Available online at www.sciencedirect.com

SCIENCE DIRECT®

Forensic Science International 150 (2005) 17–21

Forensic Science International

www.elsevier.com/locate/forsciint

Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs

António Amorim[a,b,*], Luísa Pereira[a]

[a]IPATIMUP—Instituto de Patologia e Imunologia Molecular da Universidade do Porto,
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
[b]Faculdade de Ciências da Universidade do Porto, Praça Gomes Teixeira, 4099-002 Porto, Portugal

Received 25 October 2003; received in revised form 10 June 2004; accepted 16 June 2004
Available online 9 September 2004

Abstract

Recent advances in single nucleotide polymorphisms (SNPs) research have raised the possibility that these markers could replace the forensically established short tandem repeats (STRs). In this work, we compare STRs and SNPs applicability for kinship investigation in terms of expected informative content and probability of occurrence of ‘‘difficult cases’’ (when isolated Mendelian incompatibilities between alleged father and child are found). Since SNPs have a much lower mutation rate than STRs, these difficulties were expected to occur less frequently if SNPs were used instead of STRs. The purpose of this paper is to make some simulations allowing the estimation of how often such difficult cases are expected to occur using both types of markers and how serious can be their impact in routine work. Our results demonstrate that a battery based exclusively on SNPs matching the informative power of current STR kits would be prone, if applied to routine paternity investigation, to the occurrence of cases where the statistical evidence would be inconclusive. We infer that the introduction of a SNP based strategy, as a substitute to the now classical STR approach poses statistical problems that must be carefully evaluated.

Keywords: STR; SNP; Kinship; Mutation; Null allele

1. Introduction

Forensic genetics has been currently using PCR based Short Tandem Repeat (STR) polymorphisms, also known as microsatellites, for some years with enormous success. These markers have been widely accepted and validated and have almost displaced previously used markers. Furthermore, they encouraged considerable commercial interest and research that led to nowadays-standard kits. Last, but not least, extensive efforts have produced huge databases, both for forensic and general population genetics purposes, such as, for instance, the criminal databases in the United Kingdom (Forensic Science Service, http://www.forensic.gov.uk/), and the United States (the CODIS system at the Federal Bureau of Investigation, http://www.fbi.gov/hq/lab/codis/index1.htm), or the Y-STR haplotype reference database (http://www.ystr.org/).

Recent advances in Single Nucleotide Polymorphisms (SNPs) research have however raised the possibility that this kind of marker could replace the forensically established STRs. Their relative advantages and drawbacks have been already discussed in technical terms and, namely, automation, information content, and suitability to limit samples. A very important issue contrasting the two types of markers

Abbreviations: STR, Short tandem repeat; SNP, Single nucleotide polymorphism.

[*] Corresponding author. Tel.: +351 22 5570700; fax: +351 22 5570799.

E-mail address: aamorim@ipatimup.pt (A. Amorim).

0379-0738/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2004.06.018

resides in the respective mutation rates. In fact, while in STRs they have shown to be relatively frequent, with estimates in the order of magnitude of 10^{-3} [1], for SNPs they have been considered as negligible (10^{-8}) in practical terms [2,3].

In this work, we intend to comparatively evaluate the STR and SPN use for kinship investigation in terms of expected informative content and probability of occurrence of ‘‘difficult cases’’. Common use of STRs has shown that, in a relatively high number of true paternity cases, a Mendelian incompatibility in a single locus occurs. These cases, however, are normally not considered ‘‘difficult cases’’, since the positive evidence obtained in the other loci outweighs satisfactorily the negative one provided by the isolated ‘‘exclusions’’. Since practical experience on paternity cases using exclusively SNPs is still lacking, the purpose of this paper is to make comparative simulation studies allowing to estimate how often such difficult cases are expected to occur using both types of markers and how serious can be their impact in routine paternity investigation.

2. Definitions and algorithms

All loci here considered are autosomal, and showing no gametic associations for any pair of freely recombining non-alleles. The simulated forensic situation is the most frequently encountered: a trio comprising the indubitable mother, her child and a putative father. In this work, we will assume to deal only with biologically true trios.

For simplification, we will consider a ‘‘typical’’ STR locus as containing, in the population under analysis eight equally frequent alleles ($p_1 = p_2, \ldots, = p_8 = 0.125$) and an ‘‘ideal’’ SNP as one with two, also equally frequent, alleles: $p_1 = p_2 = 0.5$, conditions that mimic the average informative power of the loci included in standard kits.

