Forensic Science International: Genetics xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

Empirical support for the reliability of DNA interpretation in Croatia

Gordan Lauc [a,b], Snjezana Dzijan [a], Damir Marjanovic [b], Simon Walsh [c], James Curran [d], John Buckleton [e,*]

[a] DNA Laboratory, University of Osijek School of Medicine, J. Huttlera 4, 31000 Osijek, Croatia
[b] Genos, Trg. Lj. Gaja 6, 31000 Osijek, Croatia
[c] Forensic & Data Centres, Australian Federal Police, GPO Box 401, Canberra, ACT 2601, Australia
[d] Department of Statistics, University of Auckland, Auckland, New Zealand
[e] ESR, PB 92021 Auckland, New Zealand

ARTICLE INFO

Article history:
Received 28 May 2008
Received in revised form 8 August 2008
Accepted 27 August 2008
Available online xxx

Keywords:
DNA interpretation
Croatia
Subpopulations

ABSTRACT

The result of empirical testing of forensic DNA match probabilities for Croatia is reported. It is concluded that if consideration is given to relatedness and subpopulation effects the model of Balding and Nichols appears to give very good predictions.

1. Introduction

In 2003, Birus et al. [1] reported data arising from their work on war victims in Croatia. Subpopulation effects were suspected in this dataset from the observation that the number of partial matches appeared to be above expectation from the product rule model. This paper has been quoted in Australian Courts as suggesting that existing population genetic models are inadequate [2].

The subject of suggested excess of matching or partial matching profiles is one that is causing considerable discussion on forensic discussion groups, web sites and newspapers. For example the observation of partial matches in the Arizona database [3] has become a topic of heated discussion in the US:

> "At stake is the credibility of the compelling odds often cited in DNA cases, which can suggest an all but certain link between a suspect and a crime scene" [4].

There is much less discussion in the refereed literature where such matters should really be settled. The issue is not the presence of partial matches per se but whether these are occurring at a rate greater than predicted by the relevant population genetic model. There have been a number of attempts to assess the population genetic models used in forensic science many centred around the pioneering work of Evett using what he termed Tippett tests [5–18]. With a dataset of size N it is possible to make $N(N-1)/2$ comparisons between pairs of profiles. A problem with comparing pairs in a database is that, if a match is observed, it is difficult to know whether the match originates from the same or different people. Matching profiles may be the same person sampled separately, twins, close relatives, or unrelated people. Investigation of these matches can be hampered by practical or legal considerations. Weir suggested focussing on a comparison of the observed and expected number of partially matching profiles [19] as a way around this problem. The Weir approach accounts for subpopulation effects. However the number of partially matching profiles could also be affected by the presence of relatives in the dataset examined and this has been suggested as an explanation, again usually in the non-refereed correspondence. However the method of Weir has been extended to account for the presence of relatives in a dataset [20] and hence the observed and expected may now be compared accounting for most plausible explanations for partial matches.

Work has continued on this identification work and a larger set of profiles (unpublished) is now available.

Given

1. The previous suspicions of subpopulation effects in the Croatian population, and
2. Their potential impact on conclusions drawn from interpretation models, and

* Corresponding author. Tel.: +64 9 8153 904; fax: +64 9 8496 046.
E-mail address: john.buckleton@esr.cri.nz (J. Buckleton).

1872-4973/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2008.08.005

3. The availability of an expanded database from Croatia, and
4. Methods for analysis of this database, and
5. The great interest in the fit of partial matches to expectation

We believe it is both timely and useful to examine a large range of population datasets for fit to the model and especially to examine the expanded Croatian dataset.

2. Methods

It is necessary to describe the full and partial matches obtained using the 15 locus Identifiler™ DNA profiling system. The nomenclature is given in Table 1 (we follow Curran et al. [20]).

For the sake of clarity in the ensuing discussion, we define at this point the 'top end of the distribution' as those partially matching pairs of profiles that are nearest to a full match. These would be the 14/1's, 14/0's, 13/2's etc. for a 15 locus system. The 'bottom end of the distribution' would then be those pairs of profiles that are nearest to complete mismatches. These are 0/0, 0/1, 0/2, etc.

The key in this procedure is not in determining the observed number of matches but in estimating the expected number.

At the top end of the distribution, we expect that any observed partial matches are more likely to be close relatives than unrelated individuals. Therefore, we would expect that the observations would be above expectation if relatedness were not taken into account. It is not unreasonable to assume, given the size of modern forensic databases, that some cousins or brothers for example would be present, and in many cases it is known that relatives are on the database. In the ensuing discussion we tend to emphasise the top end of the distribution since this is the end that is the best predictor of the point of interest, the full matches.

