Exploring and Extending Multivariate Brownian Diffusion Models of Phenotypic Evolution for Bayesian Phylogenetic Inference

By

Nick Lashinsky

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Anthropology

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Timothy D. Weaver, Chair

Shara E. Bailey

Peter C. Wainwright

Committee in Charge

2020

Dedicated to Pipsqueak . . .

Thank you for not dying during my PhD, despite your advanced aged,

and for joining me in almost every location where non-service animals were welcome.

Contents

List of Figures

List of Tables

Abstract

Exploring and Extending Multivariate Brownian Diffusion Models of Phenotypic Evolution for Bayesian Phylogenetic Inference

Paleontologists and neontologists alike desire improved methods for inferring phylogeny with morphological data, and in this work we hope to not disappoint them too greatly. An introductory section (Chapter 1) surveys our motivations for carrying out this work, and also provides a short overview of continuous character evolutionary models, as well as a short defence of the Bayesian inferential framework. Working in that framework, we explore in Chapter 2 the performance of a multivariate Brownian diffusion model (mvBM) at retrieving data-generating tree topologies through an extensive simulation study across a range of idealized and empirically realistic conditions. Following that, we characterize its performance using both standard and novel diagnostic tools, further investigating the extent to which model misspecification of different stripes affects our ability to reliably do phylogenetic inference. Then, we apply this character evolutionary model to two empirical datasets: linear measurements collected from the crania of a globally distributed sample of *Homo sapiens* (Chapter 2), and landmark data collected on the crania of 13 species of catarrhine primate (Chapter 3). The latter are matched to a well-resolved reference tree obtained through analysis of molecular sequence data, and so we compare the performance of our proposed multivariate Brownian model to that of univariate Brownian motion, two discrete character evolutionary models, and a varied set of more commonly used heuristic methods. To explore the nature of inferential errors apparent in this step, we also undertake a study of cranial character evolution using a relaxed morphological clock conditional on the molecular tree, which we then use to parameterize a short simulation study. Finally, in Chapter 4 we apply the method to a high-dimensional dataset of discrete dental traits codified using the Arizona State University Dental Anthropology System, assuming a multivariate ordinal probit filter. As joint inference of both phylogeny and tip mean liabilities proved beyond our current computational reach, we adopted two-step approach, first optimizing the locations of tip means, thresholds, and between-trait

correlations given individual-level sampling under the multivariate ordinal probit, and then conditioning on these estimates but inferring trait-specific rates in a second, phylogenetic inferential step. To assess the extent to which compromises made in the name of tractability affected our ability to infer phylogeny under this model, we performed a third and final simulation study. In addition, three appendices provide further mathematical details. The first demonstrates how a phylogenetic likelihood under mvBM — represented by a high-dimensional multivariate normal density — can be factored into the product of univariate normal densities. The second describes and validates a novel, tunable proposal distribution over correlation matrices that makes both phylogenetic likelihood calculation and mixing over this especially problematic model parameter much more efficient. The third demonstrates an easy means by which ancestral character states can be sampled from the conditional multivariate normal distribution implied by their location in a tree in conjunction with the data observed at the tree's tips. The dissertation concludes with short section reflecting upon lessons learned and describing directions for possible future work.

Acknowledgments

I owe a tremendous debt of gratitude to many parties ensconced both within and without the walls of the University of California, Davis.

Acknowledged first are my two mentors at UC-D, Tim Weaver and Brian Moore, whose long discussions — in offices, living rooms, and pubs — served as ample fuel to stoke the fires of my own scientific curiosity and dedication, while also providing no small measure of fun and enjoyment. Without their guidance and instruction in matters paleoanthropological and phylogenetic none of these projects would ever have left the ground, and I deeply thank them both for accepting me as their mentee. Alongside Tim and Brian, I'd be remiss not to also credit the Anthropology Department and Center for Population Biology to which they and I belonged.

Next, I'd like to thank the folks at *Data Science & Informatics*, and especially Pamela Reynolds, for giving me the space, funds, and guidance to host many workshops, discussion groups, and my Applied Bayesian Statistics Research Cluster. Outside my own groups, I joined several, and so am also grateful for those riveting chats emerging from paleogroup, morphogroup, pythongroup, computational molecular evolution group, machine learning group, and others.

Many additional faculty and staff aided me on my graduate career, and so also deserve my gratitude. Peter Wainwright, for highly educational walks in forests and along beaches, for ichthyological side-projects and cluster construction services, for accepting Brian's proposal to let me serve as Bodega workshop coordinator, and for serving on both my qualifying exam and dissertation committees. Mark Grote, for reigning in my *ad hoc*, shoot-from-the-hip statistical proclivities, answering hundreds of convoluted, meandering emails, and consulting on projects in fields ranging from transplant immunology to consumer dietary behavior to Bayesian phylogenetics. Nicolas Zwyns and Richard McElreath, for many fun conversations and for serving as excellent instructors during my initial TAships, as well as for freely giving their slides when it came time to teach my own courses. Teresa Steele, not only for many years of advice and insight in zooarchaeology, but also for letting me in to my first full-summer excavation. Shara Bailey, for training me

in dental anthropology during my visit to New York, for lending me her data and insight for my third main dissertation chapter, as well as for providing feedback on the broader dissertation and serving on my qualifying exam and dissertation committees. Though not (yet) faculty anywhere, Mike May also deserves special mention, for his wit, wisdom, and phylogenetic expertise, as well as for acting in the unenviable role as the first person I'd typically turn to for brainstorming project ideas / solutions & for begging for help when things went awry.

The students I instructed in my human evolutionary biology, human evolution, and primate evolution courses were all a joy to work with and teach. Thank you for putting up with my inane rambling, lame jokes, dated references, confusing labs, harsh exams, and statistical methods and theory. I wish you all the best in your future careers and endeavors!

I would not have had nearly as easy a time of it without the generous financial support of several funding bodies, and so would like to thank the National Science Foundation, the University of California, Davis, and the Department of Anthropology for their assistance, along with those who funded my undergraduate education that I need not to worry myself with debts accrued then and there.

My family, both immediate and extended, deserve much of the credit for those developmental process that shaped me into who I am today, and for chatting with me on my daily walks into work. Especially my mom, grandma, and grandpa. The lattermost of which needs to stop making excuses and return to reading his daily readings!

As my students, friends, and colleagues know all too well, a small dog and smaller cat play a big role in my life, and their emotional support and nightly cuddles have been an excellent way to unwind, 5★, would recommend to friend.

Friends outside Davis who either flew out to visit or go backpacking with me or who hosted me on their couches, in their spare bedrooms, or their empty houses provided vital distractions from the academic world, and I hope that those past habits can continue once more when conditions allow. Those who maintained a regular teleconferencing habit also deserve a shout-out — thanks for keeping in touch!

All the friends made at Davis and the Mayo Clinic also deserve mention here. Thank you for your friendship and thoughts, our trivia nights at pubs and our board game, video game, and movie nights at houses, the hikes you joined me on, the lengthy, meandering discussions, and the mutual accountability we provided one another in reading groups.

I'd also like to thank the custodial staff at UC-Davis, who kept facilities clean, took out my trash, and always greeted me with a friendly smile and wave.

Most of all, I'd like to thank my best friend and life partner, Kate Gates. We met a few days into the PhD program, and my life's been immeasurably better since. Thank you for standing as stalwart bulwark against the frustrations of this experience, and tolerating my angry mutterings at non-compliant, non-compiling code (here's looking at you, LaTeX). Coming home each night to hang out with you, be it in-person or remotely, formed the highlights of most of these days. By far, you're the best thing that's happened to me during this PhD, and I look forward to many more adventures to come.

Chapter 1

Introduction

Nikolai G. Vetr

1.1 The Need for Phylogenetic Models of Morphological Evolution

Biological variation both in the extant and fossil records is the outcome of a wide array of evolutionary and non-evolutionary processes. Darwin's "endless forms most beautiful" (Darwin, 1859) emerge proximally from the complex interplay of environment and development, but ultimately owe the bulk of their between-group differences to millions and billions of years of chance intergenerational sampling effects and reproductive asymmetries attributable to the heritable underpinnings of those organisms' morphology, physiology, behavior, biochemistry, and other traits (Mayr, 1988). Coupled with the bifurcating process by which reproductive isolation emerges and permits populations to proceed along independent evolutionary trajectories, we see these forms — these *species* — emerge and change, contract and die. By examining that variation and imagining a description of the processes that might have generated it, we can make inference of phenomena beyond the remit of immediate observation and guess at not just the tempo and mode of character evolution driving within-lineage variation, but also the order and timing of splitting events that birthed those series of independent daughter lineages.

We can represent an estimated history of these splitting events with a tree, be it one of population history (in the case of groups at the sub-specific level) or phylogeny (for group at or above the species level), which can be rooted to specify the order of events through time, or unrooted, as a means of constraining that order. With a mathematical description — or stochastic model — of the evolutionary process that changes the states of characters along branches of the tree, we can predict what sorts of observations we might expect to make at the tips of our tree, and explore how varying properties — parameters — of our tree and stochastic model might affect the distributions of those predictions. Comparing our real-world observations with the fictitious ones predicted by our model, we can find combinations of parameters more or less consistent with what we observe. The more plausible our data look under some instantiation of our model parameters, the better evidenced we might say those parameters are, allowing us to discriminate between alternative phylogenetic hypotheses. Through straightforward manipulation of the defi-

nition of conditional probability, we can combine a function describing the plausibility of our data with one describing the prior probability of our model parameters to strike a compromise between them, learning what probability we should ascribe to different hypotheses after having made our observations (Bayes, 1763), at least to the extent that we are willing to accept as given the assumptions of our data-generating process.

If our model is a poor representation of the evolutionary processes that give rise to variation, neither plausible nor probable model parameters may be very trustworthy. Tremendous efforts have gone into developing models to describe the evolution of molecular characters, of nucleotides, amino acids, proteins (Holland, 2013; Kapli et al., 2020), but comparably little attention has been paid to models of morphological evolution, especially of the hard tissues that preserve especially well in paleontological contexts. And while neontologists make ample use of increasingly sophisticated methods to infer phylogeny among extant taxa, those reliant on morphology must make do with either heuristic methods lacking in flexibility or other desirable properties (Wright and Hillis, 2014, e.g. a principled accounting of inferential uncertainty, consistency at the limit of infinite data, etc.), or else fit models of character evolution that make uncomfortable assumptions about the nature of morphological evolution (Felsenstein, 2004, e.g. independence between characters, monomorphism within lineages, etc.). And where more sophisticated models exist, their implementation might be sufficiently inefficient as to render practical inference a distant pipe dream, so slow as to be called computationally intractable.

The work presented in this dissertation does not claim much in way of novelty — the models contained herein have been known for many decades — but it does develop a few small tricks that make fitting certain stochastic models more practically efficient under the limits of modern computer hardware, as well as a few novel means of visualizing and interpreting those fitted results. In this manner it extends past work (as e.g. Felsenstein, 1973, 2005, before it), much as past work made more practical previously proposed methods and theory. Likewise, it does not address every complaint in the matter of biological realism levied against popular models of morphological evolution. But it does explore the statistical properties of some of the more satisfying models under consideration, both on their

own merits and in comparison to alternatives. Occasionally, it shifts its attentions away from its primary focus, the set of branching events collectively called the phylogenetic tree's *topology*, and towards other aspects of the phylogenetic model, such as variation in evolutionary rates between characters, across lineages, or through time, or reconstructions of character histories, in the case of truncated biogeographic diffusions. Most of its empirical focus lies with extant humans and close relatives for whom phylogenies and population histories are well-resolved, or at least halfway understood, that more experimental inferential methods may be evaluated by reference to molecular expectation. The primary model considered here is that of multivariate Brownian motion (mvBM), acting either on raw or log-transformed character data, on a vector subspace of that data, or on that data after its been passed through an ordinal probit discretization filter. The mvBM model can be derived under quantitative genetics to correspond a broad range of composite evolutionary processes (Hansen and Martins, 1996), and it is exceptionally easy to work with, making it an excellent target for exploratory phylogenetic study. An overview of the broader class of evolutionary models to which it belongs, as well as motivations for using this model over others, can be found in the section below.

1.2 A Brief Overview of Popular Continuous Character Evolutionary Models

1.2.1 Ornstein-Uhlenbeck Processes

Brownian motion can be thought of as nested in another, more parameter-rich model of character evolution called the "Ornstein-Uhlenbeck" process (Butler and King, 2004; Beaulieu et al., 2012, OU). This process resembles Brownian motion, except it incorporates an elastic, centripetal force by which quantitative traits are pulled toward some optimum value θ with magnitude of pull proportional to their distance from θ and a "strength of selection" parameter α. Where under Brownian motion the change a trait underwent in a single time step is state-independent, under OU it depends on the current state, such that $X_{t+1} = X_t + \alpha(\theta - X_t) + N(\mu,\sigma)$, or $dX_t = -\alpha(X_t - \theta)dt + \sigma dBt$ (Hansen, 1997). If X_t is greater than θ, $\alpha(\theta - X_t)$ will be negative and the trait in the next time step will be pulled "downwards" (proportional to the strength of selection coefficient α). Conversely, if X_t is less than θ, X will be pulled upwards in the next time step. Brownian motion, then, arises as an Ornstein-Uhlenbeck process where α is 0. Traits are expected to evolve toward the θ until they reach it, after which they will fluctuate around it without ever straying too far (Figure 1.1a). The larger α is, the faster traits will reach the optimum and the less they will fluctuate around it. A greater variance in the Brownian motion component of the Ornstein-Uhlenbeck process (i.e. a higher σ), meanwhile, will increase variance among the traits, leading to non-identifiability when trying to infer the values of α and σ from tip data, since an increase in one can be largely counteracted by an increase in the other.

Though the Ornstein-Uhlenbeck process does provide a measure of biological realism, it poses certain additional difficulties to the inference of phylogeny itself, difficulties that may render attempts to use it to retrieve model parameters such as tree topology futile. Under univariate OU with a single peak, the distribution of continuous character states at time t for a character with starting value X_0 is normal, with mean $\theta + e^{-\alpha t}(X_0 - \theta)$ and variance $\sigma^2(1 - e^{(-2\alpha t)}/(2\alpha)$. From this, it is easy to see that as t increases, the expression $e^{-\alpha t}(X_0 - \theta)$ goes to 0, and the long-term behavior of X centers around θ, with variance

determined by our two non-identifiable OU model parameters α and σ. Effectively, this limits the extent to which the signature of shared ancestry can persist in realizations from an OU-process at the tips of a tree. After sufficient time has passed, all tip values sample from the same, independent distribution determined by θ, α and σ, which is to say there no longer exists information about between-tip character covariation due to shared ancestry in those values. Rather than phylogenetic covariances decreasing linearly with phylogenetic distance, they instead increase exponentially (Hansen and Martins, 1996). Conversely, if α or t are sufficiently small, such as when characters are far from their soft limits in state-space over the timescales represented in the phylogeny, the OU process resembles a Brownian motion process, for which there exist more convenient and tractable multivariate likelihood functions as described later in this work (Chapter 2 and Appendix A).

To summarize, then: the more a single-peak OU process is needed to describe the character evolutionary process, the less useful realizations from that process will be for phylogenetic inference. For our purposes here, it may be best to first understand the extent to which Brownian motion, in its best-case context of little-to-no model misspecification, performs, leaving for others a greater detail exploration of those contexts in which an OU model breaks down when used to infer phylogeny. Additionally, a single-peak OU process may itself be a poor fit to the data, for which reason we might wish to posit the existence of multiple peaks. But adding peaks also adds parameters whose inference of marginalization is necessary, potentially robbing us of the power necessary to infer phylogeny and complicating our computations greatly. And at the limit, we might wish to specify as many peaks as we have independently evolving lineages, and further relax the assumption that these peaks' locations be static through time, perhaps allowing them to wander themselves in a manner analogous to fluctuating selection. But such a complex model strongly resembles a laggy Brownian motion, and indeed collapses to Brownian motion at sufficiently high α (Hansen and Martins, 1996). Similarly, uncorrelated selection models (where the magnitude and direction of selective effects are sampled from some background distribution, typically normal, across time steps) and directional selection

models (where all lineages share the same selective pressures) can also be shown to be equivalent to a Brownian motion model under certain conditions (Hansen and Martins, 1996).

This is not to say that OU-models are not useful, especially within their more usual role, serving as a candidate model in the phylogenetic comparative method. In work done in association with this dissertation but outside its scope, I explored their application in Stan for predicting reactive nitrogen concentrations across a hierarchically structured set of manure ponds across California (Figure 1.1). But their use here is limited, and so they shall not be considered further.

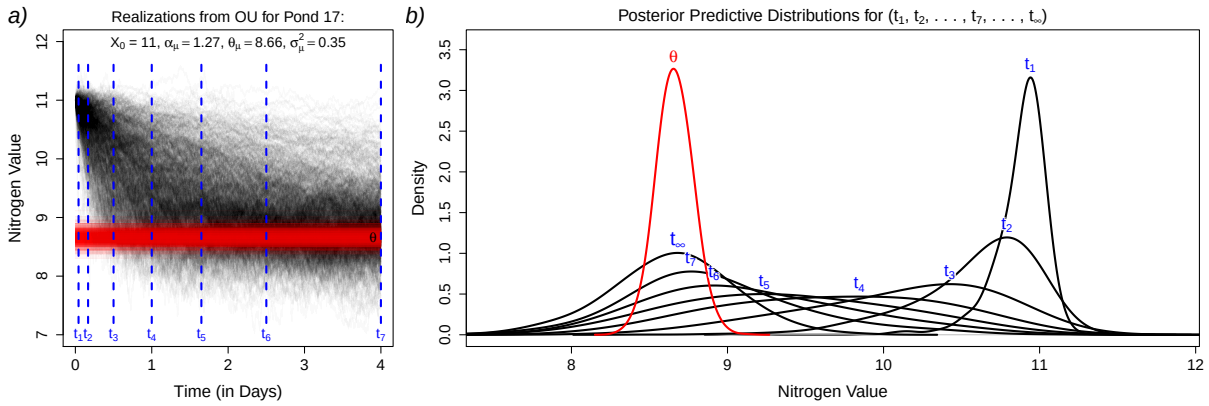

Figure 1.1: In a), realizations from an OU process over 4 days for Pond 17, averaging over posterior uncertainty. Mean values for all univariate OU model parameters are labeled, with samples from the posterior distribution for θ marked in red. The joint posterior distribution can be sampled and simulated from to generate posterior predictive distributions for arbitrary timepoints, marked with blue dashed lines. In b), kernel density plots of these posterior predictive distributions are provided, with corresponding timepoints labeled. A final timepoint, t_∞, gives the limiting distribution for the process. The model used here was a hierarchical model, and could be analogized to a star phylogeny of manure ponds whose optima evolve on the tree according to a Brownian motion process, with returning force (α_i) and rate (σ^2_i) parameters for each pond$_i$ estimated according to exponential hyperdistributions for each.

1.2.2 Morphological Clock Models

Another common modification made to the Brownian motion process in phylogenetic comparative contexts involves an allowance for the rate parameter to vary through time (Blomberg et al., 2003; Harmon et al., 2010). Specifically, one may set the Brownian

motion rate parameter at time t (σ_t^2) equal to a product of the initial rate (σ_0^2) and the base of the natural logarithm (e) raised to the power of the product of some constant (r) and time (i.e. $\sigma_t^2 = \sigma_0^2 e^{rt}$). If r is negative, rates will start off small and increase through time, leading to an Accelerating (AC) or Late Burst (LB) model of trait evolution. Meanwhile, if r is positive, rates will start off large and decrease through time, leading to a Declining (DC) or Early Burst (EB) model of trait evolution. If r is 0, the ACDC / Burst model collapses back into standard Brownian motion, as $e^{0t} = 1$ and $\sigma_t^2 = \sigma_0^2$. The inclusion of this extra layer, then, allows you to fit models that capture hypotheses pertaining to rapid diversification of form following a lengthy stasis, or vice versa.

In practice, phylogenetic likelihood calculation under these relaxed clock models works by rescaling the branch lengths of a phylogeny according to the integral of the rate function along each branch. This consideration reveals that all these named variations on Brownian motion should not *properly* be thought of as different character evolutionary models at all, but rather different *clock* models, relaxations on the assumption of a strict morphological clock dictating proportionality between morphological evolution and either time or molecular evolution. So too should be considered other transformations of phylogenetic branch lengths, such as Pagel's λ, κ, and δ (Pagel, 1999b,a), which are often interpreted to respectively measure "phylogenetic signal" (i.e., as an informal test of the strict morphological clock), association between diversification rate and morphological change, and more gradually slowing or speeding rates of evolution. For our purposes here, we typically place no strong constraints on variation in the rate of the morphological clock throughout a tree, preferring instead to set weakly informative, regularizing priors on branch lengths. As most of the analyses that follow lack an independent or jointly performed estimate of phylogeny on which morphological rate variation could be meaningfully considered, we instead infer trees whose branch lengths are in units of the square root of expected morphological change (specifically, the variance, σ^2 of a particle under Brownian motion scales linearly with branch length, and its expected absolute deviation can be found by evaluating twice the integral of the normal probability density function across (0, ∞), which equals $\sigma\sqrt{2/pi}$).

As in the preceding section, work done in association with but outside the strict scope of this dissertation explored our ability to retrieve parameters of relaxed morphological clocks — for example, centering an evolutionary burst not at the root of the tree, but some distance into its branches. Here, we examined the role mass extinction events (specifically the Cretaceous–Paleogene) played in the evolution of fish feeding morphology (Figure 1.2). In addition, a short simulation study identified the effects of such a "middle-burst" on disparity through time (Harmon et al., 2003) plots (Figure 1.3), a heuristic and difficult-to-interpret attempt to capture phylogenetic rate variation.

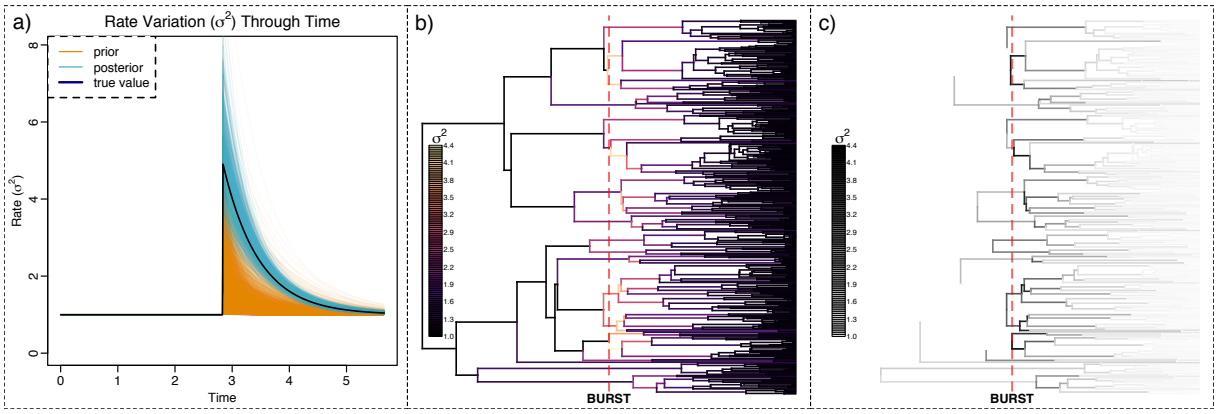

Figure 1.2: The result of a single simulation experiment attempting to infer the magnitude and decay constant of an evolutionary burst of known location given a known tree with 450 tips and univariate character data. In a), samples from both the prior and posterior are plotted and labeled, with the true value of the burst shown in dark blue. In b-c), branch-rates are plotted on the phylogeny used with color-coding and on a white-black gradient, respectively.

