TYPE Original Research
PUBLISHED 07 February 2024
DOI 10.3389/fgene.2024.1226228

OPEN ACCESS

EDITED BY
Ruzong Fan,
Georgetown University Medical Center,
United States

REVIEWED BY
Xiling Liu,
Academy of Forensic Science, China
Daniel E. Weeks,
University of Pittsburgh, United States

*CORRESPONDENCE
Shujin Li,
✉ shujinli@hebmu.edu.cn

RECEIVED 21 May 2023
ACCEPTED 10 January 2024
PUBLISHED 07 February 2024

CITATION
Ma G, Wang Q, Cong B and Li S (2024), An approach to unified formulae for likelihood ratio calculation in pairwise kinship analysis.
Front. Genet. 15:1226228.
doi: 10.3389/fgene.2024.1226228

An approach to unified formulae for likelihood ratio calculation in pairwise kinship analysis

Guanju Ma[1], Qian Wang[1], Bin Cong[1,2] and Shujin Li[1]*

[1]Hebei Key Laboratory of Forensic Medicine, Research Unit of Digestive Tract Microecosystem Pharmacology and Toxicology, Chinese Academy of Medical Sciences, College of Forensic Medicine, Hebei Medical University, Shijiazhuang, China, [2]Hainan Tropical Forensic Medicine Academician Workstation, Haikou, China

Introduction: The likelihood ratio (LR) can be an efficient means of distinguishing various relationships in forensic fields. However, traditional list-based methods for derivation and presentation of LRs in distant or complex relationships hinder code editing and software programming. This paper proposes an approach for a unified formula for LRs, in which differences in participants' genotype combinations can be ignored for specific identification. This formula could reduce the difficulty of by-hand coding, as well as running time of large-sample-size simulation.

Methods: The approach is first applied to a problem of kinship identification in which at least one of the participants is alleged to be inbred. This can be divided into two parts: i) the probability of different identical by descent (IBD) states according to the alleged kinship; and ii) the ratio of the probability that specific genotype combination can be detected assuming the alleged kinship exists between the two participants to the similar probability assuming that they are unrelated, for each state. For the probability, there are usually recognized results for common identification purposes. For the ratio, subscript letters representing IBD alleles of individual A's alleles are used to eliminate differences in genotype combinations between the two individuals and to obtain a unified formula for the ratio in each state. The unification is further simplified for identification cases in which it is alleged that both of the participants are outbred. Verification is performed to show that the results obtained with the unified and list-form formulae are equivalent.

Results: A series of unified formulae are derived for different identification purposes, based on which an R package named KINSIMU has been developed and evaluated for use in large-size simulations for kinship analysis. Comparison between the package with two existing tools indicated that the unified approach presented here is more convenient and time-saving with respect to the coding process for computer applications compared with the list-based approach, despite appearing more complicated. Moreover, the method of derivation could be extended to other identification problems, such as those with different hypothesis sets or those involving multiple individuals.

Conclusion: The unified approach of LR calculation can be beneficial in kinship identification field.

KEYWORDS

pairwise kinship testing, likelihood ratio, unified formulae, divide-and-conquer method, R package construction

1 Introduction

In kinship identification or forensic genealogy, at least two alleged relationships between participants need to be confirmed or excluded using the genetic information available. However, confirmation or exclusion can be difficult, as the alleged relationships can be much more complicated than those in well-researched parentage cases. With the development of forensic databases and sequencing technologies, increasingly complex kinship relationships need to be studied in forensic genetics and especially forensic genealogy. Several approaches can be used for identification of these complicated relationships, including the likelihood ratio (LR) method or simple application of an identity by state (IBS) score. For distant kinship identification in forensic genealogy, it has been claimed that "the traditional LR approach as a single source of classification is as good as, and in some cases even better than, the alternative approaches" (Kling and Tillmar, 2019).

In order to conduct complex kinship identification, it is essential to evaluate the practicality of specific panels or marker types (Mo et al., 2018; Li et al., 2019; Liu et al., 2020; Wu et al., 2021; Zhao et al., 2021; Du et al., 2023). Such assessments often necessitate pedigree investigations based on real cases or computer simulations. As kinship complexity increases, so too does the difficulty of investigating real cases: either the target cases are rare in the population (such as inbreeding cases); or the confirmation of the participants is challenging (for example, to confirm the relationship between a pair of first cousins, their common grandparents, and their parents who are full-siblings to each other should also be detected), making simulation a more realistic method. There are several application scenarios for simulation methods in this field, including i) assessing the sensitivity and specificity of specific markers in identification (Phillips et al., 2012; Mo et al., 2018; Liu et al., 2020; Du et al., 2023); ii) comparing different parameters in the same identification based on the same panel, e.g., LR calculated based on length information vs LR calculated based on length + sequence information (Staadig and Tillmar, 2019; Liu et al., 2020); iii) estimating the required number of markers in specific identification using curve fitting, through a quantity of simulation based on different subsets of current loci combinations, when the current combination cannot meet the identification requirements (Mo et al., 2016; Du et al., 2023).

Certain obstacles may be encountered when using the LR method in large-sample-size simulation, owing to the presentation of calculation results and the coding logic based on such results in various studies. In such studies, LR are presented as the listings of all possible genotype combinations of participants, followed by the application of different formulae in different cases. This necessitates dividing the calculation of LR into multiple types and using multi-layer logical comparison functions, such as "if (if (. . .))", during the coding process. Although current tools can eliminate the need for users to carry out the complex coding processes mentioned earlier, in research situations where such tools are not available (such as rare complex kinship cases), users still need to compute and simulate for themselves. Thus, it would be beneficial to establish a unified formula for distant or complex kinship identification that disregards the participants' genotype combinations and avoids logical comparison functions as much as possible in the coding process, mitigating the difficulty of manual coding and the time required. Egeland *et al.* devised a unified formula for pairwise non-inbred kinship testing in Egeland et al. (2017), which was used in *Familias* 3.0 (Egeland et al., 2016).

This paper presents an alternative approach that delivers equivalent results to Egeland's formula, but can be easily extended to inbreeding relationships, owing to its concise derivation methodology. A package named *KINSIMU* containing a series of newly defined functions for the R platform is provided, based on the unified formula, and can be used for large-sample-size simulation/calculation in specific kinship analysis based on independent genetic markers.

2 Methods and results

2.1 General setting

In pairwise kinship identification, there is no detectable genetic information other than the genotype combination of the two participants, who are labeled as individual A and B, respectively, in this paper. Suppose that the detected genotypes of them are ***ab*** and ***cd***, respectively, where the four alleles can be identical to each other or not. An individual is called "inbred" if his/her parents are biological relatives, or "outbred" otherwise. Mutation is only considered when constructing the paternity index calculation function in the construction of *KINSIMU* package and not in the inference process in this section, which will be discussed in section 3.

In the derivation, some symbols with specific meanings will appear, including:

I) In this article, the identity symbol "≡" is utilized to denote the identity between particular alleles or genotypes, and this will occur in two situations:
   i) It is preceded by a capital letter and followed by two lowercase letters, indicating the event that "the two alleles of the individual represented by the capital letter are those represented by these two lowercase letters";
   ii) Letters on both side of it are lowercase, meaning that the alleles or genotypes on both sides are identical to each other;

It should be noted that the "identity" status represented by the symbol can arise from genetic inheritance or by random occurrence.

