

Simon J. Walsh,<sup>1,2</sup> B.Sc.(Hons); Christopher M. Triggs,<sup>3</sup> Ph.D.; James M. Curran,<sup>4</sup> Ph.D.;  
Judi R. Cullen,<sup>1</sup> M.Sc.; and John S. Buckleton,<sup>5</sup> Ph.D.

## Evidence in Support of Self-Declaration as a Sampling Method for the Formation of Sub-Population DNA Databases

**ABSTRACT:** Well constructed sub-population databases are fundamental to the application of DNA-based forensic statistics. The size of such databases can affect the ability to examine adequately statistical or population genetic features, and the integrity of both the DNA profile and associated ethnicity information is also of importance. Use of short tandem repeat (STR) DNA profiling technology and the thoughtful construction of the governing legislation has seen large databases of DNA profiles collated for the four major sub-populations of New Zealand. Examination of the data illustrates the suitability of self-declaration as a means of categorizing samples on the basis of ethnicity.

**KEYWORDS:** forensic science, STR (short tandem repeat), population data, genetic distance, self declaration, New Zealand

### Introduction

In general, the use of DNA profiling results in the context of a criminal or legal investigation must be qualified by an estimation of the statistical weight of the DNA evidence. There are various approaches used to arrive at such an estimate but all rely on reference to databases of DNA profiles, which are usually segregated into predominant sub-populations based on the ethnic origin of sample donors. The compilation and validation of sub-population databases to support forensic DNA interpretation is a prerequisite of forensic biology laboratory accreditation (1–3). In Australia, the accuracy and adequacy of such databases has been scrutinized closely in the criminal courts (4,5). Kaska et al. (6) and Ayres et al. (7) examined the comparability of sub-population data obtained under different mechanisms for assigning ethnicity, but there has been limited research into the mechanisms themselves.

In New Zealand, the major sub-populations are from diverse evolutionary backgrounds, and have only recently mixed. The construction of accurate sub-population databases and information regarding genetic diversity gained through analysis of the data are therefore of added importance.

The contemporary New Zealand population is predominantly made up of four major sub-populations, which we have classified as Caucasian, Eastern Polynesian, Western Polynesian, and Asian.

<sup>1</sup> Institute of Environmental Science and Research Ltd., Private Bag 92-021, Auckland, New Zealand.

<sup>2</sup> Science & Justice Consulting, PO Box 83, Surry Hills, NSW, 2010, Australia.

<sup>3</sup> Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

<sup>4</sup> Department of Statistics, University of Waikato, Hamilton, New Zealand.

<sup>5</sup> The Forensic Science Service, Trident Court, Birmingham, United Kingdom and \*Institute of Environmental Science and Research Ltd., Private Bag 92-021, Auckland, New Zealand (\* denotes present address).

Received 12 Oct. 2002; and in revised form 14 Feb. 2003; accepted 5 April 2003; published 4 Aug. 2002.

To have confidence in our statistical interpretation of forensic DNA casework, we considered it important to evaluate a key paradigm underpinning the population genetic and DNA interpretation models, namely, the self-declaration of ethnicity.

### Construction of Sub-Population Databases

#### DNA Profiling

Standard methods for DNA analysis are employed, as previously described in Harbison et al. (8). The database therefore contains a mix of 6-locus SGM and 10 locus AMPF $\ell$ STR<sup>®</sup>SGM Plus<sup>™</sup> (SGM Plus) profiles. No profiles with missing alleles or loci are loaded onto the DNA database meaning the sub-population data contain only full profiles.

#### "Self-Declaration" as a Source of Ethnicity Data

In 1996 the New Zealand Government introduced specific legislation governing the use of DNA profiling (*Criminal Investigations (Blood Samples) Act, 1995*). A provision in the legislation allows an individual donor to self declare their ethnicity over four generations. This provision was specifically included for the purpose of collating sub-population databases for use in the statistical interpretation of forensic DNA profiling results. Over 65% of individuals whose DNA profiles are held on the National DNA Database have voluntarily supplied information regarding their racial pedigree.

All persons who claimed any Maori, Cook Island, or Tokelauan ancestries were included in the Eastern Polynesian database (SGM profiles  $n = 6037$ ; SGM Plus profiles  $n = 1815$ ). All persons who claimed Samoan, Tongan, or Nuiean ancestries were included in the Western Polynesian database (SGM profiles  $n = 1305$ ; SGM Plus profiles  $n = 477$ ). All persons who claimed a full European ancestry (over the designated four generations) were included in the Caucasian database (SGM profiles  $n =$

3340; SGM Plus profiles  $n = 1001$ ). All persons who claimed Asiatic descent were included in the Asian database (SGM profiles  $n = 173$ ; SGM Plus profiles  $n = 114$ ). This can be seen as somewhat arbitrary but does coincide with common usage in New Zealand. It leads to the admixed persons all being classified in the non-Caucasian ethnic groups. The majority of people in the Asian database are Vietnamese or Chinese and this group does not include persons indicating their ethnic descent as from the Indian sub-continent (as would be common usage of this term in the UK). This research focuses on those individuals in self-declared admixed populations.