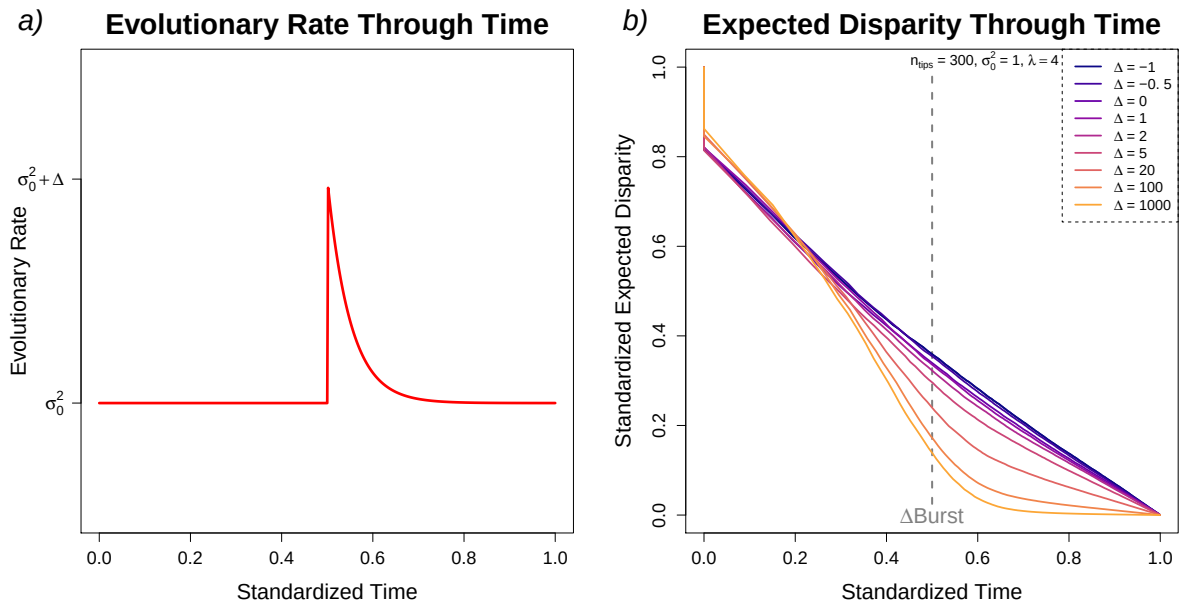

Figure 1.3: An exploration of the expected trend of a disparity through time plot for a tree with 300 taxa. In a), the relaxed morphological clock model is shown, featuring a change in evolutionary rate at the temporal middle of the tree, followed by an exponential decay to normalcy. The magnitude and direction of the change is given by Δ, and expected trends for different values of Δ are shown in panel b).

1.3 Benefits of Working in a Bayesian Inferential Framework

Model-based inference is preferred in part because of the flexibility it affords — should something in the set of considered models dissatisfy, it can be straightforward to engineer a better model, one more able to capture the suspected underlying biology responsible for generating empirical data. But having committed ourselves to the likelihood principle, we encounter a further dilemma: should the model be fitted in a (penalized) Maximum Likelihood framework, searching for the most plausible set of model parameters? Or should we instead try to reach a compromise between the information contained within our data, as perceived by our likelihood function — and our prior beliefs about the probabilities of different model parameters? As the title of this work suggests, we generally opt for the latter strategy, carrying out most inference in a Bayesian framework. This is despite objections raised by both those in favor of Bayesian approaches (Gelman and Yao, 2020) and those out of favor (Gelman, 2008). Our motivations in these regards are many and are described below:

Interpretability Though Bayes' theorem is, at heart, a trivial implication of introductory probability theory, it provides an elegant, powerful, and necessary vehicle by which to obtain probabilistic estimates of model parameters. Most scientists' statistical intuitions naturally hew close to Bayesian interpretations of probability (Kaplan, 2004), and accommodating inferential uncertainty and evaluating the degree of support in the data for competing hypotheses is much simpler with a probability than with devices such as the non-parametric bootstrap (Efron, 1979; Felsenstein, 1985a). When prompted, most practitioners default less to thinking in terms of sampling distributions of estimators and more to interpreting p-values or bootstrap support estimates as probabilities of alternative hypotheses (Greenland et al., 2016). Performing the non-parametric bootstrap in a multivariate framework would also require additional data transformation before resampling, as any data resampled with replacement would necessarily be perfectly correlated with itself, incapable of contributing additional phylogenetic information.

Prior Information Though often viewed as a detriment, Bayesian inferences allows for the incorporation of prior information into our analyses, forcing us to explicitly encode our beliefs about the model in greater detail than just those assumptions implicit to our choice of likelihood function (equivalent to point-mass priors on parameters of some higher nested model). We can specify prior distributions on model parameters as informatively or uninformatively as we feel justified, and further evaluate the extent to which results are sensitive to our choice of priors by running the model under different combinations of prior distributions and contrasting the resultant posterior distributions. When analyzing data, we rarely think that all possible supported parameter values are equally plausible: trees are unlikely to be a trillion units long, and evolutionary rates probably will not vary by dozens of orders of magnitude. Why *not* incorporate this prior information into our analysis in the most principled possible way?

Regularization Beyond, encoding our prior beliefs about the data-generating process, priors also serve an important role in preventing undue excitation of the likelihood function in regions of parameter-space that fit especially well to the data, despite not being intrinsically plausible. In so doing, priors help us to tread the line between overfitting and underfitting, balancing error resulting from the model learning from noise in the data with error from the model being insufficiently flexible. In phylogenetics, priors can help us tease apart otherwise non-identifiable parameters whose combination forms a ridge in the likelihood surface, such as clock rates and time during divergence time estimation. They also help us to avoid capture of the model by singularities in that surface, such as when estimating both the value and parameters of a normally distributed random variable (for example, if jointly estimating tip states and the branch lengths separating them under Brownian motion). Hierarchical model structure also lets the model itself learn how skeptical it should be about extreme parameter values, adaptively regularizing priors only to the extent justified by the data (Gelman et al., 2012). This strategy finds extensive application in other work, including non-phylogenetic side projects performed in association with this dissertation, but is also used in the catarrhine landmark and discrete dental analyses to flexibly tune the amount of rate variation between traits and branches

permissible in those applications. In this manner, we allow the model to partially pool information between parameters, lending aid from highly concentrated regions of the joint posterior to those more diffuse. Though less of a concern in phylogenetic applications, adaptive regularization also helps us to avoid multiple comparisons problems (Gelman et al., 2012), such as when we do inference over dozens of rates and examine the ones assessed to be more or less confidently determined.

Stability In cases where the likelihood surface is multimodal, Bayesian inference may more easily traverse the distance between peaks than a stochastic optimization algorithm, arriving at more robust estimates of model parameters than mere search for phylogenetic optima. When performing empirically parameterized simulation studies, sampling from the posterior distribution of an analysis of real-world data also helps to better represent the problematic nature of real-world data than the idealization of trees samples from parametric distributions or the simulation invariance of using a single phylogenetic Maximum Likelihood Estimate (MLE). As we use numerical algorithms to approximate joint posterior distributions, we also find ourselves endowed with the tools to assess their reliability. A purported MLE might actually represent a local maximum, with little diagnostic recourse but to compare the output of multiple independent runs, but MCMC output can be screened with many devices to ensure healthy chain mixing and convergence over and on the target distribution.

1.4 Prospectus

A table of contents appears at the start of this work, but it is also fitting to describe what is to come in prose. The ultimate goal of this work was to adapt the threshold model of quantitative genetics to the purpose of inferring phylogeny, and in particular the topologies of trees scaffolding the evolution of high-dimensional ordinal character data (Chapter 4). The primary empirical application of this work was an alignment of ordinal dental traits codified in the Arizona State University Dental Anthropology System (Turner et al., 1991) and collected by Dr. Shara Bailey. Before that could be done, we first needed to develop and implement a multivariate Brownian model of character evolution, as well as explore its

statistical properties in simulation (Chapter 2). Here, we applied the method to a publicly available, open-source dataset of craniofacial linear measurements collected by William W. Howells (Howells, 1973, 1989, 1995). In between these, we also compare the performance of our multivariate Brownian motion model to a menagerie of methods commonly used to infer phylogeny using morphology on a cranial landmark dataset collected on 13 different species of catarrhine primate by Katerina Harvati (2004) and colleagues (Chapter 3). Trees from these analyses could be compared to a well-resolved reference tree obtained from a published analysis of molecular data (Arnold et al., 2010), that the yardstick of method performance not be exclusively fictitious, in the sense of agreement with some simulated ground-truth, but also relate in some way to biological reality.

Mathematical details are scattered throughout these chapters, though hopefully not in too overwhelming or tiresome a frequency. Three supplemental appendices (A, B, and C) are also included in the closing sections detailing further details that the main text could not stand to bear. A concluding section (Chapter 5) summarizes key takeaways and provides a roadmap for follow-up research.

Chapter 2

Bayesian Phylogenetics Under Multivariate Brownian Motion

Nikolai G. Vetr, Michael R. May,
Brian R. Moore, Timothy D. Weaver

2.1 Abstract

The field of statistical phylogenetics has progressed tremendously in recent decades. However, the development and application of molecular evolutionary models stand in stark contrast to limited adoption of corresponding models of morphological evolution. Given the nature of data primarily available to paleontologists and paleoanthropologists — shapes and sizes of bones and teeth — insufficient exploration of morphological models has impeded statistical inference of phylogeny among fossil taxa too old for the preservation of ancient DNA. Here, we present and perform inference under a stochastic multivariate Brownian diffusion model informed by quantitative genetics to describe the evolution of continuous traits. Unlike previous work, our model explicitly accommodates non-independence among traits, which may more realistically reflect biological phenomena such as pleiotropy and developmental integration. We then explore the statistical behavior of this model through analysis of simulated data and evaluate its performance with respect to retrieving simulating phylogenetic model parameters, with particular focus on tree topology. Finally, we apply our method to an empirical dataset consisting of 57 craniometric measurements collected by William W. Howells on a globally distributed set of modern human populations. Posterior distributions obtained from this analysis were consistent with accepted relationships in these groups. Several extensions of this model are also in development that will allow it to flexibly accommodate a wider range of data, such as polymorphic discrete morphological traits, and join it with molecular data to better shed light on evolutionary relationships linking both fossil and extant taxa. [Statistical Phylogenetics; Morphological Evolution; Bayesian Inference; Multivariate Brownian Motion]

2.2 Introduction

Statistical phylogenetics has undergone tremendous methodological progress in recent decades (Holland, 2013). Most efforts, however, have centered on the description and implementation of discrete-state continuous-time Markov chain (CTMC) models of nucleotide substitution, with comparatively little attention paid to the development of morphological evolutionary models. When genetic data are available, both paleontologists and neontologists can fit molecular phylogenetic models in their framework of choice to make inference of the phylogenetic relationships linking taxa of interest, as well as many other phylogenetic model parameters. But for inferring phylogeny, paleontologists can only make direct use of molecular phylogenetics insofar as they can obtain viable ancient DNA – degradation quickly limits the temporal depths from which taxa can be drawn (Collins et al., 2002; Allentoft et al., 2012; Pickrell and Reich, 2014). Past those limits, they must use morphology to infer phylogeny.

Methods for inferring phylogeny using morphological characters are limited (Holland, 2013; Lee and Palci, 2015). Currently, both model-based and heuristic methods are commonly used. The former rely on mathematically explicit descriptions of evolutionary process and can be fitted in both Bayesian and Maximum Likelihood frameworks. For morphological characters, the most popular inference model is Lewis' (2001) Mk-model, which is the k-state generalization of the Jukes-Cantor (1969) model of nucleotide substitution (when $k = 4$, the models are equivalent, as the state-space of DNA is 4. For discrete binaries – morphological traits that can be present or absent, large or small, and so on – $k = 2$). Here, a discrete morphological trait changes state according to a continuous time Markov process where rate is independent of current state and waiting times between transitions are exponentially distributed. Rate heterogeneity between independent traits is straightforward to accommodate (Wright et al., 2016). Further generalization can allow for non-independence between traits by specifying states as combinations of traits and unequal rates of change between those combinations, but this rapidly swells the state-space even for modest numbers of traits (e.g. 10 non-independent binaries will have 2^{10} possible states). Correctly fitting this model under maximum likelihood produces a point estimate

of tree topology, branch lengths, and evolutionary model parameters that corresponds to the parameter values under which the observed data are most probable (the Maximum Likelihood Estimate, or MLE). Conversely, in a Bayesian framework one's target of inference is the entire joint posterior distribution of phylogenetic model parameters, often approximated with Markov chain Monte Carlo (MCMC).

However, morphological variation between groups is frequently not discrete, but continuous. While distances (e.g. Euclidean or Mahalanobis distances based on variation in the continuous characters) can easily be computed between sets of continuous characters, inference under the Mk-model (Lewis, 2001) requires that we discretize any continuous observations. This discretization can be done in a variety of ways (Garcia-Cruz and Sosa, 2006; Thorpe, 1984) in part subject to researcher preference, with some methods discarding more phylogenetically relevant information than others (Brazeau, 2011). But even discrete characters collected at the outset may just be discretizing some fundamentally quantitative feature (Wiens, 2001), relying on a researcher's present observations and prior experiences instead of an explicit algorithm run on the character alignment in its entirety. As such, we may desire statistical methods for inferring phylogeny that can make direct use of continuous characters without discretizing them.

Brownian motion represents a mathematically and computationally tractable description of continuous character change and can be shown with quantitative genetics and appeal to the central limit theorem to approximate a variety of evolutionary processes under a polygenic, additive model, including evolution at mutation-drift equilibrium, fluctuating selection, and constant directional selection (Hansen and Martins, 1996; Weaver, 2018; Harmon, 2018).

Briefly, consider that under polygenicity, continuous character values are binomially distributed, and when the number of loci underlying a polygenic trait is high, the normal approximation to the binomial holds (via the central limit theorem, as binomial random variables are sums of independent Bernoulli random variables, which have finite variance). Since the difference between two normal random variables is itself normal, a Brownian motion model (described in further detail below) can be a good approximation of an

evolutionary process acting under the above conditions. First described empirically by Brown (1828) and mathematically by Einstein (1956), it has enjoyed considerable attention in the context of phylogenetic comparative methods (Felsenstein, 1985b), where it is often contrasted with more parameter-rich Gaussian models, such as the Hansen model (Butler and King, 2004; Beaulieu et al., 2012), to better understand the evolutionary processes that structure morphological variation when the tree on which the characters evolved is assumed to be known. Recently, the statistical performance of a univariate Brownian motion model has been explored through simulation (Parins-Fukuchi, 2017). Here, moderate numbers of independent continuous traits contained sufficient information to reliably retrieve the simulating tree's topological structure under varying degrees of model misspecification, though as correlation between traits increased in the simulating model so too did error under the misspecified inference model.

As phenomena such as integration, pleiotropy, linkage disequilibrium, and correlated selection are likely to structure the evolution of continuous traits (Cheverud, 1996; Young and Badyaev, 2006; Mitteroecker and Bookstein, 2007; Klingenberg, 2008, 2014) across a variety of taxa (e.g. Parsons et al., 2011; Klingenberg et al., 2012; Neaux et al., 2018), it may not always be safe to assume each trait to be an independent realization of some univariate Brownian motion process. Moreover, explicitly working within a multivariate framework can also enable more nuanced reconstruction of both fossil (in the case of fragmentary remains – i.e. missing data) and ancestral phenotypes (e.g. see Appendix C) and permit the direct investigation of correlated evolution between traits, integrated over phylogenetic uncertainty.

Here, we describe an approach for inferring phylogeny from continuous-character data under a multivariate Brownian diffusion model that allows them to evolve non-independently. Then, we explore the statistical performance of the model and the validity of its RevBayes (Höhna et al., 2016) implementation using a comprehensive simulation study. Finally, we apply the method to well-studied empirical dataset consisting of cranial measurements collected on extant human populations (Howells, 1973, 1989, 1995).

2.3 Materials and Methods

2.3.1 Simulating from the Model

Univariate Brownian motion (uvBM) describes a stochastic process where displacements to the value of some continuous character are drawn from univariate normal distributions. When undirected, the means of these distributions are zero and their variances are equal to the product of two scalars, one representing the branch length v_i over which evolution occurs and the other representing the rate of continuous character change, σ^2 (Figure 2.1). Multivariate Brownian motion (mvBM) generalizes this process to multiple characters. Under mvBM, displacements to the values of some vector of n continuous characters are drawn from multivariate normal distributions, with mean $(0_1, 0_2, ...0_n)$ and covariance matrix given as the element-wise product $\Sigma = v_i R$, where R is a square, symmetric, positive-semidefinite covariance matrix describing the rates and correlations of evolutionary change, collectively known as the multivariate Brownian rate matrix. A univariate Brownian motion of multiple characters can be seen as a special case of mvBM where all off-diagonal elements of R are constrained to be 0; conversely, non-zero off-diagonal elements specify non-independent character evolution. It is common to decompose R into the matrix product SCS, where S is a diagonal matrix of standard deviations (constrained to be > 0) with elements $\{\sigma_1, \sigma_2, ...\sigma_n\}$ specifying the square root of the rates of evolution under mvBM, and C is a correlation matrix specifying their nonindependence. As a correlation matrix, C's diagonal elements are defined to be 1, its off-diagonal elements restricted to the range (-1,1), and the resulting square, symmetric matrix C constrained to be positive-semidefinite. As the diagonal elements of S increase, corresponding characters experience greater potential for evolution; as the off-diagonal elements of C stray further from 0, corresponding pairs (or modules) of traits evolve in an increasingly correlated fashion.

Multivariate Brownian motion can also describe correlated character evolution on a phylogenetic tree. On a phylogeny, lineages multifurcate at internal nodes and continue on independent evolutionary trajectories – displacements drawn on one lineage subsequent to multifurcation are independent of the states observed or displacements drawn at other

lineages. From some starting (e.g. root) state, one can simulate along a branch by adding to that state a draw from a normal distribution whose variance is the aforementioned product. Upon reaching an internal node, character evolution along each daughter branch can be further simulated by additional independent displacements from separate normal distributions. Alternatively, to simulate a vector containing end states across all the tips (or internal nodes, positions along branches, etc.) of the phylogeny, one can make a single draw from a multivariate normal distribution whose mean is the state at the root and whose covariance matrix is the phylogenetic covariance matrix (O), with elements representing the shared evolutionary history of each pair of tips, or the sum of those branch lengths shared by each pair of tips stretching back to the root. Diagonal elements of the O matrix are the height of corresponding tips above the root of the tree, as each tip shares the entirety of its evolutionary history with itself. For example, one can sample a vector of character data $x = \{x_A, x_B, x_C\}$ for a single trait evolving under univariate Brownian motion with rate σ^2 on the tree depicted in Figure 2.2 by sampling from a multivariate normal distribution with mean $\{x_E, x_E, x_E\}$ and covariance matrix:

$$\sigma^2 P = \begin{bmatrix} \sigma^2 v_1 & 0 & 0 \\ 0 & \sigma^2(v_4 + v_2) & \sigma^2 v_4 \\ 0 & \sigma^2 v_4 & \sigma^2(v_4 + v_3) \end{bmatrix}$$

Conversely, multivariate Brownian motion (mvBM) generalizes the above process to multiple characters evolving non-independently. Instead of sampling from univariate normal distributions when simulating along branches, one samples from multivariate normal distributions whose means are a vector of zeros and whose covariance matrices are equal to the product of each scalar branch length and some covariance matrix representing the rates and correlations of evolution in a set of continuous characters (i.e., R). To simulate a vector X containing states for all the characters across all the tips of the phylogeny, one can perform a similar pre-order traversal or, alternatively, take a single draw from a multivariate normal distribution whose mean is the state at the root and whose covariance

matrix is given by the Kronecker product of R and O. The elements of X will be in order of sets of characters or tips, depending on the order in which one took the Kronecker product (for $R \otimes O$, elements of the X are grouped first by characters and then by tips). The uvBM process, then, can be seen as a special case of mvBM where all the off-diagonal elements of R are fixed to 0.

2.3.2 Bayesian Inference

We implement the above described multivariate Brownian motion process in a Bayesian framework to approximate the joint posterior distribution of phylogenetic and evolutionary model parameters, which may include S, C, and V, described above, given observed character data X. It is common to represent the phylogenetic covariance matrix O as two distinct parameters: a discrete tree topology Ψ and a vector of appropriately indexed branch lengths V. The joint posterior density of model parameters can then be described by

$$P(S, C, \Psi, V|X) \propto P(X|S, C, \Psi, V)P(S, C, \Psi, V)$$

which states that this density is proportional to the product of the likelihood, a term that gives the plausibility of observing X given parameter values, and the joint prior density, which represents our beliefs about those parameter values prior to observing data. Computation of the likelihood is described below, and prior densities can be specified through one's choice of prior distributions, which can in turn be influenced by external information or by a desire for regularization. The sensitivity of the results of an empirical analysis can and should be evaluated with respect to one's choice from a reasonable set of prior distributions. As a probability distribution, the integral of $P(S, C, \Psi, V|X)$ across its range must equal 1, so we may divide the right-hand side of the above expression by $P(X)$, also known as the marginal likelihood, which represent the probability of the data, X, averaged over the prior $P(S, C, \Psi, V)$. In practice, closed-form solutions of $P(X)$ are difficult to express, so we instead exploit the above expression of proportionality to sample values from the distribution $P(S, C, \Psi, V|X)$ in proportion to their density using

MCMC.

2.3.3 Data

The vector of observations, X, is a $p\ x\ t$ vector of continuous characters, where p represents the number of observed nodes in the tree and t represents the number of observed characters corresponding to each node. This vector can be represented by a matrix of dimension (p, t), analogous to a nucleotide or amino acid sequence alignment in molecular phylogenetics. The order in which one constructs the vector (grouping together observations either by node or trait) is arbitrary, but must correspond to the order in which one takes the Kronecker product described earlier. It is common to specify a model of geometric Brownian motion, where the element-wise log of some vector of continuous characters undergoes (untruncated) Brownian motion, rather than the real-valued continuous characters themselves. This implies a multiplicative epistasis at the underlying gene level and is often done to restrict trait evolution to positive values only, though they can still grow arbitrarily large.

