of the mode-specific distributions. The probabilities of the identity modes, or *identity coefficients*, characterize thus the individuals' genetic relatedness succinctly.

[Figure 1 about here.]

The identity coefficients can be computed for any known pairwise genealogy (Cockerham, 1971; Lange, 1997). Hypothetical pedigrees can be thus assessed by comparing implied genotype distributions with empirical ones (Thompson, 1975; Milligan, 2003). But can identity coefficients be directly inferred from genotype distributions without genealogies? The answer depends on the number of alleles. There are 9 identity modes for a pair of diploid individuals (Figure 1), which define 8 independent identity coefficients (the ninth one is implied since the coefficients sum to one). At loci with only two alleles, nine genotype pairs are possible, but because of redundancy, there are not enough many different genotype pairs to make certain inference possible: more than one set of coefficients generate the same joint genotype distribution. Genotype distributions at loci with three or more alleles, however, convey enough information in principle to identify a single set of identity coefficients that produce it. Among molecular markers, multiallelic microsatellite loci provide in consequence high discriminatory power for a detailed characterization of genetic

relatedness, but diallelic single-nucleotide and insertion-deletion have restricted utility Weir *et al.* (2006). Here, we scrutinize the inherent ambiguity of relatedness in diallelic genotypes. Specifically, our aim is to find what aspects of coancestry result in non-identifiability and to characterize statistical measures of the identity mode distribution that can be consistently estimated from joint genotype frequencies.

THEORY AND RESULTS

Identity coefficients and biallelic genotype distributions

Identity by descent (Malécot, 1969) encapsulates the dependence between diploid genotypes due to shared parentage. Two alleles are identical by descent (IBD), if they originate from a common ancestral allele without modification. Equivalence relations for four alleles of two diploid genotypes take one of nine combinatorially distinct forms (Harris, 1964; Jacquard, 1974; Lange, 1997), or *identity modes*, as illustrated in Figure 1.

The individuals' joint pedigree determines the possible identity modes and their associated frequencies, specified by the vector of coefficients $\Delta_i\colon i = 1, \ldots, 9$ using the notation of Jacquard (1974). For instance, children of the same parents from non-overlapping lineages inherit two IBD alleles with probability $\Delta_7 = \frac{1}{4}$, one IBD set from either parent with probability $\Delta_8 = \frac{1}{2}$, and four independent alleles with probability $\Delta_9 = \frac{1}{4}$.

Suppose that the locus has two alleles, and alleles 1 and 0 (minor and major) occur with frequencies p and q, respectively. Every mode generates its own conditional distribution of joint genotypes. In mode 8, the individuals are 0/1 heterozygotes simultaneously with probability $pq^2 + p^2q$ since either the IBD alleles are the mutants, or two mutant alleles are chosen independently. In contrast, if all four alleles are sampled independently (identity mode 9) then the joint genotype $0/1 : 0/1$ occurs with probability $4p^2q^2$, accounting for two minor and two major alleles in 4 possible orderings. Table 1 lists the complete set of genotypic probabilities.

[Table 1 about here.]

Denote the distribution of joint genotypes by

$$\mathbf{f} = (f_{0000}, f_{1111}, f_{1101}, f_{0111}, f_{0101}, f_{1100}, f_{0011}, f_{0100}, f_{0001}).$$

Table 1 corresponds to the system of equations

$$
\begin{aligned}
f_{0000} &= q\Delta_1 + q^2\Big(\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7\Big) + q^3\Big(\Delta_8 + \Delta_4 + \Delta_6\Big) + q^4\Delta_9 \\
f_{1111} &= p\Delta_1 + p^2\Big(\Delta_2 + \Delta_3 + \Delta_5 + \Delta_7\Big) + p^3\Big(\Delta_4 + \Delta_6 + \Delta_8\Big) + p^4\Delta_9 \\
f_{1101} &= pq\Big(\Delta_3 + p(\Delta_8 + 2\Delta_4 + 2p\Delta_9)\Big) \\
f_{0111} &= pq\Big(\Delta_5 + p(\Delta_8 + 2\Delta_6 + 2p\Delta_9)\Big) \\
f_{0101} &= pq\Big(2\Delta_7 + q\Delta_8 + 4pq\Delta_9\Big) \\
f_{1100} &= pq\Big(\Delta_2 + q\Delta_4 + p\Delta_6 + pq\Delta_9\Big) \\
f_{0011} &= pq\Big(\Delta_2 + q\Delta_6 + p\Delta_4 + pq\Delta_9\Big) \\
f_{0100} &= pq\Big(\Delta_5 + q\Delta_8 + 2q\Delta_6 + 2q^2\Delta_9\Big) \\
f_{0001} &= pq\Big(\Delta_3 + q\Delta_8 + 2q\Delta_4 + 2q^2\Delta_9\Big),
\end{aligned}
\tag{1}
$$

or, in matrix form,

$$
\mathbf{f} = \mathbf{F} \cdot \mathbf{\Delta}, \tag{2}
$$

and Table 1 gives the transpose of $\mathbf{F}$.

The matrix $\mathbf{F}$ projects the vector of identity coefficients $\mathbf{\Delta}$ to the vector of genotype probabilities $\mathbf{f}$. Consequently, identity coefficients can be inferred from the biallelic genotype distribution if and only if the matrix $\mathbf{F}$ is invertible. The matrix rows, are however, linearly dependent.

Claim 1. *When $p+q=1$, dependencies between genotype probabilities include the following two.*

$$
f_{1101} + 2f_{1100} + f_{0100} = f_{0111} + 2f_{0011} + f_{0001}; \tag{3}
$$

$$
p = f_{1111} + \frac{3}{4}\Big(f_{1101} + f_{0111}\Big) + \frac{1}{2}\Big(f_{0101} + f_{1100} + f_{0011}\Big) + \frac{1}{4}\Big(f_{0100} + f_{0001}\Big). \tag{4}
$$

Theorem 2 below characterizes the set of identity coefficients that lead to the same distribution over joint biallelic genotypes.

Theorem 2. *Suppose that* $p+q=1$. *If* $\Delta_i\colon i=1,\ldots,9$ *satisfy* (2) *then so do the following coefficients, for all choices of* $\xi,\eta\in\mathbb{R}$.

$$\begin{aligned}
\Delta_1' &= \Delta_1 - \eta pq & \Delta_2' &= \Delta_2 + \xi - \eta pq \\
\Delta_3' &= \Delta_3 + 2\eta pq & \Delta_4' &= \Delta_4 - \xi \\
\Delta_5' &= \Delta_5 + 2\eta pq & \Delta_6' &= \Delta_6 - \xi \\
\Delta_7' &= \Delta_7 - \xi + \eta(1-2pq) & & \\
\Delta_8' &= \Delta_8 + 2\xi - 2\eta & \Delta_9' &= \Delta_9 + \eta.
\end{aligned} \tag{5}$$

Starting from an arbitrary particular coefficient set Δ_i, *Equation* (5) *generates all vector solutions to* (2).

[Figure 2 about here.]

The constraints $\Delta_i' \geq 0$ demarcate the region yielding proper distributions over identity modes; see Eq. (16) in **Methods**. Roughly, the coordinates η and ξ quantify the uncertainty about inbreeding and overall IBD level, respectively. Figure 2 illustrates the quadrilateral solution area for the example of Queen Victoria (of the United Kingdom) and Prince Albert who, in addition to be first cousins, shared multiple ancestors within seven generations.

Identifiable relatedness parameters

Despite multiple solutions, some aspects of the identity coefficients can be ascertained from the genotype distribution. In particular, if a linear combination stays the same for all sets of identity coefficients from (5), then it is computable from the biallelic genotype distribution. Theorem 3 formalizes our argument.

Definition 1. *A function of the identity distribution $\theta(\boldsymbol{\Delta})$ is called a* linear relatedness parameter *if and only if it can be written as a linear combination*

$$\theta(\boldsymbol{\Delta}) = \sum_{i=1}^{9} a_i \Delta_i, \tag{6}$$

where a_i are constants. In particular, a_i may not depend on the allele frequency p.

