BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Two molecular measures of relatedness based on haplotype sharing

David Edwards

Abstract

Background: Measuring the extent of shared ancestry between individuals or organisms is important in many fields, including forensic science, conservation genetics and animal breeding. The traditional approach is to calculate the expected degree of relatedness between individuals in a pedigree. This assumes that the founders of the pedigree are non-inbred and unrelated to each other, which is rarely the case. In contrast, molecular data allow measurement of actual relatedness without knowledge of a pedigree. Methods to do this have been proposed, but generally do not take the lengths of the genomic regions shared between individuals into account.

Results: Two measures based on the extent of haplotype sharing between genomes are proposed. The intercept measure *B* estimates the fraction of shared genome between individuals, and the product measure *C* is closely related to the numerator relationship matrix *A*. Both are based on a model for the joint distribution of markers at the haplotype level. The two measures are compared to the pedigree-based measure *A* and to vanRaden's *G*, a frequently used molecular measure, using a set of data comprising 5037 dairy cattle. The comparison criteria include the ability to capture genealogical relatedness and the prediction accuracy obtained when used in genomic prediction. Both *B* and *C* explain around 95 % of the variation in *A*, whereas *G* explains around 6 %. *G* captures genealogical relatedness poorly, particularly for distantly related individuals (second cousins or farther). Both *B* and *C* tend to be larger than *A* but this can be ascribed to the assumption of non-inbred unrelated founders. Using *C* in linear mixed models results in slightly higher prediction accuracy than *G*, and using *B* results in slightly lower prediction accuracy.

Conclusions: The two proposed measures of relatedness capture genealogical relatedness well, outperforming vanRaden's *G* in this respect. When used in genomic prediction models, the product measure leads to slightly improved prediction accuracy.

Keywords: Genomic relationship matrix, Multiset, Acyclic probabilistic finite automata, Haplotype sharing

Background

Estimating the extent of shared ancestry between individuals or organisms is central to many fields. Examples range from forensic science [1], studies of population structure [2, 3], and conservation genetics [4], to the mixed linear models used in genomic prediction and genome-wide association studies [5, 6].

Following Malêcot [7], measures of relatedness between two individuals are generally formulated in terms of the coefficient of coancestry, which is the probability that for a randomly selected gene, two alleles, one taken at random from each individual, descend from a single ancestral gene — that is, the probability that the alleles are identical-by-descent (IBD). Similarly, a measure of inbreeding for an individual is defined as the probability that the two alleles of a randomly selected gene are IBD [7]. When a pedigree is available, probabilities that two alleles are IBD from common ancestors within the pedigree can be calculated, and from these the classical measures of relatedness and inbreeding can be derived. The calculations typically assume that the founders are unrelated and not inbred, which is rarely the case. The measures depend on the choice of pedigree, and represent expected rather than realized relatedness and inbreeding, since they cannot incorporate the randomness inherent in meiosis.

Correspondence: david.edwards@mbg.au.dk
Centre for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Blichers Allé 20, DK, 8830 Tjele, Denmark

With the advent of high-scale genotyping technologies such as single nucleotide polymorphism (SNP) arrays, it is possible to estimate realized relatedness directly from molecular data without knowledge of genealogy, and a variety of ways to do this are available [8, 9]. These generally take the form of genome-wise averages of single-SNP statistics, which have the disadvantage of not taking the lengths of genomic regions shared between two individuals into account [8]. In Section "Methods," two novel measures based on the extent of haplotype sharing are described, and their properties studied. In Section "Results" the methods are applied to a set of data from dairy cattle, and compared to the classical pedigree-based measure and a measure due to vanRaden [10] that is widely used in animal breeding. Section "Software" describes the software used and Section "Discussion" gives a brief discussion.

Methods

The methods developed in this paper are based on the following conceptual framework. The genome is divided into a series of physical intervals, and the variant DNA strings that may occur in the intervals are denoted *segments*. To each interval corresponds a set of segments that may occur in the interval, and these sets do not overlap: a segment that may occur in one interval may not occur in another. With some abuse of terminology one may identify the intervals with genes, and the possible segments with alleles. A specific genome is regarded as a collection of segments, and measures of relatedness between genomes are constructed in terms of similarity between such collections. For this the concept of a multiset is needed.

Multisets

A multiset [11] is a generalization of the concept of a set that, unlike a set, allows multiple instances of its elements. The multiplicity of an element is the number of instances of the element in the multiset. For example, [2 figs, 5 pears, 3 plums] is a multiset in which the element fig has multiplicity two. Note the use of square brackets [ ] to distinguish multisets from ordinary sets using curly brackets {}. A multiset corresponds to an ordinary set if the multiplicity of every element is one or zero.

Multiset intersection is a generalization of set intersection. The intersection of two multisets is formed by taking the minima of the multiplicities of the corresponding elements in the two multisets. For example:

[2 figs, 5 pears, 3 plums] $\cap$ [1 fig, 10 pears, 0 plums] = [1 fig, 5 pears, 0 plums].

The sum and product multiset operators, represented by $+$ and $\times$ respectively, use the straightforward element-wise operations, for example:

[2 figs, 5 pears, 3 plums] $+$ [1 fig, 10 pears, 0 plums] = [3 figs, 15 pears, 3 plums], and

[2 figs, 5 pears, 3 plums] $\times$ [1 fig, 10 pears, 0 plums] = [2 figs, 50 pears, 0 plums].

The cardinality of (number of elements in) a multiset A is written $|A|$. Some useful properties of the operators include the equations

$$\begin{aligned} |A+B| &= |A|+|B|, \\ A\times[B+C] &= [A\times B]+[A\times C], \\ A+[B\cap C] &= [A+B]\cap[A+C], \end{aligned}$$

that hold for any multisets A, B and C [11].

Two measures of genomic relatedness

As described above a genome is taken to be composed of a collection of segments, taken from a larger pool of segments $\mathcal{S}$. Let G_i represent the genome of individual i, regarded as a collection of segments s in $\mathcal{S}$. Since there may be duplicates, G_i is a multiset. Let x_{is} represent the multiplicity of segment s in G_i.

Let there be p intervals (loci). For $s \in \mathcal{S}$, let $l(s) \in \{1,\ldots,p\}$ indicate the interval associated with segment s. At each interval, an individual genome has two segments, corresponding to its two haplotypes. Thus each genome has in all $2p$ segments.

