Effects of Reference Population and Number of STR Markers on Positive Evidence in Paternity Testing

Rossana Moroni1,2,*, Dario Gasbarra2, Elja Arjas1,2, Matti Lukka1 and Ismo Ulmanen1
1Forensic Genetic Paternity Testing Unit, National Institute for Health and Welfare, Helsinki, Finland
2Helsinki University, Department of Mathematics and Statistics, P.O. Box 68, FIN-00014, Finland

Abstract

Three sets of commonly used autosomal short tandem repeat (STR) markers (containing 15, 10 and 9 markers) and 14 databases from populations belonging to Africa, America, Asia and Europe were used to investigate how the selection of the population database and the number of considered markers would influence the statistical evidence that is usually produced to favour paternity. The study was based on a sample of 100 randomly chosen Finnish paternity trios collected during paternity testing case work and without any exclusion after use of 15 STR markers. Paternity Index, Probability of Paternity, Typical Paternity Index and Probability of Exclusion were computed and descriptive statistics were provided separately for trios (mother, child and putative father) and duos (obtained from trios but not considering the genetic information of the mother). This was done for all combinations of markers and databases. In trio cases the differences between results obtained are not statistically significant. However, especially in duos the cases use of 15 STR markers is recommended.

Keywords: Paternity Testing; Reference Population; Database Effect; Marker Effect

Abbreviations: STR: Short Tandem Repeat; PF: Putative Father; M: Mother; C: Child; PI: Paternity Index; W: Probability of Paternity; PE: Probability of Exclusion; TPI: Typical Paternity Index; RMNE: Random Man Not Excluded

Introduction

In recent years a growing number of disputed paternities have involved parents from different ethnic backgrounds. We examined how sensitive the Paternity Index (PI) [1], the Probability of Paternity (W) [2] and the Probability of Exclusion (PE) [1-3] are to the selection of the population STR database (‘Database Effect’) and to the number of considered STR markers (‘Marker Effect’). To evaluate this we performed paternity tests on 100 Finnish trios and duos with three sets of commonly used STR markers (containing 15, 10 and 9 markers) and 14 population databases, representing different marker allele frequencies. Such frequencies are estimated for each specific population and will generally vary between populations. As a consequence, calculating probabilities using different population data means differences in the probabilities’ values. Concerning the ‘Database Effect’ we carried out a comparative statistical analysis of PI and W which were at first calculated considering the allele frequencies of the putative father’s own population (Finnish), and then using some other reference population databases. A good measure of this effect is obtained from the analysis of Typical Paternity Index (TPI) [4]. Furthermore, we establish whether there is a correlation between a measure of genetic distance between population (Nei’s distance) [5] and the Typical Paternity Index Ratio (TPI Ratio). Some used measures will be more extensively explained in the materials and methods section. Concerning the ‘Marker Effect’ we investigated the changes in the PI when performing the test using only the Finnish allele frequencies but considering three different sets of markers. Forensic paternity testing is conducted using DNA profiles which consist of genotyping of several highly polymorphic short tandem repeat (STR) markers, chosen on different chromosome in order to ensure an independent segregation. Mutation [6] can sometimes cause an apparent discrepancy with the rules of inheritance between parent and child, but the incompatibility in at least three independent DNA markers is widely seen as sufficient evidence to exclude paternity. In the simplest case of disputed paternity a man, denoted PF, is claimed, by a mother M, to be the true father of a child C. According to [1], in a standard paternity case, two hypotheses are competing: a prosecution hypothesis (H$_p$) where PF is supposed to be the true father of C, and a defendant one (H$_d$) where some other man is the true father of C. Hence, if PF is not the true father of C, then an unknown man, randomly drawn from a population, which we call reference population, is assumed to be the true father of C, and we suppose him to be unrelated to both PF and M.

