

Allele sharing in first-degree and unrelated pairs of individuals in the Ge.F.I. AmpFISTR[®] Profiler Plus[™] database

Silvano Presciuttini^{a,1,*}, Francesca Ciampini^a, Milena Alù^b, Nicoletta Cerri^c, Marina Dobosz^d, Ranieri Domenici^e, Gabriella Peloso^f, Susi Pelotti^g, Andrea Piccinini^h, Elena Ponzanoⁱ, Ugo Ricci^j, Adriano Tagliabracci^k, J.E. Baley-Wilson^l, Francesco De Stefano^m, Vincenzo Pascali^{d,1}

^aDepartment of Biomedicine, University of Pisa, C/o Centro Retrovirus, S.S. Abetone e Brennero 2, 56127 Pisa, Italy

^bDepartment of Morphological and Forensic Sciences, University of Modena and Reggio Emilia, Modena, Italy

^cChair of Legal Medicine, University of Brescia, Brescia, Italy

^dInstitute Of Legal Medicine, Catholic University, Rome, Italy

^eChair of Legal Medicine, University of Pisa, Pisa, Italy

^fDepartment of Legal Medicine and Public Health, University of Pavia, Pavia, Italy

^gDepartment of Hygiene and Public Health, University of Bologna, Bologna, Italy

^hInstitute of Legal Medicine, University of Milan, Milan, Italy

ⁱDepartment of Medicine and Public Health, Section of Legal Medicine, University of Padua, Padua, Italy

^jCentre of Medical and Molecular Genetics, Hospital "A. Meyer", Florence, Italy

^kInstitute of Legal Medicine, University of Ancona, Ancona, Italy

^lInherited Disease Research Branch, NHGRI, Baltimore, MD, USA

^mInstitute of Legal Medicine, University of Cagliari, Cagliari, Italy

Received 22 July 2002; received in revised form 30 September 2002; accepted 15 October 2002

Abstract

Eleven Italian forensic laboratories participated in a population study based on the AB Profiler Plus[™] loci with proficiency testing. The validated database, including 1340 individuals, is available on-line. Tests for Hardy–Weinberg equilibrium, gametic unbalance, and heterogeneity of gene frequency were generally not significant. Gene frequencies at each locus were consistent with those of two previously published Italian studies, but different from a third. Individuals of each subsample were paired, and the total number of alleles shared across the nine loci was determined in each pair. The analysis was replicated over the total sample. In addition, two samples of mother–child pairs ($N = 315$) and full-sib pairs ($N = 91$) were subjected to allele sharing analysis. The resulting distributions were sufficiently distinct from the sample of unrelated pairs as to be of practical usefulness.

© 2002 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Allele sharing; Profiler Plus[™]; Validation study; Population structure

1. Introduction

In several countries, databases of DNA markers to be used for criminal casework and paternity testing have been vali-

dated through proficiency testing exercises, and are available on-line [1–4]. Most collaborative studies have been dedicated to commercially available products, such as—for example—the AmpFISTR[®] Profiler Plus[™] kit (Applied Biosystem, Foster City, CA, USA) [5]; this includes the amelogenin locus plus the following nine markers: D3S1358, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820. The same markers are included in the loci selected for the US Combined DNA Index System (CODIS) [6], and

* Corresponding author. Tel.: +39-50-555234;

fax: +39-50-560276.

E-mail address: sprex@biomed.unipi.it (S. Presciuttini).

¹ President of the Ge.F.I.

have thus become powerful tools for forensic studies and population genetics research (see e.g. [7,8]).

In this paper, we briefly illustrate the Ge.F.I. (Italian Forensic Geneticists) Profiler Plus™ collaborative study—a database-proficiency test project that involved 11 forensic laboratories, which typed 1340 individuals from 9 regions of Italy. This database is fully available on-line. As a part of the validation procedure, we determined the number of alleles shared by all pairs of individuals [9,10], both in each sub-sample and in the total database. The same analysis was replicated on 315 mother–child pairs and 91 full-sib pairs. Monte Carlo simulations of nuclear families were carried out to corroborate the results from the real data. Our purpose was to test the power of allele sharing as a method to discriminate between first-degree and unrelated pairs of individuals.

