Received: 27 May 2016 | Revised: 15 March 2017 | Accepted: 21 March 2017

DOI: 10.1111/1755-0998.12674

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Graphics for relatedness research

Iván Galván-Femenía[1,2] | Jan Graffelman[3,4] | Carles Barceló-i-Vidal[1]

[1]Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain

[2]Disease Genomics-GCAT Group, Germans Trias Health Research Institute (IGTP)-Program of Predictive and Personalized Medicine of Cancer (PMPPC), Can Ruti Campus, Badalona, Barcelona, Spain

[3]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

[4]Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence
Iván Galván-Femenía, Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain.
Email: ivan.galvan@udg.edu

Funding information
United States National Institutes of Health, Grant/Award Number: R01 GM075091; Spanish Ministry of Economy and Competitiveness, Grant/Award Number: MTM2015-65016-C2-2-R, MTM2015-65016-C2-1-R

Abstract

Studies of relatedness have been crucial in molecular ecology over the last decades. Good evidence of this is the fact that studies of population structure, evolution of social behaviours, genetic diversity and quantitative genetics all involve relatedness research. The main aim of this article was to review the most common graphical methods used in allele sharing studies for detecting and identifying family relationships. Both IBS- and IBD-based allele sharing studies are considered. Furthermore, we propose two additional graphical methods from the field of compositional data analysis: the ternary diagram and scatterplots of isometric log-ratios of IBS and IBD probabilities. We illustrate all graphical tools with genetic data from the HGDP-CEPH diversity panel, using mainly 377 microsatellites genotyped for 25 individuals from the Maya population of this panel. We enhance all graphics with convex hulls obtained by simulation and use these to confirm the documented relationships. The proposed compositional graphics are shown to be useful in relatedness research, as they also single out the most prominent related pairs. The ternary diagram is advocated for its ability to display all three allele sharing probabilities simultaneously. The log-ratio plots are advocated as an attempt to overcome the problems with the Euclidean distance interpretation in the classical graphics.

KEYWORDS
compositional data analysis, identical by state/descent, isometric log-ratio, microsatellite, relatedness, ternary diagram

1 | INTRODUCTION

Statistical methods for the analysis of the genetic relationships between individuals of a population are of great relevance for molecular ecology (Blouin, 2003). Studies of relatedness are crucial for studying population structure, evolution of social behaviour, genetic diversity, quantitative genetics, etc. It is known that the estimation of quantitative genetic parameters in wild populations is less biased and more precise if we dispose of pedigree information (Bérénos, Ellis, Pilkington, & Pemberton, 2014). The role of relatedness for selective breeding is also recognized. Loughnan, Smith-Keune, Jerry, Beheregaray, and Robinson (2016) recommend low levels of relatedness and high levels of neutral genetic diversity to form a base population for selective breeding. The exclusion of duplicated individuals and close relatives is a previous quality control filter used in studies of population structure (Gonder et al., 2015). Relatedness estimation is also important for conservation programmes, and the performance of several estimators has been compared in that context (Oliehoek, Windig, van Arendonk, & Bijma, 2006). It plays an important role in structuring societies with fusion–fission dynamics (Croft et al., 2012; Snyder-Mackler, Alberts, & Bergman, 2014; Spencer et al., 2015), can bias estimates of allele frequencies (Hansen, Nielsen, & Mensberg, 1997) and violates the assumption of independent individuals in trait-gene association studies (Foulkes, 2009). Thus, statistical methods that can verify documented or uncover undocumented family relationships in the database are important tools in molecular ecology.

Relatedness investigations can be carried out in an entirely numerical manner by inspecting estimated IBS (identity by state) and IBD (identity by descent) probabilities, likelihood ratios or confusion matrices (Boehnke & Cox, 1997; Epstein, Duren, & Boehnke, 2000). Graphics greatly facilitate the interpretation of the results of relatedness studies and are increasingly being used (Abecasis, Chemy, Cookson, & Cardon, 2001; Pemberton, Wang, Li, & Rosenberg, 2010; Rosenberg, 2006). The main aim of this article was to summarize the state of the art of the graphical methods used in relatedness research. Relatedness investigations are based on allele sharing, and we will consider techniques that use IBS alleles as well as those using IBD alleles. A plot of the means against the standard deviations of the IBS counts is a powerful tool to detect relatedness (Abecasis et al., 2001). We explore this tool in detail and establish the domain of this graphic from a mathematical point of view. Plots of the proportions of markers with 0, 1 or 2 IBS counts (p_0, p_1 or p_2) are often used to assess the existence of family relationships (Rosenberg, 2006). Nevertheless, if the researcher is interested in identifying the degree of relatedness, plotting the probabilities of sharing 0, 1 or 2 IBD alleles (k_0, k_1 or k_2) is the best strategy. The IBD probabilities depend directly on relatedness and enable us to accurately infer the type of relationship. In addition to the former graphical methods, we propose to use graphics from compositional data analysis (CoDA) for both IBS and IBD allele sharing studies. Due to the fact that the proportions (p_0, p_1, p_2) and the probabilities (k_0, k_1, k_2) are constrained to sum to one, it is possible to apply all the graphical and analytical CoDA techniques introduced by Aitchison (1986) and developed posteriorly by Pawlowsky-Glahn and Buccianti (2011). Two graphics, commonly used in CoDA, are of particular relevance for relatedness studies: the ternary diagram (also known as a *de Finetti* diagram in genetics) and a scatterplot of log-ratios. We show the ternary diagram to be useful for plotting the proportions of the IBS counts and for plotting the estimated Cotterman coefficients (IBD probabilities). Moreover, the theoretical IBD sharing probabilities for the standard family relationships can be used as reference points in the ternary diagram (Thompson, 2000). Furthermore, the CoDA techniques allow us to introduce the isometric log-ratio coordinates (ilr-coordinates) of the vectors $\boldsymbol{p} = (p_0, p_1, p_2)$ and $\boldsymbol{k} = (k_0, k_1, k_2)$, which we can represent in a scatterplot. These ilr-coordinates allow us to measure the degree of similarity between two vectors of IBS proportions or IBD probabilities. The graphics we propose are of universal value and can be used in any relatedness study that concerns diploid individuals.

The remainder of this article is organized as follows. Section 2 gives an overview of the IBS allele sharing analysis and the graphical methods used to detect family relationships. Section 3 presents the basic principles of IBD estimation and the most common graphics used for relatedness estimation in the IBD context. The former sections also detail the graphical methods from the field of CoDA used in IBS-IBD approaches: the ternary diagram and the scatterplot of log-ratios. Section 4 presents a way to enhance IBS and IBD graphics with convex hulls that express the degree of uncertainty about a relationship. Section 5 presents a case study with individuals from the Maya population. Finally, Section 6 summarizes the principal conclusions of this article and the pros and cons of each graphical method are discussed.

2 | IBS STUDIES

IBS studies disregard if the alleles for any diploid individual are derived from a common ancestor. IBS allele sharing concerns the number of matches between the alleles of the genotypes of two individuals. Two diploid individuals can share 0 (e.g., A1/A1 and A2/A2 or A1/A2 and A3/A3), 1 (e.g., A1/A1 and A1/A2 or A1/A2 and A1/A3) or 2 (e.g., A1/A1 and A1/A1) IBS alleles for a specific genetic marker, and we will refer to these as IBS counts. To detect family relationships in a given population of n individuals and m genetic markers, the number of matches between IBS alleles (the IBS counts) is considered for each pair of individuals across genetic markers. That is, we move from a data set of n individuals and m genetic markers to a data set of $\binom{n}{2}$ pairs of individuals with the information of the IBS counts for m genetic markers. There are different ways to deal with this type of data as described below. First, we focus on the plot of means and standard deviations of the IBS counts (Abecasis et al., 2001). Second, we detail the plot of the proportions of the IBS counts (Rosenberg, 2006). To conclude this section, graphics from CoDA (Aitchison, 1986; Pawlowsky-Glahn & Buccianti, 2011) are presented.