Theorem 3. *A linear relatedness parameter θ is identifiable from the joint genotype distribution only if*

$$\begin{aligned} a_2 + 2a_8 &= a_4 + a_6 + a_7; \\ a_7 + a_9 &= 2a_8; \qquad\qquad \textit{and} \\ 2a_3 + 2a_5 &= a_1 + a_2 + 2a_7. \end{aligned} \tag{7}$$

Theorem 4. *The following linear relatedness parameters are identifiable from the biallelic genotype distribution.*

$$\theta_0 = \sum_{i=1}^{9} \Delta_i \qquad \textit{(=1)} \tag{8a}$$

$$\theta_1 = \Delta_1 + \frac{1}{2}\big(\Delta_3 + \Delta_5 + \Delta_7\big) + \frac{1}{4}\Delta_8 \qquad \textit{(kinship coefficient)} \tag{8b}$$

$$\theta_{2\mathrm{A}} = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 \qquad \textit{(A's inbreeding)} \tag{8c}$$

$$\theta_{2\mathrm{B}} = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6 \qquad \textit{(B's inbreeding)} \tag{8d}$$

$$\begin{aligned} \theta_3 = {} & \Delta_1 + \Delta_2 + \Delta_3 + \Delta_5 + \Delta_7 \\ & + \frac{1}{2}\big(\Delta_4 + \Delta_6 + \Delta_8\big) \end{aligned} \qquad \textit{(triple coancestry)} \tag{8e}$$

$$\theta_4 = \frac{1}{2}\big(\Delta_4 - \Delta_6\big) \qquad \textit{(independent inbreeding difference)} \tag{8f}$$

All other identifiable parameters are linear combinations of θ_i in Equations (8).

Theorem 3 shows that, in general, the identity coefficients Δ_i are not identifiable separately. In particular, probabilities for various inbred modes $(\Delta_3, \Delta_4, \Delta_5, \Delta_6)$ are not identi-

fiable, only their differences ($\Delta_4 - \Delta_6 = 2\theta_4$ and $\Delta_3 - \Delta_5 = \theta_{2\mathrm{A}} - \theta_{2\mathrm{B}} - 2\theta_4$).

The identifiable parameters of Theorem 4 include the usual measures of inbreeding (θ_{2*}) and coancestry (θ_1) generalized to inbred parents (Harris, 1964), as well as the trivial $\sum_i \Delta_i$. Note that the matrix structure automatically guarantees $\sum_i \Delta_i = 1$ when $f_{0000} + f_{1111} + \ldots + f_{0001} = 1$ since the all-1 row vector $\mathbf{e} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}$ is a left eigenvector:

$$1 = \mathbf{e} \cdot \mathbf{f} = \mathbf{e} \cdot \mathbf{F} \cdot \boldsymbol{\Delta} = \mathbf{e} \cdot \boldsymbol{\Delta}.$$

The parameter θ_3 is the probability that there is at least one pair of IBD alleles among three randomly selected ones. A simpler three-allele parameter is

$$\theta_{3:3} = \Delta_1 + \frac{1}{2}\big(\Delta_3 + \Delta_5\big) = \theta_1 - \frac{1}{2}\theta_3 + \frac{1}{4}\big(\theta_{2\mathrm{A}} + \theta_{2\mathrm{B}}\big), \tag{9}$$

or *inbred coancestry*, which is the probability that three randomly chosen alleles are simultaneously identical by descent. By Theorem 3, $\theta_{3:3}$ is identifiable, and (9) shows how to write it as a linear combination of identifiable parameters from Theorem 4.

Linear relatedness parameters can be written either as linear combinations of identity coefficients or as linear combinations of genotypic probabilities in the linear algebraic framework of Equation (1). For the "archetypical" parameters of Theorem 4, we consider the following expressions.

$$\begin{aligned}
\tau_{1\mathrm{A}} &= \frac{f_{1101} + f_{0100}}{2} + \frac{f_{0101}}{4} + f_{1100} & \tau_{1\mathrm{B}} &= \frac{f_{0111} + f_{0001}}{2} + \frac{f_{0101}}{4} + f_{0011} \\
\tau_1 &= \frac{\tau_{1\mathrm{A}} + \tau_{1\mathrm{B}}}{2} & & \\
\tau_{2\mathrm{A}} &= \frac{f_{0111} + f_{0101} + f_{0100}}{2} & \tau_{2\mathrm{B}} &= \frac{f_{1101} + f_{0101} + f_{0001}}{2} \\
\tau_3 &= \frac{\big(f_{0100} - f_{0111}\big) + \big(f_{0001} - f_{1101}\big)}{4} & \tau_4 &= \frac{f_{1100} - f_{0011}}{2}
\end{aligned} \tag{10}$$

So,

$$\begin{aligned}
\theta_1 &= 1 - \frac{\tau_{1\mathrm{A}}}{p - p^2} = 1 - \frac{\tau_{1\mathrm{B}}}{p - p^2} = 1 - \frac{\tau_1}{p - p^2} \\
\theta_{2\mathrm{A}} &= 1 - \frac{\tau_{2\mathrm{A}}}{p - p^2} & \theta_{2\mathrm{B}} &= 1 - \frac{\tau_{2\mathrm{B}}}{p - p^2} \\
\theta_3 &= 1 - \frac{\tau_3}{p - 3p^2 + 2p^3} & \theta_4 &= \frac{\tau_4}{p - 3p^2 + 2p^3}.
\end{aligned} \tag{11}$$

Due to the linear dependencies, multiple equivalent formulas exist that relate the genotype distribution and any specific parameter. For instance, the intermediate quantities $\tau_{1\mathrm{A}}$, $\tau_{1\mathrm{B}}$ and τ_1, which weigh genotypic probabilities differently, are equal by Equation (3).

Moment-based relatedness estimators for independent sites

Suppose that minor allele frequencies $p_i \colon i = 1, \ldots, n$ are known for n independent biallelic loci with identical IBD mode distributions. Pairwise genotypes are observed across the n positions, and counted as $n_{0000}, n_{1111}, n_{1101}, \ldots, n_{0001}$. The connection between the genotype counts and the identity coefficients can be expressed in terms of the moments of the minor-allele frequency distribution by averaging Equation (1) across the loci. Write the moments for the minor allele frequency (MAF) distribution as

$$\mu = \frac{1}{n}\sum_{i=1}^{n} p_i \quad \mu_2 = \frac{1}{n}\sum_{i=1}^{n} p_i^2 \quad \mu_3 = \frac{1}{n}\sum_{i=1}^{n} p_i^3 \quad \mu_4 = \frac{1}{n}\sum_{i=1}^{n} p_i^4.$$

By averaging Equation (1), expected genotype frequencies $\mathbb{E}n_{0000}$, $\mathbb{E}n_{1111}$, ... can be related to the common identity coefficients by a matrix expressed in terms of MAF moments (see Eq. (17) in **Methods**). In order to develop estimators for identifiable relatedness parameters using genotype counts and MAF moments, we adopt the formulas of (10) and (11). The estimators assume that independent MAF moment estimates are available. First, set a

scaling parameter

$$\hat{n} = n_{1111} + \frac{3}{4}\big(n_{1101} + n_{0111}\big) + \frac{1}{2}\big(n_{0101} + n_{1100} + n_{0011}\big) + \frac{1}{4}\big(n_{0100} + n_{0001}\big).$$

By (4), $\mathbb{E}\hat{n} = \mu n$. The formulas avoid n_{0000} by normalizing the counts with $\hat{n}$, and by scaling the moments analogously with μ. (The exclusion of n_{0000} is aimed to reduce ascertainment bias caused by experiment design, site selection and genotyping errors that affect that particular tally the most.) The *normalized genotype frequencies* are calculated as

$$\begin{aligned} \hat{f}_{1111} &= \frac{n_{1111}}{\hat{n}} \quad \hat{f}_{1101} = \frac{n_{1101}}{\hat{n}} \quad \hat{f}_{0111} = \frac{n_{0111}}{\hat{n}} \quad \hat{f}_{0101} = \frac{n_{0101}}{\hat{n}} \\ \hat{f}_{1100} &= \frac{n_{1100}}{\hat{n}} \quad \hat{f}_{0011} = \frac{n_{0011}}{\hat{n}} \quad \hat{f}_{0100} = \frac{n_{0100}}{\hat{n}} \quad \hat{f}_{0001} = \frac{n_{0001}}{\hat{n}} \end{aligned} \tag{12}$$