A natural definition of the similarity of two individuals is the fraction of genome that they share. So for individuals i and j the intersect measure of their similarity is defined as the cardinality of the intersection of the two genomes, divided by the total number of segments:

$$\begin{aligned} b_{ij} &= |G_i \cap G_j|/2p \\ &= \sum_{s\in\mathcal{S}} x_{is} \wedge x_{js}/2p, \end{aligned} \tag{1}$$

where $x \wedge y$ is the minimum of x and y. Since $|G_i| = 2p$, we have $b_{ii} = 1$ for all i. For all i and j, $0 \leq b_{ij} \leq 1$. Also $b_{ij} = 0$ iff G_i and G_j have no common segments, and $b_{ij} = 1$ iff G_i and G_j are identical.

The product measure is defined similarly using the product operator:

$$\begin{aligned} c_{ij} &= |G_i \times G_j|/2p \\ &= \sum_{s\in\mathcal{S}} x_{is}x_{js}/2p. \end{aligned} \tag{2}$$

We have $0 \leq c_{ij} \leq 2$ for all i and j, with $c_{ij} = 0$ iff G_i and G_j have no common segments. In matrix terms $C = (c_{ij})$ can be written as $C = XX^T/2p$ where $X = (x_{is})$ is the $N \times |\mathcal{S}|$ matrix of multiplicities.

To relate the two measures, note that when x and y take values in $\{0, 1, 2\}$, xy is given by $\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 2 & 0 & 2 & 4 \end{array}$, and $x \wedge y$ by

$\begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 2 & 0 & 1 & 2 \end{array}$, so $xy = 2(x \wedge y) - I(x = y = 1)$, where I is the indicator function. It follows that

$$c_{ij} = 2b_{ij} - \kappa_{ij}, \tag{3}$$

where κ_{ij}, a measure of shared heterozygosity, is defined as

$$\kappa_{ij} = \#\{s \in \mathcal{S} : x_{is} = x_{js} = 1\}/2p.$$

When $x \in \{0, 1, 2\}$, $x^2 = x + 2I(x = 2)$, so

$$\begin{aligned} c_{ii} &= \sum_{s \in \mathcal{S}} x_{is}^2/2p \\ &= 1 + f_i \end{aligned}$$

where f_i, a measure of homozygosity of individual i, is

$$f_i = \#\{s \in \mathcal{S} : x_{is} = 2\}/p.$$

To relate C to the numerator relationship matrix [12], let θ_{ij} be the coefficient of coancestry between individuals i and j, that is, the probability that for a randomly selected gene, two alleles, one taken at random from each individual, are IBD [7]. Define ϑ_{ij} similarly as the probability that for a randomly selected gene, two alleles, one taken at random from each individual, are identical, that is, identical-by-state (IBS). For $k = 1, \ldots, p$, write the two segments of individual i at interval k as (s_{i1}^k, s_{i2}^k), and similarly (s_{j1}^k, s_{j2}^k) for individual j. Then for $i \neq j$,

$$\begin{aligned} \sum_{s \in \mathcal{S}: l(s) = k} x_{is} x_{js} = I\left(s_{i1}^k = s_{j1}^k\right) + I\left(s_{i1}^k = s_{j2}^k\right) + I\left(s_{i2}^k = s_{j1}^k\right) \\ + I\left(s_{i2}^k = s_{j2}^k\right), \end{aligned}$$

so from (2)

$$\begin{aligned} 2pc_{ij} = \sum_{k=1 \ldots p} I\left(s_{i1}^k = s_{j1}^k\right) + I\left(s_{i1}^k = s_{j2}^k\right) + I\left(s_{i2}^k = s_{j1}^k\right) \\ + I\left(s_{i2}^k = s_{j2}^k\right). \end{aligned} \tag{4}$$

Note that ϑ_{ij} is the probability of an event randomly chosen from the $4p$ identities on the right-hand side of Eq. (4). Hence $\vartheta_{ij} = 2pc_{ij}/4p$, and so $c_{ij} = 2\vartheta_{ij}$. Thus c_{ij} is twice the IBS-sense coefficient of coancestry ϑ_{ij}.

Similarly, let θ_i be the coefficient of inbreeding for individual i, that is, the probability that for a randomly selected gene, the two alleles are IBD [7], and let ϑ_i be the corresponding IBS-sense quantity. When $i = j$, we obtain

$$\begin{aligned} 2pc_{ii} = \sum_{k=1 \ldots p} I\left(s_{i1}^k = s_{i1}^k\right) + I\left(s_{i1}^k = s_{i2}^k\right) + I\left(s_{i2}^k = s_{i1}^k\right) \\ + I\left(s_{i2}^k = s_{i2}^k\right) \qquad (5) \\ = \sum_{k=1 \ldots p} 2I\left(s_{i1}^k = s_{i1}^k\right) + 2I\left(s_{i1}^k = s_{i2}^k\right) \qquad (6) \\ = \sum_{k=1 \ldots p} \left(2 + 2I\left(s_{i1}^k = s_{i2}^k\right)\right) \qquad (7) \end{aligned}$$

so $c_{ii} = 1 + \vartheta_i$, and f_i is the IBS-sense coefficient of inbreeding ϑ_i.

The additive, or numerator, relationship matrix is defined as $A = (a_{ij})$ where

$$a_{ij} = \begin{cases} 1 + \theta_i & \text{if i=j} \\ 2\theta_{ij} & \text{otherwise} \end{cases}$$

and as just shown

$$c_{ij} = \begin{cases} 1 + \vartheta_i & \text{if i=j} \\ 2\vartheta_{ij} & \text{otherwise} \end{cases}$$

hence C and A are conceptually closely related. An assumption behind this assertion is discussed in Section "Discussion".

Defining the segmentation

To define the segmentation a statistical model in the form of an acyclic probabilistic finite automaton (APFA) [13] is used. Such models allow the extent of haplotype sharing within and between genomes to be quantified, and underlie the Beagle program [14, 15] that is widely used for processing high-dimensional SNP data. Phase estimation, imputation and model selection are performed simultaneously, using the algorithm described in [15]. Beagle is highly efficient, taking only a few minutes to process each chromosome for the data described in Section "Data and computations", and performs well: for example, imputation accuracy rates generally exceed 97 % in cattle data [16].