To illustrate the different IBS graphics that are introduced in this Section, we use five pairs of individuals with the information of IBS counts and IBS proportions for 377 microsatellites (see Table 1). The individuals are from the Maya population which we will analyse in Section 5. We consider a parent-offspring (PO) pair, a full-sib (FS) pair, a half-sib (HS), avuncular (AV) or grandparent-grandchild (GG) pair, a pair of first cousins (FC) and a pair of unrelated individuals (UN). We discuss the different graphics in the sections below.

2.1 | $(\bar{x}, s)$-plot

Let x_{ijk} be the number (0, 1 or 2) of shared IBS alleles between individual i and j for the genetic marker k. Abecasis et al. (2001) proposed to compute the mean ($\bar{x}_{ij}$) and variance (s^2_{ij}) over K genetic markers. The plot $\bar{x}_{ij}$ versus s_{ij} reveals characteristic clusters that correspond to the different family relationships for a given population.

The statistics $\bar{x}_{ij}$ and s^2_{ij} are constrained due to the limited number of outcomes (0, 1 or 2), and we proceed to derive their range of variation (Figure 1a). As an example, we consider a table with all possible outcomes of the allele sharing counts (0, 1 or 2) for a set of 100 markers. The rows of this table represent possible pairs of individuals. There are 3^{100} combinations (rows), if the order of the outcomes is considered relevant. However, in terms of means or standard deviations, the order of the IBS counts (0, 1 or 2) over the different markers is irrelevant but their multiplicity is important. For example, a pair of individuals sharing 1 IBS allele for the first marker and 0 for all other markers will have the same mean and variance as

TABLE 1 Computations for five pairs of individuals from the Maya population. Mean and standard deviation of IBS counts, proportion of sharing 0, 1 and 2 IBS alleles (p_0, p_1, p_2) and estimated Cotterman coefficients ($\hat{k}_0, \hat{k}_1, \hat{k}_2$) are shown

	IBS studies					IBD studies		
Type of relative	Mean	Standard deviation	p_0	p_1	p_2	$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$
PO	1.34	0.48	0.002	0.650	0.348	0.009	0.991	0.000
FS	1.32	0.60	0.073	0.532	0.395	0.214	0.617	0.169
HS, AV or GG	1.09	0.64	0.160	0.581	0.259	0.447	0.553	0.000
FC	1.00	0.67	0.225	0.546	0.229	0.657	0.343	0.000
UN	0.86	0.67	0.308	0.526	0.166	0.731	0.269	0.000

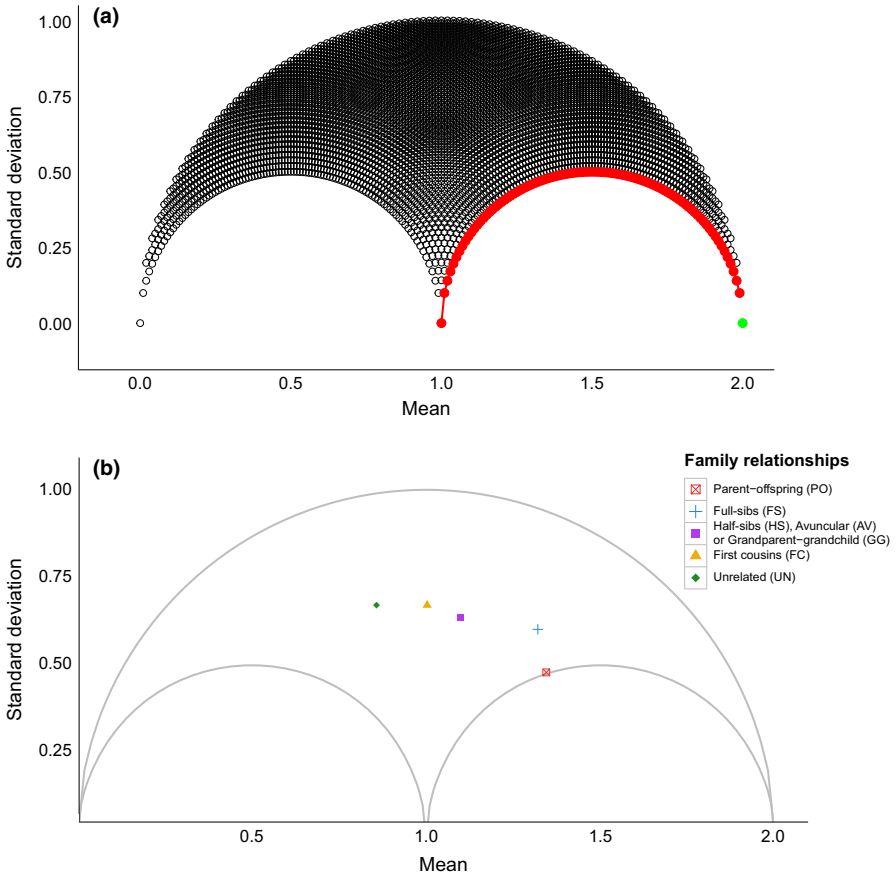

FIGURE 1 a. Plot of means and standard deviations of all possible combinations of IBS counts for a table of 100 genetic markers. The red curve shows the pairs of individuals that are parent-offspring. The green point represents a monozygotic twin pair or a pair of duplicated individuals. b. Plot of means versus standard deviations of the IBS counts for five pairs from the Maya population

a pair of individuals sharing 1 IBS allele for the k-th marker and 0 for all others. Mathematically, the combinations of the IBS counts for a pair of individuals form a multiset (Stanley, 1997, Section 1.2) of cardinality m (the number of markers) made of a basic set of cardinality $k = 3$ (the outcomes 0, 1 and 2). The possible number of $(\bar{x}, s)$ pairs in the plot can be no larger than the number of multisets of cardinality k, where the latter is given by the multiset coefficient

$$\left(\binom{k}{m}\right) = \binom{k+m-1}{k}, \quad (1)$$

Thus, for 100 genetic markers there will be at most $\left(\binom{3}{100}\right) = \binom{3+100-1}{100} = \binom{102}{100} = 5151$ different $(\bar{x}, s)$ pairs. Figure 1a shows the means and standard deviations of the 5151 combinations of IBS counts for 100 genetic markers. The figure has the shape of an umbrella and represent the domain of the $(\bar{x}, s)$-plot. For empirical data, it will be impossible to observe a $(\bar{x}, s)$ point outside the umbrella region. It is clear that the mean of the IBS counts ranges from zero to two. The maximum variance equals one and is reached when the array of IBS counts has fifty 0 IBS alleles and fifty 2 IBS alleles, whereas the minimum variance equals zero and is reached when the array of IBS counts has either one hundred 0 IBS alleles, one hundred 1 IBS allele or one hundred 2 IBS alleles.

The red points on the right hand curve of the "umbrella" correspond presumably to parent–offspring relationships for having a mean larger than 1 and low variance. The first point of the curve

with mean equal to 1 IBS allele and standard deviation equal to 0 IBS alleles corresponds to an array of one hundred ones. The second point of the curve corresponds to an array of 99 markers with 1 IBS alleles and one marker with 2 IBS alleles, and so on. In other words, this red curve represents the pairs of individuals who have a mean larger than or equal to 1 and the smallest standard deviation of all possible IBS counts. This can be related with the fact that the probability of sharing 1 IBD allele between a parent-offspring equals 1, as we will see in the next Section (Table 2). For parent-offspring pairs, we have that $\bar{x}_{ij} \geq 1$ because children inherit at least 1 IBS allele from their parents. And for monozygotic twins (MZ) or duplicated individuals, we have $\bar{x}_{ij} = 2$ and $s_{ij} = 0$ (green point in Figure 1a).

Figure 1b shows the $(\bar{x}, s)$ plot for the five Maya pairs in Table 1. The larger the mean of the IBS counts for any pair of individuals, the more likely they are to be closely related. The PO pair (red point) is located on the right hand curve of the umbrella, the FS pair (blue point) with mean larger than 1 is separated from second- and third-degree family relationships (violet and gold points respectively), whereas, the unrelated individuals have the smallest mean (green point).

