8-25-2021

A New Theoretical Approach to Ancestry Estimation as Applied to Human Crania

Michael W. Kenyhercz
Department of Defense POW/MIA Accounting Agency

A New Theoretical Approach to Ancestry Estimation as Applied to Human Crania

Michael W. Kenyhercz[1]*

[1]Central Identification Laboratory, Department of Defense POW/MIA Accounting Agency, Joint Base Pearl Harbor, Hickam, Hawaii, USA.

*Correspondence to: Michael W. Kenyhercz, Central Identification Laboratory, Department of Defense POW/MIA Accounting Agency, 590 Moffet Street, Building 4077, Joint Base Pearl Harbor, Hickam, HI 96853 USA. E-mail: michael.kenyhercz@gmail.com.

Short Title: New Theoretical Approach to Ancestry Estimation

KEY WORDS: FORENSIC ANTHROPOLOGY, DATA SCIENCE, CLUSTERING, UNSUPERVISED MODELING, RANDOM FOREST MODEL.

Abstract

Since Frank Livingstone proposed the idea that there are no races, only clines, in 1962, little has changed in the way that anthropologists study and, ultimately, estimate ancestry. The way in which we talk about the study of human variation may have changed--shifting away from "racial" labels and towards those of supposed ancestral origins--but the methods with which we label and analyze groups, however termed, has remained the same. In this paper, I suggest a new theoretical approach to ancestry estimation that does not rely on group labels using Howells Craniometric dataset as an example. In the suggested workflow, the data structure themselves into natural clusters, which I am referring to as Morphogroups, without the reliance of a group label. Each Morphogroup is explored for sub-groups and the process is repeated until no further distinctions can be made. At each level an individual is compared to the Morphogroup in a descriptive manner focusing on similarities and differences. Lastly, a multi-iteration classification procedure, using random forest modeling, is implemented to classify by Morphogroup. In this test, hierarchical clustering was used to identify the optimal number of natural clusters within the data and principal components analysis was used to explore Morphogroups. Using my suggested workflow, three main Morphogroups were identified with each having different numbers of subclusters ranging from 0-8. Morphogroup correct classifications are typically in the mid 90 percent and the accompanying sex estimations between 93-100% correct. Additionally, for anyone who has access to R, I have provided a markdown file that shows all of the code used for this paper step-by-step at https://rpubs.com/kenyhercz2/717620. I want to make it clear this is not the way I think this should be done, rather one of myriad ways it could be done. Human variation and identity are not

static and we need to stop thinking of them as such. It is on us to help one another get better at rethinking and redefining what is possible for our field.

"If a central problem of physical anthropology is the explanation of the genetic variability among human populations—and I think it is—then there are other methods of describing and explaining this variability which do not utilize the race concept" —Frank B. Livingstone, 1962

Since Frank Livingstone formally introduced the idea that there are no races, only clines in 1962, the process by which anthropologists estimate the ancestry of an individual has remained largely unchanged. Instead of treating human variation under the lens of natural selection and continuous, clinal variation of morphological features, we have done little aside from moving away from the "racial" labels that we all firmly understand as biologically meaningless, and towards the use of "ancestral" labels that supposedly relate to an individual's ancestral land of origin at some level of fidelity, such as Hungarian or European. There has been recent discussion within the field of forensic anthropology about the usefulness of estimating ancestry and if we, as a field, should even be studying human variation under this lens (Bethard and DiGangi, 2020). This is an opinion I clearly disagree with (Stull *et al.*, 2020). While Bethard and DiGangi (2020) specifically take issue with the treatment of morphoscopic methods, the issues with estimating ancestry run deeper than simply the methods with which anthropologists estimate it. The problem, as I see it, is with the mutually exclusive group label that we assume to be meaningful, regardless of the type of data used to examine it. To me, Bethard and DiGangi have resigned themselves to their own understanding of what it means to study modern human variation; but let me be clear, this is not a fault exclusively of their own design. Our field has been recapitulating the same methods and workflows (i.e., frameworks) for studying human variation since *before* Livingstone provided an exceedingly reasonable alternative. In this paper, I intend to: 1) introduce an alternative framework to studying human variation without the

reliance of group labels or even selecting which groups to compare using Howells dataset as an example; 2) demonstrate, step-by-step, how to incorporate this alternative framework; and 3) explore the possibilities of what expanding our own domain of understanding can mean for our field. Most importantly, I am not suggesting that the framework I will present is the way it should be, but rather it is one of many ways it *could* be.

Ancestry estimation within the confines of biological and forensic anthropology typically begins with an analyst choosing groups to compare, or utilizing methods that can only compare certain groups. This initial step is fundamental, particularly when utilizing discrimination techniques, because all supervised classificatory models will force the unknown into one of these pre-defined groups. Thus, if the "true" group is not chosen initially, results can be misleading, especially if the analyst does not investigate model details, such as posterior and typicality probabilities. Historically, this *a priori* procedure was justified because it allowed the analyst to create an informed model by choosing to compare only the most probable groups given each case's particular context. A discriminatory model with fewer group possibilities has more power, simply given the geometric operations of popular classification statistics, such as linear discriminant function analysis. If possible, a practitioner may employ a multi-iteration approach in which a specimen is first compared to all groups, and then tested again with the least similar group removed given some criteria (such as low posterior and typicality probabilities). Unfortunately, a multi-iteration approach is not always possible given the sample limits (either in size or composition) of many currently available methods. A notable exception to this aforementioned workflow was presented by Algee-Hewitt (2016) that utilizes unsupervised classification using a tripartite diffusion model, commonly employed in population genetic studies that attributes proportions of membership to three broad geographic groups (Asian,

African, and European). However, in my experience, this approach has been underutilized in favor of the above mentioned procedures.

In addition to the mechanical issues of initial group selection, many popular methods use inconsistent or nebulous group labels that make comparing results among different methods difficult (e.g., White vs Euro-American vs European; Hispanic vs Guatemalan). Beyond choice of group label, there is rarely any discussion of what that label actually means to the researcher as it applies to their study—and, importantly, is this group label sufficient to encapsulate the variation under study. I am guilty of the inconsistent use of group labels in my own work. For example, pooling groups, in particular, always seemed justifiable enough to increase sample sizes. One generic label does not allow for any discussion of what makes the individual under examination unique, particularly within a population of interest; and in this regard, we are doing a disservice to the deceased. After devoting so much time to studying human variation, it started to seem unreasonable to be employing such complex statistical procedures on a breadth of measurements and observations only to reduce an individual into one generic label. To me, this is an incredibly limiting, self-imposed behavior, but one that is certainly enforced by our training and tendency to conduct similar types of studies. Thinking about human variation within the confines of one generic label limits our own scientific domain; we are not learning anything new or novel and, in this regard, we are doing a disservice to ourselves. I do understand the perceived need of the medico-legal community to use terms like “White” or “Black” to aid in the identification of an unknown skeleton, but how honestly useful are these rudimentary group labels in standard practice?

When I was working on my dissertation, my advisor, Joel Irish, told me that my analyses were painting a picture of human variation using broad brushstrokes, just as one of his advisors

told him (Personal communication, J.D. Irish, 2010). I have never stopped thinking about that. If we can consider our prevailing methods for estimating ancestry as tools; they are blunt objects. Beyond the blunt tools at our disposal, the generic labeling of the samples is vague at best. As an example, imagine you have collected a dataset of craniometrics from 200 males and females (100 Black, 100 White) from the Hamann-Todd Osteological Collection and you are going to employ linear discriminant function analysis as a means to estimate ancestry of unidentified crania. You build your model and then test it on independent data and find optimistic results that you then communicate in some way (test report, conference presentation, article, etc.). My question to you is, what is this unknown cranium being compared to in your model? You might answer with the sample demographics breakdown, but my question is more generic than that.