An APFA is represented as a directed multigraph $\mathcal{A} = (V, E)$, where V is a vertex set and E an edge set: a small example is shown in Fig. 1. This is a model for the joint distribution of $p = 20$ markers at the haplotype level. To each edge in E is attached a probability, such that the sum of probabilities of the outgoing edges from each vertex in V is one. Each haplotype corresponds to a path through the graph from the root (the leftmost vertex) to the sink (the rightmost vertex). The probability of a haplotype is the product of the probabilities of the edges in its root-to-sink path. See further [17, 18].

In [14, 15] the haplotypes that traverse a given edge are known as a haplotype cluster. Here a different perspective is adopted. The edges of the APFA are taken to represent chromosomal segments, that is, we set $\mathcal{S} = E$. So if two haplotypes traverse the same edge in an interval they

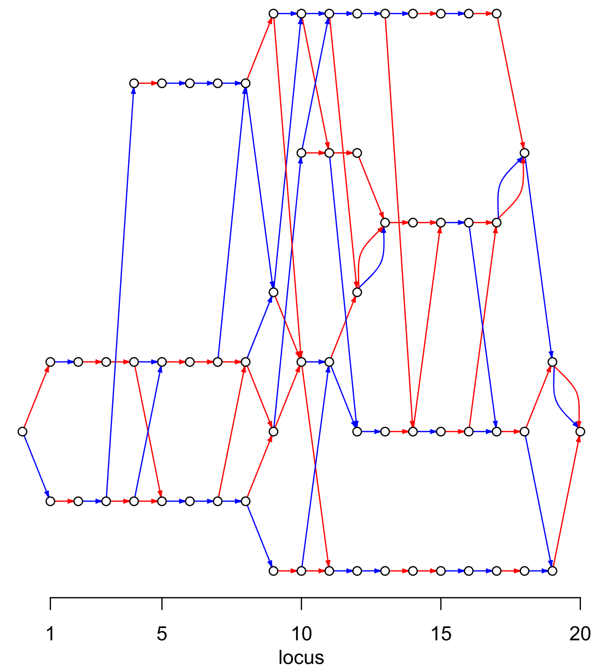

Fig. 1 A simple APFA. An APFA $\mathcal{A} = (V, E)$ for $p = 20$ markers. It has $|V| = 65$ vertices and $|E| = 91$ edges. The colour of the edges denotes the allele associated with the edge (red is "0" and blue is "1"). To each edge is attached a probability (not shown here)

are taken to share the same DNA in that interval, and if they traverse different edges, they are taken not to share DNA in that interval. The data may be represented as an $N \times |E|$ matrix X taking values in $\{0, 1, 2\}$, whose (i, j)th element specifies the multiplicity of segment j in individual i. The variables corresponding to the columns of X are called haplomarkers, and X is called the haplomarker design matrix.

Figure 2 illustrates recombination under the model of Fig. 1. The two haplotypes of an individual correspond to two root-to-sink paths in the graph, and recombination is seen as crossing-over between the paths. If the individuals represented in Fig. 2 are taken in the order mother, father and offspring, we find the relationship matrices B and C to be $B = \begin{pmatrix} 1.000 & 0.750 & 0.650 \\ 0.750 & 1.000 & 0.675 \\ 0.650 & 0.675 & 1.000 \end{pmatrix}$ and $C = \begin{pmatrix} 1.000 & 0.750 & 0.875 \\ 0.750 & 1.000 & 0.900 \\ 0.875 & 0.900 & 1.450 \end{pmatrix}$. There are, for example, nine edges shared between the two haplotypes (red dashed lines) in the offspring genome, so the homozygosity of this individual is 0.45.

Expected relatedness

Consider first three individuals, i, j, and k, where i and j are the parents of k. We examine how the parent-offspring relatedness measures depend on the relatedness of the parents. Specifically, expressions for the expectations of the parent-offspring relatedness conditional on the parental relatedness will be derived.

During meiosis the genomes G_i and G_j are first partitioned into two gametes, say $G_i = [H_i^1 + H_i^2]$ and $G_j = [H_j^1 + H_j^2]$ such that $|H_i^1| = |H_i^2| = |H_j^1| = |H_j^2| = p$. The partitioning process (segregation) is complex and stochastic, but only properties that are invariant to this are considered here. Then the genome $G_k = [H_i^* + H_j^*]$ is formed where H_i^* is either H_i^1 or H_i^2, and H_j^* is either H_j^1 or H_j^2, and the four combinations are equiprobable. Thus

$$\begin{aligned} 4E(|G_i \times G_k|) &= |G_i \times \left[H_i^1 + H_j^1\right]| + |G_i \times \left[H_i^1 + H_j^2\right]| \\ &\quad + |G_i \times \left[H_i^2 + H_j^1\right]| + |G_i \times \left[H_i^2 + H_j^2\right]| \\ &= |\left[G_i \times H_i^1\right] + \left[G_i \times H_j^1\right] + \left[G_i \times H_i^1\right] + \left[G_i \times H_j^2\right] \\ &\quad + \left[G_i \times H_i^2\right] + \left[G_i \times H_j^1\right] + \left[G_i \times H_i^2\right] + \left[G_i \times H_j^2\right]| \\ &= 2|G_i \times G_i| + 2|G_i \times G_j| \end{aligned}$$

so

$$E(c_{ik}) = (c_{ii} + c_{ij})/2. \tag{8}$$

Similarly

$$4E(G_k \times G_k) = 8p + 2|G_i \times G_j| \tag{9}$$

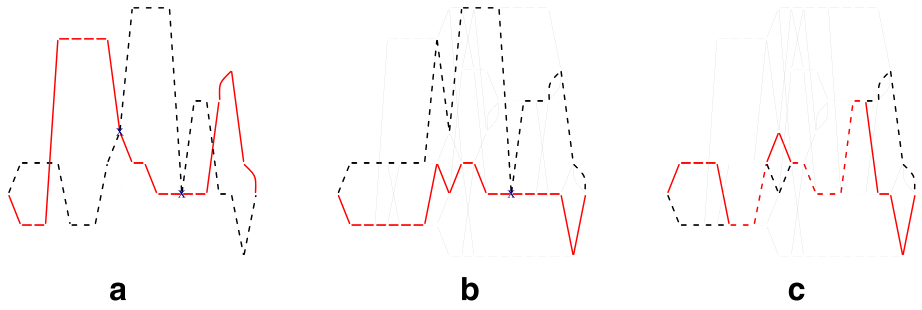

Fig. 2 Recombination in an APFA. Maternal (**a**), paternal (**b**) and offspring genomes (**c**) are shown. In each genome, the maternal haplotypes are shown as red solid lines, and the paternal as black dashed lines. The overlap is shown as red dashed lines. Two crossovers occur in the maternal line and one in the paternal line: these are marked with "x"s

so

$$E(c_{kk}) = 1 + c_{ij}/2 \tag{10}$$

and

$$E(f_k) = c_{ij}/2. \tag{11}$$

Now consider four individuals, h, i, j and k, where again i and j are the parents of k, and where the relatedness between h, i and j is known.

