Article

The Fall and Rise of Identified Reference Collection: It Is Possible and Necessary to Transition from a Typological Conceptualization of Variation to Effective Utilization of Collections

John Albanese [1,2,3,*], Abegail Dagdag [1], Carmen Skalic [1], Stephanie Osley [1] and Hugo Cardoso [2,3]

1 Department of Integrative Biology, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada; dagdag@uwindsor.ca (A.D.); carmenskalic@gmail.com (C.S.); sosley@gmail.com (S.O.)
2 Department of Archaeology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; hugo_cardoso@sfu.ca
3 Centre for Forensic Research, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
* Correspondence: albanese@uwindsor.ca; Tel.: +(519)-253-3000 (ext. 3973); Fax: +(519)-971-3621

Abstract: In some jurisdictions, race, ancestry or population affinity are part of the biological profile used in preliminary identification, for historical and political reasons. It is long overdue for forensic anthropologists to abandon this typological approach to human variation, regardless of the terms used. Using a sample (n = 105) selected from the Terry and Coimbra identified reference collections, a blind experimental approach is used to test several metric methods and versions of methods for group estimation (Fordisc 3.0 and 3.1, and AncesTrees), that rely on different statistical approaches (discriminant function analysis and random forest algorithms, respectively) derived from different reference samples (Howells' data in AncesTrees and Fordisc 3.1, and different forensic subsamples in Fordisc 3.0 and 3.1). The accuracy for matching premortem documented group designation is consistently low (36 to 50%) across testing parameters and consistent with other independent tests. The results clearly show that a change in terminology, software updates, alternative statistics, expanded reference samples, and newer collections will not solve the underlying fundamental problems. It is possible and necessary to transition from a typological conceptualization of variation to the effective utilization of identified reference collections in Forensic Anthropology. In addition to the theoretical and methodological reasons, it is unethical for forensic anthropologists to continue to use on the deceased methods that do not work and that serve only to further exclude and marginalize the living.

Keywords: forensic anthropology; race; ancestry; population affinity; typology; identified reference collection; theory; methods

1. Introduction and Background

Through an analysis of skeletal remains, a forensic anthropologist may be asked to construct a biological profile that can assist with the preliminary identification of an unknown individual. Age at death, sex, and stature form the foundation for this profile, as this information is usually available on most government issued identification documents. Therefore, there are premortem records for comparison. In some jurisdictions that have historically had a disproportionate influence on Forensic Anthropology, race or ancestry is also considered as part of personal identification and assessed for preliminary identification [1]. The perceived need for information regarding race or ancestry is jurisdiction-specific and based on historical bias rather than real patterns of variation. For example, in South Africa during the Apartheid era, racial designation was ascribed at birth by an agent of the government and appeared on official government issued identification documents. In contrast to South Africa, in the United States, racial designation is not necessarily included in official government issued identification documents. However, it is considered as part of the

identity of an individual or group. In addition, whether it is self-reported or designated by a third party, the information is considered as part of identification (see [2] on how race is categorized in missing persons database in the USA). This wide-spread racialism in the societies of the United States and South Africa has had and continues to have a disproportionate impact on Forensic Anthropology, and the construction and use of identified reference collections (IRCs) in those countries and in the rest of the world (see [3,4] for a counterpoint to racial and Eurocentric approaches).

The development and testing of all methods for constructing a biological profile have been possible due to the existence of IRCs. In the 21st century, the term IRCs is applicable to traditional osteological collections derived from anatomical or cemetery sources, collections of CT and MRI scans, and the Forensic Anthropology data bank (FDB) and similar databases (see [5] for a systematic review of different types of IRCs and their use in forensic research). Forensic Anthropology exists due to the fact that these real-bone and virtual collections are well-documented [5]. An uncritical approach to the use of IRCs has resulted in a series of sex, age at death, stature, and group estimation methods of limited utility in forensic contexts. These problems are possibly best seen in various methods for sex and stature estimation that have been developed in the last two decades, where researchers select a bone and a collection and create a "population-specific" metric method (see [5] for a critique of this approach). These problems stem from the method in which IRCs have been used in Forensic Anthropology [5]. First, the nature and impact of bias on the skeletal variation sampled *and* on documentary data that are retained and available for research are rarely considered. Second, IRCs have typically been treated as biological or statistical populations, rather than highly biased samples. Third, a typological approach that does not consider the range of human variation through time has been pursued.

This typological approach rooted in IRCs has affected the construction of a biological profile in two ways: First, group-specific methods for estimating sex, stature, and age at death; and second, group estimation methods. Theoretical frameworks and methodological approaches for dealing with typology and bias in IRCs have been described in detail elsewhere and are not re-stated here (see [5–7] for additional details). These alternative approaches for the use of IRCs to deal with the first issue have proven very successful when applied to critiques of group-specific methods and to create other alternative approaches for constructing a biological profile that can be reliably applied in a wide range of forensic contexts (for example, [8,9] on sex estimation; [10–12] on stature; and [13,14] on age). Decoupling each question from all of the other questions and certainly from group-specificity (race, ancestry, nationality, etc.) is a more robust approach to constructing a biological profile [10]. In this paper, the focus is on the second issue: The estimation of group membership. With very few exceptions (for example, [1,3,4,15]), the estimation of group membership, regardless of the term used, has not been an area of critical assessment from *within* Forensic Anthropology.

The assessment of group membership remains as one of the most contentious parts of the biological profile [16]. This persistent and unsubstantiated view is summarized by İşcan and Steyn [17] (p. 195):

Although almost all biological anthropologists agree that distinct human races do not exist and that it is impossible to classify humans into discrete race groups based on their skeletal features, most will agree that some form of geographical patterning exists. This provides some potential of providing tentative information on biological origin. However, there is currently no consensus as to how we should go about it, what the correct terminology would be or even if this type of assessment should be included at all in a forensic skeletal report.

This contentiousness is surprising considering that all of the genetic and osteological evidence is unambiguous and clearly shows that the race concept (sometimes referred to as ancestry or population affinity) is problematic at best and highly damaging at worst (for example, [1,15,18–50]). For decades, scientifically rigorous research has consistently shown that: (1) Within group genotypic and phenotypic variation is greater than between

group variation; (2) about 6% of genotypic and phenotypic variation is attributable to race or continental origin; (3) there is no concordance of human genotypic and phenotypic variation with group membership. Race or continental origin are not coarse or useful groupings for the investigation of human variation, but rather these groupings maintain existing power relations in society [22]. Despite the clear evidence, Forensic Anthropology remained resistant to these substantive critiques of the race concept until the end of the 20th century. The critiques were dismissed as coming from outside of the discipline by researchers who do not understand Forensic Anthropology (for example, [2,51], and more recently [52]).

In the 1990s, forensic anthropologists started to respond to these critiques. However, in a superficial manner with only a change in terminology from race-based to ancestry-based or continental origin terminology [1]. Terms, such as "Negroid" that had originally been adopted in *Physical* Anthropology to be scientific, had come to be perceived as problematic by the late 20th and early 21st centuries in *Biological* Anthropology [5]. Therefore, terms such as "Negroid", "Caucasoid", and "Mongoloid" became "African", "European", and "Asian", since the former terms were considered outdated and not because the typological concepts had been critically assessed. Ancestry, when not used as an euphemism for race, can be useful in studying human variation. The classic example is a well-established link between a series of hemoglobin variants, not only a sickle cell, to malarial environments on at least three continents, not only West Africa (see [43] for systematic review). In an early systematic critique from *within* Forensic Anthropology, Albanese and Saunders noted that this change in terminology without a critical assessment of the highly problematic typological framework solves nothing and is actually detrimental [1]. Changing only the terms further obfuscates the typology used in most of the field of Forensic Anthropology, which is rooted in racial concepts of human variation used to justify slavery and colonialism (see [20] for a greater historical context of terminology in Biological Anthropology; see [53] for a perspective from outside of Biological Anthropology).