What is truly represented on the axes of your canonical variate plot is a generic encapsulation of variation from a combination of assumptions based on different external influences. First, there is already an issue in the sample demographics given how ancestry was assigned (and by whom). In the case of the Hamann-Todd Collection, ancestry was assigned at autopsy via visual means. So now our available group label is based on someone else's opinion of whatever it looks like to be Black or White. Taken with the low socioeconomic status of much of the sample, your generic label represents someone's idea about someone else's identity from a population whose lack of available resources might have biological repercussions that may intrinsically bias results. Second, the mechanics of linear discriminant function analysis use these generic labels to construct a mean representation of each group based on mathematical operations to maximize the differences among and between these group means. So what you are actually comparing the unknown to is now the average of someone else's opinion on whatever it means to be Black or White and likely from a low socioeconomic status. Do you feel confident

in your model? This is what it means to paint the picture of human variation with broad brushstrokes.

This is not to say that broad brushstrokes are not useful. They are incredibly useful within our field and, especially, outside of it. My intent is not to discredit this type of research, mostly because that would discredit nearly everything I have contributed to the field (McDowell, L'Abbé, and Kenyhercz, 2012; Kenyhercz, Klales, and Kenyhercz, 2014; Klales and Kenyhercz 2014, McDowell, Kenyhercz, and L'Abbé, 2015; Stull, Kenyhercz, Tise *et al.*, 2016; Berg and Kenyhercz, 2017; Kenyhercz and Berg, 2017; Kenyhercz, Klales, Rainwater *et al.*, 2017), but to critically examine what it means to estimate ancestry. However, to do so, you must first accept the paradox of the group label—it *is* and *is not* useful. The statistical procedures can accurately classify individuals into prescribed groups, but these procedures can also find significant patterns in random noise, which is problematic if you think about what your data are encapsulating. I find it more helpful to view these generic group labels as dynamic proxy variables that *might* account for: broad ancestral heritage, socioeconomic status, time-period, access to different groups, migration, and/or social identity. This list is neither exhaustive nor static—more than anything it is not any *one* thing. With these parameters I can more easily accept the averages represented on the canonical variates plot because I am aware of what they can and cannot mean. These are our broad brushstrokes.

So what is it that we are estimating? To answer that, we must first examine why we are undertaking this task. In forensic anthropology, a common answer to "why?" is likely because it was requested as part of a biological profile to aid in the identification process; for biological anthropologists it could be contextualizing the demographics in a previously unknown cemetery or ossuary. In both of these aforementioned instances, I think what we are looking for is twofold:

1) how does this fit into what I know and 2) how is this different from what I know. Oftentimes when we estimate ancestry, we abandon the second point because the blunt tools at our disposal are designed to give a definitive answer with some associated measure of certainty. I believe it stands to reason that incorporating both the similarities and the uniqueness of an unknown as it relates to our domain will do far more for an identification effort than the simple reporting and justification of a generic label.

Still, what does it mean to estimate ancestry? I tend to agree with Sauer (1992): these social labels, for whatever reason, had biological repercussions that we can measure in the skeleton. In South Africa and the United States, those biological repercussions could partially be because of enforced segregation at a more local, microevolutionary scale, wherein at a macroevolutionary scale those biological repercussions could be because of specific adaptation strategies of a group's evolutionary lineage, contextualized by migration out of Africa and into the other continents. The prior is an opinion based on my experience working on ancestry estimation in South Africa and the U.S. Many genetic studies have demonstrated the concordance to continental-level geographic patterning of groups with accurate classifications (see Rosenberg *et al.* 2002). So what is it that we are estimating? I believe that we are estimating the biological repercussions of *something*. What that *something* is will inevitably be idiosyncratic to the specific sample and problem. What that *something is not*, is just *one thing*.

The generic group label is only useful or not useful if you believe it to be one way and are unwilling to change your mind. It is both, and it is neither. This label could encapsulate an entire lineage of biological repercussion, or an externally mediated, non-random, microevolutionary response. Alternatively, the label could have no meaning, at least any that would have a biological consequence that we, as anthropologists, could detect. Put another way,

these generic group labels are not *meaningless*, we just do not *know* what they mean, if anything. This is akin to answering a question with a question that is masquerading as an answer. Taking dozens of measurements on thousands of crania only to reduce those individuals into a single answer seems like a misuse of that data, as well as a missed opportunity to learn something new.

In data science, it is important to practice thinking about problems in very broad, conceptual terms as this allows you to see (or even identify) the big picture; getting bogged down in specific minutiae will only lead to dead-ends requiring idiosyncratic problem-solving. My problem establishes many of the parameters of what is possible for me: how can I work with data that is grouped with labels that do and do not mean anything; and if they do mean something, what exactly does it mean and how does that relate to the labels of others? The obvious answer is to both use and not use these generic group labels. The first step is the data needs to describe itself based on the variables and not the group labels (aka unsupervised learning). The second step is the data needs to partition itself into an optimal number of clusters both using and not using the group labels; I refer to these clusters as Morphogroups and during this step I am using the Howells group means under the assumption that they contain some meaningful insight. Repeat the second step on each Morphogroup until no subclusters can be identified. Use group labels to help understand each Morphogroup subcluster. With all Morphogroups and subclusters identified, employ a set of classification procedures to first classify based on Morphogroup, then tested against only groups within that Morphogroup, and repeated for as many subclusters as exist. Lastly, collate results at each scale of Morphogroup that describes the similarities and differences of each Morphogroup and the unknown individual classified within it.

In this suggested workflow, the data are described at the level of the individual, regardless of group label. That group label is then used under the assumption that it means something, but what that is remains unsure. The group labels here are used to subset the data so that the relationships among the variables can be explored for similarities and differences. Put another way, examining the means of the prescribed group labels will allow us to explore the significance of them, if any exists. The classification procedures do make use of a group label, however that label is based on observed differences within individuals, i.e., the identified Morphogroup and subcluster(s). Essentially, the group label is mostly utilized in the collating of results and for inferential analyses. If wanted, you could extend the classification procedure down to that of the prescribed Howells' population label. The end product is not just one result, but a series of results and descriptions of how both the observed variation within each Morphogroup is distinct and how the unknown individual fits within that group. This might be something like "This individual classifies into a Morphogroup characterized by larger than average crania overall, but particularly in cranial lengths, and is represented by primarily samples derived from Europe. Within these larger and longer crania, this individual classifies into a subcluster characterized by narrow midface features, found most commonly in samples derived from Europe. Within this final subcluster the unknown classifies as *X* and displays broader midface features than are observed within group *X* in this dataset." Given this alternative framework, we learn something novel about our dataset, the individuals that compose it, and those unknown on which it might be applied.

Materials and Methods

All of the analyses were conducted in R (R Core Development Team 2020) and I have provided all of the code used in this paper in the form of a markdown file at http://rpubs.com/kenyhercz2/717620. A markdown file shows all of the code and resulting output from R with annotations describing what each line is doing. I have purposely tried to use as few nested functions as possible to facilitate an analysis workflow that can be followed step-by-step. All packages and functions used are shown in the markdown file and will not be repeated here.

Howells craniometric dataset was used to test the proposed framework—simply stated, Howells dataset represents cranial measurements from populations across the globe (see Howells 1996 for a full description). It should be noted that while the quote I chose to open the paper is in reference to clinal variation, Howells dataset is not particularly useful in that regard because it was amassed to represent as geographically distinct populations as possible. Table 1 shows each of the variables used in this test. I chose to omit many of the uncommon variables and create a new variable, Gross Size (GS), that is simply GOL*XCB*BBH, that will serve as a rough approximation of an overall size metric as opposed to a linear increase in just one dimension. Further, the inclusion of GS is meant to demonstrate how researchers can further extend their own datasets. For this test, males and females have been pooled and, as a result, all variables will be centered and scaled prior to analysis, which simply means that each variable mean is centered at 0 with a standard deviation of 1. In this way, the general magnitude of different measures are on the same scale.