The normalized genotype frequencies are plugged into Equation (10) for the normalized statistics

$$\begin{aligned} \hat{\tau}_1 &= \frac{\hat{f}_{1101} + \hat{f}_{0111} + \hat{f}_{0101} + \hat{f}_{0100} + \hat{f}_{0001}}{4} + \frac{\hat{f}_{1100} + \hat{f}_{0011}}{2} \\ \hat{\tau}_{2\mathrm{A}} &= \frac{\hat{f}_{0111} + \hat{f}_{0101} + \hat{f}_{0100}}{2} \qquad \hat{\tau}_{2\mathrm{B}} = \frac{\hat{f}_{1101} + \hat{f}_{0101} + \hat{f}_{0001}}{2} \\ \hat{\tau}_3 &= \frac{\big(\hat{f}_{0100} - \hat{f}_{0111}\big) + \big(\hat{f}_{0001} - \hat{f}_{1101}\big)}{4} \qquad \hat{\tau}_4 = \frac{\hat{f}_{1100} - \hat{f}_{0011}}{2} \end{aligned} \tag{13}$$

Finally, the identity distribution parameters are estimated by

$$\begin{aligned} \hat{\theta}_1 &= 1 - \frac{\hat{\tau}_1}{1 - \mu'_2} \qquad \hat{\theta}_{2\mathrm{A}} = 1 - \frac{\hat{\tau}_{2\mathrm{A}}}{1 - \mu'_2} \qquad \hat{\theta}_{2\mathrm{B}} = 1 - \frac{\hat{\tau}_{2\mathrm{B}}}{1 - \mu'_2} \\ \hat{\theta}_3 &= 1 - \frac{\hat{\tau}_3}{1 - 3\mu'_2 + 2\mu'_3} \quad \hat{\theta}_4 = \frac{\hat{\tau}_4}{1 - 3\mu'_2 + 2\mu'_3} \end{aligned} \tag{14}$$

with the scaled moments $\mu'_2 = \frac{\mu_2}{\mu}$ and $\mu'_3 = \frac{\mu_3}{\mu}$.

Kinship inference in simulated data

The formulas of (14) employ the joint genotype counts with known or estimated MAF moments. In addition to statistical deviations of the estimated genotype frequencies, the formulas' performance is thus also affected by errors in MAF moment estimation. We assessed the power of the estimator for kinship coefficient (see $\hat{\tau}_1$ and $\hat{\theta}_1$ in Eqs. (13) and (14)) using simulated assays involving n genotyped loci for two individuals, for which MAF moments are estimated separate from a panel of N separate unrelated individuals. In one simulation step, we picked n random SNPs from the 1000 Genomes project and used their minor allele frequencies to generate simulated genotypes. Within each step, we simulated both the MAF estimation and the kinship inference. For MAF estimation, we generated $2N$ independent random alleles (representing N diploid individuals) at every locus with its allele frequency distribution. The moments for the estimated MAFs were employed with the kinship formula of (14), using simulated joint genotypes for a set of four example pedigrees. In particular, we compared the estimated kinship coefficients $\hat{\theta}_1$ for unrelated individuals, first and second cousins, as well as a more complex pedigree of first cousins sharing multiple deeper ancestors (royal cousins).

[Figure 3 about here.]

Figure 3 shows the simulation results for sample sizes $N = 10$, $N = 100$ and $N = 1000$. The plots illustrate that despite the bias of MAF moment estimation at $N = 10$, the different levels of relatedness (unrelated, second and first cousins) are well separated from 10–20 thousand sites. More accurate MAF moment estimation with $N = 100$ and $N = 1000$ results in slightly increased requirements for the necessary number of loci ($n \geq 50000$), plausibly due to the higher sensitivity which entails the inclusion of rarer SNPs. Fine aspects of relatedness, however, are more difficult to ascertain, as can be seen by the imperfect separation of estimated kinship coefficients for royal and common first cousins even with $n \geq$

200000 loci.

Moment-based relatedness estimators on stratified populations

Our framework assumes that MAF moments are the same for the two individuals. Employing the estimators of (14) with inappropriately estimated moments leads to a bias affecting the estimation of relatedness parameters. Figure 4 examines pairwise relatedness estimation assuming different MAF moments. (See additional plots in Supporting Information.) The data set consists of genotypes for 54 individuals belonging to 6 populations from the 1000 Genomes project (The 1000 Genomes Project Consortium, 2010).

[Figure 4 about here.]

The estimated kinship coefficients do not necessarily represent the true relationship between a pair of individuals. For example, both LWK (Luhya from Kenya) and CHS (Han from Southern China) pairs have seemingly high kinship coefficients (on par with siblings and cousins), when the generic MAF distribution is used. Appropriate allele frequencies alleviate the bias: LWK kinship coefficients cluster around 0 when the African MAF moments are used, and so do CHS kinship coefficients when the Asian allele frequency distribution is applied. The figure also makes it apparent that the bias affects all pairs from the same subpopulation in a similar way, so that related individuals do stand out against the base level of unrelated pairs.

[Figure 5 about here.]

In a second set of experiments, we used 733 horses of 32 breeds genotyped with the Equine SNP50 Beadchip (Illumina) from Petersen *et al.* (2013b). The comparisons of different breeds reveal known relationships, including the clustering by recognized breed groups (Petersen *et al.*, 2013a). As it is in the case with human data, using a single

MAF distribution for all pairwise comparisons reveals the population structure by the high relatedness values attributed to horses of the same breeds; see Figure 5. The absolute values of estimated kinship coefficients are misleading due to the bias of using a common MAF moment estimate.

After computing deme-specific MAF moments, relatedness parameters center around 0 as expected: see Figure 6. (See additional plots in Supporting Information.) The figure also illustrates some aspects of genetic diversity within the breeds discussed by PETERSEN *et al.* (2013a). Breeds with high diversity and large population size (including Mongolian, Paint and Tuva) exhibit small deviations in their relatedness parameters. Low within-breed diversity and inbreeding in other breeds such as Clydesdale, Exmoor, Mangalarga Paulista, Shire, and Thoroughbred are reflected in the larger support of the distributions. For these breeds, frequent positive values of inbred ancestry ($\hat{\theta}_{3:3}$) hint at selection for preferred lineages and foundational effects (the British populations decreased significantly during World War II).

[Figure 6 about here.]

A linear algebraic framework for genotypic probabilities and identity coefficients

Proof of Claim 1. The equalities can be seen by inspecting the rows of $\mathbf{F}$, but considering allele counts gives a more straightforward proof. Consider the expected number ω of minor ('1') alleles in the random joint genotype. Since it is the expectation for the sum of four indicator variables, $\mathbb{E}\omega = 4p$. Alternatively, by summing over the possible joint genotypes, $\mathbb{E}\omega = 4f_{1111} + 3\Big(f_{1101} + f_{0111}\Big) + 2\Big(f_{0101} + f_{1100} + f_{0011}\Big) + \Big(f_{0100} + f_{0001}\Big)$, and Equation (4) follows after dividing by 4. Now consider the expected number of minor alleles in the A's and B's genotype separately. Clearly, both equal $2p$. Counting by joint genotypes:

$$\underbrace{2\Big(f_{1111} + f_{1101} + f_{1100}\Big) + \Big(f_{0111} + f_{0101} + f_{0100}\Big)}_{\text{expected count in A}} = \underbrace{2\Big(f_{1111} + f_{0111} + f_{0011}\Big) + \Big(f_{1101} + f_{0101} + f_{0001}\Big)}_{\text{expected count in B}}.$$

After elimination of common terms, Equation (3) follows. □

Proof of Theorem 2. The null space of $\mathbf{F}$ is spanned by the vectors

$$\mathbf{z}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ -1 \\ -1 \\ 2 \\ 0 \end{pmatrix} \qquad \mathbf{z}_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{pmatrix} + pq \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \\ 2 \\ 0 \\ -2 \\ 0 \\ 0 \end{pmatrix} = \mathbf{z}_2^{(1)} + pq\mathbf{z}_2^{(2)} \tag{15}$$

It is straightforward to verify that $\mathbf{F}\mathbf{z}_1 = \mathbf{F}\mathbf{z}_2 = \mathbf{0}$, the null vector. Hence,

$$\mathbf{F} \cdot \Big(\boldsymbol{\Delta} + \xi \mathbf{z}_1 + \eta \mathbf{z}_2 \Big) = \mathbf{F} \cdot \boldsymbol{\Delta} = \mathbf{f}$$

for all choices of ξ, η. The rank of the 9×9 matrix $\mathbf{F}$ is 7 (established by Gaussian elimination), and therefore no other solutions exist. □