*Integrity of the Data*

Each database was screened for the presence of duplicate DNA profiles. Where duplicate profiles were identified, the origins of the sample were examined to determine whether or not the identical DNA profiles were due to a system anomaly, such as an individual donating a second databank sample under an alias, or a genuine genetic correspondence. Through police intelligence, the DNA findings were compared with other police records and fingerprint and photographic data. This process separated all spurious entries on the database from the genuine genetic matches. The investigators regard this process as vital to the integrity of the ensuing statistical data analysis.

*Analysis of Ethnicity Data*

According to the 2001 census, 80% of the New Zealand population declared themselves of Caucasian ancestry, 14.3% of Maori, 6.7% of Asian and 6.5% of "Pacific" (9). Immigration from Asia has been relatively recent and has doubled since 1991 (9). We do not expect much admixture to have occurred between Asian and New Zealand populations as yet although this process will be beginning and is expected to accelerate. Allele frequencies for the four sub-populations were presented in an earlier publication (10). As expected, there are differences in the frequencies of individual alleles at each locus tested.

*Genetic Distance*

An estimate of the pairwise genetic distance within and between certain sub-populations in the New Zealand STR profile data was made. Sub-population data from the Eastern Polynesian and Western Polynesian datasets was further divided into subsets represent-

ing varying levels of ethnicity within the selected subset. For example, New Zealand Maori samples from the Eastern Polynesian sub-population were distributed into six subsets: Maori,  $\frac{3}{4}$  Maori,  $\frac{1}{2}$  Maori,  $\frac{1}{4}$  Maori,  $\frac{1}{8}$  Maori,  $\leq \frac{1}{16}$  Maori. Similarly, Samoan samples from the Western Polynesian sub-population data were distributed into four sub-groups: Samoan,  $\frac{3}{4}$  Samoan,  $\frac{1}{2}$  Samoan,  $\leq \frac{1}{4}$  Samoan. This was possible through the self-declaration process which is based on ancestral information over four generations. Maori and Samoan samples were chosen as previous statistical assessments had shown evidence of sub-population and admixture effects in the Eastern and Western Polynesian data (results not shown, available on request). These datasets were also the only ones of sufficient size to allow subdivision by level of ethnicity, as described above.

The genetic distance,  $\theta$ , between the population subsets was calculated using the moment estimate (11,12) by use of the Genetic

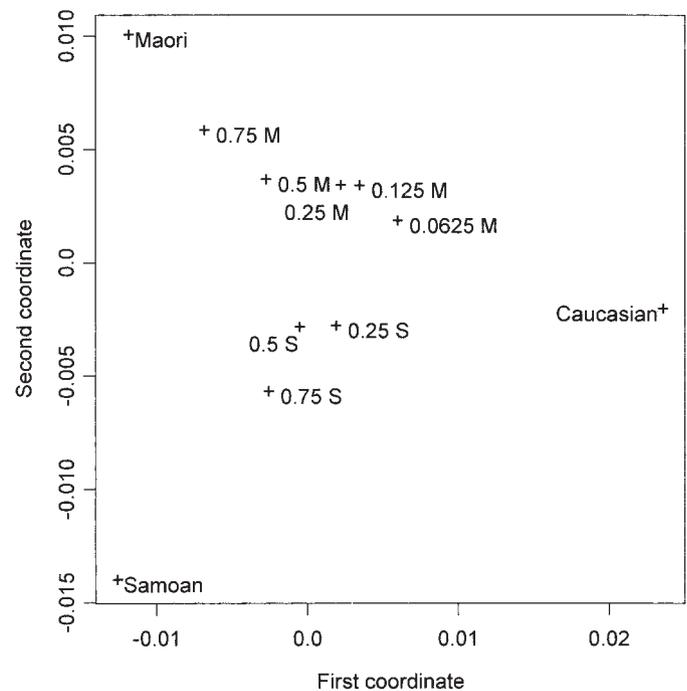


FIG. 1—Principal Coordinate representation of the inter-population genetic distances. The level of Maori (M) and Samoan (S) ethnicity for each sub-group is represented by the preceding integer, for example 0.25S indicates the  $\frac{1}{4}$  Samoan subset referred to in the text.

TABLE 1—Genetic distances ( $\times 1000$ ) between self reported ethnic groups. Negative estimates of genetic distance replaced by 0.