More recently, the terms are currently revised again and rather than estimating "race" or "ancestry", forensic anthropologists should now be estimating "population affinity" or only "affinity" (for example, [54,55]). These terms are not new, nor is the critique of them. Albanese and Saunders started their critique by stating in their first sentence, "The determination of ancestry, population affinities or ethnicity (in the past, referred to as 'race') . . . " [1] (p. 282). Furthermore, despite the claims of using "evolutionary theory", these typological approaches do not consider the arbitrary nature of group construction, and the plasticity in human phenotype resulting in a complex fluid range of human skeletal variation through time (see [20] for a critique of the historical and continued misuse of "evolutionary theory" to maintain power relations in and through Biological Anthropology). The arguments posed by Ross and Pilloud [54] in critiquing the construction of the "Hispanic" group are equally applicable to all of the racialized groups, regardless of the terms used. Additionally, as noted above, decades of research have demonstrated that none of those groupings are biologically meaningful. Another change in terms does not address any of the underlying problems, since regardless of the terms used—race, ancestry, population affinity, ethnicity, etc. —the approach used in most of the field of Forensic Anthropology involves only a change in terms, and has been and continues to be typological.

These typological views of human variation were entrenched in and reinforced by the IRCs that have been used to develop and test methods [5,7,23], beginning with the establishment of some of the most important IRCs in the USA (Terry and Hamann-Todd collections) and South Africa (Dart collection) in the 20th century [1,56]. Despite the enormous significance of the IRCs to Forensic Anthropology, very little attention has been paid to how the amassing and use of IRCs have been shaped by a typological view of variation, while at the same time the collections have reinforced the typological approach [1,5,7]. The typological approach in Forensic Anthropology assumes that variation sampled from a given collection

exists outside of all contexts, and that it is inherently and statistically representative of a given "population" with arbitrary social and political parameters for membership.

Forensic Anthropology has been shaped by views in the greater society in which forensic anthropologists live, while at the same time has reinforced those racialist and nationalist views. This typological approach to human variation manifests as racialism and/or nationalism in different regions for historical reasons [5]. In jurisdictions with a long history of racism, including the USA and South Africa, racialism is reflected in the theory and methods used in Forensic Anthropology, and is directed internally (see [30] for an analogous forensic example involving DNA databanks in the USA). In contrast, nationalism is manifested in nation-specific methods developed in forensic anthropology and is targeted externally. Nationalism is clearly seen in methods from Europe, Southeast Asia, South America, and increasingly in the USA to deal with the influx of migrants through their southern border. The return to terms, such as "population affinity" is a direct response to the failure of racialist concepts applied to the nonsensical "Hispanic" grouping when dealing with the horrific number of deaths of migrants entering the USA from its southern border. Using terms, such as "population affinity" allows for an *attempt* to distinguish between "Hispanics" born in the USA and Europe from "Hispanics" born in Mexico, Central America, and South America. In this context, Forensic Anthropology is currently used as an extension of immigration policy in the USA.

Despite the described importance of methods and theory related to race and ancestry (for example, [52,57]), the rigorous and systematic testing of methods for estimating group membership using large, independent samples have been lacking in Forensic Anthropology (see [58] for an early exception). In some cases where larger samples were used, the goal of these tests was to successfully illustrate the problems with the race concept as it is used by some biological anthropologists, including forensic anthropologists, rather than to test the forensic utility of methods (for example, [59]). In that case, an archaeological Nubian sample was used to test Fordisc. Despite the relative biocultural and temporal homogeneity and known geographic origin of the Nubian test sample, skulls were assigned to various continental origins. Unfortunately, these results have been ignored or discounted by forensic anthropologists as not forensically relevant (for example, [17]), despite the results demonstrating fundamental flaws with Fordisc as well as the problems with discriminant function analysis and the failure of post hoc probabilities to flag problematic assessments (see "Materials and Methods" section below).

Most of the research assessing the utility of methods in forensic contexts for assigning an unknown to a group have focused on the small-scale targeted testing of some methods, and more recently on highly biased historical studies of case files. Superficially, these assessments seem to suggest that some methods can provide information that may be useful in a forensic investigation. However, the historical studies are highly problematic due to the type of cases that are included in the assessment, i.e., the denominator used in the calculation of the allocation percentage (for example, [60,61]. Parsons [60] makes every effort to caution the reader about the results, and Thomas et al. clearly describe the limitations of their methodology for including cases in their analysis: " . . . although this may inflate the overall accuracy rate (sic) . . . " [61] (p. 972). However, in both cases, they still proceed to state an accuracy of over 90% for matching premortem group designations, when the actual accuracy is certainly significantly lower (see [62] for additional details).

In other cases, the tests of methods using small samples are demonstrations on how to use specific software, where the researcher has a priori knowledge of the unknown, and are not true tests of utility in a forensic context (for example, [17,63]). In other cases, larger samples are used to test the forensic utility of methods. However, the results are inflated by applying a priori knowledge of the unknown to pursue an approach that will provide better results (for example, [64–66]). Using both craniometric and geometric morphometric analyses, Duzik's [65] results of matching premortem records exceeded 90% in targeted testing using a priori information to frame the question as "Hispanic versus Japanese", but dropped to 56–57% when the analysis was not biased by knowledge of the test sample.

Similarly, Fernandes et al. [64] note how AncesTrees provides slightly higher allocation accuracies using a test sample from Brazil. One approach to applying the method using three ancestry groups works slightly better at allocating people of "European origin". Additionally, another approach of applying the method using six or two ancestry groups performs slightly better at allocating people of "African origin." In a circular fashion, the recommended best practices for estimating group membership require a priori knowledge of the group membership of the unknown (for example, [17,64–68]). A more realistic assessment suggests that AncesTress will provide an allocation to a group that matches premortem records 49–53% of the time [64].

A blind test of methods using an independent sample is the only way to have a realistic assessment of the utility of methods in a forensic context [3,4]. The test must be done without any a priori knowledge to assess how many times the method provides information that could be useful in a forensic investigation and how many times can the method be expected to provide wrong information that is detrimental to an investigation. Moreover, a blind test of methods using independent samples has clearly demonstrated the failure of metric and morphoscopic methods to provide information regarding group membership which would be useful in a forensic investigation, with a failure of the probability scores to flag erroneous cases [3,4]. These problematic results when using independent samples are not new (see [58] for a specific example, and [1] for systematic overview). "Fixes" to problems with group estimation methods have been proposed for decades and have included calls for changes in terminology, software updates, alternative statistics, and expanding samples or "updating" samples and IRCs (for example, [17,54,64,69,70]).

Using an independent sample drawn from the Terry and Coimbra collections, a blind experimental approach is used to challenge the typological assumptions of group estimation methods. First, two completely different methods are tested, specifically Fordisc and AncesTrees, that rely on different statistical approaches: Discriminant function analysis with post hoc probability calculations, and random forest algorithms, respectively. Second, several versions of methods are tested, specifically, Fordisc 3.0 and 3.1 to assess the impact of software and statistical updates. Third, the impact of different reference samples is assessed through a comparison of AncesTrees using Howells' data, Fordisc 3.1 using Howells' data, and two different sub-samples of the FDB in Fordisc 3.0 and 3.1. The goal of this paper is not a test of any specific method or to propose tweaks to fundamentally flawed concepts. Rather, building on the research of Albanese and Saunders [1], Plens et al. [3,4], and DiGangi and Bethard [15], the goal of this paper is to demonstrate that it is possible and necessary to transition from a typological conceptualization of variation to an effective utilization of IRCs in forensic contexts.

2. Materials and Methods

2.1. Fordisc and Discriminant Function Analysis

Since the 1960s, various group estimation methods have been developed for application in forensic contexts using different skeletal elements (see [1,17] for overviews of methods). Building on a key publication by Giles and Elliot [71], discriminant function analysis has been the most used statistical approach when allocating an unknown to a group using metric data [72]. Several methods have emerged that use specialized software for group estimation. One of the first and most widely available applications is Fordisc, which uses discriminant function analysis to estimate both race and sex at the same time, using a suite of standard skeletal measurements and different reference samples [73].