2. Materials and methods

Participating laboratories had to type at least 100 unrelated individuals from their own region of Italy, plus three bloodstains centrally provided by a coordinating center. Results from 11 forensic labs from Northern and Central Italy were eventually entered in the Ge.F.I. database. All laboratories used the AB technology and kits, except one [11].

All individuals of the population samples were paired to all others, and the number of alleles (0, 1, or 2) shared identical by state by pairs of individuals was determined for each marker and summed over all markers. The analysis was repeated in the total sample. Allele sharing in mother–child and sib pairs was also determined. Computations were performed in Microsoft Excel 2000. The used applications are easy to use and are available from the author (mail to S.P.).

In population genetics analysis, samples were first considered separately, testing for Hardy–Weinberg equilibrium (HWE) and heterogeneity of allele frequency. Then, samples from the same region were pooled, and the analyses were repeated by region. Finally, the total sample was tested for HWE and for heterogeneity of allele frequencies against all other Italian studies known to us. Tests were carried out using Arlequin 2000 [12].

After completion of the exercise, the 11 laboratories were asked to supply all mother–child pairs and undisputed full-sib pairs data. In parent–child and sib-pair simulations, genotypes of unrelated parental pairs were simulated for each locus, using the allele frequencies estimated from our database, and four random numbers. From these genotypes, one or two children were simulated using the Mendelian transmission rules and two or four additional random numbers, respectively.

3. Results

3.1. Population genetics analyses

Hardy–Weinberg equilibrium (HWE) and deviation from expected heterozygosity were tested first in each laboratory

(generating 99 P -values). A slight excess of P -values lower than 0.05 was observed in comparison with those expected by chance, though none was significant after applying the Bonferroni correction. The homozygosity test was also not significant after correction for the multiple comparisons, though a single lab was deviant at the nominal 0.05 significance level for one locus and at the 0.01 level for another two loci (this result was already reported using a different calculation method [13]). Then, genotypes were pooled by region, and nine sub-samples were generated. None of the observed 81 P -values exceeded the nominal (uncorrected) 0.01 significance level, for both the HWE test and the homozygosity test. Finally, the total sample of 1,340 individuals was examined, and none of the nine P -values were lower than the nominal 0.05 level, for both the HWE test and the homozygosity test.

The exact test of population differentiation was applied first to the 11 labs, and then to the nine regions (in this test, each sample is contrasted to all others). Only two loci (D13S317 and D18S51) showed a significant level of heterogeneity, both considering single laboratory data and the data aggregated by region. These two markers showed also significant levels of linkage disequilibrium (i.e. gametic unbalance at unlinked loci) with D21S11 and O21S11 ($P = 0.047$ and $P = 0.020$, respectively); markers D8S1179 and D7S820 showed also significant linkage disequilibrium ($P = 0.009$).

Three other large studies have been published in Italian populations on the Profiler Plus™ loci [14–16] (one involving six markers only [14]). Heterogeneity of allele frequencies was tested by the exact method among all four studies, the ‘Ge.F.I.’ (this study), the ‘Police’ [14] ($N = 618$ individuals), the ‘SIMEF’ [15] ($N = 2000$), and ‘SCIS’ [16] ($N = 223$). Ge.F.I., Police, and SCIS samples were homogeneous, with minor exceptions. On the contrary, all tests comparing Ge.F.I.–SIMEF, all tests of Police–SIMEF, and five out of nine tests of SCIS–SIMEF were highly significant for heterogeneity (below 0.0001), revealing that the SIMEF data are heterogeneous with respect to all others.

3.2. Allele sharing over all loci

Unrelated individuals were paired to each other separately for the 11 laboratories, and the number of shared alleles across the nine loci was determined in each pair. This allowed us to correct some minor local database errors. Then, all individuals of the total sample were paired to each other, generating 897,130 pairs. The two samples of first-degree relatives included 315 mother–child pairs and 91 full-sib pairs. We also simulated 100,000 pairs for both these relationships, using the allele frequencies estimated from the total sample of unrelated individuals. Fig. 1 shows the distributions of the number of pairs with the specified number of shared alleles in the following groups: (1) the 897,130 pairs of unrelated individuals; (2) the 315 true mother–child pairs; (3) the 100,000 simulated parent–child pairs; (4) the 91 true sib pairs; (5) the 100,000 simulated full-sib pairs. The chi square tests