2.2 | (p_i, p_j)-plots

Let x_{ij} be the vector of the IBS counts between individual i and j as large as the number of the genetic markers in the data set. Let p_0, p_1 and p_2 be the proportions of 0, 1 and 2 IBS alleles, respectively, for each pair of individuals. Rosenberg (2006) proposed a graphical method for relatedness research by plotting the proportion of sharing 2 IBS alleles (p_2) versus the proportion of sharing 0 IBS alleles (p_0) for all pairs of individuals from a given population. Similarly, Sun (2012) uses IBS proportions for relatedness research by plotting p_1 versus p_0. In fact, any combination of the three proportions could be plotted for relatedness research. We refer to these graphics as (p_i, p_j)-plots (for i, j = 0, 1, 2 and $i < j$) were p_i corresponds to the X-axis of the plot and p_j to the Y-axis.

Monozygotic twins (MZ) or duplicated individuals are easy to identify in the (p_i, p_j)-plots because they have p_2 close to 1. PO pairs have low values of p_0 and are also easy to detect visually because they are on the p_1 or p_2-axis. FS usually have large values of p_2 and are separated from unrelated individuals. Second degree and third degree are more difficult to detect because positions in the plot depend on the allele frequencies of the population under study. Figures 2a, b and c show the (p_0, p_2)-, (p_0, p_1)- and (p_1, p_2)-plots for the five Maya pairs (Table 1). Notice that the distance between pairs of individuals is not the same in the three plots. For instance, the FS pair (blue point) is most close to the PO pair (red point) in the (p_0, p_2)-plot, but closer to the HS pair (violet point) in the (p_0, p_1)-plot. If the distances between pairs of individuals are different depending on the plotted proportions, then it is not appropriate to draw conclusions about the family relationship between individuals from the (p_i, p_j)-plots.

TABLE 2 Cotterman coefficients for the different type of family relationship and degree of relatedness

Type of relative	Degree	k_0	k_1	k_2
Monozygotic twins (MZ)	0	0	0	1
Parent-offspring (PO)	1	0	1	0
Full-siblings (FS)	1	1/4	1/2	1/4
Half-siblings (HS)/avuncular (AV)/grandchild-grandparent (GG)	2	1/2	1/2	0
First cousins (FC)	3	3/4	1/4	0
Unrelated (UN)	∞	1	0	0

2.3 | Ternary diagrams

Let $\boldsymbol{p}$ be the vector (p_0, p_1, p_2) of proportions of the IBS counts. Because the three components of $\boldsymbol{p}$ sum to one ($p_0 + p_1 + p_2 = 1$), we can plot the vector $\boldsymbol{p}$ in a ternary diagram. Mathematically, the set of the vectors of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ forms the simplex, S^3. Figure 3 shows the ternary diagram for the vectors of proportions for the five Maya pairs (Table 1). The PO pair (red point) is located on the opposite side of the vertex p_0; the FS pair (blue point) has the largest value for p_2 and is the closest to the p_2 vertex. The UN pair (green point), FC pair (gold point) and the HS, AV or GG pair

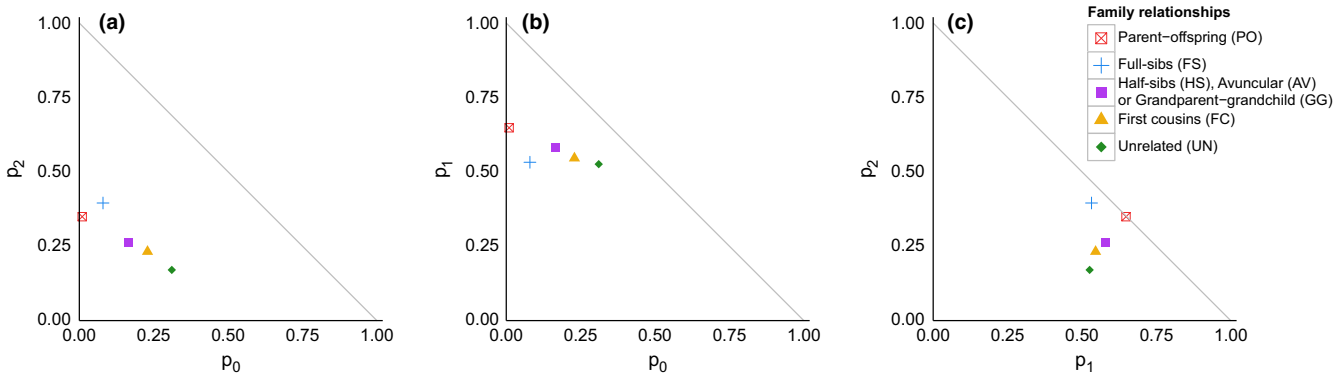

FIGURE 2 (p_i, p_j)-plots for five individuals from the Maya population. a. Plot of the proportion of sharing 0 IBS alleles (p_0) versus the proportion of sharing 2 IBS alleles (p_2): (p_0, p_2)-plot. b. Plot of the proportion of sharing 0 IBS alleles (p_0) versus the proportion of sharing 1 IBS allele (p_1): (p_0, p_1)-plot. c. Plot of the proportion of sharing 1 IBS allele (p_1) versus the proportion of sharing 2 IBS alleles (p_2): (p_1, p_2)-plot. [Colour figure can be viewed at wileyonlinelibrary.com]

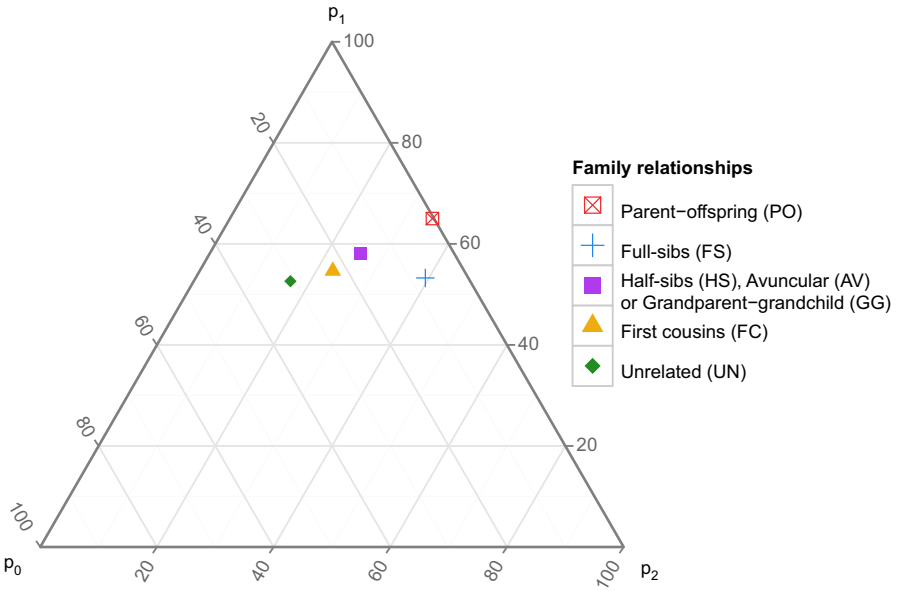

FIGURE 3 Ternary diagram of the IBS proportions for five pairs from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

(violet point) have lower values of p_2. The UN pair has the lowest values for p_2 and p_1 and is closest to the p_0 vertex. The main advantage of this graphical tool is that it represents the three proportions p_0, p_1 and p_2 simultaneously in contrast to the (p_i, p_j)-plots that represent only two of them.