II) The symbol $\boldsymbol{p}$ with subscript lowercase letter such as "p_c" denotes the frequency of allele represented by the subscript letter;
III) The symbol $\mathbb{1}$ with two subscript lowercase letters equals to 1 if the two alleles represented by the two letters are identical and to 0 otherwise. For instance, $\mathbb{1}_{ac}$ equals to 1 only if $a \equiv c$.
IV) Lowercase letters with subscript "$\boldsymbol{I}$" such as "a_I" means the allele identical by descent (IBD) to the corresponding allele. If there is no other information, the probability of an IBD allele being identical to a detected allele equals to the corresponding $\mathbb{1}$ parameter, e.g., $\Pr(a_I \equiv c) = \mathbb{1}_{ac}$;
V) The symbol "$\boldsymbol{x_I}$" and "$\boldsymbol{y_I}$" represent the alleles not IBD to none of the detected alleles of participants other than individual B, where "x_I" is unrelated to "y_I". Without

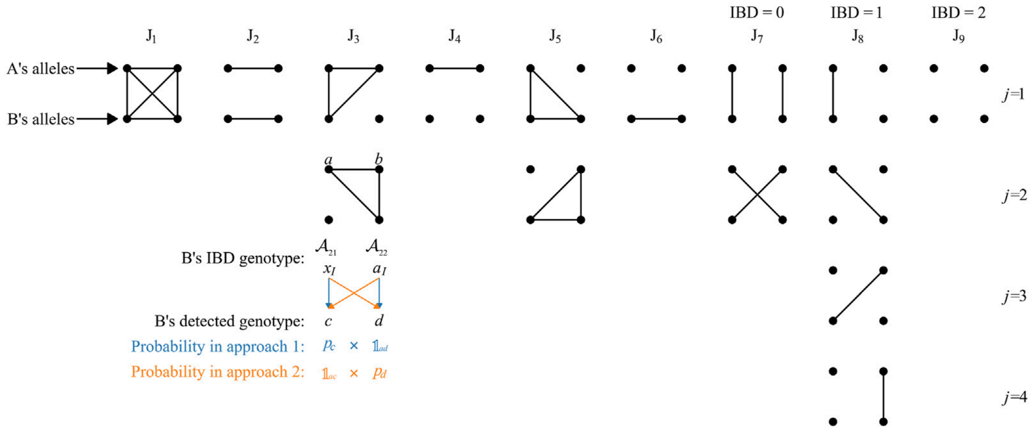

FIGURE 1
Nine Jacquard states considering inbreeding. The figure is modified from Figure 1 of Brustad et al. (2021b). Each group of 2 × 2 dots represents a pair of participants, each row of two dots represents the two alleles of an individual, and IBD alleles are connected by lines. The states J_9, J_8, and J_7 do not involve inbreeding and are sometimes denoted IBD = 0, 1, and 2, respectively. The third group of dots in J_8 column was mistakenly drawn as the second in J_7 column in Brustad et al. (2021b); the mistake is corrected in this figure.

other information, the probability of these alleles being a specific allele in individual B's genotype equals to the corresponding p parameter, e.g., $\Pr(x_I \equiv c) = p_c$.

VI) The symbol $\boldsymbol{d}$ with a subscript, which is formed by a capital letter stand for a specific individual followed by a lowercase letter denotes specific allele, such as $\boldsymbol{d}_{Ac}$, denotes the dosage of the corresponding allele in the genotype of the corresponding individual, e.g., $d_{Ac} = \mathbb{1}_{ac} + \mathbb{1}_{bc}$ if the genotype of individual A is *ab*.

2.2 Overall deduction of unified LR formulae in pairwise identification

As discussed in multiple articles (Gjertson et al., 2007; Egeland and Sheehan, 2008; Skare et al., 2009; Egeland et al., 2017; Slooten, 2020), in pairwise kinship analysis, the probabilities of different relationships existing between participants can be evaluated conditional on the genetic evidence, i.e., $\Pr(H \mid E)$, with terms H and E denoting the alleged hypothesis and the observed genetic evidence, respectively. It is difficult to calculate such probabilities directly; however, the ratio of different incompatible hypotheses can be computed according to Bayes' rule:

$$\frac{\Pr\left(H_p \middle| E\right)}{\Pr(H_d \mid E)} = \frac{\Pr\left(H_p\right)}{\Pr\ (H_d)} \times \frac{\Pr\left(E \middle| H_p\right)}{\Pr(E \mid H_d)} \tag{1}$$

In the above equation, terms H_p (Plaintiff's Hypothesis) and H_d (Defendant's hypothesis) denote the two hypotheses being compared. For all of the inference process in this paper, we define H_d as the hypothesis that "the two individuals are both outbred and unrelated to each other". Meanwhile, the definition of H_p varies depending on the scenario, such as "individual A is outbred and the biological father of outbred individual B" in paternity testing. In the kinship identification field, the LR is defined as the ratio of conditional probabilities in the formula:

$$LR = \frac{\Pr\left(E \middle| H_p\right)}{\Pr(E \mid H_d)} \tag{2}$$

The ratio $\Pr(H_p)/\Pr(H_d)$, or π_1/π_0 in several articles (Gjertson et al., 2007; Egeland and Sheehan, 2008), is called the prior odds, representing how much more likely H_p is to be true than H_d without the genetic data E. If the prior odds are considered to be 1, i.e., the two hypotheses are of equal probabilities without the genetic information, which is the most commonly used assumption in forensic practice (Slooten, 2020), then LR can represent the ratio of likelihood initially required in Eq. 1.

Further derivation can be made from Eq. 2 considering the fact that H_p can be further divided into several multiple exclusive states according to whether the four alleles of individual A and B are IBD to each other, such as the nine Jacquard states in inbreeding identification (Jacquard, 1972) ($J_1 \rightarrow J_9$, see Figure 1 of (Brustad et al., 2021b), which is reproduced with modification as Figure 1 in this work) or the three IBD states commonly be used in non-inbred ones (Egeland et al., 2017). However, there can be only one such state, i.e., J_9 or $IBD = 0$, under the H_d set, as in section 2.1. Thus, LR can be calculated as Eq. 3.

$$LR = \begin{cases} \dfrac{\sum\limits_{i=1}^{9}\left[\Pr\left(J_i \middle| H_p\right) \times \Pr\left(E \middle| J_i, H_p\right)\right]}{\Pr(E \mid J_9, H_d)} \\ \dfrac{\sum\limits_{i=0}^{2}\left[\Pr\left(IBD = i \middle| H_p\right) \times \Pr\left(E \middle| IBD = i, H_p\right)\right]}{\Pr(E \mid IBD = 0, H_d)} \end{cases} \tag{3}$$

If a specific state is set, the probability that E happens is fixed, irrespective of the hypothesis. Thus, LR can be calculated with Eq. 4, where Δ_i denotes the ith Jacquard coefficient (Brustad et al., 2021b) under H_p, equal to the probability that J_i happens between two individuals with specific relationship in the absence of their genetic information, i.e., $\Pr(J_i \mid H_p)$; and κ_i represents the IBD coefficient (Egeland et al., 2017), equal to the similar probability under non-inbred assumption, $\Pr(IBD = i \mid H_p)$.

TABLE 1 Example of verification of $\Pr(E|J_i)/\Pr(E|J_9)$ unification.