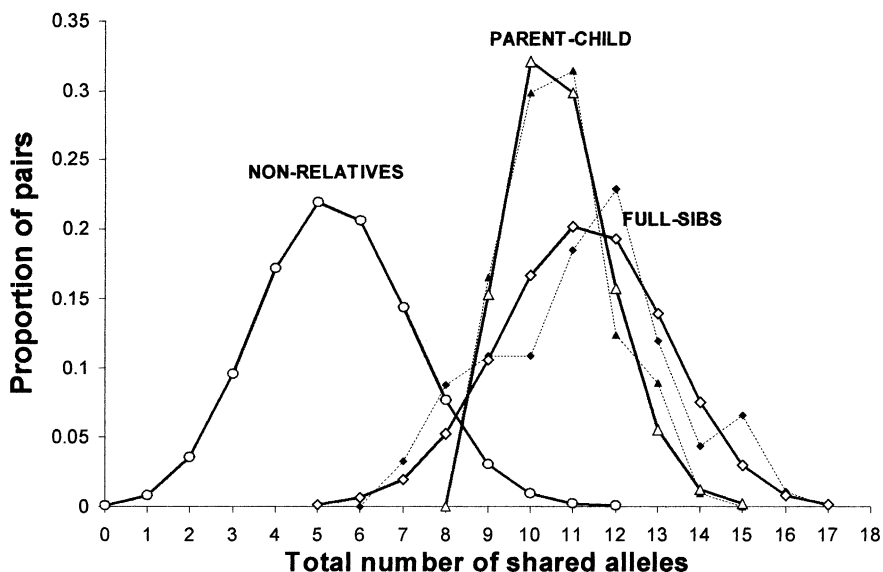


Fig. 1. Distribution of the total number of shared alleles at the Profiler Plus™ loci in different groups of individuals. Broken lines: true mother–child pairs and full-sib pairs.

Table 1

Probabilities that pairs of individuals with the given relationships share this number of alleles at the Profiler Plus™ loci, and likelihood ratios that they are first-degree relatives rather than unrelated, given the number of shared alleles

Number of shared alleles	Probabilities			Likelihood ratios	
	Non-relative	Parent–child	Full-sibs	PC/NR	FS/NR
0	8.0×10^{-4}	0	–	0	0
1	8.0×10^{-3}	0	–	0	0
2	0.036	0	–	0	0
3	0.096	0	4.0×10^{-5}	0	4.2×10^{-4}
4	0.172	0	2.1×10^{-4}	0	1.2×10^{-3}
5	0.219	0	1.5×10^{-3}	0	6.7×10^{-3}
6	0.206	0	6.5×10^{-3}	0	0.0
7	0.143	0	0.020	0	0.1
8	0.076	0	0.053	0	0.7
9	0.031	0.153	0.105	5.0	3.4
10	9.6×10^{-3}	0.321	0.166	33.4	17.3
11	2.2×10^{-3}	0.298	0.202	133.7	90.4
12	3.9×10^{-4}	0.157	0.192	405.9	496.0
13	5.1×10^{-5}	0.055	0.139	1093.3	2737.9
14	3.4×10^{-6}	0.013	0.075	3749.7	22226.2
15	1.1×10^{-6}	1.9×10^{-3}	0.030	1666.5	26442.7
16	–	1.8×10^{-4}	8.2×10^{-3}	–	–
17	–	1.0×10^{-5}	1.3×10^{-3}	–	–
18	–	–	8.0×10^{-5}	–	–

were not significant for both the comparisons of true/simulated parent–child and true/simulated full-sib distributions. Table 1 shows the probabilities that unrelated individuals, parent–child pairs and full-sib pairs share this number of alleles, respectively, and the likelihood ratio that a given pair of individuals is a parent–child pair or a full-sib pair, given the total number of alleles they share at the Profiler Plus™ loci.

