Legal Medicine 69 (2024) 102441

Contents lists available at ScienceDirect

Legal Medicine

journal homepage: www.elsevier.com/locate/legalmed

Determining the effects of genetic linkage when using a combination of STR and SNP loci for kinship testing

Da Yang [a,b,*], Sheng Xuan Ma [a], Guo Liang Zhao [a], Ao Gao [a], Zhao Kun Xu [a]

[a] Institute of Forensic Medicine And Laboratory Medicine, Jining Medical University, Shandong province, P. R. China
[b] Forensic Science Center of Jining Medical University, Shandong province, P. R. China

ARTICLE INFO

Keywords:
Linkage
Kinship
Lander-green algorithm
Paternity testing
Pedigree likelihood
DNA testing

ABSTRACT

The pedigree likelihood ratio (LR) can be used for determining kinship in the forensic kinship testing. LR can be obtained by analyzing the DNA data of Short tandem repeat (STR) and single nucleotide polymorphism (SNP) loci. With the advancement of biotechnology, increasing number of genetic markers have been identified, thereby expanding the pedigree range of kinship testing. Moreover, some of the loci are physically closer to each other and genetic linkage between loci is inevitable. LRs can be calculated by accounting for linkage or ignoring linkage ($LR_{linkage}$ and LR_{ignore}, respectively). GeneVisa is a software for kinship testing (www.genevisa.net) and adopts the Lander–Green algorithm to deal with genetic linkage. Herein, we used the simulation program of the software GeneVisa to investigate the effects of genetic linkage on 1st-degree, 2nd-degree, and 3rd-degree kinship testing. We used this software to simulate $LR_{linkage}$ and LR_{ignore} values based on 43 STRs and 134 SNPs in commercial kits by using the allele frequency rate and genetic distance data of the European population. The effects of linkage on LR distribution and LRs of routine cases were investigated by comparing the $LR_{linkage}$ values with the LR_{ignore} values. Our results revealed that the linkage effect on LR distributions is small, but the effect on LRs of routine cases may be large. Moreover, the results indicated that the discriminatory power of genetic markers for kinship testing can be improved by accounting for linkage.

1. Introduction

DNA typing methods are used to resolve several problems associated with kinship testing, such as inheritance, emigration, identification of victims in mass accidents, and finding the criminal's close relatives from a DNA database. Kinship testing can be conducted by evaluating the DNA data of different loci obtained using DNA typing methods. Short tandem repeat (STR) loci and single nucleotide polymorphism (SNP) loci are currently used for kinship testing, with STR loci being more commonly used. Forensic identification is performed using loci supplied as commercial kits. Different pedigrees can represent possible relationships between involved individuals. We can obtain pedigree likelihoods by analyzing DNA typing data of the involved individuals. Then, the likelihood ratio (LR) of two different pedigrees can be obtained to determine which relationship is true.

With the advancement of biotechnology, especially with the advent of next-generation sequencing technologies, an increasing number of loci have been developed and used for paternity testing. As the number of loci increased, some of the loci were found to be physically closer to each other on the chromosome and linkage between the loci became inevitable. Linkage indicates that alleles on two loci that are physically close to each other may be inherited as a unit during meiosis [1], that means genetic recombination did not occur between them. The degree of linkage can be measured on the basis of the recombination rate. The number of base pairs between two loci determines the physical distance between the two loci. By contrast, the genetic distance between two loci can be measured in centimorgans (cMs), with 1 cM representing a 1 % chance of recombination between the alleles of two loci. Another allelic correlation is linkage disequilibrium (LD) [2]. LD refers to the non-random correlation of alleles of different loci in a population. If the genetic distance between two loci is < 0.5 cM, whether the alleles of the two loci are in LD must be tested [3,4]. When two or more loci are in LD, haplotype frequencies instead of allele frequencies are used to obtain LR.

The International Society of Forensic Genetics (ISFG) recommends the use of autosomal genetic markers for kinship testing through biostatistics methods without accounting for linkage [5]. It has also shared guidelines regarding the use of X-STRs in kinship testing involving linkage [6]. The effects of genetic linkage of autosomal loci on

* Corresponding author.
E-mail address: yangdajnmu@163.com (D. Yang).

https://doi.org/10.1016/j.legalmed.2024.102441
Received 4 February 2024; Received in revised form 19 March 2024; Accepted 30 March 2024
Available online 4 April 2024

kinship testing has been investigated [3,7–14]. Two basic methods are used to compile the computer program for determining the LR under the condition of linkage or no linkage. In the first method, that is, the Elston–Stewart method [15], the computing time increases linearly with the number of meiosis events of the pedigree and increases exponentially with the number of genetic markers. By contrast, in the second method, that is, the Lander–Green (L–G) method [16], the computation time increases linearly with the number of genetic markers and increases exponentially with the number of meiosis events of the pedigree. Bayesian networks can analyze the probability of occurrence of multiple non-independent or interrelated events and can be used for implementing both of the aforementioned methods [17]. The software FamLink [7] and MERLIN [18] implementing the L–G algorithm and software KinBN [8] implementing the Bayesian network algorithm are used to perform linkage analysis for forensic purposes. Based on the kinship testing of full siblings versus unrelated individuals, Tillmar et al [3] determined the impact of linkage by using simulation modules of MERLIN and FamLink and a series of SNP and STR loci (27 STR + 134 SNP). Based on a series of relationship tests, Morimoto et al [8] determined the impact of linkage by using simulation modules of KinBN and 21 STR loci of the GlobalFiler kit. In addition, Zhang et al [9] investigated the impact of linkage on the discrimination power of a series of SNP and STR loci (40 STR + 91 SNP) in two commercial kits by using real genotype data from 74 full-sibling pairs, 114 uncle/aunt-nephew/niece pairs and 93 grandparent-grandson/granddaughter pairs.

In this study, we used the software GeneVisa [14] implementing the L–G algorithm to investigate the effects of genetic linkage. First, 43 STR and 134 SNP loci were selected from commercial kits. The allele frequencies of these loci in the European population were obtained from published articles and public web resources, and the genetic distance between these loci were obtained based on the genetic maps of the European population created by Bhérer C et al [19]. Second, by using a combination of STR and SNP loci in three kinship testing cases, we compared the results of the software GeneVisa with those of other software to validate this software. Finally, based on several relationship tests, the effects of genetic linkage on routine cases, LR distribution curves and discrimination power parameters were investigated using the 43 STR and 134 SNP loci in 4 commercial kits. The relationship tests discriminated individuals with first-degree, second-degree, and third-degree kinship from unrelated individuals. When the discrimination power of 43 STRs and 134 SNPs was insufficient for the relationship tests, a full sibling of the involved person was added to improve the discriminative power.