2.3.4 Likelihood

The likelihood function under multivariate Brownian motion is equivalent to the probability density function of the multivariate normal distribution described above, $\mathrm{N}(\mu, R \otimes O)$, and evaluates to a number proportional to the probability of observing some vectorized continuous character alignment conditional on matrices R and O. It is given by the expression

$$\left|2\pi\Sigma\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

where x is the stacked vector of tip characters, μ is the state at the root (corresponding to node E in the above tree), and $\Sigma = R \otimes O$.

Often, this root state is unobserved – we can attempt to infer its location and value by making strong assumptions regarding the rates of character change, by specifying a prior on the root state (including a point-mass prior, where the root state is known

without uncertainty), or by marginalizing over all possible states at the root — in effect, specifying a uniform$(-\infty, \infty)$ prior on it, as we implicitly do for states along branches and at internal nodes. As normal distributions are symmetric, the Pulley Principle holds (Felsenstein, 1981) and the probability under Brownian motion of going from state i to state j is identical to the probability of going from state j to state i. As a result, the computed likelihood is invariant to the placement of the root, and one can calculate an identical likelihood by re-rooting at any position of the tree. This allows us to perform inference under unrooted trees and root according to our criterion of choice later (e.g. along a branch leading to an outgroup or at a node we have specified an informative root-state prior on). In other words, the trees depicted in Figure 2.3 — with observed states corresponding to nodes $\{A, B, C\}$ and unobserved states to $\{D, E\}$ — have identical likelihoods.

A likelihood that marginalizes over all possible root states can be called a "restricted" likelihood (or sometimes a "residual" or "reduced" likelihood, Felsenstein 2004). Conveniently, this also shrinks the space of possible trees we need to integrate over, equivalent to dropping any of the terminal nodes of our tree.

Repeatedly computing densities of high-dimensional multivariate normals is computationally expensive, so we may wish to re-express the likelihood as a product of univariate normal densities. This can be done by diagonalizing both the O and R matrices and propagating the appropriate transformations in the likelihood function to the vector of tip states. Two algorithms will serve our purposes here. The first exploits the structure of the phylogenetic covariance matrix, O, by performing a post-order traversal of its corresponding tree and iteratively pruning off and computing densities of contrasts between neighboring tips, replacing each internal node's pair of daughter lineages with the conditional (on the values of its descendants) multivariate normal distribution of the state at that internal node (Figure 2.4). This algorithm, also called a "pruning" or "peeling" algorithm, was first described in a phylogenetic context by Felsenstein 1973; 1985b; 2004, though in itself is just a special case of the sum-product algorithm applied to trees (Höhna et al., 2014). It is directly analogous to the identically named algorithm also developed by

Felsenstein used to accelerate likelihood calculation under nucleotide substitution models. Implicit to it is the diagonalization of the O matrix by some transformation matrix that pre- and post-multiplies it, i.e.

$$Q = COC^T$$

We can solve for O and substitute $C^{-1}QC^{-T}$ in to the multivariate normal PDF above, and through simple linear algebraic manipulation move each of C^{-1} and C^{-T} to transform x into y, leaving the diagonal Q in place of Σ (where the elements of y correspond to the values of the contrasts, and the elements of Σ to those contrasts' branch lengths; see Appendix A for further mathematical details). The result of this procedure is a phylogenetic covariance matrix Q whose off-diagonal elements are 0, corresponding to variances of the transformed character vector y (i.e. the independent contrasts). This transformation subtracts out the parameter of the model corresponding to the root state, so if performing inference with a 'known' root, one must either re-root the tree elsewhere, or else include the root state in the vector X, with corresponding elements in O set to zero.

Implementing this algorithm – in effect, populating the C matrix above, but in practice constructing Q and y directly – is straightforward. Starting with some alignment of continuous characters at the tips, find a pair of sister tips. In the rooted 3-tip above, we can use any pair, but let us choose $\{B, C\}$. Compute the difference of their character states (x_B–x_C; or alternatively $x_C - x_B$) and set it as the first element(s) of y. Set the sum of the branch lengths separating ($v_2 + v_3$) as the first (diagonal) element of Q. Then, prune off the two terminal branches leading to B and C, and at the previously internal node representing their most recent common ancestor construct a new, virtual tip whose value is the normal distribution conditional on the values of B and C. As normal distributions are additive in variance, this has the effect of extending the branch length of that virtual tip by the variance of the conditional distribution, and settings its value as the conditional distribution's mean. For some multivariate normal distribution with mean and covariance

matrix partitioned as

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

and realizations partitioned as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The distribution of x_1 conditional on observed value(s) x_2 is also multivariate normal, with mean and covariance:

$$\mu_1' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\Sigma_{11}' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

In the above example, taking $x_1 = x_D$, $x_2 = x_C$, $\mu_1 = \mu_2 = x_B$, and

$$\Sigma = \begin{bmatrix} v_2 & v_2 \\ v_2 & v_2 + v_3 \end{bmatrix}$$

we can solve for $\mu_1' = x_B + v_2(v_2 + v_3)^{-1}(x_C – x_B) = (x_B v_3 + x_C v_2)/(v_2 + v_3)$ and $\Sigma_{11}' = v_2 - v_2(v_2 + v_3)^{-1}v_2 = v_1 v_2/(v_1 + v_2)$. These describe the conditional normal distribution of that ancestral node – we can add its variance to the length of its subtending branch, and set as the virtual node's value its mean. This procedure can be repeated until all the branches have been pruned and one is left with a fully populated vector y

and matrix Q (whose dimensions are all one less than those of the original x and O, an effect of marginalizing over the unobserved root state), corresponding to n–1 independent contrasts and their variances.

Through this decomposition and diagonalization of O, we transform our original $R \otimes O$ covariance matrix into a diagonal block matrix containing n–1 scalar products of R matrices, requiring us to still calculate n–1 multivariate normal densities and ensuring that compute time for a single likelihood calculation is roughly linear with tip number. We can further improve efficiency by diagonalizing R so that each of these multivariate normal densities is rewritten as the product of univariate normal densities. This can be done through many decompositions, but here we choose a Cholesky decomposition of R so that $R = LIL^T$, where L is the lower Cholesky factor of R and I is an identity matrix of appropriate dimension. Substituting this expression in for R, we can propagate L^{-1} through to premultiply x, though we must also take care to divide each contrast's density by the determinant of R (see Appendix A for mathematical details). During inference this decomposition and transformation must be reapplied each time the rate matrix changes (e.g. a proposal is made to the rate matrix during MCMC), but otherwise if most proposal weight is on parameters other than the rate matrix, compute time becomes approximately linear with trait number, plus any overhead in having to keep track of these further "transformed" tip values. In the analyses that follow in RevBayes (Höhna et al., 2016), we saw an approximately 40-fold reduction in compute time over using the pruning algorithm alone when conditioning on a particular rate matrix (for analyses of 128 characters, with more improvement under higher dimensionality and less under lower). This approach can also improve efficiency when sampling rate matrices, if R or its subcomponents are parameterized through their Cholesky factor, as can often be the case when specifying an LKJ prior on the correlations of R.

2.3.5 Priors

The algorithms described above are independent of one's choice of prior, and can indeed be applied just as well in a Maximum Likelihood inferential framework. Nevertheless, working in a Bayesian framework requires that we specify prior distributions on all unobserved

parameters of the model, both for the simulations and for the empirical analyses that follow. We describe these distributions below. In this context, priors can help us to specify our beliefs about model parameters, accommodate inferential uncertainty, and regularize inference away from singularities in the likelihood surface under more parameter-rich models (e.g. when attempting to infer ancestral states jointly with branch rates, instead of marginalizing over them).

For the two sets of simulation experiments described below, we specify two sets of priors. In the first, we place a pure-birth (Yule, 1925) prior on the tree and branch lengths with speciation rate (λ) = 1, and tree height given as a point-mass prior on the ground truth height of the tree on which the simulation was performed. Additionally, we placed a point-mass prior on the rate matrix equivalent to that used in simulation, as this initial experiment focused more on the informativeness of continuous characters on tree topology, independent of uncertainty in the rate matrix. This lets us best explore retrieval of the topology parameter under optimal inferential conditions.

For the empirical analysis, we specified a more diffuse prior on branch lengths and topology – specifically, a discrete uniform prior on topology, a log-uniform (10^{-3}, 10^{2}) prior on tree length, and a flat Dirichlet prior on branch lengths. These were chosen to avoid "long-tree" problems identified elsewhere (Brown et al., 2009), though for this dataset we found the posterior distribution with respect to topology to be largely insensitive to branch length prior specification. In contrast to the "strict morphological clock" used above, this corresponds to a relaxed clock where individual branch rates have more freedom to vary. Here, we once more used a point-mass prior on the rate matrix, fixed to the sample pooled-within-group phenotypic covariance matrix calculated from our dataset.

In the final simulation experiment, we used the same priors as in the empirical analysis described above.

2.3.6 MCMC

Inference was done in a Bayesian framework using the statistical phylogenetics environment *RevBayes* (Höhna et al., 2016). The Metropolis-Hastings algorithm (Hastings, 1970) was used to sample from the joint posterior of our phylogenetic model parameters, given

that model, priors, and simulated continuous character alignment. This machinery is fundamentally identical to that used for phylogenetic inference from nucleotide sequence data, as well as in other Bayesian inference problems. Preliminary runs with Metropolis coupling (3 heated + 1 cold chain) resulted in no improvement to MCMC mixing, so all subsequent analyses were uncoupled to save on compute costs. Longer chains were run initially to identify adequate sampling intensity; final analyses were run for 35,000,000 iterations, with the first 7,000,000 used to tune proposal mechanisms to ensure intermediate acceptance probabilities and subsequently discarded as burn-in. *RevBayes* allows for flexible specification of MCMC proposals, so a variety were used (see example script in supplemental information), with approximately half the proposal weight given to moves on topology and half given to moves on branch lengths. Two independent chains were run for MCMC diagnostics and their output later concatenated for downstream analysis. To compare inference under the simulating model to inference under a misspecified model representing univariate Brownian motion, we further replicated all analyses with off-diagonals of the inference rate matrix set to 0.

2.3.7 Simulations

We performed a series of simulation experiments to explore the statistical performance of this model, simulating continuous character alignments evolving under multivariate Brownian motion on trees drawn from several sets of trees. These first simulations were performed on trees with 30 tips sampled from a pure-birth (Yule, 1925) distribution (i.e. extinction rate = 0) with speciation rate = 1, though the scale invariance of Brownian motion would permit any birth rate, so long as differences between tips were far enough above machine precision and branch length priors during inference placed sufficient density on appropriately long branches (e.g. if we sampled trees with incredibly low speciation rates but placed low prior probability on "long" trees, we might be led astray). We then simulated sets of (2, 4, 8, 16, 32, 64, 128) characters on these trees using rate matrices of simple correlation structure – all diagonals were set to 1 and all off-diagonals were equal and set to either (0, 0.2, 0.4, 0.6, 0.8, 0.99) in order to explore the impact of the number of characters and the strength of character correlation on the phylogenetic informativeness of

continuous character data. To explore the effects of improperly assuming that tip values are known without error (e.g. measurement, taphonomic, environmental, etc.), we further perturbed the state for each tip at each trait by an i.i.d. normal random variable with mean 0 and variance 0.5, equivalent to extending each terminal branch of the tree by $\frac{1}{2}$ and evolving the trait vector via Brownian motion under a rate matrix equal to the identity matrix. We performed equivalent analyses on the unperturbed character alignments as well as these further perturbed tips without explicitly accommodating the perturbation in our inference model. To average out the effects of simulation variance, we performed 500 replicate simulations for each combination of simulating model parameters (trait #, correlation strength, and the P/A of noise). Analyses of these simulated datasets were done under two inference models – one where the character evolution process was perfectly specified and we conditioned on the true, data-generating rate matrix, and one where we conditioned on the true trait-specific rates (the diagonal of the rate matrix) but falsely conditioned on a univariate Brownian motion, fixing all off-diagonals to 0.

A second set of simulation experiments explored performance under more "empirically-realistic" simulation conditions and is described below. Simulating models here were parameterized using MCMC output from our empirical analysis, and we further explored finer-grained rate-matrix misspecification than merely performing inference under a univariate model when the simulating model was fully multivariate.

2.3.8 MCMC Diagnostics

Several diagnostic criteria were used to ensure MCMC health and identify pathology. Many MCMC diagnostics rely on visual inspection (e.g. marginal histograms, rank plots, and trace plots) which would have been onerous to evaluate across millions of possible parameter-parameterization combinations, so instead we evaluated performance according to less qualitative criteria. Specifically, using the *CODA* package (Plummer et al., 2006) in R (Team, 2013), we computed effective sample sizes on the following values, requiring that each fall above 1,000: patristic distances between all of pairs of tips, Robinson-Foulds (RF Robinson and Foulds, 1981a) and Felsenstein-Kuhner (KF; 1994) distances from the true, data-generating tree, the computed likelihood value, the posterior density,

the tree length, and Colless' Index. Additionally, we computed the squared correlation coefficients on between-chain nodal probabilities appearing at frequencies above 0.005 in either chain, requiring that they be > 0.95. MCMC analyses failing any of these criteria (approximately 5% of total runs) were rerun with triple the total run length. Those that failed these longer runs (approximately 2% of total runs) were excluded from the results and new simulated data analyzed to achieve precisely 500 replicates per condition.

2.3.9 Output Analysis

Having obtained MCMC output and evaluated chain health, we sought to evaluate inferential performance, largely through consideration of its aggregate "frequentist properties" (*sensu* Huelsenbeck and Rannala, 2004). First, we constructed calibration curves (sometimes called "reliability" curves) for nodal probabilities falling within equally spaced bins of 0.1 width. This procedure is commonplace in the machine learning literature when evaluating probabilistic classification algorithms, but phylogenetically has seen infrequent application. Specifically, we calculated the frequency with which a bipartition with probability falling in some bin was present in the true, data-generating tree, averaged across analyses. Perfect calibration would see observed frequency match inferred posterior probability, with plotted points falling along the 1:1 line, indicating that nodal probabilities are trustworthy and that the inference machinery is working correctly, conditional on the model being well-specified. Underconfidence would see true frequencies greater than posterior probabilities; overconfidence less. We examined nodal probabilities because data uninformativeness and the breadth of tree-space resulted in so diffuse a joint posterior that posterior probabilities for particular topologies were very small and unstable (e.g. the MAP tree was sampled at most a small handful of times, even with 128 characters). As the inference model grows increasingly misspecified, we expect to see greater and greater deviation from the state of perfect calibration.

Next, we evaluated how precisely tree topology could be retrieved. Traditional methods of doing so – e.g. computing distances of the true tree from some summary tree – could not be used because the data were so uninformative as to also compromise the stability of those summaries (using more than 128 traits would result in more stable sum-

mary trees, but may represent unrealistic sample sizes for data collection, especially given the need to integrate over rate matrices). Furthermore, summary trees – while useful for visualization – are not the target of Bayesian inference, whose interests instead are in the entirety of the joint posterior distribution of model parameters. Instead, we constructed a plot we have named the "Cumulative Average Resolution" (CAR) curve. Essentially, it compares across sets of analyses the average number of bipartitions that occur above some threshold probability. This corresponds to the number of nodes resolved in a consensus tree with of some probability cutoff (e.g. in the majority-rule consensus tree, that cutoff is 0.5; this is also where we start the CAR curve, as lower thresholds permit the inclusion of incompatible nodes). Examining the height and shape of the curve gives a sense of the typical informativeness of some quantity of data produced by an evolutionary model when performing inference under that or some other model. The curve resembles a staircase for a single analysis, but averaging across analyses we see greater and greater smoothness.

2.3.10 Empirical Analyses

Having explored the model through these initial simulations, we fit phylogenetic multivariate Brownian motion to an empirical dataset comprising linear craniometric measurements. These consisted of 57 linear distances between anatomical locations measured on 2,524 individuals identified as male or female by William W. Howells 1973; 1989; 1995 on crania drawn from a globally distributed sample of 30 modern human populations (Table 2.1). We chose this dataset because it is publicly available, well-known, and is often used to estimate tree-like population structure among modern humans, both by Howells but also more recently (e.g. Roseman, 2016). Conditioning on tree structure without reticulation introduces some misspecification, however; future work to extend the model and incorporate reticulation between lineages in a joint inferential framework is ongoing (following e.g. Pickrell and Pritchard, 2012). In these analyses we conditioned on a rate matrix equivalent to the maximum likelihood estimate (MLE) of the pooled within-group phenotypic covariance matrix (P), adopting Cheverud's conjecture (Cheverud, 1996; Roff, 1995; Reusch and Blanckenhorn, 1998; Waitt and Levin, 1998; Revell et al., 2007) of proportionality between P and the additive genetic covariance ma-

trix (G), which would in quantitative genetics dictate the availability of morphological change under several evolutionary models, but most commonly assumed to correspond to neutrality (Lande, 1976, 1979; Felsenstein, 1988a). The P matrix generally reflects mild-to-moderate non-independence between traits, with off-diagonals of its underlying correlation matrix having a middle 95th percentile interval of (-0.22, 0.69). These assumptions seem reasonable given that Cheverud's conjecture has recently been confirmed for a small subset of human phenotypes (Sodini et al., 2018) and observed differences in the cranium between Neandertals and modern humans seem consistent with neutral evolutionary process (Weaver et al., 2007; Weaver and Stringer, 2015). Concentrating the rate matrix' probability on a single point mass dramatically shrinks the free parameter space of the model and allows branch lengths to be interpretable in units of average within-population differentiation. This "empirical Bayesian" approach is comparable to treating individuals as tips, constraining the monophyly of designated populations, and assuming star-like pedigree structures within populations of height one. To minimize the effects of sexual dimorphism, we performed independent analyses of individuals identified male and female by Howells.

As rates of Brownian motion between lineages were thought unlikely to be constant throughout the tree, we specified a more diffuse prior on branch lengths and topology – specifically, a discrete uniform prior on topology, a log-uniform (10^{-3}, 10^{2}) prior on tree length, and a flat Dirichlet prior on branch lengths. These were chosen to avoid "long-tree" problems identified elsewhere (Brown et al., 2009), though for this dataset we found the posterior distribution with respect to topology to be largely insensitive to branch length prior specification. These would allow for scalar variation in the rate matrix throughout the tree, but not variation in the correlation structure or relative rates of particular traits, though extending the model by either sampling states at internal nodes to numerically integrate over augmented likelihoods or using the modification described by Caetano and Harmon (2019) would permit variation in the structure of the rate matrix itself and better reflect our understanding of the instability of P (Fischer et al., 2016) and G (Arnold et al., 2008).

Maximum Clade Credibility Trees from these two analyses are presented in Figure 2.7. Specifically, these conditioned on a model of arithmetic Brownian motion, where the particles movement is on the scale of the raw distances themselves, without log-transform, to better reflect the additive nature of genetic variation. A separate analysis using a geometric Brownian motion model found inferred tree topologies almost entirely insensitive to flexibility in this respect, with bipartition probabilities in the compare-trees plot falling very close to the 1-to-1 line (Figure 2.8)

2.3.11 Further Simulations

To explore the reliability of these inferences, we performed a subsequent simulation study parametrized with an eye towards biological realism. Here, we simulated according to a multivariate Brownian motion process sets of 2, 4, 8, 16, 32, and 57 characters, and performed inference under the priors described for the empirical analyses described above and a rate matrix conditioned on the empirical P matrix for Howells' data. Simulating trees were drawn from the joint posterior of the arithmetic BM, male-only analysis, as Howells' dataset has more crania that were identified as male than female (1,368 compared to 1,156) and some populations with only males identified (30 populations for males compared to 26 for females), and we favored a priori an additive polygenic model over one that was multiplicative. To assess the effects of model misspecification on the rate matrix "parameter", simulating rate matrices were either taken to be the perfectly-specified P matrix, or else a matrix constructed to reflect differences between geographic regions in Howells' data (as defined by Roseman, 2016) or in differences in the phenotypic variances and covariances of homologous characters measured in humans and chimpanzees.

Only 27 of the 57 characters collected by Howells could be observed in chimpanzees, however (Weaver and Stringer, 2015). To allow the construction of rate matrices with dimension greater than 27, we sampled from a discrete uniform without replacement sets of homologous characters from each group, and computed Mitteroecker and Bookstein (2007) distances between each of their empirical covariance matrices. We then sampled covariance matrices from a Wishart distribution with scale matrix equal to Howells' P matrix divided by some degrees of freedom (df), with df chosen such that, in expectation,

distances between Howells' matrix and sampled matrix would equal distances between subsampled empirical matrices of that dimension. For matrices with dimension greater than 27, we fit via least squares a polynomial regression to distances corresponding to values 1-27 and used the predicted distance to determine what *df* to use for higher trait numbers. It should be noted that this procedure necessarily homogenizes differences between simulating and inference matrices, but should nevertheless provide some indication as to how compromised retrieval of tree topology is under rate matrix misspecification reflecting inter- and intraspecific differences. This ad-hoc procedure may not correspond perfectly to the biological reality of misspecifying R, but can help in interpreting results from analyses performed on fossil datasets when trait non-independence is strongly informed by patterns of integration in closely related, extant taxa. As in the earlier simulations, we performed 500 replicate analyses on each combination of trait number and rate matrix condition, with two independent chains per analysis. Priors and proposals were as those described for the empirical analyses, and each chain was run for a total of 35,000,000 iterations, with the first 7,000,000 used for tuning and discarded as burn-in.

2.4 Results

Calibration curves for analyses of characters simulated on pure-birth trees were as expected, falling near the one-to-one line when simulating and inference models matched, regardless of rate matrix correlation structure (Figure 2.5a). Simulation and MCMC variance resulted in slight deviation from the state of perfect calibration, especially in intermediate probability bins where fewer bipartitions were able to inform those bins' inferred and observed values. High probability bins, meanwhile, might suffer more from the approximate nature of finite MCMC output, as the chain may get "stuck" at a particular state and appear to give higher posterior probability than it would were it allowed to run longer. When character values were perturbed by additional noise not specified in the inference model, inferences grew increasingly poorer calibrated, with miscalibration proportional to the strength of correlation between traits (Figure 2.5b). This results from i.i.d. "noise" obliterating more and more of the phylogenetically informative differences between trait values as those differences grow smaller with increasing correlation. Accommodating the possibility of noisy error is straightforward via estimating the magnitude of noise pooled across tips and sampling "true" population means, though likely at the expense of apparent confidence.

Conditioning on a univariate Brownian motion (uvBM) inference model also results in miscalibration, with greater overconfidence as non-independence between characters grows larger and increasing similarity between characters is interpreted as stronger and stronger evidence for shared evolutionary history. This is observed under both the noisy and the noiseless analyses, though the inclusion of i.i.d. noise to tip values does not affect the calibration of the univariate Brownian motion model under a given correlation structure. This is because noise in simulation actually results in the uvBM inference model better representing the simulated evolutionary process, as perturbing tip means by independent normals is equivalent to evolving them under uvBM. This fit between "noise" and uvBM does not, however, improve our ability to retrieve phylogenetic model parameters (Figure 2.6), because any "evolution" occurring along the terminal branches of a tree is evolution that is, by construction, not shared between any set of tips in the analysis. Examining

CAR curves at moderate correlation strength (e.g. 0.4, Figure 2.6), we see that the univariate inference model returns more bipartitions at higher posterior probabilities – the resolution of a consensus tree with given probability is almost always higher, regardless of trait number. But this resolution is misleading, as mismatch between simulating and inference models results in overconfident ascertainment of those bipartitions (Figure 2.5), with the numbers of "true" bipartitions resolved being substantially lower (Figure 2.6).