Identifying Morphogroups

The centered and scaled craniometric variables are then averaged into each respective Howells Population and converted into a Euclidean distance matrix. The end result is a matrix that shows the pairwise squared differences between the average of all craniometric variables among the Howells Populations, i.e., a basic dissimilarity/distance matrix. The distance matrix is then subjected to hierarchical cluster analysis using Ward's criterion, which aims to minimize the amount of observed within cluster variance; however, Ward's criterion may be influenced by highly correlated variables, giving them more weight in the clustering procedures. Several methods exist for clustering and each will have their own assumptions and outcomes. Next, the optimum number of clusters is determined with the package *NbClust*. *NbClust* allows the user to set a range of clusters to test for (in the current example between 2 and 15) and then compares 30 different indices to determine how many natural clusters, within the parameters set, exist within the data. Many of the indices offer different determinations of optimum number of clusters, so the package uses a simple majority rule in determining the final number of clusters. The optimum number of clusters then forms what I am referring to as Morphogroups and the data are subset into their respective Morphogroups.

Exploring Morphogroups

Summary statistics are then generated for each Morphogroup. Next, the craniometrics are submitted to a principal components analysis using Varimax rotation to explore the variation within and among Morphogroups. Varimax rotation was utilized because it better elucidates relationships among the data by pushing coefficients to be large or pulling them near zero. In layman's terms, Varimax rotation is trying to amplify the signal and diminish the noise. Lastly, the principal component scores are plotted to visualize the relationships among the data.

Extension to Classification

Each Morphogroup is then subjected to the same procedures to identify subclusters and this entire process is repeated until no further distinction can be observed among the Howells groups within each Morphogroup. At each clustering step, the Morphogroup membership is updated; for example, if an individual clustered into Morphogroup 3 and then into subcluster 2 of that Morphogroup, and lastly sub, subcluster 1, their final Morphogroup is 3.2.1.

The Morphogroup designators can then be used, essentially, as dynamic labels that encapsulate and compare observed variation. With these different levels, or scales, of Morphogroup labels available, a multi-iteration classification procedure can be employed wherein the first model classifies the unknown into broad Morphogroup and then additional classification models will classify into as many subclusters that exist. If you so desire, this can be done to the individual Howells Population group label. At each iteration, sex can be estimated given the parameters of the specific Morphogroup (the one time wherein variables will not be centered and scaled). Model diagnostics, such as correct classification rates and posterior probabilities, are collated with details about the observed variation within and between clusters, as well as individuals within that cluster and the unknown being compared. For this study, random forest modeling was employed for the multi-iteration classification procedure. Simply put, random forest modeling is an ensemble of decision trees that tests random sets of variables on random subsets from the data, with each decision tree resulting in a classification; a final classification is established by majority rules and the posterior probability is equal to the number of trees that led to a specific classification. In a random forest model with 500 trees and two groups (say A and B), if 300 of those trees classify the unknown as A and 200 as B, the final

classification is A with a posterior probability of 0.6 (300 classifications / 500 trees), while the posterior probability for group B is 0.4 (200 classifications / 500 trees). Each random forest model makes use of a 70% training sample, 15% validation sample, and 15% testing sample with 1000 trees and 5 variables tested at each decision node. The number of trees allows for robust results, but the choice of variables tested at each node is arbitrary.

Results

The overall procedures detailed above were used to identify as many subclusters as could be detected using the same workflow—these results are shown in the accompanying markdown file. For the sake of brevity, I have included the initial identification of Morphogroups and the subclusters within Morphogroup 1 below, but have chosen to omit describing all of the results because this is not intended as a strict research paper. In my opinion, the results are the least important part of this paper.

Identifying and Exploring Morphogroups

The NbClust function indicates the optimum number of clusters for this dataset to be 3 (Figure 1). Morphogroup 1 consists of the groups encapsulated by the red rectangle in Figure 1, Morphogroup 2 by the green rectangle, and Morphogroup 3 by the blue rectangle. Of note, Morphogroup 2 only contains Andaman and Bushman populations.

The loadings for the three derived and retained principal components indicate that component one is defined primarily by cranial and facial lengths with the top three loadings being: BPL (0.848), BNL (0.845), and GOL (0.833). The loadings for PC2 are defined primarily by measures of cranial and facial breadths: XCB (0.865), XFB (0.807), and AUB

(0.779). Lastly, the loadings of PC3 are influenced primarily by measures of craniofacial heights: OBH (0.786), NPH (0.678), and NLH (0.669). We can use the mean PC scores for each Morphogroup to infer the following: Morphogroup 1 is characterized by the longest (mean PC1 = 0.489), broadest (mean PC2 = 0.246), and tallest (mean PC3 = 0.464) crania; Morphogroup 2 is characterized by the smallest crania in each dimension (mean PC1 = -1.03; mean PC2 = -0.706; mean PC3 = -0.980); Morphogroup 3 is intermediate in each dimension in comparison to the other Morphogroups (mean PC1 = -0.132; mean PC2 = -0.047; mean PC3 = -0.124).

Identifying and Exploring Subclusters: Morphogroup 1

Within Morphogroup 1, five subclusters were identified with NbClust (Figure 2). Here, the Buriat and Eskimo comprise their own subclusters. The loadings of PC1 indicate a focus on cranial and facial breadths with the top three being: XCB (0.862), XFB (0.825), and AUB (0.809). The loadings of PC2 demonstrate an influence of cranial and facial lengths with BNL (0.872), GOL (0.865), and NOL (0.843) as the top three contributors. Lastly, PC3 is primarily loaded with craniofacial heights, the top three being OBH (0.818), NPH (0.701), and NLH (0.663). The mean PC scores for each subcluster are shown in Table 2. Using the mean PC scores and details from the loadings of each component, we can extrapolate a broad brushstroke encapsulation of the variation observed within each of these subclusters. For example, subcluster 3 (Morphogroup 1.3), comprised of the Buriat, demonstrates the relatively broadest cranial and facial features and the relatively shortest crania.

Classification

The results from classification procedures are shown graphically by the flowchart in Figure 3. The boxes in Figure 3 represent two random forest models—one that uses the scaled data to classify into Morphogroup, and another that uses the raw data to classify by sex. In total, Figure 3 represents the results of 34 individual random forest models. The line segments that extend towards subclusters show the correct classification for that Morphogroup subcluster from the preceding model. When no further subclusters were identified, the correct classification rate for each population is shown beneath the highest fidelity Morphogroup subcluster. Of note, Morphogroup 1.1 can be extended to 1.1.1–1.1.3 with the observed variation being sufficient for each Howells group to have their own subcluster; this is also true for Morphogroup 3.1 to 3.1.1–3.1.4. Morphogroup correct classifications are typically in the mid 90 percent and the accompanying sex estimations between 93–100% correct.