Fordisc is an updated and automated version of the fundamental aspects of Giles and Elliot's original method with several major additions that are intended to correct the known problems with the poor performance of this method [74]. First, a sub-sample from the FDB which includes data from the Terry collection [70,73] or Howells' [75] data can be used as the reference sample to calculate the discriminant functions used to allocate an unknown. Sample sizes by sex-race groups are provided and the general sources of FDB data that are

used in Fordisc are stated in the manual [73]. However, the exact numbers of individuals from each original source are not provided.

The second major difference with Fordisc, when compared with many other discriminant function methods, is that it calculates case-specific probability scores post hoc [74]. These scores, known as posterior and typicality probability, are intended to correct major well-known shortcomings of discriminant function analysis. As with any predictive approach, the overall allocation accuracy for a method that uses discriminant function analysis does not necessarily provide any information regarding its certainty in any one specific case where the method is applied. Furthermore, the discriminant function score for any one case may allocate an unknown to a given group. However, the score alone does not include information to assess the likelihood of a correct allocation in that one case. Finally, discriminant function analysis will force an allocation into one of the selected groups even if the unknown is not a member of any of those groups or if the parameters for group construction are arbitrary and not biologically meaningful. The probability scores are intended to assess the level of certainty in an allocation of one unknown individual in one specific case. Posterior or typicality probability scores are between 0 and 1, and there is also a threshold that needs to be considered. Posterior and typicality probability scores of less than 0.05 indicated that no confidence should be placed on the allocation of the unknown individual into a given group, since the unknown is outside of 95% of the range of the reference sample. Posterior probability is an indicator of how the unknown fits into any of the selected groups and the posterior probability scores for possible membership in all of the possible groups sum up to 1. Typicality probability addresses the "none of the above" problem using discriminant function analysis, and scores of less than 0.05 indicate that the unknown is "not typical" of the groups selected.

The typicality probability is significantly more important than the posterior probability in assessing confidence in any given allocation. For example, if the groups "White females", "Black females", and "Hispanic females" are selected for the Fordisc analysis, but the unknown is documented as male, the discriminant function will force an allocation of the unknown into one of the female groups, since no male groups were included in the analysis. The posterior probability of membership in any of the female groups would similarly support the allocation of female. The posterior probability, with a total of 1, would be divided among the three selected groups and could be 0.89 for "White females", 0.09 for "Black females", and 0.02 for "Hispanic females". Therefore, while the discriminant function will force an allocation, the typicality probability of membership in any of those groups should be less than 0.05 for all the groups selected, indicating "none of the above" groups, since the unknown is documented as male (see [59]).

In the current study, Fordisc 3.0 and 3.1 [73] were tested using the FDB reference sample. There are two main differences between these Fordisc versions that are most relevant to the current test. First, there are differences in the original source and composition of reference samples from the FDB used in Fordisc. Additionally, there was a shift toward using less data from the Terry collection with version 3.1. Second, there are updates on how the typicality probabilities are calculated. Fordisc 3.1 should provide a better allocation accuracy due to the reference samples used to calculate the discriminant functions. Moreover, it should provide more certainty in any one allocation with what should be a more statistically robust approach to calculating three different typicality probabilities.

2.2. AncesTrees and Howells' Data

AncesTrees is an internet browser-based method that relies on random forest algorithm to classify an unknown into a group using craniometric data (see [64,76] for additional details). The random forest approach is probabilistic and is intended to solve the problems with discriminant function analysis without the need for probabilities to be calculated post hoc. In contrast to Fordisc, which gives the user the option of selecting FDB or Howells' data, AncesTrees only uses Howells' data as the reference sample. Moreover, in

contrast to Fordisc, AncesTrees only assesses group membership and not sex. For each ethnic-geographic-temporal reference sample, the sexes were combined into one group.

Howells [75] collected craniometric data from archaeological samples organized into 28 groups based on ethnic, temporal, and geographic criteria, including Ainu, Andaman Island, Anyang, Arikara, Atayla, Australia, Berg, Buriat, Bushman, Dogon, Easter Island, Egypt, Eskimo, Guam, Hainan, Mokapu, Moriori, Norse, North Japan, Peru, Philippines, Santa Cruz, South Japan, Tasmania, Teita, Tolai, Zalavar, and Zulu. His approach to sampling each source was not random, and crania that were considered as "typical" of each group were selected [75]. Howells did not include crania that were "morphologically unusual for the population as a whole" [75] (p. 89), and thus created artificially distinct homogenous groups that did not sample the actual range of variation in any given group, while also exaggerating the differences between the groups. The negative impact of this approach to selecting the reference sample should be greater for methods, such as Fordisc since the discriminant function analysis is inversely related to the degree of overlap among groups [69]. However, the random forest method used in AncesTrees is a machine learning approach that should not be susceptible to the same types of problem, given the construction of the reference samples [76]. AncesTrees should outperform Fordisc using Howells' data as the reference sample.

2.3. Test Samples

Documentation alone does not make the IRC a useful source of data. The research potential for IRC is directly related to the quality and accuracy of the documentary data available for each individual and for the collection as a whole [5]. Without cross-validated and rigorously verified documentary information, an IRC is only a collection with limited use for forensic research. The samples used to test all of the methods/versions of methods were selected from two of the best documented reference collections available for research: The Terry collection [56] and the Coimbra collection [7,77]. The anatomy-derived Terry collection is curated at the Smithsonian Institution, and as with all IRCs can be a biased source of data. However, in this study, the bias can be considered positive in favor of Fordisc, since a portion of the reference sample in Fordisc comes from the Terry collection. Fordisc should work very well at allocating individuals from the Terry collection. The second major source of data came from the cemetery-derived Coimbra collection (Colecção de Esqueletos Identificados), which is curated at the University of Coimbra in Coimbra, Portugal. Group membership was based on the documentation that is available for each individual from their respective collections. For the Terry collection, race is noted as "White" or "Negro" or "Black" in the original documents reviewed by the first author [5,56]. Based on the documentary data available for the Coimbra collection, all of the individuals in the test sample were born in Portugal and by definition are European [7]. The sample (*n* = 105) included 40 Black individuals and 65 White or European individuals (see Table 1 for additional details). For consistency, this paper follows the racial terms used in Fordisc when presenting the results.

Table 1. Test sample by collection, group, and sex. The group terminology in the table is consistent with the terminology used in Fordisc; for discussion about terms, please see the text. BF: Black female; WF: White female; BM: Black male; WM: White male.

Source/Group	BF	BM	WF	WM	Total
Terry Collection	20	20	8	17	65
Coimbra Collection			20	20	40
Total	20	20	28	37	105

For the purpose of this test, the age of the Terry and Coimbra collections, i.e., when the collections were amassed, is irrelevant for two completely separate reasons. First, the samples from the Terry and Coimbra collections are consistent with samples that have

been used to develop and test methods for group membership in Forensic Anthropology. Racialized samples and collections have been used to develop and test methods in ways that assume that the biological race exists outside of all contexts. Howells' archaeological, constructed subsamples are used as reference samples in Fordisc and AncestTrees. Even when the FDB option (rather than Howells' data) is selected in Fordisc, the "American Indians" reference sample is derived from 19th century sources. Additionally, as noted above, the FDB sample includes data from the Terry collection. Furthermore, Navega et al. [76] used data from the Valle Da Gafaria collection to test AncesTrees. This collection consists of over 150 people who were enslaved in Africa and died in Lagos, Portugal. Second, the forensic relevance and research potential of "older" collections is described in detail elsewhere [5,7]. When sampled in an appropriate manner to address specific questions, the Terry and Coimbra collections have been used to develop and test methods that have been proven to perform very well in a forensic context (for example, see [8,9] for sex estimation, and [10] for stature estimation).

