Human Population History from Discrete Dental Traits Under an Approximate Multivariate Ordinal Probit

Nikolai G. Vetr[1,2,3,*], Shara E. Bailey[4], and Timothy D. Weaver[1]

[1]*Department of Anthropology, University of California, Davis, Storer Hall, E. Quad, Davis, CA 95616, USA;*

[2]*Center for Population Biology, University of California, Davis, Young Hall, One Shields Avenue, Davis, CA 95616, USA;*

[3]*Data Science and Informatics, University of California, Davis, 100 N Quad, Davis, CA 95616, USA;*

[4]*Department of Anthropology, New York University, 25 Waverly Place, New York, NY 10003, USA;*

[*]*E-mail: nlashinsky@ucdavis.edu*

Abstract.—Human dental variation is often used for the inference of population history and phylogeny in paleontological contexts. Teeth are small and hard, and so preserve well where other morphologies of the skeleton degrade. While their gross shapes and sizes likely reflect selective constraint, usually by way of their role in food processing, they're also covered in a panoply of cusps, pits, grooves, and ridges, among other structures, that vary both within and between populations and species. It is this variation that has been discretized and codified in the Arizona State University Dental Anthropology System (ASUDAS), among other expansions by later practitioners. The ASUDAS provides a lens to systematically characterize "minor" dental morphological variation into ordered sets of "quasicontinuous" dental traits, with states corresponding to greater or lesser degrees of expression. Here, we investigate the ability of these characters to retrieve plausible population trees when analyzed under an approximation of multivariate Brownian motion filtered through the multivariate ordinal probit. We further explore the reliability of this approximation at capturing the salient properties of the fuller, less tractable "latent liability" model through an empirically realistic simulation study. [ASUDAS; Human Population History; Discrete Dental Traits; Bayesian Inference; Multivariate Brownian Motion]

INTRODUCTION

Multivariate Character Evolution.—Sewall Wright (1934) first proposed the threshold model of quantitative genetics — also called the quasicontinuous model (**?**) in dental anthropology — to describe the expression of toe number on guinea pig hind feet. In the years since, it's been used to model discrete trait evolution phylogenetically (**???**), where it's also called the latent liability model (**?**) and the multivariate probit model (**?**), the latter of which enjoys an equally lengthy history as Wright's naming (**??**). Whatever its name, in this context the model supposes that the visible expression of a discrete trait is governed by the value of a hidden, continuous, polygenic character called a "liability". For some binary trait, if the liability value of an individual is greater than some threshold value, the corresponding discrete trait is expressed; if less than, not expressed. Meanwhile, for an ordinal trait, when an individual's liability falls between some pair of thresholds, a corresponding discrete trait is expressed. The locations of a set of threshold relative to the population-level distribution of liabilities, then, determines the frequencies of trait expression in that population. When these latent liabilities are determined by the actions of many alleles of small effect, their distribution across individuals within a population becomes Gaussian under the Central Limit Theorem, and if we wish to identify the location of this Gaussian, or the locations of the thresholds, we must fix its variance to some number, by convention 1. This discretization straightforwardly generalizes to multiple traits in a multivariate framework, with population liability distributions described by multivariate Gaussians with some vector of mean liabilities and correlation matrix (once more fixing variances to 1 for identifiability purposes). Instead of the assignment of discrete states emerging from the locations of univariate normal random variables falling within intervals bounded by thresholds, discrete states are instead determined at the individual-level by a liability vector's occupancy of some hypervolume in R^n, bounded by threshold hyperplanes.

Through time, evolutionary processes will cause the location of a population's latent liability distribution to wander. Under neutrality, far from its natural bounds, and at sufficiently high population size, the distribution of a sample mean across subsequent generations will itself be mul-

tivariate normal, and so can be described according to multivariate Brownian motion (mvBM), perhaps acting over a strictly bifurcating, non-reticulate population tree. Under Chevereud's conjecture, the correlation matrix of the within-population multivariate Gaussian distribution of latent liabilities will broadly reflect the additive genetic components of that matrix, which under neutrality will in turn be proportional to the mvBM rate matrix. Thus, we might wish to take as an estimate of the correlations of mvBM the pooled estimate of the within-population latent liability correlation matrix.

Simulating using the threshold model is fairly straightforward – we generate tip means by sampling from the multivariate normal distribution implied by Brownian motion, and then sample individuals or populations from multivariate normal distributioned centered on the location of each of those means, passing individual liabilities through an indicator function to determine their corresponding vector of discrete traits. In this way, we can represent the evolution of a vector of polymorphic traits with variable degrees of expression along a lineage. Working backwards, however, requires that we repeatedly take integrals of multivariate normal distributions in the dimension of however many traits are the subject of analysis (with e.g. the Genz-Bretz algorithm, **?**), or else perform data augmentation over both individual and mean liabilities, neither of which make for an appealing computational prospect. As such, while fitting this model in this work we make several compromises in the name of tractability, described in the *Materials & Methods* section below.

Relation to Other Models and Methods.—The threshold model claims many benefits over what is currently the most commonly used phylogenetic model of morphological evolution, Lewis' Mk model (**?**), which has been shown to outperform heuristic methods such as Maximum Parsimony (MP) across a range of conditions likely to be encountered in real world datasets (**??**), such as high rates of evolution or high rate heterogeneity among characters (**?**), at least when you also simulate under the Mk model. MP itself has a few other marks against it, such as statistical inconsistency when rate inequalities (heterotachy) exist between lineages, in part due to its disregard for branch lengths, as traits can only change once on any given branch, and its lack of rigorous means for

deciding between alternative implementations of parsimony and between most parsimonious trees (**?**), making it difficult to parse which clades are more or less confidently supported. MP also struggles to easily accommodate uncertainty in the data or non-independence between traits. Furthermore, while not model-based *per se*, particular implementations of MP can be shown to be equivalent to certain explicit models of character change which themselves don't seem too appealing; e.g. Fitch parsimony (**?**) always picks the same trees as the TS97 model (**?**), which, if branches have the same length for all traits, is equivalent to the Mk model (**??**).

The Mk model generalizes the simplest of the GTR family of continuous time Markov chain (CTMC) models of molecular evolution, JC69 (**?**), which can be seen as a special case of the Mk model where k=4. These rates do not change throughout the tree and are the same for all characters (though among-character rate heterogeneity can be accommodated here, too, by discretizing a gamma distribution and drawing rates from each bin; **?**), and any particular set of entries into the rate matrix can be used to calculate the likelihood of a particular set of tip outcomes given a tree with branch lengths using Felsenstein's **?** Pruning Algorithm. Unlike the threshold molde, the Mk model does not allow for polymorphism within a lineage, instead requiring that we assign tips to particular discrete states. Polymorphism, meanwhile, is a common feature of discretely coded traits, especially in those catalogued in the ASUDAS (**?**), described below. Another plausibly desirable property of the threshold model involves the frequencies of trait expression changing rapidly when they are intermediate, but more slowly once at the extremes, and slower still in expectation if they've been extreme for a large period of time. Consider, for example, a binary character – if approximately half a population expresses one state and half the other state, the mean liability is very close to the threshold, and every shift will have a large effect (as the density of a normal distribution is, of course, greatest at its center). Meanwhile, if a population has been monomorphic in some state for a long while, the liability distribution may have wandered quite far from the threshold indeed, and isn't likely to return to it any time soon. This property may capture a desirable facet of biology – populations that are split in their expression of some trait seem like they could drift this way or that, whereas populations that have uniformly expressed

some trait over long periods of time are unlikely to soon change in their frequencies (perhaps due to constraints imposed by other traits that have evolved since).