Obtaining estimates of the expected number of occurrences of each partial match follows Curran et al. [20]. We briefly reprise the Curran et al. method here. The probability of each type of full or partial match as described in Table 1 is estimated for four classes of relationship between people taking into account the specific relationship and the co-ancestry coefficient, θ. The classes of relationship modelled were unrelated, siblings, parent/child and cousins. For each relationship we assume that alleles may be identical because of chance, because they are identical by descent from the modelled relationship, or because they are identical by descent from the background co-ancestry, θ. The effect of co-ancestry is modelled using the approach of Balding and Nichols [21] which is the analogue of the National Research Council recommendation 4.2 equations 4.10 [22]. This is the model used in much of Europe and Australasia but not the US.

Relationships other than those modelled are certainly present in the data. We assume that they are of sufficiently low probability, sufficiently similar to one of the existing relationships, or of sufficiently low impact that they can be ignored without serious consequence.

Table 1
Nomenclature used to describe full and partial matches

Term	Nomenclature
15/0	A full 15 locus match
14/1	A 14 locus match with one allele of the last locus matching
14/0	A 14 locus match with no alleles of the last locus matching
13/2	A 13 locus match with one allele at each of the last two loci matching
13/1	A 13 locus match with one allele at one of the last two loci matching
13/0	A 13 locus match with no alleles at either of the last two loci matching
...	
7/2	A seven locus match with one allele matching at two of the other eight loci matching
7/1	A seven locus match with one allele matching at one of the other eight loci matching
...	

We will weight the contribution to the partial matches for each of the four classes of relationship. Since the weights for the four classes of relationship must add to one, we have only three weights to specify, the fourth being determined by subtraction. The co-ancestry coefficient, θ, is used to adjust for background relatedness in the population. Therefore, we have four free parameters; three weights and the co-ancestry coefficient. Let the number of pairwise comparisons be denoted

$$N_{\text{Comp}} = \frac{N(N-1)}{2}$$

We label the weights as

- α The fraction of comparisons that are between siblings. *Note*: This is related to the number of pairs of siblings (N_B) on the database by the relationship $\alpha = N_B/N_{\text{Comp}}$.
- β The fraction of comparisons that are between cousins.
- δ The fraction of comparisons that are between a parent and a child.
- θ The co-ancestry coefficient

The weights, α, β, δ and the value of θ are chosen so that the difference between the observed (O) and expected (E) counts in each category is minimized with respect to a distance measure. Curran et al. suggested that a statistic that fits the distributions across the full range from the bottom to the top of the distribution is given by the relative proportion of error, i.e.

$$\sum \frac{|O-E|}{E}$$

The final parameters at best fit do not represent certain knowledge of these particular fractions of relatives but suggest that these fractions, if present, would best explain the observations.

From the expanded Croatian database of relatives of missing people 295 individuals were selected, with known relatives removed. There are 259 full Identifiler™ profiles in this dataset. This gives a total of 33,411 pairs available for comparison.

3. Results

The numbers of observed and expected matches are presented at the optimal fit for the parameters for relatedness and population substructure (Fig. 1, left). There were no observed partial matches beyond 7/8 and accordingly the high end of the X-axis of the graph consists of observed counts of zero and low values of expected counts. We have therefore truncated the graph. The optimal fit parameters may be difficult to understand but translate to those estimates of relatedness in the dataset that would best fit the observations. This best fit occurs at 24 pairs of siblings, 1434 pairs of cousins and 37 parent–child pairs with the remaining 31,916 pairs being treated as unrelated. Since some effort has been made to remove all known pairs of relatives then the pairs predicted above are either unknown relationships still present in the dataset or an artefact of the optimisation process. The number seems high if they are undetected pairs of relatives. The latter effect would occur if there were some pairs of people who, although unrelated by recent ancestry, had very high level of co-ancestry because of multiple common ancestors further in the past and in many ways such relatedness would mimic subpopulation effects. Whether these pairs of relatives do indeed exist in the dataset is, surprisingly, irrelevant. This is because the observation is that a model with these parameters explains the data well. It does, however suggest that some accommodation for relatives is necessary, at least in this

Please cite this article in press as: G. Lauc, et al., Empirical support for the reliability of DNA interpretation in Croatia, Forensic Sci. Int. Gene. (2008), doi:10.1016/j.fsigen.2008.08.005

