J Forensic Sci, September 2009, Vol. 54, No. 5
doi: 10.1111/j.1556-4029.2009.01118.x
Available online at: www.blackwell-synergy.com

Joseph T. Hefner,[1] *Ph.D.*

Cranial Nonmetric Variation and Estimating Ancestry*

ABSTRACT: Historically, when predicting the ancestry of human skeletal remains, forensic anthropologists have not fully considered the variation within human populations, but instead have relied on a typological, experience-based approach. Unfortunately, reliance on observer experience has produced a method that is as much an art as it is a science. This research focuses on the frequency distribution and inter-trait correlations of 11 common morphoscopic traits to demonstrate that the experience-based approach to ancestry prediction is indeed an art that is unscientific, because it is unreplicable, unreliable, and invalid. Ten of 11 traits examined had frequency distributions with significant differences ($p < 0.001$) between groups, but the range in variation of these traits far exceeds previous assumptions. Such within group variation clearly demonstrates that extreme trait expressions are not reliable for estimating ancestry through visual observation alone, but instead that these traits should be analyzed within a statistical framework.

KEYWORDS: forensic science, morphoscopic traits, forensic anthropology, nonmetric traits, ancestry prediction

An integral part of the biological profile constructed by forensic anthropologists for unknown human skeletal remains is a prediction of the peer-perceived ancestry of an individual. This prediction is usually accomplished through visual inspection of morphological variants of the cranium and mandible (i.e., nonmetric or morphoscopic traits) or through an analysis of measurements of the cranial and postcranial skeleton. However, predicting ancestry using nonmetric traits is not straightforward and often relies on years of experience and a remarkable understanding of human variation.

Interpreting nonmetric traits has traditionally involved qualifying a bone's shape (e.g., the nasal bones have a "Quonset hut" shape), a suture's course (e.g., the zygomaticomaxillary suture [ZS] is *S*-shaped), a feature's presence or absence (e.g., a postbregmatic depression [PBD] is present), or a feature's degree of expression (e.g., the anterior nasal spine [ANS] is pronounced). Ousley and Hefner (1) have noted a discrepancy between the traits employed by biological anthropologists using binary variables (e.g., pterygoalar bridging) of a single individual for biological distance studies and the variables of the skull (e.g., nasal bone structure) used by forensic anthropologists to predict ancestry. They suggested the term *morphoscopic* traits to describe the nonmetric traits used in a forensic context and having historical ties to E.A. Hooton (1887–1954). The traits established by Hooton in his laboratory are used today by the majority of forensic anthropologists (2,3), and his contributions to skeletal biology have had a major impact on the current philosophies and methodologies for ancestry prediction in a forensic context.

While Hooton did not use morphoscopic traits explicitly for the purpose of predicting ancestry, he recognized their utility for classification purposes. "They [morphoscopic traits] are capable of classification according to presence or absence, [and] grade of development and form, if the observer is experienced and is able to maintain a consistent standard for his morphological appraisals" (4). Clearly, Hooton recognized the need for standardizing observations among observers (4–6). This conviction was made even more clear, when he tested his own students and found such a low level of observer agreement that he was almost certain of the complete unreliability of these traits (6). Fortunately, he did not abandon morphoscopic traits completely, but instead sought to standardize the descriptions and illustrations of the traits to reduce some of the subjectivity and interobserver errors he had noted. Eventually, he developed the "Harvard List," a series of cranial nonmetric traits and observations for skeletal analysis. The illustrations Hooton created for the Harvard List (Fig. 1) were never published; however, along with the Harvard List, they have greatly influenced the traits used in predicting ancestry in a forensic context (2,3,7,8). The intent of the Harvard List was to reduce the subjectivity inherent in the process of trait observation, but the historical reliance within forensic anthropology on elements of the Harvard List without additional refinement has diminished the value of ancestry prediction using morphoscopic traits: a method still viewed by Rhine as much an art as it is a science (3). Hooton also noted that "we cannot be sure of the phenotypic composition of a population until the type combinations have been observed and tabulated in individuals as combinations and not as isolated traits" (4). Unfortunately, noone addressed the issue of subjectivity or standardization until recently. Walker (9) recently confirmed the usefulness of illustrations for reducing interobserver error when sexing the skull visually. He also clearly demonstrated that the results of such an analysis can be easily put into a statistical framework. In such a framework, it is the combinations of traits that work best, rather than single traits alone.

The lack of a methodological approach and, more importantly, the fact that there are no error rates associated with ancestry prediction using the morphoscopic method, suggest that they have not been investigated with appropriate scientific and legal considerations in mind. Minimizing subjectivity is certainly one of the goals of the scientific method (10). In light of the *Daubert* ruling (11), forensic anthropologists need to standardize and test the methods they apply to all aspects of skeletal analysis. Mastery of the "art"

[1]Statistical Research, Inc., Tucson, AZ 85712.
*This research was partially funded by a Forensic Sciences Foundation Acorn Grant in 2002.

Received 25 July 2008; and in revised form 23 Oct. 2008; accepted 25 Oct. 2008.

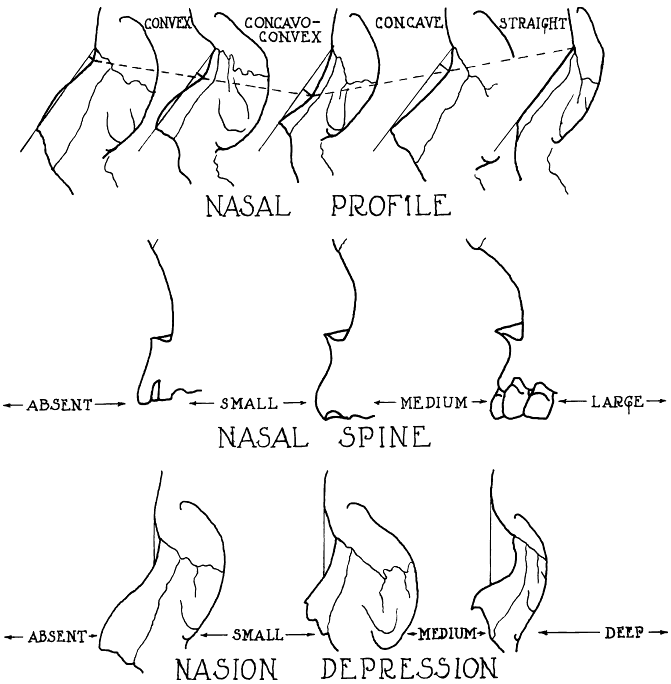

FIG. 1—*A representative example of Hooton's line drawings.*

of ancestry prediction is proportional to the experience of the analyst. Experienced observers are often able to predict ancestry correctly for the populations with which they regularly work, citing one trait or another in support of their assessment, but when pressed for an explanation, they often have trouble detailing their methods (12). This is the traditional approach in ancestry prediction. The emphasis is on experience, simplified trait scales, and extreme trait values rather than on patterns of trait distribution. When ambiguous or discordant trait values are encountered, admixture (3,8), or individual idiosyncrasy (3) is invoked without any consideration of the actual distribution of the traits in the reference population. Such conclusions are inevitable when using lists of traits supposedly representative of each ancestry as presented in textbooks and research articles (13,14). Hefner (15,16) and Hefner and Ousley (17) demonstrate that these traits cannot be viewed in such a simple manner. In fact, scoring only six traits (ANS, inferior nasal aperture [INA], nasal bone structure, nasal aperture width [NAW], PBD, transverse palatine suture [TPS]; see below for a description of these traits) in a large, diverse sample, Hefner and Ousley (17) noted that the percentage of individuals with all of the expected trait values based on race ranged from only 17% to 58%. Their results, and the findings of Hefner et al. (12), highlight the key paradox associated with the use of morphoscopic traits: forensic anthropologists claim they can assess ancestry to a high degree of accuracy using these traits, but the actual trait frequencies of these traits are much lower than assumed. So, if forensic anthropologists can predict ancestry accurately using a visual approach, what traits are they using? Hefner and Ousley (17) have suggested that forensic anthropologists estimate ancestry using the cranial *Gestalt* (the overall morphological impression) and, after a positive identification, choose traits *post hoc* that support their subjective evaluation. Thus, the traits will appear to be objective and valid indicators of ancestry although the process is subjective. In such a process, the forensic anthropologist will not discover combinations of traits that are more valid in estimating ancestry, and the science cannot progress. This subjective method of analysis has repercussions for the witness stand and in forensic reports (18).