Example

As an example of the above suggested workflow, I will use the first individual from Howells' test sample that is also included in the *bioanth* package, who is described as a Zalavar male. The code for these procedures is available in Appendix A. From the provided markdown file, I can use the random forest models constructed for classification purposes to first classify into one of the three identified Morphogroups with the *predict* function. In this first step, this individual is classified into Morphogroup 3 (posterior probability = 0.778). To examine the uniqueness of the individual, I can subtract the craniometric values from individual one from the mean of Morphogroup 3. This procedure indicates that this individual is larger than the mean values for Morphogroup 3 in nearly all dimensions except BPL, NLB, MAB, and OBH in which they are slightly lower than the respective means. So, in the first step, this individual classifies into a

Morphogroup with intermediate cranial dimensions, but this individual is larger than average in nearly every dimension, with smaller than average midfacial features. From the three subclusters identified in Morphogroup 3, the individual is then classified into subcluster 3.2 (posterior probability = 0.612). Examining the difference between test individual one and the mean for subcluster 3.2 indicates this individual is again larger in most observable dimensions, primarily GOL and NOL with only OBH demonstrating a shorter OBH than the mean for the group. Next, the individual is tested against the four subclusters identified within Morphogroup 3.2, and is classified into subcluster 3.2.2 (posterior probability = 0.772), which comprises three European populations (Norse, Berg, and Zalavar). Within subcluster 3.2.2, this individual again has larger than average craniometric dimensions except NLH and OBH. The next step moves into the Howells population labels and classifies this individual as Zalavar (posterior probability = 0.658). At this population level, this individual is again larger than all average dimensions with the exception of NLH and OBH. Finally, this individual is classified as a male (posterior probability = 0.96) and demonstrates longer than average lengths, but narrower breadths than other males within this Morphogroup, particularly in XCB and XFB, while still displaying shorter than average OBH.

If we were to use a popular program like *Fordisc 3* (Jantz and Ousley 2005) to estimate the ancestry for this individual, we would be forced to first select all of the groups with which we wanted to compare the cranium. Sometimes that decision can be informed by the context of the specific case, which may or may not be justified. Perhaps you could eliminate one sex based on the results of other testing, such as using Walker's method for estimating sex from skulls (Walker 2008) and thus only have half of the groups remaining. The problem of which groups to include still remains. As mentioned in the introduction, this issue of group selection could be

addressed by a multi-iteration approach in which all groups are selected and then iteratively removed based on a decision-making criterion such as lowest posterior probability and a typicality probability that does not meet some sort of significance threshold (e.g., $p < 0.05$ or $p < 0.01$). At each iteration, you could make note of how the individual classified, but with each iteration, the model parameters change and it is not unlikely to get different classification results among each run. How should we handle and report these flip flopping classifications? If the answer is "based on experience", then why bother reporting any quantitative measures of certainty about the classification? Regardless of how the test is setup, in my experience the reporting is boiled down to a final classification with different metrics that inform our confidence, be it model performance as total correct classification, posterior, and/or typicality probabilities, or all three, but nothing about how the individual under examination fits into these prescribed labels or how they might be unique within them.

Discussion

I believe that this is a more useful framework for viewing and studying human variation, particularly when it comes to estimating ancestry. Given this workflow, we learn about what makes groups and individuals unique within a relational context. The end result, then, is a series of statements that we can make about the individual under examination, as well as the results of their classifications to examine their overall relationship to all that we know (which here is just limited to variation observed within the Howells dataset and the craniometric variables I chose to retain).

Clearly using Howells dataset, the variables tend to align themselves geographically, thus demonstrating a component of spatial autocorrelation being captured by the Howells group

labels. However, if you read into the details of each of the samples studied by Howells, then you will know that there has to be some sort of temporal component, but this is much harder to tease out the relative effect because not all samples are contemporaneous. It is interesting to pay attention to how the inter-Howells group label relationships change at different scales of Morphogroup subcluster identification (Figure 4). Why these changes have occurred are outside of the scope of this paper, except to point out that you *can* and *should* explore this variation. Maybe it is related to language group, religious preference, temporally appropriate socioeconomic status, diet, etc. The main point here is that our understanding of human variation is based on the relationships among individuals and populations and not in a vacuum. We are trying to boil down the entirety of an individual's evolutionary and developmental processes into some sort of generic label and that will always be rife with uncertainty.

In using the suggested workflow, a practitioner can negate the need for having to preselect groups for comparison and thus avoid explaining away the decisions that were made. Instead, we can provide a descriptive account of how the individual under examination fits into all we know. Given the test example above, instead of reporting "European" or even "Zalavar" male, we now have much more information about how this individual fits into the observed variation at different scales. For an identification effort, it seems to me that being able to say the individual classified into a Morphogroup subcluster composed of only European populations with intermediate cranial sizes overall, but is longer and narrower than many of the comparison samples is a more holistic approach that utilizes both similarities and differences to all that we know.

The clear spatial component to the Howells dataset can help us understand the geographic repercussions of the group label used, but how might one work with modern populations who are

not as bounded by geography? What about combining samples from time periods that had vastly different standards of living? If I have a combined dataset of individuals from the Terry Collection and the Bass Donated Collection, what it means to be a White male is wildly different because the individuals who comprise these samples had very different life histories that inform their biology differently, but we are labeling them the same. I do not find any issue in amassing datasets this way. I take issue with basing strong conclusions on these types of data because the label that is being used cannot possibly mean the same thing to every individual within the dataset with that particular label.

Other Possibilities

Throughout the course of preparing this thought experiment, I have identified several ways in which this could be done differently. For example, you could ignore the group labels until the very end and examine the proportion of individuals from each Howells group that have clustered into their respective Morphogroups, which then is truly letting the data describe itself. You could choose to use different clustering methods to tease out different information from the same dataset, or different classification procedures. The point is, this framework opens up so many more research threads, just by admitting that we do not know as much as we think we do about our data.

It is important to keep in mind that the relationships identified here are only based on craniometric variables. What if we also included morphoscopic features? Odontometrics? Dental morphology? Isotopes? Imagine we had all of this information for all of the individuals within our dataset. How might observed relationships change based on more information? It's hard to say, but I would imagine there would be substantial overlap on what it can tell you with

each additional variable set also showing something new or unique. This type of framework for studying variation will always teach you something new or novel about your data, and, I believe, offer richer results that we can say something about.

Allowing the data to partition itself is not something that could be used exclusively in ancestry estimation. Having the data to partition itself at multiple scales allows for the researcher to identify blunt, major variation at lower resolution scales and more nuanced variation within it. The workflow employed for ancestry estimation above can be easily adjusted for age estimation by first allowing the data to identify natural clusters and then examine what the individuals within those clusters have in common. Say the data partitions itself into roughly "young" "middle aged" and "old" populations; now you can examine the variation within each of these in a much more nuanced way that could lead to much more precise estimates. Importantly, you learn something unique about your problem and dataset, or the limitation of the variables you are using to try to capture variation. This aforementioned example could also be easily tailored to stature estimation in the same way.

Conclusions

I do not feel that forensic anthropology has a "race" problem; our problem is we have become stuck redoing the same types of research only dressed up with fancier analytics. Evolutionary theory provides all the components to understand the complexities of identity and phenotypes, however, the same complexities have not been embraced in our analytical tools or methodological frameworks. Essentially, we stopped exploring the boundaries of possibility within our domain. Human variation is dynamic and so is the act of acquiring knowledge; therefore, if we want to continue learning, we cannot treat human variation as static. Our

understanding of human variation is always relative to what we know and as we learn more, we need to recontextualize what it is we think we know.

For the purpose of this paper, recontextualizing what we know means to fully accept that the way we are labeling data is both meaningless and (potentially) meaningful. We can only detect meaning based on observed relationships in our dataset combined with our current knowledge and understanding. The study of human variation will inevitably become more complex as once disparate peoples have increased access to one another and our ability to study it will always be bounded by our methods and workflows. It is on us to help one another get better at rethinking and redefining what is possible, to challenge what we know, and to learn something new.

Acknowledgments

Thank you to Drs Joel Irish, Alexandra Klales, Bridget Algee-Hewitt, and Kyra Stull for reviewing earlier drafts and all of their suggestions, I am indebted to you all. Additionally, many thanks to the two anonymous reviewers, your suggestions and edits were incredibly helpful.

Received 10 February 2021; accepted for publication 26 June 2021.

Literature Cited

Algee-Hewitt, B. F.B. 2016. Population inference from contemporary American craniometrics. *Am. J. Phys. Anthropol.* 160:604–624.

Berg, G. E., and M. W. Kenyhercz. 2017. Introducing human mandible identification [(hu)MANid]: A free, web-based GUI to classify human mandibles. *J. Forensic Sci.* 62:1,592–1,598.