2.4. Measurements and Analysis

It is widely reported that cranial data are the best source of information for estimating group membership [17], and thus the focus of this research is the cranium. AncesTrees or Fordisc provide almost no guidance for which and how many measurements should be used to assign an unknown to a group. Including too few variables in the analysis may not capture the size and shape of a cranium, while including too many variables is statistically problematic due to the correlation among predictor variables. In this research, we pursued a practical approach and selected nine cranial measurements that capture cranial size and shape, but that are relatively resistant to premortem changes, such as tooth loss, perimortem changes, such as trauma, taphonomic processes due to burial and recovery (Coimbra collection), and handling for research purposes (Terry collection). All of the nine measurements were used in the Fordisc analyses using the FDB data, and eight were used for the analyses involving Howells' data. Upper facial height was not collected by Howells. The measurements are listed and described in Table 2 [78]. Crania that were sectioned during autopsy or dissection were not included, since cranial height and other measurements taken across the cuts would be affected by the unknown thickness of the saw used for sectioning. The sample is a subset of the data collected by the first author to sample a wide range of human variation in *Homo sapiens* rather than any one collection (see [9] for details on sampling).

All of the analyses were carried out using a blind experimental approach. The first author anonymized the test sample of 105 individuals. First, all of the information regarding sex, collection, and group membership were removed from the spreadsheet with cranial data. Second, a random case number was assigned to each individual, in order that the case number did not reveal any information regarding the source IRC. Third, the order of the cases was randomized in order that the individuals were not clustered by source or other criteria. The second and third authors conducted the analysis using the FDB reference sample in Fordisc 3.0 and 3.1, and Howells' reference sample in Fordisc 3.1 and AncesTrees. The analysis was focused on one issue: Utility. In other words, a count of how often the methods tested provide information that would be useful in a forensic investigation, and how often the methods would provide information that would be detrimental to an investigation.

Table 2. Cranial measurements used in the analysis [78].

Maximum cranial length (GOL)	The distance of Glabella from opisthocranion in the mid-sagittal plane in a straight line; measured with spreading caliper
Maximum cranial breadth (XCB)	The maximum width of the skull perpendicular to the mid-sagittal plane wherever it is located with the exception of the inferior temporal line and the immediate area surrounding the latter; measured with spreading caliper
Bizygomatic breadth (ZYB)	The direct distance between both zygia located at their most lateral points of the zygomatic arches; measured with sliding caliper
Basion-bregma height (BBH)	The direct distance from the lowest point on the anterior margin of the foramen magnum, basion, to bregma; measured with spreading caliper
Cranial base length (BNL)	The direct distance from nasion to basion; measured with spreading caliper
Basion-prosthion length (BPL)	The direct distance from basion to prosthion; measured with spreading caliper or sliding caliper
Maxillo-alveolar breadth (MAB)	The maximum breadth across the alveolar borders of the maxilla measured on the lateral surfaces at the location of the second maxillary molars; measured with spreading caliper
Upper facial height * (UFHT)	The direct distance from nasion to prosthion; measured with sliding caliper
Minimum frontal breadth (WFB)	The direct distance between the two frontotemporale; measured with sliding caliper

* Not used in the analysis involving Howells' data.

3. Results

The results of the blind test of AncesTrees and various versions and options in Fordisc are presented in Table 3. A match with premortem records for group membership ranges from 36% using Fordisc 3.0 with Howells' reference sample to 50% for Fordisc 3.0 using the FDB reference sample. The results from the tests of Fordisc 3.0 and 3.1 using the FDB reference sample are further deconstructed in Table 3. Both versions of Fordisc force the user to select groups based on race and sex. Estimating only sex is not an option. As presented in Table 4, the results for sex and/or race would have been misleading in forensic investigation 60% of the time for Fordisc 3.0 and 61% of the time for Fordisc 3.1.

Table 3. Utility by trial using identified cases from the Terry and Coimbra collections (*n* = 105).

	Fordisc 3.0	Fordisc 3.1	Fordisc 3.1	AncesTrees
Reference Sample	FDB *	FDB *	Howells	Howells
Statistical Approach	Discriminant Function	Discriminant Function	Discriminant Function	Random Forest
Matches Documentation	50%	47%	36%	37%

* FDB: Forensic Anthropology databank.

Table 4. Accuracy by sex and race (*n* = 105) for Fordisc 3.0 and 3.1 using the FDB reference sample. Both versions of Fordisc do not allow for the separate estimation of sex and group membership. Linking sex to group, reduced the utility of sex estimation, and increased the number of cases with wrong information.

Correct Allocations	Race Only	Sex Only	Both Race *and* Sex
Fordisc 3.0	50%	73%	40%
Fordisc 3.1	47%	75%	39%

One of the important statistical updates with Fordisc 3.1 was the inclusion of typicality probabilities that are calculated in three different ways. Low typicality probabilities are

invaluable for assessing the level of certainty in any one allocation and flagging and excluding wrong assessments. As noted above, discriminant function analysis may force an allocation. However, typicality probabilities of less than 0.05 should flag the problematic cases. Therefore, the typicality scores for each of the 64 cases (61%), where Fordisc 3.1 did not match the documented sex and/or race, were reviewed. In 21 erroneous cases, at least one of the three typicality scores indicated that no confidence could be placed in the allocation. In other words, the typicality probability worked as intended in about one third of cases to flag allocations that did not match premortem records. However, in about two thirds (42 out of 64) of erroneous cases, the typicality probability indicated confidence in the assessment even though it did not match premortem records.

4. Discussion

Theoretical, methodological, and ethical problems persist with the continued use of a typological approach to investigating human variation using IRCs in Forensic Anthropology. These problems are most evident in the continued attempts at developing methods for assessing group membership, regardless of the terms used. Whether the people in the test sample are referred to as "Negro" (as apparent in some original documents), Black or of African origin or whether they are referred to as "Caucasian," White, European or of European origin, the groupings do not capture patterns of variation. Furthermore, the methods used to assign people to a group have consistently failed. Changing terms from race to ancestry to population affinity does not solve any of the problems. Additionally, it only serves to further obfuscate the underlying power relations, where the racialization of the deceased further contributes to the marginalization of the living [62].

Methodologically, the results from this study are consistent with blind independent forensically relevant tests of methods for assessing group membership, where accuracy is about 50%. Additionally, the probability scores fail to identify cases with erroneous assessments (for example, [3,4]). Moreover, the results presented are consistent with the test of AncesTrees by Fernandes and colleagues [64] when they assessed accuracy (49–53%) without a priori assumptions of group membership, which approximates a blind trial. Similarly for Fordisc, the results are consistent with other independent tests [59,66,79,80]. Furthermore, the figure of about 50% is consistent with the first major systematic test of Giles and Elliot's [71] foundational method that used discriminant function analysis to estimate group membership. Birkby [58] found that the allocation accuracy for matching the documented group was 52%.

The results presented in this paper are clear and unambiguous. With a systematic, experimental, and blind approach to testing, the methods and the versions of methods using various options performed poorly. Using data from the FDB rather than Howells' data did result in an improvement in the overall allocation accuracy for matching premortem records. Part of this improvement could possibly be attributed to a larger portion of the test sample that is selected from the Terry collection. However, the best accuracy for matching premortem records was still low at 50%. When holding the reference source constant using Howells' data, the results for AncesTrees (37%) were not substantively different from Fordisc 3.1 (36%). The statistical approach in AncesTrees does not correct any of the problems with discriminant function analysis or the sampling problems with Howells' data. Software updates and some changes in the reference samples used between Fordisc 3.0 (50%) and Fordisc 3.1 (47%) did not result in any meaningful difference in allocation accuracy. Overall, these results are consistent with Urbanová et al. [66] who tested the same sample with various methods, including Fordisc 3.0, and found that the highest accuracy for matching premortem records was about 50%.

