International Journal of Legal Medicine (2021) 135:1369–1374
https://doi.org/10.1007/s00414-020-02495-7

ORIGINAL ARTICLE

Selection and evaluation of bi-allelic autosomal SNP markers for paternity testing in Koreans

Soyeon Bae[1] · Sohyoung Won[2] · Heebal Kim[1,2,3]

Received: 6 July 2020 / Accepted: 17 December 2020 / Published online: 28 April 2021

Abstract

Due to the advantages of single-nucleotide polymorphisms (SNPs) in forensic science, many forensic SNP panels have been developed. However, the existing SNP panels have a problem that they do not reflect allele frequencies in Koreans or the number of markers is not sufficient to perform paternity testing. Here, we filtered candidate SNPs from the Ansan-Ansung cohort data and selected 200 SNPs with high allele frequencies. To reduce the risk of false inclusion and false exclusion, we calculated likelihood ratios of alleged father-child pairs from simulated families when the alleged father is the true father, the close relative of the true father, and the random man. As a result, we estimated that 160 SNPs were needed to perform paternity testing. Furthermore, we performed validation using Twin-Family cohort data. When 160 selected SNPs were used to calculate the likelihood ratio, paternity and non-paternity were accurately distinguished. Our set of 160 SNPs could be useful for paternity testing in Koreans.

Keywords Single-nucleotide polymorphism · Paternity testing · Korean · False inclusion

Introduction

In modern forensic science, DNA profiling has become an important tool for human identification and paternity testing. Short tandem repeat (STR) markers, usually composed of 13-17 loci, and recently expanded to 21 or more loci, have generally been used for DNA profiling [1–3]. However, advances in sequencing technologies have enabled the production of large amounts of single-nucleotide polymorphism (SNP) data, and this led to a discussion about the availability of SNP markers in the field of forensic science.

Compared to STRs, SNP markers have the advantage of low mutation rate, small amplicon size, which is advantageous for analysis of degraded samples, and fast and automated analysis [4, 5]. On the other hand, more SNPs are needed to approach the match probability of STR panels since bi-allelic SNPs are less polymorphic than STRs. Krawczak [6] and Gill [7] reported that 50-60 SNPs with allele frequencies close to 0.5 are required to have the same discriminatory power as STR panels. Ayres [8] suggested that the number of SNPs with allele frequencies in the range [0.3, 0.5] required for the standard trio (father-mother-child) case and duo (father-child) case is 50-60 and 70-80, respectively. However, these studies assumed the use of independent markers. When the number of markers increases, the probability of genetic linkage increases. Since the use of markers that are not independently transmitted can affect the results of the forensic analysis, linkage should be considered in forensic calculations [9, 10].

Several bi-allelic autosomal marker panels, such as the SNPforID multiplex (52 SNPs) [11] and the IISNP panel (86 SNPs) [12–14], were developed for human identification and paternity testing. However, if the alleged father (AF) is the close relative of the true father (TF), there may be cases where the number of SNP loci used in existing panels is not enough to perform paternity testing [15]. In addition, these panels were selected based on allele frequencies of various human populations. As allele frequencies may vary by population, markers selected based on allele frequencies of a certain population may not sufficiently reflect allele frequencies of another population. Paternity testing using incorrect allele frequencies can lead to erroneous results [16, 17]. Thus, several studies have developed forensic SNP panels for a specific

✉ Heebal Kim
heebal@snu.ac.kr

[1] Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Republic of Korea

[2] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Republic of Korea

[3] eGnome, Inc., Seoul, Republic of Korea

population [18, 19]. Lee et al. [20] and Kim et al. [21] selected highly informative SNPs from Koreans for forensic purposes and provided a database, but the number of markers was 24 and 30, respectively, which was insufficient to perform paternity testing.

In this study, we aimed to select bi-allelic autosomal SNP markers for paternity testing for Korean individuals based on likelihood ratio (LR) principles, where genetic evidence is evaluated by calculating the LR [22]. Korean SNP data were screened to collect candidate markers. Allele frequencies of retained SNPs were calculated, and based on this information, we selected the appropriate number of markers using simulated family data. Moreover, we examined the performance of final set of SNPs in real cases.