Bethard, J. D., and E. A. DiGangi. 2020. Letter to the editor—Moving beyond a lost cause: Forensic anthropology and ancestry estimates in the United States. *J. Forensic Sci.* 65:1,791–1,792.

Howells, W. W. 1996. Howells' craniometric data on the internet. *Am. J. Phys. Anthropol.* 101:441–442.

Kenyhercz, M. W., and G. E. Berg. 2017. Evaluating mixture discriminant analysis to classify human mandibles with (hu)MANid, a free, R-based GUI. In *New Perspectives in Forensic Human Identification*, K. E. Latham, E. J. Bartelink, and M. Finnegan, eds. London: Academic Press, 35–44.

Kenyhercz, M. W., A. R. Klales, and W. E. Kenyhercz. 2014. Molar size and shape in the estimation of biological ancestry: A comparison of relative cusp location using geometric morphometrics and interlandmark distances. *Am. J. Phys. Anthropol.* 153:269–279.

Kenyhercz, M. W., A. R. Klales, C. W. Rainwater et al. 2017. The optimized summed scored attributes method for the classification of U.S. Blacks and whites: A validation study. *J. Forensic Sci.* 62:174–180.

Klales, A. R., and M. W. Kenyhercz. 2014. Morphological assessment of ancestry using cranial macromorphoscopics. *J. Forensic Sci.* 60:13–20.

Livingstone, F. B. 1962. On the non-existence of human races. *Curr. Anthropol.* 3:279–281.

McDowell, J. L., M. W. Kenyhercz, and E. N. L'Abbé. 2015. An evaluation of nasal bone and aperture shape among three South African populations. *Forensic Sci. Int.* 252:181.e1–181.e7.

McDowell, J. L., E. N. L'Abbé, and M. W. Kenyhercz. 2012. Nasal aperture shape evaluation between Black and white South Africans. *Forensic Sci. Int.* 222:397–403.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber et al. 2002. Genetic structure of human populations. *Science* 298:2,381–2,385.

Sauer, N. J. 1992. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? *Soc. Sci. Med*. 34:107–111.

Stull, K. E., E. J. Bartelink, A. R. Klales et al. 2020. Commentary on: Bethard JD, DiGangi EA. Letter to the editor—Moving beyond a lost cause: Forensic anthropology and ancestry estimates in the United States. J Forensic Sci. 2020;65(5):1791–2. doi: 10.1111/1556-4029.14513. *J. Forensic Sci.* 65:1,791–1,792.

Stull, K. E., M. W. Kenyhercz, M. L. Tise et al. 2016. The craniometric implications of a complex population history in South Africa. In *Forensic and Bioarchaeological Perspectives on Biological Distance*, M. A. Pilloud and J. T. Hefner, eds. London: Academic Press, 245–264.

Appendix A

```
htestpop<-howelltest[1:3] ## selecting just demographics from the howelltest file
htestcran<-howelltest[8:27] ## selecting just the craniometric variables under analysis
htest<-cbind(htestpop, htestcran) ## combining the demographics with craniometrics into a new
data frame
test<-htest[1,] ## selecting the first row of the test set, a Zalavar male
test$GS<-(test$GOL*test$BBH*test$XCB) ## creating the GS variable for the test individual
test ## checking to make sure everything is alright
test<-test[4:24] ## selecting just the columns with the craniometric variables for the predict
functions
test ## double-checking everything is alright
rf_target<-"Morphogroup"
rf_data<-as.data.frame(hwlr)
nobs<-nrow(rf_data)
rf_data[5:25]<-rf_data[5:25]
rfsamp<-rftrain<-sample(nrow(rf_data), 0.7*nobs)
rfvalidate<-sample(setdiff(seq_len(nrow(rf_data)), rftrain), 0.15*nobs)
rftest<-setdiff(setdiff(seq_len(nrow(rf_data)), rftrain),rfvalidate)
cranio<-names((rf_data[5:25]))
rf1<-randomForest::randomForest(Morphogroup ~ ., data=rf_data[rfsamp,c(cranio,
rf_target)],ntree=1000, mtry=5, importance=TRUE,
na.action=randomForest::na.roughfix,replace=FALSE)
pr<-predict(rf1, newdata=na.omit(rf_data))
ct<-table(pr, rf_data$Morphogroup)
caret::confusionMatrix(ct, reference = rf_data$Morphogroup)
predict(rf1, newdata = test, type="response") ##using predict with the random forest model to
show classification
predict(rf1, newdata = test, type="prob") ## using predict function with random forest model to
show posterior probabilities
mg3mean<-aggregate(.~Morphogroup, hwlr, mean) ## aggregating the means for each
morphogroup
test-mg3mean[6:26] ##subtracting the craniometric values from the test individual from the
mean of Morphogroup 3.

rf_target<-"Morphogroup"
rf_data_1<-as.data.frame(hwl_c3)
nobs<-nrow(rf_data_1)
rf_data_1[4:24]<-rf_data_1[4:24] ##data scaled.
rfsamp<-rftrain<-sample(nrow(rf_data_1), 0.7*nobs)
rfvalidate<-sample(setdiff(seq_len(nrow(rf_data_1)), rftrain), 0.15*nobs)
rftest<-setdiff(setdiff(seq_len(nrow(rf_data_1)), rftrain),rfvalidate)
cranio<-names((rf_data_1[4:24]))
rf1<-randomForest::randomForest(Morphogroup ~ ., data=rf_data_1[rfsamp,c(cranio,
rf_target)],ntree=1000, mtry=5, importance=TRUE,
na.action=randomForest::na.roughfix,replace=FALSE)
```

```
pr<-predict(rf1, newdata=na.omit(rf_data_1))
ct<-table(pr, rf_data_1$Morphogroup)
caret::confusionMatrix(ct, reference = rf_data_1$Morphogroup)
predict(rf1, newdata = test, type="response")
predict(rf1, newdata = test, type="prob")
mg3.2mean<-aggregate(.~Morphogroup, hwl_c3, mean)
mg3.2mean<-mg3.2mean[2,] ##selecting the means for Morphogroup 3.2 only
test-mg3.2mean[5:25]

rf_target<-"Morphogroup"
rf_data_1<-as.data.frame(hwlr_c3.2)
rf_data_1[4:24]<-rf_data_1[4:24]
nobs<-nrow(rf_data_1)
rfsamp<-rftrain<-sample(nrow(rf_data_1), 0.7*nobs)
rfvalidate<-sample(setdiff(seq_len(nrow(rf_data_1)), rftrain), 0.15*nobs)
rftest<-setdiff(setdiff(seq_len(nrow(rf_data_1)), rftrain),rfvalidate)
cranio<-names((rf_data_1[4:24]))
rf1<-randomForest::randomForest(Morphogroup ~ ., data=rf_data_1[rfsamp,c(cranio,
rf_target)],ntree=1000, mtry=5, importance=TRUE,
na.action=randomForest::na.roughfix,replace=FALSE)
pr<-predict(rf1, newdata=na.omit(rf_data_1))
ct<-table(pr, rf_data_1$Morphogroup)
caret::confusionMatrix(ct, reference = rf_data_1$Morphogroup)
predict(rf1, newdata = test, type="response")
predict(rf1, newdata = test, type="prob")
mg3.2.2mean<-aggregate(.~Morphogroup, hwlr_c3.2, mean)
mg3.2.2mean<-mg3.2.2mean[2,] ##selecting the means for Morphogroup 3.2.2 only
test-mg3.2.2mean[5:25]

rf_target<-"Population"
rf_data_1<-as.data.frame(c3.2.2)
rf_data_1[4:24]<-rf_data_1[4:24]
nobs<-nrow(rf_data_1)
rfsamp<-rftrain<-sample(nrow(rf_data_1), 0.7*nobs)
rfvalidate<-sample(setdiff(seq_len(nrow(rf_data_1)), rftrain), 0.15*nobs)
rftest<-setdiff(setdiff(seq_len(nrow(rf_data_1)), rftrain),rfvalidate)
cranio<-names((rf_data_1[4:24]))
rf1<-randomForest::randomForest(Population ~ ., data=rf_data_1[rfsamp,c(cranio,
rf_target)],ntree=1000, mtry=5, importance=TRUE,
na.action=randomForest::na.roughfix,replace=FALSE)
pr<-predict(rf1, newdata=na.omit(rf_data_1))
ct<-table(pr, rf_data_1$Population)
caret::confusionMatrix(ct, reference = rf_data_1$Population)
predict(rf1, newdata = test, type="response")
predict(rf1, newdata = test, type="prob")
mg3.2.2.popmean<-aggregate(.~Population, c3.2.2, mean)
```

```
mg3.2.2.popmean<-mg3.2.2popmean[3,]
test-mg3.2.2.popmean[4:24]

rf_target<-"Sex"
rf_data_1<-as.data.frame(c3.2.2)
rf1<-randomForest::randomForest(Sex ~ ., data=rf_data_1[rfsamp,c(cranio,
rf_target)],ntree=1000, mtry=5, importance=TRUE,
na.action=randomForest::na.roughfix,replace=FALSE)
pr<-predict(rf1, newdata=na.omit(rf_data_1))
ct<-table(pr, rf_data_1$Sex)
caret::confusionMatrix(ct, reference = rf_data_1$Sex)
predict(rf1, newdata = test, type="response")
predict(rf1, newdata = test, type="prob")
mg3.2.2.sexmean<-aggregate(.~Sex, c3.2.2, mean)
mg3.2.2.sexmean<-mg3.2.2.sexmean[2,]
test-mg3.2.2.sexmean[4:24]
```