Expert witness testimony based on subjective analyzes can be dismissed today in light of recent interpretations of the *Daubert* ruling (11). The reliability of a method must now be demonstrated with scientific findings rather than some level of certainty based on experience (19). Predicting ancestry using morphoscopic traits does not currently address the guidelines established by *Daubert*, in part because these traits have not been established as reliable or valid, they have not been subjected to appreciable peer review, and they have no known error rates. Although the use of morphoscopic traits in predicting ancestry have been accepted by the forensic anthropological community, the method relies too heavily on experience (12).

The experience-based method of ancestry prediction using morphoscopic traits indeed is an art: an art that is intuitive, untestable, unempirical, and consequently unscientific. The purpose of this paper is to address the issues Hooton noted over 60 years ago, and to refine existing standards for morphoscopic trait analysis and explore the distribution of nonmetric traits in a large sample of modern human crania.

Materials and Methods

Samples

To explore trait variation among groups, 11 common morphoscopic traits (Table 1) were collected for 747 individuals. Table 2 presents the sample sizes by group. Hefner (15) previously found no significant sex differences in morphoscopic traits, so males and females were pooled within groups for analysis. Following the typical forensic model, populations are grouped according to geographic ancestry and a pooled, four-group model is used for all subsequent analyses.

African Sample—The African sample consists of native Africans and American Blacks. The native African sample comprises individuals from East and West Africa ($n = 15$ and $n = 17$, respectively) collected during the Smith African Expedition (1909) and purchased by the Smithsonian Institution through the efforts of Frederick Muller & Co., Amsterdam, Holland in 1910. They are currently housed at the National Museum of Natural History (NMNH), Smithsonian Institution, in Washington, DC. The American Black sample is derived from the Robert J. Terry Collection

TABLE 1—*Morphoscopic traits used in this study.*

Morphoscopic Trait	References
Anterior nasal spine	Gill (34); Hefner (15); Rhine (3)
Inferior nasal aperture	Gill (34); Hefner (15); Krogman and Iscan (35); Rhine (3)
Interorbital breadth	Bass (36); Gill (34); Gill and Rhine (37); Hefner (2); Rhine (3)
Malar tubercle	Hauser and De Stefano (21); Hefner (15); Rhine (3)
Nasal aperture width	Bass (36); Hefner (15); Rhine (3); Stewart (38);
Nasal bone contour	Brues (2); Gill (34); Hefner (15)
Nasal overgrowth	Hefner (15); Rhine (3)
Postbregmatic depression	Bass (36); Hefner (15); Krogman and Iscan (35); Rhine (3)
Supranasal suture	Hauser and De Stefano (21); Hefner (15)
Transverse palatine suture	Gill (34); Hauser and De Stefano (21); Hefner (15); Rhine (3)
Zygomaticomaxillary suture	Gill (34); Hauser and De Stefano (21); Hefner (15); Rhine (3)

TABLE 2—*Materials used in this study (male and female samples).*

Sample	Female (*n*)	Male (*n*)	Age
Native North Americans (prehistoric and protohistoric)			
Arikara	18	24	1550–1700
Hawikuh	16	24	1000–200 BP
Doyon Eskimo	24	15	1600+
Pastolik Eskimo	8	4	1600+
Pueblo Bonito	4	3	1000–200 BP
Santa Barbara	27	30	1000–200 BP
Almeda	17	9	1000–200 BP
Perico Island	10	7	1000–200 BP
Canaveral	6	13	1000–200 BP
St. Lawrence Eskimo	4	5	1800+
Terry Collection			
American Whites	89	81	19th Cent.
American Blacks	61	89	19th Cent.
W.M. Bass Collection			
American Blacks	4	34	Contemporary
Japanese	5	10	Contemporary
Dutch	3	4	Contemporary
German	4	4	Contemporary
Chinese	6	53	Contemporary
African			
East	8	7	Contemporary
West	7	10	Contemporary
Total	321	426	

(*n* = 150) and the William M. Bass Donated Skeletal Collection (*n* = 38).

Asian Sample—The Asian sample consists of individuals from Japan (*n* = 15), China (*n* = 59), and North America housed at the NMNH. The Japanese specimens are from the Tokyo prefecture and were all donated by Tokyo Imperial University to the Smithsonian Institution. The Chinese crania were obtained by Ales Hrdlička from a cannery cemetery in Karluk, Alaska. They represent Chinese individuals who worked in the salmon canneries in the late 1800s.

European Sample—The European sample consists of native Europeans and American Whites. The European sample (*n* = 15) represents individuals purchased by the NMNH from anatomical houses or other museum trades. The American White sample is derived from the Robert J. Terry Collection (*n* = 170).

Native American—The native American sample consists of the following groups (all housed at the NMNH): Arikara, SD (*n* = 42), Hawikuh, NM (*n* = 40), Doyon Eskimo (*n* = 39), Pastolik Eskimo (*n* = 12), Pueblo Bonito, NM (*n* = 7), Santa Barbara Island, CA (*n* = 57), Almeda, CA (*n* = 26), Perico Island, FL (*n* = 17), Cape Canaveral, FL (*n* = 19), and St. Lawrence Eskimo (*n* = 9). These groups are pre- and protohistoric, ranging in chronological age from 1000 years before present to the early 19th century.

Morphoscopic Traits

Data were collected by the author using a data entry program (20) designed for the collection of morphoscopic trait data (Fig. 2) and available from the author. The definitions for each trait are modified from several sources (2,3,13,14,21). Each trait definition is presented along with individual character states and line drawings. Previous definitions and line drawings did not encompass the range of variability in skeletal populations or were considered uninformative. These new definitions were used for this analysis and are recommended for future studies.