4. Discussion

This study was initially designed to establish a nationwide database controlled by a proficiency test and available online. The current database now includes 1340 unrelated individuals from nine Italian regions of Northern, Central and Southern Italy, and may be accessed at the Ge.F.I. website

(www.gefi-forensicDNA.it). Hardy–Weinberg equilibrium and homozygosity tests were non-significant for deviation after correction for the multiple comparisons. Two loci (D13S317 and D18S51) showed significant heterogeneity of allele frequency among Italian regions; both loci were also involved in two out of three tests (among a total of 72 pairwise comparisons) that resulted in significance for gametic unbalance between loci. This suggests the presence of a genetic structure across Italy; whereas this result would be worthy of a closer investigation in population genetics studies, the strength of the whole phenomenon is not so marked as to prevent using these loci the forensic practice in Italy. Comparison of allele frequencies among different Italian studies showed consistency of the present work and two published results [14,15], whereas a fourth compilation [16] was highly significant for heterogeneity with all others. This finding strongly emphasizes the need of coordinating all studies aimed at generating databases, and shows the positive effect of quality control tests. As a part of the validation procedure, the distribution of the number of matches between all pairs of individuals at the nine loci was determined (interestingly, we identified by this approach five subjects, including two pairs of spouses, who were typed independently in two different labs). The results of this analysis prompted us to investigate the same issue in first-degree relatives. We collected a total of 315 mother–child pairs and 91 full-sib pairs from the 11 participating groups. As the total number of pairs was too small to obtain accurate empirical distributions of the number of shared alleles, we performed computer simulations of these relationships. The distributions observed in 100,000 replicates of both parent–child and full-sibs pairs were not significantly different from those observed in the true pairs. This allowed us to use the simulated pairs to compute the likelihood ratio of a particular relationship (parent–child versus non-relatives and full-sibs versus non-relatives) between any two individuals, given the total number of alleles they share at the Profiler Plus™ loci. This approach may be useful to obtain a first insight about the probability that two specific samples typed at these loci are first-degree relatives rather than non-relatives. For example, if two subjects share 12 total alleles, the likelihood ratio that they are siblings rather than non-relatives is 496 (Table 1), meaning that it is about 500 times more likely that they are siblings rather than unrelated; the likelihood ratio that they are a parent–child pair is about 400 in this case (of course, with the additional condition that none of the markers showed incompatibility, i.e. zero matches; for the distribution of the number of loci excluding paternity using 13 markers see [17]).

The easy discrimination between first-degree relatives and unrelated individuals obtained by the allele sharing test comes out clearly from consultation of Table 1. If considering 9 shared alleles, it appears that all parent–child pairs and 92% of full-sib pairs are correctly identified at a significance level <0.05 (these values are obtained by summing together the probabilities of the three relationships for numbers of shared alleles ≥ 9 ; in the present example the exact values are 0.043,

1.000, and 0.919, respectively). With 10 shared alleles, the significance level (equivalent to the ratio of false positives) is 0.012, and the proportion of first-degree relatives correctly identified is 85% for parent–child pairs and 81% for full-sib pairs. With 11 shared alleles, the significance level becomes 0.003, and percentages of true positives are 53% of parent–child pairs and 65% of full-sib pairs are correctly identified. Thus, more than 50% of true first-degree relatives are correctly identified with probability higher than 99.5%.

These results strictly apply to the collection of loci included in the Profiler Plus™ kit. However, the number of shared alleles between two individuals at a locus (0, 1, or 2) depends on locus heterozygosity rather than on the particular distribution of allele frequencies (Presciuttini et al., submitted). This means that any locus with heterozygosity equal to that of a locus included in the kit may be substituted for it without changing the results. Therefore, the above results apply to any collection of nine loci, provided that the distribution of heterozygosity among them is equivalent to that of the Profiler Plus™ loci. The statistics of allele sharing is therefore a quick and versatile way to approach the problem of relatedness between pairs of individuals in the forensic context.

Acknowledgements

The following persons contributed to the present work: P. Grignani and C. Previderè (Dipartimento di Medicina Legale e Sanità Pubblica, Università di Pavia); G. Beduschi (Dipartimento di Scienze Morfologiche e Medico Legali, Università di Modena e Reggio Emilia); I. Sani (Ospedale Anna Meyer, Firenze); L. Buscemi (Istituto di Medicina Legale, Università di Ancona); S. Prandoni (Istituto di Medicina Legale, Università di Milano); L. Casarino and S. Verdiani (Dipartimento di Medicina Legale, del Lavoro, Psicologia Medica e Criminologia, Università di Genova); L. Caenazzo (Dipartimento di Medicina Ambientale e Sanità Pubblica, Università di Padova); I. Spinetti and C. Toni (Cattedra di Medicina Legale, Università di Pisa).