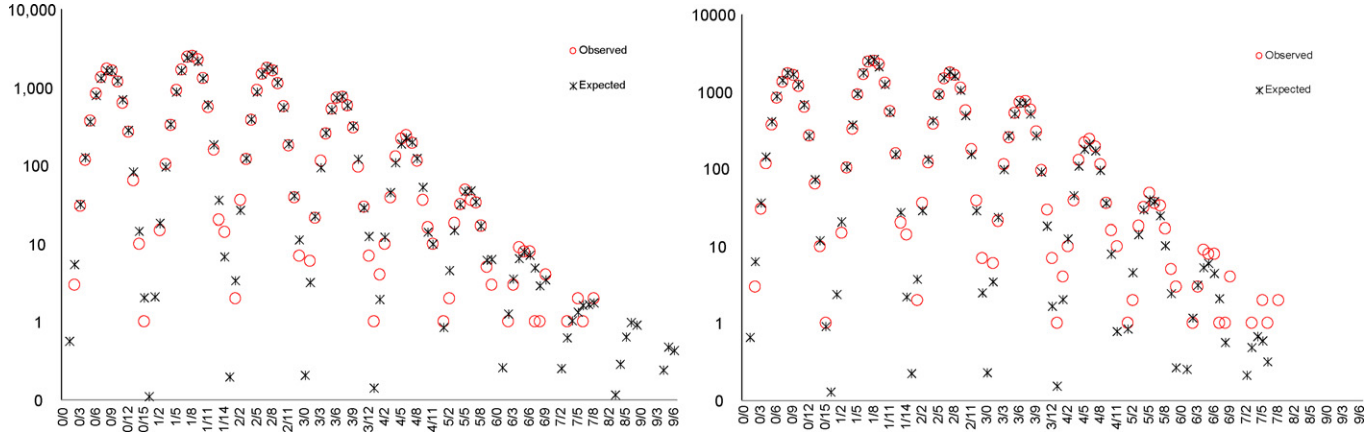

Fig. 1. Observed vs. expected for the partially matching profiles at the optimised fitting parameters (left) and using the raw product rule with no accommodation for substructure or relatedness (right). The fitting parameters at optimisation were: $\alpha = 7 \times 10^{-4}$, $\beta = 4.3 \times 10^{-2}$, $\delta = 1.1 \times 10^{-3}$, $\theta = 0.003$.

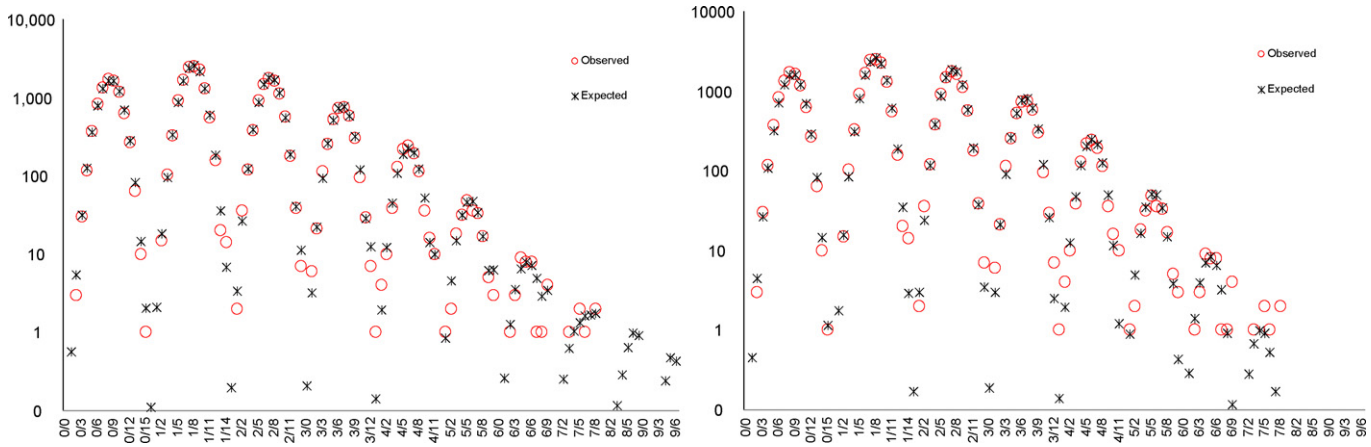

Fig. 2. Observed vs. expected for the partially matching profiles using the subpopulation model with $\theta = 0.01$ and no accommodation for relatedness (right) with the fit at the optimised fitting parameters (left) for comparison.

dataset, to explain the partial matches.The value for θ at optimisation was 0.003 which is quite low and does not suggest excessive substructure once relatedness has been accommodated.