Materials and Methods

A total of 100 Finnish standard ‘Trios’ paternity cases (putative father, mother and child) with no genetic inconsistencies between mother and child or between father and child, were collected during paternity testing case work. The Finnish origin of fathers and mothers was based on surnames. The genotyping of the STR-loci was performed with AmpFISTR Profiler kit (9 markers), AmpFISTR SGM Plus kit (10 markers) and with combination of Profiler and SGM kits (15 markers: TH01, D3S1358, FGA, TPOX, CSF1PO, D5S818, D13S317, VWA, D8S1179, D21S11, D18S51, D19S433, D16S539, D7S820, D22S138). A database of allele frequencies for 14 populations (Finland, Poland, Turkey, Vojvodina, Extremadura, Italy, Belgium, Kosovo, Mexico, Taiwan, Korea, El Salvador, Somalia, Mozambique) was collected from studies that have been already published. Probabilities for paternity were computed, in every trio and in every duo (motherless cases,
obtained from trios but not considering the genetic information of the mother) for all combinations of markers and databases. In particular Paternity Index, Probability of Paternity, Typical Paternity Index and Probability of Exclusion were computed and descriptive statistics performed. In order to have a measure of the Database Effect we compared the TPI of different populations with Finnish TPI. TPI is the harmonic mean of PIs and it is considered to be a good measure of typical performance [4]. Nei’s distance was calculated between the Finnish population and all the other 13 populations to investigate if the PI is influenced by the genetic distance between reference population and the putative father’s population.

Population databases

Database of allele frequencies for the 15 STR loci for 14 populations was collected from studies that have been already published in journals. The considered populations are the following: Poland (Central Europe) [7], Turkey (Southwestern Asia) [8], Vojvodina (Autonomous Province in Serbia – Southeastern Europe) [9], Extremadura (Autonomous Community in Spain – Southern Europe) [10], Italy (Southern Europe) [11-13], Belgium (Northwestern Europe) [14], Kosovo (Province in Serbia-Southeastern Europe) [15], Mexico (Northern America) [16], Taiwan (Eastern Asia) [17], Korea (Eastern Asia) [18], El Salvador (Central America) [19], Somalia (Eastern Africa) [20], Mozambique (Southeastern Africa) [21] and Finland (Northern Europe), whose alleles frequencies have been kindly supplied by THL.

Paternity testing and related measures

As the PI and W have been widely explained elsewhere [1-3], we are not going to discuss them in this article. Calculation of the PI has been performed using uniform prior probabilities. The Probability of Exclusion [1-3] was calculated from the following relation:

\[ PE = 1 - RMNE \]

where RMNE denotes the Random Man Not Excluded, that is the proportion, in the reference population, of men that present all the obligate alleles, where the ’obligate allele’ is defined as the one that must come from the biological father, under the hypothesis that M is the mother of the disputed child C and no mutations occurred. For a given marker j, with j = 1, ..., N, the mother M and the child C can have 1 or 2 common alleles. If M and C have 1 common allele, denoted with I, then there is an obligate allele that the biological father must contribute. Then we let

\[ RMNE_i = 1 - (1 - p_j)^2 \]

If M and C have two common alleles i and k, then there are two obligate alleles and the true father could contribute either i or k. In this case we define

\[ RMNE_{jk} = 1 - (1 - p_j - p_k)^2 \]

The overall RMNE is obtained as follows, where it is assumed that the markers are independent:

\[ RMNE = \prod_{j=1}^{N} RMNE_j \]

RMNE is considered to be the power of a paternity test, because it is strictly related to the Probability of Exclusion (PE) that represents the probability of excluding a falsely accused man. The Typical Paternity Index (TPI) is the harmonic mean of the PIs obtained after performing the paternity test, and it is considered to be a good measure of typical performance [4]. TPI has been here calculated for every population P considering all the cases c = 1, ..., 100 as follows:

\[ TPI_c = \frac{100}{\sum_{j=1}^{N} \frac{1}{PI_j}} \]  

In order to have a measure of the ‘Database Effect’ we compared the TPI of different populations with Finnish TPI. Particularly, given a population P and its TPI, denoted with \( TPI_P \), we calculated the following TPI Ratio:

\[ TPI \text{ Ratio} = \frac{TPI_P}{TPI_{\text{Finnish}}} \]