2. Materials and method

2.1. Basic principle

We aimed to identify the relationship between individuals. Different relationships were described by different hypotheses. The pedigree likelihood of different hypotheses was evaluated based on genetic evidence, and then, LR was used to determine the most acceptable hypothesis. To determine the effects of genetic linkage, two types of LR were considered: LR_{ignore} without considering linkage and $LR_{linkage}$ considering linkage.

The software GeneVisa use the Lander–Green (L–G) algorithm to deal with genetic linkage. Herein, we describe $LR_{linkage}$ and LR_{ignore} based on the L–G algorithm. Let $G = [G_1, G_2 \ldots G_m]$ represent the genetic evidence of m loci, G_n represent the genetic evidence at the n-th loci, $v = [v_1, v_2 \cdots v_m]$ represent the inheritance vector of m loci,v_n represent the inheritance vector at the n-th loci, $H = [H_1, H_2]$ represent two opposing hypotheses describing two types of kinship, and $Z = [Z_1, Z_2]$ represent the inheritance vector established based on the different hypotheses, for example, Z_1 represents the inheritance vector established based on the kinship described by hypothesis H_1. Then, LR_{ignore} based on m loci that are independent of each other can be expressed as follows:

$$LR_{ignore} = \frac{P(G|H_1)}{P(G|H_2)} = \prod_{n=1}^{m} \frac{\sum_{v_n \in Z_1} P(G_n, v_n|H_1)}{\sum_{v_n \in Z_2} P(G_n, v_n|H_2)} = \prod_{n=1}^{m} \frac{\sum_{v_n \in Z_1} P(G_n|v_n, H_1) \times P(v_n|H_1)}{\sum_{v_n \in Z_2} P(G_n|v_n, H_2) \times P(v_n|H_2)}$$

$LR_{linkage}$ values for m linked loci can be expressed using the recursive formulas:

If $m = 1$

$$LR_{linkage} = LR_{ignore} = \frac{P(G|H_1)}{P(G|H_2)} = \frac{\sum_{v_1 \in Z_1} P(G_1, v_1|H_1)}{\sum_{v_1 \in Z_2} P(G_1, v_1|H_2)} = \frac{\sum_{v_1 \in Z_1} P(G_1|v_1, H_1) \times P(v_1|H_1)}{\sum_{v_1 \in Z_2} P(G_1|v_1, H_2) \times P(v_1|H_2)}$$

Herein, $P(G_1, v_1|H) = P(G_1|v_1, H) \times P(v_1|H)$

If $m > 1$

$$LR_{linkage} = \frac{P(G|H_1)}{P(G|H_2)} = \frac{\sum_{v_m \in Z_1} P(G_1 \cdots G_m, v_m|H_1)}{\sum_{v_m \in Z_2} P(G_1 \cdots G_m, v_m|H_2)}$$

$$= \frac{\sum_{v_m \in Z_1} P(G_m|v_m, H_1) \times \sum_{v_{m-1} \in Z_1} P(v_m|v_{m-1}, H_1) \times P(G_1 \cdots G_{m-1}, v_{m-1}|H_1)}{\sum_{v_m \in Z_2} P(G_m|v_m, H_2) \times \sum_{v_{m-1} \in Z_2} P(v_m|v_{m-1}, H_2) \times P(G_1 \cdots G_{m-1}, v_{m-1}|H_2)}$$

The basic idea of the aforementioned formulas is similar to that of previous studies [14,20]. $P(v_n|v_{n-1}, H)$ depends on the recombination rate r, and $P(v_n|v_{n-1}, H) = r^d \times (1-r)^{s-d}$, where s is the amount of meiotic events in the pedigree described by the hypothesis and d represents differences between the vectors v_m and v_{m-1}. If we ignore the effects of the subpopulation and mutation and assume f founders to be present in the pedigree, let $a = [a_1, a_2 \ldots a_{2f}]$ represent a collection of founder alleles compatible with genotypes and vectors, then we have $P(G_m|v_m, H) = \sum_a \prod_{i=1}^{2f} P(a_i)$, where $P(a_i)$ represents the frequency of allele a_i.

We generally use threshold t to determine the relationship. If $LR > t$, H_1 is considered to be true, that is, the individuals have the relationship described by H_1; otherwise, H_2 is true.

2.2. Genetic markers, allele frequencies, and recombination rate

A total of 44 STRs were selected from three commercial kits, namely PowerPlex® Fusion kit (Promega, America), Investigator HDplex kit (Qiagen, Germany), and SureID® 23comp Human DNA Identification Kit (Ningbo Health Gene Technologies, China). These STR loci are commonly used in kinship testing. SNP loci can be used as supplements to improve the system power for kinship testing. GeneRead DNAseq 140 IISNP (Qiagen, Germany) is a relatively commonly used SNP kit, so we selected 140 SNPs in this panel for our research. There is genetic linkage between these loci [4,21], and when using two or more of the aforementioned kits for kinship testing, the linkage will affect the LR values, so these STRs and SNPs were used to investigate the effects of genetic linkage. Allele frequencies of the STR loci in the European population were taken from previous studies [22–24]. The allele frequency of the Northern Italian population is not significantly different from those of almost all of the European neighboring populations [24]. Therefore, we used the allele frequency of the Northern Italian population instead of that of the European population. Allele frequencies of SNP loci in the European population were obtained from the 1000 Genomes Project (https://www.internationalgenome.org).

The physical positions of STRs and SNPs were obtained from [21] and [4], respectively, which were from dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and the BioProject (https://www.ncbi.nlm.nih.gov/bioproject/). The genetic positions of these genetic markers were obtained based on the genetic maps of the European population created by Bhérer C et al [19]. If the physical position of a genetic marker matched directly in these European genetic maps, the genetic position of

this marker was directly obtained. If no match occurred, a linear calculation was performed to infer the genetic position of this marker based on its closest markers. Finally, the Kosambi mapping function was used to determine the recombination rate based on the genetic distance between genetic markers.