		Results				Results	
J_i	Unified ratio	Unified[a]	Listed[b]	J_i	Unified ratio	Unified[a]	Listed[b]
J_1	$\frac{\mathbb{1}_{ab}\mathbb{1}_{ac}\mathbb{1}_{cd}}{p_a^3}$	$\frac{1\times1\times1}{p_a^3}=\frac{1}{p_i^3}$	$\frac{p_i}{p_i^4}$	J_2	$\frac{\mathbb{1}_{ab}\mathbb{1}_{cd}}{p_a p_c}$	$\frac{1\times1}{p_i\times p_i}=\frac{1}{p_i^2}$	$\frac{p_i^2}{p_i^4}$
J_3	$\frac{\mathbb{1}_{ab}(\mathbb{1}_{ac}+\mathbb{1}_{ad})}{2p_a^2}$	$\frac{1\times(1+1)}{2\times p_i^2}=\frac{1}{p_i^2}$	$\frac{p_i^2}{p_i^4}$	J_4	$\frac{\mathbb{1}_{ab}}{p_a}$	$\frac{1}{p_i}$	$\frac{p_i^3}{p_i^4}$
J_5	$\frac{\mathbb{1}_{cd}(\mathbb{1}_{ac}+\mathbb{1}_{bc})}{2p_c^2}$	$\frac{1\times(1+1)}{2\times p_i^2}=\frac{1}{p_i^2}$	$\frac{p_i^2}{p_i^4}$	J_6	$\frac{\mathbb{1}_{cd}}{p_c}$	$\frac{1}{p_i}$	$\frac{p_i^3}{p_i^4}$
J_7	$\frac{\mathbb{1}_{ac}\mathbb{1}_{bd}+\mathbb{1}_{ad}\mathbb{1}_{bc}}{2p_c p_d}$	$\frac{1\times1+1\times1}{2\times p_i\times p_i}=\frac{1}{p_i^2}$	$\frac{p_i^2}{p_i^4}$	J_8	$\frac{d_c p_d+d_d p_c}{4p_c p_d}$	$\frac{2\times p_i+2\times p_i}{4\times p_i\times p_i}=\frac{1}{p_i}$	$\frac{p_i^3}{p_i^4}$
J_9	1	1	$\frac{p_i^4}{p_i^4}$				

[a]Notes:* $\mathbb{1}_{ab}=\mathbb{1}_{ac}=\mathbb{1}_{ad}=\mathbb{1}_{bc}=\mathbb{1}_{bd}=\mathbb{1}_{cd}=1$, $d_c=d_d=2$ and $p_a=p_c=p_d=p_i$ when both individual A and B are homozygotes *ii*.

[b]Results calculated according to the first row of Table 1 in Brustad et al. (2021b), in which the alleles of the two individuals were "*a*" and have been adjusted to "*i*" to avoid confusion with the parameters in the unified formulae.

$$LR=\begin{cases}\sum_{i=1}^{9}\left[\Delta_i\times\frac{\Pr(E\,|\,J_i)}{\Pr(E\,|\,J_9)}\right]\\ \sum_{i=0}^{2}\left[\kappa_i\times\frac{\Pr(E\,|\,IBD=i)}{\Pr(E\,|\,IBD=0)}\right]\end{cases}\tag{4}$$

Therefore, the unification of LR can be divided into two problems: the calculation of the ratio of conditional probabilities under each state according to the participants' genotypes (as discussed in section 2.3); and the Δ/κ distribution according to the alleged relationship (Pinto et al., 2010), which remains unchanged if the identification purpose is set (e.g., $\kappa=\{1/4,1/2,1/4\}$ for full-sibling identification) and can be obtained from the pedigree tree using existing tools (Vigeland, 2022). Thus, it is not necessary to infer the distribution of the two types of coefficients per case if there is a recognized result for the target identification.

2.3 Unification of the ratio of conditional probabilities under each state

2.3.1 When at least one of the two individuals is inbred

LR calculation when at least one of the two individuals is inbred has been well discussed in previous publications (Jacquard, 1972; Brustad et al., 2021a; b), and a method has been developed considering the Jacquard states. The probability $\Pr(E\,|\,J_i)$ can be listed in a 9×9 table, as in Table 1 in (Brustad et al., 2021b). Herein, we improve the method by eliminating the difference between the genotype combinations of the two individuals, i.e., "E"; thus, the probability ratio $\Pr(E\,|\,J_i)/\Pr(E\,|\,J_9)$ can be listed in a 1×9 table, and LR can be calculated with a unified formula for specific identification. The ratio can be further deducted based on the fact that the event E can be understood as the event that the following two events happened simultaneously: $A\equiv ab$ and $B\equiv cd$:

$$\begin{aligned}\frac{\Pr(E\,|\,J_i)}{\Pr(E\,|\,J_9)}&=\frac{\Pr(A\equiv ab,B\equiv cd\,|\,J_i)}{\Pr(A\equiv ab,B\equiv cd\,|\,J_9)}\\&=\frac{\Pr(A\equiv ab\,|\,J_i)}{\Pr(A\equiv ab\,|\,J_9)}\times\frac{\Pr(B\equiv cd\,|\,A\equiv ab,J_i)}{\Pr(B\equiv cd\,|\,A\equiv ab,J_9)}\end{aligned}\tag{5}$$

The former ratio in Eq. 5 can be calculated differently depending on whether the two alleles of individual A are IBD. If the two alleles of individual A are not IBD to each other and no other information is considered, the probability that the individual's genotype is *ab* should equal the frequency of the genotype, i.e., $\Pr(A\equiv ab\,|\,J_i)/\Pr(A\equiv ab\,|\,J_9)=1$ when $i\in[5,9]$. Otherwise, if the two alleles of individual A are IBD to each other, and mutation is not considered, the probability of individual A being a specific genotype equals p_a when $a\equiv b$ and 0 otherwise, i.e., $\Pr(A\equiv ab\,|\,J_i)=\mathbb{1}_{ab}p_a$ when $i\in[1,4]$. Thus, the former ratio of Eq. 5 can be calculated as Eq. 6 when $i\in[1,4]$.

$$\frac{\Pr(A\equiv ab\,|\,J_{1\to4})}{\Pr(A\equiv ab\,|\,J_9)}=\left\{\begin{matrix}\frac{\mathbb{1}_{ab}p_a}{p_a^2} & , & a\equiv b\\ 0 & , & a\neq b\end{matrix}\right\}=\frac{\mathbb{1}_{ab}}{p_a}\tag{6}$$

In summary, if consider all Jacquard states, the former ratio of Eq. 5 can be calculated as Eq. 7.

$$\frac{\Pr(A\equiv ab\,|\,J_i)}{\Pr(A\equiv ab\,|\,J_9)}=\begin{cases}\frac{\mathbb{1}_{ab}}{p_a} & , \quad i\in[1,4]\\ 1 & , \quad i\in[5,9]\end{cases}\tag{7}$$

The latter ratio in Eq. 5 can be derived with individual B's genotype from the perspective of individual A's IBD alleles (defined as "IBD genotype" in this work). Under each Jacquard state, the IBD genotype of individual B should be set, there is only one possible type when i = 1, 2, 4, 6, or 9, e.g., it must be $a_I a_I$ under J_1; otherwise, there can be multiple possible types under a same state, e.g., $a_I x_I$ or $x_I a_I$ under J_3 (see Figure 1). If we set "$\mathcal{G}_j$" as individual B's *j*th possible IBD genotypes under a specific Jacquard state. It can be derived according to the Law of Total Probability that,

$$\begin{aligned}\Pr(B\equiv cd\,|\,A\equiv ab,J_i)=\sum_j\Big[&\Pr\big(B\equiv\mathcal{G}_j\,|\,A\equiv ab,J_i\big)\\&\times\Pr\big(B\equiv cd\,\big|\,B\equiv\mathcal{G}_j,A\equiv ab,J_i\big)\Big]\end{aligned}\tag{8}$$

where $\Pr(B\equiv\mathcal{G}_j\,|\,A\equiv ab,J_i)$ denotes the probability that individual B's IBD genotype is $\mathcal{G}_j$ under the corresponding state before detecting his/her actual genotype. It can be seen that, if there are multiple possible types of such IBD genotype, this probability should be equivalent for each type owing to the absence of other genetic information, i.e., Eq. 8 can be further derived as Eq. 9.

$$\Pr\left(B \equiv \mathcal{G}_j \mid A \equiv ab, J_i\right) = \begin{cases} 1 & , \quad i \in \{1,2,4,6,9\} \\ \frac{1}{2} & , \quad i \in \{3,5,7\} \\ \frac{1}{4} & , \quad i = 8 \end{cases} \tag{9}$$

And the probability $\Pr(B \equiv cd \mid B \equiv \mathcal{G}_j, A \equiv ab, J_i)$ stands for the probability that individual B's actual genotype is detected as cd given that his/her IBD genotype is $\mathcal{G}_j$. If the IBD genotype is set, this probability should be fixed regardless of other conditions, i.e.,

$$\Pr\left(B \equiv cd \,\middle|\, B \equiv \mathcal{G}_j, A \equiv ab, J_i\right) = \Pr\left(\mathcal{G}_j \equiv cd\right) \tag{10}$$