To help visually interpret results from our empirical analysis, we construct and show MCC (Figure 2.7) trees for each of the analyses performed on Howells' craniometric dataset, with inference done separately on male and female individuals (as identified by Howells) to avoid confounding sexual dimorphism with differences due to unshared population history. Sex-specific posterior distributions appear to be quite distinct for each within-sex analysis, with almost entirely non-overlapping multidimensional scaling (metric principle coordinates) plots and compare-trees plots far from the 1-to-1 line (Figure 2.8). Additionally, geometric and arithmetic models produced very similar trees, suggesting insensitivity to this modeling decision. In cases where MCMC output differs substantially between models (e.g. in analysis of a broader range of continuous character data), formal model comparison and averaging tools such as Bayes Factors and reversible jump MCMC can be applied. Both superficial and deep internal nodes were often well resolved, but intermediate nodes frequently saw low posterior probabilities, indicating uncertainty about population structure at those levels.

Additional simulations involving simulating trees drawn from the joint posterior of Howells' arithmetic mvBM of identified males showed reduced resolvability (Figure 2.10) at each corresponding trait number, likely due to the increased flexibility of our branch length and topology priors. The empirical cumulative resolution staircase from the above described empirical analysis plotted below the mean curve observed for simulated data (Figure 2.10), likely due to misspecification of either the bifurcating process of population diversification or the Brownian evolutionary model (e.g. Figure 2.5). Misspecification of the rate matrix to reflect intra- and interspecific differences in the P matrix induced miscalibration in the posterior, though it was not too large in magnitude and far less than

that produced by assuming a univariate model (Figure 2.9). Furthermore, the simulating model conditioned on the "perfectly specified" rate matrix was itself slightly overconfident, though we would not expect perfect calibration here, because simulating and inference models still differed in the distributions with which they specified trees and branch lengths under all conditions.

2.5 Discussion

We have described above and in RevBayes implemented an efficient calculation of the phylogenetic multivariate Brownian likelihood function, and in a Bayesian inferential framework explored its performance at retrieving phylogenetic model parameters under a range of simulating model conditions. Here, we have validated both the theory and our implementation of mvBM by observing calibration curves that lie close to the 1-to-1 line under matching simulation and inference models, indicating that nodal posterior probabilities are trustworthy. Overall, mvBM performs well in phylogenetic contexts using moderate numbers of continuous characters and appears to be fairly robust to violations of modeling assumptions when correlations between traits are not too great. Conversely, a much greater penalty to calibration is produced when multivariate characters are assumed to be independent realizations of a univariate Brownian process, especially when correlations between characters during simulation are high. Otherwise, we see diminishing returns to the inclusion of additional characters in our ability to resolve the true, underlying phylogeny, as expected, but with plenty of room for additional characters to inform reconstruction. Assuming proportionality between the mvBM rate matrix and the within-group phenotypic covariance matrix, as we did in the empirical analysis, inherently accommodates finite within-group sampling by extending terminal branch lengths by a small amount, with some minor penalty to the resolution of internal nodes and with negligible cost to calibration. Meanwhile, failing to accommodate noise in the tip data during inference harms both calibration and resolution, and especially calibration under a multivariate Brownian model, because simulating noise from i.i.d. normal distributions is equivalent to evolving tip characters under univariate Brownian motion. Thinking carefully about what measurement error model to layer over the evolutionary process one conditions on for inference may be invaluable for our ability to trust Bayesian output.

In fitting the model to human craniometric data, it is likely that multivariate Brownian motion on a strictly bifurcating, non-reticulate tree is an incomplete description of the processes that gave rise to modern morphological variation. But in light of recent arguments in favor of cranial neutrality and P-to-G proportionality and past investiga-

tions into the "treeness" of Howells' data, we find that conditional on the model we obtain sensible results – a rediscovery of geographical proximity and broad consistency with trees reported from microsatellite (e.g. Pemberton et al., 2013), allele frequency (e.g. Pickrell and Pritchard, 2012), and nucleotide sequence (e.g. Li et al., 2008; Mallick et al., 2016) data. Populations from Africa, Asia, Oceania, Europe, and the Americas all appear to cluster together with moderate-to-high posterior probabilities, as seen in studies using genetic data, though internal branches linking these regions are less certain in their structure. In the trees reported here, we artificially rooted using the Sān (coded "BUSHMAN" in Howells' original naming) population as outgroup, despite inference being performed under unrooted trees. We did this for ease of interpretation, lest readers be unfamiliar with unrooted trees. We also attempted rooting with a Neandertal outgroup, but missing data and ambiguous sex assignment coupled with low sample sizes resulted in long branches leading to the Neandertal tip and low posterior probabilities for Neandertal inclusion in any bipartition, though when each Neandertal specimen was allowed its own tip they confidently grouped together. Although the boundlessness of Brownian motion ensures that characters never reach saturation, the placement of relatively long branches still remains ambiguous if all other branches are substantially shorter, as the multivariate normal distribution describing evolution on that long branch is so diffuse as to no longer affect the likelihood with respect to alternative placements. However, our Neandertal-inclusive analysis was done with complete cases only, effectively discarding 20 of the 57 traits not able to be measured on Neandertals (Weaver and Stringer, 2015). Further work to phylogenetically impute those missing values may better resolve a Neandertal root.

Inference in this empirical case was done using males' and females' sub-datasets in independent analyses and using both males and females in the same analysis, treating each (sex, population) pair as its own tip. Generally, inferred population trees between these analyses broadly agreed, and in the latter analysis male and female tips of the same population almost always formed high-probability sister pairs. Nevertheless, we urge caution when interpreting trees inferred on this dataset, given aforementioned concerns about model misspecification.

Future applications of mvBM in phylogenetic contexts are plenty. The model can be ideally applied towards paleontological and paleoanthropological datasets, where molecular data is fundamentally unobtainable. It can be used in a partitioned analysis featuring both continuous morphological and molecular sequence data for extant taxa, so that the latter might be used to inform our inference of the processes describing the former, such as the recent application of multivariate Brownian motion to divergence time estimation (Álvarez-Carretero et al., 2019). Multivariate methods are especially relevant with the recent proliferation of automated and semi-automated landmarking technologies, which tend to produce large, highly correlated morphological datasets that are relatively cheap to obtain, allowing for intensive sub-population sampling to estimate pooled within-group P matrices. The phylogeny itself need not even be a focal parameter under mvBM models — its application can be motivated by questions regarding patterns of covariation in continuous character evolution or between-lineage rate heterogeneity, as arbitrary levels of non-independence or rate heterogeneity are trivial to accommodate.

Working within a Bayesian framework allows for straightforward accommodation of phylogenetic uncertainty and is reflected in lower nodal probabilities where the data are uninformative, and these probabilities are trustworthy insofar as the multivariate Brownian model is a decent approximation of the true, data-generating process. While other methods can also produce distributions of trees – maximum parsimony can generate sets of nearly most parsimonious trees, and nonparametric bootstrapping can be performed in a maximum likelihood or distance matrix framework to explore consistency within the dataset under resampling with replacement, none are quite so interpretable as posterior probabilities. Furthermore, multivariate Brownian data is fundamentally incapable of being resampled, as each observation within the alignment is just a single, multivariate point, and any resampled individual traits would by definition be perfectly correlated with themselves. This sampling procedure, meanwhile, may be more consistent and objective than that used to collect less well defined discrete binary or ordinal morphological data, where researcher judgment may play a role in evaluating the degree of expression of a given feature, and instead automated or semi-automated measurement technologies may

more easily be brought to bear.

Extensive exploration of the model's performance and its extensions remains, providing fertile ground for further work. For both computational and analytical simplicity we here conditioned on a particular mvBM rate matrix, equivalent to loading all the prior density for those parameters on a single point. Empirically, this can be valid under the assumption of Cheverud's conjecture when sample sizes are large, as pedigree structure near the tips of the tree will dominate the likelihood relative to the far fewer branches connecting separate populations. Further simulations will explore how well topology can be retrieved when the rate matrix is not assumed to be known without error – e.g. how much resolution is lost when the correlation structure of the rate matrix is given a diffuse LKJ prior? Or when an informative (e.g. Wishart) prior is used, parameterized using the MLE P matrix? Perhaps the tree can be treated as a nuisance parameter, and inference instead focus on inferring the structure of the mvBM rate matrix, integrating over phylogenetic uncertainty. Alternatively, non-scalar rate variation may be allowed throughout the tree, and one may attempt to infer the locations and magnitudes of shifts in the rate matrix, potentially with tip-specific P matrices serving to inform priors on the rate matrix in local regions of the tree? Also, how well can continuous data be used to infer reticulation on the tree or other departures from a strictly bifurcating model? Recent attentions have been paid to the inference of phylogenetic networks and hybridization edges (e.g. Wen et al., 2016; Wen and Nakhleh, 2018), including under Brownian motion (Pickrell and Pritchard, 2012), and allowing the possibility of reticulation in a joint inferential framework could both improve model specification and allow for the inference of parameters of great interest to paleontologists and paleoanthropologists.

The overarching multivariate Brownian framework presents many additional opportunities for empiricists. Simultaneously modeling morphological and nucleotide sequence evolution enables more nuanced reconstruction of both divergence times and phylogeny. In the analyses presented here, all datasets were complete and assumed to be known without error. In a Bayesian framework, sampling unobserved values through data augmentation is straightforward – since fossil fragmentation is just a special case of missing

data, imputation of missing fossil morphology can be done in a phylogenetically sensible way. Studies of character coevolution can integrate over phylogenetic uncertainty by estimating parameters of the mvBM rate matrix without conditioning on any particular topology, with inferred correlations between traits able to be read off the joint posterior directly. Minor extensions of the model can also allow us to jointly estimate both ancestral states and histories, as well as between-lineage rates of character evolution under contrasting morphological clock models or as functions of adjacent models (e.g. those describing mass-extinction events). A multivariate probit model can be layered atop multivariate Brownian motion to extend the univariate threshold model (e.g. Felsenstein, 2005, 2012; Revell, 2014) beyond independent discrete characters. Bounded phylogeographic diffusion can occupy a row and column of the rate matrix to not only reconstruct historical migrations informed by morphological patterning but also better investigate the association between geography and body shape and size. These and many other opportunities await further development in the modeling framework of phylogenetic multivariate Brownian motion, and we are optimistic that this and subsequent work will help further expand the applicability of model-based Bayesian inference of phylogeny to continuous-character datasets.

2.6 Figures

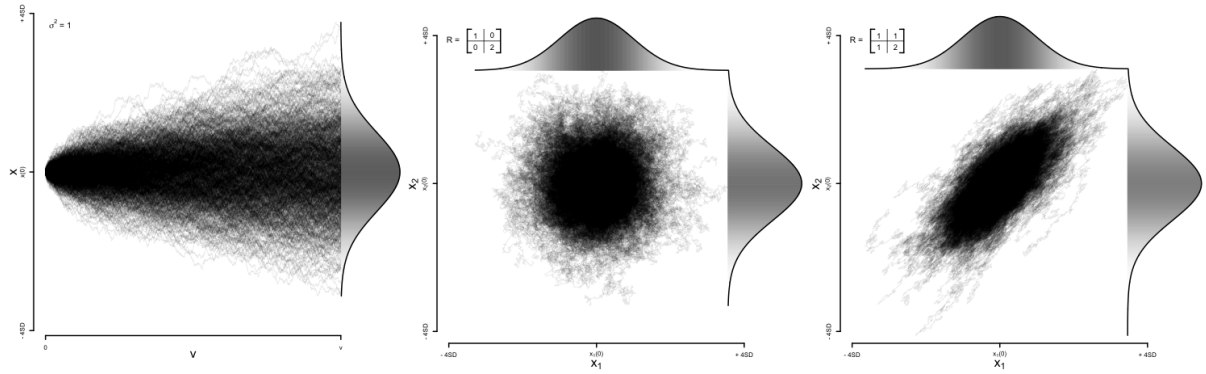

Figure 2.1: Visualizing Brownian motion in one and two dimensions. In a), we see 1,000 realizations from a uvBM process with rate given by $\sigma^2 = 1$ over branch length v. The vertical axis is centered on the starting value of the process, $x(0)$, with bounds +/- 4 standard deviations of the normal distribution describing the ending state. This normal distribution is drawn in the right margin, with shade proportional to probability density at each value x. In b), we see this process generalized to two dimensions for characters x_1 and x_2, with rate matrix R in the upper left corner. As off-diagonal elements of R were set to 0, this is equivalent to a uvBM process of two independent characters. In c), we observe realizations from a bivariate Brownian motion where off-diagonals are non-zero – here, the correlation between characters is $\frac{1}{\sqrt{2}} \approx 0.71$.

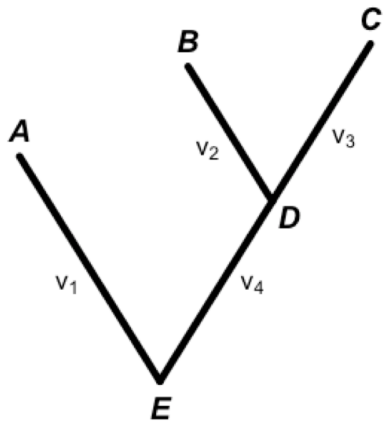

Figure 2.2: A 3-tip, rooted phylogenetic tree with branches $\{v_1, v_2, v_2, v_4\}$ and nodes $\{A, B, C, D, E\}$ labeled.

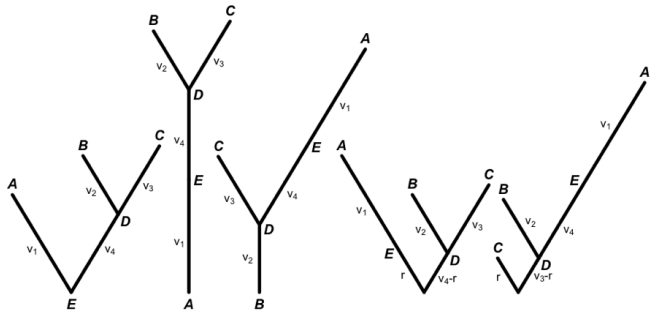

Figure 2.3: Given character data at $\{A, B, C\}$ and labeled edge lengths, all of the above trees have identical likelihoods, marginalizing over unobserved states.

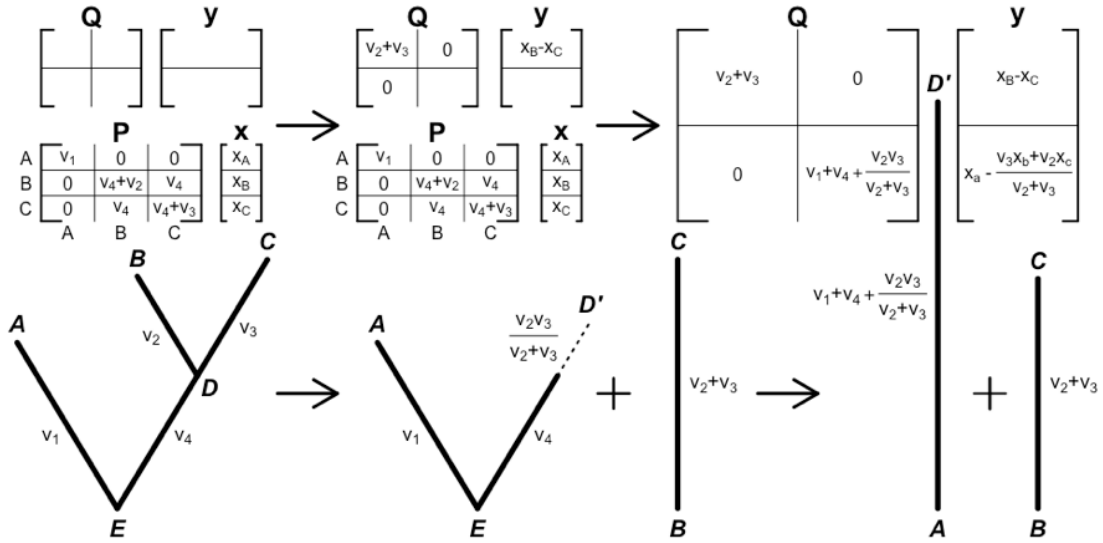

Figure 2.4: The pruning algorithm applied to the tree in Figure 2.2 and used to construct matrix $Q = CPC^T$ and vector $y = Cx$.

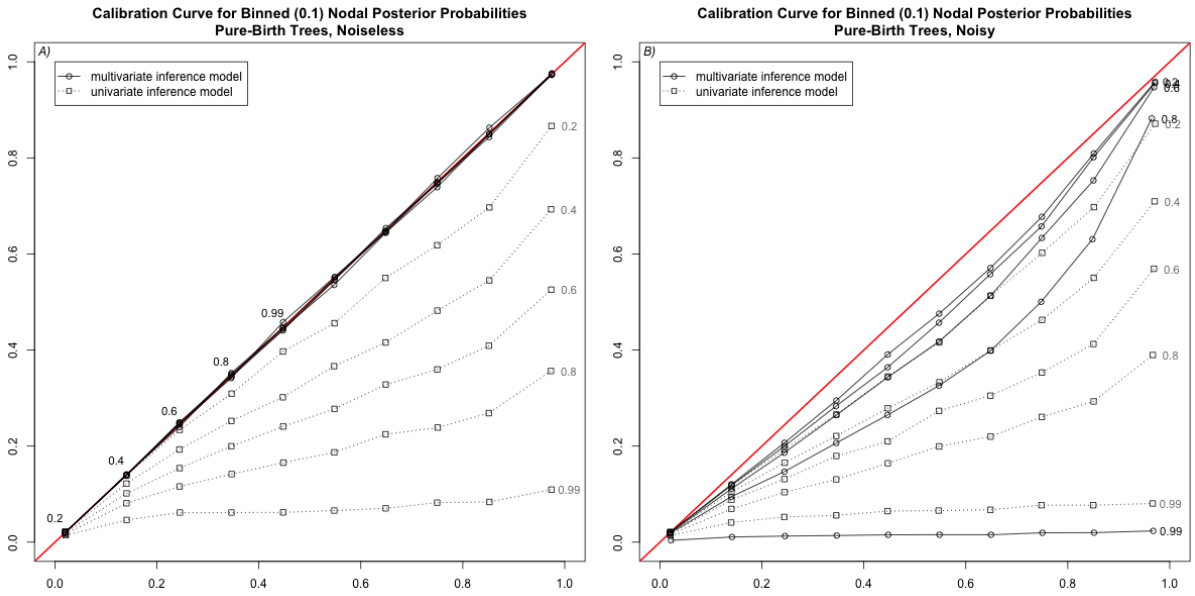

Figure 2.5: Calibration curves for the analyses of pure-birth trees, averaged across replicates and trait numbers. Solid lines represent inference under a multivariate Brownian motion model, with individual lines representing the strength of correlations between traits in the rate matrix. Dotted lines represent calibration curves under the univariate inferential model, with correlation strengths labeled. A line through the origin with a slope of 1 represents perfect calibration. The horizontal axis marks the average probability of bipartitions falling within bins of 0.1 width, and the vertical axis marks the frequency with which those bipartitions appeared in their corresponding data-generating trees. The left plot a) depicts analyses where no noise was added to the tip means; the right plot b) shows analyses where tip values were perturbed by added noise (described in text).

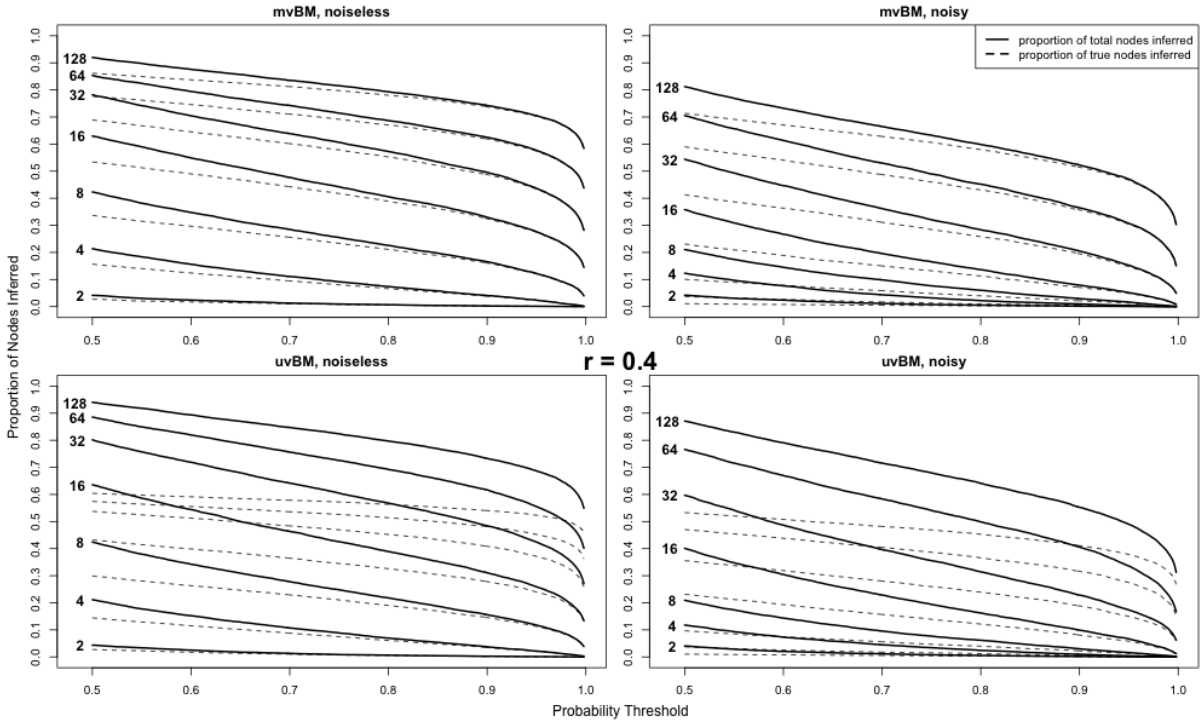

Figure 2.6: Cumulative Average Resolution (CAR) curves for analyses of character alignments simulated under multivariate Brownian motion on pure-birth trees. Here, we see curves corresponding to moderate correlation structure (correlations of 0.4 in the multivariate Brownian motion rate matrix); plots for other correlation structure conditions can be seen in Supplemental Figure 1. The vertical axis represents the proportion of nodes occurring above some probability threshold, shown on the horizontal axis, averaged across replicates for each trait number. Trait numbers corresponding to each curve appear in the left portion of the figure. Solid lines correspond to the ratio of the total number of nodes inferred above each probability threshold to the maximum number of nodes appearing in the true, unrooted tree, while dashed lines represent only those inferred nodes which are actually present in the true, unrooted tree.