Updates to the statistical approaches in Fordisc 3.1, including three different calculations of typicality probability, lead to incorrect results in all of the calculations in two thirds of the erroneous cases. These results are consistent with the Fordisc 2.0 tests when typicality was calculated only one way [59]. Additional statistics further buried the lack of typicality of an unknown, and thus contributed to the perceived certainty in an allocation when the

result did not match premortem records. When considering the typicality probabilities for race and sex, the Fordisc 3.1 results suggest confidence in an allocation in 80% of the cases (correct matches as well as wrong allocations that were indicated as certain by the typicality score). However, the results match premortem records for both race and sex in only 39% of the test cases. Different methods, different group terminologies, alternative statistics, modified reference samples, software updates, etc. do not provide better results, since the underlying issues are theoretical problems with conceptualizing human variation and the use of IRCs in most of the field of Forensic Anthropology.

The fundamental problems are theoretical on at least two levels. On one level, as noted in the "Introduction and Background" section of this paper, human variation does not cluster into racial categories, continental origin, ethnicity, population affinity or arbitrarily defined IRCs. Regardless of the terms used, the evidence is overwhelming: The groupings are based on social and political criteria for exclusion and marginalization, and not based on genotypic and phenotypic variation. On another level, a major shift in Forensic Anthropology is required to conceptualize patterns of human variation using IRCs. In contrast to the previously used approaches, IRCs must be considered as highly biased samples and not as statistical or biological populations [5]. Stating the likelihood of an unknown's skeletal variation matching a constructed social group by extrapolating from a biased sample selected from an IRC is highly problematic. It is relatively easy to find significant statistical correlations between skeletal variation and arbitrary typological categories. Albanese and Saunders [1] demonstrated how pubic bone length seems to be highly correlated with "race" in the Terry collection. However, the pattern of variation is due to mortality bias, where adult females who were described as Black died prematurely with compromised growth.

The correlation of variation with a racialized group is an artifact of the collection process, which is not unique to the Terry collection and affects all of the IRCs. The skeletal manifestations reflect the impact of poverty, racism, and economic disparity, and this impact on the skeleton varies [34,81]. In other words, assessing the impact of racism is not a useful approach for estimating race, since it is not consistent through time and does not manifest in the same way on the skeleton. Variation is not static and does not exist outside of economic and political contexts. Furthermore, the parameters for group membership vary. Socially defined racial identity, as in the USA, and bureaucratically defined racial designation, as in South Africa, vary through time, by region, and jurisdiction (see [1] for a discussion concerning social and bureaucratic races). One can easily change their race by crossing a political boarder [82]. A method using pelvic dimensions based on a sample from the Terry collection will seem promising for estimating race, ancestry or affinity. However, the failure to assess the complex causes of variation, racism and not race, results in a significantly poorer understanding of human variation, and methods that cannot possibly work when applied to forensic cases. When used in a similar manner, Howells' data can and should be considered as a collection that is similar to almost any other IRC. As with all of the other IRCs, these biased samples were constructed through the efforts of collectors working within historically specific contexts. IRCs are highly biased samples of human variation and must be sampled accordingly to address specific research questions. The arbitrary, artificial homogeneity of Howells' data that also exaggerates group differences is not a unique problem. It should not be a surprise that AncesTrees (37%) did not make any difference in allocation accuracy for matching premortem records when compared with the Fordisc (36%) trial with Howells' data. Fordisc, which uses racial terminology, and AncesTrees which uses ancestry terminology, are based on the theoretically erroneous assumption that social categories are fixed and unchanging and that race or ancestry groups capture at least some human variation. Switching to terms, such as "population affinity" or only "affinity" does not solve any of the fundamental problems: All of the real-bone and virtual IRCs are biased samples and not populations; variation is not static; groups are not homogeneous and distinct; and group membership is fluid and based on social, economic, historical, and political criteria that do not account for phenotypic or genotypic variation.

In this paper, the focus has been on metric approaches to assess the key issues related to software updates, alternative statistics, and expanded or updated reference samples. However, the results are relevant regardless of the types of data used. The debate and discussion regarding metric versus morphoscopic data span theoretical and methodological problems. The difference in accuracy is marginal, and consistently poor, when testing metric (50%) versus morphoscopic (53%) methods for estimating group membership [3,4,79]. As Albanese and Saunders [1] have demonstrated, the fundamental problems with typological approaches are the same regardless of the types of data used, including metric, non-metric, morphoscopic, and genetic data. Furthermore, geometric morphometric analyses do not perform in a better way than traditional metric methods [65]. Different approaches to data collection can capture similar patterns of variation [9]. The pubic bone is an important source of information for estimating sex. Additionally, as noted above, it can be erroneously linked to race. Variation in the pubic bone is directly attributable to differential growth at the symphyseal end of the bone, which will result in the presence or absence of the ventral arc [83,84]. The ventral arc itself cannot be measured with calipers, and thus it is scored. However, the differential growth of the pubic bone can be assessed with the measurement of the superior ramus of the pubic bone (SPRL), a more reproducible alternative to the traditional measurement of the pubic bone. Measuring the pubic bone and scoring the manifestation of the ventral arc capture the same variation. There are no "magical" types of data that will solve fundamental problems with conceptualizing human variation and poor sampling strategies for the construction of reference samples from IRCs [5].

Moreover, spanning theoretical and methodological problems, the linking of sex, stature, and age estimation to arbitrary group membership undermines the utility of these methods in a forensic context. Group-specific methods for estimating sex, stature, and age at death, are more difficult to apply since group membership must be first assessed, while at the same time, these group-specific methods do not provide better results when independently tested. The importance of decoupling all of the questions in the biological profile from each other, and especially from race/ancestry/affinity and nationality, has been well illustrated in a series of publications on stature estimation in a forensic context [10–12]. Regardless of how the groups were defined (age, sex, race, continental origin, nationality, year of birth, etc.) the group-specific methods did not provide better estimates of stature. Sex-specific equations provide marginally better results, but only if sex was certain. An error in estimating sex would result in a noticeable decrease in precision and utility for estimating stature. The method for stature estimation, developed in part from a subsample of individuals described as "Black females" from the "old" Terry collection, worked as well or better than a Portuguese-specific and males-specific method when tested on a sample from Portugal. The generic equations for stature estimation worked best most often without the need to estimate group membership. Better results are possible when addressing each question concerning the biological profile independently of each other. Additionally, erroneous results from one question will not have an impact on answers to the other questions. The forced linking of sex estimation to race estimation in Fordisc is more explicit evidence in support of uncoupling group membership from sex estimation. Fordisc undermines its own potential usefulness for estimating sex. If only sex could be estimated, the two versions of Fordisc could provide useful information regarding sex that matched premortem records in 73–75% of cases. By linking sex to race with a typological approach to human variation, Fordisc provides information that will compromise an investigation in 60–61% of cases.

There are at least two interconnected levels to an ethical argument against the continued use of a typological approach in Forensic Anthropology [1,3,4,15]. First, forensic anthropologists, as with all researchers and scholars, have an ethical obligation to conduct rigorous research and use the best theories and methods available for research and practice. As demonstrated in the current research, a typological approach to conceptualizing human variation that racializes the dead will result in a poor understanding of human variation, and produce methods that do not work as intended. Current methods in Forensic

Anthropology for estimating group membership provide wrong information that would compromise a forensic investigation in 50–64% of cases. After 60 years, Fordisc, the updated and improved automated version using discriminant function analysis, does not perform any better than the original Giles and Elliot [71] method when tested by Birkby [58] on a large, independent sample. The utility of AncesTrees to provide information that would be useful in a forensic investigation is even lower. Furthermore, linking sex, stature, and age at death estimation to group membership (race, ancestry, affinity, nationality) places ideology ahead of utility and compromises the answers to those questions, as well. Forensic anthropologists have an ethical obligation to pursue better research and practice.