$$\begin{aligned}4E(|G_k \times G_h|) &= |[H_i^1 + H_j^1] \times G_h| + |[H_i^1 + H_j^2] \times G_h| \\ &\quad + |[H_i^2 + H_j^1] \times G_h| + |[H_i^2 + H_j^2] \times G_h| \\ &= 2|G_i \times G_h| + 2|G_j \times G_h|,\end{aligned}$$

so

$$E(c_{kh}) = (c_{ih} + c_{jh})/2. \tag{12}$$

An expression for the expectation of the intersect measure may be derived in a similar fashion:

$$E(b_{ik}) = E(b_{jk}) = 1/2 + c_{ij}/4. \tag{13}$$

but I have not been able to derive an expression for $E(b_{kh})$ corresponding to (12).

Expressions (8), (10) and (12) are identical to those used in the calculation of the numerator relationship matrix using the algorithm of [19]. When only a subset of individuals in a pedigree are genotyped, a hybrid expected/realized relationship matrix $R = (r_{ij})$ exploiting both pedigree and genomic information can be obtained using the following simple modification to the algorithm.

Order the individuals so that parents precede their offspring and label them $1, \ldots, N$. Write the subset of genotyped individuals as S, and for all $i, j \in S$, set $r_{ij} = c_{ij}$, the realized genomic relatedness described above. For $k = 1, \ldots N$, derive the relatedness between an individual k and the preceding individuals as follows. When $k \notin S$, set

$$r_{kk} = 1 + r_{ij}/2$$

where i and j are the parents of k. If either or both i and j are unknown (i.e., not in the pedigree) assume they are unrelated, that is, use $r_{ij} = 0$ in this calculation. For each $h \in \{1, \ldots, k-1\}$, if $\{h, k\} \not\subseteq S$, set

$$r_{hk} = r_{kh} = (r_{ih} + r_{jh})/2,$$

where again i and j are the parents of k. If i is unknown, use $r_{ih} = 0$, and if j is unknown, use $r_{jh} = 0$ in this calculation.

This algorithm adjusts the expected relationships downstream of S in the pedigree. An alternative method that adjusts all the relationships outside of S is sketched in the following subsection.

Modifying *A*

This subsection describes an established technique that is useful in various contexts. The individuals are partitioned into two groups, so that $A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right)$ and $A^{-1} = \left(\begin{array}{c|c} A^{11} & A^{12} \\ \hline A^{21} & A^{22} \end{array}\right)$. We regard A as a covariance matrix and wish to construct a new $\tilde{A}$ for which the marginal covariance of group one is set to A_{11}^*, and the conditional (co)variance of group two, given group one, is kept the same. That is to say, such that A^{12}, $A^{21} = (A^{12})^T$ and A^{22} are retained. So we require that

$$\tilde{A} = \left(\begin{array}{c|c} A_{11}^* & * \\ \hline * & * \end{array}\right) = \left(\begin{array}{c|c} A^{11} + E & A^{12} \\ \hline A^{21} & A^{22} \end{array}\right)^{-1}$$

where $*$ denotes unspecified and E is an increment matrix to be found. Using standard results on inverses of partitioned matrices we obtain

$$A_{11}^* = \left(A^{11} + E + A^{12} \left(A^{22}\right)^{-1} A^{21}\right)^{-1}$$

and so

$$\begin{aligned}E &= \left(A_{11}^*\right)^{-1} - \left(A^{11} + A^{12}(A^{22})^{-1}A^{21}\right) \\ &= \left(A_{11}^*\right)^{-1} - (A_{11})^{-1}\end{aligned}$$

is the required increment. Write the matrix obtained in this way $\tilde{A} = A|A_{11}^*$. This technique is used when a subset of individuals in a pedigree are genotyped, to compute a hybrid expected/realized relationship matrix exploiting both pedigree and genomic information [20, 21].

Results

In this section empirical comparisons are made between the proposed relationship matrices B and C, the numerator relationship matrix A derived from the pedigree, and the matrix G of vanRaden [10].

Data and computations

The data used in this analysis are genotypes and complex traits for 5037 Nordic Holstein bulls. The 5037 bulls were genotyped using a 50K chip and then imputed with Impute2 [22] to 777K (HD), using a reference panel of 1197 HD genotyped bulls. Five traits are examined below: protein, fat, yield, body and mastitis. These are de-regressed proofs (DRP) derived from genetic evaluations in December 2013. A detailed description of their definitions and derivations is available from the Danish Agricultural Advisory Centre (`https://www.landbrugsinfo.dk`). The year of birth of the bulls ranged from 1974 to 2009. The 3914 bulls born until 2004 were taken to comprise the training set, and the remaining 1394 bulls born from 2005 to 2009 were taken to comprise the test set.

The pedigree-based relationship matrix (A) for the 5037 bulls was derived from the Nordic Holstein pedigree of year 2013 (which contains a total of 134832 animals) in the standard way.

The genomic relationship matrix (G) following [10] was calculated from the marker data, using

$$g_{ij} = \frac{\sum_{k=1\ldots p}(m_{ik} - \bar{m}_k)(m_{jk} - \bar{m}_k)}{\sum_{k=1\ldots p} 2\bar{m}_k(1 - \bar{m}_k)} \tag{14}$$

where $M = (m_{ik})$ is the $N \times p$ marker design matrix whose elements take values in $\{0, 1, 2\}$, and $\bar{m}_k$ is the mean allele frequency of the relevant allele of the kth marker, that is, $\bar{m}_k = \sum_{i=1\ldots N} m_{ik}/N$. Thus the allele frequencies are set to those in the current sample.

Beagle version 3.3.2 was applied to the unphased marker data from the Holstein bulls, and the B and C matrices were derived from the Beagle output files. Beagle uses two tuning parameters, m and b. The larger the parameters, the simpler the selected APFA. The settings $m = 1$ and $b = 0$, suggested in [15, 18], were used below in the following sections, except Section "Prediction": here the settings $m = 4$ and $b = 0.2$ suggested in [23] were used, since they result in slightly better prediction accuracy.