D5S2500 and D5S2800 exhibited a very close genetic distance of less than 0.5 CM. We removed D5S2500 to avoid LD. In accordance with Ran Li et al [4], four pairs of genetic markers and one group of three genetic markers in the 140 SNP genetic markers were in LD; hence, we removed six SNP genetic markers to avoid LD. Of the remaining 43 STR and 134 SNP loci, pairs of loci with a genetic distance of less than 0.5 CM were tested for LD, and the results revealed no LD in these loci [4].

2.3. Validation of the kinship program

The software GeneVisa (https://www.genevisa.net) was used for kinship testing in this study. The primary version of GeneVisa can only handle STR genetic markers [14]. However, this version can also handle SNP genetic markers at present. The software has been validated using STR genetic markers [14]. We further used three kinship cases in Fig. 1 to validate the software by using a combination of 14 STR and SNP markers. The results without linkage were compared using Familias3.3 [25] and those involving linkage were compared using FamLink 2.1 [7].

2.4. Simulations

The effects of linkage were evaluated based on the distinction between first-, second-, and third-degree relatives from unrelated individuals. The corresponding kinship testing is described in terms of H_1 and H_2 as follows:

i. H_1: two full siblings and H2:two unrelated.
ii. H1: nephew–uncle and H2: two unrelated.
iii. H1: two first cousins and H2: two unrelated.

For the kinship tests ii and iii, the discrimination ability of 43 STRs and 134 SNPs was inferred insufficient based on previous study results [26]. We, therefore. added a full sibling of the involved person to improve the discriminative ability. The corresponding kinship testing is described as follows:

iv. H1: uncle–two nephew and H2: two full siblings and one unrelated.
v. H1: two full siblings and one first cousin and H_2: two full siblings and one unrelated.

The kinship tests iv and v corresponded to cases 1 and 2 in Fig. 1, respectively, in which the relationship between individual 1 and individual 2 had to be determined. The full sibling of individual 1, that is,

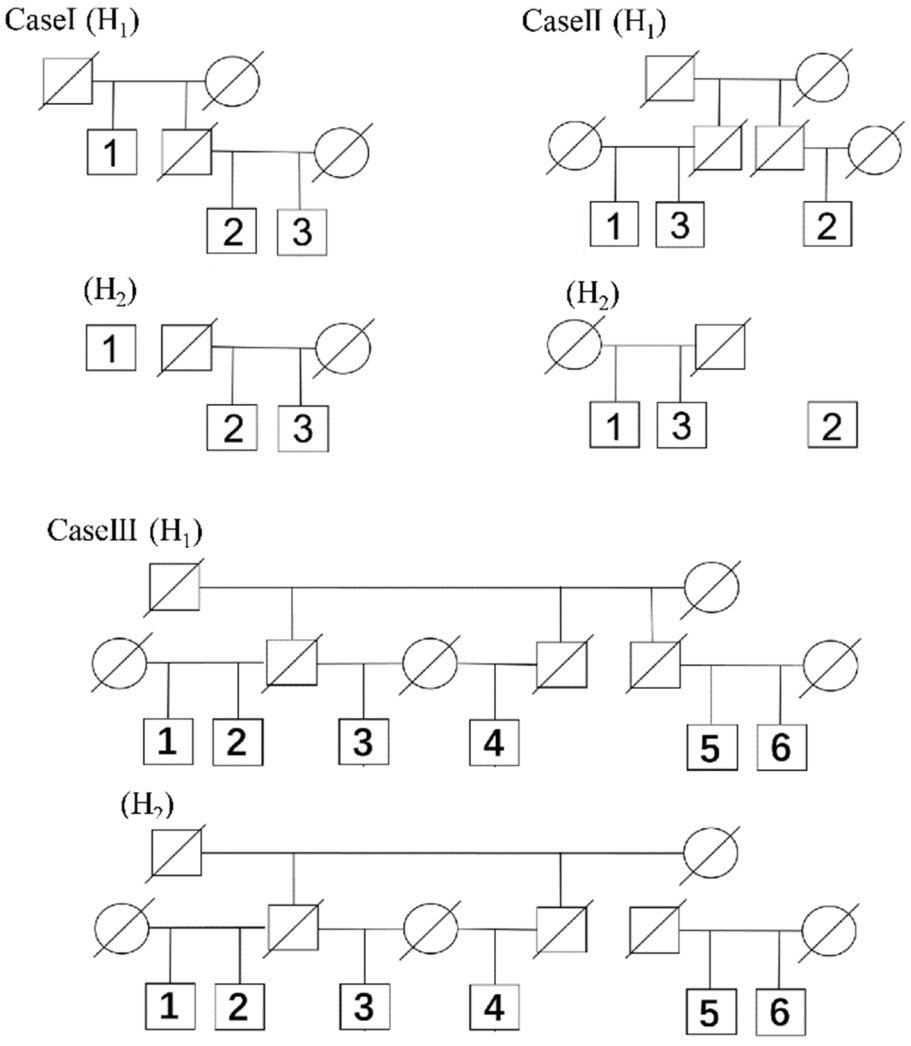

Fig. 1. Relationships of H_1 and H_2 in three kinship cases. Typed individuals are numbered.

individual 3, was also involved in the kinship testing.

The simulation program of GeneVisa can simulate LR distributions. In this study, 10,000 simulations based on the kinship testing i–v were performed considering H_1 and H_2 true respectively. In each simulation process, the program first simulated the genotype of individuals based on the allele frequency, recombination rate, allelic mutation rate, and Mendelian laws of inheritance. Allele frequency and recombination rate are consistent with Section 2.2. The average mutation rate of STR was adopted from a previous study [27]. For STR loci, the mutations were programmed to occur between two genes existing in the frequency file [22–24] and 99 % of the mutations were programmed to occur with a minimum number of steps between two genes, meaning that most of the mutations were one-step and very few were multi-step ($\geq$ 2-step). The other 1 % is directly set to multi-step mutation. For SNP loci, mutations are ignored. Then, the L–G algorithm was used to obtain $LR_{linkage}$ and LR_{ignore} values based on the allele frequency and recombination rate, as no genetic inconsistency was presented in the kinship tests i–v.