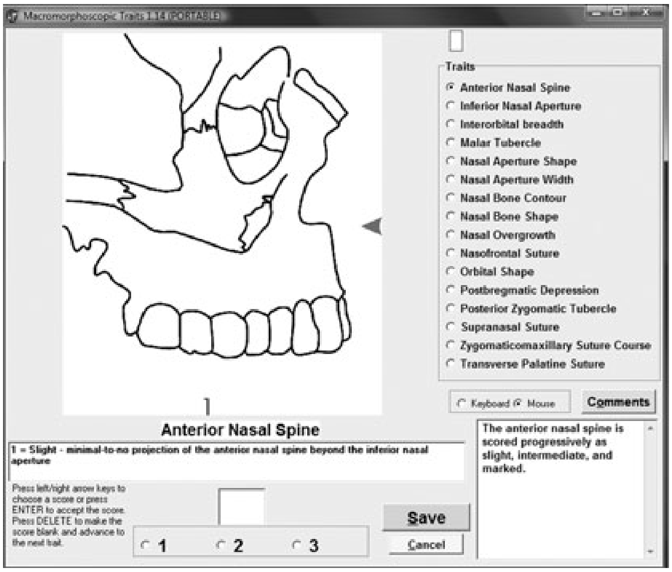

FIG. 2—*Screen-capture of the computer program Macromorphoscopics.*

ANS—One commonly encountered problem when assessing an ANS is the extreme fragility of this area, which is often damaged either peri- or postmortem. Those crania exhibiting trauma, pathology (including alveolar resorption), or postmortem damage to the overall inferior nasal margin were excluded from the analysis. The ANS (Fig. 3) is scored progressively as slight, intermediate, and marked: (i) Slight: minimal-to-no projection of the ANS beyond the INA, (ii) intermediate: a moderate projection of the ANS beyond the INA, and (iii) marked: a pronounced projection of the ANA beyond the INA.

INA—Inferior nasal morphology is defined as the most inferior portion of the nasal aperture, which, when combined with the lateral alae, constitutes the transition from nasal floor to the vertical portion of the maxillae, superior to the anterior dentition. INA is an assessment of the shape of the inferior border of the nasal aperture. Bilateral asymmetry was noted. In those instances where bilateral asymmetry did occur, the left side was used. INA (Fig. 4) is scored as follows: (i) an inferior sloping of the nasal floor which begins within the nasal cavity and terminates on the vertical surface of the maxilla, producing a smooth transition. The morphology is distinct from INA 2 regarding the more posterior origin and the greater slope of INA 1; (ii) sloping of the nasal aperture beginning more anteriorly than in INA 1, and with more angulation at the exit of the nasal opening; (iii) the transition from nasal floor to the vertical maxilla is not sloping, nor is there an intervening projection, or sill. Generally, this morphology is a right angle, although a more blunted form may be observed; (iv) any

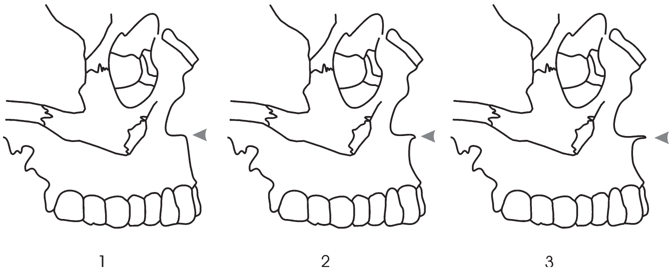

FIG. 3—*Character states for the anterior nasal spine morphology.*

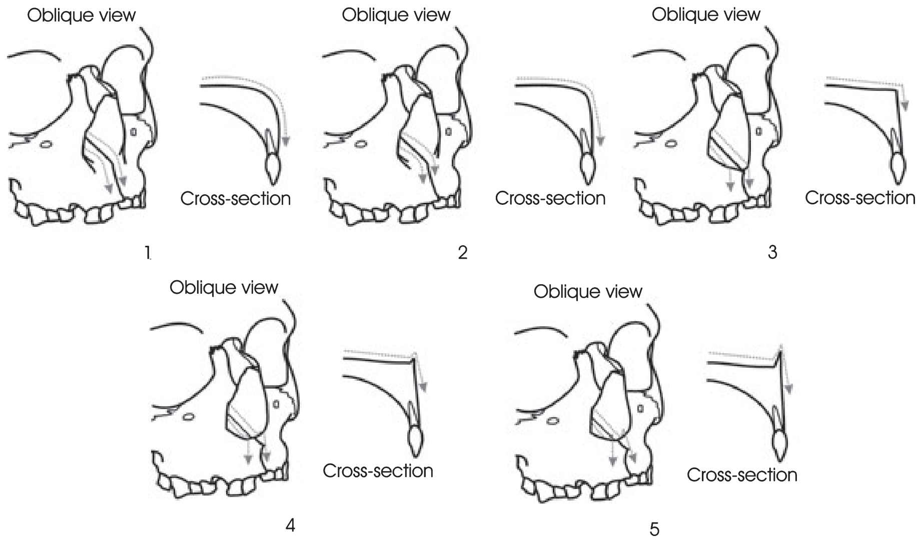

FIG. 4—*Character states for the inferior nasal aperture morphology.*

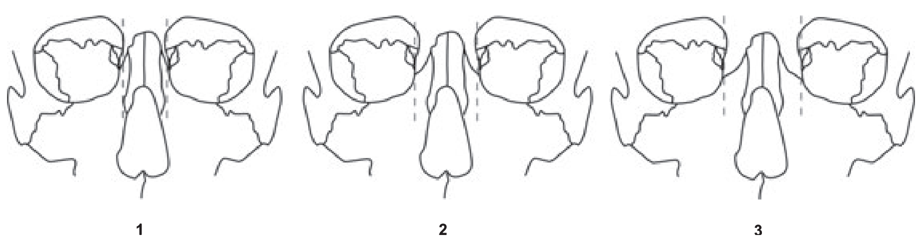

FIG. 5—*Character states for the interorbital breadth.*

superior incline of the anterior nasal floor, creating a weak (but present) vertical ridge of bone that traverses the inferior nasal border (partial nasal sill); and (v) a pronounced ridge (nasal sill) obstructing the nasal floor-to-maxilla transition.

IOB—Interorbital breadth (IOB) is a morphoscopic trait that could be measured with calipers using the defined measurement dacryon to dacryon (22,23), rather than scored nonmetrically. For this study, IOB (Fig. 5) is assessed as: (i) narrow, (ii) intermediate, and (iii) broad. This assessment is made relative to the facial skeleton.

MT—The malar tubercle (MT) (Fig. 6) is a caudally protruding tubercle located on the inferior margin of the maxilla and zygomatic bone in the region of the ZS. MT is scored following Hauser and De Stefano (21), who recommend placing a transparent ruler at the intersection of the ZS and the inferior margin of the malar to the deepest point on the curvature of the maxilla. An assessment is then made on the extent of protrusion beyond the ruler's edge. In instances where the suture is directly on the tubercle, the ruler is placed from the deepest curvature of the maxilla to the deepest anterior curvature on the zygomatic. It should be noted that a MT may be present on the maxilla, the zygomatic, or along the ZS. Observers should not consider the tubercles on the lateral zygomatic arch. A completely absent MT is rare. MT is scored as follows: 0—no projection of bone; 1—a trace tubercle below the ruler's edge (roughly 2 mm or less); 2—a medium protrusion below the ruler's edge (roughly 2–4 mm); 3—a pronounced tubercle below the ruler's edge (roughly 4 mm or more).

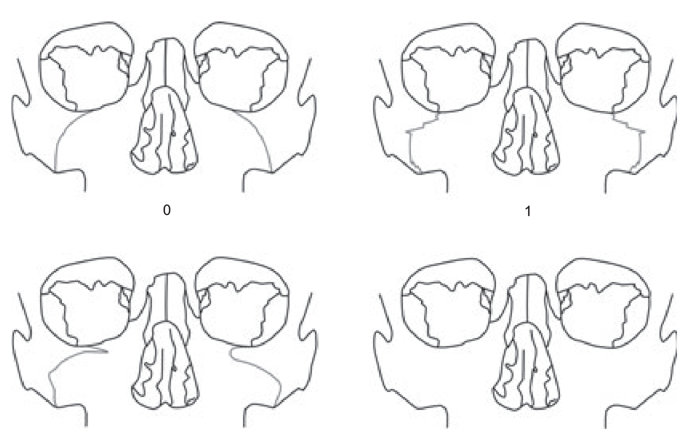

FIG. 6—*Character states for the malar tubercle.*

NAW—The width of the nasal aperture width (NAW) (Fig. 7) is assessed relative to the facial skeleton. It is scored as 1—narrow; 2—medium; or 3—broad.