Materials and methods

Sample collection

We used SNP genotyping data from the Ansan-Ansung cohort and the Twin-Family cohort, which were part of the Korean Genome and Epidemiology Study (KoGES) [23]. DNA was extracted from blood samples collected from individuals. The participants of the Ansan-Ansung cohort study were 10,030 adults aged 40 to 69, who live in Ansan or Ansung. Among them, 8840 individuals were genotyped with the Affymetrix Genome-Wide Human SNP Array 5.0. The Twin-Family cohort study, consisting of 3202 twins and their families, collected SNP data from 1716 individuals using the Affymetrix Genome-Wide Human SNP Array 6.0.

Quality control and SNP selection

In this study, the selection of candidate SNP markers and calculation of allele frequencies were based on the Ansan-Ansung cohort data, and the performance of markers was evaluated using the Twin-Family cohort data. The Ansan-Ansung cohort data included 352,228 bi-allelic autosomal SNPs. Among them, SNPs not included in the Twin-Family cohort data were excluded. To avoid the influence of selection pressure, SNPs within the range of a gene list (hg19) were discarded. Quality control (QC) steps were performed as follows: (1) Samples with individual missing rates higher than 0.05 were filtered. (2) SNPs with missing genotype rates higher than 0.01 were removed. (3) SNPs that deviated from the Hardy-Weinberg equilibrium (p value < 10^{-5}) were removed. Next, we estimated kinship coefficients to identify potential relatives. Samples were filtered until kinship coefficients of all pairs of individuals were lower than 0.0884, meaning that all pairs were treated as third-degree or more distant relationships. The fixation index (F_{ST}) was calculated between Ansan and Ansung populations. Then, we performed linkage disequilibrium (LD)-based SNP pruning with the following parameters: window size = 500, step size = 50, and r^2 threshold = 0.01. Finally, 200 candidate SNPs with the highest minor allele frequencies (MAFs) were selected among the retained SNPs. The minimum distance between candidate SNPs in each chromosome was 10 Mbp. PLINK v1.90 was used to conduct QC steps and LD pruning and to calculate F_{ST} and MAF [24]. Kinship coefficient was estimated using KING v2.2.4 [25].

Testing in simulated pedigrees

In order to select the appropriate number of markers needed for the paternity test, we simulated 10,000 pedigrees using MERLIN v1.1.2 [26]. Since this program requires centimorgan (cM) position for each SNP, we obtained this information from the genetic map of the CHB (Han Chinese in Beijing, China) population of the 1000 Genomes Project (available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/). If there was no information about the genetic position of the SNP, the genetic position of the nearest SNP was used. The structure of simulated pedigrees is shown in Fig. 1. Founder genotypes were randomly generated based on the previously calculated allele frequencies of candidate SNPs. Then, it was assumed that each parent contributes one allele to the offspring. Alleles spaced less than 25 cM were clustered and passed to the offspring based on the estimated haplotype frequencies.

In paternity testing, we evaluated genetic evidence by comparing likelihoods of hypotheses using the equation: LR = $\Pr(G|H_{\text{parent-child}})/\Pr(G|H_{\text{unrelated}})$, where G is the observed genotype data, $H_{\text{parent-child}}$ is the hypothesis that two tested people are in a parent-child relationship, and $H_{\text{unrelated}}$ is the hypothesis that two tested people are unrelated [22]. We calculated LRs for each marker and multiplied them all to obtain final LRs for a set of markers using a "likelihoodMerlin" function of "pedprobr" package in R [26, 27]. This function