To determine the effects of genetic linkage, $LR_{linkage}$ values were compared to LR_{ignore} values in 10,000 simulations. First, we compared the distributions of $Log_{10}^{LR_{linkage}}$ with those of $Log_{10}^{LR_{ignore}}$. The discriminatory ability of kinship testing is determined by various parameters including sensitivity, specificity, and accuracy [28]. We here defined sensitivity as the proportion of cases in which involved individuals with kinship H_1 are correctly identified. Specificity was defined as the proportion of cases in which involved individuals with kinship H_2 are correctly identified. Accuracy is the proportion of cases in which individuals with relationship H_1 or H_2 are correctly identified. To study the impact of gene linkage on LR distribution, we determined sensitivity, specificity, and accuracy at different thresholds based on $LR_{linkage}$ and LR_{ignore} values. The threshold recommended by ISFG is 100 or 1000 [5]. Herein, we used 1, 10, 100, and 1000 as thresholds of LR values, corresponding to log_{10}LR values of 0, 1, 2, and 3, respectively. Finally, we obtained $Log_{10}^{LR_{linkage}/LR_{ignore}}$ to determine the effects of linkage on LRs of routine cases.

3. Results

Without considering genetic linkage, the calculation results of software GeneVisa and software Familias for the cases in Fig. 1 are consistent. When considering linkage, the calculation results of software GeneVisa and software FamLink for the cases in Fig. 1 are consistent. The following are the simulation results for the effects of genetic linkage on 1st-degree, 2nd-degree, and 3rd-degree kinship testing.

3.1. First-Degree kinship analysis

When the number of genetic markers increased, the ability to identify kinship improved (Table 1 and Fig. 2). For example, under the condition involving linkage, when 134 SNPs were added (from 43 STRs to 43 STRs + 134 SNPs), the accuracy increased from 99.75 % to 100 % at threshold 1000, and the impact of linkage also increased. For 43 STRs, the impact of linkage on the LR distribution was not apparent. (Fig. 2a). Comparing Fig. 2a with Fig. 2b, we observed that the addition of 134 SNPs increased the impact of linkage on the LR distributions and rendered it more apparent. Accuracy and sensitivity increased under the condition involving linkage. For example, at threshold 1000, when 22 STRs were used, sensitivity and accuracy increased from 90.16 % to 90.40 % and 95.07 % to 95.20 %, respectively. When 43 STRs were used, accuracy increased further from 99.66 % to 99.75 %.

Furthermore, with an increase in the number of genetic markers, the impact of linkage on the LR of cases significantly increased (Fig. 3a). In 10,000 simulations considering H_1 to be true, when 22 STRs were used, the number of $LR_{linkage}/LR_{ignore} > 4$ was 3, with a maximum value of 4.5118. When 43 STRs were used, this number became 2123, and the number of $LR_{linkage}/LR_{ignore} > 10$ was 393, with a maximum value of 67.8561. When 134 SNPs + 43 STRs were used, this number was 7286, and the number of $LR_{linkage}/LR_{ignore} > 10{,}000$ was 518, with a maximum value of 3018270.66. A similar situation occurs in 10,000 simulations considering H_2 true. In addition, if linkage was ignored, individuals in the three cases with full sibling relationships were misclassified as unrelated individuals at threshold 100 or 1000, no such case was observed when linkage was considered.

3.2. Second-Degree kinship Analysis

With an increase in the number of genetic markers or the addition of a full sibling, the ability to identify kinship from unrelated relatives increased (Table 2). For example, at threshold 1000, when 134 SNPs were added, LR_{ignore} accuracy increased from 80.12 % to 92.00 %. When a full sibling was added using 134 SNPs + 43 STRs, LR_{ignore} accuracy increased from 92.00 % to 97.89 %. During this process, the impact of linkage also increased.

For 43 STRs, the impact of linkage on the LR distribution was not very obvious (Fig. 2c). On comparing Fig. 2c with Fig. 2d, we found that the impact of linkage on the LR distribution based on H_1 true became more apparent with the addition of a full sibling. Accuracy and sensitivity increased when considering linkage. For example, at threshold 10 (log_{10}LR threshold is 1), with the use of 143 SNPs + 43 STRs, sensitivity increased from 96.47 % to 97.52 % under the condition involving linkage. When a full sibling of the nephew was added, under the condition of with linkage, sensitivity increased from 99.07 % to 99.46 % and accuracy increased from 99.52 % to 99.71 %. However, specificity mostly decreased when linkage was considered. For instance, at threshold 10, with the use of 143 SNPs + 43 STRs, specificity decreased from 99.85 % to 99.81 %. When a full sibling was added, specificity

Table 1
The parameter values based on the kinship testing of full siblings versus unrelated individuals.

Loci	Thresholds (Log_{10}LR)	Sensitivity(%)		Specificity(%)		Accuracy (%)	
		$LR_{linkage}$	LR_{ignore}	$LR_{linkage}$	LR_{ignore}	$LR_{linkage}$	LR_{ignore}
22STR	0	99.46	99.40	99.42	99.45	99.44	99.43
	1	98.23	98.14	99.85	99.87	99.04	99.01
	2	95.61	95.45	99.96	99.96	97.79	97.71
	3	90.40	90.16	99.99	99.98	95.2	95.07
43STR	0	99.97	99.96	100	100	99.99	99.98
	1	99.89	99.85	100	100	99.95	99.93
	2	99.75	99.69	100	100	99.88	99.85
	3	99.50	99.31	100	100	99.75	99.66
134SNP + 43STR	0	100	99.99	100	100	100	100.00
	1	100	99.99	100	100	100	100.00
	2	100	99.97	100	100	100	99.99
	3	100	99.97	100	100	100	99.99

22STR: 20 Codis loci, Penta D and Penta E, or loci in PowerPlex® Fusion kit.