Nei’s distance

Genetic Nei’s distance (1972) was calculated between the Finnish population and all the other 13 populations (see Table 1). Let A and B be two populations and denote with \( j = 1, ..., N \) the markers and with \( i = 1, ..., n_j \) the alleles; Nei’s distance is:

\[ D(A, B) = -\ln \left( \frac{\sum_{i=1}^{n_j} \sum_{j=1}^{N} p^A_{ij} \cdot p^B_{ij}}{\left( \sum_{i=1}^{n_j} \left( p^A_{ij} \right)^2 \right) \left( \sum_{j=1}^{N} \left( p^B_{ij} \right)^2 \right)} \right) \]

where \( p^A_{ij} \) is the frequency of the \( i \)-th allele of the \( j \)-th marker in the population A and \( p^B_{ij} \) is the frequency of the \( i \)-th allele of the \( j \)-th marker in the population B.

Results

100 paternity cases were investigated and in all cases the PI was greater than 10000; moreover no genetic inconsistencies were observed between mother and child or between father and child.

Database effect

‘Database Effect’ has been defined as the effect on PI when the paternity test is performed using a reference population different from the one the putative father is from. Results for standard trio and 15 markers (Table 2) show that there is a ‘Database Effect’ on PI, but not strong enough to change the final conclusion of the paternity test. In fact, the minimum value of PI is 1.4558E+05 (obtained with Kosovo’s allele frequencies) which gives a Probability of Paternity of 0.99999. The maximum value for the same parameter is 6.3693 E+17 and it is obtained with El Salvador. TPI gave us an idea of the typical performance of the PI. The minimum value (2.7222E+06) is related to Finland and it is obtained with El Salvador. TPI gave us an idea of the typical performance of the PI. The minimum value (2.7222E+06) is related to Poland and the maximum (2.8747E+07) is related to Mozambique. It seems that the lowest values of the considered parameters are related to populations genetically close to Finnish one. We can observe a similar behaviour in the TPI Ratios, that we consider as a good measure of the ‘Database Effect’. All the TPI Ratios values, with the exception of the one related to Poland, which is the lowest (0.9281) are bigger than 1; in particular the two highest values are obtained with Mozambique and Finland respectively.

\[ TPI_{\text{Ratio}} = \frac{TPI_P}{TPI_{\text{Finnish}}} \]

| Populations | Distance | Populations | Distance |
|-------------|----------|-------------|----------|
| Finland_Finland | 0 | Finland_Kosovo | 0.0467 |
| Finland_Poland | 0.0223 | Finland_Mexico | 0.0473 |
| Finland_Turkey | 0.0254 | Finland_Taiwan | 0.0917 |
| Finland_Vojvodina | 0.0258 | Finland_Korea | 0.0972 |
| Finland_Extremadura | 0.0295 | Finland_El Salvador | 0.1009 |
| Finland_Italy | 0.0335 | Finland_Somalia | 0.1199 |
| Finland_Belgium | 0.0466 | Finland_Mozambique | 0.1473 |

Table 1: Nei’s distance between Finnish population and all the others considered.
(9.8010) and Somalia (8.9809). These two populations present the two highest Nei’s distance values.

Among motherless cases (Table 3) PIs are generally smaller than in standard trios, but again the final conclusion of the test is not changed. As already seen for trio cases, the lowest values of the parameters are related with Finnish population or with some other population genetically close to the Finnish one. In particular, the minimum PI value (3.2900E+02) is obtained considering Finnish as reference population and the corresponding Probability of Paternity is equal to 0.9970. The maximum value of PI (1.9706E15) is related to El Salvador. TPI takes this minimum value (7.6197E+03) with Finnish allele frequencies and the maximum (7.8919E+04) with Mozambique. TPI Ratios show the presence of a ‘Database Effect’, and results are slightly less variable than what we have observed in standard trio cases. All the ratios are bigger than 1 and the maximum TPI Ratios value (10.3573) is obtained with Mozambique.