Relatedness and pedigree distance

In Sections "Two measures of genomic relatedness" and "Expected relatedness" it was seen that there is a close conceptual relationship between C and the numerator relationship matrix A. This section examines empirically the extent to which the measures capture the genealogical relationships between individuals. This is done in several ways.

In a crude but informative approach, the distance between each pair of bulls may be calculated as the length of the shortest path between the bulls in the pedigree. For example, full- and half-sibs are at distance two, first cousins are at distance four, and second cousins are at distance six. Figure 3 shows sample densities of the four measures broken down by distance for all pairs of animals. Corresponding summary statistics are shown in Table 1.

The parent-offspring pairs are clearly identified by all four measures, but for the more distant pairs there appears to be least separation for the G measure. For the A measure, distances of four and above are less than 0.2, with a spike close to zero at distance 9, reflecting the assumption of unrelated founders. There are distinct peaks for most distances, indicating good separation between these.

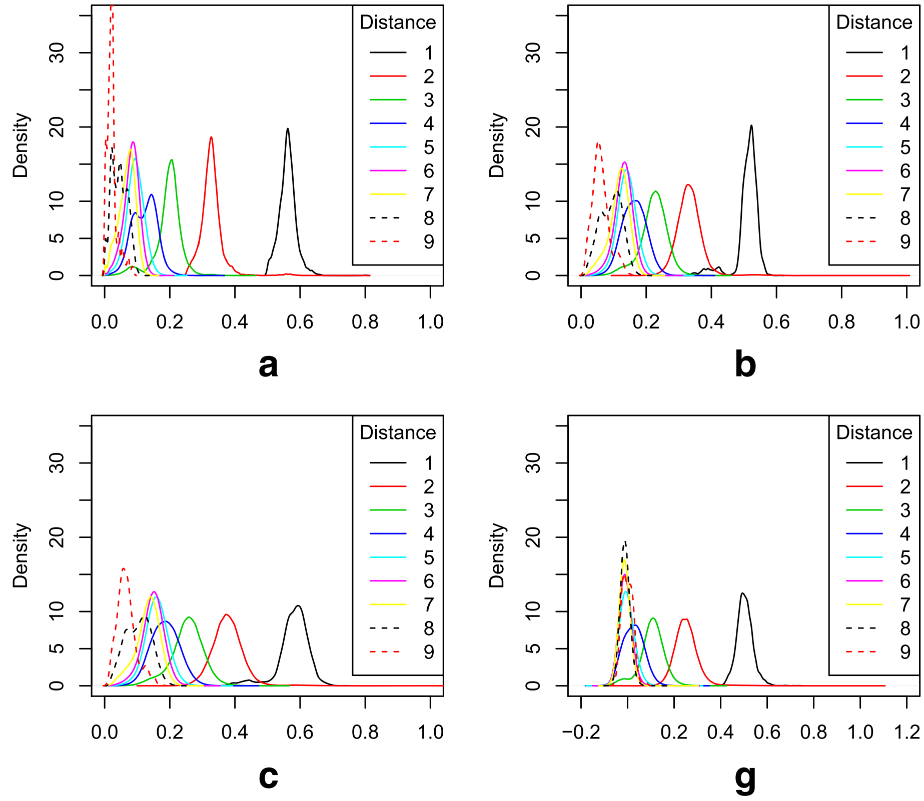

Fig. 3 Sample densities of relationship measures by pedigree distance. Sample densities broken down by pedigree distance for four measures: **a** the numerator relationship matrix; **b** the intersect measure; **c** the product measure and **g** the genomic relationship measure of [10]

Table 1 Summary statistics of relationship measures broken down by distance

	A		B		C		G	
Distance	mean	sd	mean	sd	mean	sd	mean	sd
1	0.562	0.026	0.510	0.037	0.579	0.048	0.504	0.035
2	0.328	0.035	0.332	0.040	0.377	0.049	0.252	0.052
3	0.195	0.042	0.223	0.042	0.253	0.050	0.104	0.054
4	0.122	0.037	0.163	0.037	0.186	0.044	0.025	0.046
5	0.094	0.027	0.140	0.030	0.160	0.036	-0.002	0.033
6	0.082	0.024	0.130	0.029	0.148	0.034	-0.010	0.027
7	0.068	0.025	0.117	0.031	0.132	0.037	-0.012	0.024
8	0.046	0.023	0.092	0.034	0.103	0.039	-0.011	0.021
9	0.023	0.016	0.061	0.026	0.067	0.029	0.000	0.024

The B and C measures at distance four and above are larger, being less than 0.3, also with good separation. The G measures for distance five and above are centered around zero, so about half of the values are negative, and there is poor separation. This suggests that G performs relatively poorly for distantly related individuals, say second cousins or farther.

Comparison with pedigree-based relationships

The distance measure just described is crude since pairs of animals at a given distance may be more or less related, due to varying numbers and lengths of lineage paths between them and common ancestors in the pedigree. The A matrix takes this into account and is the natural pedigree-based measure. The upper three subplots of Fig. 4 show smoothed scatterplots of A versus G, B, and C, for all distinct pairs of animals in the data. It is seen that the bulk of the points lie above the identity line in the A versus G plot, and under the line in the A versus B, and A versus C plots: that is, G tends to underestimate A whereas B and C tend to overestimate A. See also Table 1. Adjusted R^2 statistics based on simple linear regression models with no intercept, as shown in Fig. 4, indicate that G only explains around 6 % of the variation of A, whereas both B and C explain around 95 %.

Comparison using consistency

The tendency for B and C to be larger than A could be due to the assumption of non-inbred, unrelated founders that underlies A: if this is false, deflated estimates of relatedness and inbreeding would result. To examine this possibility we use the technique described in Section "Modifying A" to examine the consistency of B, C and G with A, in the following way. A random sample of 1000 animals from the genotyped Holstein bulls is taken, and called group one. The matrices $A|G_{11}$, $A|B_{11}$ and $A|C_{11}$ are derived and compared with the realized relationships, that is, the off-diagonals of $(A|G_{11})_{22}$ are compared with those of G_{22} and so forth. The process is repeated for 10 random samples of size 1000. The three lower subplots in Fig. 4 show the results. It is seen that B is highly consistent with A: the realized relatedness measures in B_{22} are very close to the adjusted values $(A|B_{11})_{22}$. The same is true of C. But G shows poor consistency with A: the realized relatedness measures in G_{22} tend to exceed the adjusted values $(A|G_{11})_{22}$.