2.4 | ilr-plots

Aitchison (1986) stated that it is not meaningful to interpret the distances between two vectors of proportions in the ternary diagram as if we were in an Euclidean space. Aitchison (1986) defines a new distance based on the log-ratio of the components of the vectors of proportions. This distance, jointly with the perturbation and powering operators (analogous to translation and scalar multiplication in the real space, respectively), forms the structure of the simplex in a two-dimensional metric space (Aitchison, Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Pawlowsky-Glahn & Buccianti, 2011). Thereby, the vectors of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ can be expressed in coordinates using any orthonormal basis defined in the simplex (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, & Barceló-Vidal, 2003). These coordinates are called isometric log-ratio coordinates (ilr-coordinates). The distance between two vectors of proportions is calculated as the Euclidean distance between their ilr-coordinates. The ilr-coordinates of a vector of proportions depend on the orthonormal basis used in the simplex. The most commonly used ilr-coordinates $\boldsymbol{z}_0$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ of a vector of proportions (p_0, p_1, p_2) are given by

$$\boldsymbol{z_0} = \begin{cases} z_{01} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_2}{p_1}\right) \\ z_{02} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_1 p_2}{p_0^2}\right) \end{cases} \quad \boldsymbol{z_1} = \begin{cases} z_{11} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_2}{p_0}\right) \\ z_{12} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_0 p_2}{p_1^2}\right) \end{cases} \quad \boldsymbol{z_2} = \begin{cases} z_{21} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_1}{p_0}\right) \\ z_{22} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_0 p_1}{p_2^2}\right) \end{cases}, \quad (2)$$

Figures 4a, b and c plot the ilr-coordinates for the five Maya pairs (Table 1). Notice that the distance between any pair of points is exactly the same in the three graphics, irrespective of the ilr-coordinates ($\boldsymbol{z_0}$, $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$) that are plotted. The PO pair (red point) in Figures 4a–c is an outlying pair. The FS pair (blue point) is also isolated from pairs of second and third degree of relationships. The degree of relationship decreases with the z_{02}, z_{11} and z_{21} ilr-coordinates (close relatives with a first-degree relationship (PO, FS) have larger values for these coordinates than second-degree relationships (HS, AV, GG)).

3 | IBD STUDIES

Studies of relatedness based on IBD alleles are based on the probabilities that a pair of individuals shares 0, 1 or 2 IBD alleles. These probabilities are commonly referred to as Cotterman's coefficients (Cotterman, 1941) and denoted by the vector of proportions $\boldsymbol{k} = (k_0, k_1, k_2)$. Table 2 shows the values of the Cotterman coefficients for some standard relationships. Cotterman's coefficients can be estimated by the maximum-likelihood method (Milligan, 2003; Weir, Anderson, & Hepler, 2006). The maximum-likelihood estimates reveal the most likely relationship for a pair given the observed genotype data. Let R represents a possible relationship between two individuals with genotypes G_1 and G_2, respectively. The likelihood of R is defined by the probability of observing G_1 and G_2 given relationship R. This probability depends on the allele frequencies of the population under study and is conditioned by the Cotterman coefficients. This likelihood is calculated across loci to obtain the most likely values (estimates) of the Cotterman coefficients. These estimates provide a first indication of the possible relationship between a pair of individuals. A hypothesis test is recommended to confirm or refute this relationship (García-Magariños, Egeland, López-de-Ullibarri, Hjort, & Salas, 2015). More details are explained by Wagner, Creel, and Kalinowski (2006). Under the assumption of absence of inbreeding, the inequality $k_1^2 \geq 4k_0k_2$ applies and constrains the Cotterman coefficients (Thompson, 1991).

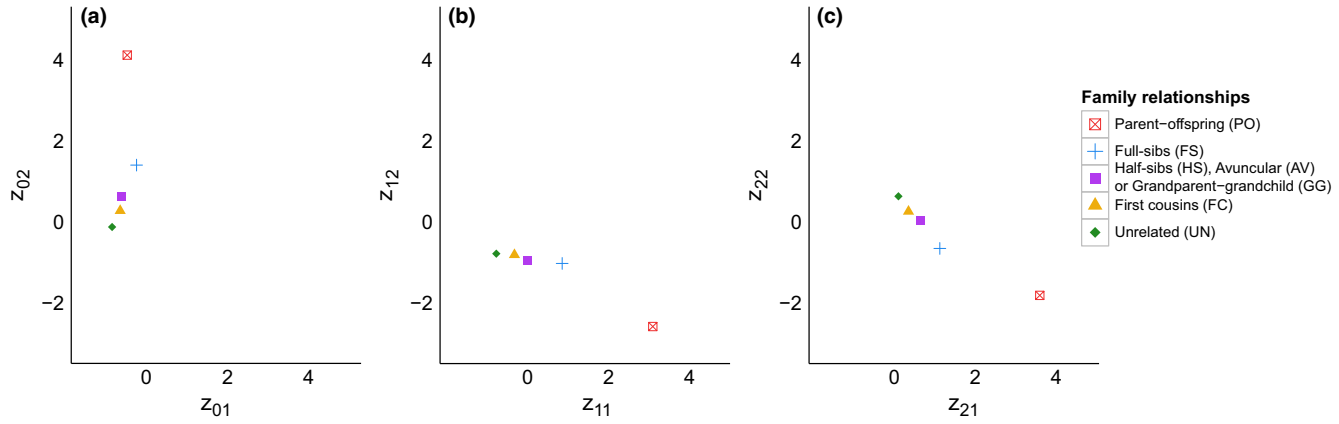

FIGURE 4 Ilr-coordinates of the IBS proportions for five pairs of individuals from the Maya population. a. $\boldsymbol{z}_0 = (z_{01}, z_{02})$. b. $\boldsymbol{z}_1 = (z_{11}, z_{12})$. c. $\boldsymbol{z}_2 = (z_{21}, z_{22})$. [Colour figure can be viewed at wileyonlinelibrary.com]

Analogously to the vector of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ of the IBS counts, Cotterman's coefficients also satisfy $k_0 + k_1 + k_2 = 1$. We can use the same graphical techniques described for $\boldsymbol{p} = (p_0, p_1, p_2)$ to identify relatedness from the estimated Cotterman coefficients $\hat{\boldsymbol{k}}$. The Cotterman coefficients can be represented in a $(\hat{k}_i, \hat{k}_j)$-plot, in a ternary diagram or in an ilr-plot with the ilr-coordinates $\boldsymbol{z}_0$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, defined in the Equation (2), substituting p_i for $\hat{k}_i$. With the aim of describing each graphical method used in IBD studies, we compute maximum-likelihood estimates of the Cotterman coefficients for the five Maya pairs (Table 1).

3.1 | $(\hat{k}_i, \hat{k}_j)$-plots

In the literature, the estimated Cotterman coefficients are plotted in different ways to identify relatedness. Nembot-Simo, Graham, and McNeney (2013) use the $(\hat{k}_0, \hat{k}_1)$-plot. Similarly, Moltke and Albrechtsen (2014) use the $(\hat{k}_1, \hat{k}_2)$-plot. The remaining possibility, the $(\hat{k}_0, \hat{k}_2)$-plot, could be also considered. Figure 5a shows the plot for the five Maya pairs (Table 1). The grey curve in the $(\hat{k}_0, \hat{k}_1)$-plot corresponds to the equation $k_1^2 = 4k_0k_2$. This curve jointly with the hypotenuse and the vertical axis delimits the feasible region $k_1^2 \geq 4k_0k_2$. PO pairs are points located on the k_1-axis with values close to 1, FS pairs are located close to the centre of the grey curve according to the theoretical IBD probabilities (Table 2) and second and third degree pairs are located around the centre of the hypotenuse. UN pairs theoretically have $k_0 = 1$ and are located between the hypotenuse and the grey curve, near to the vertex $\hat{k}_0 = 1$. Finally, the origin of the $(\hat{k}_0, \hat{k}_1)$-plot is the position for any MZ pair. As previously shown for IBS studies with the (p_i, p_j)-plots, only two of the three Cotterman coefficients are plotted and the relative positions and distances between points vary depending on the $(\hat{k}_i, \hat{k}_j)$-plot used. For this reason, we propose graphics from CoDA.