Since exclusion power is a biased summary statistic in this context [5,6] we will use heterozygosity as a measure of the informative content of a locus ($h = 1\text{-}\Sigma p_i^2$; where p_i is the frequency of the ith allele). For the combination of a set of markers, cumulative heterozygosity will be employed, i.e., $1\text{-}\Pi\ (1-h_l)$, over the l loci under analysis, where h_l is the heterozygosity of each locus. This measure will then mean the probability, for a random individual, of not being homozygous over all loci considered.

In the tradition of Landsteiner’s rules, we will define two different kinds of ‘‘exclusion’’ situations in a paternity case where the mother is also typed: those in which the incompatibility can be alleviated by the putative presence of a silent gene (apparent opposite homozygosity between alleged father and the child) and those in which the child exhibits a gene (or gene product) absent from both mother and putative father.

In the first case, (Landsteiner’s second rule), the paternity index in a locus where such observation is registered amounts to $s/[(q+s)(p+2s)]$, where p and q stand for the frequencies of the codominant alleles and s to the silent gene. The derivation of the formula can be explained as shown in the following table, where B stands for any codominant allele distinct from A.

Father: A	**Mother: AB**	**Child: B**
X	***Y***	***X/Y***
$2ps$ 1/2 $2pq$ 1/2	$(2ps + p^2)$ $2pq$ 1/2 $(q + s)$	$s/[(q + s)(p + 2s)]$
Father: A	**Mother: B**	**Child: B**
X	***Y***	***X/Y***
$2ps$ 1/2 q^2	$(2ps + p^2)$ $q^2(q + s)$	$s/[(q + s)(p + 2s)]$

Silent genes have been found at PCR based loci and their corresponding reported frequencies (although good estimates require large sample sizes) are in the range of 0.005–0.001 [7–9]. Furthermore, the (negative) evidential value of those exclusions, as measured by a paternity index, depends on the frequency of the codominant alleles. In the case of binary system with equally frequent alleles, it amounts to 2×10^{-2} while for the ‘‘typical’’ STR locus defined above it takes the value of 3×10^{-1}.

The probability of observing such trios, will be given by

$$\mathrm{Pn} = \sum_{i=1}^{n} p_i s(1 - p_i - s)$$

where n is the number of alleles at that locus, p_i the frequency of the ith allele and s the frequency of the silent gene. Thus the probability of observing at least one of these exclusions in a set of l identical loci can be given by $1-(1-\mathrm{Pn})^l$ and the probability of observing at least two of these exclusions in a set of l identical loci will be $1-[(1-\mathrm{Pn})^l + l\ \mathrm{Pn}\ (1-\mathrm{Pn})^{l-1}]$.

As for the case of Landsteiner’s first rule, we will distinguish the cases for STRs and SNPs as follows. Single-step mutation rate in STRs (m_s) will be defined as the frequency of +/− one-step transitions per gametogenesis. Average values in the order of magnitude of 10^{-3} (for both increase and decrease of repeat number) have been reported [1].

The paternity index in a locus where the observation of a paternal single step mutation is registered depends on the mutation rate and the frequency of the allele, which, in a true trio, has arisen by mutation:

Father	Mother	Child	*X*	*Y*
12–18	18	18–19	$2pq$ 1/2 m_s q^2	$2pq$ q^2 r

In this case it amounts to $m_s/2r$, which means that for common alleles ($r \geq 10^{-1}$) is in the order of magnitude of 10^{-2} while for rare ones ($r \leq 10^{-2}$) it can be higher than 10^{-1}.

Mutation rate in SNPs will be defined as the frequency of the apparent change in allelic state of the binary marker per gametogenesis. In contrast with the case of STR, no direct experimental data are available and only phylogenetic

Table 1
Probability of occurrence of father/child ‘‘exclusions’’ by Landsteiner’s second rule among true trios using batteries of SNP or STR loci

Number of loci	Probability of occurrence							
	One ‘‘exclusion’’				Two ‘‘exclusions’’			
	s = 0.001		s = 0.005		s = 0.001		s = 0.005	
	SNPs	STRS	SNPs	STRS	SNPs	STRS	SNPs	STRS
15	0.0075	0.0130	0.0365	0.0630	0.0000	0.0000	0.0006	0.0019
20	0.0099	0.0173	0.0484	0.0832	0.0000	0.0001	0.0011	0.0034
40	0.0197	0.0343	0.0944	0.1594	0.0002	0.0006	0.0045	0.0131
45	0.0222	0.0386	0.1055	0.1774	0.0002	0.0007	0.0057	0.0164
50	0.0246	0.0427	0.1165	0.1951	0.0003	0.0009	0.0069	0.0200

The expected frequencies of single and double (at two loci) paternal incompatibilities were made assuming null allele frequencies (s) of 0.005 and 0.001.

estimates [2,10,11] have been reported (10^{-8}). On the other hand, since many of the SNP typings are just binary, most of the practical cases will be formally indistinguishable from opposite homozygosities.