Looking at Figure 1 and Figure 2, a positive linear association can intuitively be seen between Nei’s distance and TPI Ratio and moreover the variability of TPI Ratio appears to increase with increasing Nei’s distance. In particular, there is a strong positive correlation between TPI Ratio and Nei’s distance, in both duo and trio cases, using 15 markers. This result is confirmed by the high value of the Multiple Correlation Coefficient R (0.893 for trios and 0.905 for duos). R-square values (0.798 for trios and 0.818 for duos) show that the model is able to explain 79% of the variation in TPI Ratio.

Concerning W and PE, for trio cases and for every choice of reference population, these two probabilities take values greater than 0.9999. On the other hand, for duo cases W and PE can also take values in the interval (0.99-0.9999). Using the 10 markers set all the considered measures’ values are smaller than the values obtained with the 15 markers set. This is due to a reduced amount of information involved in the test. In standard trio cases, we can still observe the presence of a ‘Database Effect’ but, as we have seen for the 15 markers set, the influence of a different reference population does not substantially changing the final result of the paternity test. The minimum PI is related, once again, to Finland (1.0290E+03) and the Probability of Paternity is 0.999. The maximum value for PI is 1.7618E+14 and it is obtained with El Salvador. Using the 10 markers set, as when using the 15 markers set, the lowest values of the parameters are obtained with Finland or using allele frequencies from populations genetically close to Finnish. The minimum value for TPI (2.7645E+04) is related to Finland and the maximum value (1.7555E+05) obtained with Mozambique. ‘Database Effect’ is well represented by TPI Ratios. Once again the maximum value, which is 6.3502, is linked with a population (Mozambique) which is genetically far from the Finnish one. Duo cases’ values, for every parameter, are lower then those observed in trio cases. The minimum value for PI (1.0900E+01) related with Finland, implies a Probability of Paternity of 0.95. The minimum value for TPI (3.2232E+02, Finland) gives a Probability of Paternity of 0.9981, the maximum value is obtained with Somalia (3.1026E+03). The minimum
value for TPI Ratio is 1 and it is related to Finland, the maximum value is 5.8295 and it is obtained using Somalian allele frequencies.

Using the 10 markers set with motherless cases leads to a reduction in the amount of information such that the Probability of Paternity (W) does not reach the commonly used threshold value 0.99.

A good positive linear association between Nei’s distance and TPI Ratio exists both in trio and duo cases. The Multiple Correlation Coefficient R is quite high (0.926 for trios and 0.869 for duos). Moreover the model is able to explain 85% and 75% of the variation in TPI Ratio, for trios and duos respectively. The variability of TPI Ratio appears to increase with increasing Nei’s distance.

Reducing the amount of markers considered creates lower values for W and PE in both trio and duo cases for every choice of reference population. In particular, for trio cases W and PE takes values bigger than 0.99. For duo cases, most of the results are bigger than 0.95, only one case, using Polish frequencies, produced a W value smaller than 0.95.