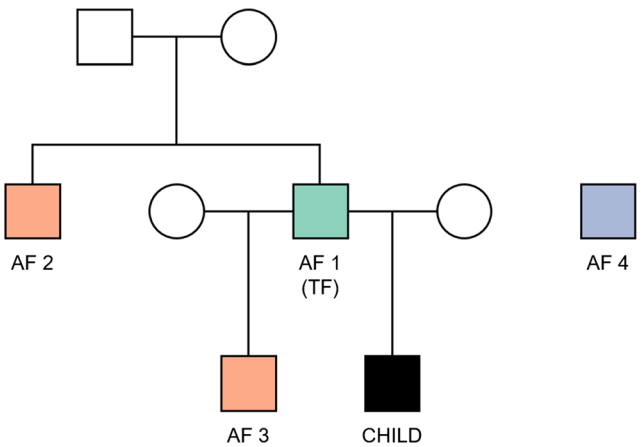

Fig. 1 Structure of simulated pedigrees (AF 1: TF, AF 2: brother of TF, AF 3: child of TF, AF 4: random man)

can take into account linkage when calculating LRs. Ayres [8] estimated that at least 70 SNP loci were required for paternity testing in the motherless duo case, so the number of markers used in the analysis was increased from 70 to 200 by 10. When the $\log_{10}$LR was greater than or equal to 5, which provides very strong support for $H_{\text{parent-child}}$ [9], AF and child were judged to be in a parent-child relationship. The accuracy is defined as the percentage by which the relationship of two individuals is correctly determined as parent-child or non-parent-child. The false positive is an error in which a person who is not the TF is falsely included as the TF. In contrast, the false negative means that the TF is falsely excluded.

Validation in real cases

The Twin-Family cohort data was used to examine the performance of selected markers. Individuals with missing genotype data of these SNPs were filtered out. We collected all pairs of two individuals who were in parent-child or unrelated relationships. We also used second-degree relative pairs, such as uncle-nephew or grandparent-grandchild, to check LR values when the AF is the close relative of the TF. The methods used to decide paternity and calculate the accuracy, false positive, and false negative were written above. The method of maternity testing was the same as that of paternity testing, except that only the gender of the typed person is female. Thus, we calculated LRs for all pairs of individuals, ignoring their gender.

Results

Candidate SNP selection

The number of bi-allelic autosomal SNPs included in both data was 280,905. Since functional markers have a possibility that the selection pressure affects the allele frequency [9], 125,485 SNPs in the gene region were excluded to avoid this. A total of 12,238 SNPs with a genotyping rate lower than 0.99 were removed. The total genotyping rate of the retained SNPs from the Ansan-Ansung cohort was 0.9987. One hundred ninety-three SNPs were removed due to failure to pass the Hardy-Weinberg exact test. Among 8840 participants with a genotyping rate per individual higher than 0.95, 8621 unrelated samples were selected. All SNPs had F_{ST} values lower than 0.01, so it could be considered that there was no significant genetic difference between the populations in these regions. Finally, 10,538 independent (pairwise $r^2 < 0.01$) SNPs were retained by the LD pruning method.

To select highly informative SNPs, we calculated MAFs of 10,538 SNPs from 8621 unrelated samples and selected 200 SNPs located far from each other (> 10 Mbp) with a high MAF. These SNPs had an MAF in the range [0.49, 0.5].

Paternity testing in simulated pedigrees

A total of 10,000 families were generated to determine the appropriate number of markers for paternity testing. Within each family, we were able to collect four types of AF-child pairs as shown in Fig. 1: The AF was the TF (AF 1), the AF was the close relative of TF (AF 2 and 3), and the AF was the unrelated person (AF 4). LRs were calculated for each pair, using 70, 80, …, and 200 SNPs.

Table 1 shows the accuracy, false-positive rate, and false-negative rate in simulated duo cases. Since we used a high value of LR cutoff, the false-negative rate was very high with 70-80 loci, which were suggested in a previous study [8]. When the number of loci was 150, no TF was falsely excluded. However, there were some cases that the first-degree relative of the TF was judged to be the TF, so the false-positive rate was 0.0033%. One hundred percent accuracy was achieved when 160 or more SNPs were used. As a result, we selected a set of 160 SNPs for paternity testing without errors. The details of the selected SNPs are shown in Supplementary Table 1.

Figure 2 shows the distribution of $\log_{10}$LR values for true parent-child pairs. The average, minimum, and maximum values of the $\log_{10}$LR for parent-child pairs were 12.05, 6.02, and 18.98, respectively. In non-parent-child pairs, LRs were all zero due to SNP mismatches between the two tested people.