References

- [1] C.M. Ruitberg, D.J. Reeder, J.M. Butler, STRBase: a short tandem repeat DNA database for the human identity testing community, *Nucleic Acids Res.* 29 (2001) 320–322.
- [2] W. Huckenbeck, K. Kuntze, H.G. Scheil, *The Distribution of the Human DNA–PCR Polymorphisms—a Worldwide Database*, vol. II, Verlag, Berlin, 2002.
- [3] J. Gomez, A. Carracedo, The 1998–1999 collaborative exercises and proficiency testing program on DNA typing of the Spanish and Portuguese Working Group of the International Society for Forensic Genetics (GEP-ISFG), *Forensic Sci. Int.* 114 (2000) 21–30.
- [4] L. Roewer, M. Krawczak, S. Willuweit, M. Nagy, C. Alves, A. Amorim et al., Online reference database of European

- Y-chromosomal short tandem repeat (STR) haplotypes, *Forensic Sci. Int.* 118 (2001) 106–113.
- [5] Baylor College of Medicine, inventors. PE Biosystems Applied Biosystems Corporation, assignee. AmpFISTR[®] Profiler Plus[™] PCR Amplification Kit. US Patent 5,364, 759.
- [6] B. Budowle, B. Shea, S. Niezgodna, R. Chakraborty, CODIS STR loci data from 41 sample populations, *J. Forensic Sci.* 46 (2001) 453–489.
- [7] H. Geadia, R.M. Brito, T. Ribeiro, R. Espinheira, Portuguese population and paternity investigation studies with a multiplex PCR-the AmpFISTR[®] Profiler Plus[™], *Forensic Sci. Int.* 108 (2000) 31–37.
- [8] B. Bramanti, S. Hummel, T. Schultes, B. Herrmann, STR allelic frequencies in a German skeleton collection, *Anthropol. Anz.* 58 (2000) 45–49.
- [9] R. Chakraborty, L. Jin, Determination of relatedness between individuals using DNA fingerprinting, *Hum. Biol.* 65 (1993) 875–895.
- [10] S. Presciuttini, L. Casarino, S. Verdiani, A. Mannucci, L. Peirano, J.E. Baileys-Wilson, F. De Stefano, Distribution of the number of alleles identical by state (IBS) summed over several loci among full-sib pairs and unrelated individuals, *Prog. Forensic Genet.* 8 (2000) 88–90.
- [11] U. Ricci, I. Sani, S. Guarducci, C. Biondi, S. Pelagatti, V. Lazzarini, A. Brusaferrri, M. Lapini, E. Andreucci, L. Giunti, M.L. Giovannucci Uzielli, Infrared fluorescent automated detection of thirteen short tandem repeat polymorphisms and one gender-determining system of the CODIS core system, *Electrophoresis* 21 (2000) 3564–3570.
- [12] S. Schneider, D. Roessli, L. Excoffier, Arlequin (version 2000): a software for population genetics data analysis, Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.
- [13] U. Ricci, I. Sani, L. Giunti, S. Guarducci, S. Coviello, M.L. Giovannucci Uzielli, Analysis of 13 tetrameric short tandem repeat loci in a population of Tuscany (central Italy) performed by means of an automated infrared sequencer, *Forensic Sci. Int.* 125 (2002) 83–85.
- [14] R. Biondo, A. Spinella, P. Montagna, P.S. Walsh, C. Holt, B. Budowle, Regional Italian allele frequencies at nine short tandem repeat loci, *Forensic Sci. Int.* 115 (2001) 95–98.
- [15] L. Garofano, M. Pizzamiglio, C. Vecchio, G. Lago, T. Floris, G. D'Errico, G. Brembilla, G. Romano, B. Budowle, Italian population data on 13 short tandem repeat loci: HUMTH01, D21S11, D18S51, HUMVWFA31, HUMFIBRA, D8S1179, HUMTPOX, HUMCSFIPO, D16S539, D7S820, D13S317, D5S818, D3S1358, *Forensic Sci. Int.* 97 (1998) 53–60.
- [16] A. Barbaro, P. Cormaci, G. Falcone, A. Barbaro, Population genetic study of 9 STR loci using AmpFISTR[®] Profiler Plus[™] kit and capillary electrophoresis: distribution of allelic frequencies in 20 Italian towns, *Prog. Forensic Genet.* 8 (2000) 184–189.
- [17] F. Calafell, The probability distribution of the number of loci indicating exclusion in a core set of STR markers, *Int. J. Legal Med.* 114 (2000) 61–65.