biology letters
Phylogeny

Biol. Lett. (2009) **5**, 849–852
doi:10.1098/rsbl.2009.0462
Published online 8 July 2009

FORDISC and the determination of ancestry from cranial measurements

Marina Elliott and Mark Collard*

Laboratory of Human Evolutionary Studies, Department of Archaeology, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6
**Author for correspondence* (*mark.collard@sfu.ca*).

Determining the ancestry of unidentified human remains is a major task for bioarchaeologists and forensic anthropologists. Here, we report an assessment of the computer program that has become the main tool for accomplishing this task. Called FORDISC, the program determines ancestry through discriminant function analysis of cranial measurements. We evaluated the utility of FORDISC with 200 specimens of known ancestry. We ran the analyses with and without the test specimen's source population included in the program's reference sample, and with and without specifying the sex of the test specimen. We also controlled for the possibility that the number of variables employed affects the program's ability to attribute ancestry. The results of the analyses suggest that FORDISC's utility in research and medico-legal contexts is limited. FORDISC will only return a correct ancestry attribution when an unidentified specimen is more or less complete, and belongs to one of the populations represented in the program's reference samples. Even then FORDISC can be expected to classify no more than 1 per cent of specimens with confidence.

Keywords: bioarchaeology; forensic anthropology; unidentified human remains; ancestry determination; cranial variation; FORDISC

1. INTRODUCTION

FORDISC is a popular computer program designed to determine the ancestry of modern human skeletal specimens through discriminant function analysis (Jantz & Ousley, 2005). Recently, the utility of FORDISC has become a subject of debate (Kosiba 2000; Leathers *et al.* 2002; Ubelaker *et al.* 2002; Freid *et al.* 2005; Williams *et al.* 2005; Naar *et al.* 2006; Hubbe & Neves 2007; Keita 2007; Ousley *et al.* 2009). A number of researchers have applied the program to specimens of known ancestry and concluded from the high numbers of incorrect attributions that the program is flawed (Kosiba 2000; Leathers *et al.* 2002; Williams *et al.* 2005). In response, FORDISC's developers have argued that the program's poor performance results from mistakes in the implementation of the program and/or misinterpretation of its output (Freid *et al.* 2005). Given the importance in bioarchaeology and forensic anthropology of determining the ancestry of unidentified skeletal specimens and the frequency with which FORDISC is used for this purpose, there is a pressing need to determine which of these claims is correct.

Here we report a study that employs a larger sample of test specimens than previous studies and addresses the main points of contention in the ongoing debate. One of the latter is the presence/absence of the target specimen's source population in the reference sample. Most previous evaluations of the program have employed test specimens from populations not represented in its reference datasets. FORDISC's developers have rejected this approach on the grounds that the nature of discriminant function analysis is such that the program can only be expected to perform adequately if a specimen's source population is represented in the reference sample (Freid *et al.* 2005). With this in mind, we analysed each test individual with and without their source population included in the reference sample. A second controversial issue relates to variable number. Some researchers contend that FORDISC's performance can be expected to improve as variable number increases (Hubbe & Neves, 2007), while others argue that using too many variables reduces FORDISC's reliability (Jantz & Ousley 2005). We dealt with this issue by carrying out analyses with different numbers of variables. In addition, because several studies have found that FORDISC's ancestry attributions change when the sex of the test specimen is altered (e.g. Williams *et al.* 2005), we carried out analyses in which both male and female specimens were included in the reference sample as well as analyses that only used reference specimens of the same sex as the test specimen.

2. MATERIAL AND METHODS

The test dataset consists of values for 56 variables recorded on 10 male and 10 female crania from each of the following populations: Berg (Europe), Hokkaido Japanese (Asia), Santa Cruz (Americas), Tasmanians (Australia and Pacific) and Zulus (Africa). The values were obtained from Howells' (1996) global craniometric dataset, which is one of the two reference datasets that FORDISC uses to generate discriminant functions. The variables represent the largest set for which all the test groups have values. Further details of the variables are given in table S1 of the electronic supplementary material.