3.2 | Ternary diagrams

The theoretical IBD probabilities for the standard family relationships can be represented in a ternary diagram (Thompson, 2000). These probabilities form reference points against which the empirical estimates can be compared. Figure 5b shows the ternary diagram for the estimated Cotterman coefficients for the five Maya pairs

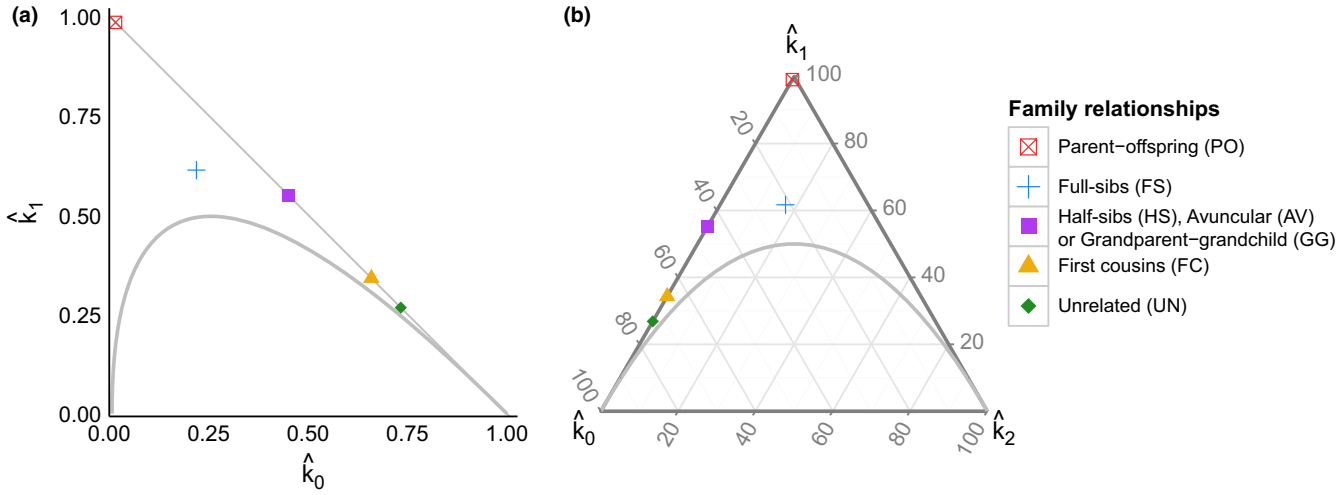

FIGURE 5 $(\hat{k}_0, \hat{k}_1)$-plot (a) and ternary diagram (b) for five pairs of individuals from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

(Table 1). Most pairs in Table 1 are close to their theoretical IBD probabilities given in Table 2. However, values of k_1 are larger than expected for the FS, HS, AV and notably, the UN pair (see the Discussion section). The domain has the shape of an arrowhead inside the ternary diagram. The curve delimiting the arrowhead from below corresponds to the inequality $k_1^2 \geq 4k_0k_2$.

3.3 | ilr-plots

It has been shown that the maximum-likelihood estimates of the Cotterman coefficients in the simplex are the same as the estimates obtained by maximizing the likelihood in ilr-coordinates (Graffelman & Galván-Femenía, 2016). With the aim of establishing reference zones for the standard family relationships in the ilr space, we compute the maximum-likelihood estimates of the Cotterman coefficients from the ilr-coordinates defined by the Equation (2) and we plotted the $\boldsymbol{z}_1 = (z_{11}, z_{12})$ ilr-coordinates as is shown in Figure 6. All the family relationships have values lower than $-\sqrt{(2/3)}\ln(2)$ for z_{12} which corresponds to the grey line in the graph. This line corresponds to the curve shown in the former graphs (Figure 5a and b). Due to the fact that some Cotterman coefficients equals 0, some of the (or both) ilr-coordinates tend to $+/-$ infinity. Thus, given that it is impossible to represent the point, we are limited to indicate the direction of the infinity in the ilr-plot for each type of family relationship. Regarding Figure 6, PO pairs have a large variability of values, either positive or negative for z_{11}; FS have values close to 0 for z_{11} and $-\sqrt{(2/3)}\ln(2)$ for z_{12}. HS, AV, GG and FC are located between PO, FS and UN. UN pairs have negative values of z_{11} which correspond to the green point of the left hand. If present, MZ pairs are points with positive values of z_{11} located on the right hand side of the plot.

4 | UNCERTAINTY IN IBS/IBD GRAPHICS

With the previously described graphics, one can try to infer the relationship of a pair for which the relationship is not documented, or try to confirm the documented relationships. Such graphical inference is hampered by the fact that the statistics represented in the graphs (means and standard deviations of the IBS counts, p_0, p_1, p_2, k_0, k_1, k_2) are subject to uncertainty. For a given sample, relationships are not represented by points, but by zones. Some insight into this uncertainty and the corresponding zones can be obtained by simulation. Ideally, this would require a large sample for which a subset of unrelated individuals can be identified. From these individuals, by sampling alleles across markers according to Mendelian laws, the reproductive process can be simulated allowing us to generate artificial children, leading to artificial PO pairs, FS pairs and artificial pairs of any other desired relationship. For example to simulate a PO pair we sample two UN individuals at random without replacement from the database. From each UN individual, we sample one allele at random from each marker and join the alleles to form a child. The process of sampling UN pairs and child generation is repeated many times, generating many artificial PO pairs. We can calculate the IBS/IBD statistics of the artificial pairs, and add these to the graphics of the previous sections by representing them individually or with a convex hull. A convex hull for a given set of points X is the unique convex polygon whose vertices are points from X and that contains all points of X (de Berg, van Kreveld, Overmars, & Schwarzkopf, 2000). By generating a large number of artificial pairs and representing these in the IBS/IBD graphics of interest, the zones corresponding to the different relationships can be approximated. Such simulations are conditional on the observed allele frequencies and can quantify the uncertainty in a graphical assessment of the relationship to some extent. We illustrate this with examples in the next section where all graphics are enhanced with hulls based on 80 PO, 48 FS, 120 second degree, 36 FC and 1256 UN artificially generated pairs.

5 | CASE STUDY

We applied all the graphical methods detailed in the previous sections using empirical data extracted from a world-wide data set from the Noah A. Rosenberg Research lab at Stanford University (Rosenberg et al., 2002). This world-wide database is derived from the Human Genome Diversity Cell Line Panel (HGDP, Cavalli-Sforza, 2005). The genetic information is given by 377 microsatellites genotyped for 52 human populations around the world. We used all 25 available individuals of the Maya sample to illustrate all graphical methods for relatedness research. All the family relationships present in this sample were reported by Rosenberg (2006). All the Figures

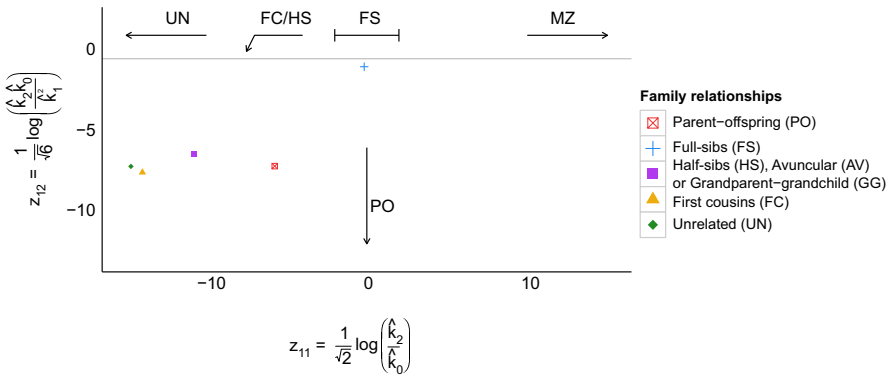

FIGURE 6 Ilr-coordinates $\boldsymbol{z}_1 = (z_{11}, z_{12})$ of the estimated Cotterman coefficients $(\hat{k}_0,\hat{k}_1,\hat{k}_2)$ for five pairs of individuals from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

presented throughout this article are made with the R software (R Core Team, 2015) using the R packages **ggplot2** (Wickham, 2009) and **ggtern** (Hamilton, 2015).