Table 1. Craniometric Variables Used in the Current Study

Abbreviation	Variable
ASB	Biasterionic breadth
AUB	Biauricular breadth
BBH	Basion-Bregma height
BNL	Basion-Nasion length
BPL	Basion-Prosthion length
GOL	Glabello-Occipital legnth
GS	Gross size
JUB	Bijugal breadth
MAB	Palate breadth
MDB	Mastoid breadth
MDH	Mastoid height
NLB	Nasal breadth
NLH	Nasal height
NOL	Nasio-Occipital length
NPH	Nasion-Prosthion height
OBB	Orbital breadth
OBH	Orbital height
WCB	Minimum cranial breadth
XCB	Maximum cranial breadth
XFB	Maximum frontal breadth
ZYB	Bizygomatic breadth

Table 2. Mean PC Sore by Morphogroup 1 Subcluster

Morphogroup	PC1	PC2	PC3
1.1	-0.291	0.465	-0.521
1.2	0.352	-0.113	-0.589
1.3	1.371	-1.124	0.367
1.4	-0.853	-0.061	0.922
1.5	-0.239	0.153	0.703

Figure Captions

Figure 1. Dendrogram based on hierarchical clustering depicting the three Morphogroups identified via NbClust. Morphogroup 1 is encapsulated by the red box, Morphogroup 2 is encapsulated by the green box, and Morphogroup 3 is encapsulated by the blue box.

Figure 2. Dendrogram based on hierarchical clustering depicting the five subclusters within Morphogroup 1 identified via NbClust. Subcluster 1 is shown in the purple box, 2 by the dark blue, 3 the red, 4 the light blue, and 5 the green.

Figure 3. Flowchart depicting correct classifications for each model. Boxes represent classification models and show the total correct classification by Morphogroup and Sex. The line segments show the correct classification for that Morphogroup from the model from which it extends. Howells population labels are provided underneath Morphogroup subclusters when no further natural clusters could be identified.

Figure 4. Dendrogram from Figure 1 with final Morphogroup subcluster identification added to each Howells group.

Figure 1.

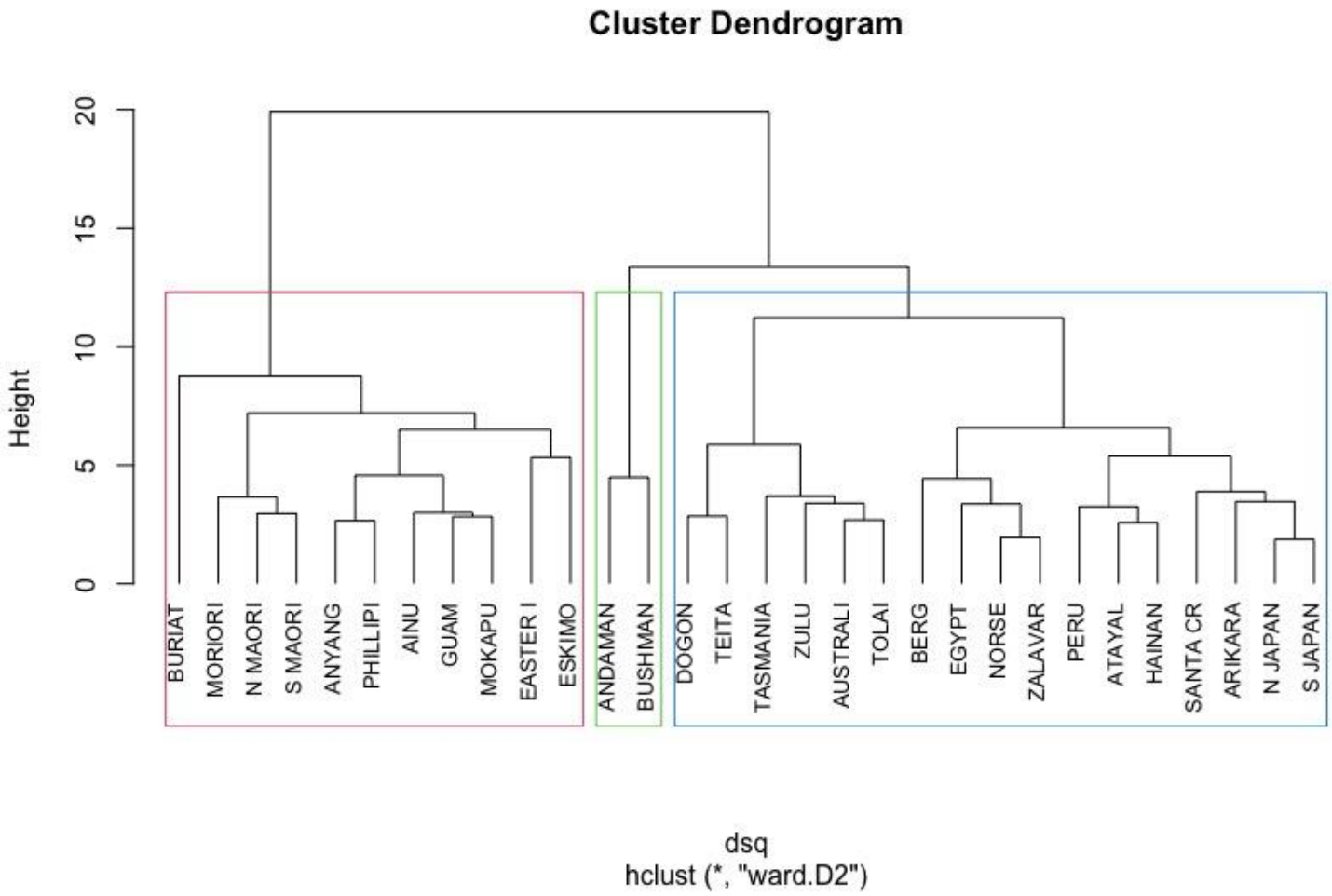

Cluster Dendrogram
Height
0
5
10
15
20
BURIAT
MORIORI
N MAORI
S MAORI
ANYANG
PHILLIPI
AINU
GUAM
MOKAPU
EASTER I
ESKIMO
ANDAMAN
BUSHMAN
DOGON
TEITA
TASMANIA
ZULU
AUSTRALI
TOLAI
BERG
EGYPT
NORSE
ZALAVAR
PERU
ATAYAL
HAINAN
SANTA CR
ARIKARA
N JAPAN
S JAPAN
dsq
hclust (*, "ward.D2")

Figure 2.