We began by analysing each test individual with and without their source population included in the reference sample. All 56 variables were used in these analyses, and the sex of the test specimens was left unspecified. Next, we conducted a series of analyses designed to control for the possibility that using too many variables negatively affects FORDISC's performance (Jantz & Ousley 2005). These analyses were based on three non-overlapping sets of 10 variables (table S2 of the electronic supplementary material). The latter number was calculated from a formula provided by Jantz & Ousley (2005) for determining the number of variables that should be used in a FORDISC analysis. We used non-overlapping sets of variables to control for the possibility that cranial regions differ in their usefulness for determining ancestry (Harvati & Weaver 2006). We carried out six 10-variable analyses. In the first three, all populations were included in the reference samples. In the others, a test specimen was analysed only after its source population was excluded from the reference sample. As before, the sex of the test specimens was not specified. Finally, we ran a series of analyses to control for the possibility that FORDISC's performance is affected by the sex of the target specimen. In these analyses, we used 56 variables again but compared the test specimens only to reference specimens of the same sex.

In the source-population-included analyses, we evaluated FORDISC's performance on the basis of the percentage of test specimens correctly assigned to their source population. In the source-population-excluded analyses, we assessed the program's performance on the basis of the percentage of test specimens assigned to the most closely related population in the reference

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2009.0462 or via http://rsbl.royalsocietypublishing.org.

Received 10 June 2009
Accepted 18 June 2009

sample. After reviewing the available evidence, we selected the Norse (Europe), Kyushu (Asia), Yauyos (Americas), mainland Australian Aborigines (Australia and Pacific) and Teita (Africa) as the closest relatives of the Berg, Hokkaido Japanese, Santa Cruz, Tasmanian and Zulu, respectively (table S3 of the electronic supplementary material).

The FORDISC manual recommends that a population attribution should be accepted only if the posterior probability (PP) is greater than 0.5 and the typicality probability (TP) greater than 0.01. However, at a FORDISC workshop in February 2007 the program's designers suggested that determinations with a PP less than 0.8 are more likely to be incorrect than correct. Accordingly, we calculated the number of correctly classified specimens once accepting an attribution if $PP > 0.5$ and $TP > 0.01$, and once if $PP > 0.8$ and $TP > 0.01$. Incorrect attributions and correct attributions with PPs and TPs lower than the sectioning points were grouped together and designated 'failed attributions'.

3. RESULTS

The results of the source-population-included/excluded analyses are summarized in the first two rows of table 1. When PP was set at greater than 0.5 and TP at greater than 0.01, FORDISC correctly classified 80 per cent of the test specimen's in the source-population-included analyses, and 24.5 per cent in the source-population-excluded analyses. When PP and TP were set at greater than 0.8 and greater than 0.01, respectively, the corresponding figures were 69.5 per cent and 12 per cent.

Rows three to eight of table 1 summarize the results of the six 10-variable analyses. In the source-population-included analyses, FORDISC correctly assigned between 2 per cent and 17 per cent of the test specimens when PP was set at greater than 0.5 and TP at greater than 0.01, and between 0.5 per cent and 4.5 per cent when the PP was set at greater than 0.8 and TP at greater than 0.01. In the source-population-excluded analyses, FORDISC correctly assigned between 1 per cent and 3 per cent when PP was set at greater than 0.5 and TP at greater than 0.01, and between 0 per cent and 0.5 per cent when PP was set at greater than 0.8 and TP at greater than 0.01.

The results of the analyses carried out to assess the effect of specifying the sex of the test specimen are summarized in rows nine and 10 of table 1. In the sex-specified, source-population-included analyses, FORDISC correctly classified 85.5 per cent of the test specimens when PP was set at greater than 0.5 and TP at greater than 0.01, and 78 per cent of the test specimens when PP was set at greater than 0.8 and TP at greater than 0.01. In the sex-specified, source-population-excluded analyses, FORDISC correctly classified 27 per cent of the test specimens when PP was set at greater than 0.5 and TP at greater than 0.01, and 19 per cent of the test specimens when PP was set at greater than 0.8 and TP at greater than 0.01.

Table 1. Results of tests of FORDISC using 200 specimens of known ancestry. PP, posterior probability; TP, typicality probability. Twenty-eight populations were included in the source-population-included analyses, and 27 in the source-population-excluded analyses.

analysis	$PP > 0.5$/ $TP > 0.01$	$PP > 0.8$/ $TP > 0.01$
56 variables, source population included, sex unspecified	160 (80.0%)	139 (69.5%)
56 variables, source population excluded, sex unspecified	49 (24.5%)	24 (12.0%)
10 basicranium variables, source population included, sex unspecified	4 (2.0%)	1 (0.5%)
10 basicranium variables, source population excluded, sex unspecified	2 (1.0%)	0 (0.0%)
10 neurocranium variables, source population included, sex unspecified	34 (17.0%)	9 (4.5%)
10 neurocranium variables, source population excluded, sex unspecified	5 (2.5%)	1 (0.5%)
10 face variables, source population included, sex unspecified	29 (14.5%)	5 (2.5%)
10 face variables, source population excluded, sex unspecified	6 (3.0%)	0 (0.0%)
56 variables, source population included, sex specified	171 (85.5%)	156 (78.0%)
56 variables, source population excluded, sex specified	54 (27.0%)	38 (19.0%)