Additionally and despite the caveats mentioned above, it is far easier to accommodate correlated evolution under the threshold model by incorporating covariances into our model of multivariate Brownian motion. It is also possible to accommodate correlated evolution in an instantaneous rate model (**??**), but with far worse scaling at high dimension, requiring a kn x kn instantaneous rate matrix for k traits, which quickly becomes unwieldy (consider two binary traits – instead of having to only model changes $0 \leftrightarrow 1$, you need to model $01 \leftrightarrow 00 \leftrightarrow 10 \leftrightarrow 11$, $10 \leftrightarrow 01 \leftrightarrow 11$, and $00 \leftrightarrow 11$). Alternative approaches exist (**???**), but have yet to be thoroughly explored in the context of morphological evolution. Finally, the threshold model has been invoked to explain the expression of traits in the Arizona State University Dental Anthropology System (ASUDAS) (**?**) before, so there exists precedent in applying it to that suite of traits (**?**).

Discrete Dental Traits.—ASUDAS traits represesent a common material for the inference of both human population history (**???**) and hominin phylogeny (**??**), though the latter may benefit from typologies better able to capture nonmetric dental variation across species (**??**). Due to their high mineral content, overall hardness, and small size, teeth preserve especially well in the fossil record, their variability examined and used for inference in many other paleontological contexts, as well as for neontological forensic applications (**?**). The majority of them are scored on an ordinal scale, but are often dichotomized into presence / absence for use in analysis (e.g. **?**), as they would necessarily be for the basic, single threshold model described above. Genetically, many appear to follow threshold-like patterns of inheritance, with high positive associations between trait incidence and expressivity within populations (**?**), and they are frequently treated as such (e.g. **?**). Complex segregation analysis accepts a quasicontinous, polygenic model for many of the discrete dental traits hitherto considered (**?**), and to date not a single dental trait has been found to have simpler genetic architecture (**?**). ASUDAS traits appear to broadly track neutral patterns of human genetic variation (**??**), and so may well fit a multivariate Brownian model of character evolution on the underlying latent liability scale. However, many adaptive explanations have

been proposed for ASUDAS traits, typically invoking mechanical advantage during mastication, resilience to attrition, mate atraction and social signaling, and sundry other benefits (**?**). To the extent that selection is fluctuating, Brownian motion may provide an adequate fit to these data, but exploration of other stochastic processes better able to capture directional phenomena may yield conflicting results. Finally, the evolution of the mammalian dentition is not characterized by independence between characters (**?**), and so its study would benefit from a principled accounting of non-independence. For these reasons, it is precisely a collection of discrete dental characters collected on a set of globally distributed human populations that form the empirical focus of this work.

MATERIALS AND METHODS

Empirical Data.—The data used here are a collection of 722 individual-level samples of 137 discrete dental traits collected on a globally distributed set of human populations assigned to the groups Neandertal, Oceanian, European, West Asian, South Asian, Northeast Asian, Sub-Saharan African, and American. Pooling was done at this level and not with a finer grain to ensure adequate sample sizes across populations. Initially, all teeth across both upper and lower dentitions were represented in this work, though as not all traits were scored on all teeth for all populations, data were subsequently filtered to ensure stable estimation of population mean liabilities. When possible, the right side of the mouth was used for scoring. To minimize the effects of interobserver error, all dental traits were scored by Shara Bailey (SB) with reference to ASUDAS dental plaques. As the within-population expression of ASUDAS traits shows minimal sexual dimorphism (**?**), sexes were pooled for this analysis. Despite the ubiquity of dichotomization in studies of ASUDAS traits, we chose not to split traits into discrete binary presence / absence categories, both to avoid introducing further researcher degrees of freedom with respect to breakpoint selection, and because preliminary analysis of simulated data showed that the recovery of population means could be much more reliably performed with multistate characters than with binary ones.

Data Filtration.—Before data analysis could begin, several preprocessing steps were performed to ensure both the data's compatibility with the inference model, as well as to identify traits with insufficient observations for stable estimation within an optimization framework. First, all non-binary, non-ordinal characters were removed from consideration. It is possible to model unordered character evolution with a threshold model by positing the action of multiple, coevolving liabilities, but we chose not to do so here. Some traits in the dataset, such as those corresponding to premolar lingual cusp (PLC) variation, were scored on an ordinal scale that included additional information regarding non-ordinal character states. This additional information was discarded, as we collapsed PLC scores into ordinal categories corresponding to 1, 2, and 3 cusps.

At first pass, we examined patterns of missingness in the raw data, noting how many traits

were present in how many individuals, as well as how many individuals were present in how many traits (figure 1a-b). Subsequently, we plotted the number of traits present in some number of individuals in at least some number of populations (figure 1c). Noting a horizontal stretch followed by a sharp inflection downward in this figure, we additionally filtered traits that were not represented in at least 6 populations by at least 8 individuals. Ultimately, 118 traits across 684 individuals and 8 populations were included in the final analysis, though only 34% of the entries in this alignment were unambiguously scored, with 65% missing entirely and 1% scored with ambiguity codes. Additional information regarding the composition of these data, including population-specific sample sizes for each trait, can be found in Supplemental Appendix I.

Hierarchical Phylogenetic Likelihood.—The full phylogenetic likelihood of an ordinal discrete character alignment at the individual level whose group mean liability vectors evolve on a tree according to multivariate Brownian motion can be given by two distributions, the first describing the evolution of those means on a phylogeny with some branch lengths and rate matrix, and the second describing the distribution of individual-level character vectors under those same means and correlation matrix, which yield each population's individual-level liability distribution, and a set of threshold locations that parameterize the indicator function that transforms each individual's liability vector into a vector of discrete characters according to which hypervolume it's contained in. The former distribution can be given by the usual multivariate normal probability density function, whose mean is the root state (marginalized out by the Felsenstein Pruning Algorithm), and whose covariance matrix is the kronecker product of the phylogenetic covariance matrix, with diagonal entries the height of each tip above the root and off-diagonal entries the sum of shared branch lengths from the root between each pair of tips, and the mvBM rate matrix R, which can further be decomposed into a matrix product SCS, where S is a diagonal matrix of standard deviations (the square roots of each trait's evolutionary rate) and C the correlation matrix describing non-independence in the collection of traits' within-lineage evolutionary trajectories. The latter distribution, meanwhile, is a very high dimension multinomial, whose tip-specific probabilities are given by integrating the hypervolumes of a set of multivariate normal distributions whose

means are tip-specific vectors of mean liabilities and whose covariance matrix is a correlation matrix by Chevereud's conjecture equal to the aforementioned C, and whose bounds of integration are defined by matrices of adjacent thresholds. Where there are d traits each with k thresholds, the multinomial for each tip is described by a vector of probabilities with length $(k+1)^d$, which can be very large for even reasonably small k and d, though one really only needs to compute those probabilities for which one has unique site patterns (vectors of ordinal traits) within-each populations. In this sense, the likelihood can be thought of as the probability mass function of a multinomial, whose bin probabilities are partially determined by a hyperdistribution with phylogenetic structure, though for our purposes here, it is a parameter of the hyperprior (i.e. the topology of the tree) that is focal.