Paternity testing in real cases

The family genotype data included 1716 samples. Among them, 816 samples, whose genotypes of finally selected SNPs were observed, were used for validation. Out of a total of 332,520 pairs, those with unknown or uncertain relationships were excluded from the results. Finally, there were 295, 19, and 331,778 pairs of parent-child, second-degree relative,

Table 1 Summary of simulated results

Number of markers	Accuracy (%)	False-positive rate (%)	False-negative rate (%)
70	89.72	0.18	40.58
80	95.0725	0.22	19.05
90	98.0975	0.17	7.1
100	99.3625	0.1033	2.24
110	99.7825	0.08333	0.62
120	99.945	0.03333	0.12
130	99.975	0.01667	0.05
140	99.9925	0.006667	0.01
150	99.9975	0.003333	0
160 ≤	100	0	0

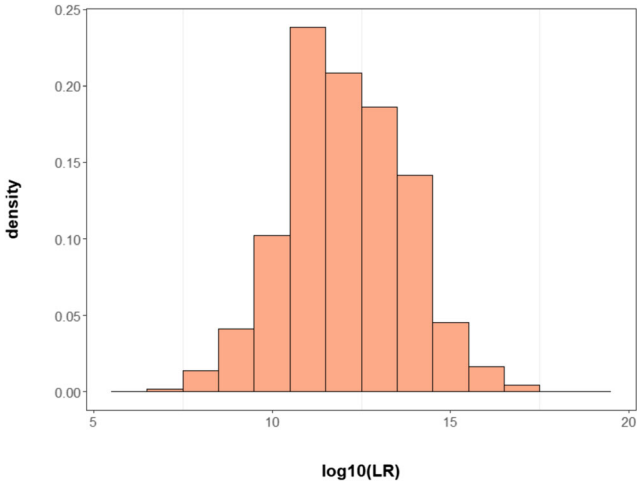

Fig. 2 Distribution of $\log_{10}$LR values for true parent-child pairs in simulation results using 160 SNPs

and unrelated individuals, respectively. When paternity testing was performed using 160 selected SNPs, the accuracy reached 100%. The LR values for non-parent-child pairs were zero, and the average, minimum, and maximum values of the $\log_{10}$LR for parent-child pairs were 12.18, 8.42, and 19.26, respectively (Fig. 3). We also identified the number of opposite homozygosity for 160 SNPs, which means the child and the AF are homozygous and have different alleles; for example, the child has allele AA and the AF has allele BB [15, 28]. The more mismatches are observed, the less likely the two individuals are in parent-child relationships. No mismatch was found in any of the 160 loci in all true parent-child relationships. When the AF was the first-degree relative of the TF and the unrelated man, the average number of mismatches between the AF and the child was 10.95 and 20, respectively.

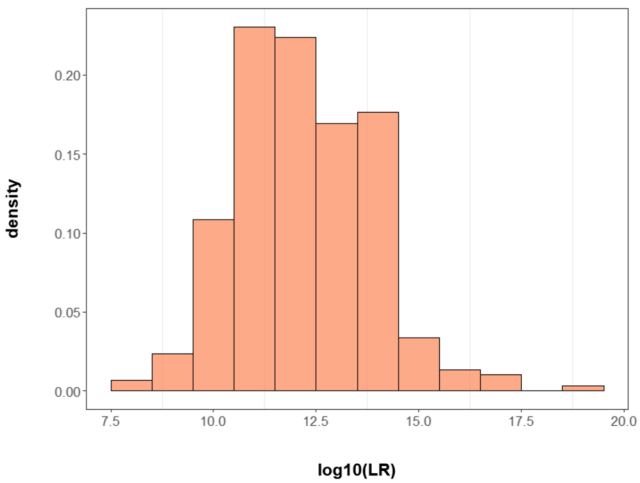

Fig. 3 Distributions of $\log_{10}$LR values for true parent-child pairs in real cases using 160 SNPs