It is not trivial to assess the quality of fit in these graphs. We suggest that the readers compare visually the observed and expected for partial matches at the top end of the scale as this is the end of the most evidential interest, being the nearest to full matches. The reader may, for example, concentrate in the 7/2 to 7/8 range. We will also present the distance measure which is a composite statistic across the whole range.

The fit of observed and expected for the raw product rule with no accommodation for relatedness was inferior (see Fig. 1, right, distance function 323) to the optimal fit with both relatedness and population substructure (see Fig. 1, left, distance function 70.9).

Table 2
The parameters for relatedness at subpopulation effects at optimal fit for various databases

	α	β	δ	θ
Australian Caucasians	7.80×10^{-6}	9.38×10^{-4}	1.14×10^{-8}	0.00
Australian Aborigines	5.95×10^{-5}	4.99×10^{-2}	3.65×10^{-11}	0.03
West. Aust. Caucasians	6.91×10^{-6}	3.08×10^{-3}	2.95×10^{-7}	0.00
West. Aust. Aboriginals	8.32×10^{-5}	1.19×10^{-15}	2.05×10^{-5}	0.00
NZ Caucasians	7.09×10^{-6}	6.51×10^{-10}	2.77×10^{-8}	0.00
NZ Eastern Polynesians	5.43×10^{-5}	3.06×10^{-2}	5.64×10^{-14}	0.00
NZ Western Polynesian	9.02×10^{-5}	7.20×10^{-3}	1.36×10^{-6}	0.00
Croatian Caucasians	7×10^{-4}	4.3×10^{-2}	1.1×10^{-3}	0.003

The product rule with relatedness gave a graphic virtually identical to Fig. 1, left, and is not given due to its similarity. The distance measure for this was 71.6, similar to that for the full optimisation with subpopulation effects and relatedness.

This agrees with earlier observations [1] that there appear to be more partial matches than predicted by the product rule. This is a key finding of both this and the earlier work but the work here suggests that the key modelling parameters, at least for this dataset and multiplex, are a consideration of relatedness, not of the subpopulation effect.

The fit to the subpopulation model with the commonly used value $\theta = 0.01$ and no accommodation for relatedness (Fig. 2, right, distance function 209) was also inferior to the subpopulation model with relatedness (Figs. 1 and 2, left, distance function 70.9) and was, in fact, only a moderate improvement in fit over the raw product rule (Fig. 1, right, distance function 323).

4. Conclusions

The overlap of observed and expected appears to be exceptionally good to us as long as relatedness and a small level of substructure are successfully modelled. The raw product rule with no accommodation for relatedness and the subpopulation model with no accommodation for relatedness seem to predict too few matches. However the product rule with an accommodation for relatedness gives only a marginally poorer fit that the subpopulation model with relatedness.

Please cite this article in press as: G. Lauc, et al., Empirical support for the reliability of DNA interpretation in Croatia, Forensic Sci. Int. Gene. (2008), doi:10.1016/j.fsigen.2008.08.005

We conclude from this that the primary factor in modelling match probabilities, in this Croatian dataset at least, is to adequately model relatedness.

Previous work has investigated this type of fit to Australian Caucasians, Australian Aborigines, New Zealand Caucasians, New Zealand Eastern Polynesians, and New Zealand Western Polynesians. [20] Direct comparison of the fit of observed and expected using the distance measure for these different populations is hampered by the fact that our distance measure is not independent of the size of the database and the number of loci. The databases in the earlier work were between 8000 and 17,000 in size hence random effects in the number of observed partial profiles would be much less in those databases than in the current work. This may be seen by looking at the observed values on the *Y*-axes which are typically larger by orders of magnitude in the earlier work. However the modelling parameters at optimal fit may be compared between the different population datasets (Table 2).

In all cases the co-ancestry coefficient, θ, at optimal fit is small with the possible exception of Australian Aborigines. This indicates that subpopulation effects are usually minimal. However the relatedness parameters at optimal fit are small but non-negligible in that they have a significant effect on the expected number of partial matches at the top end of the distribution. This is completely in line with the expectation that most of the excess of partial matches would be explained if a small number of pairs of relatives were present in the datasets.