Each tip's mean liability vector is *latent* — unobserved — and so needs to be sampled through data augmentation, its own probability given by the mvBM likelihood function. If tips are monomorphic (i.e., the thresholds are located so far apart and evolutionary rates so high that each lineage's liability distribution spends all its time wandering the interiors of each hypervolume, rather than near its edges), one only needs to ensure each sampled tip liability is within the appropriate hypervolume, i.e. the probability of each tip-wise vector of discrete traits is 1 inside and 0 elsewhere. If all traits are binary, the location of each trait's threshold can be fixed to some arbitrary value, typically 0. But with ordinal traits, one also need to perform inference over the locations of all later thresholds. With polymorphic traits and information at the sub-population level, one could, in principle, extend the data-augmentation strategy to each individual, taking densities of each individual's augmented liability vector in their corresponding tip's multivariate normal, alongside augmenting that tip's mean vector and using a similar indicator function to ensure each individual's liability vector is in the appropriate space. Data augmentation over so many individuals multiplied by equally many of their traits, however, would introduce orders of magnitude more parameters into our inference model, and so such a strategy was quickly deemed computationally infeasible. Instead, we sought to evaluate the integral of each multivariate normal distribution corresponding to each individual in our character alignment. Unfortunately,

multivariate normal integrals have no solution in closed form, and so after exploring various numerical approximations we settled on the transformation and Monte Carlo integration algorithm described by Alan Genz (**?**) and implemented in the function `pmvnorm` in the package *mvtnorm* (**?**)genzPackageMvtnorm2020) in *R* (**?**). This proved efficient and stable over alternatives, but still too slow for our purposes, taking integrals of dimension on the order 10^2 many hundreds of times per single likelihood calculation. Instead, we used a further approximation to this integral, evaluating choose(d, 2) bivariate normal integrals, finding their geometric mean, and rescaling it to the appropriate dimension by taking its square root and raising it to the power of the full dimensionality. This appears to produce a value roughly proportional to that of the true integral (figure 2) while imposing a computational burden many orders of magnitude smaller at high dimension. Though it appears to hold less well at extreme correlations (figure 1b), for our purposes it is only the slope of the relationship that matters, as multiplying all likelihoods by a constant (equivalent to adding or subtracting a value on the log scale) does not distort the relative distances between peaks and valleys on the likelihood surface. We were further able to vectorize these computations using a reimplementation of *mvtnorm* code in the R package *pbivnorm* (**?**). To avoid underflow, all calculations were performed on the log-scale.

Two-Step Algorithm.—As a further concession to computational tractability, we separated the inferential procedure into two steps, in a manner vaguely analogous to sequence alignment and conditioning used in molecular contexts. The first iteratively optimized the locations of tip means, threshold locations, and correlations between traits independent of phylogenetic structure using a bounded form of the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm implemented in the `optim` function in base-R, and the second conditioned on those tip means and between-trait correlations per Cheverud's conjecture to infer phylogeny using the Metropolis-Hastings algorithm to approximate the joint posterior distribution of phylogenetic model parameters in a Bayesian inferential framework. In the latter analysis, we fixed the correlation components of the mvBM rate matrix and inferred only its rates, as well as branch lengths — in which rate and time are confounded, as we specified no explicit morphological clock model here — and topology. Both

steps of these analyses are described in turn below.

Iterative Optimization.—In preliminary simulated contexts, we found that we were able to reliably retrieve data-generating values of population mean liabilities, between-trait correlations, and threshold-locations by iterating through each model parameter and maximizing the probability of observing the ordinal observations its change affected. Additionally, because we approximated the full multivariate normal integral as a product of bivariate normal integrals, we only had to optimize those components of the overall function that the current parameter affected, drastically reducing our overall computational burden. Thus, we iterated through each individual (tip, trait) mean liability, constrained along the real number line; each correlation parameter, constrained between $(-1, 1)$, and each vector of distances between thresholds, which were constrained to be positive over $(0, \infty)$ to ensure that threshold locations were monotonic increasing for each subsequent ordinal state. This iterative optimization was random with respect to the order of parameters whose values were to be maximized, and proceeded for a sufficiently many rounds until parameter values converged onto some stable set, typically within 6-8 rounds of optimization. Because we performed stochastic imputation over missing values, model parameters never truly converged, and so we stopped the algorithm after a dozen rounds and took as our final estimate the arithmetic mean of model parameters across four independent runs.

To regularize parameters away from extreme values (e.g. means from $\pm\infty$ when discrete states are at their maximal or minimal states and invariant within a tip, a problem long-recognized in the context of probit models, **?**), several forms of regularization were used, i.e. penalties to the likelihood function over which we were performing optimizations, analogous to priors in a Bayesian inferential context. For correlations, we used a Beta(10,10) penalty, adding to our log-likelihood the log-density of our correlation in a Beta distributions stretched to the (-1,1) range with shape parameters equal to 10. Attempting to also optimize the magnitude of these shape parameters resulted in a singularities over the likelihood surface that drew each shape parameter to $+\infty$ and each correlation to 0, even with highly informative hyperpenalties on each shape parameter. The marginal distribution of each correlation parameter implied by a flat LKJ(1) was

also considered, but judged too aggressive, as it would give each shape parameter a value of $\eta - 1 + d/2 = 59$ for our 118 traits, which was entirely too difficult to overcome with the information contained in our dataset. Instead, shape parameters equal to 10 could be interpreted to imply modularity between packages of 20 traits at a time, which seemed appropriate for the 4 types of tooth and 8 teeth per quadrant found in the human dentition, as 118 traits / mean(4, 8) $\approx$ 20. To regularize the strictly positive spacings between adjacent thresholds, an exponential distribution was used, whose rate parameter λ was itself optimized during each round of optimization and constrained to $(0, \infty)$. To regularize means, we used a univariate Brownian motion process acting on a tree with constant rates, whose rates x branch length product was itself optimized over $(0, \infty)$. Univariate Brownian motion was used due to possible instability in the correlation matrix over the earlier rounds of iterative optimization. Mean estimates were highly insensitive to the shape of tree used, be it a star phylogeny or different varieties of distance tree. We found this reassuring in light of our adopted two-step approach, that most of the information regarding the locations of tip mean liabilities could be found in the individual-level data, rather than in the structure of the tree. As such, final analyses used star phylogenies to regularize means, in order to not double-count whatever phylogenetic signal might be found in the means themselves.

Output from the algorithm was also insensitive to parameter values used for initialization, be they cleverly chosen (e.g. to their analytically solvable expected univariate values, or Pearson correlation coefficients thereof), neutrally chosen (e.g. the identity for a correlation matrix, the origin for means, and values of 0.5 for each threshold spacing) or randomly chosen (e.g. a sample from an LKJ(1), in the case of correlations, from samples from an exp(1), in the case of threshold spacings, or means from uniform between the maximum and minimum thresholds).

Individual-level data in the discrete character alignment were both partially and wholly missing. Data that were wholly missing lacked an observation for that (individual, trait); observations for partially missing data, meanwhile, were coded with one of three ambiguity codes: a number followed by a +, indicating states $\geq$ than the supplied state; a number followed by a -, indicating states $\leq$ than the supplied state; and two adjacent numbers separated by a period, indicating that

either state could judged appropriate in that instance. Additionally, data were thought to be plausibly missing not at random but instead in a state-dependent manner, for example, with larger cusps or deeper grooves harder to obliterate through dental wear processes, or else for more robust teeth to be harder to lose due to mechanical strain or tooth decay. As such, we required an algorithm to impute missing values that could be Missing Not At Random (MNAR), lest we bias our inference of population means and artificially conflate convergence in the processes that give rise to state-dependent missingness for evidence of shared dental ancestry.