If defined $\mathcal{A}_{j1}$ and $\mathcal{A}_{j2}$ as the first and the second allele of $\mathcal{G}_j$, there are two exclusive approach for IBD genotype $\mathcal{G}_j$ to be detected as cd when $c \neq d$ and not considering mutation: **i)** $\mathcal{A}_{j_1} \equiv c,\ \mathcal{A}_{j_2} \equiv d$, and **ii)** $\mathcal{A}_{j_1} \equiv d,\ \mathcal{A}_{j_2} \equiv c$. Thus, according to the Rule of total probability,

$$\begin{aligned}\Pr\left(\mathcal{G}_j \equiv cd\right) = \Pr\left(\mathcal{A}_{j_1} \equiv c\right) \times \Pr\left(\mathcal{A}_{j_2} \equiv d \,\middle|\, \mathcal{A}_{j_1} \equiv c\right) \\ + \Pr\left(\mathcal{A}_{j_1} \equiv d\right) \times \Pr\left(\mathcal{A}_{j_2} \equiv c \,\middle|\, \mathcal{A}_{j_1} \equiv d\right)\end{aligned} \tag{11}$$

As discussed in section 2.1, the probabilities $\Pr(\mathcal{A}_{j_1} \equiv c)$ and $\Pr(\mathcal{A}_{j_1} \equiv d)$ equal the corresponding $\mathbb{1}$ parameter if $\mathcal{A}_{j_1}$ is IBD to one of individual A's alleles, or the corresponding $\boldsymbol{p}$ parameter when $\mathcal{A}_{j_1} \equiv x_I$. The probabilities $\Pr(\mathcal{A}_{j_2} \equiv d \mid \mathcal{A}_{j_1} \equiv c)$ and $\Pr(\mathcal{A}_{j_2} \equiv c \mid \mathcal{A}_{j_1} \equiv d)$ equal $\mathbb{1}_{cd}$ if the two alleles $\mathcal{A}_{j_1}$ and $\mathcal{A}_{j_1}$ are IBD to each other (under $J_{1,2,5,6}$), otherwise, the probability can be calculated similarly to the former two probabilities, i.e.,

$$\Pr\left(\mathcal{A}_{j_2} \equiv d \,\middle|\, \mathcal{A}_{j_1} \equiv c\right) = \begin{cases} \mathbb{1}_{cd} & , \quad J_1, J_2, J_5, J_6 \\ \Pr\left(\mathcal{A}_{j_2} \equiv d\right) & , \quad other \end{cases} \tag{12}$$

When $c \equiv d$, the calculation result obtained with the above equation would be double of the actual probability, and the latter ratio in Eq. 5 would remain unchanged.

Taking the situation J_3 as an example, under this state, one of individual B's allele is IBD to both of individual A's while the other one is not IBD to none of A's alleles. As mentioned above, there can be two possible $\mathcal{G}_j$, i.e., $a_I x_I$ (the first row of the third column in Figure 1) or $x_I a_I$ (the second row). The two possible approaches of the second type of IBD genotype to be detected as cd are listed with blue and orange colors in Figure 1 that,

$$\Pr\left(\mathcal{G}_j \equiv cd\right) = \Pr\left(x_I a_I \equiv cd\right) = p_c \mathbb{1}_{ad} + \mathbb{1}_{ac} p_d \tag{13}$$

Using a similar derivation, the same result can be achieved for the first type of IBD genotype, i.e., $\Pr(E \mid J_3) = p_c \mathbb{1}_{ad} + \mathbb{1}_{ac} p_d$. And for J_9, $\Pr(\mathcal{G}_j \equiv cd) = 2 p_c p_d$. Thus,

$$\begin{aligned}\frac{\Pr\ (E \mid J_3)}{\Pr\ (E \mid J_9)} &= \frac{\mathbb{1}_{ab}}{p_a} \times \left(\frac{\mathbb{1}_{ad}}{2p_d} + \frac{\mathbb{1}_{ac}}{2p_c}\right) = \frac{\mathbb{1}_{ab}}{p_a} \times \left(\frac{\mathbb{1}_{ad}}{2p_a} + \frac{\mathbb{1}_{ac}}{2p_a}\right) \\ &= \frac{\mathbb{1}_{ab}\left(\mathbb{1}_{ac} + \mathbb{1}_{ad}\right)}{2p_a^2}\end{aligned} \tag{14}$$

The detailed inference process for all states is shown in Section 1 of File S1 in Supplementary Materials, and the LR calculation in such cases can be unified as Eq. 15:

$$\begin{aligned} LR = & \\ & + \frac{\Delta_5 \mathbb{1}_{cd}\left(\mathbb{1}_{ac} + \mathbb{1}_{bc}\right)}{2p_c^2} + \frac{\Delta_6 \mathbb{1}_{cd}}{p_c} + \frac{\Delta_7\left(\mathbb{1}_{ac}\mathbb{1}_{bd} + \mathbb{1}_{ad}\mathbb{1}_{bc}\right)}{2p_c p_d} \\ & + \frac{\Delta_8\left[\left(\mathbb{1}_{ac} + \mathbb{1}_{bc}\right)p_d + \left(\mathbb{1}_{ad} + \mathbb{1}_{bd}\right)p_c\right]}{4p_c p_d} + \Delta_9 \end{aligned} \tag{15}$$

Therefore, 9 elements (p_a, p_c, p_d, $\mathbb{1}_{ab}$, $\mathbb{1}_{cd}$, $\mathbb{1}_{ac}$, $\mathbb{1}_{ad}$, $\mathbb{1}_{bc}$, and $\mathbb{1}_{bd}$) need to be calculated in the determination of LR, which can be calculated uniformly based on the genotype data of the two individuals and brought into a unified formula.

2.3.2 When both individual A and B are outbred

If an individual is outbred, his or her two alleles should not be IBD alleles, i.e., there is no possibility of one of Jacquard states J_1 to J_6 occurring when the two participants are both outbred. As discussed above, if the genotype $\mathcal{G}_j$ is set, the probability that it is cd is independent of the hypothesis; thus, the calculation of the conditional probabilities' ratios under J_7–J_9, i.e., the ratios when $IBD = 2$, 1, or 0, is the same, and the LR calculation can be simplified as follows:

$$LR = \frac{\kappa_2\left(\mathbb{1}_{ac}\mathbb{1}_{bd} + \mathbb{1}_{ad}\mathbb{1}_{bc}\right)}{2p_c p_d} + \frac{\kappa_1\left[\left(\mathbb{1}_{ac} + \mathbb{1}_{bc}\right)p_d + \left(\mathbb{1}_{ad} + \mathbb{1}_{bd}\right)p_c\right]}{4p_c p_d} + \kappa_0 \tag{16}$$

Here, the number of elements needed in the unified calculation would be reduced to 6 (the calculation of p_a, $\mathbb{1}_{ab}$ and $\mathbb{1}_{cd}$ are not needed). Furthermore, for identification where $\kappa_2 = 0$, if we define d_{Ac} and d_{Ad} as section 2.1, Eq. 16 can be further simplified as follows and the number of elements needed would be reduced to 4 (d_{Ac}, d_{Ad}, p_c and p_d):

$$LR = \frac{\kappa_1}{4}\left(\frac{d_{Ac}}{p_c} + \frac{d_{Ad}}{p_d}\right) + \kappa_0 \tag{17}$$