Span of equivalent solutions: Values of (ξ, η) for which Eq. (5) produces a proper distribution are precisely those where $\Delta_i' \geq 0$ for all i:

$$\begin{array}{llll} \eta \leq \dfrac{\Delta_1}{pq} & \eta \leq \dfrac{\Delta_2}{pq} + \dfrac{\xi}{pq} & \eta \leq \dfrac{\Delta_8}{2} + \xi & \\ \eta \geq -\dfrac{\Delta_3}{2pq} & \eta \geq -\dfrac{\Delta_5}{2pq} & \eta \geq -\dfrac{\Delta_7}{1-2pq} + \dfrac{\xi}{1-2pq} & \eta \geq -\Delta_9 \\ \xi \leq \Delta_4 & \xi \leq \Delta_6 & & \end{array} \tag{16}$$

Proof of Theorem 3. In order to be identifiable, $\theta(\boldsymbol{\Delta})$ must remain the same for all distributions satisfying (2). The vector of coefficients $(a_i \colon i = 1, \ldots, 9)$ then has to be orthogonal

to the null space of $\mathbf{F}$. The identities of (7) express the orthogonality with the vectors $\mathbf{z}_1$, $\mathbf{z}_2^{(1)}$ and $\mathbf{z}_2^{(2)}$: the latter two are used separately since orthogonality must be maintained for all p. □

Proof of Theorem 4. By Theorem 3, identifiable parameters satisfy three independent linear equations. The theorem lists a maximal set of 6 linearly independent parameters. □

Estimating relatedness from independent sites

By (1), the average genotype frequencies are

$$\begin{aligned}
f_{0000} = \frac{\mathbb{E}n_{0000}}{n} = \quad & (1-\mu)\Delta_1 \\
& +(1-2\mu+\mu_2)\Big(\Delta_2+\Delta_3+\Delta_5+\Delta_7\Big) \\
& +(1-3\mu+3\mu_2-\mu_3)\Big(\Delta_8+\Delta_4+\Delta_6\Big) \\
& +(1-4\mu+6\mu_2-4\mu_3+\mu_4)\Delta_9 \\
f_{1111} = \frac{\mathbb{E}n_{1111}}{n} &= \mu\Delta_1+\mu_2\Big(\Delta_2+\Delta_3+\Delta_5+\Delta_7\Big)+\mu_3\Big(\Delta_4+\Delta_6+\Delta_8\Big)+\mu_4\Delta_9 \\
f_{1101} = \frac{\mathbb{E}n_{1101}}{n} &= (\mu-\mu_2)\Delta_3+(\mu_2-\mu_3)(\Delta_8+2\Delta_4)+2(\mu_3-\mu_4)\Delta_9 \\
f_{0111} = \frac{\mathbb{E}n_{0111}}{n} &= (\mu-\mu_2)\Delta_5+(\mu_2-\mu_3)(\Delta_8+2\Delta_6)+2(\mu_3-\mu_4)\Delta_9 \\
f_{0101} = \frac{\mathbb{E}n_{0101}}{n} &= 2(\mu-\mu_2)\Delta_7+(\mu-2\mu_2+\mu_3)\Delta_8+4(\mu_2-2\mu_3+\mu_4)\Delta_9 \\
f_{1100} = \frac{\mathbb{E}n_{1100}}{n} &= (\mu-\mu_2)\Delta_2+(\mu-2\mu_2+\mu_3)\Delta_4+(\mu_2-\mu_3)\Delta_6+(\mu_2-2\mu_3+\mu_4)\Delta_9 \\
f_{0011} = \frac{\mathbb{E}n_{0011}}{n} &= (\mu-\mu_2)\Delta_2+(\mu-2\mu_2+\mu_3)\Delta_6+(\mu_2-\mu_3)\Delta_4+(\mu_2-2\mu_3+\mu_4)\Delta_9 \\
f_{0100} = \frac{\mathbb{E}n_{0100}}{n} &= (\mu-\mu_2)\Delta_5+(\mu-2\mu_2+\mu_3)(\Delta_8+2\Delta_6)+2(\mu-3\mu_2+3\mu_3-\mu_4)\Delta_9 \\
f_{0001} = \frac{\mathbb{E}n_{0001}}{n} &= (\mu-\mu_2)\Delta_3+(\mu-2\mu_2+\mu_3)(\Delta_8+2\Delta_4)+2(\mu-3\mu_2+3\mu_3-\mu_4)\Delta_9,
\end{aligned} \tag{17}$$

where $\Delta_i : i = 1, \ldots, 9$ are the common identity coefficients across the loci.

Data sets

1000 Genomes: Data on human genome variants was downloaded from the May 2011 release of SNP calls along chromosome 12 in the 1000 Genomes project The 1000 Genomes Project Consortium (2010) at `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/`. Inferred ("cryptic") blood relationships were annotated alongside the data. We selected the following samples to study the behavior of estimators.

Population	N	**Samples**
ASW (African ancestry in Southwest US)	7	NA19713, NA19818, NA19819, NA19982, NA19985, NA20359, NA20363
CEU (Utah residents with Northern and Western European ancestry)	5	NA06984, NA06989, NA12340, NA12341, NA12342
CHS (Han Chinese South)	14	HG00404, HG00406, HG00407, HG00418, HG00419, HG00427, HG00500, HG00512, HG00524, HG00656, HG00657, HG00671, HG00672, HG00702
LWK (Luhya in Webuye, Kenya)	13	NA19312, NA19313, NA19331, NA19334, NA19350, NA19351, NA19380, NA19381, NA19382, NA19384, NA19385, NA19390, NA19391
MXL (Mexican ancestry in Los Angeles)	6	NA19660, NA19661, NA19663, NA19664, NA19684, NA19685
TSI (Tuscans from Italy)	9	NA20502, NA20503, NA20504, NA20505, NA20506, NA20507, NA20508, NA20509, NA20510

Horses: Horse SNP data of Petersen *et al.* (2013b) was downloaded from `http://www.animalgenome.org/repository/pub/UMN2012.1130/`. The data set consists of 733 horse genomes belonging to 32 breeds: Akhal Teke (AH), Andalusian (AND), Arabian (ARR), Belgian (BEL), Caspian Pony (CS), Clydesdale (CL), Exmoor (EX), Fell Pony (FELL), Finnhorse (FINN), Franches-Montagnes (FM), French Trotter (FT), Hanoverian (HAN), Icelandic (ICE), Mangalarga Paulista (MNGP), Miniature (MINI), Mongolian (MON), Morgan

(MOR), New Forest Pony (NF), North Swedish Horse (NSWE), Norwegian Fjord (NORF), Paint (PT), Percheron (PR), Peruvian Paso (PERU), Ouerto Rican Paso Fino (RP), Quarter Horse (QH), Saddlebred (SB), Shetland (SHET), Shire (SH), Standardbred (STBD), Swiss Warmblood (SZWB), Thoroughbred (TB) and Tuva (Tu).

Simulated data: The following procedure was used to generate genotype data for simulated inference. In one simulation run, parametrized by the number of desired loci n, sample size for MAF estimation N, a minimum allele frequency $p_0 = 1/N$ and a set of identity coefficients Δ_i, n loci with MAF $> p_0$ were selected uniformly along chromosome 12 from the 1000 Genomes project. At each selected locus, a "true" minor allele frequency p was set by perturbing the annotated MAF value (adding a uniformly distributed offset $-p_0 \leq \delta \leq p_0$ to the annotated value). The estimated MAF $\hat{p}$ for the locus was computed assuming $2N$ independent alleles, i.e., $\hat{p} = \frac{B_{2N,p}}{2N}$ where $B_{2N,p}$ is a binomial random variable with parameters $2N$ and p. Only those loci with $0 < \hat{p} < \frac{1}{2}$ were retained for use in the relatedness formulas. After computing $\hat{p}$ at a locus, joint genotypes were generated for two individuals with the given identity coefficients and the "true" MAF p. The simulation procedure was repeated independently for every n, N, and relation 10 times (see Figure 3).

Joint parentage for royal cousins: We traced back the well-documented ancestry (see Supporting Information for the annotated family tree) of Queen Victoria and Prince-Consort Albert to founders within 200 years, going back up to 7 generations using `peerage.com` as well as the English and German editions of Wikipedia. Identity coefficients were computed from the joint pedigree using the program `idcoefs` of Abney (2009).