Considering the third set of markers, the one containing 9 markers, results are quite similar to those obtained with 10 markers. Final results of paternity test are not changed for standard trio cases, but among motherless cases it is possible to have PIs not high enough to produce a Probability of Paternity greater than or equal to 0.99. This is probably due to the joint effect of a small number of considered markers and lack of maternal information. Particularly, in trio cases the minimum observed PI value is 1.6900E+02 corresponding to a Probability of Paternity of 0.994 and it is obtained withVoivodina allele frequencies. The maximum observed PI value is 2.7757E+08 obtained with Korea. The average value of the PI is once again described by TPI, the minimum obtained value for this parameter is 2.7382E+03 (Poland) with a Probability of Paternity of 0.9996, the maximum value is 7.3046E+03 (Somalia). The TPI Ratios show a weak ‘Database Effect’ with a minimum value of 0.9570 (Poland) and a maximum of 2.5530 obtained with Somalia. It is important to underline that even Voivodina and Belgium present a TPI Ratio smaller than 1. Results for duo cases present a minimum value for PI of 3.0000E+00 obtained with two populations, Poland and Voivodina which lead a Probability of Paternity of 0.75. The highest value for PI is 8.1235E+06 and it is related to Korea. Concerning the TPI Ratio, this parameter has its minimum with Poland (0.8804) and its maximum (3.3337) with Mozambique. It seems that using Finnish allele frequencies with the 9 markers set does not grant the best results. Studying the relation between Nei’s distance and TPI Ratio, considering 9 markers, gives some interesting results. It seems that the model fits very well the ‘duo cases-9 markers’ set of data, which is the one with the biggest lack of information. The coefficient R is quite high for trios and duos, 0.854 and 0.928 respectively. R-squared shows that the model is able to explain 72.9% and 86.2% of the variation in TPI Ratio in trios and duos respectively. The variability of TPI Ratio appears to increase with increasing Nei’s distance, especially for trio cases.

Results for W and PE are decreasing, especially in duo cases. In detail, for trio cases and for every choice of reference population, W and PE take values bigger than 0.99. For duo cases, on the other hand, W and PE can take values even smaller than 0.95. Once again, values obtained with motherless cases are smaller than those obtained in trios.

**Marker effect**

The following results have been obtained considering only the Finnish population as reference, in order to investigate the effect on some paternity measures of using different sets of markers. Table 4 shows that trio cases are weakly influenced by the number of considered markers, in fact reducing their amount produces a decreasing in values both for W and PE, but the final conclusion of the paternity test does not change using a different set of markers. Moreover W and PE follow the same distribution in trios. Duo cases present a different situation: results are strongly influenced by the number of considered markers. In particular, values for W and PE are smaller than those presented by trios and for some cases the final conclusion of the test is changed, presenting a value of W smaller than the commonly used threshold 0.99. Results suggest that when dealing with a reduced amount of information, typical of duo cases, it is advisable to use a 15 markers set.

Analyzing the TPI we have an idea of the typical performance of PI. Looking at Table 5 it is easy to notice that, both for trios and duos, TPIs become smaller. As we reduce the number of markers considered; once again duo cases are the most strongly influenced.

**Conclusion**

The results suggest that for trio cases there is no need to use a specific population database, no matter what set of markers is in use, as the Probability of Paternity is always greater than 0.99, 0.999 and 0.9999 for 9, 10 and 15 markers, respectively. The Probability of Paternity and the Probability of Exclusion display similar distribution. Duo cases present a notable database effect and values of Probability of Paternity and Probability of Exclusion being significantly smaller than what we observed in the trio cases. It is recommendable to use a kit with the highest number of markers, in order to achieve reliable results. Analysis of TPI Ratios shows, for every set of markers and for both trio and duo cases, a positive linear relation between Nei’s distances and TPI Ratios. So the more the reference population is genetically far from the Finnish one, the bigger TPI values will be. In particular if the putative father is the true father of the disputed child, then using Finnish allele frequencies will produce a smaller value of PIs than those
obtained using allele frequencies of a population which is genetically far from the Finnish one. It could be interesting to investigate if this behaviour holds in general for every population or whether it is typical of the Finnish population. The ‘Marker Effect’ is the most critical one as it influences both PI’s results and Probability of Paternity, especially in motherless cases, where using the 9 markers set can lead to an exclusion from paternity. It is evident that results obtained with a reduced amount of information (reduced number of markers) are smaller than those obtained with the 15 markers set. It would be fruitful to perform the same experiment using a different set of markers, composed by those markers that present a high level of entropy.