2.3.3 LR when H_d set as section 2.1 is ruled out in advance

Unified formulae can be applied even if H_d, i.e., "both individual A and B are outbred and unrelated to each other" is ruled out in advance and the identification takes place between two hypotheses H_p and H_d', under each of which the two individuals are relatives. For that case, LR can be calculated as the ratio of the two LRs with $\Pr(E \mid H_p)$ and $\Pr(E \mid H_d')$ being the numerators, i.e.,

$$LR_{H_p,H_d'}(E) = \frac{LR_{H_p,H_d}(E)}{LR_{H_d',H_d}(E)} \tag{18}$$

For example, in father–daughter incest cases, if the mother–offspring relationship is confirmed and the genetic information of the alleged father is unavailable, an LR called the incest index (II) (Wenk et al., 1994; Wenk, 2007) can be calculated according to the genetic information of the mother–offspring pair to measure the probability of the incest event. For this index, H_p and H_d' are the hypotheses that "individual B is the offspring of an outbred female (individual A) with her outbred father (i.e., $\Delta = \{0, 0, 0, 0, 0.25, 0, 0.25, 0.5, 0\}$ between the mother–offspring pair)" and "individual B is the offspring of an outbred female (individual A) with a random outbred male (i.e., $\kappa = \{0, 1, 0\}$ between the mother–offspring pair)". Thus, it can be unified as follows (see details in Section 2.1 of File S1 in Supplementary Materials):

$$II = \frac{d_{Ac} d_{Ad}}{2d_{Ac} p_d + 2d_{Ad} p_c} + \frac{1}{2} \tag{19}$$

Furthermore, Eq. 19 can be generalized to a more comprehensive scenario, in which H_p represents the hypothesis "individual B is the offspring of two outbred relatives with $\kappa = \{\kappa_0, \kappa_1, \kappa_2\}$, only one of which participated as individual A". If we define $\varphi = \kappa_2/2 + \kappa_1/4$ as the kinship coefficient (Brustad et al., 2021b) between the two alleged parents, it can be inferred that $\Delta = \{0, 0, 0, 0, \varphi, 0, \varphi, 1 - 2\varphi, 0\}$ between individual A and B under this circumstance (see Section 2.2 of File S1 Supplementary Materials). Therefore, a more comprehensive form of II (II_φ) can be computed as follows

$$II_\varphi = \frac{2\varphi d_{Ac} d_{Ad}}{d_{Ac} p_d + d_{Ad} p_c} + (1 - 2\varphi) \tag{20}$$

2.4 The extension of the aforementioned calculation method to kinship identification involving multiple participants

2.4.1 The basic method

The method of LR derivation in this work can be extended to identification involving multiple individuals. In such cases, LR can be calculated by considering the probability of a specific individual being a specific genotype from the perspective of IBD alleles of other individuals, and the probability of the genotype being the actual genotype detected. If this individual is labeled as individual B (with actual genotype of *cd*) and the others individuals R, then

$$LR = \frac{\Pr\left(R \mid H_p\right) \times \Pr\left(B \mid R, H_p\right)}{\Pr\left(R \mid H_d\right) \times \Pr\left(B \mid R, H_d\right)} \tag{21}$$

The latter probabilities in both Numerator and Denominator can be calculated as follows:

$$\begin{aligned}\Pr(B \mid R, H) &= \sum_j \left[\Pr\left(B = \mathcal{G}_j \mid R, H\right) \times \Pr\left(B \equiv cd \mid B \equiv \mathcal{G}_j, R, H\right)\right] \\ &= \sum_j \left[\Pr\left(B \equiv \mathcal{G}_j \mid R, H\right) \times \Pr\left(\mathcal{G}_j \equiv cd\right)\right]\end{aligned} \tag{22}$$

The second step in this equation is derived based on the same logic in section 2.3.1 when deriving Eq. 10, and probability $\Pr(\mathcal{G}_j \equiv cd)$ can be calculated with Eq. 11. Moreover, similar to pairwise cases, the calculation results remain unchanged regardless whether ***c*** and ***d*** are identical, due to the constant factor canceling from the numerator and denominator.

2.4.2 An example: "Standard" non-inbred trio cases

Consider the following non-inbred situation: 3 individuals participated the identification: a child (labeled as C, with detected genotype *cd*), one of his/her biological parents whose parentage has been confirmed (labeled as TP, with detected genotype *ab*), and an individual (labeled as AR, with detected genotype *ef*) who is unrelated to TP and alleged to be related to C under H_p. LR can be calculated by taking the null hypothesis as H_d, i.e., AR is unrelated to both TP and C. For such identification, C can be regarded as individual B mentioned in Eq. 21, and the other two participants as R, i.e.,

$$LR = \frac{\Pr\left(TP \equiv ab, AR \equiv ef \mid H_p\right)}{\Pr\left(TP \equiv ab, AR \equiv ef \mid H_d\right)} \times \frac{\Pr\left(C \equiv cd \mid TP \equiv ab, AR \equiv ef, H_p\right)}{\Pr\left(C \equiv cd \mid TP \equiv ab, AR \equiv ef, H_d\right)} \tag{23}$$

The relationship between TP and AR, which is unrelated in both hypotheses, remains constant. Therefore, LR equals the latter ratio in the above equation. The two probabilities in that ratio can still be calculated from the perspective of IBD alleles. If AR is unrelated to TP, the allele C inherited from TP must not be IBD to any of AR's alleles, i.e., $\kappa_2 = 0$ between C and AR. Thus their relationship under H_p can be described with κ_1 between them, for example, $\kappa_1 = 1$ if AR is alleged to be the other parent of C. If there is no other information, there can be 3 types of IBD allele C inherited from the other parent, e_I, f_I, and x_I, with probabilities of $\kappa_1/2$, $\kappa_1/2$, and $1 - \kappa_1$, respectively. Considering that TP must pass a_I or b_I to C with equal probabilities, there can be 6 IBD genotypes of C under H_p: a_Ie_I, a_If_I, b_Ie_I, b_If_I, each with a probability of $\kappa_1/4$, as well as a_Ix_I and b_Ix_I, each with a probability of $1/2 - \kappa_1/2$; Meanwhile, there can be 2 type of C's IBD genotype under H_d: a_Ix_I and b_Ix_I, each with a probability of 1/2. Thus, according to Eq. 22, LR can be calculated as follows:

$$LR = \frac{\frac{\kappa_1}{4}\left[\Pr(a_I e_I \equiv cd) + \Pr(a_I f_I \equiv cd) + \Pr(b_I e_I \equiv cd) + \Pr(b_I f_I \equiv cd)\right]}{\frac{1}{2}\left[\Pr(a_I x_I \equiv cd) + \Pr(b_I x_I \equiv cd)\right]} + (1 - \kappa_1) \tag{24}$$

Each probability in the equation can be calculated according to Eq. 11 considering the fact that no IBD relationship should exist among the four alleles of TP and AR, if the non-inbred assumption is accepted. In summary, LR in non-inbred trio cases can be calculated as follows:

$$LR = \frac{\kappa_1 d_{TPc} d_{ARd} + \kappa_1 d_{TPd} d_{ARc}}{2 d_{TPc} p_d + 2 d_{TPd} p_c} + (1 - \kappa_1) \tag{25}$$

Two more examples of LR calculation in identifications involving multiple participants are given in Sections 4.3.7 and 4.3.8 of File S1 in Supplementary Materials.

2.5 Verification of the unification results

2.5.1 For inbred identification

The 1 × 9 results of $\Pr(E \mid J_i)/\Pr(E \mid J_9)$ unification can be verified by comparison with Table 1 of Brustad et al. (2021b) (which is in 9 × 9 form) under the nine possible genotype combination types of the participants. The simplest situation, in which both individual A and B are homozygous, *ii*, is given as example in Table 1; as shown in the table, the results obtained with the two methods were identical under every Jacquard state. This identity persisted for the other eight genotype combinations, as shown in Section 3 of File S1 in Supplementary Materials.