NBC—Nasal bone contour (NBC) (Fig. 8) is defined as the contour of the midfacial region (particularly the contour of the nasal bones and the frontal process of the maxilla) *c.* 1 cm below nasion. Visual interpretation of nasal contour is not the most effective manner of analysis due to high inter- and intraobserver error. The use of a contour gage permits a more reliable and consistent assessment

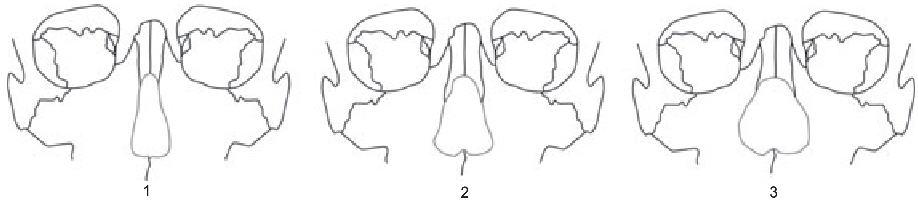

FIG. 7—*Character states for the nasal aperture width.*

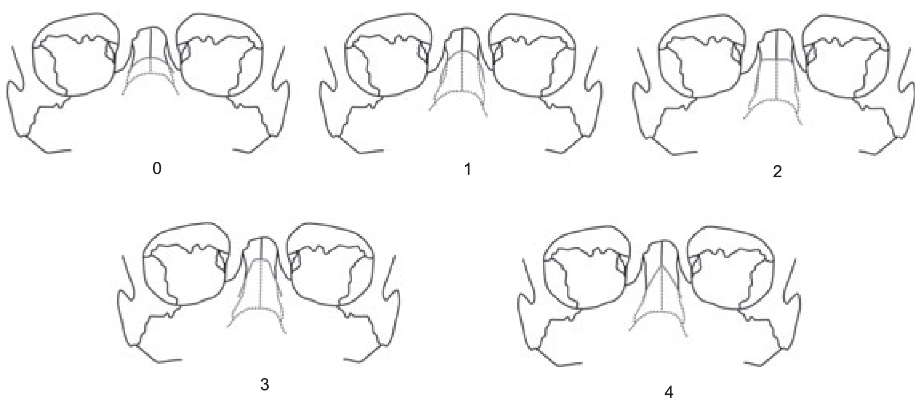

FIG. 8—*Character states for the nasal bone contour.*

of nasal contour (Fig. 9). To assess NBC, the cranium is placed in a position that allows the observer to gently, but with consistent and balanced pressure, place the contour gage directly on the nasal bones *c.* 1 cm inferior to nasion, while maintaining the gage roughly perpendicular to the palate and parallel to the orbits. NBC is scored as follows: 0—low and rounded NBC; 1—an oval contour, with elongated, high, and rounded lateral walls; NBC 1 presents a circular shape and lacks steep walls. Brues (2) suggests the term Quonset-hut to describe this shape, although the term is somewhat dated; 2—steep lateral walls and a broad (roughly 7 mm or more), flat superior surface "plateau," noted on the contour gage as a flat cluster of needles in the midline; 3—steep-sided lateral walls and a narrow superior surface "plateau"; 4—triangular cross-section, lacking a superior surface "plateau."

NO—Nasal overgrowth (NO) (Fig. 10) is defined as an inferior projection of the lateral border of the nasal bones beyond the maxillae at nasale inferious. Assessment of NO does not include anterior bulging of the nasal bones. Observations should be made on the left side. If the left side is damaged, the right side may be substituted. If both nasal bones are missing or fractured (anti-, peri-, or postmortem), the trait is not scored. It is often useful to run a finger along the borders of the maxilla and nasal bones near nasale inferious to determine whether a projection is present. NO is scored dichotomously as 0—absent or 1—present.

PBD—Postbregmatic depression (Fig. 11) is a slight to broad depression along the sagittal suture, posterior to bregma that is not the result of pathology (e.g., premature synostosis). Observed in lateral profile, the trait is scored as either 0—absent (no depression) or 1—present.

SPS—In adult crania, a secondary complex suture may persist, which is generally referred to as the supranasal suture (SPS), or sutura supranasalis (Fig. 12). This suture does not represent the nasal

FIG. 9—*A typical contour gage.*

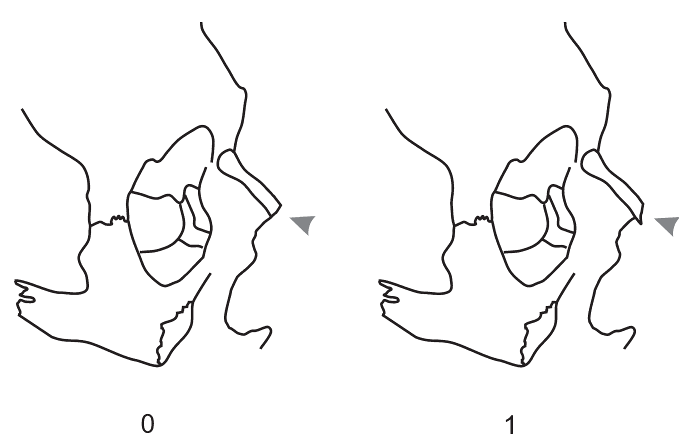

FIG. 10—*Character states for the nasal overgrowth.*

portion of a persistent metopic suture, which is usually a single, non-oscillating line. The SPS is the fusion of the nasal portion of a frontal suture that appears as a complex of interlocking bony spicules at glabella. SPS is scored as follows: 0—completely obliterated; 1—open (unfused); 2—closed, but visible.

TPS Shape—The course of the TPS (Fig. 13) is highly variable, although certain themes persist. TPS is not scored unilaterally, although asymmetrical sutures are not uncommon. The entire suture is observed, but the medial one-half in the region of the palatine suture is most closely scrutinized. When an asymmetrical suture is present (the two branches of the suture do not come into contact at midline) the general theme is recorded (e.g., straight or jagged). Slight undulations of the suture should not be considered when making a determination. If the suture is obliterated, it is not scored. TPS is scored as follows: (i) the suture crosses the palate perpendicular to the median palatine suture, with no significant anterior or posterior deviations. If the right and left halves of the suture do not contact each other at midline, but the suture is otherwise straight, score the suture as a 1; (ii) the suture crosses the palate perpendicular to the median palatine suture, but near this juncture a significant anterior deviation, or bulging, is present. If the right and left halves of the suture do not contact each other, but the suture is otherwise bulging anteriorly, a score of 2 is used; (iii) the suture crosses the palate, but deviates anteriorly and posteriorly (e.g., *M*-shaped) in the region of the median palatine suture. If the right and left halves of the suture do not contact each other, but the suture is otherwise jagged, a score of 3 is used; and (iv) the suture crosses the palate perpendicular to the median palatine suture, but near this juncture a posterior deviation, or bulging, is present.