Stochastic MNAR Imputation.—If the data were Missing Completely At Random (MCAR), one could envision cheaply sampling missing states from their conditional probabilities, Pr(state | individual, trait, population parameters): given the current values preferred by the iterative optimization algorithm for each tip mean, correlation matrix, and threshold locations, what are the probabilities for observing each possible state at a particular missing index? One could expensively imputing these values on an individual or population-wide scale, though combinatorial difficulties quickly arise, even with modest numbers of traits, individuals, and missing values. However, for MNAR data, this is insufficient, as not all states are equally likely to have been rendered missing, and we instead desire Pr(state | individual, trait, population parameters, missing) to sample from. Thus, an estimate of Pr(missing | state) is required, the compromise of which with Pr(state | individual, trait, population parameters) can be easily found by rote application of Bayes' theorem.

For the former probability, we simply evaluate our approximation to the multivariate normal integral across all the possible states a particular trait can take in that individual, conditional on all the other traits also observed in that individual. We then divide these by their sum to ensure they equal one. For the latter, we count up all the observed states for a particular trait across all the individuals in our sample, and then, knowing the multinomial distribution of these states marginal of all the other traits, find the conditional distribution of the unobserved states, conditional on the vector of states already observed. Rather than sample from this distribution and take the raw $n_{missing}/(n_{missing} + n_{observed})$ as our estimate of Pr(missing | state) we further regularize

by computing the expectation of this conditional distribution of unobserved states and using it, as well as the observed counts, to update a flat beta distribution, from which we sample a Pr(missing | state). To find the expected count of the unobserved component of a multinomial distribution, we initially rejection sample draws from the multinomial distribution according to the Monte Carlo method until we produce 500 state vectors compatible with the observed component. In cases where the observed states are highly incompatible with the current means, correlations, and thresholds, Monte Carlo simulation and rejection sampling becomes highly inefficient, and we instead use the Metropolis algorithm to approximate this distribution with a stopping rule such that every nonzero, state-specific difference from the observed component needs to have an effective sample size (ESS) of at least 500, which we compute using the *CODA* package (**?**) in R.

We then weigh state conditional probabilities by Pr(missing | state) and divide by their weighted sum, sampling states for these missing values according to the calculated state-specific probabilities, conditional on missingness, the observed states at other traits in that individual, and all other model parameters. For partially missing states, we simply re-weight these state-specific probabilities by a vector with ones for each state compatible with a given ambiguity code and zeros elsewhere. As these imputed values are sampled one trait at a time, marginal of other imputed values in any given individual, they are inappropriate to use during iterative optimization steps of each pairwise correlation, and so we forego their inclusion there. In estimating these values, we pool across populations and not traits, but wish to note that this does not imply that the probabilities of particular states going missing are equal across populations. Rather, the assumption of consistency across populations only applies up to odds — or the ratios of probabilities — as it is only through these relative measures that the state conditional probabilities are affected, given the normalization constant found in the denominator of Bayes' theorem.

As mentioned before, our stochastic imputation algorithm precludes convergence to some optimal set of values, as new missing states are sampled after each round of optimization, resulting subsequently in slightly new optima. To obtain a more stable estimate of optimal values, averaging over stochastic imputation variance, we take the arithmetic average of model parameters

from four independent chains. Additionally, we assume the data are MCAR for the first 4 rounds of iterative optimization, excluding missing values from the procedure, in order for the algorithm to first attain a plausible set of values before attempting to estimate missing state probabilities.

Additional Correlation Matrix Processing.—The space of positive semi-definite (PSD) correlation matrices is far smaller than the space of square matrices with unit diagonals and off-diagonal elements in the range (-1,1), and so despite averaging four independent runs and regularizing correlation coefficients by a beta(10,10), the correlation matrix estimated from the above algorithm is nevertheless improper. To obtain the nearest positive semi-definite correlation matrix, we use an algorithm that minimizes the distance — measured as a weighted Frobenius norm — between our improper, non-PSD correlation matrix and a proper PSD correlation matrix (**?**), as implemented in the `nearPD(corr = T` function in the *Matrix* package (**?**) in R, also used in similar contexts elsewhere (**?**). The largest change to a single pairwise correlation resulting from this procedure is 0.116, and the median change 0.012. For numerical stability when computing determinants (otherwise $-\infty$) and inverse lower Cholesky factors of this and related matrices in the next stage of inference, we then weight this matrix with the identity in a 50:1 ratio, resulting in a further maximum change to the previous matrix of 0.017, and a median change of 0.0016.

Bayesian Inference.—Have obtained an estimate of optimal values from the first step of this analysis, we now turn to the second step: Bayesian phylogenetic inference. Here, we specify a multivariate Brownian motion model of character evolution acting over a strictly bifurcating phylogeny with 8 tips, realizing our estimated means. As mvBM is insensitive to the location of the root, inference is done under unrooted trees. We use a discrete uniform prior over tree topologies, $\log_{10}$Normal(1,1) prior over total tree length, and a flat dirichlet(1,1,...) over branch length proportions, which multiply tree length to obtain branch lengths. For correlation components of the rate matrix, we specify a point-mass prior on the above within-group, between-liability correlation matrix, fixing it to that value. For the rates, we specify a regularizing dirichlet(α, α, ...) prior on the relative rates, and an offset $\log_{10}$normal(0,1) + 1 hyperprior on α, to allow the model to

learn the extent of between-trait rate variation justified by the data. These relative rates multiply the total number of traits used in this analysis — 118 — to constrain the rate matrix to an average rate of 1 and allow trait-specific rates to be identifiable alongside phylogenetic branch lengths. We approximate the joint posterior distribution of these five sets of parameters — tree topology, trait rates, tree length, branch lengths, and α — using the Metropolis-Hastings algorithm (**?**), making NNI and SPR proposals to tree topology, beta proposals to all simplex variables, and sliding window proposals to all other parameters — in approximately a 4:16:2:4:1 ratio, respectively, using the `rNNI` and `rSPR` functions from the *phangorn* (**?**) package for tree proposals but otherwise implementing the remainder in base-R. We ran four independent chains initialized from the prior for 1E7 iterations each, thinning every 5E3 iterations. The first 40% of the each chain was discarded as burnin. To diagnose MCMC performance, we assessed the effective sample size and Gelman-Rubin Convergence Diagnostic (**?**) of several explicit and implicit model parameters, both implemented in the R-package *CODA* (**?**), and requiring that each be above 1,000 in the former case and have a upper 95% value below 1.01 in the latter. This criterion is applied in each of the four independent chains as well as in all four chains concatenated. The parameters examined here included all rate parameters, α, tree length, terminal branch lengths, and Robinson-Foulds Distance (**?**) from a reference tree. Additionally, we required that the correlation between all pairwise comparisons of bipartition probabilities between chains be $>$0.99.