DISCUSSION

Identity coefficients (Harris, 1964; Jacquard, 1974) encapsulate the dependencies between the alleles of two diploid individuals that determine the joint genotype distribution. If the individuals are not inbred, only three of the coefficients may be positive ($\Delta_7, \Delta_8, \Delta_9$), corresponding to Cotterman's k-gene coefficients (Thompson, 1975) for the individuals sharing $k = 0$, 1, or 2 alleles between them. The three coefficients can be retrieved from sampled genotypes using well-established methods relying on likelihood maximization (Thompson, 1975; Milligan, 2003) or allele frequency moments (Ritland, 1996; Lynch and Ritland, 1999).

In general, it may be of interest to estimate all nine identity coefficients simultaneously. In particular, all nine IBD modes may occur if the individuals have inbred coancestries (Harris, 1964), or come from a structured population (Wang, 2011). Biallelic genotypes, however, do not convey enough information about the generic IBD structure, since different identity coefficients can generate the same joint genotype distribution. Theorem 2 scrutinizes the inherent ambiguity about the identity coefficients, describing the linear subspace in which all solutions are found. One particular source of the ambiguity (corresponding to the null vector $\mathbf{z}_1$ in (15)) is that symmetric mixtures of simultaneous inbreeding and coancestry (modes Δ_7-Δ_4-Δ_6 vs. Δ_2-Δ_8-Δ_8) have identical effects in the genotype distribution. Importantly, these equivalent solutions varying only the ξ coordinate remain equivalent for any minor-allele frequency. The uncertainties about the identity coefficients are within the same magnitude as the inbreeding levels when both individuals are inbred ($\Delta_4, \Delta_6 > 0$) and also share ancestors ($\Delta_8 > 0$). Indeed, the real-life example of Figure 2 shows that the subtle details of coancestry can be irretrievable from the genotype distribution.

Consistent estimation is thus impossible since even as the number of independent sampled loci n goes to infinity and genotype frequencies concentrate around their true probabilities,

the identity coefficients stay ambiguous regardless of the estimation method used. The decomposition of the solution space (Theorem 3) shows the aspects of the IBD structure that can instead be inferred from biallelic genotypes. Specifically, Theorem 4 lists five non-trivial relatedness parameters, deconvolving the IBD structure to the maximum degree that is attainable. Principal aspects of genetic relatedness, quantified by the coefficients of kinship and inbreeding, are identifiable. Other identifiable attributes are the probabilities for three-allele joint IBD and the asymmetry of inbreeding modes with and without simultaneous coancestry. In contrast, parameters that do not weigh the identity coefficients properly (Eq. (7)) are not identifiable from the biallelic genotypes. Ill-defined relatedness parameters include the probabilities of separate identity modes (e.g., the probability Δ_1 of fourwise IBD), the fraternity coefficient ($\Delta_1 + \Delta_7$) and other generalizations of Cotterman's k-gene coefficients.

If allele frequencies are known in advance, relatedness parameters can be readily inferred from the genotype distribution using simple linear estimators (Eq. (14)). The formulas can even accommodate small biases of the MAF moment estimation due to sparse population sampling (Figure 3). More problematic is the discovery of genetic relatedness in structured populations (Anderson and Weir, 2007; Astle and Balding, 2009; Wang, 2011). For instance, if there is a clear subpopulation structure, the model's assumption that non-IBD alleles are identically distributed does not hold. The imposition of a common MAF distribution thus entails a different bias specific to each subpopulation (Figures 4 and 5–6).

Our non-identifiability results assumes the classic model of independent loci. In practice, genetic linkage can be powerfully exploited to infer pairwise relatedness. Identity modes change along a chromosome due to ancestral crossover events, partitioning the genomes into *IBD segments* Browning and Browning (2012) formed by consecutive loci sharing the same combined inheritance history. The segmentation can be explored with mapped markers that are spaced densely enough to display linkage. IBD modes along a chromosome

are conveniently captured by states of a hidden Markov model (Thompson, 2008), which can simultaneously incorporate pedigrees (Kyriazopoulou-Panagiotopoulou *et al.*, 2011) and linkage disequilibrium (Han and Abney, 2011) in its state transition rates.

In the case of human genomes, the limits of inferring relatedness are dominantly determined by linkage and finite genome size, and not identifiability (Skare *et al.*, 2009). The mean length of a segment with the same particular history involving m meioses decreases linearly with m. In human whole genome sequences, IBD segments of length 0.4 cM can be demarcated (Su *et al.*, 2012) with confidence by high-coverage sequencing. The detection of shared ancestry is thus constrained by the fact that descendants inherit a common ancestor's allele simultaneously with exponentially small probability in the number of meioses separating them (2^{-m+1}). As Browning and Browning (2012) point out, fifth cousins ($m = 12$) simultaneously inherit 1/2048 of their genome on expectation from the shared great-great-great-great grandfather or great-great-great-great grandmother each, which amounts to about 1.5 cM in an entire human genome, while the average IBD segment length is 8.3 cM. Then, by Markov's inequality, there is at least one IBD segment between the two cousins' genomes with probability at most $\frac{2\times 1.5}{8.3} = 0.35\ldots$. Identity modes with more than two IBD alleles ($\Delta_1, \Delta_2, \Delta_3, \Delta_5, \Delta_7$) usually involve even more distant ancestries and are, thus, almost certainly undetectable (Thompson, 2008). For example, if both individuals are children of fourth cousins (as the royal cousins here), the simultaneous inbreeding mode Δ_2 appears in segments of average length 4.2 cM, covering $1/2^{20}$ of their genome (about 0.003 cM); so, no such segments are seen in at least 99.93% of the cases.

The ambiguity of identity coefficients (outlined by Theorem 2) complements well-known results on equivalent pedigrees (Donnelly, 1983; Skare *et al.*, 2009). In the absence of linkage information, when, for instance, background relatedness is investigated in an inbred population based on a few genotyped loci in many individuals (Anderson and Weir, 2007; Wang, 2011), the non-identifiable mode combinations represent the theoretical limits

of dissecting the IBD structure. By our results , only two more distribution parameters can be inferred in addition to the usual two-gene coefficients for coancestry and inbreeding: one for three-gene IBD, and another measuring asymmetry in inbreeding modes.

LITERATURE CITED

Abney, M., 2009 A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. Bioinformatics **25**: 1561–1563.

Anderson, A. D. and B. S. Weir, 2007 A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics **176**: 421–440.

Astle, W. and D. J. Balding, 2009 Population structure and cryptic relatedness in genetic association studies. Statist. Sci. **4**: 451–471.

Browning, S. R. and B. L. Browning, 2012 Identity by descent between distant relatives: Detection and applications. Annu. Rev. Genet. **46**: 617–633.

Cockerham, C. C., 1971 Higher order probability functions of identity of alleles by descent. Genetics **69**: 235–246.

Donnelly, K. P., 1983 The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. **23**: 34–63.

Han, L. and M. Abney, 2011 Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. **35**: 557–567.

Harris, D. L., 1964 Genotypic covariances between inbred relatives. Genetics **50**: 1319–1348.

Jacquard, A., 1974 *Genetics of Human Populations*. Springer, New York.

Kyriazopoulou-Panagiotopoulou, S., D. K. Haghighi, S. J. Aerni, A. Sundquist, S. Bercovici, and S. Batzoglou, 2011 Reconstruction of genealogical relationships with applications to Phase iii of HapMap. Bioinformatics **27**: i333–i341.

Lange, K., 1997 *Mathematical and Statistical Methods for Genetic Analysis*. Springer, New York.

Lynch, M. and K. Ritland, 1999 Estimation of pairwise relatedness with molecular markers. Genetics **152**: 1753–1766.

Malécot, G., 1969 *The Mathematics of Relationship*. W. H. Freeman, San Francisco.

Milligan, B. G., 2003 Maximum-likelihood estimation of relatedness. Genetics **163**: 1153–1167.

Petersen, J. L., J. R. Mickelson, E. G. Cothran, L. S. Andersson, J. Axelsson, *et al.*, 2013a Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLoS ONE **8**: e54997.

Petersen, J. L., J. R. Mickelson, A. K. Rendahl, S. J. Valberg, L. S. Andersson, *et al.*, 2013b Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. **9**: e1003211.

Ritland, K., 1996 Estimators for pairwise relatedness and individual inbreeding coefficients. Genet. Res. (Camb.) **67**: 175–185.

Skare, Ø., N. Sheehan, and T. Egeland, 2009 identification of distant family relationships. Bioinformatics **25**: 2376–2382.