5.1 | IBS graphics

Figure 7 shows all IBS graphics for all pairs of the Maya population. In the $(\bar{x}, s)$-plot (Figure 7a), the points with the smallest standard deviation close to the grey curve are two PO pairs. The relationships of first and second degree are the points with a mean above 1. Note that some pairs of FC are mixed with UN pairs. Figure 7b (the (p_0, p_2)-plot) clearly separates the family relationships of first and second degree from the UN pairs. In the ternary diagram (Figure 7c), PO pairs are points on the opposite side of the vertex p_0, meaning that the p_0 is close to 0. The FS pair is the point closest to the vertex p_2, which has the largest p_2; the violet points represent the family relationships of second degree are separated from the green points representing UN pairs. In Figure 7d, the first ilr-coordinate (z_{11}) clearly discriminates first-degree relatives from UN pairs. Pairs with larger values for z_{11} are more likely to correspond to related individuals. PO pairs are extreme outliers because they have p_0 values close to 0 which increase the first coordinate of the corresponding log-ratio. The scatterplot of the log-ratios is seen to produce a larger degree of separation between FS and PO pairs, and between first-degree relationship pairs and all other pairs. The convex hulls for the simulated related pairs in Figure 7 are seen to enclose the sample estimates of the PO, FS, HS and FC pairs and so confirm the assigned relationships.

5.2 | IBD graphics

We estimated IBD probabilities for all pairs of the Maya population. All IBD graphics are shown in Figure 8. The $(\hat{k}_0, \hat{k}_1)$-plot (Figure 8a) separates the first, second and some pairs of third degree of relatedness. In the ternary diagram of $\hat{\boldsymbol{k}}$ (Figure 8b), it is easy to identify PO pairs at the vertex of $\hat{k}_1$, a FS pair close to the barycenter of the triangle and other family relationships of second degree on the opposite side of the $\hat{k}_2$ vertex. UN pairs are on the $k_0 - k_1$ edge and tend towards the k_0 vertex. Third-degree pairs are mixed with unrelated individuals. In the ilr-plot (Figure 8c), the pairs with a close family relationship tend to have larger values of z_{11}. The family relationships of the first degree (FS and PO) are located according to the directions indicated in Figure 6. The ilr-plot clearly separates out these FS and PO relationships from all other pairs. Notice that Figure 8a and b show only one pair with a second degree relationship (the violet point), whereas in Figure 8c, there are two visible violet pairs. The IBD graphics were also amplified with convex hulls of artificially generated related pairs to show the approximate expected

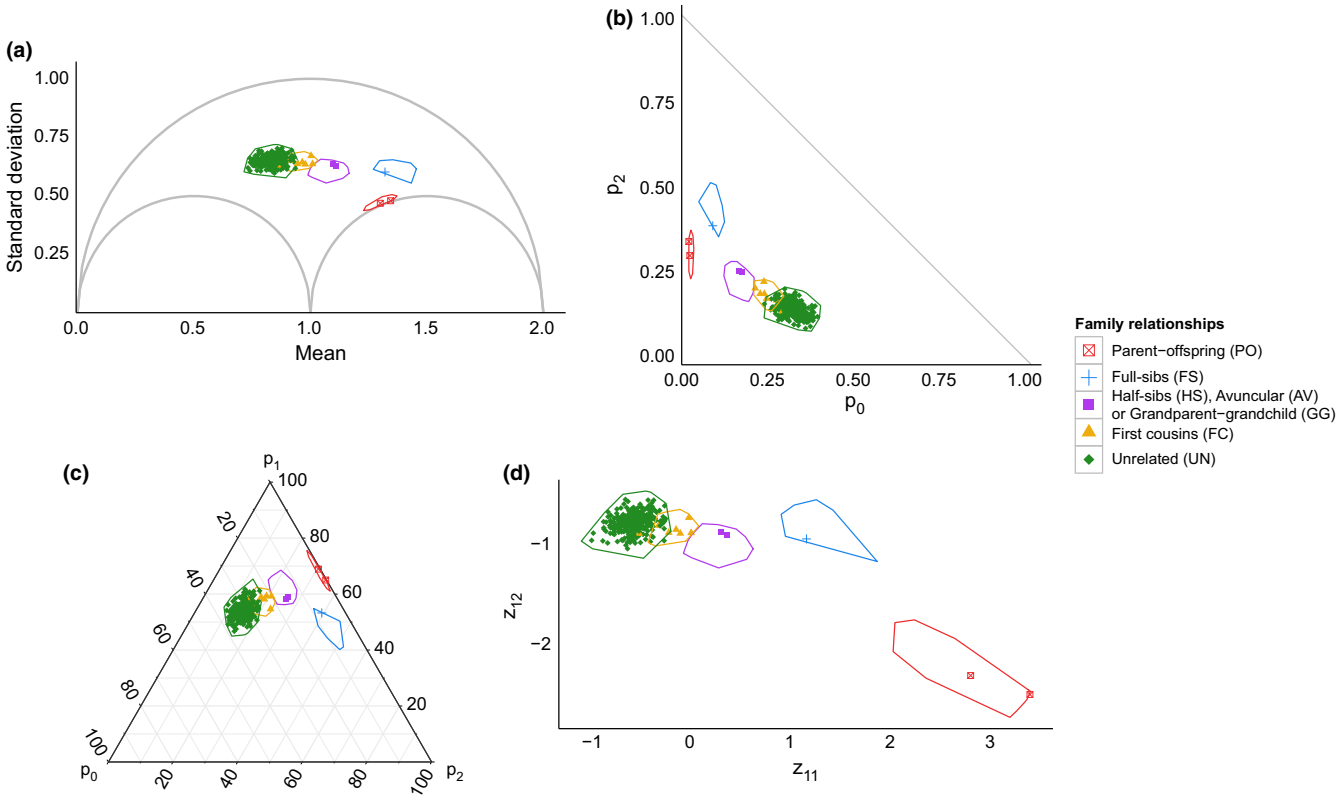

FIGURE 7 Identical by state (IBS) alleles for all the pairs of individuals from the Maya population. a. Plot of means versus standard deviations. b. (p_2, p_0)-plot. c. Ternary diagram. d. Ilr-coordinates: $\boldsymbol{z}_1 = (z_{11}, z_{12})$. The convex hulls are obtained by simulating artificial children from a subset of unrelated individuals from the Maya population and each hull is based on 80 PO, 48 FS, 120 second degree, 36 FC and 1256 UN artificial pairs