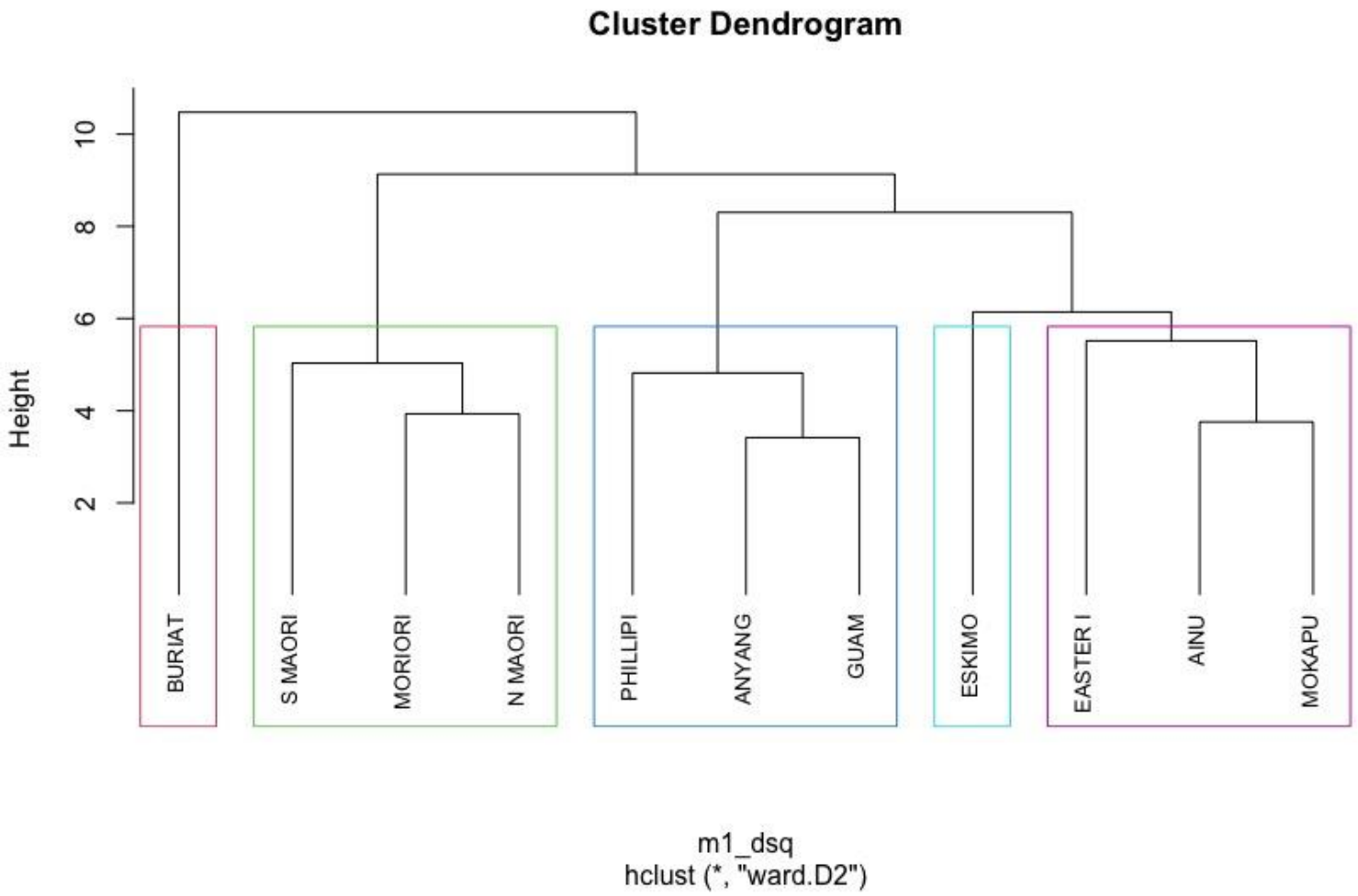

Cluster Dendrogram
Height
10
8
6
4
2
BURIAT
S MAORI
MORIORI
N MAORI
PHILLIPI
ANYANG
GUAM
ESKIMO
EASTER I
AINU
MOKAPU
m1_dsq
hclust (*, "ward.D2")

Figure 3.