Several computational tricks were used to accelerate likelihood computation, mostly with respect to storage of the rate matrix and exploiting basic identities in linear algebra. As the correlation components of the rate matrix were fixed, and information regarding the structure of the rate matrix stored in the form of its inverse lower Cholesky factor L^i and determinant, perturbations to individually indexed rates of the rate matrix required only that we multiply the columns of the former by the square root of the factors by which their corresponding rates changed, and the latter by the product of those factors' inverses. These could then be used to update the transformed trait values, which could then be transformed by the appropriate factor of the phylogenetic covariance matrix, which we diagonalized using a linear algebraical implementation of Felsenstein's Prun-

ing Algorithm (**?**), rather than the postorder traversal through which it's usually implemented. Information regarding the tree, then, could be stored in the form of a transformation matrix and vector of contrasts' branch lengths, which could then be cheaply updated following proposals to the tree, tree length, and branch length proportions, and used to further transform raw tip means into a series of iid standard normal variables, the densities of which could be altogether far more easily evaluated to produce the same likelihood values as more computationally cumbersome approaches commonly implemented in standard phylogenetic software.

Simulation Experiments.—Having inferred a strictly bifurcating population history using our empirical dental dataset, we sought to better understand the statistical properties of our two-step approximate restricted multivariate Brownian ordinal probit (TSAR-MBOP) model, having made several concessions in the names of tractability and practicality. Thus, we conducted a short simulation study in which the performance of the method at retrieving simulating trees and rates with well-calibrated posterior distributions could be assessed under empirically realistic data-generating conditions. First, we take our estimate of the matrix of thresholds and correlations from the empirical step one above. Then, we sample at uniform from step two's joint posterior output a vector of trait rates and tree with vector branch lengths, using the former to recompose a rate matrix with our estimated correlation matrix. We then midpoint root our sampled unrooted tree and, using our estimated tip means, sample from the multivariate Brownian bridge coursing through the root an ancestral state by the closed-form expression of multivariate normal conditional distributions, which we obtain via Schur complements of the covariance matrix by which a mvBM likelihood may be written in its Kronecker product form. This is itself a multivariate normal distribution representing the distribution of states at the root, conditional on the tree, tip data, rate matrix, and stochastic process, though the procedure is far more general and can be used to jointly sample character histories throughout the entire tree. We then simulate forward in time tip liability mean vectors according to the mvBM process, which we then use alongside our estimated correlation matrix to sample individual liability vectors in count equal to that of our processed empirical dataset, with population sizes (119, 51, 84, 40, 40, 17, 135, 198) corre-

sponding to the (Neandertal, Oceanian, European, West Asian, South Asian, Northeast Asian, Sub-Saharan African, American) tips, respectively. With our estimated threshold matrix, we then convert these individual liability vectors into ordinal characters, and simulate state-dependent missingness with the inverse-logit function, assigning state 1 a probability of missiningness equal to 0.69 and other states a monotonic decreasing or increasing probability of missingness 0.5 away in either direction on the logit scale, corresponding to state-dependent missingness probabilities of (0.79, 0.69, 0.57, 0.45, 0.33, 0.23, 0.15, 0.10, 0.06) for states 0 through 8. Applying this function to our simulated alignment, we render approximately between 65% and 70% of the data missing, targeting the empirical missing probability of 65.4%, and further specify partial missingness by simulating presence in each ambiguous assignment category in proportion to its empirical frequency. Having thus constructed an individual-level discrete ordinal alignment matrix similar to that obtained after data pre-processing in our empirical application, we analyze it using the two-step procedure described above. These simulations and analyses are repeated 500 times in order to disentangle the properties of our method from simulation variance.

RESULTS

Fitting the ordinal probit model to our empirical data according to the first step of our two step procedure produces highly similar estimates across four independent runs (figure 3), providing reassurance that these model parameters are being estimated reliably. Averaging these output and adjusting the correlations as described earlier, we analyze them in a Bayesian phylogenetic framework and, upon assuring ourselves of MCMC health, sort the posterior distribution of trees according to their posterior probability. For eight tips there exist 10,395 unique topologies, and despite a relatively diffuse posterior distribution we are still able to consistently find a most probable set of trees across chains. The four most probable trees are visible in figure 4, with nodal bipartition probabilities plotted. Branch lengths on these trees are posterior means for only those trees in the posterior distribution that shared their particular topology.

In addition to tree topology, other phylogenetic model parameters may also be of interest. From our iterative optimization step, we obtained within-group estimates of between-liability correlations for each of our dental traits. Partitioning these into correlations within individual teeth, within the same trait across teeth, and remaining components, we can assess the nature of modularity across the human dentition (figure 5a). Our phylogenetic analysis also provides estimates of trait-specific rates under a mvBM process of dental evolution. Examining these, we can see whether particular traits or teeth are evolving at unusual rates across the entire tree (figure 5b), with the caveat that these rates are confounded with the degree of separation between thresholds, itself influenced by within-tip variability in discrete state, especially at intermediate degrees of expression. The posterior mean of our α-concentration parameter used to regularize trait rates was 3.37, with a 90% credible interval of (2.42, 4.60), suggesting substantial variation in the rates of trait-specific evolution.

Having inferred the population history of our seven populations of *Homo sapiens* and one Neandertal tip, we assessed how reliably our method could recover simulating model parameters under empirically realistic conditions, given the approximate nature of the compromises made along the way. To evaluate our ability to retrieve between-trait / within-population correlations,

population liability means, and threshold locations, we generated scatterplots (figure 6a-c) of estimated vs. true values for all three sets of model parameters, as well as examined the distribution of R^2 values for these across our 500 runs, also restricted to the subset of traits that were not invariant in our simulated data (figures 6d-f). To examine the success of our stochastic MNAR imputation algorithm, we generated violin plots for the probabilities used in our final round of iterative optimization across runs, comparing them to the known Pr(state | missing) used to simulate state-dependent missingness (figure 7). Finally, to see the extent of error introduced by our two-step procedure when inferring trees conditional on estimated means and correlations, we produced calibration curves for both topological and bipartition probabilities (figure 8a-b), as well as histograms of quantiles for true, data-generating rates, branch lengths, and tree lengths in the marginal posterior distributions of inferred rates, branch lengths, and tree lengths (figure 8c).

Discussion

Empirical trees (figure 4) inferred appear to be broadly consistent with both prior work (**?**) and molecular expectation (**?**). Though no explicit outgroup was specified for rooting, midpoint rooting resulted in trees mostly leading to the Neandertal tip. Across the entire posterior output, 60% of trees specified a first bifurcation between Neandertals and all *Homo sapiens* populations when midpoint rooted, despite Neandertal extinction robbing the extinct species of tens of thousands of years of dental evolution experienced by the other tips. Curiously, the next tip to split off from the *Homo sapiens* stem appears to be that corresponding to South Asian populations, rather than Sub-Saharan African populations, despite the latter representing the earliest divergent human groups in molecular studies. Instead, Sub-Saharan Africans appear to cluster with the European tip with moderately high probability (0.68), potentially due to paraphyly in the former tip caused by our lumping of multiple populations into one. In contrast, American and non-South Asian tips appear to cluster together with intermediate probabilities, along with populations from Oceania, consistent with molecular expectation. In the *maximum a posteriori* (MAP) tree, northeast Asian populations and American populations appear to bifurcate last of any pair of tips in the tree, potentially a signature of the later peopling of the Americas by the latter group according to a northeast Asian disperal across the Bering land bridge (**?**).