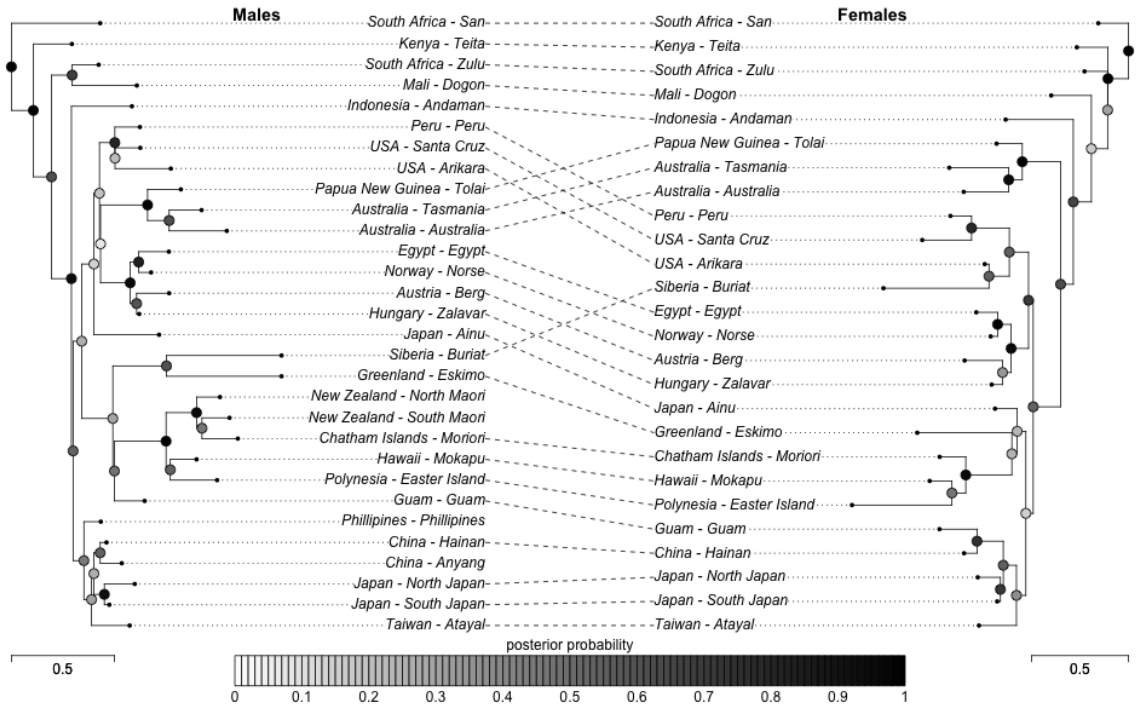

Figure 2.7: Two Maximum-Clade Credibility (MCC) trees are shown, each computed from MCMC output of Howells' craniometric measurements, with the left corresponding to population means of individuals identified as male by Howells, and the right identified female. Nodes have been rotated in each to minimize the mismatch of vertical tip ordering using the `cophylo()` function in Phytools (Revell, 2012). Trees have been rooted with the "South Africa - Sān" population serving as the outgroup for viewing convenience. Posterior probabilities at each internal node are represented by grayscale circles, with key shown. Scale bars are included for each tree to allow visual comparison of branch lengths in units of average within-population differentiation (a single unit of branch length corresponds morphological evolution comparable to the average amount of variation within each tip population; i.e., a draw from the same multivariate normal distribution).

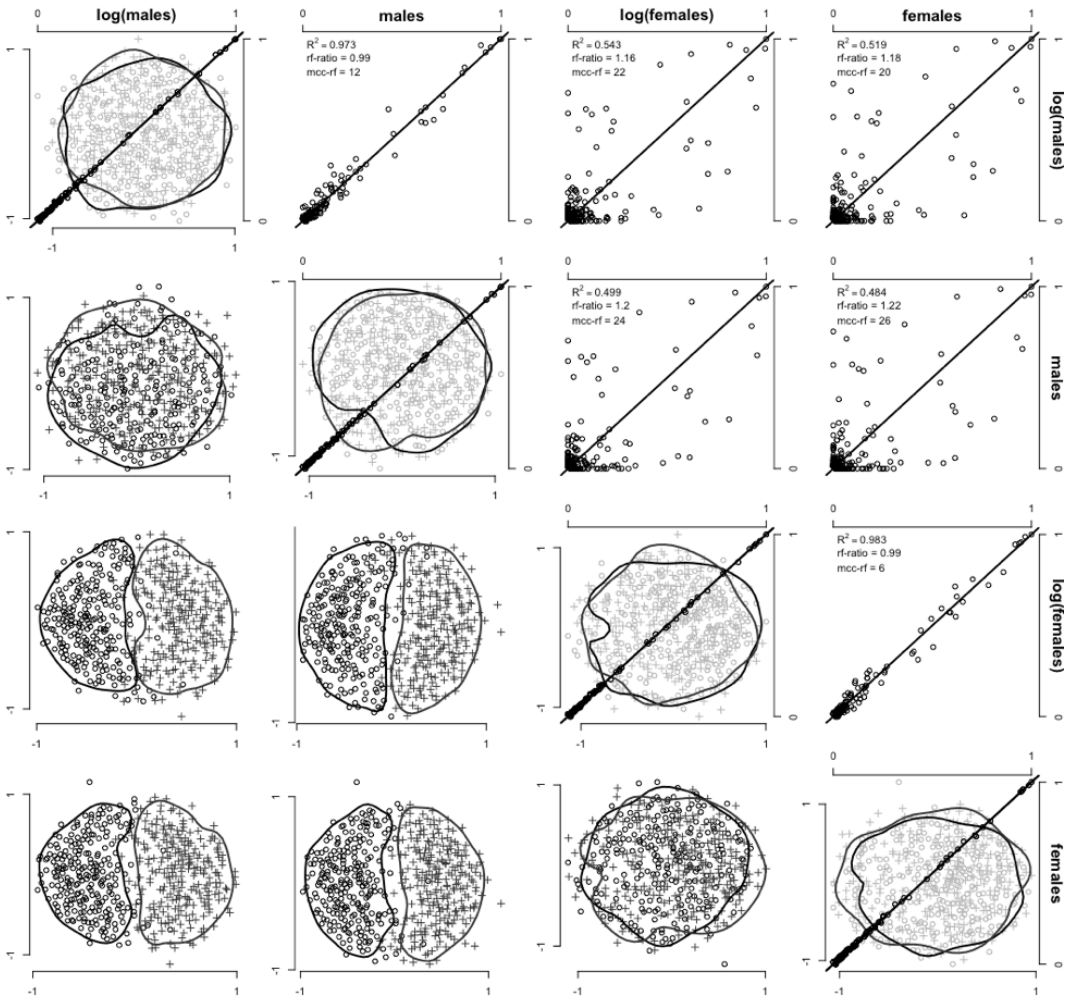

Figure 2.8: Sensitivity to choice of dataset (male vs. female) and model (arithmetic vs. geometric mvBM) is compared through multidimensional scaling (MDS) and compare-trees plots. The former find a set of 2-dimensional coordinates whose distances are closest to a distance matrix provided, in this case a matrix of pairwise RF distances between samples from the MCMC output of analyses represented by the rows and columns. Here, we used 500 trees from each analysis to compute MDS coordinates, but for ease of viewing only plotted 250 points from each analysis. Circles represent the row analysis and + signs the column analysis, with 95% contour envelopes drawn. The latter compare posterior probabilities for the same bipartition (with probability threshold equal to or greater than 0.01) between each analysis, with complete agreement between analyses finding points that fall along the 1-to-1 line. Diagonals represent both types of plot overlain for two independent chains of the same analysis.

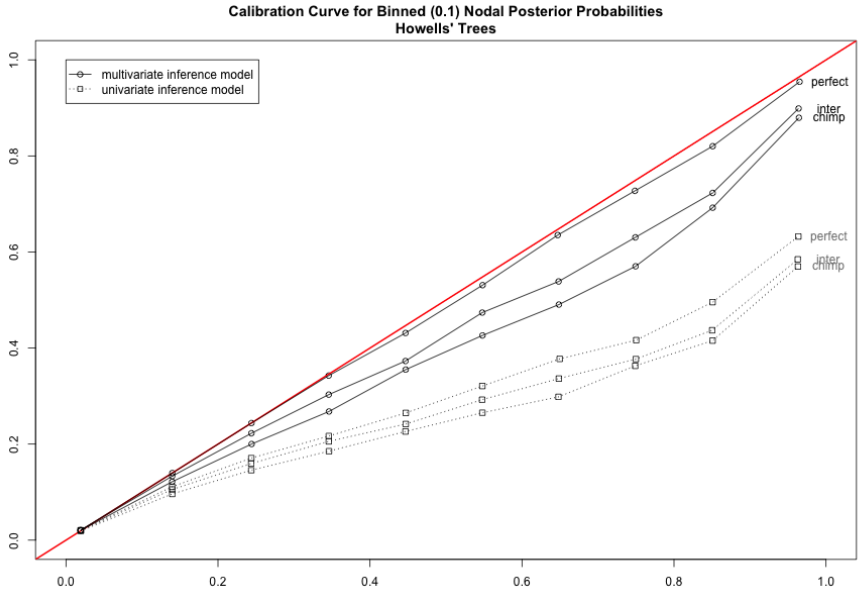

Figure 2.9: Calibration curves for the analyses of Howells' simulations, with rate matrix specification conditions labeled. See main text and caption for Figure 2.5 for further information.

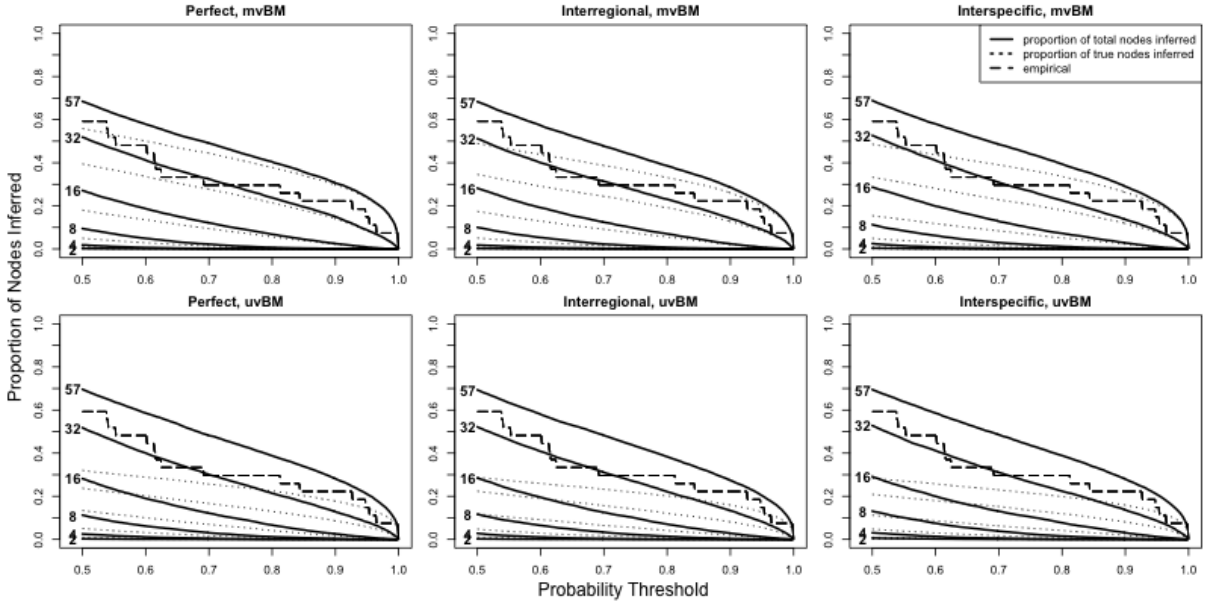

Figure 2.10: CAR curves for the analyses of Howells' simulations, with rate matrix specification conditions labeled. A stairstep-like line for the arithmetic analysis of Howells' 57 craniometric measurements collected on individuals estimated to be male is overlain, though no qualitative difference in the location or shape of this curve across different datasets or models could be seen. See main text and caption for Figure 2.6 for further information.

2.7 Tables

	Females	Males	Total
Japan - Ainu	38	48	86
Indonesia - Andaman	35	35	70
China - Anyang	0	42	42
USA - Arikara	27	42	69
Taiwan - Atayal	18	29	47
Australia - Australia	49	52	101
Austria - Berg	53	56	109
Siberia - Buriat	54	55	109
South Africa - Sān	49	41	90
Mali - Dogon	52	47	99
Polynesia - Easter Island	37	49	86
Egypt - Egypt	53	58	111
Greenland - Eskimo	55	53	108
Guam - Guam	27	30	57
China - Hainan	38	45	83
Hawaii - Mokapu	49	51	100
Chatham Islands - Moriori	51	57	108
Japan - North Japan	32	55	87
New Zealand - North Maori	0	10	10
Norway - Norse	55	55	110
Peru - Peru	55	55	110
Phillipines - Phillipines	0	50	50
Japan - South Japan	41	50	91
New Zealand - South Maori	0	10	10
USA - Santa Cruz	51	51	102
Australia - Tasmania	42	45	87
Kenya - Teita	50	33	83

Papua New Guinea - Tolai	54	56	110
Hungary - Zalavar	45	53	98
South Africa - Zulu	46	55	101

Table 2.1: A table detailing the composition of Howells' linear measurement dataset used in our empirical analysis. Elements of the table represent numbers of individuals in each population corresponding to each estimated column sex. Further details regarding the nature of this sample can be found in (Howells, 1995).

Chapter 3

Primate Phylogenetics with Landmark Data: Exploring Model-Based and Heuristic Approaches

Nikolai G. Vetr, Timothy D. Weaver

3.1 Abstract

Paleontologists and paleoanthropologists have long sought to better understand evolutionary relationships linking fossil and extant taxa. While modern genetic and genomic methods offer ample opportunity for phylogenetic inference, age and degradation quickly limit adequate retrieval of ancient DNA from fossil specimens. New approaches extending phylogenetic models common in the comparative methods literature could conceivably improve inference of fossil phylogeny over more commonly used heuristic methods developed prior to recent decades' advances in computation and theory. Here, we explore the application of a stochastic multivariate Brownian diffusion model to a set of landmark data collected across a range of catarrhine primate species. We contrast results from this fitted model to those obtained from Bayesian analysis of matched molecular data, as well as to results from more conventional, heuristic inference procedures, such as distance-based methods and Maximum Parsimony. A short simulation study is also performed to explore the behavior of the multivariate Brownian method under empirically parameterized conditions. With the advent of increasingly sophisticated continuous character data capture technologies, improved understanding of how different methods perform will help to inform researchers' phylogenetic analyses and better infer the evolutionary relationships connecting both fossil and extant taxa. [Primate Evolution; Bayesian Inference; Multivariate Brownian Motion]

3.2 Introduction

Phylogenies are inferred representations of the branching process of speciation, the splitting of ancestral populations into reproductively isolated daughter lineages. Evolution makes little sense except in their light. But far from merely depicting genealogical relationships among taxa of interest, their inference is essential to the investigation of many evolutionary questions. Perhaps most importantly, they serve as a scaffold for the analysis of character evolution (Felsenstein, 1985b), as shared ancestry confounds relationships between traits observed in taxa evolving along the branches of phylogenetic trees. They also provide a framework for divergence time estimation (Glazko and Nei, 2003; Heath and Moore, 2014), ancestral trait reconstruction (Schluter et al., 1997; Joy et al., 2016), species delimitation and taxonomic classification (Rannala, 2015), historical biogeography (Donoghue and Moore, 2003; Ronquist and Sanmartín, 2011), analysis of lineage diversification (Nee et al., 1994; Morlon, 2014), and much more. They can help to generate, constrain, and discriminate between paleontological hypotheses, guiding our interpretation of the fossil record. Whether some species is ancestor, sister, or descendant to some other species; whether the same character evolved many times, convergently, or just once; whether suites of characters coevolved simultaneously or appeared piecewise; these and other questions can be explicitly explored within a phylogenetic framework.

Likely owing to our own taxonomic affinities and the wide availability of both morphological and molecular data, primates – especially the catarrhine primates – serve as a common target for the application of new phylogenetic methods or the exploration of existing methods (e.g. Reis et al., 2018) and as examples in tutorials demonstrating the use of phylogenetic software (e.g. in Beast or RevBayes; Bouckaert et al. 2019; Höhna et al. 2016). Though most recent analyses of primate phylogeny focus on molecular data (e.g. Perelman et al., 2011; Springer et al., 2012), historical attempts at inference made use primarily of morphological characters, despite their arguable unreliability at retrieving trees compatible with molecular results (e.g. Collard and Wood, 2000; Gibbs et al., 2000; Varón-González et al., 2020). When inferring phylogeny among fossil taxa, paleontologists are often limited to the use of morphological characters, as degradation

quickly limits the ages of fossils from which ancient DNA can be successfully extracted, sequenced, and assembled (Collins et al., 2002; Allentoft et al., 2012; Pickrell and Reich, 2014). Among hominins, for example, attempts to infer phylogeny have made use of a panoply of algorithms – described briefly below – to transform some alignment of morphological character data into a point estimate or set of phylogenetic trees. There exists a long history of these attempts, stretching many decades into the past (e.g. Chamberlain and Wood, 1987; Stringer, 1987; Skelton and McHenry, 1992; Strait et al., 1997), but also in recent years (e.g. Irish et al., 2013; Dembo et al., 2015, 2016; Argue et al., 2017). To further inform practice in these contexts, it is useful to explore the performance of newer, model-based methods for inferring phylogeny using morphological characters, contrasting their output with results from more conventional approaches, such as Maximum Parsimony or distance-based methods. A natural system for such a comparison can be found, of course, in the catarrhine primates.

The most popular inference model for phylogenetically structured morphological data is Lewis' (2001) Mk-model, which can be fit in both Bayesian and Maximum Likelihood frameworks. For four-state characters, it is identical to the Jukes-Cantor (1969) model of nucleotide substitution (as the canonical state-space of DNA is 4), and otherwise generalizes the Jukes-Cantor model to arbitrary numbers of states. This model is well explored in phylogenetic contexts, both in the nucleotide substitution case and for discrete morphological characters (e.g. Wright et al., 2016). Fitting this model under maximum likelihood produces a point estimate of tree topology, branch lengths, and evolutionary model parameters that corresponds to the parameter values under which the observed data are most plausible (the Maximum Likelihood Estimate, or MLE). Conversely, the Bayesian target of inference is the entire joint posterior distribution of phylogenetic model parameters, often approximated numerically with Markov chain Monte Carlo (MCMC).

Other phylogenetic inferential methods commonly applied to morphological datasets include tree-search under the Maximum Parsimony (MP) criterion and some cost matrix, as well as clustering algorithms (e.g. neighbor joining or UPGMA) that accept as input some distance matrix and produce as output a tree with particular properties. Both

typically also produce point estimates, though one can envision procedures (e.g. the non-parametric bootstrap; Efron 1979; Felsenstein 1985a) that incorporate sampling or measurement uncertainty, as one might also do with Maximum Likelihood inference. For MP methods, often there can be obtained a set of distinct most parsimonious or nearly-most parsimonious trees, but how one probabilistically interprets the frequency of clade membership in this set is unclear. Parsimony can also mislead under certain conditions even at the limit of infinite data, such as when evolutionary rates are high or heterogenous (Wright and Hillis, 2014; Wright et al., 2016; O'Reilly et al., 2018), and typically requires the use of independent, discrete characters. Distance methods can calculate distances that incorporate evolutionary models of character change, but also struggle to accommodate inferential uncertainty in a principled manner, especially in a multivariate framework where resampling characters cannot, by construction, supply any further phylogenetically meaningful information.

Frequently, morphological variation between groups is not discrete, but continuous. While distances can easily be computed between sets of continuous characters, inference under the Mk-model (Lewis, 2001) and most parsimony-based methods (Goloboff et al., 2006) often require that we discretize any continuous observations. Continuous forms of MP algorithms do exist, such as Squared-Change Parsimony (Maddison, 1991; Rohlf, 2002), Linear Parsimony (Kluge and Farris, 1969), Manhattan-Metric Parsimony (Swofford and Berlocher, 1987), and others (see Rogers, 1991), but these are almost always applied to quantitative characters in the service of ancestral state reconstruction and not the inference of tree structure itself, and so will not be directly considered here. Discretization can be done in a variety of ways (Garcia-Cruz and Sosa, 2006; Thorpe, 1984) in part subject to researcher preference, with some methods retaining more phylogenetic information than others (Brazeau, 2011; Worthington, 2017). Even discrete characters collected at the outset may just be discretizing some fundamentally quantitative feature (Wiens, 2001), relying on a researcher's present observations and prior experiences instead of an explicit algorithm run on the character alignment in its entirety. Recent decades have also witnessed the rise of new techniques for the collection of continuous morpholog-

ical data, such as surface and CT scanning (Mitteroecker and Gunz, 2009; Adams et al., 2013; Rein and Harvati, 2014). As such, we may desire to better understand statistical methods for inferring phylogeny that can make direct use of continuous characters without needing to discretize them.

Inference under a multivariate Brownian motion model of character evolution may satisfy this desire. Multivariate Brownian motion (mvBM) is a straightforward extension of univariate Brownian motion (Felsenstein, 1973), modifying only that displacements to character states be drawn from multivariate normal distributions, rather than univariate normal distributions. This allows for the representation of phenomena such as pleiotropy, correlated selection, linkage, and integration, which may structure the evolution of quantitative traits. Here, we investigate how well an mvBM model fit in a Bayesian statistical framework is able to retrieve catarrhine phylogeny using 15 3-Dimensional craniofacial landmarks sampled across a collection of 13 catarrhine primate species (Harvati et al., 2004). We compare this inference to those obtained from other parametric models of character evolution, both distance-based and Maximum Parsimony methods, as well as to relationships in this group inferred from molecular data (Arnold et al., 2010). We then perform a short, empirically realistic simulation study under mvBM to explore the performance of the method in its ability to retrieve true, data-generating trees and clades when the inference model is well specified.