Comparison of inbreeding coefficients

Finally, Fig. 5 shows smoothed scatter plots comparing the inbreeding coefficients obtained from A with those from G and C. It is seen that C explains 89 % of the variation in A whereas G explains only 42 %. The inbreeding coefficients from A are non-negative, but negative values occur in G. The inbreeding coefficients from C are consistently larger than those from A, which may reflect assumptions of non-inbred unrelated founders underlying A. To examine this, the consistency of the inbreeding coefficients from C and G are compared using the method just described: the results are shown in Fig. 5. The coefficients from C_{22} are slightly smaller than those from $(A|C_{11})_{22}$. Thus the difference may at least in part be due to assumptions of non-inbred unrelated founders. The consistency of the inbreeding coefficients from G with those from A is very poor. Estimates of inbreeding coefficients based on G may be sensitive to choice of allele frequencies in the base population [10].

Prediction

To compare the use of the relatedness measures in prediction, breeding values were predicted using a genomic restricted maximum likelihood (G-REML) model of the form

$$y = \mathbf{1}\mu + g + e \tag{15}$$

where y is the response vector, μ is the overall mean, $\mathbf{1}$ is a vector of 1's, g is a vector of breeding values, and e

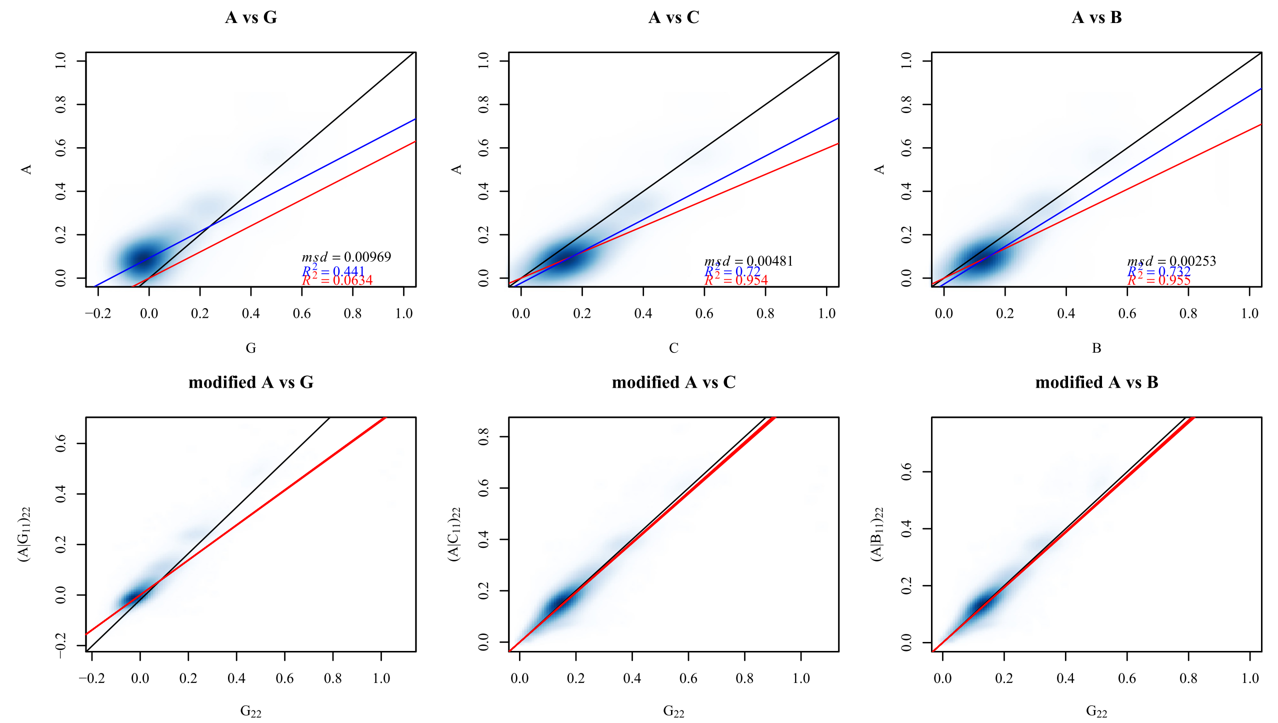

Fig. 4 Comparison of relationship coefficients of *A* with those of *G*, *C* and *B*. The three upper subplots show smoothed scatter plots of the off-diagonals of A with those of G, B, and C. Identity lines are shown in black, and linear regression lines with and without intercept terms are shown in blue and red, respectively. The mean squared deviation (average of $\sum_{i \neq j}(x_{ij} - y_{ij})^2$) and the adjusted R^2 values corresponding to the regressions are also shown. The three lower subplots display the consistency of *G*, *C* and *B* with *A*, showing smoothed scatter plots of $(A|G_{11})_{22}$ versus G_{22}, $(A|C_{11})_{22}$ versus C_{22} and $(A|B_{11})_{22}$ versus B_{22}, respectively. Identity lines are shown in black, and linear regression lines in red: these are overlaid from 10 replicates

is a vector of residuals. It is assumed that $g \sim N(0, V\sigma_g^2)$ and independently $e \sim N(0, D\sigma_e^2)$, where V is a relationship matrix (i.e. A, B, C, or G), and D is a diagonal matrix with elements $d_{kk} = (1 - r_k^2)/r_k^2$ to account for heterogeneous residual variances due to varying reliability r_k^2 of the complex trait y.

The analysis was performed with the package DMU [24] applied to data from the training set, using the four relationship matrices. To examine prediction using less related individuals, a reduced training set was constructed by excluding all sires and grandsires of any animal in the test set from the training set. The prediction accuracy, that is, the correlation between the predicted and observed values in the test set, are shown in Table 2. It is seen that *C* consistently has the highest prediction accuracy, though the improvement over *G* is modest, of the order of 0.4 % when the full training set is used, and 0.6 % when closely related animals are excluded from the training set. In contrast, *B* has slightly less prediction accuracy than *C* for both the full and the reduced training set.