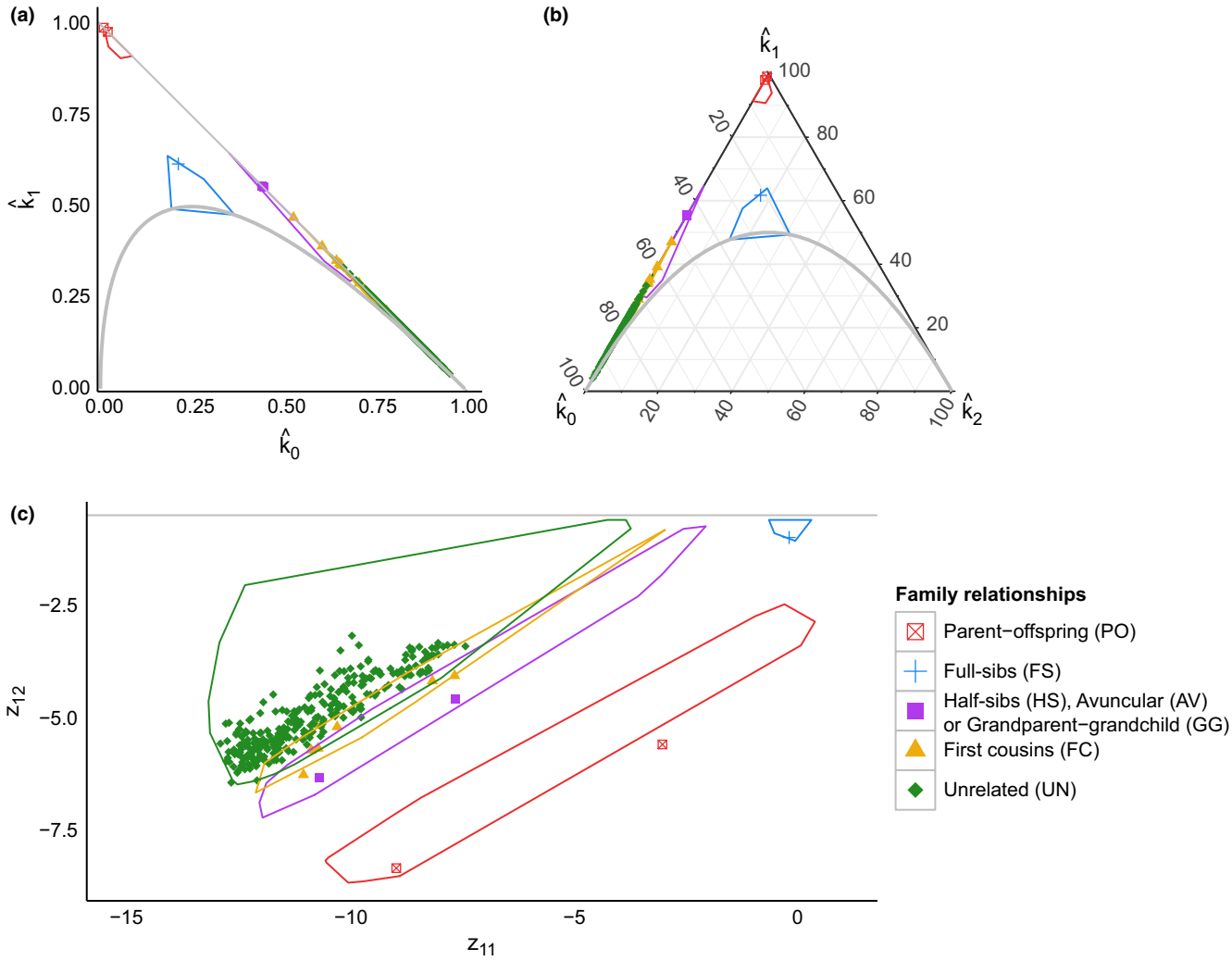

FIGURE 8 Identical by descent (IBD) alleles for all the pairs of individuals from the Maya population. a. $(\hat{k}_0, \hat{k}_1)$-plot. b. Ternary diagram. c. Ilr-coordinates: $\boldsymbol{z}_1 = (z_{11}, z_{12})$

positions for the different relationships. These hulls mainly confirm the assigned relationships. In ilr-coordinates, PO hulls do not capture all observed PO pairs (see Discussion).

6 | DISCUSSION

The main aim of this article was to review all graphical methods used in relatedness research. We have distinguished graphics based on IBS and IBD allele sharing. Plotting means versus standard deviations of the IBS counts allows us to detect monozygotic twins (MZ), parent-offspring (PO) and full-sibs (FS) pairs. However, higher degree relationships are more difficult to detect visually. The distances between unrelated and related pairs depend on the allele frequency distribution of the markers under study. The larger the heterozygosity in a population, the larger the distances between related and unrelated individuals are. A disadvantage of this mean-variance plot is that there are no fixed reference points for the standard relationships. Such reference points could eventually be found by calculating expectations of the mean and the variance of the IBS counts. These do depend on the allele frequency distribution and will therefore depend on the population that has been sampled, and on the distribution of the allele frequencies in that population. The (p_i, p_j)-plots allow easy detection of MZ pairs (or duplicated individuals) because they have p_2 values close to 1, and PO pairs have low values of p_0 and are also easy to detect. FS pairs are located between PO pairs and the pairs with large values of p_0. However, it remains hard to detect relationships of the second and third degree. The (p_i, p_j)-plots neither have a fixed reference position for the standard relationships. Moreover, as has been noted in Section 2, the Euclidean distance between two pairs in a (p_i, p_j)-plot is not invariant with respect to the chosen index (0, 1 or 2), for example, is not the same in a (p_0, p_1) and a (p_0, p_2)-plot. $(\hat{k}_i, \hat{k}_j)$-plots have, in comparison with (p_i, p_j)-plots, the advantage that fixed reference positions for the standard relationships exist, as given in Table 2. This is of great practical value when inferring relationships. Moreover, IBD plots are more reliable for classifying relationships because they show a larger degree of separation between the different relationships than their

IBS counterparts. This is clearly visible when one compares Figures 2 with 5a, 3 with 5b, 7b with 8a and 7c with 8b. However, the IBD-based $(\hat{k}_i, \hat{k}_j)$-plots suffer from the same problem as their IBS counterparts: the Euclidean distances between pairs (and reference points) depend on the index (0, 1 or 2) that is used.

We comment on some peculiarities of the HGDP-CEPH database analysed in the article. We found the high estimate of k_1 (0.27) in Table 1 for the reported UN pair to be not too unusual for Maya UN pairs, being the median of k_1 0.17 for UN pairs of this population. The relatively high k_1 estimates are probably to some extent due to inbreeding, as the South American populations had the largest medians of k_1 for UN pairs. However, for many other less inbred populations k_1 estimates of UN pairs had a large median too, in the range 0.1–0.2. We suggest the database could be affected by a certain degree of sample contamination, as this will increase the number of heterozygote calls, leading to overestimated IBD (Andoh, Sato, Sakamoto, Yoshida, & Ohtaki, 2010).

We continue with some remarks on the graphics from CoDA proposed in this article. We advocate the ternary diagram as an alternative for the (p_i, p_j)-plots because it clearly shows all three proportions simultaneously. MZ twins are close to the vertex p_2; PO pairs are easy to identify on the opposite side of the vertex p_0. FS pairs usually have large values of p_2 and are separated from unrelated pairs which have lower values of p_2. We also advocate the ternary diagram for IBD studies for the same reasons: all three estimated IBD probabilities are represented in one single graph with all three $\hat{k}_i$ axes. The theoretical IBD probabilities (Table 2) are easily added for use as reference points. The ternary diagram resolves the indeterminacy of the Euclidean distances between pairs due to the choice of axes observed above in (p_i, p_j) and (k_i, k_j) scatterplots. However, the interpretation of Euclidean distances in the ternary diagram remains a tricky issue, because the simplex is a constrained space. We note that the Euclidean distance is regarded inadequate for the comparison of compositions, and for this reason, we have considered isometric log-ratio coordinates of IBS and IBD probabilities. The Euclidean distances between the pairs in ilr-coordinates correspond to Aitchison distances between (p_0, p_1, p_2) (or (k_0, k_1, k_2)) compositions. The Aitchison distance is considered to be an adequate metric for representing compositions (Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015, Chapter 3). Plotting the ilr-coordinates of the IBS proportions is useful for detecting related individuals because usually unrelated individuals are concentrated in a cloud of points and most outlying individuals correspond to related pairs. Plotting the ilr-coordinates of the estimated Cotterman coefficients gives reference zones over the ilr space for the different relationships (Figure 6). Standard family relationships can be inferred depending on the values of z_{11} and z_{12}. UN pairs are mainly represented in the scatterplot of the isometric log-ratios of IBD probabilities by a central cloud of points around $(-10, -5)$ (Figure 8c) but also by points close to the upper limit of the second ilr-coordinate $(-\sqrt{(2/3)}\ln(2))$. A small change in the tolerance or the initial point of the maximization algorithm can greatly influence the final position of an UN pair. Both IBS- and IBD-based log-ratio plots show a strong discrimination of PO and FS pairs which typically appear as outliers in these plots. We also note that all inference on relationships in all presented graphical methods relies on the judgement of the analyst, who interprets distances between points in a graph. Depending on the sample size of the study, the number of markers used for the genotyping and the distributions of their allele frequencies, those distances will be subject to some degree of uncertainty which complicates graphical inference on relationships. By simulating artificial related pairs using the genotypes of unrelated pairs of the database, convex hulls for the expectation of the standard relationships can be obtained, which are conditional on the observed sample allele frequencies. These convex hulls assess the degree of uncertainty that can be expected for the different related pairs and are helpful for confirming putative relationships. In the present work, the convex hulls are limited by the fact that they assumed independent markers. This may explain why some related pairs are outlying with respect to their corresponding convex hulls. The accuracy of the convex hulls depends on the sample size, and in particular on the number of UN individuals in the sample from which it is generated. More accurate convex hulls may be obtained if linkage disequilibrium is taken into account and artificial pairs are generated by sampling from haplotypes instead of by sampling individual markers independently. Convex hulls of PO pairs in ilr-coordinates often do not capture all observed PO pairs (Figure 8). We suggest this might be due to a small sample size combined with numerical instability. The position of a PO pair in ilr-coordinates has a high variability and depends on the tolerance and initial point used in the maximization of the likelihood (Graffelman & Galván-Femenía, 2016). If the sample size is small, or the number of simulated pairs is small, the PO hull many not cover the full area compatible with PO pairs. It is worth remarking that PO and FS convex hulls do not intersect each other and do not overlap with the rest of the hulls, having a valuable discrimination power (Figures 7 and 8). We think the current simulated convex hulls are helpful to assess uncertainty but of limited value and see a clear need for methods of formal statistical inference on relationships by means of hypothesis testing and confidence regions (García-Magariños et al., 2015).