2.5.2 For non-inbred identification

We reproduce the unified formula derived by (formula (2.19) in Egeland et al. (2017)) as Eq. 26 in this paper. If mutation is not considered, $m_{ij}^{(n)} = \mathbb{1}_{ij}$, where $m_{ij}^{(n)}$ represents the probability that

TABLE 2 Comparison between the unified formula with listed ones in the calculation of incest index.

Genotype			Unified calculation as Eq. 19				
Mother	Child	Listed result derived by Wenk	d_c	d_d	p_c	p_d	Result
A/A	A/A	$(0.5a + 0.5)/a$	2	2	a	a	$1/2 + 1/2a$
	A/B	0.5	2	0	a	b	$1/2$
A/B	A/A	$(0.5a + 0.25)/a$	1	1	a	a	$1/2 + 1/4a$
	A/C	0.5	1	0	a	c	$1/2$
	A/B	$(0.5a + 0.5b + 0.5)/(a + b)$	1	1	a	b	$1/2 + 1/(2a + 2b)$

Note: Symbols here are adjusted in the form used by Table 1 of Wenk (2007): alleles in the genotypes of the two participants are represented with uppercase letters "A", "B" and "C", which are different to each other; The corresponding lowercase letters "a", "b" and "c" denote the frequency of these alleles in the population.

TABLE 3 Comparison between the unified LR formula with listed ones in standard trio paternity testing.

Genotype			Listed result derived by Fung *et al*	Unified calculation as Eq. 27						
C	M	AF		d_{Mc}	d_{Md}	d_{AFc}	d_{AFd}	p_c	p_d	Result
A_iA_i	A_iA_i	A_iA_i	$1/p_i$	2	2	2	2	p_i	p_i	$8/(4p_i + 4p_i) = 1/p_i$
		A_iA_j	$1/(2p_i)$	2	2	1	1	p_i	p_i	$4/(4p_i + 4p_i) = 1/(2p_i)$
	A_iA_j	A_iA_i	$1/p_i$	1	1	2	2	p_i	p_i	$4/(2p_i + 2p_i) = 1/p_i$
		A_iA_j	$1/(2p_i)$	1	1	1	1	p_i	p_i	$2/(2p_i + 2p_i) = 1/(2p_i)$
		A_iA_k	$1/(2p_i)$	1	1	1	1	p_i	p_i	$2/(2p_i + 2p_i) = 1/(2p_i)$
A_iA_j	A_iA_i	A_iA_j	$1/(2p_j)$	2	0	1	1	p_i	p_j	$2/(0p_i + 4p_j) = 1/(2p_j)$
		A_jA_j	$1/p_j$	2	0	0	2	p_i	p_j	$4/(0p_i + 4p_j) = 1/p_j$
		A_jA_k	$1/(2p_j)$	2	0	0	1	p_i	p_j	$2/(0p_i + 4p_j) = 1/(2p_j)$
	A_iA_j	A_iA_i	$1/(p_i + p_j)$	1	1	2	0	p_i	p_j	$2/(2p_i + 2p_j) = 1/(p_i + p_j)$
		A_iA_j	$1/(p_i + p_j)$	1	1	1	1	p_i	p_j	$2/(2p_i + 2p_j) = 1/(p_i + p_j)$
		A_iA_k	$1/[2(p_i + p_j)]$	1	1	1	0	p_i	p_j	$1/(2p_i + 2p_j) = 1/[2(p_i + p_j)]$
	A_iA_k	A_iA_j	$1/(2p_j)$	1	0	1	1	p_i	p_j	$1/(0p_i + 2p_j) = 1/(2p_j)$
		A_jA_j	$1/p_j$	1	0	0	2	p_i	p_j	$2/(0p_i + 2p_j) = 1/p_j$
		A_jA_k	$1/(2p_j)$	1	0	0	1	p_i	p_j	$1/(0p_i + 2p_j) = 1/(2p_j)$
		A_jA_l	$1/(2p_j)$	1	0	0	1	p_i	p_j	$1/(0p_i + 2p_j) = 1/(2p_j)$

Note: Symbols here are adjusted in the form used by Table 4.1 of Fung and Hu (2007): alleles in the genotypes of the three participants are represented with "$\boldsymbol{A}_{\ldots}$" and the corresponding frequencies in the population are denoted with "$\boldsymbol{p}_{\ldots}$", where different subscripts means different alleles.

allele i becomes j after n cycles of meiosis. Thus, Eq. 16 is equivalent to Eq. 26 when we do not consider mutation.

$$LR = \kappa_0 + \kappa_1 \frac{\left(m_{ac}^{(n)} + m_{bc}^{(n)}\right)p_d + \left(m_{ad}^{(n)} + m_{bd}^{(n)}\right)p_c}{4p_c p_d} + \kappa_2 \frac{m_{ac}^{(n)} m_{bd}^{(n)} + m_{ad}^{(n)} m_{bc}^{(n)}}{2p_c p_d} \tag{26}$$

2.5.3 For alleged father-daughter incest cases

II was calculated and listed according to different genotype combination of the two individuals in Table 1 of Wenk (2007). The comparison of results calculated by Eq. 19 with these listed results is provided in Table 2, showing that the two methods are equivalent to each other under every possible combination type of the mother-offspring pair.

2.5.4 For trio paternity testing

"Standard" trio paternity testing is a routine procedure in forensic genetic practice, involving the evaluation of the paternity of a male (AF) to a child (C) with the genetic information of C's mother (M). This testing can be considered as a specific scenario of standard non-inbred trio cases discussed in section 2.4.2, where AR is AF, TP is M, and $\kappa_1 = 1$. Consequently, the paternity index in these cases, denoted as PI_{trio}, can be calculated as follows:

$$PI_{trio} = \frac{d_{Mc} d_{AFd} + d_{Md} d_{AFc}}{2d_{Mc} p_d + 2d_{Md} p_c} \tag{27}$$

The comparison of results calculated by Eq. 27 with those calculated by the recognized list formed methods (Fung and Hu, 2007) is listed in Table 3, showing that the two methods are equivalent to each other under every possible combination type of the three participants.

2.6 *KINSIMU*: A new series of simulation tools to promote research in kinship analysis

In order to evaluate the efficacy of specific panels in kinship identification, simulation can be a useful tool. Based on the unified formulae of LR calculation in pairwise kinship analysis, a series of newly defined functions is constructed for kinship simulation and calculation on the R (4.2.1) platform, an archive file is provided in File S2 in Supplementary Materials. Simulation can be divided into two steps: i) generation of genotype data for a specific number of participants (i.e., the sample size, labeled as "ss") with specific relationships; and ii) calculation of specific parameters for each case, e.g., identity by state score (IBS), or different type of LR. If independence among all the markers is assumed, the simulations at each locus should not affect each other, and the calculation results at each locus can be directly accumulated or multiplied. Thus, the whole code can be written in a one-layer loop, as shown in Algorithm 1. In each loop, the genotype data are replaced and the IBS/$\log_{10}$LR results are accumulated with the previous calculation. The above process can be completed using function "*testsimulation* ()". Detailed instructions for use of the functions is given in Section 4 of File S1 in Supplementary Materials.

```
Input: The allele frequency data
Output: A data frame containing calculation results
Input the allele frequency by hand or import the frequency data with function "EvaluatPanel()";
Extract the number of loci ("nl")
Set the sample size ("ss"), the true relationship between/among the individuals ("tdelta"), and the alleged relationship in LR calculation ("adelta");
Create data frame containing the results with initial values of 0;
for $i$ = 1:$nl$ do
    Simulate genotype combination of ss pairs/groups of individual cosidering tdelta with function "pairsimu()" or "pedisimu()", based on the frequency data of ith marker;
    Calculate LR or identity by state (IBS) score for each pair/group with function "LRparas()", "IICAL()", "TrioPI()", etc., based on unified equations. Note that the base 10 logarithms of LRs calculated will be output;
    Accumulate output results to the result data frame;
end
```

Algorithm 1. Typical process of simulation with *KINSIMU* package.