Software

Beagle version 3.3.2 was used to select APFA on which the measures are based. A C++ program, available from the author, was written to construct the *B* and *C* relationship matrices from Beagle output files. DMU [24] was used to perform the REML analyses. The remaining computations were performed using R: in particular, the *A* matrix was computed using the pedigree package, and Figs. 1 and 2 were constructed using the gRapfa package [18].

Discussion

Two novel measures of relatedness based on shared haplotypes were introduced in Section "Methods". The intersect measure (*B*) is an estimate of the fraction of shared genome, and the product measure (*C*) is closely related to the numerator relationship matrix (*A*) [12].

The framework underlying the measures is that of a diploid genome divided into intervals and segments (or genes and alleles) in which it is assumed that segments are not shared between intervals, so that the multiplicity of each segment in a genome is in $\{0, 1, 2\}$. It would be interesting to examine whether the measures can be extended to polyploid genomes, and whether the assumption of no shared segments, which cannot accommodate phenomena such as gene duplication, can be relaxed.

The close conceptual relation between *C* and *A* rests implicitly on an assumption that whenever two haplotypes

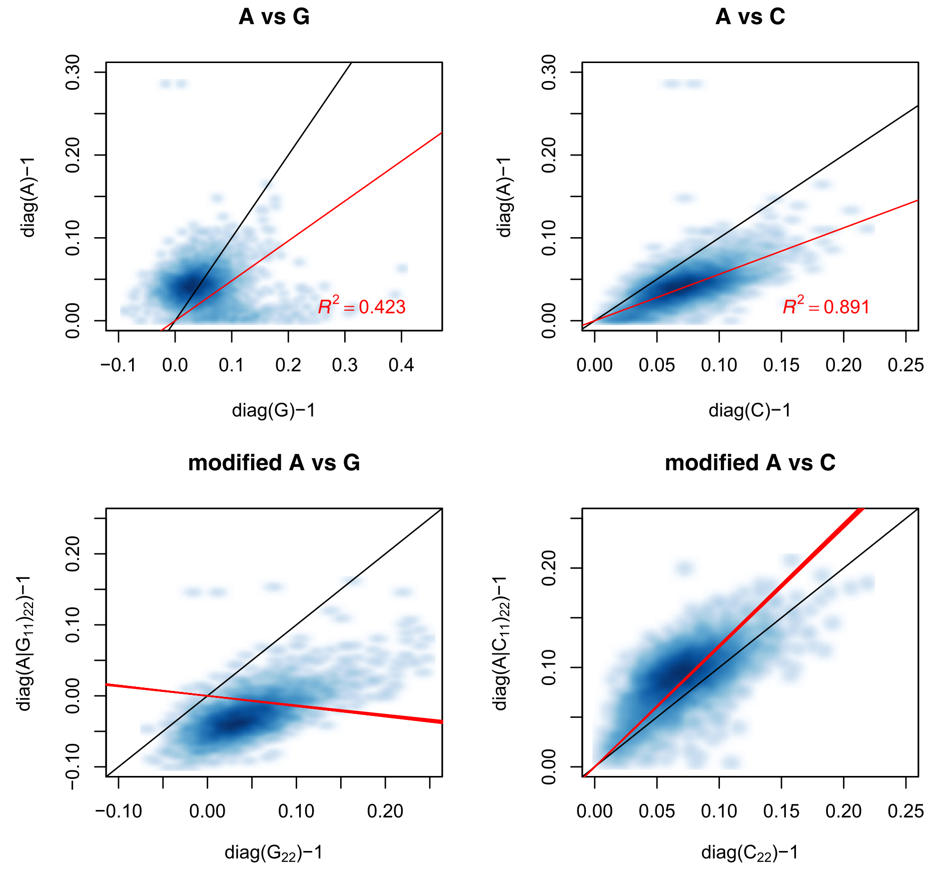

Fig. 5 Comparison of inbreeding coefficients from *A* with those from *G* and *C*. The two upper subplots show smoothed scatter plots comparing the inbreeding coefficients from *A* and *G*, and from *A* and *C*, respectively. Identity lines are shown in black and linear regression lines (with zero intercept) in red. The corresponding adjusted R^2 values are also shown. The two lower subplots display the consistency of the estimates of inbreeding coefficients from *G* and *C*, respectively, with those from *A*. Smoothed scatter plots of inbreeding coefficients from $(A|G_{11})_{22}$ versus G_{22}, and $(A|C_{11})_{22}$ versus C_{22} are shown. Identity lines are shown in black, and linear regression lines (with zero intercept) in red: these are overlaid from 10 replicates

share the same segment (in the APFA context, traverse the same edge) their chromosomal segments are indeed identical (IBS). This is a strong assumption, and most likely only approximately true. Probably the reason that the proposed measures capture genealogical relatedness well here is that the APFA-based segmentation is in reasonable accordance with this assumption. If the sample size were small, a overly simple APFA would be selected, leading to over-estimation of haplotype sharing. Similarly, if segments were defined directly using the marker alleles (say, with two segments per interlocus interval corresponding to the alleles of a flanking marker), the assumption would be violated, and the resulting measures may be expected to capture relatedness poorly. If full-sequence data are available, it would in principle be possible to verify the assumption for the APFA-based segmentation, or perhaps to develop an improved segmentation method.

Table 2 Prediction accuracy (correlation) for G-REML using the four relationship matrices

	Full training set				Reduced training set			
Trait	A	B	C	G	A	B	C	G
protein	0.498	0.661	0.670	0.667	0.219	0.556	0.562	0.559
fat	0.474	0.625	0.651	0.643	0.220	0.526	0.556	0.549
body	0.490	0.567	0.570	0.565	0.377	0.518	0.526	0.512
mast	0.454	0.572	0.585	0.581	0.298	0.504	0.513	0.514
yield	0.513	0.656	0.667	0.663	0.203	0.543	0.554	0.549

In Section "Results" it was shown that the tendency for B and C to be larger than A can be ascribed to the assumptions of non-inbred, unrelated founders that underlie A. An alternative explanation could be that the segmentation method chosen here tends to overestimate the extent of haplotype sharing. Further research into this would be useful.

A comparison of the prediction accuracy in a mixed linear model using the relationship matrices as covariances found that C performed consistently better than G, with an improvement of about 0.4 % when all available animals were used in the training set, increasing to 0.6 % when close relatives were excluded. There is intense interest in methods to improve prediction accuracy in genomic selection programmes [25], since small improvements may represent substantial economic gains for the breeding company, and the present methods may contribute to this goal.