Discussion

After the usefulness of SNP-based human identification and paternity testing was discussed, several sets of forensic SNP markers were developed. SNPforID [11] and IISNP [14] are universal forensic SNP panels for various populations. However, SNPforID panel consists of 52 loci, which is an insufficient number of markers to perform paternity testing of duo cases [8]. Børsting et al. [28] observed that false association occurred in some duo cases when using SNPforID panel. In addition, according to NCBI dbSNP (https://www.ncbi.nlm.nih.gov/snp/), 19 and 7 SNPs in SNPforID and IISNP, respectively, had an MAF value lower than 0.3 in East Asians based on 1000 Genomes Project data. SNPs with a low MAF are less informative and may not be the best choice for forensic analysis in East Asians (Supplementary Fig. 1). Furthermore, since there are various populations within East Asians, it is unclear whether the existing allele frequency database is accurate for Koreans. It is important to accurately estimate allele frequencies of the population to reduce errors in forensic analysis [16, 17]. Although several studies have selected forensic SNP marker sets for Koreans and provided allele frequency information [20, 21], these panels are expected to be unsuitable for paternity testing because they consist of fewer than 50 SNPs, which are suggested to be needed for the analysis of trio cases [8].

In the present study, we selected and tested the appropriate number of bi-allelic autosomal markers for paternity testing in Korean individuals. We considered difficult cases when choosing the number of markers. There were special cases where false inclusion occurred when the TF was a close relative of the AF [29, 30]. These problems were solved by supplementing additional markers [31, 32]. We aimed to solve these problems with only autosomal SNPs by selecting a sufficient number of loci and focus on the duo case because there are special cases where genotype of one of the parents is not available.

Of 352,228 SNPs, 200 candidates were selected from 8621 unrelated Korean samples after filtering processes. These markers were non-functional, and had a high MAF (≥ 0.49) and an F_{ST} (< 0.01) value between Ansan and Ansung. To minimize the effects of genetic linkage and LD, we selected only SNPs located far from each other with a low level of LD ($r^2 < 0.01$) between different loci in the population. However, it was still not far enough to assume that these markers were transmitted independently. Thus, we calculated LRs by considering genetic distances from the genetic map of the East Asian population (Han Chinese in Beijing, China). Based on allele frequencies and genetic positions of 200 candidate SNPs, we randomly generated 10,000 families and calculated the LR for parentage. Based on our simulation results, we finally selected highly informative 160 SNP loci to remove falsely included cases. Using these final set of 160 SNPs, all

332,092 comparisons in real cases were determined for paternity and non-paternity.

In summary, we selected 160 SNPs for paternity testing based on allele frequencies in Koreans. Our study showed that using 160 autosomal SNPs with an MAF close to 0.5 in paternity testing would be sufficient to remove the risk of false inclusion. Considering that SNP has a lower mutation rate, which reduces the probability of false exclusion, our final set of SNPs seems to be useful for paternity testing.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-020-02495-7.

Acknowledgments This study was conducted with bioresources from the National Biobank of Korea, the Centers for Disease Control and Prevention, Republic of Korea (KBN-2019-003).

Data availability The datasets analyzed during the current study are available on request from the National Institute of Health, http://www.nih.go.kr/contents.es?mid=a50401010100.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval We used de-identified samples from the National Biobank of Korea, the Centers for Disease Control and Prevention, Republic of Korea. All participants provided written informed consent. This study was granted exemption by the Institutional Review Board of Seoul National University (IRB No. E2003/001-001) and approved by the National Biobank of Korea (KBN-2019-003).

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Not applicable.