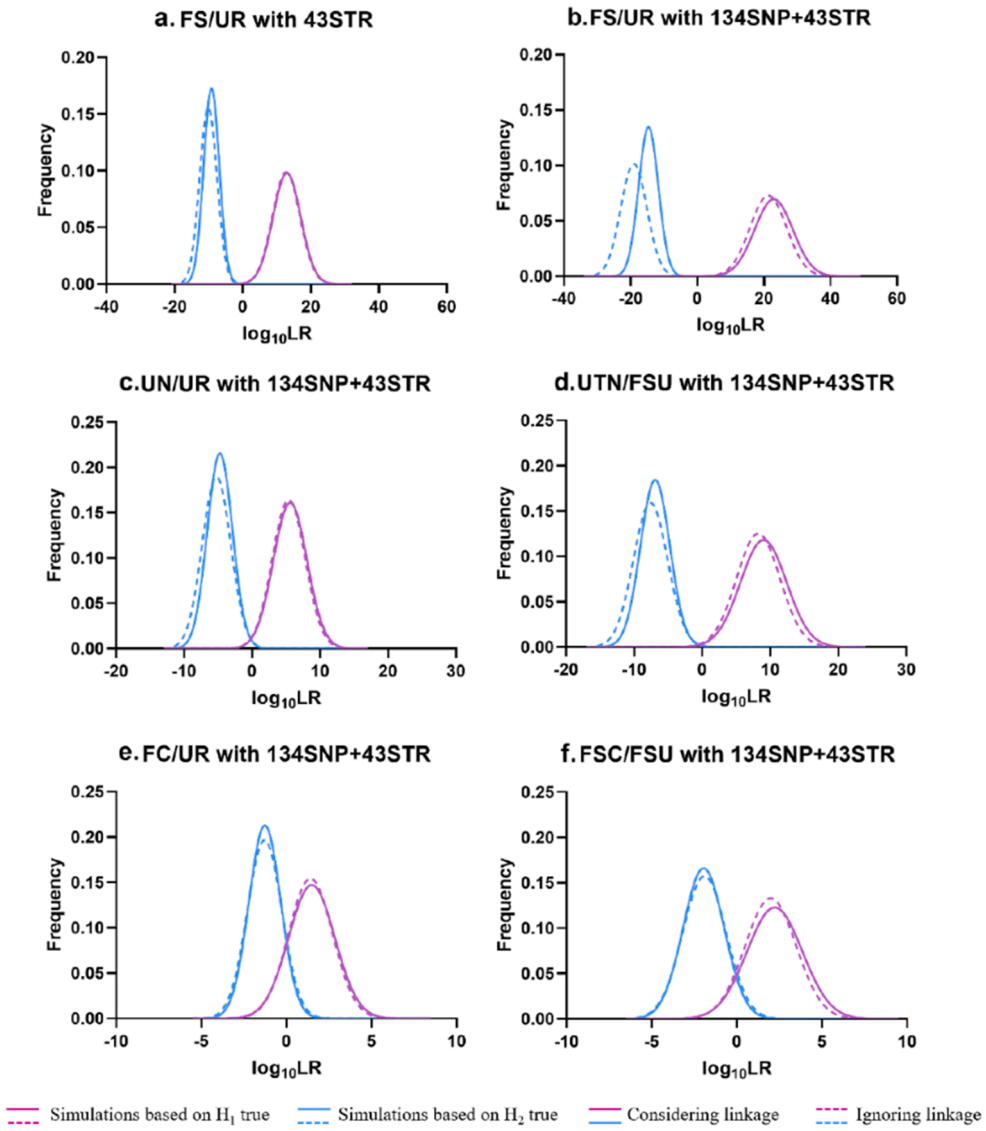

Fig. 2. Distributions of $LR_{linkage}$ and LR_{ignore} values for different kinship tests. FS: two full sibs UN: uncle–nephew FC: two first cousins UR: two unrelated UTN: uncle–two nephew FSC: two full siblings and one first cousin FSU: two full siblings and one unrelated.

decreased from 99.97 % to 99.95 %. Overall, accuracy increased, for example, at threshold 10. Accuracy increased from 98.16 % to 98.67 % with the use of 43 STRs + 134 SNPs, and when a full sibling was added, accuracy increased from 99.52 % to 99.71 %.

Additionally, for LR values of specific cases, as a full sibling was involved or the number of genetic markers increased, the impact of linkage on LR values significantly increased (Fig. 3c and d), which indicates that linkage had a greater impact on the LR values of cases. In 10,000 simulations based on H_1 true, when 43 STRs were used, the number of $LR_{linkage}/LR_{ignore} > 5$ was 16, with a maximum value of 8.0882. with the use of 143 SNPs + 43 STRs, this number became 1873. When a full sibling was involved, the number of $LR_{linkage}/LR_{ignore} > 5$ increased to 5571, and the number of $LR_{linkage}/LR_{ignore} > 100$ was 1021, with a maximum value of 34193.18. A similar trend was shown in 10,000 simulations based on H_2 true.

3.3. Third-Degree kinship analysis

When the number of genetic markers or the number of individuals involved increased, the ability for third-degree kinship analysis and the impact of linkage on LRs also increased. For 43 STRs, the impact of linkage on the distribution was not significant, with the differences between $LR_{linkage}$ accuracy and LR_{ignore} accuracy being < 0.50 % at all thresholds (1, 10, 100, and 1000) (Table 2). For 134 SNPs + 43 STRs, the impact of linkage on the LR distribution was more significant, with differences between $LR_{linkage}$ accuracy and LR_{ignore} accuracy at thresholds 10, 100, and 1000 being > 1 % (Table 2). When a full sibling was further added, the impact of linkage on the LR distribution based on H_1 true became more obvious, similar to the second-degree kinship (Fig. 2e and f). The differences between $LR_{linkage}$ accuracy and LR_{ignore} accuracy at thresholds 10, 100, and 1000 became > 2 % (Table 2). Sensitivity increased from 50.01 % to 56.37 % at threshold 100 and from 25.53 % to 33.08 % at threshold 1000, corresponding to differences of 6.36 % and 7.55 %, respectively. The differences between $LR_{linkage}$ parameters and LR_{ignore} parameters of third-degree kinship tests are more significant than those of second-degree kinship tests.

Fig. 3e and f reveals that the impact of linkage on LR values increased with the addition of genetic markers or relatives. In 10,000 simulated cases based on H_1 true, when 134 SNPs + 43 STRs were used, the number of $LR_{linkage}/LR_{ignore} > 20$ was 7, with a maximum value of 46.2892. With the addition of a full sibling, this number became 300, with a maximum value of 382.5946.