4. DISCUSSION

The results of the analyses suggest that the use of FORDISC for ancestry determination is problematic. The program correctly classified more than 70 per cent of the test specimens in some analyses. But the analyses in question were the ones in which not only were 56 variables employed but also the test specimen's source population was included in the reference sample. In all the other analyses, less than 40 per cent of the test specimens were classified correctly. This suggests that FORDISC is only likely to be useful when an unidentified specimen is more or less complete and belongs to one of the populations represented in its reference samples. Importantly, a specimen must belong to a population in the reference sample and not just be closely related to one of them. The program's poor performance in the analyses in which the test specimen's source population was excluded from the reference sample suggests that it cannot be relied on to assign an unidentified specimen to a closely related population in the absence of its own group. Because FORDISC's reference datasets contain fewer than 30 populations, the chances that an unidentified specimen's group will be represented in them are low. Given this, and the fact that complete crania are uncommon in archaeological and forensic contexts, there is reason to believe that FORDISC will only rarely identify the ancestry of an unidentified specimen.

In fact, FORDISC may be even less useful than our results suggest. During the course of our study, it became apparent that the PP/TP sectioning points

Table 2. Minimum and maximum probabilities returned for correct and incorrect assignments in 'best' analyses (source population included, 56 variables, sex-specified). PP, posterior probability; TP, typicality probability.

	min and max PP/TP for correct assignments				min and max PP/TP for incorrect assignments			
test population	PP min	PP max	TP min	TP max	PP min	PP max	TP min	TP max
Berg	0.646	1.000	0.000	0.947	0.521	0.876	0.000	0.643
Santa Cruz	0.752	1.000	0.077	0.942	–	–	–	–
Hokkaido Japanese	0.593	1.000	0.196	0.964	0.447	0.850	0.000	0.952
Tasmania	0.873	1.000	0.043	0.935	0.436	0.991	0.327	0.690
Zulu	0.546	1.000	0.000	0.964	0.389	0.939	0.440	0.482

recommended by FORDISC's designers do not in fact separate correct and incorrect attributions. We found that there were always some incorrect attributions among the attributions with PPs and TPs that exceeded the sectioning points. With this in mind, we calculated new PP and TP sectioning points from the results of our 'best' analysis (source population included, 56 variables, sex-specified). In the latter analysis, the PPs associated with incorrect assignments ranged from 0.389 to 0.991, while the TPs associated with incorrect assignments ranged from 0 to 0.952 (table 2). Thus, our best analysis suggests that the sectioning points for PP and TP should be 0.991 and 0.952, respectively. If we had used these sectioning points, 198 of 200 (99%) determinations in our 'best' analysis would have been classified as failed attributions. This suggests that even in favourable conditions—when the focal specimen's source population is present in the reference sample, the focal specimen is nearly complete and its sex is known—FORDISC has no more than a 1 per cent chance of success.

There are several reasons for suspecting that even this may overstate FORDISC's usefulness. First, Howells' collection strategy for his populations was not random. Rather, he 'carefully selected' crania that he considered to be typical of each group (Howells 1995, p. 3). Crania that were 'morphologically unusual for the population as a whole' (Howells 1989, p. 89) were not included, even if there were no obvious pathological changes to account for the differences. For some, this meant that only a small percentage of the available individuals were measured. For example, 50 ancient Egyptian crania were selected from a sample of nearly 1800. Thus, the degree of overlap among the reference populations is likely to be artificially low. Given that classification success in discriminant function analysis is inversely related to the degree of overlap among groups, the analyses reported here probably overestimate FORDISC's ability to attribute ancestry. Second, a number of the specimens Howells measured were sexed on the basis of cranial morphology alone (Howells 1989). Although Howells attempted to corroborate his estimates with those of other researchers who had examined the remains, he admitted that some of the skulls of known sex 'would certainly have been assigned to the wrong sex if it had been done by inspection' (Howells 1989, p. 94). This suggests that the sexual attributions used by Howells may not be a reliable guide to the actual range of variability within the sexes, because he excluded crania he could not be certain of. Thus, Howells' dataset probably exaggerates the differences between the sexes. The corollary of this is that the success rate of FORDISC in the analyses in which sex was specified may have been artificially high. Finally, Jantz & Ousley (2005) have suggested that secular change may affect FORDISC's ability to attribute ancestry. If this is indeed the case, then not only must a target specimen's source population be represented in the reference sample, but also the representatives of the source population in the reference sample must be from the same time period as the target specimen. Needless to say, this is likely to further reduce the chances that FORDISC will identify the ancestry of an unidentified specimen.