3.3 Materials and Methods

Empirical data used in this analysis were drawn from the (2004) work of Harvati and colleagues and consisted of 15 3-dimensional craniofacial landmarks sampled on 13 species of catarrhine primate. Samples were also partitioned at the subspecific level, with 7 populations of *Homo sapiens*, 6 subspecies of *Papio hamadryas*, 3 subspecies of *Pan troglodytes*, and 2 subspecies of *Gorilla gorilla* present in the dataset, for 27 total tips at maximum taxonomic resolution. Given the extent of hybridization known to occur between subspecies and populations — for example, in *Papio hamadryas* (Rogers et al., 2019) — we lumped together all divisions in the data below the species level, that their phylogeny might be better represented by a non-reticulate, strictly bifurcating tree. Sample sizes for extant species ranged from 17 for *Macaca hecki* to 400 for *Papio hamadryas*; the only extinct species in the analysis, *Homo neanderthalensis*, was also present in the fewest number, with 5 sampled individuals. Information regarding the composition of this dataset can be found in Table 3.1, with more information on the precise sample composition and locations of specific landmarks able to be found in the aforementioned paper (Harvati et al., 2004). These data also contained information on each individual's estimated sex, and to avoid the confounding effects of sexual dimorphism we treated sex-specific observations as representing separate, independently evolving partitions, pooling information across partitions with respect to topology but otherwise independently estimating phylogenetic model parameters. In other words, we specified an evolutionary model where male and female morphologies were able to evolve independently from one another, constrained only in that the branching order of speciation events had to be identical for both. Given the non-independent nature of within-species male and female morphology, this may induce some degree of pseudo-data-duplicatory effect, but additional pooling of model parameters was judged inappropriate given further transformations described below.

This dataset has several features that make it desirable for the analyses performed here. The species included are not only highly studied but also of intrinsic interest, with well ascertained molecular phylogenies available for comparison. Additionally, the landmark nature of the data makes exploring the statistical performance of modern multivariate phy-

logenetic methods especially enticing, given the projected proliferation of these datasets as morphologists continue to develop semiautomated landmark capture techniques. Past attempts to infer phylogeny using craniofacial morphology have met mixed success (e.g. Collard and Wood, 2000), so assessing the performance of different methods with respect to both empirical and simulated data may help to inform researchers' decisions when analyzing landmark data from other systems.

Before the data could be analyzed, several pre-processing steps were required. First, we performed Procrustes transformation (Gower, 1975; Dryden and Mardia, 2016) on all data irrespective of sex simultaneously, scaling the set of each individual's landmarks to unit centroid size and iteratively rotating and translating until the Procrustes distances across all individuals' landmarks converged to some minimum value (Bookstein, 1997). This was done using the `gpagen()` function in the *geomorph* package (Adams et al., 2019) in *R* (Team, 2013). Following Mitteroecker and colleagues (2004) and to allow size and not just shape to inform subsequent analyses, log-centroid size was reintroduced as a single additional variable, bringing the total number of traits used to 46 (3 x 15 shape coordinates + size).

However, the evolution of these 46 traits is not independent, and so we might attempt to correct for this nonindependence before attempting any phylogenetic analysis. To do this, we compute the sex-specific pooled within-group covariance matrix of our transformed landmark data, P. Under Cheverud's conjecture (Cheverud, 1996), this should be proportional to the additive genetic covariance matrix, G, which describes the variances and covariances of morphological evolution at mutation-drift equilibrium (Weaver, 2018). However, as Procrustes transformation removes 7 degrees of freedom (1 for translation and rotation in each of the three dimensions, 1 for scaling), the resulting P matrix is rendered non-positive-semidefinite — with seven negative eigenvalues trailing — and therefore uninvertible, which makes it unsuitable for use in, for example, an informative prior for the multivariate Brownian motion rate matrix. To circumvent this limitation, we decompose P into its eigensystem, with columns of the eigenvector matrix ordered according to monotonically decreasing eigenvalues. We then project the Procrustes transformed landmarks

onto those eigenvectors corresponding only to the positive eigenvalues, and further divide them by the square root of each corresponding eigenvalue, effectively transforming the 46 original characters into 39 whose pooled within-group covariance matrix equals the identity matrix. As a further dimensionality reduction step, we examine the scree plot of our eigensystem and discard those later axes comprising the final 1% of the total variance, restricting ourselves to the minimum set of axes corresponding to 99% of the total within-group variance, and reducing the number of characters to be analyzed to 19 in the female partition and 20 in the male. Thus, subsequent analyses assume that evolution can only occur within a lower dimension subspace along these principal axes, which do not span the entire space of our 46 characters, but may nevertheless allow for it to realize most of the variation observed in the empirical data. We relax this assumption and allow for greater non-independence in the Brownian motion model, but otherwise treat these normalized scores as independent outcomes of the evolutionary data-generating process.

To incorporate information from both sexes into the heuristic analyses performed here, we were not able to specify separate data partitions with greater or lesser degrees of parameter pooling, as heuristic methods lack a formal description of the data-generating process. Instead, we treated scores on each of the axes as independent observations for each tip, concatenating them into a single morphological character alignment.

Finally, no female Neandertals were present in the morphological dataset. To allow for the joint analysis of both partitions in phylogenetic software, we required equality between the sets of tips in either partition. Preliminary male-specific analysis found that Neandertals and *Homo sapiens* clustered together with high confidence regardless of method or model used, so to construct a fictitious female Neandertal we found the vector difference between the two species means of male genus *Homo* and displaced the female *Homo sapiens* mean by that difference. This was done after the joint Procrustes transformation but before projection of species means onto each respective P's eigenvectors, with the fictitious female Neandertal then projected onto the female P's eigenvectors and included in the female partition for analysis.

Some methods required additional pre-processing, the details of which can be found

below. In total, we inferred catarrhine phylogeny using three heuristic methods: UPGMA, Neighbor-Joining, and Maximum Parsimony; as well as under four models in a Bayesian framework: univariate Brownian motion, multivariate Brownian motion, Lewis' Mk model, and an ordered Continuous Time Markov Chain (CTMC) model. Each is described in turn.

3.3.1 Distance-Based Methods

Two clustering algorithms served our purposes here – the Unweighted Pair Group Method with Arithmetic mean algorithm (Sokal and Michener, 1958) and the Neighbor-Joining algorithm (Saitou and Nei, 1987). Both were implemented in R as the functions `upgma()` and `nj()` in the *phangorn* (Schliep, 2011) and *ape* (Paradis et al., 2004) packages, respectively. These methods require that we provide them a distance matrix, which they then use to construct either a rooted ultrametric tree or an unrooted tree with identical patristic distances (with elements of the distance matrix equal the sums of branch lengths separating tips). To compute a distance matrix, we needed a measure of distance between pairs of taxa. For this, we used the Euclidean distances between our transformed traits, similar to using a squared Mahalanobis distance (Mahalanobis, 1936) with P serving to standardize distances between tips, as under our transformation of the character data, P became the identity matrix.

3.3.2 Data Discretization

Both the maximum parsimony analyses and many popular parametric evolutionary models require that data be discrete, rather than continuous. A number of approaches were available here, and we explored two. For the first, we followed the divergence coding procedure favored by Collard and Wood (2000, 2001) and described by Thorpe (1984). This assigned each tip a discrete character in the range $(1, 2, ..., 2n_{tips} - 1))$ and may better accommodate differences in the spacing between tips along a continuous scale, with larger differences corresponding to greater required amounts of evolutionary change. Conversely, the second discretized our continuous character ranges into intervals according to Jenks Natural Breaks Optimization Method (Jenks, 1967), implemented via the

`classIntervals()` function in the *classInt* package (Bivand et al., 2020) in *R*. Jenks' algorithm identifies a pre-specified number of breakpoints along a continuous distribution so as to optimally divide that distribution into categories such that total within-category variance is minimized. As a compromise between flexibility and computational convenience, we set a threshold Goodness of Variance Fit (GVF) value such that the average number of categories across traits was closest to four. This allowed different traits to have different numbers of categories, depending on their need for them. If a character could be discretized into two or three states and exceed the requisite GVF, it could effectively "donate" its opportunity for further discretization to a different character not yet meeting the threshold.

3.3.3 Maximum Parsimony

Maximum Parsimony based methods describe a set of algorithms that explore tree-space in search of a tree that minimizes the amount of evolutionary change needed to produce some observation of character data. Typically, evolutionary change is represented in a parsimony score that measures the smallest number of discrete steps compatible with a given discrete character alignment, and taking Collard and Wood (2000) for inspiration, we performed inference under the maximum parsimony criterion using traits discretized under divergence coding.

To search for the optimally short tree, we used the Parsimony Ratchet (Nixon, 1999), implemented as the `pratchet()` function in the *phangorn* package in *R*. Given the ordinal nature of divergence-coded traits, we also supplied a Wagner Cost Matrix whose entries $W_{i,j}$ corresponded to the absolute value of the difference between the i^{th} row and j^{th} column, representing the penalty to the parsimony score of particular transitions (e.g. moving from state 25 to state 13 would contribute a penalty of $25 - 13 = 12$ to the parsimony score). As the parsimony ratchet may be susceptible to capture by local minima, we ran it 50 times from independent starting trees (generated with the `rtree()` function in *ape*), recording the final state of each chain after 500 iterations had passed without the discovery of a lower-scoring tree. The shortest tree or set of trees across these replicates was taken to be the global minimum.

3.3.4 Bayesian Inference

Four evolutionary models were assumed for the Bayesian analyses performed in *RevBayes*, which uses the Metropolis-Hastings algorithm, a form of Markov chain Monte Carlo (MCMC), to approximate the joint posterior distribution of phylogenetic model parameters. These models include the univariate and multivariate Brownian motion model of continuous character evolution, the Mk-model — which generalizes Jukes-Cantor — and an ordered Continuous Time Markov Chain model with equal instantaneous rates of change between adjacent states and rates of 0 elsewhere. These models were used to analyze the transformed landmark data in the case of uvBM and mvBM, and the Jenks-coded traits in the case of the two CTMC models.

In the Brownian motion analyses, we specified an informative prior on the multivariate Brownian rate matrix about the identity, effectively centering it around the empirical P, as expected under neutrality or fluctuating selection by quantitative genetic theory. In the multivariate Brownian analysis, we set independent priors on the correlation and variance components of the rate matrix, with an LKJ($\eta = 20$) for the former and a Dirichlet multiplied by the number of characters in each partition for the later. This had the effect of fixing the *average* rate of each rate matrix to one, allowing their identifiability when also attempting to estimate branch lengths. A $\log_{10}$normal(1, 0.25) + 1 offset hyperprior was placed on the concentration parameters of this Dirichlet distribution to adaptively regularize the degree of rate heterogeneity along each principal component while still specifying an *a priori* preference towards equal rates, relaxing slightly the Cheverud's conjecture assumption. In the *univariate* Brownian analysis, we fixed the correlation component of the rate matrix to the identity, inferring only the diagonal variances, or rates. Strictly speaking, this latter analysis does not represent a univariate Brownian motion over the original landmark data, but rather corresponds to one involving simultaneously greater or lesser degrees of correlation and rate in the traits that load more on each eigenvector. When the estimated rate matrix is precisely the identity, the implied rate matrix is equal to P, to the extent that it is well approximated by recomposition using only those first eigenvectors corresponding to 99% of the sum of the eigenvalues.

The multivariate Brownian analysis, meanwhile, allows for finer-tuned inference of trait correlations and rates in directions not reflected by the eigenvectors of P. A discrete uniform prior was specified for the topology parameter, a $\log_{10}$normal(1, 1) prior for the overall tree length, and a flat Dirichlet(1,1,1,...) for the branch length proportions.

Following convention (Yang, 1996), we allowed gamma-distributed among site rate variation (ASRV) in the CTMC models, approximated by averaging over 4 discrete rate categories specified according to equally spaced quantiles and fixing both of the gamma's shape parameters to equality (thereby constraining the mean rate to be 1). A $\log_e$normal(e, 1) prior was used for the one shape parameter of our gamma-ASRV model. The same lognormal and flat Dirichlet priors for tree length and branch length proportions were used here as for the mvBM analysis.

Analyses under mvBM and the Mk-models were run for 20,000 iterations in each of two independently seeded chains, with a preceding 5,000 iterations discarded as burn-in and used for proposal distribution tuning. Each iteration consisted of 5,370 proposals. Thus, analyses were run for 26,850,000 moves each. Metropolis-coupling (Geyer, 1991; Altekar et al., 2004) with one heated chain was used to improve mixing.

A number of diagnostics were used to diagnose the health of MCMC output. Beyond terminal branch lengths, tree length, the computed likelihood value, and the posterior density, the pairwise correlations, trait-specific rates, and all regularizing hyperparameters, several "fictitious" parameters were coerced from the Newick strings recorded by *RevBayes*. These included patristic distances between all of pairs of tips, Robinson-Foulds (RF; 1981a) and Kuhner-Felsenstein (KF; 1994) distances from an arbitrary reference tree, and the presence or absence of the twenty most common bipartitions with frequency < 0.99. Visual inspections of marginal histograms, rank plots, trace plots, compare-trees plots, and multidimensional scaling plots were performed at first pass; then, effective sample sizes were computed using the `effectiveSize()` function in the *R*-package *CODA* (Plummer et al., 2006) and ensured to fall above 500 for each real and generated parameter, using first each independent chain and then both concatenated. Additionally, R^2 values for all bipartition frequencies in either chain above 0.01 were required to fall above

0.95.

3.3.5 Inverse Analysis

In a follow-up analysis, we explored the inverse problem — given a time-calibrated phylogeny of these 13 catarrhine species and their 45 Procrustes-transformed landmarks and log-centroid size, to what extent do we see morphological evolutionary rate variation in the mvBM process, and how well does the pooled within-group covariance matrix approximate the inferred multivariate Brownian motion rate matrix? Here, we require a well-ascertained, preferably time-calibrated phylogeny of those taxa present in the morphological dataset. Rather than conduct such an analysis ourselves, we instead relied upon trees drawn from the *10kTrees Project* online repository (Arnold et al., 2010), which will also serve as an external check on method reliability. These represent samples drawn from the joint posterior of a Bayesian analysis of primate phylogeny performed on 17 nuclear and mitochondrial genes under a partitioned scheme of GTR-family models selected by *JModelTest* (Posada, 2008). All taxa included in the Harvati dataset were present here up to species designation, with only one subspecies — the Kinda baboon, *Papio hamadryas kindae*, absent, so we subsetted the dataset to include only those species for whom morphological data was present. Then, we generated a Maximum Clade Credibility (MCC) tree via *phangorn* (Schliep, 2011) for these molecular trees, conditioning on it for this follow-up analysis. On this tree we fit an uncorrelated relaxed morphological clock model with parameterization similar to that used to infer topology earlier, involving a flat LKJ($\eta = 1$) prior on the correlation component of the rate matrix, a Dirichlet prior on the relative variance components with $\log_{10}$normal(0,1) + 1 hyperprior on its concentration parameters, scaled by the number traits that the average rate equal 1, a $\log_{10}$normal(1,1) prior on the total branch-specific rates, and another Dirichlet prior with $\log_{10}$normal(0,1) + 1 hyperprior on the concentration parameters to multiply the total branch-specific rates and obtain individual branch-rates. Independent analyses were performed on the male and female tips in the sample, so we did not include the fictitious female Neandertal tip in the female analysis. Four independent, Metropolis-coupled chains were run with 1 heated and 1 cold chain each to achieve adequate sampling, and all the aforementioned MCMC

diagnostics applied to model parameters including branch rates, with the exception of those involving tree topology and branch lengths, as both were fixed in these analyses.

3.3.6 Simulation Experiment

To help identify the degree to which the error observed in these analyses could be expected or else attributed to model misspecification or the various transformations we performed in the name of tractability or convenience, we performed a short simulation experiment in which 46 x 2 = 92 characters were simulated to evolve under multivariate Brownian motion on the time-calibrated MCC tree obtained from the *10kTrees Project*. Branch rates and mvBM rate matrices were drawn from the posterior distribution of the Inverse Analysis described above. To simulate the pseudoreplication inherent to treating male and female tip means as outcomes of independent data-generating process, we averaged sampled rate matrices R_F and R_M to obtain a single matrix, R_{MF}, which we Kronecker multiplied by a 2 x 2 correlation matrix with off-diagonal 0.9, used to represent the coincident evolutionary trajectories of each sex within the same evolutionary lineage. This procedure induced strong covariation in male and female values for the same character within a given lineage, but allowed for empirically realistic degrees of covariation between characters. Sampled branch rates were drawn from the male joint posterior distribution, for convenience, and continuous tip data simulated under the mvBM process. Then, we projected each tip's sex-specific outcomes onto the eigenvectors of their corresponding sex-specific rate matrices, and used as data normalized scores on the first 19 F and 20 M eigenvectors, which we then analyzed under the same priors as the empirical mvBM analysis described above. This was done to more realistically reflect possible mismatch between phenotypic covariances and the covariances of evolutionary change. These eigenvectors, incidentally, comprised on average 94% and 92% of the variance, respectively, suggesting greater non-independence in characters than that found in the sample P matrix, a fact further observed in our querying of the posterior correlation matrix distribution below. As there were no tremendous differences in the performance of the order-preserving model-based methods considered here, and because a major motivation for this work entailed more detailed exploration of the mvBM model, we restricted the results we reported to only those obtained from the

mvBM analysis. This procedure was then replicated 100 times to help disentangle the effects of procedural error from simulation variance. MCMC diagnostic procedure here was identical to that used in the empirical analysis for all diagnostics that did not rely upon visual inspection, i.e. those that came after our first heuristic pass. Any analyses that failed any of these criteria ($\sim 40\%$ of all runs) were re-run with fivefold the number of iterations, which proved sufficient for these remaining runs to pass.

3.3.7 Proposal Distribution

Finally, we describe a substantially more efficient, tunable Metropolis-Hastings proposal distribution for correlation matrices, the motivations for and implementation and validation of are detailed in Appendix B.

3.4 Results

The primary standard by which multivariate Brownian motion is to be judged is empirical: how well it performs at retrieving topologies consistent with those obtained from better explored CTMC models of nucleotide substitution, both absolutely and relative to other methods. Here, we compared the MCC tree from the *10kTrees Project* trees with point estimates provided by the heuristic methods and MCC trees by our Bayesian model-based methods. For the Bayesian methods, we visually compared each morphological MCC tree with the molecular MCC tree using cophylo plots (Revell, 2012, visible in Figure 3.1a). Alternative tree summarization methods either produced qualitatively identical results (e.g. greedy consensus trees, constructed using PHYLIP, Felsenstein, 1993), or were not able to be stably estimated (e.g. MAP trees, as the morphological joint posteriors were too diffuse for even the highest probability trees to appear more than a small handful of times). When available, nodal posterior probabilities were indicated on both trees, but as the molecular analysis concentrated probability on a very small subset of unique topologies, these represent results from the morphological analysis exclusively. However, as these still discard uncertainty with respect to nodes not included in the morphological MCC tree, we also generated compare trees plots to evaluate concordance between morphology and molecules (Figure 3.1b). Owing to the aforementioned high resolution of the *10kTrees Project* posterior (with almost all bipartitions having probabilities very near 0 or 1), these were hard to interpret, and so we additionally examined histograms of morphological bipartition probabilities corresponding to probable ($p > 0.5$) or improbable ($p < 0.5$) bipartitions of the molecular analysis. Finally, we desired to evaluate the degree to which each posterior distribution of trees contained trees neighboring the molecular MCC tree. To do so, we plotted histograms of each posterior distributions normalized Robinson-Foulds distances (Figure 3.1c).

None of the heuristic methods used included a principled means of representing uncertainty about the tree satisfying each respective optimality criterion. For these, we simply generated cophylo plots of each tree with molecular MCC tree opposing, marking those nodes on each tree with their presence or absence on the opposing tree (Figure 3.2).

To further explore how well the mvBM process might capture the evolutionary phylodynamics of cranial landmark variation in catarrhines, we queried the output of our simulation experiment. First, we once more coerced posterior distributions to MCC trees and recorded the RF distance of these point estimates to the true, data-generating tree, a measure representing the minimum number of nearest neighbor interchange (NNI) rearrangements to move between the two. As in the empirical analysis, these were rescaled according to the maximum such number to fall in the interval (0,1). Across simulation replicates, these distances form a distribution of how close a point estimate falls from its target, which can be compared to the empirical result. Histograms of these distributions are shown in Figure 3.3a. We could also visualize the distribution of compare-tree plots and probability histograms relative to the empirical result, which we do to produce 3.3b. Finally, we can explore the distribution of RF-distance distributions from the simulation experiment replicates — the result of that visualization can be found in 3.3c.

To investigate possible drivers of phylogenetic error, we assumed a known time-calibrated tree and fit to it an uncorrelated mvBM relaxed clock model, inferring both the shape of the multivariate Brownian motion rate matrix, as well as independent branch rates. To check whether poorly identified nodes corresponded in some way higher rates of evolution, we visualized branch rate variation in both male and female partitions using cophylo plots (Figure 3.4a), also plotting both against one another as a check of general concordance (Figure 3.4b). To see whether pooled within-group variance-covariance patterns agreed with the inferred and correlations of evolution per Cheverud's conjecture and model specification, we visually contrasted pooled phenotypic variances with inferred rates (Figure 3.5a-b), as well as pooled phenotypic pairwise correlations with the correlation elements of our estimated rate matrix (Figure 3.5c-d).

Finally, we sought to better understand the extent to which our priors — and their lack of, for example, a more sophisticated heterotachy model, relying instead on the adaptive regularization of multilevel model structure — might capture or fail to capture phenomena such as the branch rate variation observed in Figure 3.4. Thus, we identified *fractious* bipartitions in our empirical mvBM analysis, here defined as those bipartitions whose

estimated probability was greater than 0.75 away from the probability of the corresponding bipartition in the molecular posterior distribution. We then examined the distribution of probabilities estimated for these nodes across our 100 simulation replicates, plotting the result in Figure 3.6.

3.5 Discussion

The analyses presented here are limited, but may nevertheless shed light on both the evolutionary dynamics governing variation in catarrhine cranial shape and form, as well as the phylogenetic informativeness of that variation. Though we had initially expected more unambiguous success for the multivariate Brownian model, the empirical results were far more equivocal, with both model-based and heuristic methods performing comparably well as measured by the Robinson-Foulds distance of each analysis' point estimate from the molecular MCC tree. As a whole, all analyses of morphological data were able to successfully recover cercopithecoid and hominoid monophyly (Figure 3.1-3.2), despite the lack of any explicit constraint to that effect, which is reassuring given longstanding recognition of morphological affinities in those groups (Darwin, 1896). Moving shallower in the tree, first into the ape clade, we find the Brownian methods to correctly and confidently resolve the Chimp-Gorilla-Human trichotomy (Bradley, 2008) by first diverging *Gorilla gorilla* from the remainder, where all other methods prefer Chimp-Gorilla monophyly to varying degrees of confidence. However, both Brownian analyses fail to recover the *Pan* clade, grouping *Pan paniscus* in with *Homo*, potentially due to convergent, paedomorphic developmental trajectories in both taxa (Shea, 1983; Williams et al., 2001) or unaccounted-for allometry. Both sets of taxa are subtended by the fastest branch rates on the tree (Figure 3.4a), though this alone does not appear sufficient to account for the degree of phylogenetic error observed (Figure 3.6a). No other model-based or heuristic analyses considered here were able to identify *Pan-Homo* monophyly, though both Maximum Parsimony analyses (Figure 3.2c-d) correctly identified the *Pan* clade, grouping it with *Gorilla*. Conversely, all analyses were able to successfully unite Neandertals and *Homo sapiens*: indeed, their union was the most confident of all in these trees, appearing in every single of the 10,000 sampled by the mvBM analysis.