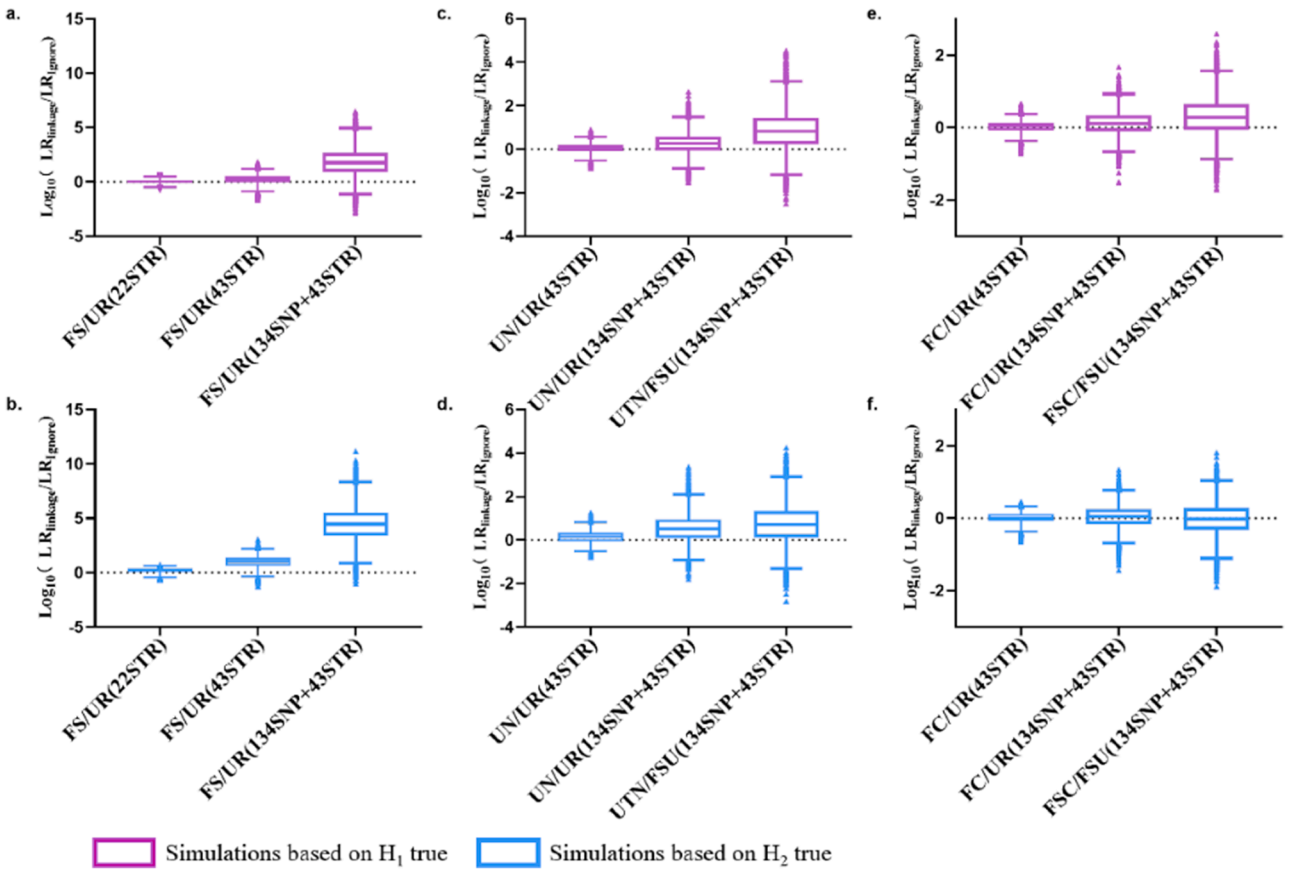

Fig. 3. Boxplots of $LR_{linkage}/LR_{ignore}$ for different kinship tests. FS: two full sibs UN: uncle–nephew FC: two first cousins UR: two unrelated UTN: uncle–two nephew FSC: two full siblings and one first cousin FSU: two full siblings and one unrelated.

Table 2
The parameter values based on differentiating second-degree relatives and third-degree relatives from unrelated individuals.

Loci (Relationship)	Thresholds ($Log_{10}LR$)	Sensitivity(%)		Specificity(%)		Accuracy (%)	
		$LR_{linkage}$	LR_{ignore}	$LR_{linkage}$	LR_{ignore}	$LR_{linkage}$	LR_{ignore}
43STR	0	97.10	96.71	97.25	97.34	97.18	97.03
(UN/UR)	1	90.88	90.59	99.52	99.51	95.2	95.05
	2	78.65	78.05	99.92	99.93	89.29	88.99
	3	61.01	60.24	99.99	99.99	80.50	80.12
134SNP + 43STR	0	99.04	98.57	99.26	99.31	99.15	98.94
(UN/UR)	1	97.52	96.47	99.81	99.85	98.67	98.16
	2	93.25	91.85	99.97	99.97	96.61	95.91
	3	86.37	83.99	100	100	93.19	92
134SNP + 43STR	0	99.84	99.6	99.89	99.92	99.87	99.76
(UTN/FSU)	1	99.46	99.07	99.95	99.97	99.71	99.52
	2	98.82	97.96	100	100	99.41	98.98
	3	97.28	95.78	100	100	98.64	97.89
43STR	0	82.19	81.95	85.43	84.95	83.81	83.45
(FC/UR)	1	46.80	46.07	98.54	98.55	72.67	72.31
	2	17.19	16.34	99.97	99.96	58.58	58.15
	3	3.88	3.38	100	100	51.94	51.69
134SNP + 43STR	0	88.33	87.23	90.35	89.89	89.34	88.56
(FSC/FSU)	1	64.78	62.81	98.96	98.81	81.87	80.81
	2	35.97	32.87	99.91	99.9	67.94	66.39
	3	14.89	11.62	100	100	57.45	55.81
134SNP + 43STR	0	93.86	91.71	94.71	93.77	94.29	92.74
(FSC/FSU)	1	79.44	75.41	99.20	99.02	89.32	87.22
	2	56.37	50.01	99.96	99.95	78.17	74.98
	3	33.08	25.53	99.99	100	66.54	62.77

UN: uncle-nephew UR: two unrelated UTN: uncle-two nephew FSU: two full sibs and one unrelated FC: two first cousins UR: two unrelated FSC: two full siblings and one first cousin FSU: two full siblings and one unrelated.

4. Discussion

In this study, we utilized the simulation modules of the software GeneVisa to investigate the linkage impact of 134SNP + 43STR. The effect of genetic linkage on the overall distribution of LR values is relatively small, but the effect on LR values of routine cases may be large. According to the study by Morimoto et al [8], based on various kinship scenarios, linkage between 21 STR loci of the GlobalFiler kit had a significant effect on routine cases, but has little impact on the distribution of LR values. According to this study, with the increase in the number of genetic markers and individuals involved in the kinship testing, the linkage impact on routine cases will increase (see Fig. 3), and eventually the linkage will have an impact on the distribution of LR values (see Fig. 2). For example, for the kinship testing of full siblings versus unrelated individuals, when we added 143 genetic markers, the linkage effect become larger (see Fig. 2a and Fig. 2b), for the kinship testing of uncle/nephew versus unrelated, when we add a full sibling of an involved person, the linkage effect on LR distribution of H_1 true simulations become more significant (see Fig. 2c and Fig. 2d).