It appears, then, that FORDISC's utility is limited. Even in favourable circumstances it can be expected to classify no more than 1 per cent of specimens with confidence. One implication of this is that many of the ancestry determinations that have already been obtained using FORDISC are likely to be unreliable. Another is that there is a pressing need for bioarchaeologists and forensic anthropologists to develop more reliable methods for determining the ancestry of unidentified human remains. Recent work suggests that human cranial variation fits a model of African origin followed by repeated bottlenecking events as humans spread across the rest of the world (von Cramon-Taubadel & Lycett 2008). This implies that the similarities and differences in cranial shape among human populations are hierarchically structured. If this is the case, then distinguishing between shared-primitive and shared-derived similarities may improve our ability to determine the ancestry of unidentified human crania. While this and other possibilities are evaluated, care should be taken when interpreting FORDISC's output. In particular, an attribution should only be accepted if the PP and TP exceed 0.991 and 0.952, respectively.

Research was funded by SSHRC-CGS no. 766-2007-1077 and the Canada Research Chairs Programme. We are grateful to Alan Cross, Stephen Lycett, Yoel Rak, Mark Skinner, Bernard Wood and two anonymous reviewers for their advice.

Freid, D., Spradley, M. K., Jantz, R. L. & Ousley, S. D. 2005 The truth is out there: how NOT to use FORDISC. *Am. J. Phys. Anthropol.* **S40**, 103.

Harvati, K. & Weaver, T. D. 2006 Human cranial anatomy and the differential preservation of population history and climate signatures. *Anat. Rec.* **288A**, 1225–1233. (doi:10.1002/ar.a.20395)

Howells, W. W. 1989 *Skull shapes and the map*. Cambridge, MA: Harvard University Press.

Howells, W. W. 1995 *Who's who in skulls: ethnic identification of crania from measurements*. Cambridge, MA: Harvard University Press.

Howells, W. W. 1996 Howells' craniometric data on the internet. *Am. J. Phys. Anthropol.* **101**, 3. (doi:10.1002/ajpa.1331010302)

Hubbe, M. & Neves, W. A. 2007 On the misclassification of human crania. *Curr. Anthropol.* **48**, 285–288. (doi:10.1086/512985)

Jantz, R. L. & Ousley, S. D. 2005 *FORDISC*, version 3.0. Knoxville, TN: University of Tennessee.

Keita, S. O. Y. 2007 On Meroitic Nubian crania, FORDISC 2.0 and human biological history. *Curr. Anthropol.* **48**, 425–427. (doi:10.1086/507184)

Kosiba, S. 2000 Assessing the efficacy and pragmatism of 'race' designation in human skeletal identification: a test of FORDISC 2.0 program. *Am. J. Phys. Anthropol.* **S30**, 200.

Leathers, A., Edwards, J. & Armelagos, G. J. 2002 Assessment of classification of crania using FORDISC 2.0: Nubian X–group test. *Am. J. Phys. Anthropol.* **S34**, 99–100.

Naar, N. A., Hilgenberg, D. & Armelagos, G. J. 2006 FORDISC 2.0 the ultimate test: what is the truth? *Am. J. Phys. Anthropol.* **S42**, 136.

Ousley, S., Jantz, R. & Fried, D. 2009 Understanding race and human variation: why forensic anthropologists are good at identifying race. *Am. J. Phys. Anthropol.* **139**, 68–76. (doi:10.1002/ajpa.21006)

Ubelaker, D. H., Ross, A. H. & Graver, S. M. 2002 Application of forensic discriminant functions to a Spanish cranial sample. *For. Sci. Comm.* **4**, 1–5. Available at: http://www.fbi.gov/hq/lab/fsc/backissu/july2002/ubelaker1.htm

von Cramon-Taubadel, N. & Lycett, S. J. 2008 Human cranial variation fits iterative founder effect model with African origin. *Am. J. Phys. Anthropol.* **136**, 108–113. (doi:10.1002/ajpa.20775)

Williams, F. L., Belcher, R. L. & Armelagos, G. J. 2005 Forensic misclassification of Ancient Nubian crania: implications for assumptions about human variation. *Curr. Anthropol.* **46**, 340–346. (doi:10.1086/428792)