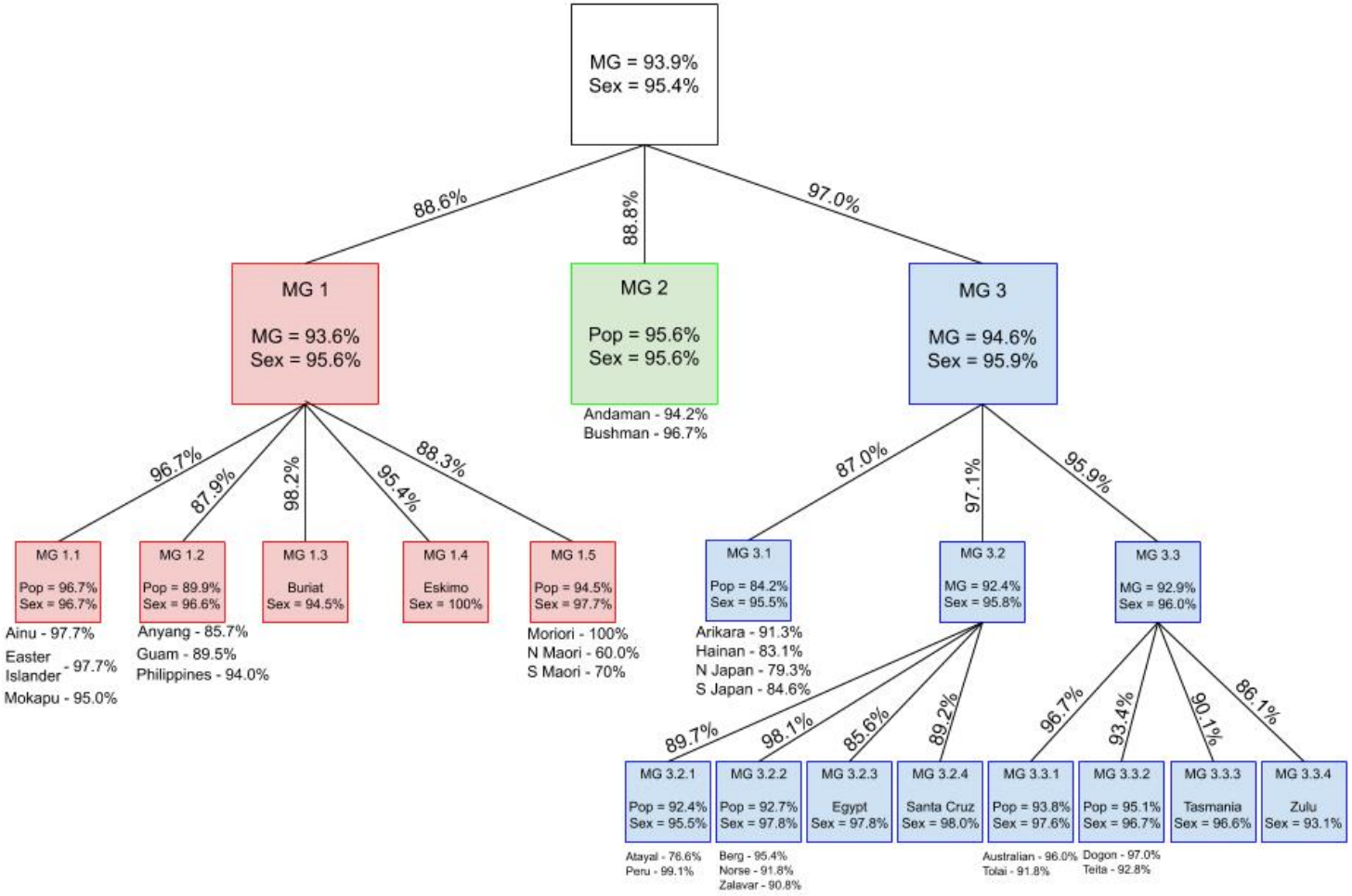
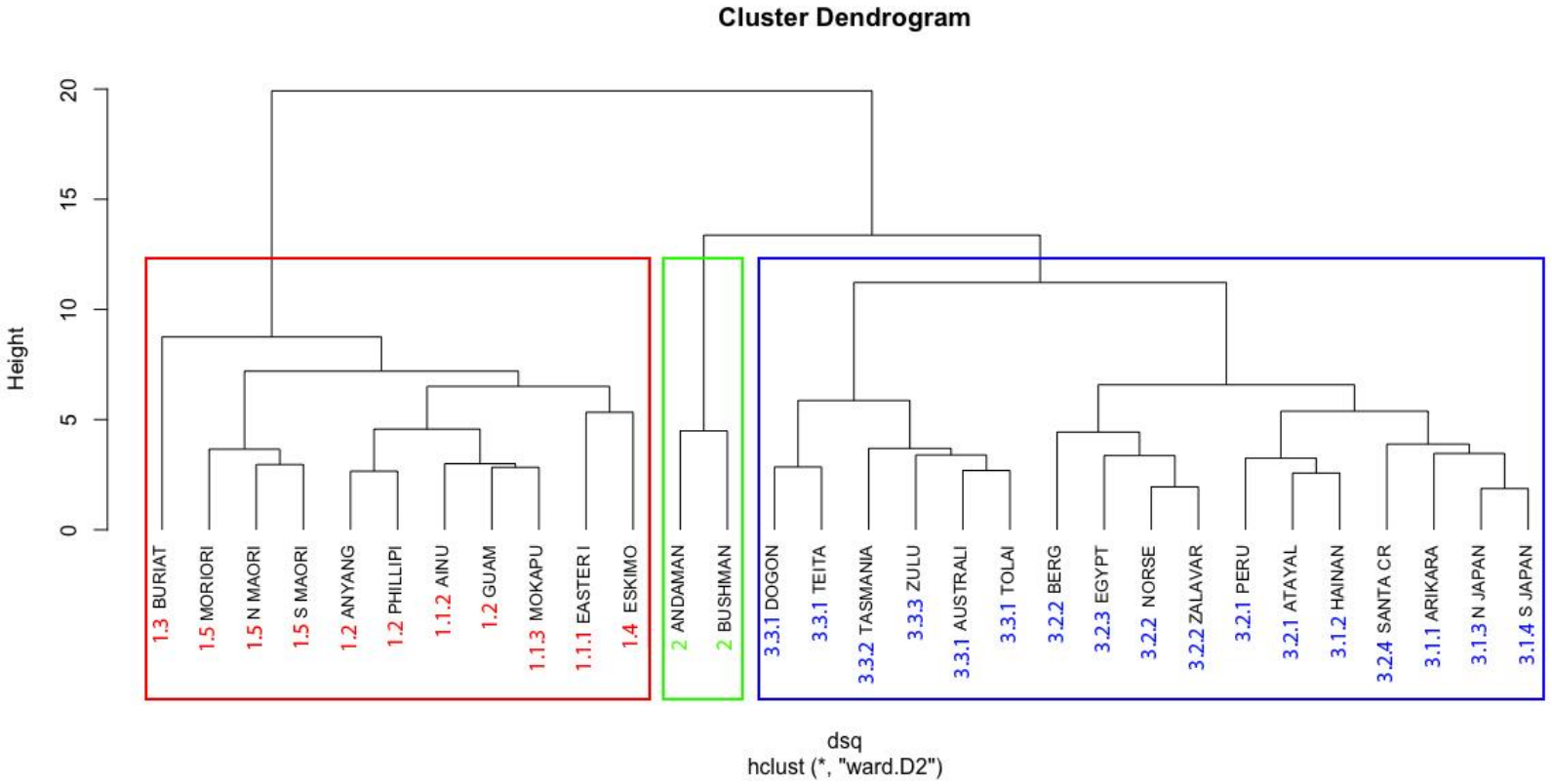

MG = 93.9%
Sex = 95.4%
88.6%
88.8%
97.0%
MG 1
MG = 93.6%
Sex = 95.6%
MG 2
Pop = 95.6%
Sex = 95.6%
Andaman - 94.2%
Bushman - 96.7%
MG 3
MG = 94.6%
Sex = 95.9%
96.7%
87.9%
98.2%
95.4%
88.3%
87.0%
97.1%
95.9%
MG 1.1
Pop = 96.7%
Sex = 96.7%
MG 1.2
Pop = 89.9%
Sex = 96.6%
MG 1.3
Buriat
Sex = 94.5%
MG 1.4
Eskimo
Sex = 100%
MG 1.5
Pop = 94.5%
Sex = 97.7%
MG 3.1
Pop = 84.2%
Sex = 95.5%
MG 3.2
MG = 92.4%
Sex = 95.8%
MG 3.3
MG = 92.9%
Sex = 96.0%
Ainu - 97.7%
Easter Islander - 97.7%
Mokapu - 95.0%
Anyang - 85.7%
Guam - 89.5%
Philippines - 94.0%
Moriori - 100%
N Maori - 60.0%
S Maori - 70%
Arikara - 91.3%
Hainan - 83.1%
N Japan - 79.3%
S Japan - 84.6%
89.7%
98.1%
85.6%
89.2%
96.7%
93.4%
90.1%
86.1%
MG 3.2.1
Pop = 92.4%
Sex = 95.5%
MG 3.2.2
Pop = 92.7%
Sex = 97.8%
MG 3.2.3
Egypt
Sex = 97.8%
MG 3.2.4
Santa Cruz
Sex = 98.0%
MG 3.3.1
Pop = 93.8%
Sex = 97.6%
MG 3.3.2
Pop = 95.1%
Sex = 96.7%
MG 3.3.3
Tasmania
Sex = 96.6%
MG 3.3.4
Zulu
Sex = 93.1%
Atayal - 76.6%
Peru - 99.1%
Berg - 95.4%
Norse - 91.8%
Zalavar - 90.8%
Australian - 96.0%
Tolai - 91.8%
Dogon - 97.0%
Teita - 92.8%

Figure 4.


Cluster Dendrogram
Height
0
5
10
15
20
1.3 BURIAT
1.5 MORIORI
1.5 N MAORI
1.5 S MAORI
1.2 ANYANG
1.2 PHILLIPI
1.1.2 AINU
1.2 GUAM
1.1.3 MOKAPU
1.1.1 EASTER I
1.4 ESKIMO
2 ANDAMAN
2 BUSHMAN
3.3.1 DOGON
3.3.1 TEITA
3.3.2 TASMANIA
3.3.3 ZULU
3.3.1 AUSTRALI
3.3.1 TOLAI
3.2.2 BERG
3.2.3 EGYPT
3.2.2 NORSE
3.2.2 ZALAVAR
3.2.1 PERU
3.2.1 ATAYAL
3.1.2 HAINAN
3.2.4 SANTA CR
3.1.1 ARIKARA
3.1.3 N JAPAN
3.1.4 S JAPAN
dsq
hclust (*, "ward.D2")