	Caucasian	Maori	$\frac{3}{4}$ Maori	$\frac{1}{2}$ Maori	$\frac{1}{4}$ Maori	$\frac{1}{8}$ Maori	$\frac{1}{16}$ Maori	Samoan	$\frac{3}{4}$ Samoan	$\frac{1}{2}$ Samoan
Maori	36.57									
$\frac{3}{4}$ Maori	29.65									
$\frac{1}{2}$ Maori	23.06	2.41	0.29							
$\frac{1}{4}$ Maori	14.09	6.60	3.33	1.63						
$\frac{1}{8}$ Maori	10.03	6.46	3.23	1.31	0.00					
$\leq \frac{1}{16}$ Maori	3.06	12.94	8.73	5.52	0.00	0.00				
Samoan	36.81	23.17	19.51	17.57	19.28	19.39	20.89			
$\frac{3}{4}$ Samoan	21.19	13.79	9.16	7.97	7.79	8.95	6.25	0.00		
$\frac{1}{2}$ Samoan	13.96	3.61	0.62	0.08	0.00	0.00	0.00	0.00	0.00	
$\frac{1}{4}$ Samoan	1.46	0.79	0.00	0.00	0.00	0.00	0.00	3.01	6.02	0.00

Data Analysis (GDA©) software (13). The matrix of genetic distances (Table 1) is represented using principal coordinates (14) by finding a set of points in two dimensions whose interpoint distances approximate the genetic distances (displayed graphically in Fig. 1). The genetic distance from the Caucasian sub-population increases as the level of Maori or Samoan ethnicity increases. This provides support for the effectiveness of self declaration as a means of segregating reference samples by ethnicity. The points corresponding to small reported fractions of Maori and Samoan ancestry are closer to each other than they are to the point representing the Caucasian population. We believe that this is because the admixture is complex, and a person reporting a small fraction of, say, Samoan ancestry may also have some Maori as well as Caucasian ancestors.

### Conclusions

As it is the role of forensic scientists to provide the most accurate estimate of the evidential significance of DNA profiling results presented to the courts, examination and understanding of certain population genetic features of the people within their jurisdiction is required. New Zealand's geographical association with the Pacific Islands has seen the growth of a diverse population, heavily influenced by the founding Eastern Polynesian peoples. Statistical scrutiny of large pools of DNA profile data generated through forensic analysis has highlighted some features for consideration. In particular, the compilation of large sub-population databases can be facilitated through DNA-based legislation, with accurate assignment of ethnicity achieved through self declaration.

### Acknowledgments

We wish to thank Dr. Peter Gill, of the Forensic Science Service for comments that greatly improved this manuscript. This research was funded in part by the New Zealand Police Capability Development Program.

### References

1. Technical Working Group on DNA Methods. Guidelines for a quality assurance program for DNA analysis. *Crime Laboratory Digest* 1995;22:18–43.
2. U.S. Department of Justice. Federal Bureau of Investigation DNA Advisory Board. Quality assurance standards for forensic DNA testing laboratories. October 1998. *Forensic Sci Communications* 2000;2(3).
3. National Association of Testing Authorities, Australia Accreditation Requirements. See "Supplementary requirements for accreditation in the field of Forensic Science." ISO/IEC 17025 Application Document, 2000; Version 1.
4. *R v Tran* (1990) 50 A Crim R 233.
5. *R v Pantoja* (1996) 88 A Crim R 554.
6. Kaska DE, van Oorschot RA, Mitchell RJ. Variation at three short tandem repeat (STR) loci in Australians: forensic and ethnic considerations. *Electrophoresis* 1997;18:1620–3.
7. Ayres KL, Chaseling J, Balding DJ. Implications for DNA identification arising from an analysis of Australian forensic databases. *Forensic Sci Int* 2002;129:90–8.
8. Harbison SA, Hamilton JF, Walsh SJ. The New Zealand DNA databank: its development and significance as a crime solving tool. *Sci Justice* 2001;41(1):33–7.
9. Statistics New Zealand (available at [www.stats.govt.nz](http://www.stats.govt.nz) accessed on 29 January 2003).
10. Walsh SJ, Cullen JR, Harbison SA. Allele frequencies for the four major sub-populations of New Zealand at the 10 AMPFLSTR SGM Plus loci. *Forensic Sci Int* 2001;122(2–3):189–95.
11. Reynolds J, Weir BS, Cockerham C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 1983;105:767–79.
12. Weir BS. *Genetic data analysis II*. Sinauer, Sunderland, Massachusetts, 1996.
13. Lewis PO, Zaykin D. *Genetic data analysis: computer program for the analysis of allelic data*. 2001; Version 1.0 (d16c). Free program distributed by the authors at <http://lewis.eeb.uconn.edu/lewishome/software.html>.
14. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 1966;53:325–38.

Additional information and reprint requests:

Mr. Simon J. Walsh  
 Science and Justice Consulting  
 P.O. Box 83  
 Surry Hills, NSW, 2010  
 Australia