The importance of relatedness has been predicted and is expected to become greater relative to subpopulation effects as we add more loci to our multiplexes [15]. However it is clear from articles such as the LA Times article quoted in the introduction [4] that there is still room for surprise at the potential magnitude of this anticipated genetic effect. It may be timely to start discussion about further practical ways, beyond the substantial efforts already made by some laboratories, in which the importance of relatedness may be appropriately and proportionately accommodated in casework and reported to court.

Acknowledgments

We gratefully thank Sally Coulson, Susan Vintiner, and three anonymous referees for comments that have greatly improved this paper.

References

[1] I. Birus, M. Marcikic, D. Lauc, S. Dzijan, G. Lauc, How high should paternity index be for reliable identification of war victims by DNA typing? Croat. Med. J. 44 (3) (2003) 322–326.
[2] R.v. Bropho, WADC, 182 (2004).
[3] L. Muller, Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? J. Genet. 87 (2) (2008), http://www.ias.ac.in/jgenet/Vol87No2/temp/jgen00133.pdf (accessed 10th August 2008).
[4] J. Felch, M. Dolan, How reliable is DNA in identifying suspects? Los Angeles Times, July 20, 2008, http://www.latimes.com/news/local/la-me-dna20-2008jul20.0.1506170.full.story (accessed 23 July, 2008).
[5] J. Buckleton, S. Walsh, S. Harbison, The fallacy of independence testing and the use of the product rule? Sci. Justice 41 (2001) 81–84.
[6] C.M. Triggs, J.S. Buckleton, Logical implications of applying the principles of population genetics to the interpretation of DNA profiling evidence, Forensic Sci. Int. 128 (2002) 108–114.
[7] K. Walsh, J.S. Buckleton, A discussion of the law of mutual independence and its application to blood-group frequency data, J. Forensic Sci. Soc. 28 (2) (1988) 95–98.
[8] P.D. Gill, L.A. Foreman, J.S. Buckleton, C.M. Triggs, H. Allen, A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2003) 184–196.
[9] J.M. Curran, J.S. Buckleton, C.M. Triggs, What is the magnitude of the subpopulation effect? Forensic Sci. Int. 135 (1) (2003) 1–8.
[10] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL., 2004.
[11] J.S. Buckleton, C.M. Triggs, The effective of linkage on the calculation of DNA match probabilities for siblings and half siblings, Forensic Sci. Int. 160 (2006) 193–199.
[12] J. Buckleton, J. Curran, S. Walsh, How reliable is the sub-population model in DNA testimony? Forensic Sci. Int. 157 (1–2) (2006) 144–148.
[13] S.J. Walsh, R.J. Mitchell, F. Torpy, J.S. Buckleton, Use of subpopulation data in Australian forensic DNA casework, Forensic Sci. Int.: Genet. 1 (3–4) (2007) 238–246.
[14] S.J. Walsh, R.J. Mitchell, N. Watson, J.S. Buckleton, A comprehensive analysis of microsatellite diversity in Aboriginal Australia, J. Hum. Genet. 52 (2) (2007) 712–728.
[15] J. Buckleton, C. Triggs, Relatedness and DNA: are we taking it seriously enough? Forensic Sci. Int. 152 (2005) 115–119.
[16] I.W. Evett, P.D. Gill, J.K. Scranage, B.S. Weir, Establishing the robustness of short-tandem-repeat statistics for forensic applications, Am. J. Hum. Genet. 58 (2) (1996) 398–407.
[17] I.W. Evett, J.A. Lambert, J.S. Buckleton, B.S. Weir, Statistical analysis of a large file of STR profiles of British Caucasians to support forensic casework, Int. J. Legal Med. 109 (1996) 173–177.
[18] I.W. Evett, B.S. Weir, Interpreting DNA Evidence–Statistical Genetics for Forensic Scientists, Sinauer Associates, Inc., Sunderland, 1998.
[19] B.S. Weir, Matching and partially matching profiles, J. Forensic Sci. 49 (5) (2004) 1009–1014.
[20] J.M. Curran, S.J. Walsh, J.S. Buckleton, Empirical testing of estimated DNA frequencies, Forensic Sci. Int.: Genet. 1 (3–4) (2007) 267–272.
[21] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[22] NRC_II, National research council committee on DNA forensic science, in: The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.

Please cite this article in press as: G. Lauc, et al., Empirical support for the reliability of DNA interpretation in Croatia, Forensic Sci. Int. Gene. (2008), doi:10.1016/j.fsigen.2008.08.005