Within-group correlations partitioned *within* named sets of dental traits *between* teeth are overall more positive and stronger than those within teeth between traits or those between traits between teeth (figure 5), though correlations between traits within teeth appear to be more variable overall, with the strongest correlations of any in the matrix found there (figure 5a). However, given the results of our empirically parameterized simulation study (figure 6), correlation parameters appear to be the least reliably estimated of all within-population parameters during our first optimization step. This may be partly attributable to low sample sizes within tips limiting the extent to which the model could learn correlation patterns in the data, given that information thereof lies in paired variation throughout the dataset. Because of the long trees, variable rates, low sample sizes, uncertain ancestral states, and high proportions of missingnness used to parameterize our

simulations, simulated data frequently lacked this paired variation at the ordinal trait level. For example, the median number of wholly monomorphic traits in the observed subset of our simulated discrete character alignments was 2, with over 10% of simulations having 5 or more entirely invariant traits. There is fundamentally no information regarding correlations between liabilities within populations for data such as these, and so in an optimization framework the only value possible for correlations between these invariant traits and all others is 0, the mode of our regularizing Beta(10,10) distribution. Furthermore, a median of 12 additional traits were not represented in more than one state by at least 10 individuals (with over 10% having an additional 18 traits so impoverished), suggesting that their correlations would be hard-estimated indeed, as those few individuals would need to covary in their trait expression at other locations in the alignment for there to be information regarding correlations that optimization could learn from.

These issues highlight aspects of the simulating process that did not accurately reflect the mechanisms by which the ASUDAS was constructed, as well as broader concerns over ascertainment bias that afflict any phylogenetic study of morphology. Unlike continuous traits, discrete traits may easily be invariant within populations, and systems such as the ASUDAS were explicitly designed to characterize variation within and between human populations. Furthermore, commensurability between traits is itself questionable. In molecular sequence alignments, there's a sense in which the evolutionary processes acting upon different loci are comparable, allowing us to adaptively regularize inference across loci by pooling information between sites in a principled manner. For quantitative characters evolving under geometric Brownian motion, perhaps a similar pooling might be justified. But discrete characters — such as dental cusps or grooves — hardly seem to be so fundamentally equivalent, though we may still wish to specify weakly informative priors that allow them the opportunity to regularize, as was done here, should there be sufficient hints of consistency in the between-character evolutionary process to vindicate our allowance.

Despite these caveats, it would appear that tip mean liabilities and threshold locations may still be reliably estimated with data such as these, likely because there is no more need for paired variation in the dataset. Instead, the only tip mean liabilities our optimization procedure truly strug-

gled with were those that had drifted to extreme values, especially those that resulted in within-tip invariance at the maximal or minimal ordinal state. When individuals within a tip are invariant for some trait in this manner, the most compatible location of its mean liability is at positive or negative infinity, respectively, and almost equally plausible are all values between those extremes and some short distance away from the largest and smallest thresholds. It falls, then, to one's choice of regularization to pull estimates away from their extremes, penalizing the likelihood function for growing too excited about invariance. As we regularized under an constant-rate, univariate Brownian process acting on a star phylogeny, it fell to the overall variation observable between tips on a liability scale to reign in optimization's desire to supply the most ostensibly plausible values. But plenty of information was ignored here, specifically pertaining to covariances in the evolutionary process generating variation between tips and phylogenetic structure itself. Joint inference, which simultaneously traverses only PSD correlation matrices and bifurcating trees, is likely the solution needed to improve estimates for troublesome, invariant traits.

Our MNAR imputation algorithm appeared to be reasonably successful at recovering patterns of state-dependent missingness (figure 7), with pooled probabilities across traits recovering the appropriate monotonic decreasing order, despite that assumption never having been explicitly baked into the algorithm. For estimation, however, these probabilities were evaluated and incorporated on a per-trait basis, given commensurability concerns. This proved far less reliable than pooling across traits, considering how much more information lies in the cumulative signal of 118 traits observed in 8 populations and just one. Small probabilities at high degrees of expression were not as well estimated despite pooling done for figure 7, likely because the extent of pooling was far weaker. While all traits could contribute to the estimation of Pr(missing | state) for states 0 or 1, only a single trait in these data could occupy a 9th state of expression, only 5 additional traits the eighth, 8 more the seventh, and so on. With less data available, our flat beta could not be so reliably updated, and so despite its uninformativeness, it nevertheless appears to have shrunk estimates towards intermediate values. Still, despite our imputation algorithm not having quite recovered the true probabilities of state-dependent missingness at these sample sizes, it appears

to have proved sufficient to unbias mean estimates away from their otherwise positively biased, MCAR values (figure 6).

Overall, it appears that our use of a two-step algorithm as a concession to tractability did not impact our ability to infer phylogeny too greatly. Despite poor estimation of correlations of the mvBM rate matrix, bipartition probabilities (figure 8a) and tree probabilities (figure 8b) were still reasonably well calibrated and true values of continuous model parameters appeared to be as drawn from their respective posterior distributions. Strictly speaking, our empirically minded simulation study parameterization necessarily supposes posterior probability miscalibration, as simulating model parameters were not drawn from the prior distributions used for Bayesian inference. Further work may try to disentangle the extent to which the more tractable multivariate normal integral approximator and two-step optimization-inference procedure results in miscalibration, versus error due to mismatch in simulating and prior distributions.

Many additional opportunities to improve the approach adopted here remain. As mentioned, exploring the statistical properties of the high-dimensional phylogenetic multivariate ordinal probit model in a joint inferential framework could yield easy improvements. Greater mathematical rigor or more clever computational approaches to approximating multivariate normal integrals may allow us to do away with dissatisfying approximations, and, combined with novel algorithms to traverse difficult parameter spaces (Apendix II), may allow for the exploration of higher dimensional character evolutionary processes than currently feasible. Investigating the impact of ascertainment bias on the collection of discrete morphological character data is likely to reveal similar biases as found in regions of statistical inconsistency under Maximum Parsimony based methods, which also disregard information at invariant, parsimony-uninformative sites. As our ability to more easily record greater amounts of information on population distributions of morphological characters improves, there likewise grows a greater need for more sophisticated inferential models, and an even greater need to render the fitting of those models tractable under the limits of current computer hardware.