ZS Shape—The ZS (Fig. 14) is the suture between the maxilla and the zygomatic. The course of the suture is best observed in the anterior view. In instances of asymmetrical manifestations, the left side is preferred. The infraorbital suture should be ignored when making a determination. Assessment of ZS is based primarily on the approximate location of greatest lateral projection of the suture, and also on the number of major angles present. ZS is scored as follows: 0—A suture with no angles and greatest lateral projection at the inferior margin of the malar. Sutures having greatest lateral projection at the inferior margin, but a slight angle near the midpoint of the suture should be scored as 0; 1—a suture with one angle and greatest lateral projection near the midline; 2—a suture with two or more angles (presenting a jagged and/or *S*-shaped appearance) and a variable position for greatest lateral projection. The figure shows both *S*-shaped and jagged courses of the suture.

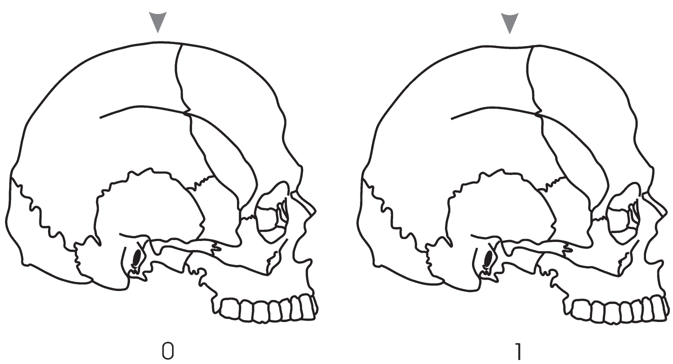

FIG. 11—*Character states for the postbregmatic depression.*

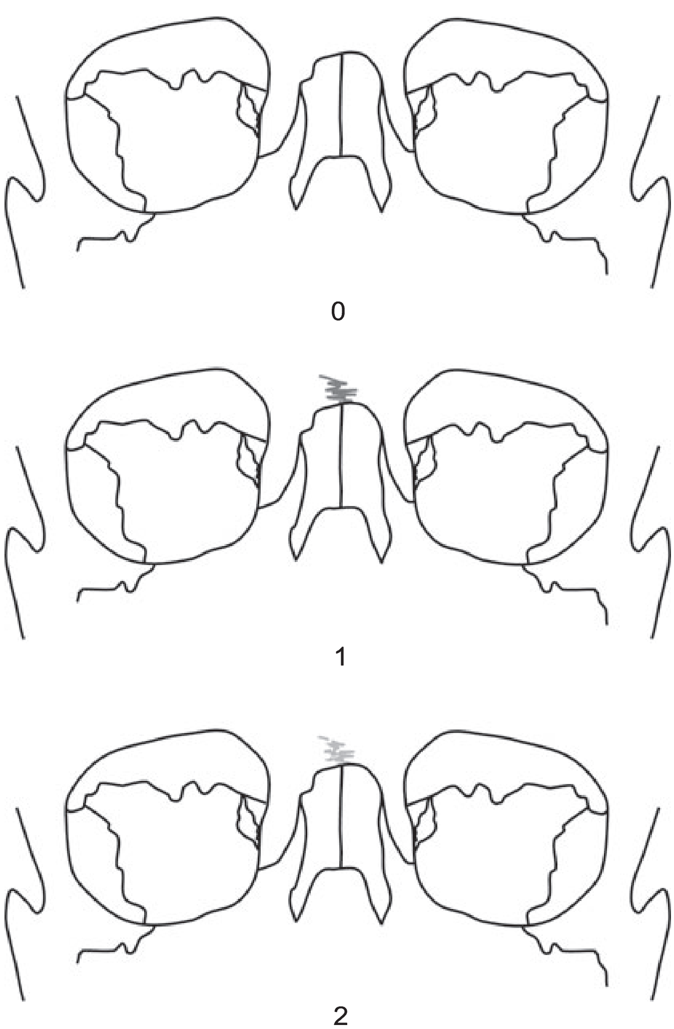

FIG. 12—*Character states for the supranasal suture.*

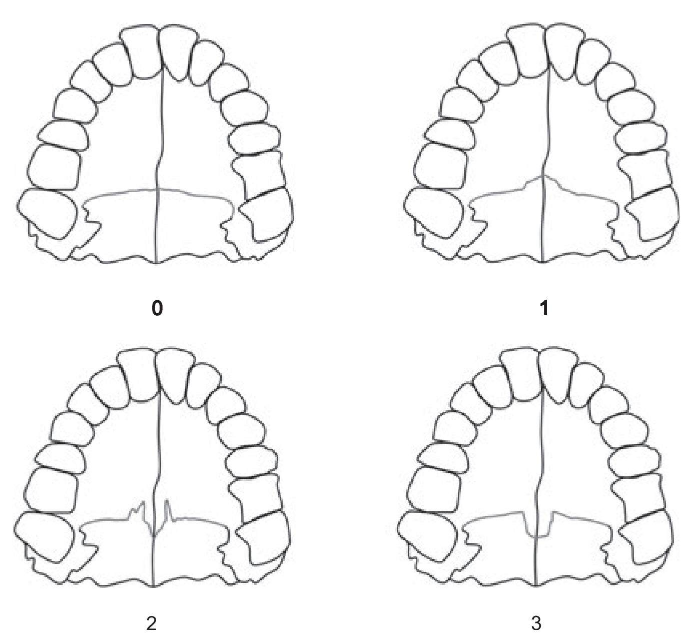

FIG. 13—*Character states for the transverse palatine suture.*

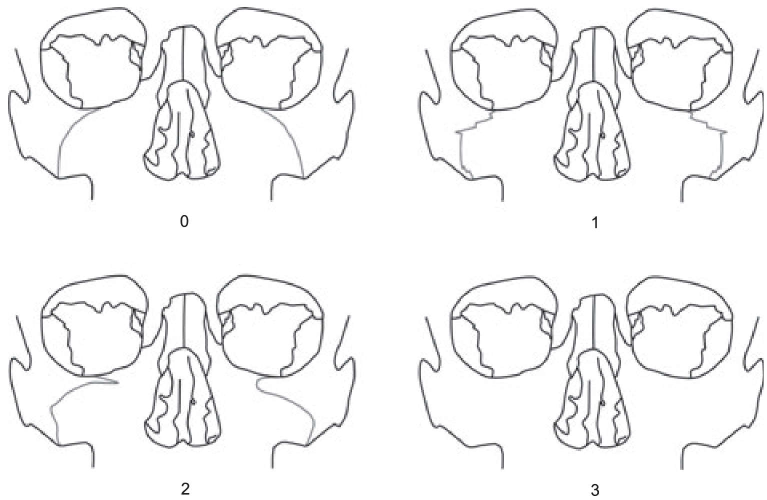

FIG. 14—*Character states for the shape of the zygomaticomaxillary suture.*

Statistical Methods

Frequency Distributions—Frequency distributions were calculated using SYSTAT 9.0 (24). Two-way cross-tabulation tables were created and Fisher's exact test based on the chi-squared statistic was used to determine whether significant differences were observed between groups.

Correlations—The polychoric correlation coefficient was used to measure the association of ordinal variables (e.g., 0, 1, 2, 3) and is appropriate because variable correlations are not affected when the latent continuous variable underlying the trait is compressed into a categorical response (25). Tetrachoric correlation coefficients are the recommended measures of association for binary (0, 1) variables. *LISREL* 7 (26) was used to estimate both the polychoric and the tetrachoric correlation coefficients.