Su, S.-Y., J. Kasberger, S. Baranzini, W. Byerley, W. Liao, *et al.*, 2012 Detection of identity by descent using next-generation whole genome sequencing data. BMC Bioinformatics **13**: 121.

The 1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. Nature **467**: 1061–1073.

Thompson, E. A., 1975 The estimation of pairwise relationships. Ann. Hum. Genet. **39**: 173–188.

Thompson, E. A., 2008 The IBD process along four chromosomes. Theor. Popul. Biol. **73**: 369–373.

Wang, J., 2011 Unbiased relatedness estimation in structured populations. Genetics **187**: 887–901.

Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. **7**: 771–780.

LIST OF FIGURES

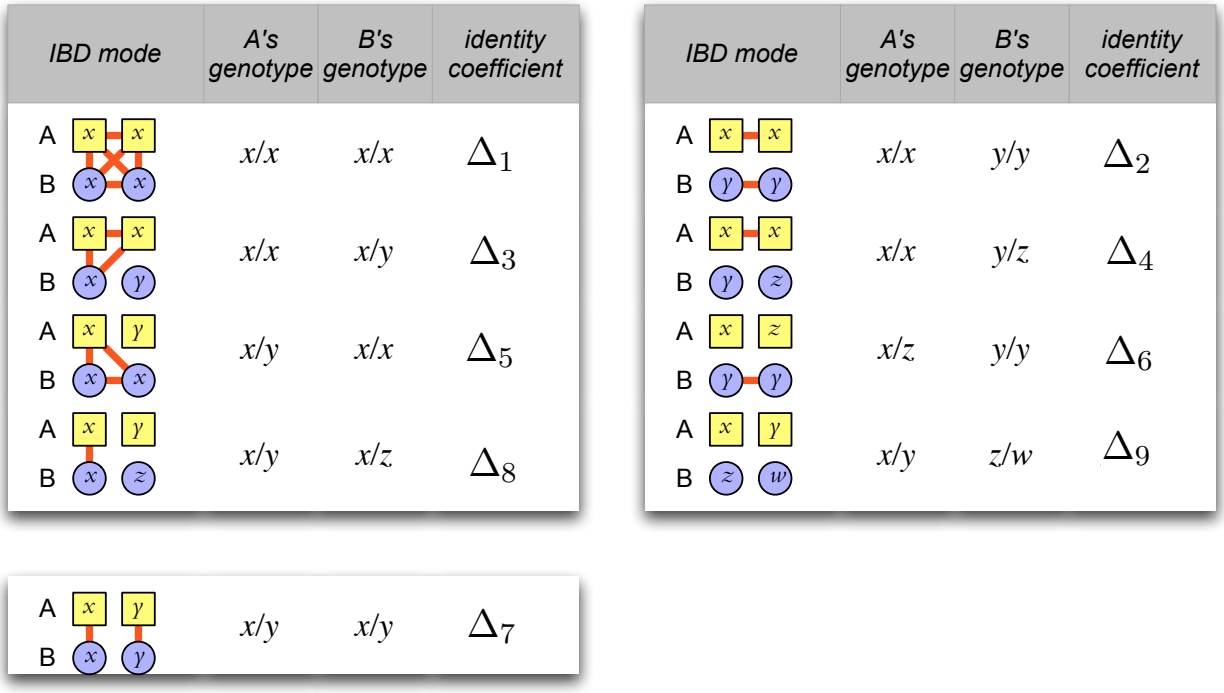
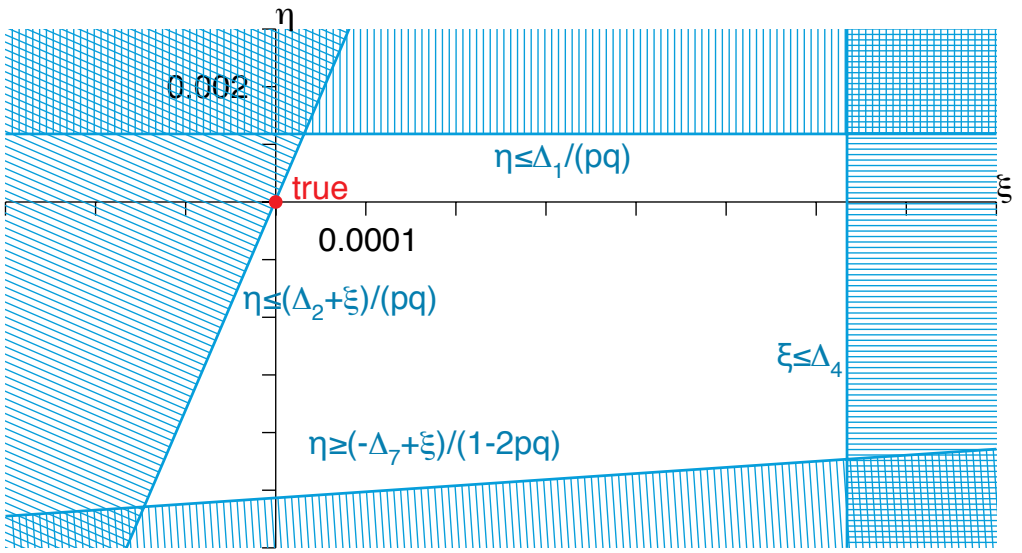


Figure 1. Identity modes for a diploid genotype pair. Identity by descent is marked by thick red lines. Alleles $x, y, \ldots$ observed in the genotypes may be equal. Probabilities for different modes are denoted by the identity coefficients Δ_i.



Identity coefficient	$\times 2^{-20}$	**decimal value**	**ambiguity range**	
			min	max
Δ_1	34	$3.24 \cdot 10^{-5}$	0	$1.78 \cdot 10^{-4}$
Δ_2	1	$9.54 \cdot 10^{-7}$	0	$7.58 \cdot 10^{-4}$
Δ_3	324	$3.09 \cdot 10^{-4}$	$1.85 \cdot 10^{-5}$	$3.74 \cdot 10^{-4}$
Δ_4	665	$6.34 \cdot 10^{-4}$	0	$7.80 \cdot 10^{-4}$
Δ_5	3140	$2.99 \cdot 10^{-3}$	$2.70 \cdot 10^{-3}$	$3.06 \cdot 10^{-3}$
Δ_6	9113	$8.69 \cdot 10^{-3}$	$8.06 \cdot 10^{-3}$	$8.84 \cdot 10^{-4}$
Δ_7	5087	$4.85 \cdot 10^{-3}$	0	$5.94 \cdot 10^{-3}$
Δ_8	278698	0.266	0.263	0.276
Δ_9	751514	0.717	0.711	0.718
Relatedness parameter				
θ_1 (kinship)	73984	0.071		
θ_{2A} (Victoria's inbreeding)	1024	$9.77 \cdot 10^{-3}$		
θ_{2B} (Albert's inbreeding)	12288	0.0117		
θ_3 (triple coancestry)	152824	0.146		
$\theta_{3:3}$ (inbred ancestry)	1766	$1.68 \cdot 10^{-3}$		
θ_4 (independent inbreeding difference)	-8448	$-8.06 \cdot 10^{-3}$		

Figure 2. Null space identity coefficients. The intersection of the constraints from Equation (16) defines the convex polygonal area within which (ξ, η) values plugged into Eq. (5) yield valid identity mode distributions that generate the same genotypic distribution. The example is based on the joint parentage of Queen Victoria and her spouse Prince Albert, for which the relevant parameters are listed below the plot. The illustration assumes $pq = 0.02746\ldots$ reflecting typical allele frequency moments in humans (from the 1000 Genomes project). Extremal values of possible Δ_i listed under "ambiguity range" are attained in the corners of the shaded area.