Then, if the probability of observing a mutation at a specific locus in a gametogenesis is m, the probability of observing no mutations in a biologically true trio (both in the paternal and the maternal line) is Pm = $(1-m)^2$. Consequently, the probability of observing no mutations over l loci will be given by $(1-\mathrm{Pm})^l$, the probability of observing at least one by $1-(1-\mathrm{Pm})^l$ and of observing at least two $1-[(1\text{-}\mathrm{Pm})^l + l\ \mathrm{Pm}\ (1-\mathrm{Pm})^{l-1}]$. If we assume, for simplification purposes, an identical paternal and maternal mutation rate (which is probably an underestimate) half of these mutations would be of parental origin. For the same reason (which now leads to an overestimation) we will consider that in all cases the parental origin of the mutation can be discerned.

3. Results

3.1. Comparison of informative content

We have compared the informative power of STR and SNP simulated batteries of loci. For STRs we used as reference the currently used STR based commercial kits, such as AmpFLSTR® Identifiler™ or Powerplex™ 16, with 15 and 16 autosomal loci respectively, which individual heterozygosities are, in average, around 80%. For SNPs we considered sets of ‘‘optimal’’ SNPs (with 50% frequency for each allele). Even so, the number of SNP loci needed to match the informative power provided by the STR kits is over 40 (in our simulations it amounts to 44).

3.2. Probability of occurrence of ‘‘difficult paternity cases’’

Beginning with the cases of father/child ‘‘exclusion’’ by Landsteiner’s second rule among true trios, we have studied the expected frequencies of single and double (at two loci) paternal incompatibilities, assuming frequencies for the null alleles of 0.005 and 0.001. The results for the most relevant cases summarised in Table 1. The important facts to retain are: (a) isolated ‘‘exclusions’’ are expected to occur in a SNP battery equally powerful to the current STR ones in a frequency between 2.0 and 2.5% (for s = 0.001; or between 9.4 and 11.7% for s = 0.005) while for the equivalent STR set the corresponding figures are 1.3–1.7% or 6.3–8.2%, respectively; and (b) the frequency of ‘‘exclusions’’ at two loci in the SNP battery is estimated between 2–3/10,000 (for s = 0.001 or 45–69 for s = 0.005) while in a corresponding STR set the values are 0–1 or 19–34.

We have also computed for STRs the expected frequency of paternal incompatibilities due to (a) isolated ‘‘exclusions’’ or to (b) two ‘‘exclusions’’ due to single-step mutations. Assuming single step paternal mutation rates of 10^{-3} and a battery of 15 loci, those cases are expected to occur once in, approximately, each 70 and 9600 true trios, respectively.

4. Discussion and conclusions

Concerning the informative power of both types of markers, SNPs and STRs, our results confirm in general those already advanced [4], namely, that to match the currently available STR kits, about 50 SNPs would be required.

We then tried to evaluate their behaviour concerning the expected frequency of ‘‘difficult’’ paternity cases when using batteries of STR or SNP loci of identical informative power. Indeed, the previous comparison in theoretical informative power ignores both mutation and null alleles. As most of the workers in the field are well aware, disagreeably often we are faced with cases where, along with very strong evidence in favour of paternity, an apparent opposite homozygosity between putative father and child shows up or a mutation seems to have occurred.

When analysing both the expected frequency of those ‘‘difficult’’ cases among true trios and weighting their evidential value, some unexpected differential behaviour of the two types of markers have been disclosed.

First of all, considering ‘‘exclusions’’ by Landsteiner's second rule, and assuming (we have no reason whatsoever to think differently if both are PCR based) that null alleles are equally frequent in SNP or STR systems, we have found that (a) isolated ‘‘exclusions’’ in a SNP based approach are expected to occur twice as often (2–12%) as in a equally powerful STR set, and (b) two exclusions (depending much more on the frequency of the silent genes) two to three times more frequently (2–70/1000).

However, while for SNPs, as justified above, we have not considered the case of mutation, for STR, they can not be ignored, and we expect to observe, in true trios, an ‘‘exclusion’’ attributable to a single-step change once in 70 cases and with two once in every 9600.

But the comparison would be incomplete without evaluating the evidential value of the ‘‘exclusions’’ in each case. Beginning with those by the 2nd rule, in the case of binary systems the resulting paternity index amounts to 2×10^{-2} while for a ‘‘typical’’ STR locus it takes the value of 3×10^{-1}. With respect to the ‘‘exclusions’’ by the 1st. rule, while for a STR locus, the values are in the range of 10^{-2}, for SNPs, although extremely rare, they have a much more negative evidential value, around 10^{-8}.