Inter- and Intraobserver Error—The rate of interobserver error is particularly high when scoring morphoscopic traits without standard scales (27). To assess inter- and intraobserver variation, five individuals scored seven crania using the trait definitions outlined above. The author scored the same crania on two occasions separated by a 2-week period to assess intraobserver error. Two separate but related statistics were used to measure observer agreement. The statistical tests used for continuous, quantitative data usually comprise a standard ANOVA mixed model to measure differences in the mean response among observers. However, these methods are not appropriate when the dataset is composed of categorical, qualitative traits. Intraobserver variability was assessed using Cohen's kappa statistic (28). Cohen's kappa measures the agreement between two observations, while taking into account any agreement that would occur by chance. Cohen's kappa is calculated as:

$$k = \Pr(\mathrm{a}) - \Pr(\mathrm{e})/1 - \Pr(\mathrm{e})$$

where Pr(a) is the relative observed agreement among raters and Pr(e) is the probability that the agreement was due to chance. If the observations are in perfect agreement, a kappa value of 1 is expected.

Interobserver variability cannot be measured using Cohen's kappa, because it is only appropriate for two observers. Fortunately, a similar (and related) measure exists for multiple observers. Fleiss' kappa (29) works for any number of observers and is appropriate for categorical, qualitative traits. Like Cohen's kappa, Fleiss' kappa is a measure of observer agreement expressed as a value between 0 and 1. Fleiss' kappa can be expressed as:

$$k = \bar{P} - \bar{P}_{\mathrm{e}}/(1 - \bar{P}_{\mathrm{e}})$$

where $1 - \bar{P}_{\mathrm{e}}$ is the agreement that can be attained above chance and $\bar{P} - \bar{P}_{\mathrm{e}}$ is the observed agreement. Thus, Fleiss' kappa is roughly equivalent to the ratio of observed versus expected values. The interpretation of both Cohen's and Fleiss' kappa has been debated (30,31). For the current study, the table of Landis and Koch's (32) significance values for the individual kappa score was used.

Results

Frequency Distributions

Frequency distributions are presented in Tables 3–13. Significant differences between groups were noted at the $p < 0.001$ level with the exception of the MT, which did not demonstrate significant differences ($\chi^2 = 13.158$; d.f. = 12; $p = 0.358$). The range in variation of the remaining traits among and between groups far exceeds previous assumptions (2,3,13,14). Again, Hefner (15,16) and Hefner and Ousley (17) found that the percentage of each group presenting all of the expected trait values ranged from only 17% to 51%, percentages that drop even lower as more traits are added to the analysis. In the current analysis, no single individual had all 11 expected trait values. These frequency distributions suggest that the compiled trait lists for ancestry ignore a substantial amount of variation within groups. Additionally, concluding that an individual is of "mixed" ancestry due to discordant traits would only be tenable if all of these groups have been "mixed" for some time—groups which include native Africans, precontact native Americans, and Europeans—because these groups include many individuals with traits from different ancestral groups. If the published trait lists are indeed indicative of ancestry, then we can correctly conclude that most every decedent can be interpreted as having "mixed" ancestry.

Correlations

The correlation analysis examined the relationship among all 11 morphoscopic traits. The results are presented in Table 14 along with significance values. The results of the correlation analysis

TABLE 3—*Anterior nasal spine (ANS) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
ANS	*n*	%	*n*	%	*n*	%	*n*	%
1	152	69.7	178	67.9	60	80	53	36.3
2	44	20.2	56	21.4	10	13.3	38	26
3	22	10.1	28	10.7	5	6.7	55	37.7

TABLE 4—*Inferior nasal aperture (INA) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 283)		Asian (*N* = 75)		European (*N* = 146)	
INA	*n*	%	*n*	%	*n*	%	*n*	%
1	64	29.4	10	3.8	9	12	1	0.7
2	63	28.9	63	24	13	17.3	5	3.4
3	47	21.6	149	56.9	48	64	35	24
4	29	13.3	39	14.9	3	4	60	41.1
5	15	6.9	1	0.4	2	2.7	45	30.8

TABLE 5—*Interorbital breadth (IOB) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
IOB	*n*	%	*n*	%	*n*	%	*n*	%
1	21	9.6	155	59.2	31	41.3	45	30.8
2	75	34.4	96	36.6	39	62	92	63
3	122	56	11	4.2	5	6.7	9	6.2

TABLE 6—*Malar tubercle (MT) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
MT	*n*	%	*n*	%	*n*	%	*n*	%
0	110	50.5	107	40.8	32	42.7	75	51.4
1	60	27.5	98	37.4	25	33.3	47	32.2
2	32	14.7	40	15.3	10	13.3	18	12.3
3	16	7.3	17	6.5	8	10.7	6	4.1

TABLE 7—*Nasal aperture width (NAW) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
NAW	*n*	%	*n*	%	*n*	%	*n*	%
1	8	3.7	22	8.4	2	2.7	79	54.1
2	89	40.8	204	77.9	65	86.7	48	32.9
3	121	55.5	36	13.7	8	10.7	19	13.1

TABLE 8—*Nasal bone structure (NBS) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
NBS	*n*	%	*n*	%	*n*	%	*n*	%
0	114	52.3	30	11.5	19	25.3	11	7.5
1	50	22.9	67	25.6	17	22.7	23	15.8
2	22	10.1	65	24.8	29	38.7	27	18.5
3	23	10.6	90	34.4	9	12	37	25.3
4	9	4.1	10	3.8	1	1.3	48	32.9

TABLE 9—*Nasal overgrowth (NO) frequencies in four ancestral groups.*

	African (*N* = 207)		American-Indian (*N* = 220)		Asian (*N* = 75)		European (*N* = 146)	
NO	*n*	%	*n*	%	*n*	%	*n*	%
0	141	68.1	97	44.1	51	68	77	52.7
1	66	31.9	123	55.9	24	32	69	49.2

TABLE 10—*Postbregmatic depression (PBD) frequencies in four ancestral groups.*

	African (*N* = 218)		American-Indian (*N* = 253)		Asian (*N* = 72)		European (*N* = 184)	
PBD	*n*	%	*n*	%	*n*	%	*n*	%
0	115	52.8	235	92.9	65	90.3	121	82.9
1	103	47.2	18	7.1	7	9.7	25	17.1

TABLE 11—*Supranasal suture (SPS) frequencies in four ancestral groups.*

	African (*N* = 215)		American-Indian (*N* = 262)		Asian (*N* = 75)		European (*N* = 146)	
SPS	*n*	%	*n*	%	*n*	%	*n*	%
0	92	42.8	90	34.3	9	12	57	39
1	67	31.2	82	31.3	23	30.7	57	39
2	56	26	90	34.4	43	57.3	32	22

TABLE 12—*Transverse palatine suture (TP) frequencies in four ancestral groups.*

	African (*N* = 180)		American-Indian (*N* = 260)		Asian (*N* = 75)		European (*N* = 145)	
TP	*n*	%	*n*	%	*n*	%	*n*	%
0	33	18.3	165	63.5	34	45.3	42	29
1	85	47.2	72	27.7	25	33.3	40	27.6
2	45	25	14	5.4	11	14.7	49	33.8
3	17	9.4	9	3.5	5	6.7	14	9.7

indicate a moderate and significant correlation among the morphoscopic traits, with the exception of the MT, NO, and the ZS, which are not significantly correlated with any other traits. Most of the midfacial structures were moderately-to-strongly correlated. This analysis indicates that the implicit assumption of independence among morphoscopic traits is incorrect. In metric analyses, correlations among variables generally improve classification rates (33). Statistical models assuming independence among the variables

TABLE 13—*Zygomaticomaxillary suture (ZS) frequencies in four ancestral groups.*

	African (*N* = 177)		American-Indian (*N* = 242)		Asian (*N* = 75)		European (*N* = 135)	
ZS	*n*	%	*n*	%	*n*	%	*n*	%
0	9	5.1	7	2.9	4	5.3	2	1.5
1	56	31.6	92	38	21	28	50	37
2	88	49.7	127	52.5	38	50.7	57	42.2
3	24	13.6	16	6.6	12	16	26	19.3

should be carefully considered. These results also indicate that multivariate models would provide better classification rates.