Note that as described above C takes the form $XX^T/2p$, where X is the $N \times |E|$ haplomarker design matrix. Hence the use of C in (15) is equivalent to the model $y = \mathbf{1}\mu + Xh + e$ with random haplomarker effects $h \sim N(0, I_{|E|}\sigma_g^2)$ and independent error $e \sim N\left(0, D\sigma_e^2\right)$. From Eq. (14), G takes the form $(M-\delta)(M-\delta)^T/\eta$ where M is the $N \times p$ marker design matrix, and δ and η are shift and scale constants. Although different shift and scale transformations of the (haplo)marker variables would lead to different relationship matrices, they would not affect the predictive ability of the models [26]. So in this sense the comparison between C and G in the prediction context is between the predictive power of X and M, rather than between the relatedness measures *per se*.

It is straightforward to construct weighted versions of the measures. Let $w_1, \ldots, w_p$ be a set of apriori given non-negative numbers such that $\sum_{i=1\ldots p} w_i = 1$. These could for example be proportional to inter-marker distances, or to probabilities of the existence of a quantitative trait locus (QTL) in the respective interval in order to quantify trait-specific relatedness. Expressions (1) and (2) are replaced by $\sum_{s\in\mathcal{S}} w_{l(s)}(x_{is} \wedge x_{js})$ and $\sum_{s\in\mathcal{S}} w_{l(s)} x_{is} x_{js}$, respectively.

The present methods have a certain similarity of approach to that of the Chromopainter program [27]. This seeks to explore admixture in SNP data sampled from multiple populations. Given SNP data for a set of recipient chromosomes, and for a set of donor chromosomes, it forms each recipient chromosome as a mosaic of donor chromosomes, by applying the haplotype copying model [28] in a hidden Markov model framework. This has been used to explore human migratory history [29]. The present methods provide an alternative modelling approach in which it is not necessary to prescribe a donor/recipient ordering. A review of genomic similarity measures from the population structure perspective is given in [30].

A natural way to display patterns of relatedness is to apply principal coordinates analysis ([31], Chapter 14), using $-\log(b_{ij})$ as a distance measure between individuals i and j. Also the length of shared regions is informative: on average, the longer the shared regions, the more recent the ancestor(s). The location of the shared regions may sometimes also be of interest, for example, when there is knowledge of the location of genetic variants influencing a complex trait.

Conclusions

Two novel molecular measures of relatedness based on haplotype sharing are described. The intersect measure estimates the fraction of shared genome between individuals, and the product measure has a close conceptual relationship with the coefficient of coancestry. Both capture genealogical relatedness well, outperforming vanRaden's G in this respect. When used in genomic prediction models, the product measure leads to slightly improved prediction accuracy.

Availability of data and materials

The cattle data are the property of the Danish Cattle Federation (Aarhus, Denmark), Faba Co-op (Helsinki, Finland), Seges (Aarhus, Denmark), Växa Sverige (Uppsala, Sweden), Swedish Dairy Association (Stockholm, Sweden), and Nordic Cattle Genetic Evaluation (Aarhus, Denmark), and are not publicly available.

Competing interests

The author declares that he has no competing interests.

Acknowledgements

Helpful conversations with Rasmus Brøndum, Ole Christensen, Luc Janss, Bernt Guldbrandtsen and Just Jensen are gratefully acknowledged. I thank the Danish Cattle Federation (Aarhus, Denmark), Faba Co-op (Helsinki, Finland), Seges (Aarhus, Denmark), Växa Sverige (Uppsala, Sweden), Swedish Dairy Association (Stockholm, Sweden), and Nordic Cattle Genetic Evaluation (Aarhus, Denmark) for providing data. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.

Received: 26 July 2015 Accepted: 29 October 2015
Published online: 11 November 2015

References

1. Balding DJ, Nichols RA. Dna profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int. 1994;64(2):125–40.
2. Norder H, Hammas B, Lee SD, Bile K, Couroucé A-M, Mushahwar IK, et al. Genetic relatedness of hepatitis b viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen. J Gen Virol. 1993;74:1341–1341.
3. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):1002453.
4. Oliehoek PA, Windig JJ, Van Arendonk JA, Bijma P. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genet. 2006;173(1):483–96.

5. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54.
6. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–6.
7. Malécot G. Mathématiques de l'Hérédité. Paris: Masson; 1948.
8. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16(1):33–44.
9. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;7(10):771–80.
10. VanRaden P. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–423.
11. Syropoulos A. Mathematics of multisets In: Calude C, editor. Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View. New York: Springer; 2001. p. 347–58.
12. Wright S. Coefficients of inbreeding and relationship. Am Nat. 1922:330–338.
13. Ron D, Singer Y, Tishby N. On the learnability and usage of acyclic finite automata. J Comput Syst Sci. 1998;56:133–52.
14. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84(2):210–23.
15. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol. 2007;31(5):365–75.
16. Brøndum RF, Ma P, Lund MS, Su G. Short communication: Genotype imputation within and across nordic cattle breeds. J Dairy Sci. 2012;95(11):6795–800.
17. Edwards D, Ankinakatte S. Context-specific graphical models for discrete longitudinal data. Stat Model. 2014;15(4):301–25.
18. Ankinakatte S, Edwards D. Modelling discrete longitudinal data using acyclic probabilistic finite automata. Comput Stat Data Anal. 2015;88: 40–52.
19. Emik LO, Terrill CE. Systematic procedures for calculating inbreeding coefficients. J Hered. 1949;40(2):51–5.
20. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92(9):4656–663.
21. Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 2010;42(2):1–8.
22. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499–511.
23. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
24. Madsen P, Sørensen P, Su G, Damgaard LH, Thomsen H, Labouriau R. DMU-a package for analyzing multivariate mixed models. In: 8th World Congress on Genetics Applied to Livestock Production; 2006:247.
25. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Gen. 2013;193(2):327–45.
26. Strandén I, Christensen OF. Allele coding in genomic evaluation. Genet Sel Evol. 2011;43:25.
27. Hellenthal G, Auton A, Falush D. Inferring human colonization history using a copying model. PLoS Genetics. 2008;4(6). doi:10.1371/annotation/da6e20fe-f8eb-4e44-8669-6c20c7102b3d.
28. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213–233.
29. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172): 747–51.
30. Lawson DJ, Falush D. Population identification using genetic data. Annu Rev Genomics Hum Genet. 2012;13:337–61.
31. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic press; 1979.