FIGURES

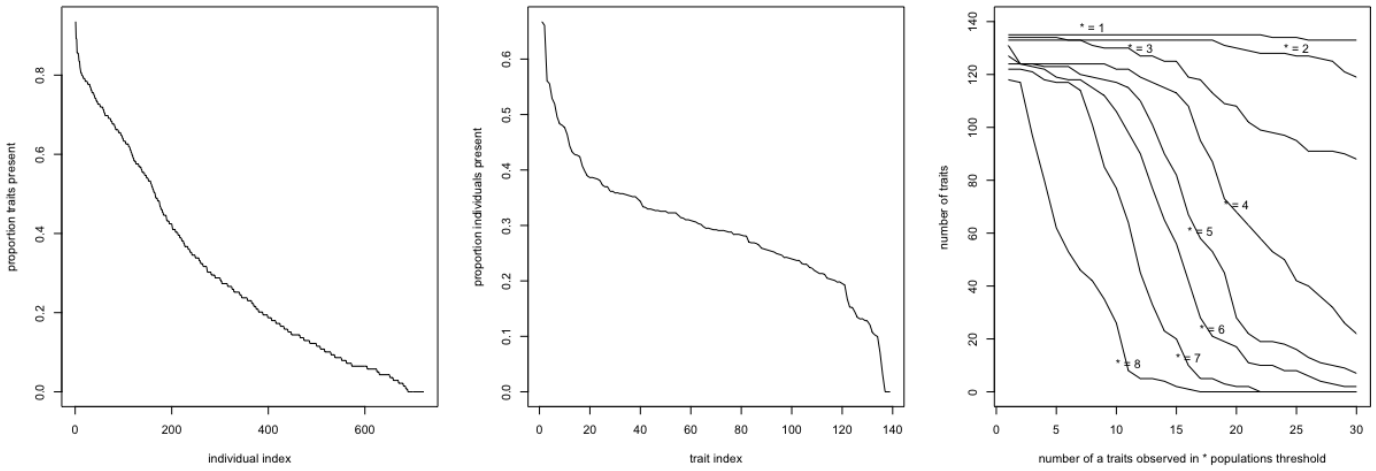

Figure 1: A vizualization of missingness in the dataset. In a), the proportion of traits present in the sorted, decreasing set of individuals represented in the sample. Colors represent different tooth types, stacked according to their mesio-distal progression within the dention. In b), the number of individuals available to represent each set. Colors represent populations, stacked according to total population size. In c), information in these figures is combined to produce a graph depicting how criteria pertaining to the minimum number of individuals in a minimum number of populations affects the number of traits ultimately present in the sample.

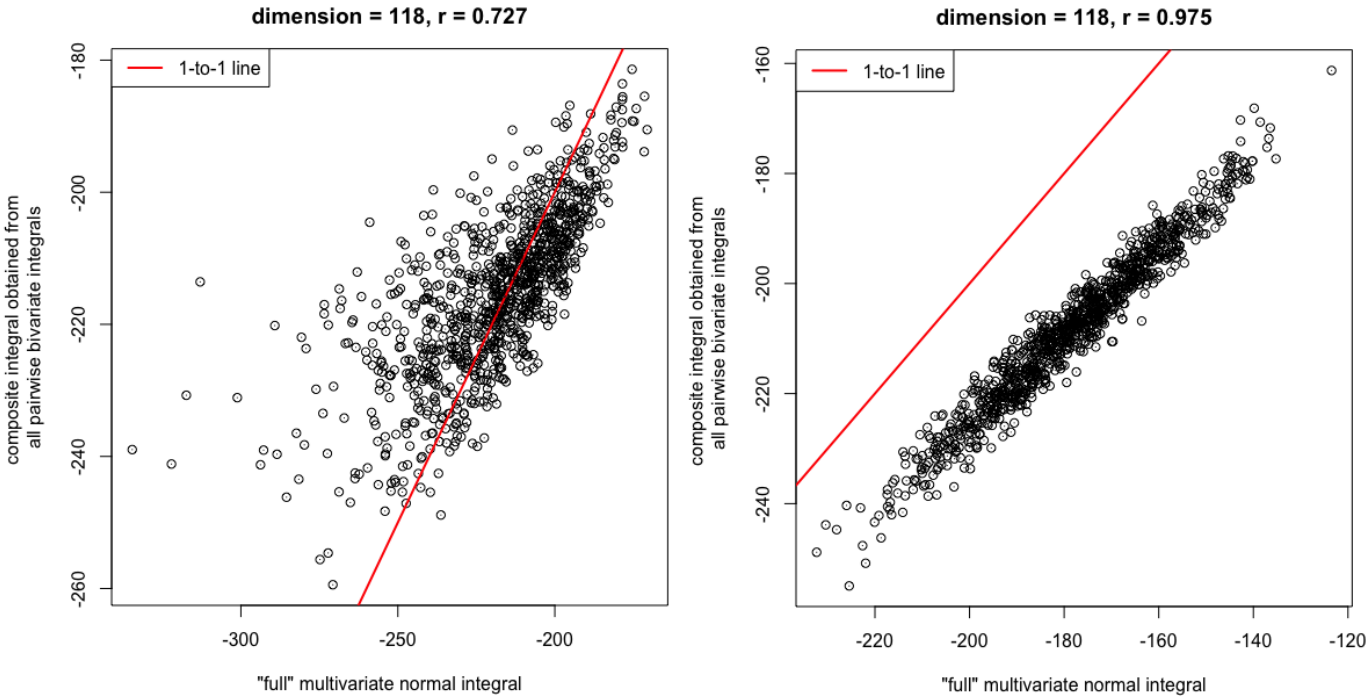

Figure 2: Visualizing relationships between the transformed bivariate integral of a multivariate normal and its full evaluation. In a), the integral of a multivariate normal with mean at the origin and 118 x 118 correlation matrix sampled from an LKJ(1) was evaluated with both methods between pairs of lower and upper bounds sampled at uniform and sorted from the (-1,1) range. The $\log_e$ scale output of 1,000 such simulations is shown, with 1-to-1 line marked and correlation between the two labeled. In b), the procedure is repeated, except with the correlation matrix to have all off-diagonal elements equal to 0.9.

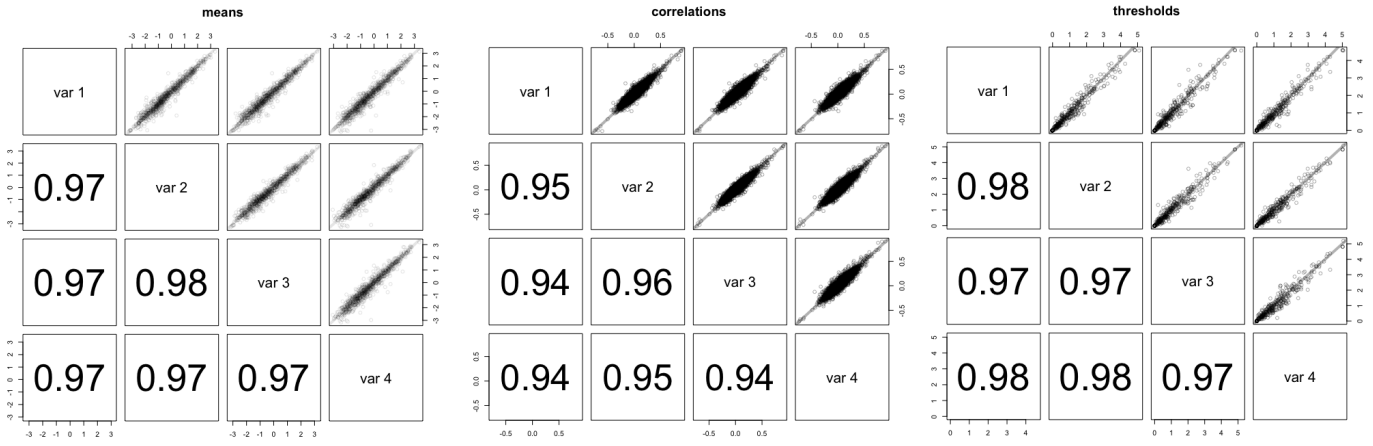

Figure 3: Output from the iterative optimization step of our two-step algorithm across four independent runs. In a), means are plotted in the upper right panels of the figure, with correlations between runs in the lower left panels. In b), within-group, between-liability correlation parameters are plotted. In c), threshold locations.