Harm to the living is the direct result of the second level for an ethical argument against a typological approach. Forensic anthropologists have a privileged place at the intersection of law and science in popular and scholarly discourse. How something is said can be as important as what is said (see [81] for a case study applying critical discourse analysis in Forensic Anthropology). Regardless of the terms used by forensic anthropologists, in highly racialized societies (for example, South Africa, USA, Brazil), authorities are asking about race, a concept used to marginalize and exclude, and historically to justify slavery and colonialism. Discussions of group membership in this context by privileged speakers reinforce racial stereotypes, which promote an Eurocentric world view and allow for third parties to be racist.

Two papers, published almost 20 years apart with similar titles, illustrate the persistence of these problematic issues with the discourse of methods for estimating group membership in Forensic Anthropology. Sauer [2] authored "Forensic Anthropology and the concept of race: if races don't exist, why are forensic anthropologists so good at identifying them?". Seventeen years later, Ousley and colleagues' [85] coauthored an article with a similar title: "Understanding race and human variation: why forensic anthropologists are good at identifying race". The focus of these two articles is to illustrate the persistent harm to the living that can be caused through a typological discourse of the dead [62,81]. This critique is directed at the discipline for the continued use of a typological approach in the study of human variation, and not any one forensic anthropologist (see Brace [26] on Sauer [2]). Third parties looking for "evidence" to justify overt racism and inexperienced forensic scholars can easily find provocative titles to articles published by forensic anthropologists in major journals. Detailed and contextualized discussions are more difficult to find in forensic literature. The more nuanced three-page discussion on race and Fordisc does not begin until page 72 of the current Fordisc manual (V 1.53) in the section "Race, Races, and 'Biological Race'", where statements, such as the following can be found: "FORDISC does not define, redefine, or justify any racial classifications . . . ".

In the titles for both publications [2,85], the respective authors establish the uncontested view of the reality of "race" without defining the meaning of the concept: Biological? social? bureaucratic? As stated in this manner, readers can default to their own meaning of race to reinforce their world view. Furthermore, those authors assert the authority of their discipline and frame the question as to *why* forensic anthropologists are very good at identifying "them." Framing a question in this way dismisses the value of research that is critical of race by scholars from outside of the discipline and precludes the need for independent assessments of methods through rigorous testing from within the discipline. Despite the biased framing through these provocative titles, the results presented in this paper are consistent with the results from previous research: Forensic anthropologists are not good at and have never been good at identifying race. After 60 years of refinement, different methods, different group terminologies, alternative statistics, modified reference samples, software updates, and updated IRCs have not made a difference. It is long overdue for forensic anthropologists to abandon a typological approach to human variation, regardless of the terms used.

5. Conclusions

The problems with a typological approach to conceptualizing human variation are theoretical, methodological, and ethical, regardless of the terms used (race, ancestry, population affinity, ethnicity, etc.). These concepts are firmly rooted in a misconception of variation among the general public. Additionally, they are reinforced by the way forensic anthropologists have used and continue to use IRCs.

Methodologically, the results from the current research are consistent with previous blind, independent tests of methods for assigning an unknown individual to a group. The methods consistently perform very poorly (36–50%). Software updates, alternative statistical approaches, expanded reference samples, etc. have not and cannot solve the fundamental theoretical problems with a typological approach to human variation. Since they are based on flawed assumptions regarding the variation sampled in IRCs, these methods do not and cannot provide information that would be useful in a forensic investigation. Forensic anthropologists can expect wrong or misleading results involving a mismatch with premortem descriptions in at least half of the cases, with a failure of probability statistics to flag problematic assessments. Linking sex, stature, and age estimation to an estimated group membership, only ensures that the answers to these questions concerning the biological profile will also be compromised.

Theoretically, it is to be expected that group estimation methods cannot possibly work, since evidence of over 100 years is unambiguous and demonstrates that phenotypic and genotypic variation does not cluster into racial categories, continental origin or population affinities. This typological approach is firmly rooted in pseudo-scientific concepts of human variation that have been used to racialize, marginalize, and exclude individuals and groups. Sanitizing terminology does not address the wide-spread misconceptions of human variation in scholarly context or with the general public. The continued use of IRC data in a typological framework in Forensic Anthropology will contribute to a poorer understanding of human variation.

Ethically, forensic anthropologists have an obligation to conduct rigorous research and to use the best available methods. No information in a forensic investigation is better than wrong information, which will result in taking an investigation in the wrong direction. Pursuing a typological approach places ideology regarding human differences ahead of utility in a forensic context. Furthermore, forensic anthropologists have a very privileged place in the scholarly and public discourse regarding human variation. Although not intended by forensic anthropologists, sanitized terms serve as dog whistles that are heard by some segments of society with a long history of white supremacy. In 2006, Albanese and Saunders asked a question with the title: "Is it possible to escape racial typology in forensic identification?". With rigorous theoretical and methodological frameworks, it is possible and necessary to transition from a typological conceptualization of variation to the effective utilization of IRCs in Forensic Anthropology, and to discuss human variation in a manner that does not further marginalize the living.