In this study, several parameters were used, as described in reference [28], including sensitivity, specificity, and accuracy, to illustrate LR distributions. These parameters can also be used to evaluate the discriminatory power of genetic markers for various kinship tests. According to a previous study's description [28], sensitivity is equal to the positive rate that can be defined as the rate of LR > t in simulations based on H_1 true, whereas specificity is equal to the negative rate that can be defined as the rate of LR < t in simulations based on H_2 true. Further, the false positive rate (the rate of LR > t in simulations based on H_2 true) and the false negative rate (the rate of LR < t in simulations based on H_1 true) were obtained, and hence, we derived false positive rate is equal to 1 minus specificity and false negative rate is equal to 1 minus sensitivity.

For example, for the kinship testing i with two full siblings and two unrelated conducted using 43 STRs, the false negative rate was 0.31 % based on the LR_{ignore} values and 0.25 % based on the $LR_{linkage}$ values at threshold 100.

Additionally, based on the parameters sensitivity, specificity, and accuracy, two other parameters related to system power assessment the positive predictive value (PPV) and the negative predictive value (NPV) can be obtained. PPV represents the proportion of participants correctly determined as kinship H_1 and NPV represents the proportion of participants correctly determined as kinship H_2. In case of full sibling testing (two full siblings and two unrelated individuals), if we assume the predictive value depended on the proportion of full siblings in the population [28], these data cannot be obtained easily and may vary among different populations. In kinship testing, the statement of the involved individuals can offer some clues for kinship determinations. Thus, the proportion of full siblings in kinship testing is different from that in the population. During kinship testing, we assumed that the proportion of full sibling pairs and unrelated pairs are the same. We, thus, performed 10,000 simulations based on each H_1 true and H_2 true. According to the same number of simulations considering H_1 and H_2 true, we obtained the following expressions:

$$PPV = \frac{positive\ rate}{positive\ rate + \ false\ positive\ rate}$$

$$NPV = \frac{negative\ rate}{negative\ rate + false\ negative\ rate}$$

For example, for full sibling testing performed using 43 STRs, NPV was 99.31 % and 99.5 % based on LR_{ignore} and $LR_{linkage}$ values at threshold 1000, respectively.

The discriminatory power of 43 STRs and 134 SNPs in kinship testing were evaluated using the aforementioned parameters. For the full sibling testing, 43 STRs demonstrated a high discrimination power. When 43 STRs and 134 SNPs were used in a combination, the full sibling pairs were effectively identified from unrelated pairs. The sensitivity, specificity, and accuracy were 100 % at all thresholds under the condition with linkage. The 43 STRs + 134 SNPs also exhibited a high discriminatory power for second-degree kinship testing. For the kinship testing with uncle–nephew and two unrelated individuals, the accuracy exceeded 90 % at all thresholds. When we further added a relative (full sibling), the accuracy exceeded 97 % at all thresholds. However, for third-degree kinship testing, the discriminatory power of 43 STRs and 134 SNPs was relatively insufficient (Table 2). In this case, the distribution differences between $LR_{linkage}$ parameters and LR_{ignore} parameters at common thresholds is more significant than that in second-degree kinship testing for which 43 STRs and 134 SNPs exhibited a high discriminatory power.

Because the probability of mutation occurrence is very low, the hidden mutation is usually ignored in the routine kinship case under the condition that the genetic data are consistent with the inheritance law. In this study, mutations of STR loci were considered while simulating individual genotypes, but not while obtaining LR. In addition, this study does not consider LD and the population substructure. The impact of genetic linkage needs to be further investigated in the presence of hidden mutation [29] and LD and based on the population substructure [30].

In conclusion, as the number of available genetic markers increased, linkage between genetic markers became inevitable. GeneVisa is a useful tool for dealing with linkage for kinship testing. In this study, involving genetic linkage had no significant effect on LR distributions, but may have large effect on routine cases. For a specific kinship testing, the linkage effect on LR distributions and routine cases tended to increase with the number of genetic markers used and that of individuals involved. Linkage was considered to improve the discriminatory power of genetic markers for kinship testing. Our study provides a guideline for considering linkage in kinship testing.

CRediT authorship contribution statement

Da Yang: Writing – original draft, Writing – review & editing, Project administration, Software, Methodology, Funding acquisition. **Sheng Xuan Ma:** Writing – original draft, Formal analysis, Resources. **Guo Liang Zhao:** Formal analysis, Data curation, Methodology. **Ao Gao:** Formal analysis, Resources. **Zhao Kun Xu:** Funding acquisition, Validation, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to JiNing Medical University for supporting this study. This work was supported by Scientific Research Support Fund for teachers of JiNing Medical University with the number JY2017FY010, Education and Teaching Research Project of JiNing Medical University, with the number Y2020042 and the Student's Platform for Innovation Training Program with the number cx2022215.

References

[1] J. Ott, Analysis of Human Genetic Linkage, JHU, Press, Baltimore, Maryland, 1985.
[2] M. Slatkin, Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future, Nat Rev Genet. 9 (2008) 477–485, https://doi.org/10.1038/nrg2361.
[3] A.O. Tillmar, C. Phillips, Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets, Forensic Sci. Int. Genet. 26 (2017) 58–65, https://doi.org/10.1016/j.fsigen.2016.10.007.