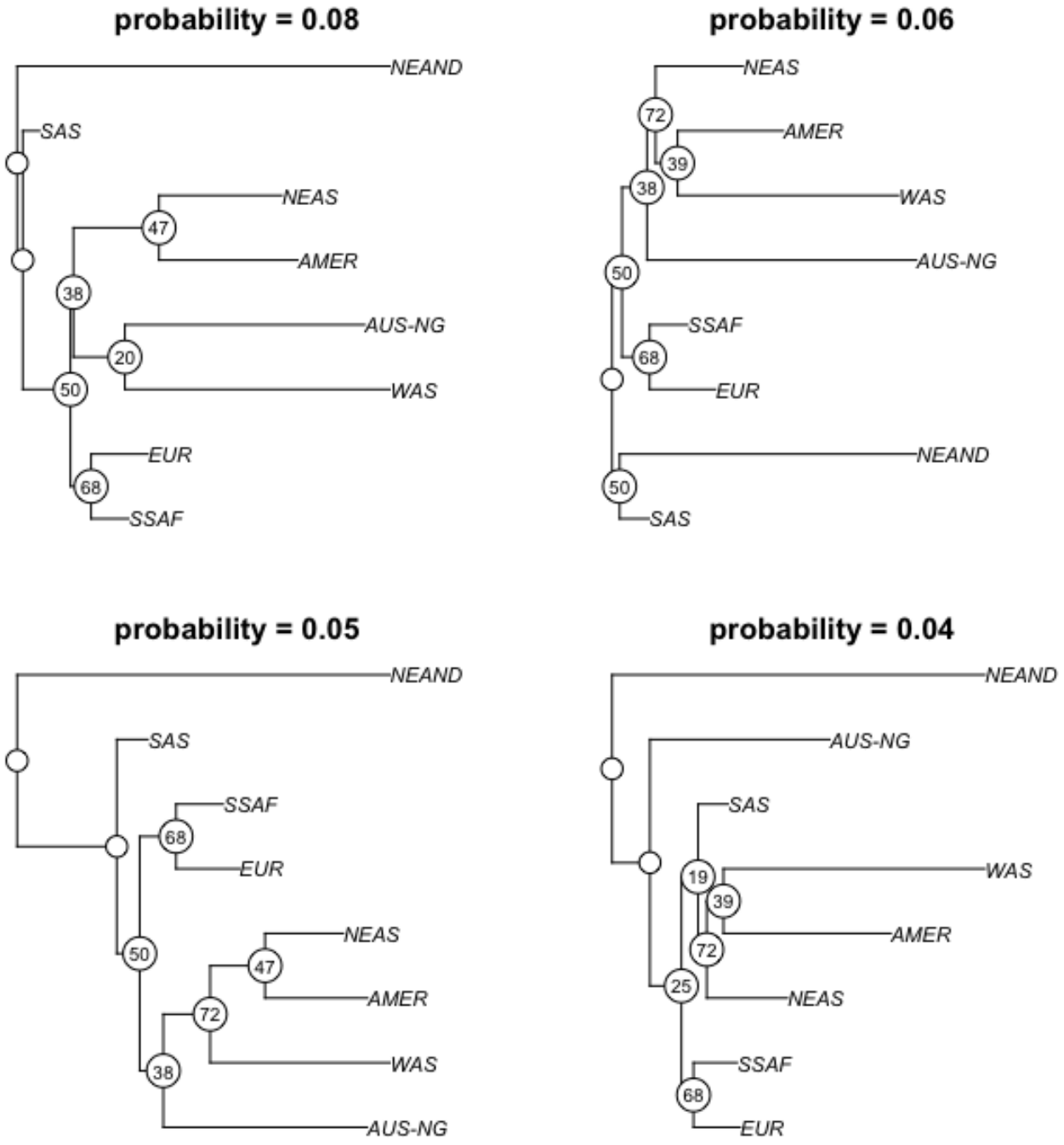

Figure 4: The four most probable trees from the posterior distribution of our Bayesian phylogenetic analysis, with nodal posterior probabilities plotted. Branch lengths are posterior mean estimates conditional on each tree topology and are proportional to the extent of morphological evolution on each branch. For visual clarity, midpoint rooting was performed, though may be unreliable given the lack of explicit accounting for tip ages, specifically with respect to the reduced opportunity for evolution on the Neandertal branch.

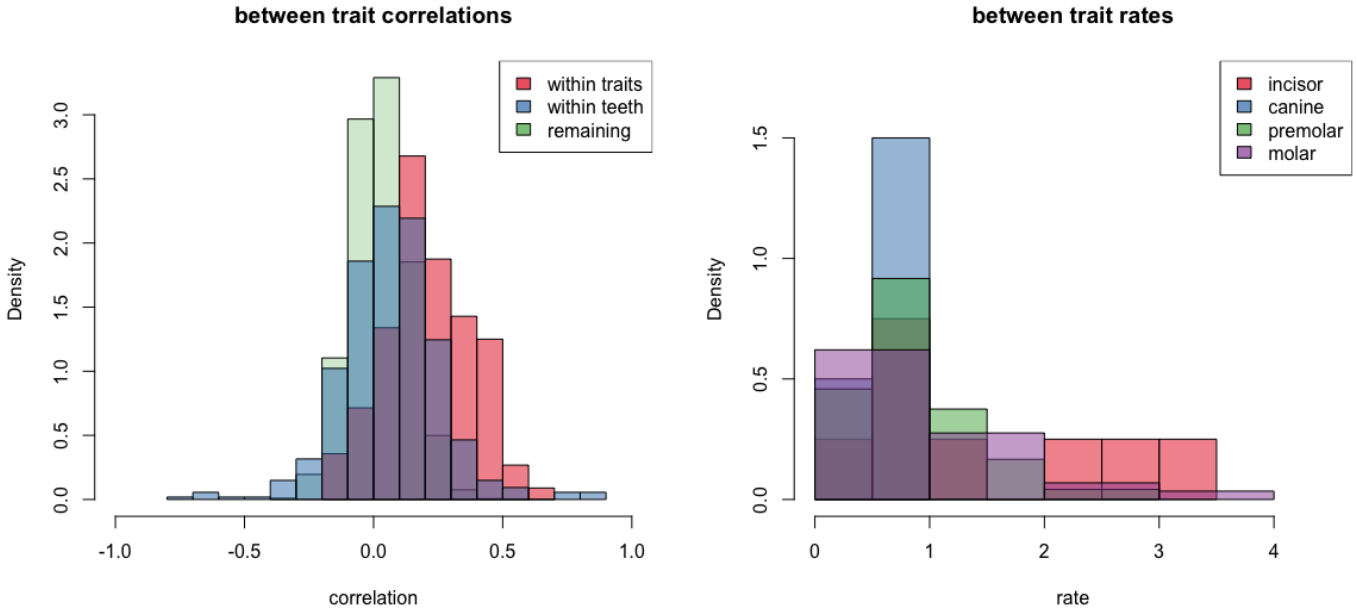

Figure 5: In a), histograms of correlations between traits within individual teeth, within traits across teeth, and for the remaining elements of the correlation matrix are plotted. Correlations are those from the evolutionary rate matrix, and so are interpretable as correlations of the evolutionary process, though they were estimated from within-population data per Cheverud's conjecture. In b) posterior means of trait-specific rates are partitioned across types of tooth within the human dentition, and tooth-specific histograms are plotted.