Of the Old World Monkeys present in these data, all analyses confidently recovered the *Papio-Mandrillus* clade, as well as *Mandrillus* monophyly. Within the genus *Macaca*, however, results were far more mixed. The *fascicularis* group (*Macaca fascicularis*, *Macaca mulatta*) is identified with intermediate probability in the uvBM and ordered CTMC anal-

yses (Figure 3.1), as well as in both Maximum Parsimony analyses (Figure 3.2), with other analyses preferring to intersperse *Macaca sylvanus* in with one or the other, despite it being the earliest divergent macaque lineage in the molecular tree for these data. UPGMA and the ordered CTMC were the only analysis that successfully identified *Macaca* monophyly, with all other analyses preferring a split between *Macaca tonkeana* (and, in the Mk analysis, *Macaca tonkeana-Macaca hecki*) predating that between the other macaques and baboons-drills/mandrills. Tonkean macaques are socially unlike most other members of the genus, their behavior characterized by greater amiability and reduced reliance on strict dominance hierarchies (Thierry, 1985, 2007). Given the role craniofacial morphology plays in social signaling (Brecht and Freiwald, 2012), as well as the recognized divergence in significance of Tonkean macaque expression (Thierry et al., 1989; Pellis et al., 2011), it may be that morphological divergence in this taxon could be misinterpreted as evidence for more distantly shared ancestry. Further evidence for this can be found in the elevated evolutionary rates present in the macaque clade (Figure 3.4), which appear to trouble recovery of the *Macaca hecki-Macaca tonkeana* lineage in simulation (Figure 3.6c).

Both heuristic and model-based analyses were performed in the course of this research. Treating phylogenetic inference as a peculiar, highly constrained and high-dimensional binary classification problem over the presence or absence of key bipartitions, we might initially be tempted to use a standard loss-function to compare approaches, such as cross entropy. However, as the heuristic methods do not provide distributions of trees, and MCMC proves unstable at finite chain lengths in estimating the probabilities of improbable bipartitions, we must instead rely on less principled measures, such as RF-distances to a molecular target tree. In this regard, it's not clear that either model-based or heuristic methods strictly outperform one another — both Maximum Parsimony over Jenks-coded traits and UPGMA achieved degrees of error equivalent to the best-performing model, the ordered CTMC. But model-based methods present some advantages, especially in a Bayesian context, in that they provide probabilistic estimates of bipartitions that do not constrain one to a single point within tree-space (Figure 3.1b). And to the extent that they are combined with outside analyses, they move over a variety of trees, some of which

are closer to the molecular reference tree than any single point estimate (Figure 3.1c). Examining the posterior distribution of RF-distances, there is no clear victor in which falls closer to the molecular tree, though Lewis' Mk model does appear to be a clear loser, with a distribution shifted over 15% higher than the three alternatives. These, in turn, performed similarly in expectation, though the Brownian analyses sampled a narrower distribution of trees, likely because a continuous character carries more phylogenetic information than its discretization.

All analyses failed to retrieve a tree perfectly consistent with the molecular result, and sometimes assigned high probabilities on the order 0.9 to nodes absent in the molecular tree. But the extent to which this represents a *failure* of morphology to estimate phylogeny (Collard and Wood, 2000; Gibbs et al., 2000) is questionable, insofar as simulation and subsequent inference under a multivariate Brownian model presents a degree of error broadly consistent with that observed in comparison to our reference tree (Figure 3.3). While this is not a formal test of model adequacy, it nevertheless provides some reassurance that the error observed in our empirical analysis may not be attributed purely to model misspecification. That said, the analyses presented here made several assumptions regarding the nature of the evolutionary process governing craniofacial variation. Working in a fully multivariate framework allowed us to relax plausibly restrictive assumptions regarding the nature of integration across traits beyond simple correction vis-à-vis Cheverud's conjecture for character non-independence (Álvarez-Carretero et al., 2019; Varón-González et al., 2020), but we still assumed that the face and cranium evolved according to a process that could be well approximated by Brownian motion — in other words, some flavor of neutral evolution or fluctuating selection sufficiently within the bounds of developmental or geometric constraint that lineages would not have much opportunity to struggle against those bounds over the timescales and ecological contexts represented here (see Chapter 1 for more detail). Exploring the extent to which alternative models are better or worse able to retrieve focal phylogenetic model parameters remains an opportunity for future work.

In our subsequent, inverted analysis, we explored how well the pooled within-group

phenotypic covariance matrix approximates a multivariate Brownian rate matrix up to some constant of proportionality, and found that there is indeed some resemblance between them (Figure 3.5). Cheverud's conjecture seems to hold well for narrow subsets of traits in humans (Sodini et al., 2018) and other animals (Roff, 1996), and here we find that posterior mean evolutionary rates and within-group variances in male and female catarrhines do appear to be roughly proportional (Figure 3.5a), with an R^2 on the log-log scale of 0.65 in males and 0.55 in females. However, much of this is driven by the one outlying non-shape related variable, log(centroid-size) appearing in the upper right corner of Figure 3.5a. Removing this variable, R^2 shrinks to 0.52 and 0.39, respectively.

However, posterior means fail to incorporate the full breadth of Bayesian uncertainty in Brownian rate, and under perfect concordance the quantile distribution of within-group variances should appear uniform in the marginal posterior of each rate. It does not (Figure 3.5b), but there does not appear to be much bias in the inference of trait rates in either males or females. Inferred correlations, meanwhile, appear to be similarly related to within-group phenotypic correlations, with R^2 values of 0.60 in males and 0.66 in females, respectively. The scale of these is far reduced in the mvBM rate matrix joint posterior (Figure 3.1c), however, potentially due to the implicit regularizing effect of constraining ourselves to prior uniformity over the space of PSD correlation matrices. At high dimension, there are far more correlation matrices near the identity than far away from it, and 13 taxa's values for those characters over a phylogeny are less able to inform inference of a correlation matrix than hundreds of pairwise observations. The sigmoidal shape seen in Figure 3.5c suggests stronger regularization at low correlation than at high, though the entire range did shrink approximately fivefold from almost the entire (-1,1) space. This is further seen in the quantile plots of Figure 3.5d, though at least the sign of the estimated correlation appears to be consistently estimated and there is no within-group bias visible in favor of positive or negative correlations. Overall, it seems like the use of the P matrix as an informative prior for the mvBM rate matrix is partially justified, as is its use to correct for character non-independence in methods less able to accommodate non-independence in the evolutionary process.

We do not present here a full exploration of the use of multivariate Brownian or other models in the context of landmark data, nor even any sort of definitive statement regarding catarrhine cranial evolution. To the extent that our analyses failed to retrieve trees compatible with molecular result, we see opportunity not to discard the whole endeavor of phylogenetic inference from morphological data, but to develop and specify more biologically realistic models of morphological evolution. Sometimes, however, we may lack the power to do inference over increasingly large sets of phylogenetic model parameters. Phylogenetic signal may not persist long in the face of stabilizing selection (Varón-González et al., 2020), such as described by an Ornstein-Uhlenbeck model, as the signature of shared ancestry is obliterated by lineages stretching against their limits in shape and form. But reality is probably better captured by multiple character optima whose locations drift through time — a process that can itself be well approximated by Brownian motion. Some low-hanging follow-up analyses could, upon noting heterotachy in the morphological clock (Figure 3.4), see if rates throughout the tree covary alongside some plausible discrete character (May and Moore, 2020), such as one encoding dietary behavior or mating system. This might help to prevent error due to short-branch attraction (Philippe et al., 2005), as taxa diverging earlier in time resemble each other more due to lower evolutionary rates on the branches leading to them. If the high rates on some nested branch can be learned from their association with a discrete character, true phylogenetic structure might better be revealed.

Though we restricted ourselves to analysis of tips at or above the species level unlikely to be affected too strongly by incomplete lineage sorting or hybridization, recent work in extending reticulate and coalescent models to character evolution under Brownian motion might benefit analysis at the subspecific levels (Bastide et al., 2018; Mendes et al., 2018), such as if we had chosen to not collapse *Papio hamadryas*' six subspecies to a single tip in light of their tangled population histories (Rogers et al., 2019). Our analyses also assumed constancy of the mvBM rate matrix throughout the tree, up to some constant of proportionality. But as the genotypic covariance matrix is known to change through time (Arnold et al., 2008), we might wish to reflect that change in a model that allows for

structural variation in the mvBM rate matrix (Caetano and Harmon, 2019). Finally, many choices — often with no principled best practice to guide us — were made in the course of these analyses, taking us through a garden of forking paths replete with both obvious and unobvious researcher degrees of freedom (Simmons et al., 2011). How sensitive the results presented here are to the combinatoric explosion of alternative possible decisions is unclear, and so we urge caution in interpreting this work with too much confidence, lest a slight left at its start have taken us in entirely different directions.

3.6 Figures

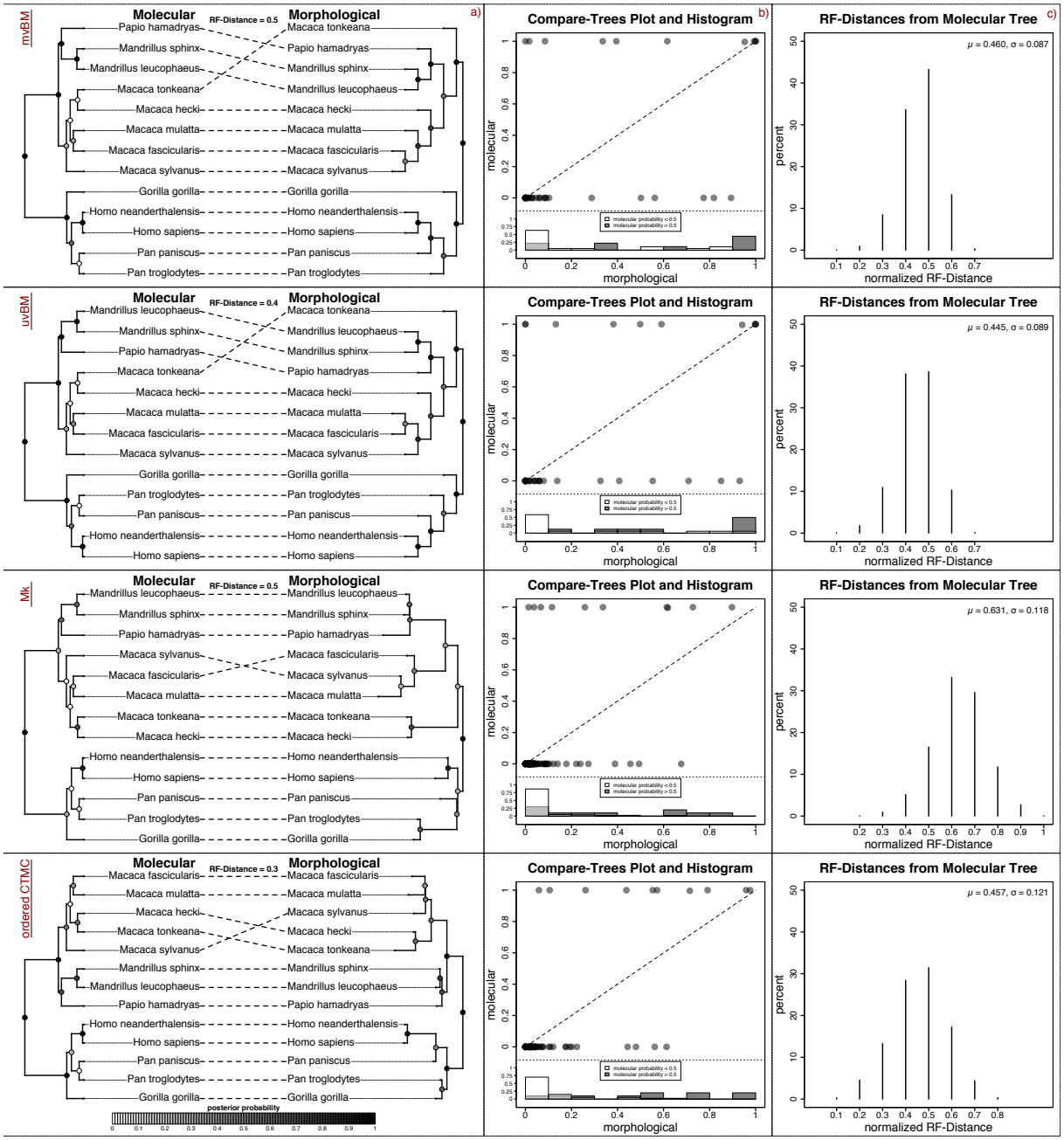

Figure 3.1: In a), cophylo plots of molecular MCC trees opposing those from the four morphological analyses specified above. In cophylo plots, nodes of two opposing trees are rotated as to match tips along the vertical dimension. Both trees were rooted using cercopithecoids as an outgroup. Nodal posterior probabilities from the morphological analyses are marked along a white-to-black gradient. In b), compare-trees plots of bipartition probabilities from the morphological analyses (horizontal axis), and from the 10kTrees Project trees. Histograms of bipartitions from the morphological analyses appear in the lower portion of the graph. In c), normalized Robinson-Foulds distances from the morphological posterior distributions to the Molecular MCC tree.

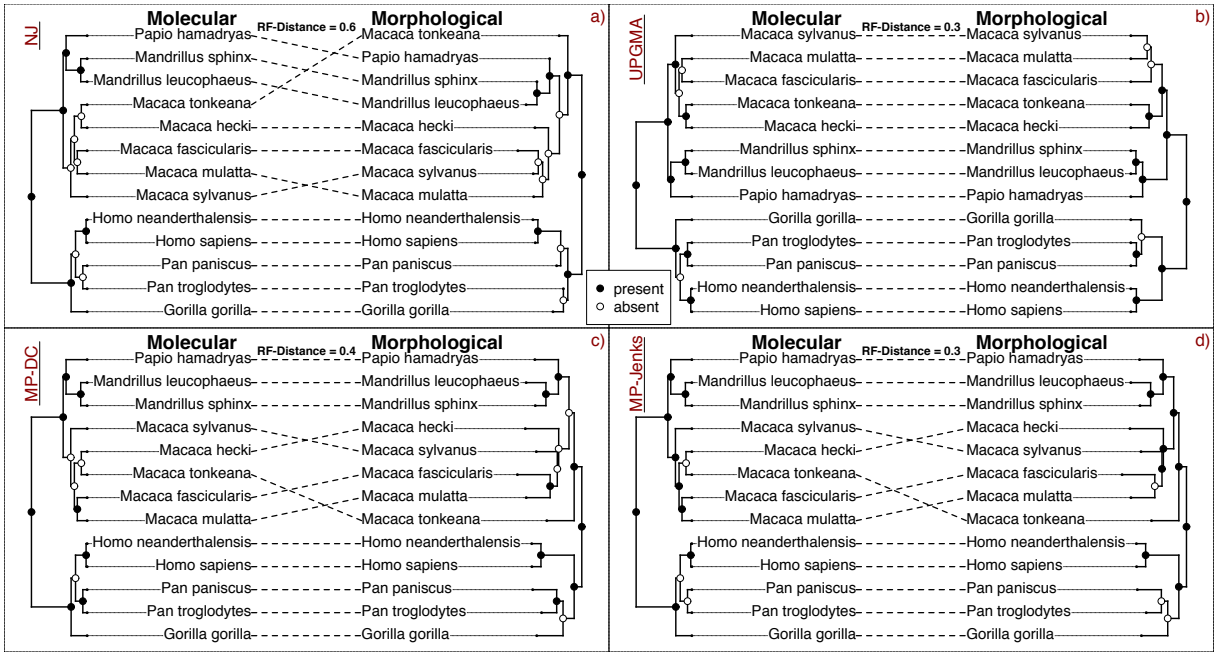

Figure 3.2: Cophylo plots of molecular MCC trees opposing those from the four morphological analyses specified above, with normalized Robinson-Foulds distances shown. Here, each of the morphological trees produced only point estimates, so nodes are marked according to bipartition presence and absence in the opposing tree. In a), the neighbor-joining tree is shown. In b), the UPGMA tree. In c), the maximally parsimonious tree obtained from PC scores discretized using divergence coding. And in d), the maximally parsimonious tree obtained from PC scores discretized using the Jenks Natural Breaks algorithm with 4 expected categories.

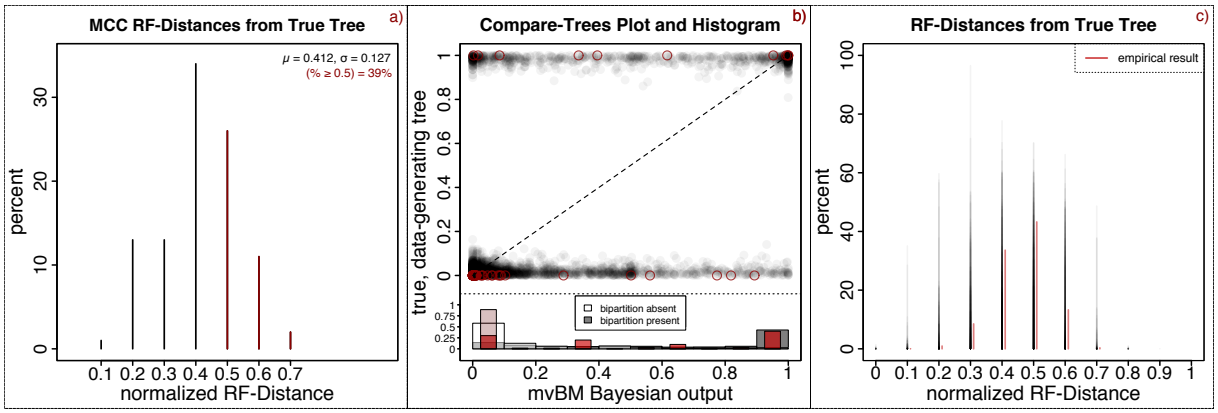

Figure 3.3: In a), we see the distribution of normalized RF-distances of the Bayesian MCC tree from the true, data-generating tree across the 100 replicates of our simulation experiment. The mean and standard deviation of this distribution are given, with distances greater than or equal to our empirical mvBM distance from the molecular tree (0.5) marked in dark red. In b), we construct similar plots across simulation replicates as in figure 1b, jittering probabilities inwards to better represent simulation variance. In the bottom section of the plot, two histograms of the average (across replicates) posterior bipartition probabilities are constructed, one corresponding to bipartitions that are present and the other to bipartitions that are absent. As in a), the empirical result is marked in red. In c) we visualize the distribution of RF-distributions from the data-generating tree across simulation replicates, depicting also the empirical result for each normalized RF-distance to the right of the corresponding simulation result.

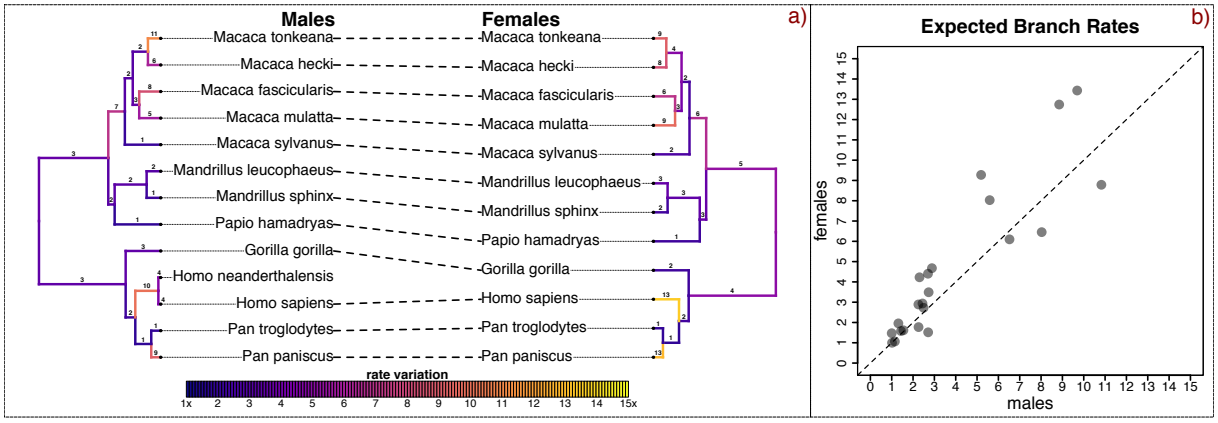

Figure 3.4: In a), cophylo plots of molecular MCC trees with male and female mean branch rates colored according to the heatmap at the bottom of the figure and labeled above each branch. Branch rates arithmetically average over the Bayesian marginal posterior, discarding inferential uncertainty. To allow rates to be compared on a common scale, male and female branch rates were divided by the slowest branch rate in their respective partition, in both cases corresponding to the branch subtending the cercopithecoids. We more directly compare these branch rates in b), where relative branch lengths are plotted for both partitions. As the female tree lacked a Neandertal tip, we matched the branch rate of the branch subtending the Neandertal-*Homo sapiens* MRCA in the male tree to the terminal *Homo sapiens* branch of the female tree, to better reflect the cranial differentiation present along the hominin lineage after its split from that of the chimpanzee.

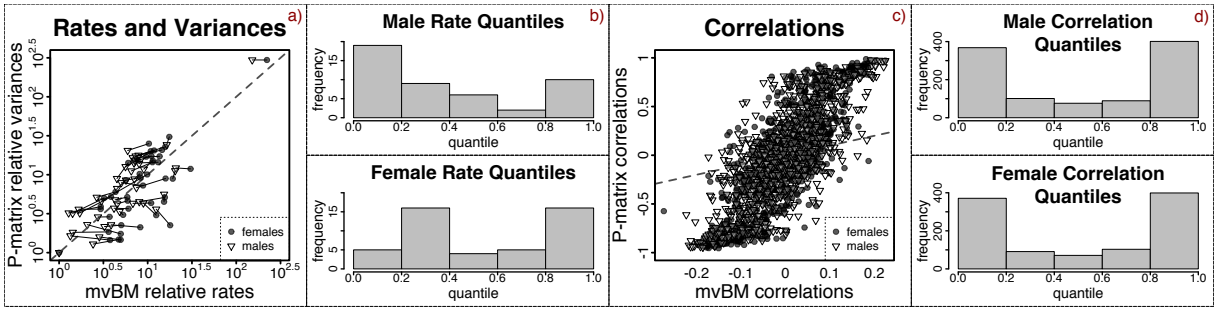

Figure 3.5: In a), a scatter plot of posterior mean trait-specific rates of evolution on the horizontal axis and pooled within-group phenotypic variances on the vertical axis. Males and female estimates for the same trait are connected by thin solid line black lines, and the 1-to-1 line is shown by a dashed line. For these to be comparable on a common scale, both were standardized by dividing all rates by the smallest rate, which in both sets corresponded to the same coordinate (the Z^{th} position of the 7^{th} landmark). Rescaling the entire marginal posterior by each sample-specific instance of this value, in b) we depict the distribution of quintiles into which the rescaled pooled within-group variance for each trait fell. In c), we compare posterior mean pairwise correlations with their corresponding pooled-within group (P) correlations, once more drawing a dashed 1-to-1 line. In d), the distribution of quintiles of each pairwise P correlation in their matched marginal posterior distribution is shown.