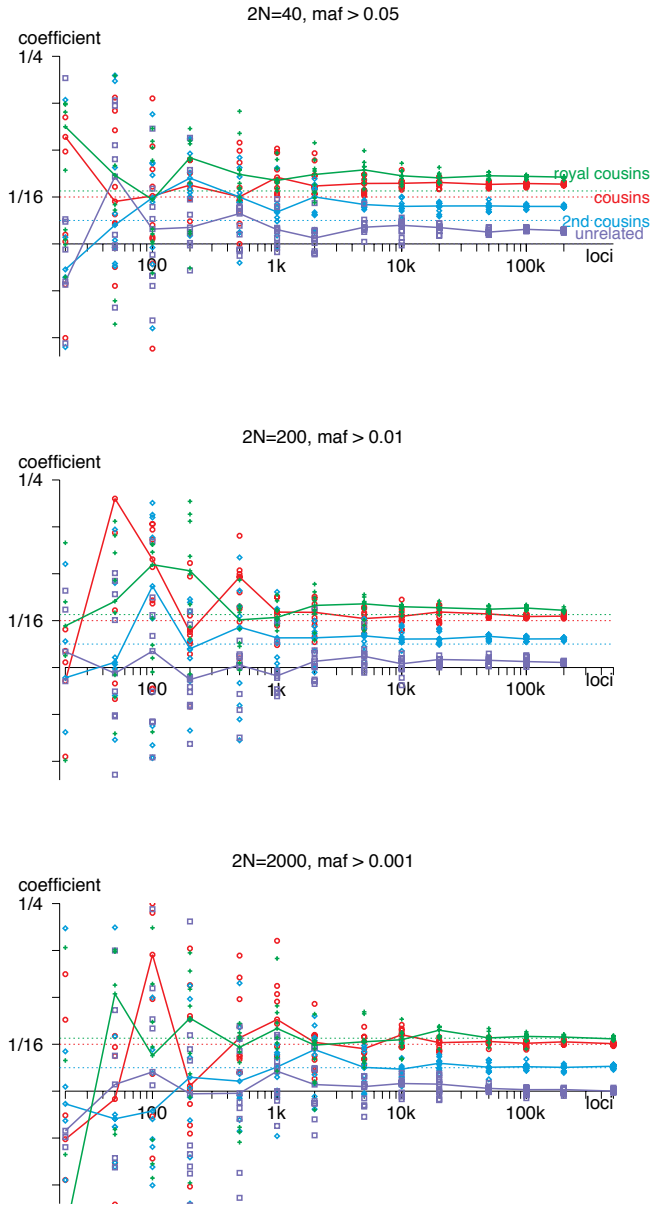


Figure 3. Estimated kinship coefficients in simulations. The Y axis shows the estimated kinship coefficient $\hat{\theta}_1$ of Eq. (14) from random genotypes for unrelated individuals ($\theta_1 = 0$), or simulated along simple pedigrees of first ($\theta_1 = \frac{1}{16}$) and second cousins ($\theta_1 = \frac{1}{32}$), as well as a complex pedigree ("royal cousins" with $\theta_1 = \frac{289}{4096} = \frac{1}{16} + \frac{33}{4096}$). Every point plots $\hat{\theta}_1$ for simulated genotype data across n loci (along the X axis), with random MAF values (based on randomly picked sites from the 1000 Genomes project). For every data set of random loci, empirical minor allele frequencies are calculated assuming N genotyped diploid individuals. The formulas $\hat{\theta}_1$ are employed with MAF moments estimated from these empirical frequencies. The procedure is repeated 10 times for every n and N; the solid lines connect the medians of the ten replicates.

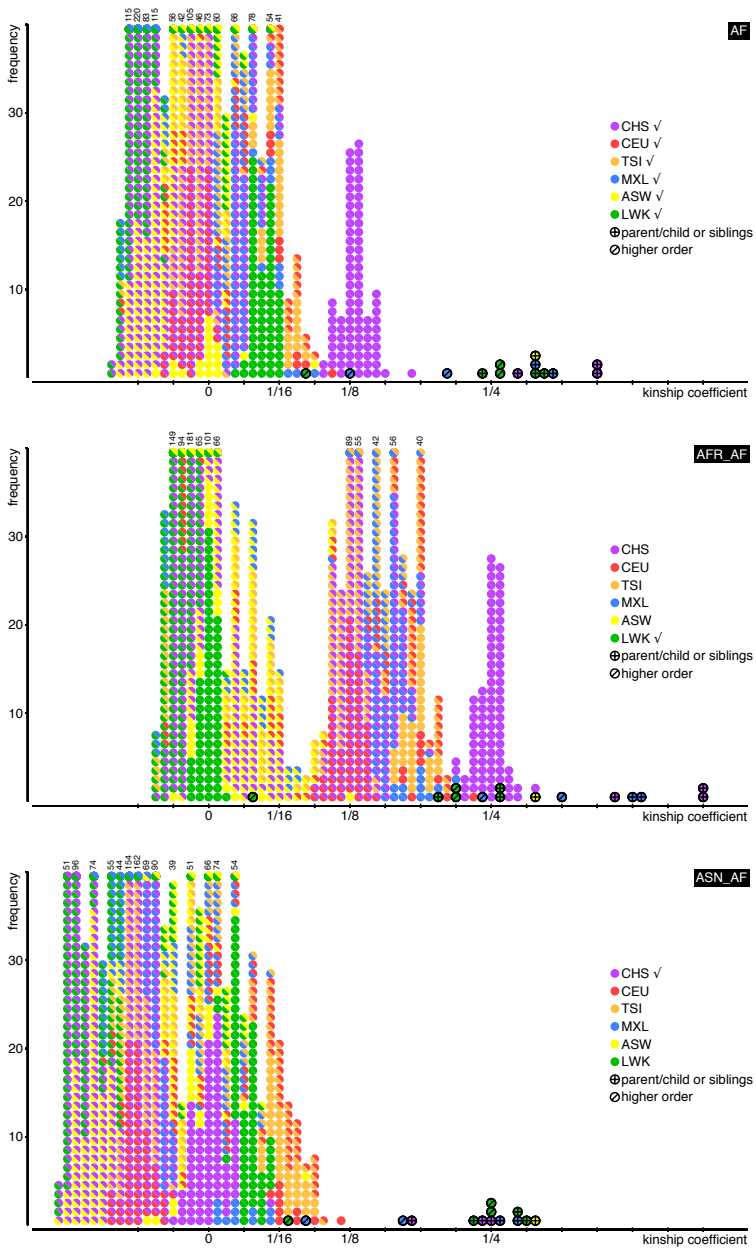


Figure 4. Distribution of estimated kinship coefficients in pairs between 54 samples from the 1000 Genomes project. The three histograms illustrate the distribution of the coefficients when estimated using the allele frequency moments of all samples (AF), African samples (AFR_AF) and Asian samples (ASN_AF). On every plot, estimated pairwise coefficients are binned by rounding to the nearest multiple of $\frac{1}{128}$. Bins are shown along the X axis, and bin sizes (how many pairs have $\hat{\theta}_1$ falling into the bin) plotted along the Y axis. Numbers above the truncated bars show the true bin size. Every disk corresponds to a pairwise coefficient, colored by the subpopulations the two individuals belong to. Marked disks denote known relationships. Checkmarks indicate that a subpopulation was used to calculate allele frequencies.