References

1. Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 51(2): 253–265. https://doi.org/10.1111/j.1556-4029.2006.00046.x
2. Ensenberger MG, Hill CR, McLaren RS, Sprecher CJ, Storts DR (2014) Developmental validation of the PowerPlex® 21 System. Forensic Sci Int Genet 9:169–178
3. Kraemer M, Prochnow A, Bussmann M, Scherer M, Peist R, Steffen C (2017) Developmental validation of QIAGEN Investigator® 24plex QS Kit and Investigator® 24plex GO! Kit: two 6-dye multiplex assays for the extended CODIS core loci. Forensic Sci Int Genet 29:9–20
4. Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5(10):739–751
5. Butler JM, Coble MD, Vallone PM (2007) STRs vs. SNPs: thoughts on the future of forensic DNA testing. Forensic Sci Med Pathol 3(3):200–205
6. Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. Electrophoresis 20(8):1676–1681
7. Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med 114(4-5):204–210
8. Ayres KL (2005) The expected performance of single nucleotide polymorphism loci in paternity testing. Forensic Sci Int 154(2-3): 167–172
9. Buckleton J, Triggs C, Walsh S (2005) Forensic DNA evidence interpretation. CRC Press, Boca Raton
10. Bright J-A, Curran JM, Buckleton JS (2013) Relatedness calculations for linked loci incorporating subpopulation effects. Forensic Sci Int Genet 7(3):380–383
11. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27(9):1713–1724
12. Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SL, Karoma NJ, Kungulilo S, Kim J-J, Lu R-B, Odunsi A (2006) Developing a SNP panel for forensic identification of individuals. Forensic Sci Int 164(1):20–32
13. Pakstis AJ, Speed WC, Kidd JR, Kidd KK (2007) Candidate SNPs for a universal individual identification panel. Hum Genet 121(3-4): 305–317
14. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, Kidd KK (2010) SNPs for a universal individual identification panel. Hum Genet 127(3):315–324
15. Børsting C, Morling N (2012) Reinvestigations of six unusual paternity cases by typing of autosomal single-nucleotide polymorphisms. Transfusion 52(2):425–430
16. Rohlfs RV, Fullerton SM, Weir BS (2012) Familial identification: population structure and relationship distinguishability. PLoS Genet 8(2):e1002469
17. Karlsson AO, Holmlund G, Egeland T, Mostad P (2007) DNA-testing for immigration cases: the risk of erroneous conclusions. Forensic Sci Int 172(2-3):144–149
18. Augustinus D, Gahan ME, McNevin D (2015) Development of a forensic identity SNP panel for Indonesia. Int J Legal Med 129(4): 681–691
19. Sarkar A, Nandineni MR (2017) Development of a SNP-based panel for human identification for Indian populations. Forensic Sci Int Genet 27:58–66
20. Lee HY, Park MJ, Yoo J-E, Chung U, Han G-R, Shin K-J (2005) Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. Forensic Sci Int 148(2-3):107–112
21. Kim J-J, Han B-G, Lee H-I, Yoo H-W, Lee J-K (2010) Development of SNP-based human identification system. Int J Legal Med 124(2):125–131
22. Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, Lessig R, Mayr WR, Pascali VL, Prinz M (2007)

ISFG: recommendations on biostatistics in paternity testing. Forensic Sci Int Genet 1(3-4):223–231
23. Kim Y, Han B-G, Group K (2017) Cohort profile: the Korean genome and epidemiology study (KoGES) consortium. Int J Epidemiol 46(2):e20–e20
24. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4(1):s13742-13015-10047-13748
25. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867–2873
26. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30(1):97–101
27. Vigeland MD (2020) pedprobr: probability computations on pedigrees. R package version 0.4.0. https://CRAN.R-project.org/package=pedprobr
28. Børsting C, Sanchez JJ, Hansen HE, Hansen AJ, Bruun HQ, Morling N (2008) Performance of the SNPforID 52 SNP-plex assay in paternity testing. Forensic Sci Int Genet 2(4):292–300
29. von Wurmb-Schwark N, Mályusz V, Simeoni E, Lignitz E, Poetsch M (2006) Possible pitfalls in motherless paternity analysis with related putative fathers. Forensic Sci Int 159(2-3):92–97
30. Dogan M, Murat Canturk K, Emre R, Kara U, Filoglu G (2017) Demonstration of false inclusion risks of duo parentage analyses in the Turkish population in light of parentage acceptance criteria. Aust J Forensic Sci 49(3):326–331
31. Junge A, Brinkmann B, Fimmers R, Madea B (2006) Mutations or exclusion: an unusual case in paternity testing. Int J Legal Med 120(6):360–363
32. Tomas C, Sanchez JJ, Castro JA, Børsting C, Morling N (2010) Forensic usefulness of a 25 X-chromosome single-nucleotide polymorphism marker set. Transfusion 50(10):2258–2265