References

1. Albanese, J.; Saunders, S.R. Is it Possible To Escape Racial Typology In Forensic Identification? In *Forensic Anthropology and Medicine: Complementary Sciences from Recovery to Cause of Death*; Schmitt, A., Cunha, E., Pinheiro, J., Eds.; Humana Press: Totowa, NJ, USA, 2006; pp. 281–315.
2. Sauer, N.J. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? *Soc. Sci. Med.* **1992**, *34*, 107–111. [CrossRef]
3. Plens, C.R.; De Souza, C.D.; Albanese, J.; Lopez Capp, T.T.; De Paiva, L.A.S. Reflections on methods to estimate race and ancestry on reference osteological samples in the Brazilian context. *Ethics Med. Public Health* **2021**, *18*, 100680. [CrossRef]
4. Plens, C.R.; De Souza, C.D.; Albanese, J.; Lopez Capp, T.T.; De Paiva, L.A.S. Reflexões sobre métodos de estimativa de ancestralidade em amostras osteológicas de referência no contexto brasileiro. *Braz. J. Forensic Anthropol. Leg. Med.* **2021**, *3*, 2675–6951.
5. Albanese, J. Approaches for Constructing and Using Reference Samples to Maximize the Utility of New Forensic Methods. *Eur. J. Anat.* **2021**, *25*, 63–82.
6. Sharman, J.; Albanese, J. Bioarchaeology and identified skeletal collections: Problems and potential solutions. In *Identified Skeletal Collections: The Testing Ground of Anthropology*; Henderson, C., Alves Cardoso, F., Eds.; Archaeopress: Oxford, UK, 2018; pp. 83–114.
7. Albanese, J. Strategies for Dealing with Bias in Identified Reference Collections And Implications For Research in the 21st Century. In *Identified Skeletal Collections: The Testing Ground of Anthropology*; Henderson, C.Y., Cardoso, F.A., Eds.; Archaeopress: Oxford, UK, 2018; pp. 59–82.
8. Albanese, J. A method for determining sex using the clavicle, humerus, radius and ulna. *J. Forensic. Sci.* **2013**, *58*, 1413–1419. [CrossRef]
9. Albanese, J. A metric method for sex determination using the hipbone and femur. *J. Forensic. Sci.* **2003**, *48*, 263–273. [CrossRef]
10. Albanese, J.; Tuck, A.; Gomes, J.; Cardoso, H.F.V. An alternative approach for estimating stature from long bones that is not population- or group-specific. *Forensic Sci. Int.* **2016**, *259*, 59–68. [CrossRef]
11. Albanese, J.; Osley, S.E.; Tuck, A. Do century-specific equations provide better estimates of stature? A test of the 19th–20th century boundary for the stature estimation feature in Fordisc 3.0. *Forensic Sci. Int.* **2012**, *219*, 286–288. [CrossRef]
12. Albanese, J.; Osley, S.E.; Tuck, A. Do group-specific equations provide the best estimates of stature? *Forensic Sci. Int.* **2016**, *261*, 154–158. [CrossRef]
13. Cardoso, H.F.V.; Spake, L.; Rios, L.; Albanese, J. Population variation in diaphyseal growth and age estimation of juvenile skeletal remains. In *Remodelling Forensic Skeletal Age*; Algee-Hewitt, B., Kim, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2021; pp. 99–138.
14. Spake, L.; Cardoso, H.F.V. Are we using the appropriate reference samples to develop juvenile age estimation methods based on bone size? An exploration of growth differences between average children and those who become victims of homicide. *Forensic Sci. Int.* **2018**, *282*, 1–12. [CrossRef]
15. Digangi, E.A.; Bethard, J.D. Uncloaking a lost cause: Decolonizing ancestry estimation in the United States. *Am. J. Phys. Anthropol.* **2021**, *175*, 422–436. [CrossRef]
16. Komar, D.A.; Buikstra, J.E. *Forensic Anthropology: Contemporary Theory and Methods*; Oxford University Press: New York, NY, USA, 2008.
17. İşcan, M.Y.; Steyn, M. *The Human Skeleton in Forensic Medicine*; Charles C Thomas: Springfield, IL, USA, 2013.
18. Edge, M.D.; Ramachandran, S.; Rosenberg, N.A. Celebrating 50 years since Lewontin's apportionment of human diversity. *Phil. Trans. R. Soc. B* **2022**, *377*, 20200405. [CrossRef]
19. Albanese, J. Albanese, J. A critical Review of The Methodology for The Study of Secular Change Using Skeletal Data. In *Papers in Honour of Michael Spence*; Publication No 9; Ellis, C., Ferris, N., Timmins, P., White, C., Eds.; Ontario Archaeological Society Occasional: London, ON, Canada, 2010; pp. 139–155.
20. Armelagos, G.J.; Van Gerven, D.P. A century of skeletal biology and paleopathology: Contrasts, contradictions, and conflicts. *Am. Anthropol.* **2003**, *105*, 53–64. [CrossRef]
21. Armelagos, G.J.; Carlson, D.S.; Van Gerven, D.P. The theoretical foundations and development of skeletal biology. In *A History of American Physical Anthropology 1930–1980*; Spencer, F., Ed.; Academic Press: New York, NY, USA, 1982; pp. 305–328.
22. Armelagos, G.J.; Goodman, A.H. Race, racism, and anthropology. In *Building a New Biocultural Synthesis: Political-Economic. Perspectives on Human Biology*; Goodman, A.H., Leatherman, T.L., Eds.; University of Michigan Press: Ann Arbor, MI, USA, 1998; pp. 359–378.
23. Blakey, M.L. Skull doctors: Intrinsic social and political bias in the history of American physical anthropology with special reference to the work of Aleš Hrdlička. *Crit. Anthropol.* **1987**, *7*, 7–35. [CrossRef]
24. Blakey, M.L. Beyond European enlightenment: Towards a critical and humanistic human biology. In *Building a New Biocultural Synthesis: Political-Economic Perspectives on Human Biology*; Goodman, A.H., Leatherman, T.L., Eds.; University of Michigan Press: Ann Arbor, MI, USA, 1998; pp. 379–406.
25. Brace, C.L. The roots of the race concept in American physical anthropology. In *A History of American Physical Anthropology 1930–1980*; Spencer, F., Ed.; Academic Press: New York, NY, USA, 1982; pp. 11–29.
26. Brace, C.L. Region does not mean "race"—Reality versus convention in forensic anthropology. *J. Forensic Sci.* **1995**, *40*, 171–175. [CrossRef]

27. Brown, R.A.; Armelagos, G.J. Apportionment of racial diversity: A review. *Evol. Anthropol.* **2001**, *10*, 34–40. [CrossRef]

28. Corcos, A.F. *The Myth of Human Races*; Michigan University Press: Lansing, UK, 1997.

29. Duster, T. The concept of race in science. genetic nature/culture. In *Genetic Nature/Culture, Anthropology and Science Beyond the Two-Culture Divide*; Goodman, A.H., Heath, D., Lindee, M.S., Eds.; University of California Press: Berkeley, CA, USA, 2003; pp. 258–277.

30. Duster, T. Selective Arrests, An Ever-Expanding DNA Forensic Database, and the Specter of an Early-Twenty-First-Century Equivalent Of Phrenology. In *The Technology of Justice: DNA and the Criminal Justice System*; Lazer, D., Ed.; MIT Press: Cambridge, MA, USA, 2004; pp. 159–169.

31. Excoffier, L.; Hamilton, G. Comment on "genetic structure of human populations". *Science* **2003**, *300*, 1877. [CrossRef]

32. Goodman, A.H. Bred in the bone? *Science* **1997**, *37*, 20–25.

33. Gould, S.J. *The Mismeasure of Man*, 2nd ed.; Norton and Company: New York, NY, USA, 1996.

34. Gravlee, C. How race becomes biology: Embodiment of social inequality. *Am. J. Phys. Anthropol.* **2009**, *139*, 47–57. [CrossRef]

35. Keita, S.O.Y.; Kittles, R.A.; Royal, C.D.; Bonney, G.E.; Furbert-Harris, P.; Dunston, G.M.; Rotimi, C.N. Conceptualizing human variation. *Nat. Genet.* **2004**, *36*, S17–S20. [CrossRef]

36. Keita, S.O.Y.; Kittles, R.A. The persistence of racial thinking and the myth of racial divergence. *Am. Anthropol.* **1997**, *99*, 534–542. [CrossRef]

37. Keita, S.O.Y.; Boyce, A.J. "Race": Confusion about zoological and social taxonomies, and their places in science. *Am. J. Hum. Biol.* **2001**, *13*, 569–575. [CrossRef] [PubMed]

38. Kittles, R.; Keita, S.O.Y. Interpreting African genetic diversity. *Afr. Archaeol. Rev.* **1999**, *16*, 87–91. [CrossRef]

39. Latter, B. Genetic differences within and between populations of the major human subgroups. *Am. Natural.* **1980**, *116*, 220–237. [CrossRef]

40. Lewontin, R. The apportionment of human diversity. *Evol. Biol.* **1972**, *6*, 381–398.

41. Lieberman, L. How "Caucasoids" got such big crania and why they shrank—From Morton to Rushton. *Curr. Anthropol.* **2001**, *42*, 69–95. [CrossRef]

42. Livingstone, F. On the non-existence of human races. *Curr. Anthropol.* **1962**, *3*, 279–281. [CrossRef]

43. Molnar, S. *Human Variation: Races, Types, and Ethnic Groups*, 6th ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2005.

44. Ossenberg, N.S. Within and between race distances in population studies based on discrete traits of the human skull. *Am. J. Phys. Anthropol.* **1976**, *45*, 701–716. [CrossRef]

45. Relethford, J.H. Global analysis of regional differences in craniometric diversity and population substructure. *Hum. Biol.* **2001**, *73*, 629–636. [CrossRef]

46. Smay, D.; Armelagos, G.J. Galileo wept: A critical assessment of the use of race in forensic anthropology. *Transform. Anthro.* **2000**, *9*, 19–29. [CrossRef]

47. Templeton, A.R. Human races: A genetic and evolutionary perspective. *Am. Anthropol.* **1998**, *100*, 632–650. [CrossRef]

48. Tishkoff, S.A.; Kidd, K.K. Implications of biogeography of human populations for 'race' and medicine. *Nat. Gen.* **2004**, *36*, S21–S27. [CrossRef] [PubMed]

49. Washburn, S.L. The study of race. *Am. Anthropol.* **1963**, *65*, 521–531. [CrossRef]

50. Cobb, W.M. Race and runners. *J. Health Phys. Ed.* **1936**, *7*, 1–9. [CrossRef]

51. Kennedy, K.A. But Professor, why teach race identification if races don't exist? *J. Forensic. Sci.* **1995**, *40*, 797–800. [CrossRef]

52. Ousley, S.D.; Jantz, R.L.; Hefner, J.T. From Blumenbach to Howells: The Slow, Painful Emergence of Theory Through Forensic Race Estimation. In *Forensic Anthropology: Theoretical Framework and Scientific Basis*; Boyd, C.C., Boyd, D.C., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2018; pp. 67–97.