References

1. Evett IW, Weir BS (1998) Interpreting DNA evidence. Statistical Genetics for Forensic Scientists. Sinauer.
2. Brenner CH (1997) Assumption in the “Probability of Paternity”, http://www.dna-view.com/assump.htm.
3. Scarpetta MA, Staub RW, Einum DD (2007) Assessing exclusionary power of a paternity test involving a pair of alleged grandparents. Transfusion 47: 335-340.
4. Brenner C, Morris JW (1989) Paternity Index calculations in single locus hypervariable DNA probes: validation and other studies. Proceedings for the International Symposium on Human Identification. Promega Corporation.
5. Nei M (1972) Genetic distance between populations. The American Naturalist 106: 283-292.
6. Dawid AP, Mortera J, Pascali VL (2001) Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. Forensic Science International 124: 55-61.
7. Jacewicz R, Berent J, Prośniak A, Galecki P, Florkowski A, et al. (2004) Population genetics of the Identifier system in Poland. International Congress Series 1261: 229-232.
8. Yavuz I, Sarikaya AY (2005) Turkish Population Data for 15 STR Loci by Multiplex PCR. J Forensic Sci 50: 737-738.
9. Veselinović I, Kubat M, Furac I, Skavić J, Martinović Klarić I, et al. (2004) Allele frequencies of the 15 AmpFISTR Identifier loci in the population of Vojvodina Province, Serbia and Montenegro. Int J Legal Med 118: 184-186.
10. Garcia-Hirschfel J, Farfan MJ, Prieto V, Lopez-Soto M, Torres Y, et al. (2003) Allele distribution of 15 STRs in a population from Extremadura (Central-Western Spain). International Congress Series 1239: 165-169.
11. Garofano L, Pizzamiglio M, Vecchio C, Lago G, Floris T, et al. (1998) Italian population data on thirteen short tandem repeat loci: HUMTH01, D21S11, D18S51, HUMVWA31, HUMFIBRA, D8S1179, HUMTPOX, HUMCSF1PO, D16S539, D7S820, D13S317, D5S818, D3S1358. Forensic Sci Int 97: 53-60.
12. Garofano L, Pizzamiglio M, Bizzaro GP, Donato F, Rossetti M, et al. (1999) Italian population data on two new short tandem repeat loci: D6S477 and D19S433. Forensic Sci Int 101: 203-208.
13. Garofano L, Pizzamiglio M, Donato F, Biondi F, Rossetti M, et al. (1999) Italian population data on two new short tandem repeat loci: D2S1338 and Penta E. Forensic Sci Int 105: 131-136.
14. Decorte R, Engelen M, Larno L, Nelissen K, Glissien A, et al. (2004) Belgian population data for 15 STR loci (AmpFISTR SGM Plus and AmpFISTR profiler PCR amplification kit). Forensic Sci Int 139: 211-213.
15. Kubat M, Skavić J, Behluli I, Nuraš B, Bekteshi T, et al. (2004) Population genetics of the 14 AmpFISTR Identifier loci in Kosovo Albanians. Int J Legal Med 118: 115-118.
16. Hernández-Gutiérrez S, Hernández-Franco P, Martínez-Tripp S, Ramos-Kuri M, Rangel-Villalobos H (2005) STR data for 15 loci in a population sample from the central region of Mexico. Forensic Sci Int 151: 97-100.
17. Wang CW, Chen DP, Chen CY, Lu SC, Sun CF (2003) STR data for the AmpFISTR SGM Plus and Profiler loci from Taiwan. Forensic Sci Int 138: 119-122.
18. Kim YL, Hwang JY, Kim YJ, Lee S, Chung NG, et al. (2003) Allele frequencies of 15 STR loci using AmpFISTR Identifier kit in a Korean population. Forensic Sci Int 136: 92-95.
19. Morales JA, Monterrosa JC, Puente J (2004) Population genetic data from El Salvador (Central America) using AmpFISTR Identifier PCR Amplification Kit. International Congress Series 1261: 223-225.
20. Alves C, Gusmão L, Damasceno A, Soares B, Amorim A, (2004) Contribution for an African autosomic STR database (AmpFISTR Identifier and Powerplex 16 system) and a report on genotypic variations, Forensic Science International 139: 201-205.