2.6.1 Evaluation of the package

Multiple types of evaluation were performed on the function "*testsimulation*()" based on allele frequency data for 42 autosomal short tandem repeat (STR) markers (Liu et al., 2020) in a Chinese Han population, given as data "*FortytwoSTR*" in the package. Similar simulations were carried out using two existing tools, *Familias* 3.3 (Egeland et al., 2016) and R package *relSim* (Curran, 2023), for comparison. All simulations were performed on the same personal device with Intel® Core™ i7-9700F CPU and 16 GB RAM, and the code used in the evaluation is given in File S3 in Supplementary Materials.

In essence, the main purpose of simulation (as well as the integration of LR formulae) is not only to compute precise results for individual real-life cases, but also to reveal the overall pattern of specific parameters within a particular type of identification. Therefore, in addition to the accuracy of the calculation results for a single case, the stability of result distribution in a large sample size is also crucial for the simulation package. Given that we have previously proven exhaustively in section 2.5 that the unified formula produces consistent results with the list format for specific single cases, our attention in this comparative analysis is primarily directed towards the following aspects: i) Identify any disparities in the distribution of calculated results between *KINSIMU* with the two existing tools, after a significant number of simulations. In other words, analyze whether there is a bias towards simulating specific genotype combinations with *KINSIMU*; ii) Assess the speed of simulation execution. As shown in Section 5 of File S1 in Supplementary Materials, the package *KINSIMU* can stably simulate and calculate multiple kinship cases with a speed of about 6,000,000 loci per second (as shown in Figure S4 of File S1 in Supplementary Materials). When simulating individual pairs with the same relationships and calculating the same parameters, the running speed of the *KINSIMU* package is approximately 10 times that of *relSim* (see Figure S3D of File S1 in Supplementary Materials) and at least 80 times that of *Familias* 3.3 (see Section 5.2 of File S1 in Supplementary Materials).

2.6.2 Application of the package

The above tools have been used in the construction of next-generation sequencing kits containing multiple single-nucleotide polymorphism (Zhao et al., 2021) or microhaplotype (Du et al., 2023) markers. In the relevant studies, family surveys were carried out at the same time, and the results of the two types of research were found to be similar; see Fig. 4 of (Zhao et al., 2021) and Fig. 4 of (Du et al. 2023), which illustrate the effectiveness of the simulation tools in real-life cases.

3 Discussion

In this work, a new approach in the unification of LR formulae has been introduced, the core of which is based on the perspective of IBD alleles and IBD genotypes. Based on the formulae, a package named *KINSIMU* is constructed for large-sample-size simulation research. Although some prototype software is already available for pedigree simulations (such as the two methods we compared with *KINSIMU*), any such tool (including *KINSIMU*) has limitations and may not take every possibility into account. Therefore, it may be more convenient for researchers to write their own simulation and calculation code for specific complex situations. In such cases, the coding logic and concepts of existing tools can be used for reference. The aim of this paper was to develop unified formulae for LR calculation and to simplify the coding process. Based on these formulae, we are making our tools for kinship simulation open-source and suggest that other researchers do the same to help the development of the discipline.

In simulation studies involving LR methods, the commonly used list-form presentation tends to obstruct the self-coding process. By using the most simplified formulae provided by the list-form method, the LR calculation process would involve two multiple-to-one choosing tasks: i) determining the type of the participants' genotype combination at a specific locus from multiple possible ones; and ii) judging the position of each allele in the specific combination. These tasks would vary at different loci, requiring locus by locus or type by type calculations for LR, necessitating logical comparison functions like "if()", leading to increased coding complexity. Additionally, all possible participant combinations must be considered in the coding process, or errors may occur in certain cases. In contrast, no logical comparison function would be required when using unified formulae; based on the genotype data, nine or six parameters (depending on whether inbreeding factors are considered) could be uniformly calculated and then brought into a unified formula for identification.

In most studies that utilize *Familias* (Gonçalves et al., 2017; Li et al., 2019; Wu et al., 2021; Pilli et al., 2022), the sample size for simulations is often less than 10,000 due to the substantial increase in memory requirements when simulating larger samples. This limitation may be attributed to the tool's operation method, wherein genotype data is retained until the end of the simulation, leading to unnecessary memory occupation. Contrarily, R packages like *KINSIMU* or *relSim* replace the data per loop (or per "Block") after it is used for calculation. It is somewhat "unfair" to compare the running time of our R tools with the stand-alone software *Familias* instead of its R version (which is no longer directly accessible from CRAN, and the available installation package is not compatible with our R version). However, some insights can be gained from this comparison. For instance, with *Familias*, the more complex the relationship between participants, the longer the running time, even with the same sample size. This phenomenon is not observed with *KINSIMU*, possibly because we directly generate the participants' genotypes rather than determining them through parent-child inheritance, resulting in a significant reduction in the needed random numbers during simulation. However, the parent-child inheritance approach may be necessary in cases with complicated pedigrees or when more individuals need to be considered for calculating LR. Therefore, we provide an approach for such cases in the form of the function "*pedisimu*()". Meanwhile, it is important to note that the tools *Familias* and *relSim* offer numerous functions besides simulation.

The direct-generation approach is used in the package *relSim*, and the simulation process is divided into 100 sub-fractions (so-called "Blocks"), each of which is replaced by the next. As a result, it can simulate at least 10,000,000 pairs in a single process on our device. When the sample size is no longer an obstacle, the running time becomes the main bottleneck in large-sample simulation research. In the reference manual of *relSim* package, the authors state that it would take at least 30 h on a personal device to perform a simulation of 300,000,000 individual pairs on 13 CODIS loci (Balding et al., 2013). Even allowing for decreased running time with ongoing updates to devices, the latest version of *relSim* would still take twice as long as *KINSIMU* to carry out the same simulation. Examination of the functions of *relSim* shows that genotype data are cached per case but generated per locus. Therefore, an extra "for" loop layer is required to allocate the data. Another extra loop layer exists for the calculation of CIBS or CLR with *relSim*, which would take place per case per locus when applying list-form formulae. Furthermore, at least two layers of "if (. . .)" functions are needed in these calculations to classify the genotype combinations in different situations and to calculate the parameters with different equations. It is difficult to apply such methods to multiple cases per locus, as can be done with the unified functions in *KINSIMU*; this may be the main cause of the difference in running time between the two packages. A side-by-side comparison of the codes used in PI calculation is given in Section 6 of File S1 in Supplementary Materials.

The method introduced in this paper for LR unification in kinship testing can be regarded as a divide-and-conquer method. More specifically, the problem is essentially about the genotype inference of individual B from the perspective of IBD under different hypotheses. This problem can be solved with the following approach: i) division of the original problem into several independent easy-to-solve sub-problems, i.e., the division of different IBD states; ii) conquering each sub-problem, i.e., calculating Δ/κ and the ratio of conditional probabilities under each state; and iii) combining the solutions to the sub-problems into the solution for the original problem. The use of sub-scripted letters (e.g., a_I) for individual B's genotype from the perspective of IBD alleles eliminates the differences in the detected genotype combinations of the two individuals. Therefore, the derivation in step ii of each sub-problem can be done in a unified way; this is simpler in some ways than the method Egeland *et al*. Introduced in Appendix A of (Egeland et al., 2017), in which the combinations had to be listed exhaustively. Thus, the derivation can be easily extended to identifications considering inbreeding factors or involving multiple individuals.

Mutation is not considered in the inference process in the present work, resulting in underestimation of the LR value in some cases. However, we argue that the impact of mutation on the LR value is relatively small in identifications where LR cannot be 0. For example, in pairwise identification using STR markers, as discussed in (Egeland et al., 2017), the probability of mutation occurring is relatively low in parentage identification and increases with the number of cycles of meiosis. In other words, the probability is larger when the relationship between the two individuals is more distant. In that case, κ_0 would be larger, and the other two κ parameters (i.e., the coefficients of parts in the LR formula related to the mutation) would be smaller, which limits the impact of mutation on the LR value to a relatively low level. Furthermore, the mutation factor can be introduced into the calculation if $\mathbb{1}$ parameters are replaced by corresponding m ones.