It can be argued that our analyses depend a lot on the frequency of silent genes, and that the range we used (0.001–0.005) to make our simulations overestimates it. However, empirical data, such as those provided in Table 1 of the AABB annual report for the 2001 Paternity Testing Program [12], just show the opposite: most of the PCR based loci analysed have shown to host nulls (for those who have not, samples sizes are still small) and some of the observed frequencies are, indeed, much higher. It can be further discussed the suitability of current estimates (obtained with STR loci) for newly introduced SNP PCR based loci. We agree that it is possible to design more ‘‘robust’’ primers, but this possibility applies equally to both types of markers.

Another line of reasoning that could be raised against our analyses could be formulated as follows: if we would use a battery of, say, 45 SNPs (a theoretical minimum as we concluded above), then the probability of observing an isolated ‘‘true’’ exclusion (i.e., when a man is wrongfully accused) among such a set of loci should be vanishingly small, and, therefore, isolated exclusions could be safely equated to true paternity. A simple calculation using the parameters' values stated above ($p = q = 0.5$) demonstrates that it is not exactly the case, even assuming total absence of nulls. Indeed, for binary systems the exclusion chance, per locus, is $pq(1-pq) = 0.1875$. Thus, the probability of facing a situation in which a false father is excluded in a single system out of 45 is $45 \times 0.1875 \times (1–0.1875)^{44} = 9.09 \times 10^{-4}$. So, even under these very unrealistic assumptions an isolated exclusion is shown to be not a practical indication of non-paternity since it would be just less frequent (some 10 times) than ‘‘false’’ exclusions in true trios.

Weighting all these considerations, we will try to summarise the results of our analyses stating that a SNP battery of loci as a substitute for the STR based ones showed up to bring some unexpected drawbacks. Indeed, it was demonstrated that such a battery would be prone, if applied to routine paternity investigation, to the occurrence of a higher frequency of cases where the statistical evidence is inconclusive.

In conclusion, the prospect of the introduction, in a near future, of a SNP based strategy for kinship analysis, is shadowed with predictable statistical problems that must be properly evaluated and taken into account before considering the substitution of the now classical STR approach.

Acknowledgements

This work was partially supported by a research grant to LP (SFRH/BPD/7121/2001) from Fundação para a Ciência e a Tecnologia and IPATIMUP by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

References

[1] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Huhne, B. Rolf, Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat, Am. J. Hum. Genet. 62 (1998) 1408–1415.

[2] M.W. Nachman, S.L. Crowell, Estimate of the mutation rate per nucleotide in humans, Genetics 156 (2000) 297–304.

[3] A.S. Kondrashov, Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases, Hum. Mutat. 21 (2003) 12–27.

[4] P. Gill, An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, Int. J. Legal Med. 114 (2001) 204–210.

[5] M.P. Baur, R.C. Elston, H. Gurtler, K. Henningsen, K. Hummel, H. Matsumoto, W. Mayr, J.W. Moris, L. Niejenhuis, H. Polesky, et al. No fallacies in the formulation of the paternity index, Am. J. Hum. Genet. 39 (1986) 528–536.

[6] A. Amorim, J. Rocha, The bias of the dichotomy exclusion/non-exclusion and the evidential value of L (or W), Adv. Forensic Haemogen. 2 (1988) 603–606.

[7] E.A. Cotton, R.F. Allsop, J.L. Guest, R.R. Frazier, P. Koumi, I.P. Callow, A. Seager, R.L. Sparkes, Validation of the AMPFlSTR SGM plus system for use in forensic casework, Forensic Sci. Int. 112 (2000) 151–161.

[8] C. Alves, A. Amorim, L. Gusmão, L. Pereira, VWA STR genotyping: further inconsistencies between Perkin-Elmer and Promega kits, Int. J. Legal Med. 115 (2001) 97–99.

[9] C. Leibelt, B. Budowle, P. Collins, Y. Daoudi, T. Moretti, G. Nunn, D. Reeder, R. Roby, Identification of a D8S1179 primer binding site mutation and the validation of a primer designed to recover null alleles, Forensic Sci. Int. 133 (2003) 220–227.

[10] M.J. Lercher, L.D. Hurst, Human SNP variability and mutation rate are higher in regions of high recombination, Trends Genet. 18 (2002) 337–340.

[11] D.E. Reich, S.F. Schaffner, M.J. Daly, G. McVean, J.C. Mullikin, J.M. Higgins, D.J. Richter, E.S. Lander, D. Altshuler, Human genome sequence variation and the influence of gene history, mutation and recombination, Nat. Genet. 32 (2002) 135–142.

[12] American Association of Blood Banks. Annual Report for Testing in 2001. Prepared by the Parentage Testing Program Unit. October 2002.