TABLE 15—*Inter- and intraobserver error analysis.*

Trait	Interobserver Error Fleiss' *k*	Intraobserver Error Cohen's *k*
Anterior nasal spine	0.506*	0.422*
Transverse palatine suture	0.700*	1.000*
Supranasal suture	0.650*	0.468*
Postbregmatic depression	0.232	0.820*
Nasal overgrowth	1.000*	1.000*
Nasal bone contour	0.231	0.810*
Nasal aperture width	0.732*	0.929*
Malar tubercle	0.470*	0.929*
Interorbital breadth	0.325	0.857*
Zygomaticomaxillary suture	0.541*	0.857*
Inferior nasal aperture	0.376	0.964*
Mean	0.524*	0.837*

*Moderately significant or higher following Landis and Koch (28).

Inter- and Intraobserver Error

Table 15 presents the results of the inter- and intraobserver error analysis. Each of the Cohen's kappa values were moderately significant for the intraobserver test, suggesting very small levels of intraobserver variation. Only two traits (ANS and SPS) were significantly lower than the others (ANS $k = 0.422$; SPNS $k = 0.468$). Assessing ANS morphology relies on a measure of the length of the ANS relative to the facial skeleton. Several explanations are possible for the low level of agreement for this trait, including error introduced when recording the score, difficulty assessing the length of a small structure relative to the face, or inadequacies in the morphological variants included in the Morphoscopic program. The SPS may be difficult to assess reliably, in part because the observer has to determine whether the suture is obliterated, closed, but visible, etc. The difficulty inherent in this trait may suggest removing it from further analyses, at least until it has been more systematically defined and illustrated.

The interobserver analysis indicates some variance among observers. One observer (JTH) had previously used the Morphoscopic program, so unfamiliarity with the application may have introduced some of the error. However, with the standard definitions and illustrations provided in the Morphoscopic program, this may not have been the issue.

Four traits were only moderately agreed on among observers, including PBD ($k = 0.232$), NBC ($k = 0.231$), INA ($k = 0.376$), and IOB ($k = 0.325$), indicating some difficulty in trait assessment. The discordance among observers for two of these traits is unexpected. PBD is scored on a binary scale (presence/absence), which is normally thought to reduce observer error. However, PBD has a continuous underlying threshold, so slight expressions may be misinterpreted by the novice and the expert alike. The low level of agreement among observers for NBC is also surprising. The Morphoscopic program guides the observer on the use of a contour gage to assess NBC, unlike the traditional approach, which is not clearly defined and has no method for standard assessment. The use of a contour gage to assess NBC is a relatively recent phenomenon (15), so there may be a learning curve associated with the use of this tool. Intraobserver variation for NBC was low ($k = 0.810$). Conceivably, as the use of a contour gage becomes more widely accepted, observer error will decrease for the NBC. The interobserver error analysis indicates that the morphoscopic traits can be scored with high reliability and replicability.

Classification

Based on these and other results, the classification of an individual to an ancestry group using morphoscopic traits in a statistical framework shows great promise (1,17). Several statistical methods that work well for morphoscopic traits, including logistic regression, naïve Bayesian, and *k*-Nearest Neighbor, produce classification rates ranging between 84% and 93%, depending on the combination of traits used and the statistical method applied to the classification. While an in-depth discussion of these classification methods is beyond the scope of the current research, several points are worthy of mention. First, these methods produce high classification accuracies and low error rates, similar to those obtained using metric data (33). Second, these statistical methods indicate that combinations of traits work better than individual ones. Until recently, however, proper methods of variable selection did not

TABLE 14—*Correlation coefficient of 11 morphoscopic traits.*

	ANS	INA	IOB	MT	NAW	NBC	NO	PBD	SPS	TPS	ZS
ANS	–										
INA	0.423*	–									
IOB	−0.092	−0.237*	–								
MT	0.035	−0.041	0.012	–							
NAW	−0.360*	−0.521*	0.448*	−0.036	–						
NBC	0.359*	0.467*	−0.399*	0.010	−0.589*	–					
NO	−0.007	0.061	−0.182	0.019	−0.103	0.183	–				
PBD	0.052	−0.093	0.263*	−0.030	0.170	−0.149	−0.164	–			
SPS	−0.210*	−0.063	0.023	−0.039	0.073	−0.144	0.046	0.006	–		
TPS	0.081	−0.033	0.283*	−0.028	0.016	−0.025	−0.101	0.084	−0.021	–	
ZS	−0.016	0.014	0.021	−0.034	−0.094	0.050	0.021	0.055	0.081	0.020	–

ANS, anterior nasal spine; INA, inferior nasal aperture; IOB, interorbital breadth; MT, malar tubercle; NAW, nasal aperture width; NBC, nasal bone contour; NO, nasal overgrowth; PBD, postbregmatic depression; SPS, supranasal suture; TPS, transverse palatine suture; ZS, zygomaticomaxillary suture.

*Significant at the $p < 0.05$ level or below.

exist for this type of data, so refinement and trait selection was not feasible. As a final point, these statistical methods confirm the importance of trait inter-correlations, because some combinations of traits work better than others. Future research will draw on several statistical approaches for classification using morphoscopic traits.

Conclusions

This research has empirically investigated the "typical" morphoscopic traits thought to be indicative of African-, Amerindian-, Asian-, and European-derived groups and has shown that they are not found at the frequencies suggested by earlier studies (2,3,13,14). Extreme trait expressions are not reliable for estimating ancestry in a valid manner. Because the distribution of these traits does not conform to traditional assumptions (i.e., extreme trait values linked to specific races), it seems likely that if forensic anthropologists can accurately assess ancestry using visual methods, they do so based on the cranial *gestalt* or by employing *post hoc* trait selection after positive identification.

The frequently cited apparent evidence of admixture can be explained by a lack of knowledge of the variation within groups. Every individual in the total sample size of 747 showed at least one trait thought to be typical of a different continental group. When the actual distribution of these traits are taken as a whole, however, the discordance of multiple traits should come as no surprise and should not be treated as evidence of admixture or hybridity. On the contrary, a combination of supposedly ancestrally diagnostic traits in many individuals shows the fallacy of the typological approach to ancestry prediction and reveals variation in morphoscopic traits within ancestral groups. The variation observed within groups does not mean that these traits are not useful; morphoscopic traits show great promise in a statistical framework, particularly for discriminating between American Blacks and Whites (1,17,18).

The correlation analysis in this study demonstrated a much higher degree of trait association than previously assumed, particularly for traits relatively close to one another on the cranium. Until further research has been carried out on the correlations among morphoscopic traits, any assumptions of trait independence should be carefully considered.