[4] R. Li, B. Budowle, H.Y. Sun, J.Y. Ge, Linkage and linkage disequilibrium among the markers in the forensic MPS panels, J Forensic Sci. 66 (2021) 1637–1646, https://doi.org/10.1111/1556-4029.14724.
[5] D.W. Gjertson, C.H. Brenner, M.P. Baur, A. Carracedo, F. Guidet, J.A. Luque, et al ISFG: recommendations on biostatistics in paternity testing. Forensic Sci Int Genet. 1 (2007) 223–231. doi: 10.1016/j.fsigen.2007.06.006.
[6] A.O. Tillmar, D. Kling, J.M. Butler, W. Parson, M. Prinz, P.M. Schneider, T. Egeland, L. Gusmão, DNA Commission of the International Society for forensic genetics (ISFG): guidelines on the use of X-STRs in kinship analysis, Forensic Sci. Int. Genet. 29 (2017) 269–275, https://doi.org/10.1016/j.fsigen.2017.05.005.
[7] D. Kling, T. Egeland, A.O. Tillmar, FamLink-a user friendly software for linkage calculations in family genetics, Forensic Sci. Int. Genet. 6 (2012) 616–620, https://doi.org/10.1016/j.fsigen.2012.01.012.
[8] C. Morimoto, H. Tsujii, S. Manabe, S. Fujimoto, E. Hirai, Y. Hamano, K. Tamaki, Development of a software for kinship analysis considering linkage and mutation based on a Bayesian network, Forensic Sci. Int. Genet. 47 (2020) 102279, https://doi.org/10.1016/j.fsigen.2020.102279.
[9] Q.Z. Zhang, Z. Zhou, L. Wang, C. Quan, Q.Q. Liu, Z.B. Tang, L.Y. Liu, Y.C. Liu, S. Q. Wang, Pairwise kinship testing with a combination of STR and SNP loci, Forensic Sci. Int. Genet. 46 (2020) 102265, https://doi.org/10.1016/j.fsigen.2020.102265.
[10] T. Tamura, M. Osawa, Y. Kakimoto, E. Ochiai, T. Suzuki, T. Nakamura, Combined effects of multiple linked loci on pairwise sibling tests, Int. J. Leg. Med. 131 (2017) 95–102, https://doi.org/10.1007/s00414-016-1491-4.
[11] C. Morimoto, S. Manabe, C. Kawai, S. Fujimoto, K. Tamaki, Effect of linkage on sibship determination based on likelihood ratio, Forensic Sci. Int. Genet Suppl. 5 (2015) 126–127.
[12] J.A. Bright, J.M. Curran, J.S. Buckleton, Relatedness calculations for linked loci incorporating subpopulation effects, Forensic Sci. Int. Genet 7 (2013) 380–383, https://doi.org/10.1016/j.fsigen.2013.03.002.
[13] P. Gill, C. Phillips, C. McGovern, J.A. Bright, J. Buckleton, An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees, Forensic Sci. Int Genet. 6 (2012) 477–486, https://doi.org/10.1016/j.fsigen.2011.11.001.
[14] D. Yang, Pedigree likelihood formulae based on founder and founder couple symmetry and validation of DNA testing software, Forensic Sci. Int. Genet. 62 (2023) 102782, https://doi.org/10.1016/j.fsigen.2022.102782.
[15] R.C. Elston, J. Stewart, A general model for the genetic analysis of pedigree data, Hum Hered. 21 (1971) 523–542, https://doi.org/10.1159/000152448.
[16] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. u. s. a. 84 (1987) 2363–2367, https://doi.org/10.1073/pnas.84.8.2363.
[17] M. Fishelson, D. Geiger, Exact genetic linkage computations for general pedigrees, Bioinformatics 18 (1) (2002) S189–S198, https://doi.org/10.1093/bioinformatics/18.suppl_1.s189.
[18] G.R. Abecasis, S.S. Cherny, W.O. Cookson, L.R. Cardon, Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (2002) 97–101, https://doi.org/10.1038/ng786.
[19] C. Bhérer, C.L. Campbell, A. Auton, Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales, Nat. Commun. 8 (2017) 14994, https://doi.org/10.1038/ncomms14994.
[20] D. Kling, A. Tillmar, T. Egeland, P., Mostad a general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations, Int. J. Legal. Med. 129 (2015) 943–954, https://doi.org/10.1007/s00414-014-1117-7.
[21] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, Forensic Sci. Int. Genet. 29 (2017) 193–204, https://doi.org/10.1016/j.fsigen.2017.04.011.
[22] S. Iyavooa, O. Afolabia, B. Boggia, A. Bernotaiteb, T. Haizel, Population genetics data for 22 autosomal STR loci in european, south asian and african populations using SureID® 23comp human DNA identification kit, Forensic Sci. Int. 301 (2019) 174–181, https://doi.org/10.1016/j.forsciint.2019.05.033.
[23] C. Phillips, L. Fernandez-Formoso, M. Gelabert-Besada, M. García-Magariños, J. Amigo, A. Carracedo, M.V. Lareu, Global population variability in qiagen investigator HDplex STRs forensic sci, Int. Genet. 8 (2014) 36–43, https://doi.org/10.1016/j.fsigen.2013.07.006.
[24] S. Turrina, M. Ferrian, S. Caratti, D.D. Leo, Evaluation of genetic parameters of 22 autosomal STR loci (PowerPlex® fusion system) in a population sample from northern Italy, Int. J. Leg. Med. 128 (2014) 281–283, https://doi.org/10.1007/s00414-013-0934-4.
[25] D. Kling, A.O. Tillmar, T. Egeland, Familias 3 - extensions and new functionality, Forensic Sci. Int. Genet. 13 (2014) 121–127, https://doi.org/10.1016/j.fsigen.2014.07.004.
[26] Q. Zhang, X. Wang, P. Cheng, S. Yang, W. Li, Z. Zhou, S. Wang, Complex kinship analysis with a combination of STRs, Snps, and Indels, Forensic Sci Int Genet. 61 (2022) 102749, https://doi.org/10.1016/j.fsigen.2022.102749.
[27] E.M. Dauber, A. Kratzer, F. Neuhuber, W. Parson, M. Klintschar, W. Bär, W. R. Mayr, Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions, Forensic Sci Int Genet. 6 (2012) 381–386, https://doi.org/10.1016/j.fsigen.2011.07.015.
[28] R. Gaytmenn, D.P. Hildebrand, D. Sweet, I.A. Pretty, Determination of the sensitivity and specificity of sibship calculations using AmpFISTR profiler plus, Int. J. Leg. Med. 116 (2002) 161–164, https://doi.org/10.1007/s00414-001-0273-8.
[29] K. Slooten, F. Ricciardi, Estimation of mutation probabilities for autosomal STR markers, Forensic Sci Int Genet 7 (2013) 337–344, https://doi.org/10.1016/j.fsigen.2013.01.006.
[30] D.J. Balding, R.A. Nichols, A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, Genetica 96 (1995) 3–12, https://doi.org/10.1007/BF01441146.