7 | SOFTWARE

R functions for making the graphics in this manuscript are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2532d.

ACKNOWLEDGEMENTS

This study was supported by grants MTM2015-65016-C2-2-R and R01 GM075091 from the United States National Institutes of Health (JG) and by MTM2015-65016-C2-1-R (2015-2017) of the Spanish Ministry of Economy and Competitiveness (IGF and CBV). Part of this work was presented at the 6th International Workshop on

Compositional Data Analysis in L'Escala, Spain in 2015. The authors thank Noah Rosenberg for making the HGDP-CEPH diversity panel used in this manuscript publicly available on the website of the Rosenberg Research lab at Stanford University (https://rosenberglab.stanford.edu/). We also thank the editors and all anonymous referees whose comments have helped us to improve the article.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of this article. I.G.F. analysed data.

DATA ACCESSIBILITY

The data analysed in this article are freely available on the web of the Rosenberg lab at Stanford University (https://rosenberglab.stanford.edu/diversity.html\#2002).

REFERENCES

Abecasis, G. R., Chemy, S. S., Cookson, W. O., & Cardon, L. R. (2001). GRR: graphical representation of relationship errors. *Bioinformatics*, *17*, 742–743.

Aitchison, J. (1986). *The statistical analysis of compositional data*. UK: Chapman & Hall.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology*, *32*, 271–275.

Andoh, M., Sato, Y., Sakamoto, H., Yoshida, T., & Ohtaki, M. (2010). Detection of inappropriate samples in association studies by and IBS-based method considering linkage disequilibrium between genetic markers. *Journal of Human Genetics*, *55*, 436–440.

Bérénos, C., Ellis, P. A., Pilkington, J. G., & Pemberton, J. M. (2014). Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular Ecology*, *23*, 3434–3451.

de Berg, M., van Kreveld, M., Overmars, M., & Schwarzkopf, O. (2000). *Computational geometry: algorithms and applications*. Berlin:Springer. 2–8.

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, *18*, 503–511.

Boehnke, M., & Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics*, *61*, 423–429.

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. *Nature Reviews, Genetics*, *6*, 333–340.

Cotterman, C. W. (1941). Relative and human genetic analysis. *The Scientific Monthly*, *53*, 227–234.

Croft, D. P., Hamilton, P. B., Darden, S. K., Jacoby, D. M. P., James, R., Bettaney, E. M., & Tyler, C. R. (2012). The role of relatedness in structuring the social network of a wild guppy population. *Oecologia*, *170*, 955–963.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*, 279–300.

Epstein, M. P., Duren, W. L., & Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, *67*, 1219–1231.

Foulkes, A. S. (2009). *Applied statistical genetics with R*. New York, NY: Springer.

García-Magariños, M., Egeland, T., López-de-Ullibarri, I., Hjort, N. L., & Salas, A. (2015). A parametric approach to kinship hypothesis testing using identity-by-descent parameters. *Statistical Applications in Genetics and Molecular Biology*, *14*(5), 465–479.

Gonder, M. K., Mitchell, M. W., Locatelli, S., Ghobrial, L., Pokempner, A. A., Sesink-Clee, P. R., … Hahn, B. H. (2015). The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evolutionary Biology*, *15*, 3.

Graffelman, J., & Galván-Femenía, I. (2016). An application of the isometric log-Ratio transformation for relatedness research. In: J. A. Martín-Fernández & S. Thió-Henestrosa (Eds.), *Compositional Data Analysis, Springer Proceedings in Mathematics & Statistics*, Vol *187*. (pp. 75–84). Cham: Springer International Publishing.

Hamilton, N. (2015). ggtern: An extension to `ggplot2', for the creation of ternary diagrams. R package version 2.1.1. http://CRAN.R-project.org/package=ggtern

Hansen, M. M., Nielsen, E. E., & Mensberg, K. L. D. (1997). The problem of sampling families rather than populations: relatedness among individuals in samples of juvenilebrown trout Salmo trutta L. *Molecular Ecology*, *6*, 469–474.

Loughnan, S. R., Smith-Keune, C., Jerry, D. R., Beheregaray, L. B., & Robinson, N. A. (2016). Genetic diversity and relatedness estimates for captive barramundi (Lates calcarifer, Bloch) broodstock informs efforts to form a base population for selective breeding. *Aquaculture Research*, *47*, 3570–3584.

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

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*, 1027–1028.

Nembot-Simo, A., Graham, J., & McNeney, B. (2013). CrypticIBDcheck: an R package for checking cryptic relatedness in nominally unrelated individuals. *Source Code for Biology and Medicine*, *8*, 5.

Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M., & Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, *173*, 483–496.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: theory and applications*. Chichester: Wiley.

Pawlowsky-Glahn, V., Egozcue, J.J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester, United Kingdom: Wiley & Sons.

Pemberton, T. J., Wang, C., Li, J. Z., & Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap phase III. *The American Journal of Human Genetics*, *87*, 457–464.

R Core Team (2015). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, *70*, 841–847.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*, 2381–2385.

Snyder-Mackler, N., Alberts, S. C., & Bergman, T. J. (2014). The sociogenetics of a complex society: female gelada relatedness patterns mirror association patterns in amultilevel society. *Molecular Ecology*, *23*, 6179–6191.

Spencer, P. B. S., Hampton, J. O., Pacioni, C., Kennedy, M. S., Saalfeld, K., Rose, K., & Woolnough, A. P. (2015). Genetic relationships within social groups influence the application of the Judas technique: a case study with wild dromedary camels. *The Journal of Wildlife Management*, *79*, 102–111.

Stanley, R. P. (1997). *Enumerative combinatorics*. Vol. *I*. New York, NY: Cambridge University Press.

Sun, L. (2012). Statistical human genetics: methods and protocols. Chapter 2, 25–46.

Thompson, E. A. (1991). Estimation of relationships from genetic data. In: C. R. Rao & R. Chakraborty (Eds.), *Handbook of Statistics*, Vol *8*. (pp. 255–269). Amsterdam: Elsevier Science.

Thompson, E. A. (2000). Statistical inference from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. *6*. Chapter 3, 29–46.

Wagner, A. P., Creel, S., & Kalinowski, S. T. (2006). Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, *97*(5), 336–345.

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews*, *Genetics*, *7*, 771–780.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

How to cite this article: Galván-Femenía I, Graffelman J, Barceló-i-Vidal C. Graphics for relatedness research. *Mol Ecol Resour*. 2017;17:1271–1282. https://doi.org/10.1111/1755-0998.12674