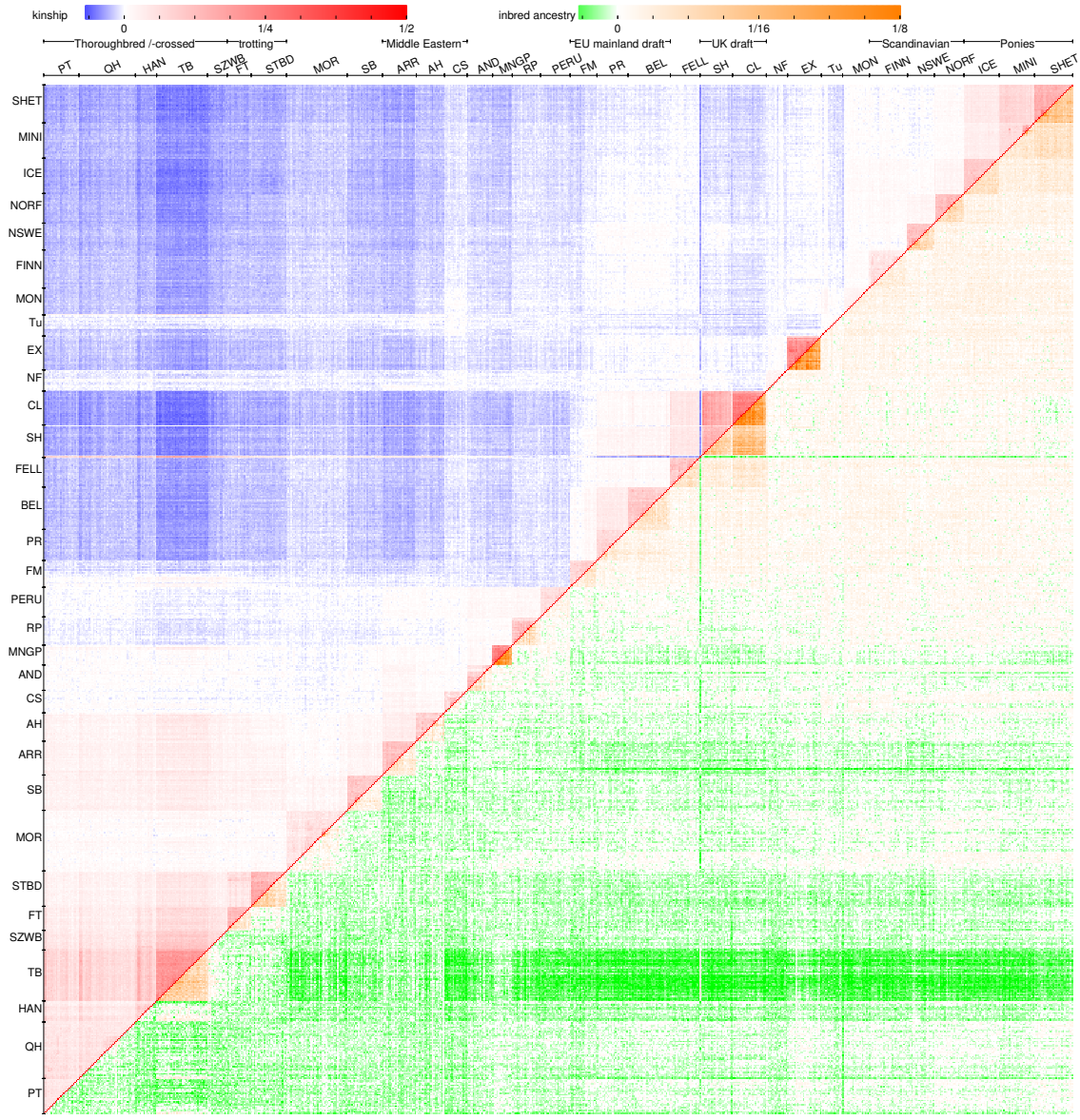


Figure 5. Estimated kinship and inbred ancestry coefficients between 733 horses. Rows and columns correspond to individual horses in the same order from left to right and bottom to top. Horses belong to one of 32 breeds (see **Methods** for the breed codes). Cells are colored by kinship coefficients above the diagonal and by inbred ancestry below, following the color scale shown above the heatmap. MAF moments were estimated from combining all the breeds in the data set.

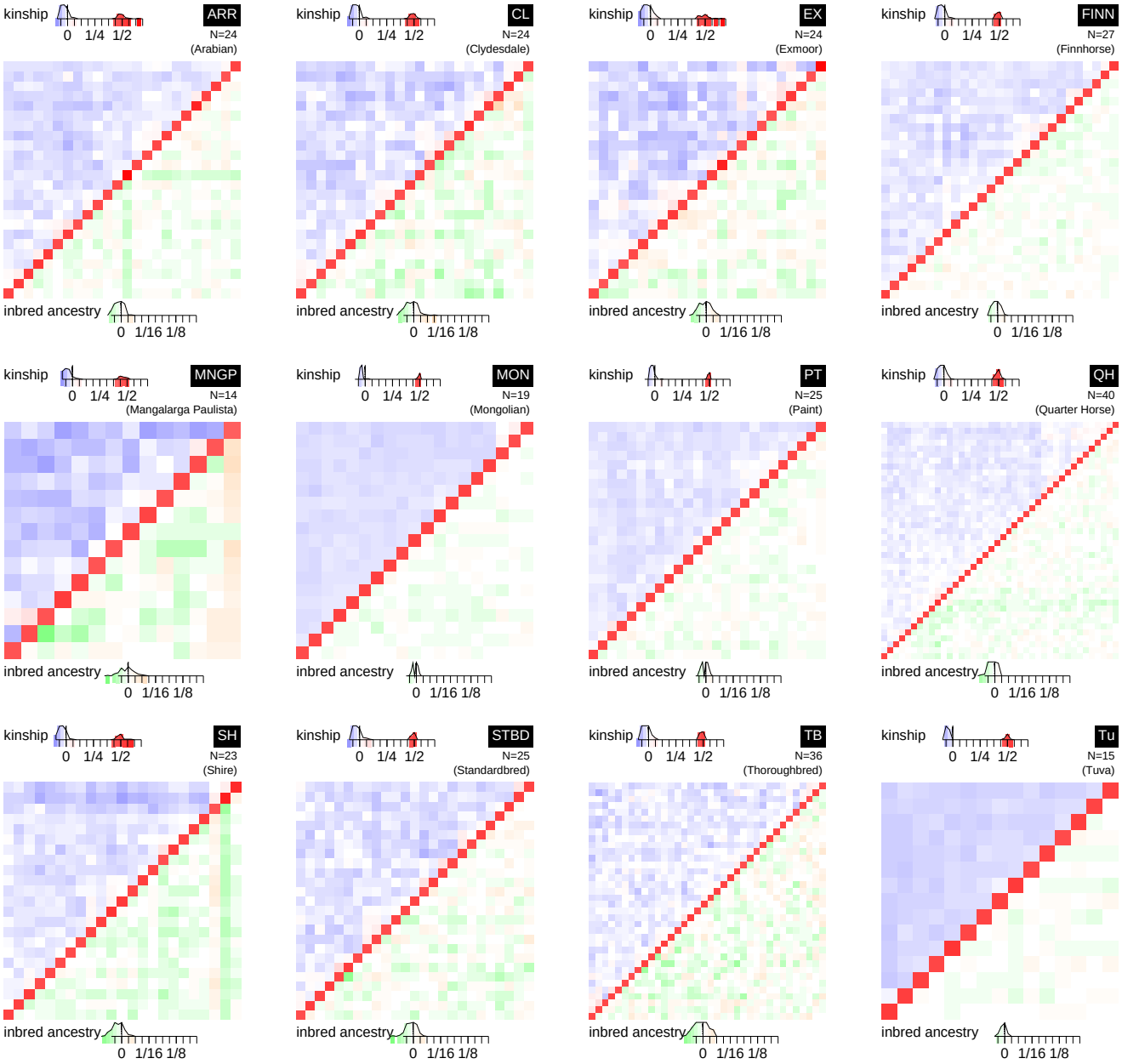


Figure 6. Kinship and inbred ancestry coefficients within selected horse breeds. Cells are colored by inferred relatedness parameters as in Figure 5, but MAF moments are estimated separately for each breed here. Cells along the diagonal show the kinship coefficient of each horse with itself ($= 1/2$ if not inbred). The histograms above and below the heatmaps plot the distributions for the relatedness parameters. (The histograms bin $\hat{\theta}_1$ and $\hat{\theta}_{3:3}$ by rounding to the nearest multiple of $\frac{1}{32}$ and $\frac{1}{128}$, respectively. The histograms' Y axis uses the transformation $y = 1 - e^{-m}$ to project the bin size m onto a unit interval.)

LIST OF TABLES

Table 1. Distribution of biallelic genotypes by identity mode

Mode	0/0:0/0	1/1:1/1	1/1:0/1	0/1:1/1	0/1:0/1	1/1:0/0	0/0:1/1	0/1:0/0	0/0:0/1
1	q	p	0	0	0	0	0	0	0
2	q^2	p^2	0	0	0	pq	pq	0	0
3	q^2	p^2	pq	0	0	0	0	0	pq
4	q^3	p^3	$2p^2q$	0	0	pq^2	p^2q	0	$2pq^2$
5	q^2	p^2	0	pq	0	0	0	pq	0
6	q^3	p^3	0	$2p^2q$	0	p^2q	pq^2	$2pq^2$	0
7	q^2	p^2	0	0	$2pq$	0	0	0	0
8	q^3	p^3	p^2q	p^2q	$pq(p+q)$	0	0	pq^2	pq^2
9	q^4	p^4	$2p^3q$	$2p^3q$	$4p^2q^2$	p^2q^2	p^2q^2	$2pq^3$	$2pq^3$

Genotypic probabilities for every mode are given by assuming that alleles are chosen independently for each IBD group, with probability p for allele 1 (minor allele) and with probability q for allele 0 (major allele). Genotypes are unordered (1/0 and 0/1 are considered equivalent).