Finally, the inter- and intraobserver analysis in this study suggest that the level of reliability for trait scores is moderate-to-high with only a few, somewhat surprising, exceptions, when using standard definitions, and drawings of individual character states. Even the exceptions, however, had significantly lower interobserver error rates than those observed by Hefner et al. (12). Standardization of trait definitions and further refinement of the morphoscopic method will undoubtedly help to reduce interobserver error in the future.

The morphoscopic approach to ancestry determination using traits pioneered by Hooton has been tested within a statistical framework to assess the validity and reliability needed to address the *Daubert* challenge. The problem of subjectivity in scoring traits (reliability) can be largely solved by using standard drawings, following Hooton (5) and Walker (9). The optimal weighing of the traits seen in an individual to produce the best prediction of ancestry (validity) can be accomplished through statistical methods and reference group trait distributions. Only when these are used together can nonmetric traits can be reliable, replicable, and valid indicators of ancestry.

Acknowledgments

I would like to thank Stephen D. Ousley for his critical eye and tireless support. I would also like to thank Michael Warren for all of his help with various aspects of this research. The following individuals granted permission to study crania in their collections: David Hunt, National Museum of Natural History, Washington, DC; Lee Meadows Jantz, University of Tennessee, Knoxville, TN; Dennis C. Dirkmaat, Mercyhurst College, Erie, PA. Further editorial comments were provided by Natalie M. Uhl, M. Kate Spradley, Nicholas V. Passalacqua, and two anonymous reviewers.

References

1. Ousley SD, Hefner JT. The statistical determination of ancestry. Proceedings of the 57th Annual Meeting of the American Academy of Forensic Sciences; 2005 Feb 21–26; New Orleans, LA. Colorado Springs, CO: American Academy of Forensic Sciences, 2005.
2. Brues AM. The once and future diagnosis of race. In: Gill G, Rhine S, editors. Skeletal attribution of race: methods for forensic anthropology. Maxwell Museum of Anthropological Papers No. 4. Albuquerque, NM: University of New Mexico, 1990;1–7.
3. Rhine S. Nonmetric skull racing. In: Gill G, Rhine S, editors. Skeletal attribution of race: methods for forensic anthropology. Maxwell Museum of Anthropological Papers No. 4. Albuquerque, NM: University of New Mexico, 1990;9–20.
4. Hooton EA. Up from the ape, 2nd ed. New York: The Macmillan Co., 1946.
5. Hooton EA. Lecture notes of EA Hooton. Cambridge, MA: Peabody Museum, 1926.
6. Hooton EA. The test. In: The papers of Earnest A. Hooton, 1930–1950. Accession Number: 995-1 Box# 20. Cambridge, MA: Peabody Museum.
7. Birkby W, Fenton T, Anderson B. Identifying southwest Hispanics using nonmetric traits and the cultural profile. J Forensic Sci 2008;53(1):29–33.
8. Caldwell L, Tersigni MT, Manhein M. Racial admixture: a test of nonmetric ancestry estimation. Proceedings of the 60th Annual Meeting of the American Academy of Forensic Sciences; 2008 Feb 18–22; Washington, DC. Colorado Springs, CO: American Academy of Forensic Sciences, 2008.
9. Walker PL. Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthrop 2008;136:39–50.
10. Rao CR. Statistics and truth: putting chance to work, 2nd ed. London: World Scientific, 1989.
11. United States Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 1993.
12. Hefner JT, Emanovsky PD, Byrd J, Ousley SD. The value of experience, education, and methods in ancestry prediction. Proceedings of the 59th Annual Meeting of the American Academy of Forensic Sciences; 2007 Feb 19–23; San Antonio, TX. Colorado Springs, CO: American Academy of Forensic Sciences, 2007.
13. Burns K. Forensic anthropology training manual. New Jersey: Prentice-Hall, 1999.
14. Byers SN. Introduction to forensic anthropology: a textbook. Boston, MA: Allyn and Bacon, 2001.
15. Hefner JT. Assessing nonmetric cranial traits currently used in the forensic determination of ancestry. Master's Thesis, Gainesville, FL: The University of Florida, 2003.
16. Hefner JT. The statistical determination of ancestry using cranial nonmetric traits. Doctoral Dissertation, Gainesville, FL: The University of Florida, 2007.
17. Hefner JT, Ousley SD. Morphoscopic traits and the statistical determination of ancestry II. Proceedings of the 58th Annual Meeting of the American Academy of Forensic Sciences; 2006 Feb 20–25; Seattle, WA. Colorado Springs, CO: American Academy of Forensic Sciences, 2006.
18. Hefner JT, Ousley SD, Warren M. An historical perspective on nonmetric skeletal variation: Hooton and the Harvard list. Proceedings of the 56th Annual Meeting of the American Academy of Forensic Sciences; 2004 Feb 16–21; Dallas, TX. Colorado Springs, CO: American Academy of Forensic Sciences, 2004.
19. Christensen A. The impact of Daubert: implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). J Forensic Sci 2004;49:1–4.
20. Hefner JT, Ousley SD. Morphoscopics [computer program]. Beta version. Tucson, AZ: Hefner and Ousley, 2005.
21. Hauser G, De Stefano GF. Epigenetic variants of the human skull. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung, 1989.
22. Howells WW. Cranial variation in man: a study by multivariate analysis of patterns of difference among recent human populations. Papers of the

Peabody Museum of Archaeology and Ethnology, Harvard University, No. 67. Cambridge, MA: The Museum, 1973.
23. Howells WW. Skull shapes and the map: craniometric analyses in the dispersion of modern homo. Papers of the Peabody Museum of Archaeology and Ethnology, Harvard University, No. 79. Cambridge, MA: The Museum, 1989.
24. SYSTAT 9.0. [computer program]. Windows standard version. Richmond, CA: SPSS, Inc., 1998.
25. Coenders G, Saris WE. Categorization and measurement quality. The choice between Pearson and polychoric correlations. In: Saris WE, Münnich Á, editors. The multitrait–multimethod approach to evaluate measurement instruments. Budapest: Eötvös University Press, 1995;71–87.
26. Joreskorg KG, Sorbom D. LISREL 7: a guide to the program and applications. Chicago, IL: SPSS Publications, 1989.
27. Hooton EA. Some notes on observer variation. In: The papers of Earnest A. Hooton, 1930–1950. Accession Number: 995-1 Box# 20. Cambridge, MA: Peabody Museum.
28. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
29. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378–82.
30. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.
31. Feinstein AR, Cicchetti DV. High agreement but low kappa: the problems of two paradoxes. J Clin Epidemiol 1990;43:543–8.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
33. Jantz RL, Ousley SD. FORDISC 3: computerized forensic discriminant functions. Version 3.0. Knoxville, TN: The University of Tennessee-Knoxville, 2005.
34. Gill G. Craniofacial criteria in the skeletal attribution of race. In: Reichs K, editor. Forensic osteology. Springfield, IL: Charles C. Thomas 1998;293–317.
35. Krogman WM, Iscan MY. The human skeleton in forensic medicine. Springfield, IL: Charles C. Thomas, 1986.
36. Bass W. Human osteology: a laboratory and field manual. Missouri Archaeological Society, No. 2. 1987.
37. Gill GW, Rhine S. Skeletal attribution of race: methods for forensic anthropology. Albuquerque, NM: Maxwell Museum of Anthropology, 1990.
38. Stewart TD. Essentials of forensic anthropology. Springfield, IL: Charles C. Thomas, 1979.

Additional information and reprint requests:
Joseph T. Hefner, Ph.D.
Statistical Research Inc.
6099 E. Speedway Blvd.
Tucson
AZ 85712
E-mail: jhefner@sricrm.com