This is a preliminary study on the concept of unification of LR calculation in kinship identification; further work based on our findings could include the inference of LR in cases involving multiple individuals or where linked markers are available.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

BC and SL proposed the idea of unifying LR calculation; GM performed the deduction process, constructed and evaluated the R package; QW verified the unification results; GM and QW wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (82072118), Natural Science Foundation of Hebei Province (H2021206451), and S&T Program of Hebei (225A5602D).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1226228/full#supplementary-material

References

Balding, D. J., Krawczak, M., Buckleton, J. S., and Curran, J. M. (2013). Decision-making in familial database searching: KI alone or not alone? *Forensic Sci. Int. Genet.* 7, 52–54. doi:10.1016/j.fsigen.2012.06.001

Brustad, H. K., Colucci, M., Jobling, M. A., Sheehan, N. A., and Egeland, T. (2021a). Strategies for pairwise searches in forensic kinship analysis. *Forensic Sci. Int. Genet.* 54, 102562. doi:10.1016/j.fsigen.2021.102562

Brustad, H. K., Vigeland, M. D., and Egeland, T. (2021b). Pairwise relatedness testing in the context of inbreeding: expectation and variance of the likelihood ratio. *Int. J. Leg. Med.* 135, 117–129. doi:10.1007/s00414-020-02426-6

Curran, J. M. (2023). *relSim: a tool for simulating related DNA profiles and mixtures. R package version 1.0.0.*

Du, Q., Ma, G., Lu, C., Wang, Q., Fu, L., Cong, B., et al. (2023). Development and evaluation of a novel panel containing 188 microhaplotypes for 2nd-degree kinship testing in the Hebei Han population. *Forensic Sci. Int. Genet.* 65, 102855. doi:10.1016/j.fsigen.2023.102855

Egeland, T., Kling, D., and Mostad, P. (2016). *Relationship inference with Familias and R.* Massachusetts, United States: Academic Press. doi:10.1016/C2014-0-01828-X

Egeland, T., Pinto, N., and Amorim, A. (2017). Exact likelihood ratio calculations for pairwise cases. *Forensic Sci. Int. Genet.* 29, 218–224. doi:10.1016/j.fsigen.2017.04.018

Egeland, T., and Sheehan, N. (2008). On identification problems requiring linked autosomal markers. *Forensic Sci. Int. Genet.* 2, 219–225. doi:10.1016/j.fsigen.2008.02.006

Fung, W. K., and Hu, Y. Q. (2007). *Parentage testing.* 4. New Jersey, United States: John Wiley and Sons, Ltd, 47–78. doi:10.1002/9780470727041.ch4

Gjertson, D. W., Brenner, C. H., Baur, M. P., Carracedo, A., Guidet, F., Luque, J. A., et al. (2007). ISFG: recommendations on biostatistics in paternity testing. *Forensic Sci. Int. Genet.* 1, 223–231. doi:10.1016/j.fsigen.2007.06.006

Gonçalves, J., Conde-Sousa, E., Egeland, T., Amorim, A., and Pinto, N. (2017). Key individuals for discerning pedigrees belonging to the same autosomal kinship class. *Forensic Sci. Int. Genet.* 29, 71–79. doi:10.1016/j.fsigen.2017.03.018

Jacquard, A. (1972). Genetic information given by a relative. *Biometrics* 28, 1101–1114. doi:10.2307/2528643

Kling, D., and Tillmar, A. (2019). Forensic genealogy—a comparison of methods to infer distant relationships based on dense SNP data. *Forensic Sci. Int. Genet.* 42, 113–124. doi:10.1016/j.fsigen.2019.06.019

Li, R., Li, H., Peng, D., Hao, B., Wang, Z., Huang, E., et al. (2019). Improved pairwise kinship analysis using massively parallel sequencing. *Forensic Sci. Int. Genet.* 38, 77–85. doi:10.1016/j.fsigen.2018.10.006

Liu, Q., Ma, G., Du, Q., Lu, C., Fu, L., Wang, Q., et al. (2020). Development of an NGS panel containing 42 autosomal STR loci and the evaluation focusing on secondary kinship analysis. *Int. J. Leg. Med.* 134, 2005–2014. doi:10.1007/s00414-020-02295-z

Mo, S.-K., Liu, Y.-C., Wang, S.-q., Bo, X.-C., Li, Z., Chen, Y., et al. (2016). Exploring the efficacy of paternity and kinship testing based on single nucleotide polymorphisms. *Forensic Sci. Int. Genet.* 22, 161–168. doi:10.1016/j.fsigen.2016.02.012

Mo, S.-K., Ren, Z.-L., Yang, Y.-R., Liu, Y.-C., Zhang, J.-J., Wu, H.-J., et al. (2018). A 472-SNP panel for pairwise kinship testing of second-degree relatives. *Forensic Sci. Int. Genet.* 34, 178–185. doi:10.1016/j.fsigen.2018.02.019

Phillips, C., García-Magariños, M., Salas, A., Carracedo, Á., and Lareu, M. V. (2012). SNPs as supplements in simple kinship analysis or as core markers in distant pairwise relationship tests: when do SNPs add value or replace well-established and powerful STR tests? *Transfus. Med. Hemotherapy* 39, 202–210. doi:10.1159/000338857

Pilli, E., Tarallo, R., Riccia, P. L., Berti, A., and Novelletto, A. (2022). Kinship assignment with the ForenSeq$^{\text{TM}}$ DNA signature prep kit: sources of error in simulated and real cases. *Sci. Justice* 62, 1–9. doi:10.1016/j.scijus.2021.10.007

Pinto, N., Silva, P. V., and Amorim, A. (2010). General derivation of the sets of pedigrees with the same kinship coefficients. *Hum. Hered.* 70, 194–204. doi:10.1159/000316390

Skare, Ø., Sheehan, N., and Egeland, T. (2009). Identification of distant family relationships. *Bioinformatics* 25, 2376–2382. doi:10.1093/bioinformatics/btp418

Slooten, K. (2020). Likelihood ratio distributions and the (ir)relevance of error rates. *Forensic Sci. Int. Genet.* 44, 102173. doi:10.1016/j.fsigen.2019.102173

Staadig, A., and Tillmar, A. (2019). An overall limited effect on the weight-of-evidence when taking STR DNA sequence polymorphism into account in kinship analysis. *Forensic Sci. Int. Genet.* 39, 44–49. doi:10.1016/j.fsigen.2018.11.020

Vigeland, M. D. (2022). QuickPed: an online tool for drawing pedigrees and analysing relatedness. *BMC Bioinforma.* 23, 220. doi:10.1186/s12859-022-04759-y

Wenk, R., Chiafari, F., and Houtz, T. (1994). Incest diagnosis by comparison of alleles of mother and offspring at highly heterozygous loci. *Transfusion* 34, 172–175. doi:10.1046/j.1537-2995.1994.34294143949.x

Wenk, R. E. (2007). Incest indices from microsatellite genotypes of mother-child pairs. *Transfusion* 48, 341–348. doi:10.1111/j.1537-2995.2007.01528.x

Wu, R., Chen, H., Li, R., Zang, Y., Shen, X., Hao, B., et al. (2021). Pairwise kinship testing with microhaplotypes: can advancements be made in kinship inference with these markers? *Forensic Sci. Int.* 325, 110875. doi:10.1016/j.forsciint.2021.110875

Zhao, G.-B., Ma, G.-J., Zhang, C., Kang, K.-L., Li, S.-J., and Wang, L. (2021). BGISEQ-500RS sequencing of a 448-plex SNP panel for forensic individual identification and kinship analysis. *Forensic Sci. Int. Genet.* 55, 102580. doi:10.1016/j.fsigen.2021.102580