53. Martin, B.L. From Negro to Black to African American: The power of names and naming source. *Political Sci. Q.* **1991**, *106*, 83–107. [CrossRef]

54. Ross, A.H.; Pilloud, M. The need to incorporate human variation and evolutionary theory in forensic anthropology: A call for reform. *Am. J. Phys. Anthropol.* **2021**, *176*, 672–683. [CrossRef]

55. Spradley, K.; Jantz, R.L. What Are We Really Estimating in Forensic Anthropological. Practice, Population Affinity or Ancestry? *Forensic Anthropol.* **2021**, *4*, 171–180. [CrossRef]

56. Hunt, D.R.; Albanese, J. The history and demographic composition of the Robert J. Terry Anatomical Collection. *Am. J. Phys. Anthropol.* **2005**, *127*, 406–417. [CrossRef]

57. Cunha, E.; Ubelaker, D.H. Evaluation of Ancestry from Human Skeletal Remains: A Concise Review. *Forensic Sci. Res.* **2020**, *5*, 89–97. [CrossRef]

58. Birkby, W.H. An evaluation of race and sex identification from cranial measurements. *Am. J. Phys. Anthropol.* **1966**, *24*, 21–27. [CrossRef] [PubMed]

59. Williams, F.L.; Belcher, R.L.; Armelagos, G.A. Forensic misclassification of ancient Nubian crania: Implications for assumptions about human variation. *Curr. Anthropol.* **2005**, *46*, 340–346. [CrossRef]

60. Parsons, H.R. An Evaluation of Forensic Anthropology Reports in the United States. *Forensic Anthropol.* **2021**, *4*, 192–202.

61. Thomas, R.M.; Parks, C.L.; Richard, A.H. Accuracy Rates of Ancestry Estimation by Forensic Anthropologists Using Identified Forensic Cases. *J. Forensic Sci.* **2017**, *62*, 971–974. [CrossRef]

62. Albanese, J. Anthropologia Forense para os vivos: Movendose além da análise da casos na luta pelos direitos humano universais. In *Violência de Estado na America Latina: Direitos Humanos, Justiça de Transição e Anthropologia Forense*; Amadeo, J., Ed.; UNIFESP: São Paulo, Brazil, 2019; pp. 425–434.

63. Ubelaker, D.H. Chapter 2 Forensic Anthropology: Methodology and Diversity of Applications. In *Biological Anthropology of The Human Skeleton*, 2nd ed.; Katzenberg, M.A., Saunders, D.R., Eds.; John Wiley and Sons: Hoboken, NJ, USA, 2008; pp. 41–69.

64. Fernandes, L.C.C.; Bento, M.I.C.; Rabello, P.M.; Soriano, E.P.; Navega, D.; Júnior, E.D.; Cunha, E. Analysis of the Accuracy of AncesTrees Software in Ancestry Estimation in Brazilian Identified Sample. *Adv. Anthropol.* **2021**, *11*, 163–178. [CrossRef]

65. Dudzik, B. Examining cranial morphology of Asian and Hispanic populations using geometric morphometrics for ancestry estimation. *Forensic Anthropol.* **2019**, *2*, 304–315. [CrossRef]

66. Urbanová, P.; Ross, A.H.; Jurda, M.; Nogueira, M.I. Testing the reliability of software tools in sex and ancestry estimation in a multi-ancestral Brazilian sample. *Legal Med.* **2014**, *16*, 264–273. [CrossRef]

67. Konigsberg, L.W.; Algee-Hewitt, B.F.B.; Steadman, D.W. Estimation and evidence in forensic anthropology: Sex and race. *Am. J. Phys. Anthropol.* **2009**, *139*, 77–90. [CrossRef]

68. Sauer, N.J.; Wankmiller, J.C. The assessment of ancestry and the concept of race. In *Handbook of Forensic Anthropology and Archaeology*; Blau, S., Ubelaker, D.H., Eds.; Lest Coast Press: Walnut Creek, CA, USA, 2009; pp. 187–200.

69. Elliott, M.; Collard, M. FORDISC and the determination of ancestry from cranial measurements. *Bio. Lett.* **2009**, *5*, 849–852. [CrossRef]

70. Ousley, S.D.; Jantz, R.L. The forensic data bank: Documenting skeletal trends in the United States. In *Forensic Osteology*; Reichs, K., Ed.; Charles C Thomas: Springfield, IL, USA, 1998; pp. 441–458.

71. Giles, E.; Elliot, O. Race identification from cranial measurements. *J. Forensic Sci.* **1962**, *7*, 147–157.

72. Van Vark, G.N. Advances in The Quantitative Analysis of Skeletal Morphology. In *Skeletal Biology of Past Peoples: Research Methods*; Saunders, S.R., Katzenberg, M.A., Eds.; Wiley-Liss: Hoboken, NJ, USA, 1992; pp. 225–257.

73. Jantz, R.L.; Ousley, S.D. *FORDISC 3: Computerized Forensic Discriminant Functions*, Version 3.1; The University of Tennessee: Knoxville, TN, USA, 2005.

74. Ousley, S.D.; Jantz, R.L. Fordisc 3 and Statistical Methods for Estimating Sex and Ancestry. In *A Companion to Forensic Anthropology*; Dirkmaat, D.C., Ed.; Wiley-Blackwell: London, UK, 2012; pp. 311–329.

75. Howells, W.W. *Skull Shapes and the Map: Papers of the Peabody Museum of Archaeology and Ethnology*; Harvard University: Cambridge, MA, USA, 1989; Volume 79.

76. Navega, D.; Coelho, C.; Vicente, R.; Ferreira, M.T.; Wasterlain, S.; Cunha, E. AncesTrees: Ancestry estimation with randomized decision trees. *Int. J. Legal. Med.* **2015**, *129*, 1145–1153. [CrossRef] [PubMed]

77. Rocha, A. Les collections ostéologiques humaines identifiées du Musée Anthropologigue de l'Université de Coimbra. *Antropol. Port.* **1995**, *13*, 7–38.

78. Moore-Jansen, P.H.; Ousley, S.D.; Jantz, R.L. *Data Collection Procedures for Forensic Skeletal Material*; Report No. 48; Department of Anthropology, The University of Tennessee: Knoxville, TN, USA, 1994.

79. Sierp, I.; Henneberg, M. Can ancestry be consistently determined from the skeleton. *Anthropol. Rev.* **2015**, *78*, 21–31. [CrossRef]

80. Ubelaker, D.H.; Ross, A.H.; Graver, S.M. Application of forensic discriminant functions to a Spanish cranial sample. *Forensic Sci. Com.* **2002**, *4*, 1–6.

81. Albanese, J.; Halliday, J. Muitas maneiras de desaparecer: Defendendo os Direitos Humanos Por Meio De Padrões Éticos E Pesquisas Mais Rigorosas Em Uma Antropologia Forense Crítica. In *Direito Humanos Sob a Perspectiva Do Direito À Vida, Da Antropologia Forense E Da Justiça No Caso De Violações*; Plens, C., Ed.; Editora Annablume/HHRRC (AAFS): São Paulo, Brazil, 2022; pp. 263–300.

82. Templeton, A.R. Biological races in humans. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2013**, *44*, 262–271. [CrossRef]

83. Anderson, B.E. Ventral arc of the os pubis: Anatomical and developmental considerations. *Am. J. Phys. Anthropol.* **1990**, *83*, 449–458. [CrossRef]

84. Budinoff, L.C.; Tague, R.G. Anatomical and developmental bases for the ventral arc of the human pubis. *Am. J. Phys. Anthropol.* **1990**, *82*, 73–79. [CrossRef]

85. Ousley, S.D.; Jantz, R.L.; Freid, D.L. Understanding Race and Human Variation: Why Forensic Anthropologists are Good at Identifying Race. *Am. J. Phys. Anthropol.* **2009**, *139*, 68–76. [CrossRef]