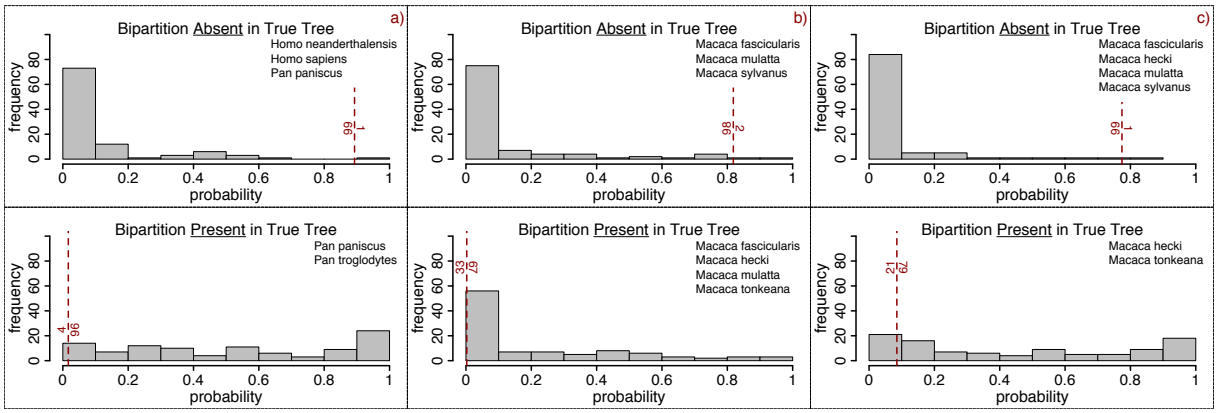

Figure 3.6: Each panel corresponds to a *fractious* bipartition in the empirical tree whose estimated probability in the mvBM morphological analysis was > 0.75 away from the corresponding bipartition's molecular probability. The morphological probability is marked by a dark red vertical line, and the histograms represent that bipartition's probability across the 100 replicates of our simulation experiment. Numbers on either side of the dashed line represent the proportion of simulated runs falling on either side. The set of taxa comprising the bipartition is given in the upper right-hand corner of each panel. Presence and absence of particular bipartitions in the tree can be paired, as all trees are strictly bifurcating and one mismatch necessarily implies another. As such, we have organized the figure into three P/A pairs, labeled *a)*, *b)*, and *c)*.

3.7 Tables

	Males	Females	Total
Gorilla gorilla	48	36	84
Homo neanderthalensis	5	0	5
Homo sapiens	118	112	230
Macaca fascicularis	30	21	51
Macaca hecki	10	7	17
Macaca mulatta	11	12	23
Macaca sylvanus	10	10	20
Macaca tonkeana	9	11	20
Mandrillus leucophaeus	22	15	37
Mandrillus sphinx	19	12	31
Pan paniscus	21	28	49
Pan troglodytes	55	74	129
Papio hamadryas	264	136	400

Table 3.1: A table detailing the composition of the landmark dataset used in our empirical analysis. Elements of the table represent numbers of individuals in each row species corresponding to each estimated column sex. Further details regarding landmarks used can be found in (Harvati et al., 2004).

Chapter 4

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

Nikolai G. Vetr, Shara E. Bailey, Timothy D. Weaver

4.1 Abstract

Human dental variation is often used for the inference of population history and phylogeny in paleontological contexts. Teeth are hard and compact, 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 are 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]

4.2 Introduction

4.2.1 Multivariate Character Evolution

Sewall Wright (1934) first proposed the threshold model of quantitative genetics — also called the quasicontinuous model (Grüneberg, 1955) in dental anthropology — to describe the expression of toe number on guinea pig hind feet. In the years since, the threshold model has been used to model discrete trait evolution phylogenetically (Felsenstein, 2005, 2012; Revell, 2014), where it is also called the latent liability model (Cybis et al., 2015) and the multivariate probit model (Zhang et al., 2019), the latter of which enjoys an equally lengthy history as Wright's naming (Bliss, 1934; Chib and Greenberg, 1998).

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 (presence / absence) trait, if the liability value of an individual is greater than some threshold value, the corresponding discrete trait is expressed; if less than, it is 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 thresholds 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, they are normally distributed across individuals under the Central Limit Theorem, and if we wish to identify the location of this normal distribution, or the locations of the thresholds, we must fix its variance to some number, by convention unity. This discretization straightforwardly generalizes to multiple traits in a multivariate framework, with population liability distributions described by multivariate normals with some vector of mean liabilities and correlation matrix (once more fixing variances to unity 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. Animated visualizations of the effect of varying population means, thresholds, and between-trait correlations on

ordinal trait frequencies in one and two dimensions can be found below (Supplemental Figures).

Through time, evolutionary processes will cause the location of a population's multivariate normal 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 multivariate normal, and so can be described according to multivariate Brownian motion (mvBM), perhaps acting over a strictly bifurcating, non-reticulate population tree. Taking as given Cheverud's conjecture (Cheverud, 1996; Sodini et al., 2018), the correlation matrix of the within-population multivariate normal 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 distributions 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; Genz and Bretz 2002), 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.

4.2.2 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 (Lewis, 2001). The Mk

model 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 (Wright and Hillis, 2014; Wright et al., 2016), such as high rates of evolution or high rate heterogeneity among characters (Wagner, 2012), at least when data is also simulated under a so-parameterized Mk model. MP itself has a few other drawbacks, such as 1) statistical inconsistency when rate inequalities exist between lineages (heterotachy), in part due to its disregard for branch lengths, as traits can only change once on any given branch, and 2) its lack of rigorous means for deciding between alternative implementations of parsimony and between most parsimonious trees (Felsenstein, 2004), 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 do not seem too appealing; e.g. Fitch parsimony (Fitch, 1971) always picks the same trees as the TS97 model (Tuffley and Steel, 1997), which, if branches have the same length for all traits, is equivalent to the Mk model (Lewis, 2001; Steel and Penny, 2000).

The Mk model generalizes the simplest of the GTR family of continuous time Markov chain (CTMC) models of molecular evolution, JC69 (Jukes and Cantor, 1969), 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; Yang 1994), 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 (1973) Pruning Algorithm. Unlike the threshold model, 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 (Scott et al., 2018b), described below. Another plausibly desirable property of the threshold model involves the frequencies of trait expression changing rapidly when they are inter-

mediate, but more slowly once at the extremes, and slower still in expectation if they have been extreme for a long 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 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 is not 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 (Pagel, 1994; Pagel and Meade, 2006), but with far worse scaling at high dimension, requiring a $n^k \times n^k$ instantaneous rate matrix for k traits with n degrees of expression, 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 $00 \leftrightarrow 11$), though constraining elements of this rate matrix to 0 helps to limit its dimensionality somewhat. Alternative approaches exist (Robinson et al., 2003; Rodrigue et al., 2005, 2006), 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; Turner et al., 1991) before, so there exists precedent in applying it to that suite of traits (Scott et al., 2018b).

4.2.3 Discrete Dental Traits

ASUDAS traits represent a common material for the inference of both human population history (Hubbard et al., 2015; Rathmann et al., 2017; Reyes-Centeno et al., 2017) and

hominin phylogeny (Irish et al., 2013, 2018), though the latter may benefit from typologies better able to capture nonmetric dental variation across species (Bailey, 2002; Carter et al., 2014). Due to their high mineral content and overall hardness, 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 (Scott et al., 2018a). The majority of dental traits are scored on an ordinal scale, but are almost always dichotomized into presence / absence for use in analysis (e.g. Irish et al. 2013), 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 (Scott, 1973), and they are frequently treated as such (e.g. Rathmann and Reyes-Centeno 2020). Complex segregation analysis accepts a quasicontinuous, polygenic model for many of the discrete dental traits hitherto considered (Nichol, 1989), and to date not a single dental trait has been found to have simpler genetic architecture (Scott et al., 2018b).

ASUDAS traits appear to broadly track neutral patterns of human genetic variation (Hanihara, 2008; Rathmann and Reyes-Centeno, 2020), 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 attraction and social signaling, and sundry other benefits (Scott et al., 2018b). To the extent that selection is fluctuating or universally directional, Brownian motion may provide an adequate fit to these data, but exploration of other stochastic processes better able to capture adaptive evolutionary processes may yield conflicting results. Finally, the evolution of the mammalian dentition is not characterized by independence between characters (Brocklehurst and Benevento, 2020), 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 forms the empirical focus of this work.

4.3 Materials and Methods

4.3.1 Empirical Data

The data used here come from 722 individuals from 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*, with 137 discrete dental traits in total scored by Shara Bailey. 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 (Scott et al., 2018b), 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.

4.3.2 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 4.1a-b). Subsequently, we plotted the number of traits present in some number of individuals in at least some number of populations (Figure 4.1c). 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 eight 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 Table 4.1.

4.3.3 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 of these describes the evolution of those means on a phylogeny with some branch lengths and rate matrix (Appendix A), integrated over their uncertainty. The second, meanwhile, describes the distribution of individual level character vectors under those same means and correlation matrix. These yield each population's individual level liability distribution, and coupled with a set of threshold locations that parameterize an indicator function, transform each individual's liability vector into a vector of discrete characters according to which hypervolume contains it. 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 and the mvBM rate matrix, R. The former has diagonal entries corresponding to the height of each tip above the root and off-diagonal entries corresponding to the sum of shared branch lengths from the root between each pair of tips. The

former, meanwhile, is often further decomposed into a matrix product SCS, where S is a diagonal matrix of standard deviations (the square roots of each trait's evolutionary rate, σ_i^2) 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 Cheverud's conjecture equal to the aforementioned C, with bounds of integration 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 of the populations under consideration. 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 integrated over or sampled through data augmentation, its own plausibility defined 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, with the probability of each tip-wise vector of discrete traits equal to 1 inside it 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, also 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 (Genz, 1992) and implemented in the function `pmvnorm` in the package *mvtnorm* (Genz et al., 2020) in *R* (Team, 2013). 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 for d traits, 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 (d). This appears to produce a value roughly proportional to that of the true integral (Figure 4.2a) while imposing a computational burden many orders of magnitude smaller at high dimension. Though it appears to hold less well at extreme correlations (Figure 4.2b), 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 by modifying a reimplementation of *mvtnorm* code in the R package *pbivnorm* (Kenkel, 2015). To avoid underflow, all calculations were performed on the log-scale.

4.3.4 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 struc-

ture 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, and a flowchart visualizing the order of analysis can be seen in Figure 4.3.

4.3.5 Iterative Optimization

In preliminary simulative 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 touched, 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 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 dis-

crete states are at their maximal or minimal states and invariant within a tip, a problem long-recognized in the context of probit models, Fisher 1935), several forms of regularization were used, i.e. penalties to the likelihood function over which we were performing optimization, 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 distribution stretched to the (-1,1) range with shape parameters equal to 10. Attempting to also optimize the magnitude of these shape parameters resulted in 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($\eta = 1$) 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 initial-

ization, 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 be 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.

4.3.6 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 impute these values on an individual or population-wide scale, though combinatorial difficulties quickly arise in the latter case, 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 use rejection sampling from the unconditional multinomial distribution 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, rejection sampling is 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 (Plummer et al., 2006) 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 four 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.

4.3.7 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 (Higham, 2002), as implemented in the `nearPD(corr = T)` function in the *Matrix* package (Bates and Maechler, 2019) in R, also used in similar contexts elsewhere (Blows et al., 2015). 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.

4.3.8 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($\alpha, \alpha,$ ...) 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 (Hastings, 1970), making NNI and SPR proposals to tree topology, truncated sliding window 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* (Schliep, 2011) 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 each chain was discarded as burnin. To diagnose MCMC performance, we assessed the effective sample size and Gelman-Rubin Convergence Diagnostic (Gelman and Rubin, 1992) of several explicit and implicit model parameters, both implemented in the R-package *CODA*

(Plummer et al., 2006), 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 (Robinson and Foulds, 1981a) from a reference tree. Additionally, we required that the squared 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 Pruning Algorithm (Felsenstein, 1973), 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 i.i.d. standard normal variables, the densities of which could together be far more easily evaluated to produce the same likelihood values as more computationally cumbersome approaches commonly implemented in standard phylogenetic software.

4.3.9 Simulation Experiments

Having inferred a human 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 (see Appendix C). 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 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) corresponding to the (*Neandertal, Oceanian, European, West Asian, South Asian, Northeast Asian, Sub-Saharan African, American*) tips, respectively. With our estimated threshold matrix, we convert these individual liability vectors into ordinal characters, and simulate state-dependent missingness with the inverse-logit function, assigning state 1 a probability of missingness 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) for states 0 through 6. 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.

To explore the effects of low within-population sampling and error introduced by TSAR-MBOP, we perform two follow-up sets of simulation experiments. In the first, we simulate ordinal character data with no missingness and dramatically inflated within-population sample-sizes, giving each population twice the number of individuals as our most populous (*American*) empirical population. Effectively, this increases our total sample size by approximately 15-fold (taking us from 684 individuals with ≈33% data presence to 3,168 individuals with 100% data presence). As before, we perform 500 replicate analyses of these newly simulated data with our two-step procedure. To explore the effects of estimation error introduced and conditioned upon during our first optimization step, we then perform just the second step of inference — fitting a phylogenetic multivariate Brownian diffusion model with adaptively regularized trait-specific rates — using the true population means and correlation components of our rate matrix. Again, this is done across the 500 replicates of our original simulation study, reusing the same simulated mean liabilities and estimated correlations, and using a phylogenetic model identical to that used to analyze our empirical data. All analyses of simulated data were required to adhere to those same convergence and other diagnostic criteria as were used in our empirical analysis.

4.4 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 4.4), 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 unrooted 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 shown in Figure 4.5, with nodal bipartition probabilities labeled. 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 4.6a). 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 4.6b), 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 4.7a-f) of estimated vs. true values for all three sets of model parameters across both sets of sample-size conditions, as well as examined the distribution of R^2 values for these over our 500 replicates (Figure 4.7g-i). 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 4.8). 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 bipartition probabilities (Figure 4.9a) across all three sets of simulating conditions, as well as histograms of quantiles for true, data-generating rates in the marginal posterior distributions of inferred rates (Figure 4.9b-d), along with kernel density plots of the distribution across replicates of R^2 values for posterior mean rates against true, data-generating rates (Figure 4.9e).

4.5 Discussion

Empirical Results. Trees inferred from our empirical analysis (Figure 4.5) appear to be broadly consistent with both prior work (Scott et al., 2018b) and molecular expectation (Mallick et al., 2016). Midpoint rooting resulted in trees most often leading to the Neandertal tip at the first bifurcation, with over four-fold as many as to any other single terminal node. Across the entire posterior output, however, these comprised only 7% of trees. This may partially be driven by Neandertal extinction 40ka (Higham et al., 2014) and the even older ages of several of our scored Neandertal specimens robbing them of opportunity for dental evolution available to the other tips (e.g. the modal specimen originates from Krapina and dates to around 130ka; Rink et al., 1995). With a population split time of 600ka (Nielsen et al., 2017; Schlebusch et al., 2017), a Neandertal tip age of 100ka, and a *Homo sapiens* split time of 300ka, the Neandertal branch should be approximately 12.5% longer, assuming a homogenous within-lineage evolutionary rate on the branches leading to Neandertals and the node ancestral to *Homo sapiens* populations (though not a tree-wide strict clock; Gómez-Robles, 2019). Lengthening the Neandertal branch by this amount raises the proportion of midpoint-rooted trees to 11%, eleven-fold as many as the next most common single-tip to split off first as a result of midpoint rooting. As such, we rooted trees along the Neandertal branch $\frac{5}{8}^{ths}$ of the way towards its connecting internal node to reflect these estimated population split times.

Curiously, the next tip to split off from the *Homo sapiens* stem appears to be that corresponding to Oceanian populations (native Australians and Papua New Guineans), rather than Sub-Saharan African populations (SSAF), despite the latter representing the earliest divergent human groups in molecular studies. Instead, the SSAF appear to cluster with the European tip with intermediate probability (0.43), potentially due to paraphyly in the former tip caused by our lumping of multiple SSAF populations into one. However, our analysis finds Sub-Saharan Africa to be the next to split off in the second most probable tree, with nodal probability equal to an initial Oceanian split (0.34). Additionally, the branch length leading to the SSAF-European group is very short, almost polytomous, indicating little dental evolution along their ancestral lineage. In contrast,

American and Asian tips appear to cluster together with intermediate-high probabilities, consistent with molecular expectation. The Oceanian tip, however, is absent from this group, despite its molecular affinities lying there. 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, likely a signature of the later peopling of the Americas by the latter group according to a northeast Asian dispersal across the Bering land bridge (Mulligan and Szathmáry, 2017).

Estimated trait-specific rates appear to be fairly uniform in canines, premolars, and molars, but especially elevated in incisors (Figure 4.6b), potentially due to the latter's greater role in social signaling (Demir et al., 2017), speech production (Howell, 1987), and grasping / clamping (Trinkaus, 1987), or because of pleiotropy affecting incisal form as a result of selection on unrelated traits (Hlusko et al., 2018). Meanwhile, 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, though correlations between traits within teeth appear to be more variable overall, with the strongest correlations of any in the matrix found there (Figure 4.6a).

Simulation Experiments. However, given the results of our empirically-parameterized simulation study (Figure 4.7), correlation parameters appear to be the least reliably estimated of all within-population parameters during our first optimization step, especially in the empirically parameterized simulating condition (Figure 4.7h). 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 missingness 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 two, with over 10% of simulations having five 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. In the analyses performed with 15-fold sampling at the individual level, correlations between-traits within-populations were much more reliably estimated (Figure 4.7e,h)

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 that allowance.

Despite these caveats, it would appear that tip mean liabilities and threshold locations may still be reliably estimated with data such as these (Figure 4.7a,c,g,i), likely because there is no need in their estimation for paired variation in the dataset. Instead, the only tip mean liabilities our optimization procedure truly struggled 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 that invariance not result in pathological overfitting. As we regularized under a 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 unchecked tendency to supply the most ostensibly plausible, if ridiculous 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 4.8), with pooled probabilities across traits recovering the appropriate monotonic decreasing order, despite that assumption never having been explicitly specified in our implementation of 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 eight populations than in just one. Small probabilities at high degrees of expression were not as well estimated, contrary to the apparent success evident in Figure 4.8, 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 single digit numbers of traits could occupy the later degrees of expression. With less data available, our flat beta could not be so reliably updated, and so despite its uninformativeness, it broadly 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 4.7a).

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* catastrophically. Despite poor estimation of correlations of the mvBM rate matrix, increasing bipartition probabilities (Figure 4.9a), while not especially well calibrated, did nevertheless associate with increasing frequencies of true bipartitions. However, estimated bipartition probabilities conditional on tip means and between-trait correlations are nevertheless quite untrustworthy for both TSAR-MBOP simulating conditions, with a marked bias upwards. In other words, high probability bipartitions emitted during inference do not represent high-frequency bipartitions, but rather medium-frequency bipartitions, and improved estimates of population means and between-trait correlations (Figure 4.7d-i) does not appear to be of terribly much help here. One possible reason for this may involve convergence in our estimation of population mean liabilities for invariant discrete traits occupying extreme ordinal states, for which all extreme liabilities, no matter how they may differ on the underlying latent scale, are estimated to have similar values (Figure 4.7a,d). This, in turn, may be unduly interpreted as phylogenetic evidence for shared ancestry — further simulation experiments conditioning on true means *or* true correlations but not both may help to disentangle the source of this error. As described earlier, the structure of ASUDAS data, which conditions on within-population polymorphism, may not present as great a difficulty to inference as that simulated with no such constraint here. But we nevertheless urge caution when interpreting estimated bipartition probabilities as representative of those probabilities truly implied by the multivariate ordinal threshold model.

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. As a result, phylogenetic inference using the true means and correlations did not correspond to a calibration curve falling along the one-to-one line (Figure 4.9a), instead deviating slightly from it. Similarly, there appears to be a slight inferential bias towards depressed evolutionary rates

(Figure 4.9b), likely due to our trait-specific rate regularization prior concentrating probability at greater degrees of similarity than observed in the posterior rate distribution. In our empirically parameterized and high-sample simulation experiments (Figure 4.9c-d), regularization appears to have an even stronger effect, with true rates often falling in the tails of their respective marginal posterior distributions. This is undoubtedly a result of the invariance problem mentioned earlier: mean liabilities are free to vary along the (-∞, ∞) scale, but estimates of those mean liabilities are fundamentally constrained by finite sample sizes and regularized towards estimates obtained for those same traits at other tips. As a result, inferred rate variation is reduced, which results in a corresponding increase in the inferred value of our regularizing hyperparameter α, pulling slower-evolving trait rates upwards in light of the more informative Dirichlet prior. Unfortunately, these rates are themselves quite poorly estimated under the two-step algorithm (Figure 4.9e), so strong caution should also be urged when interpreting rate variation results from our empirical analysis. 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, rather than 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 (Appendix B), 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.

4.6 Figures

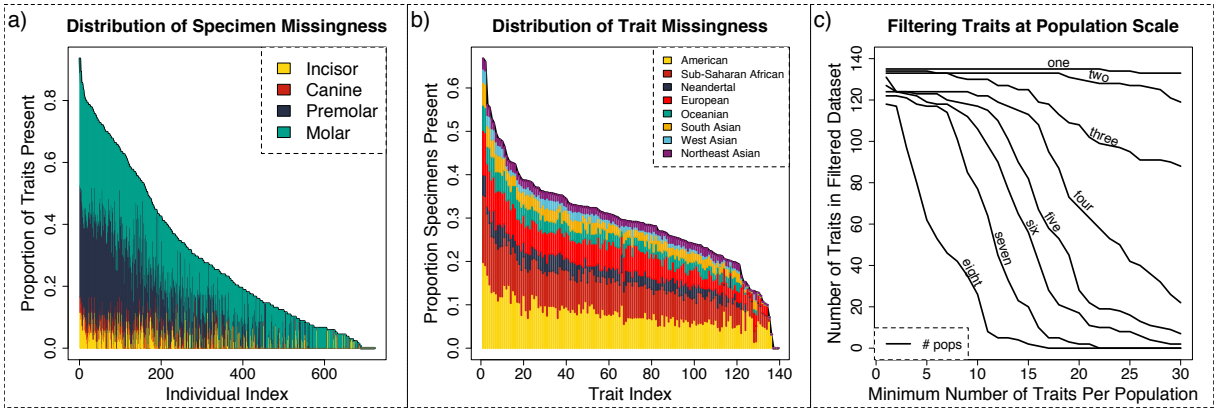

Figure 4.1: A visualization of missingness in the unfiltered dataset. In a), the proportion of traits present in the sorted, decreasing set of individuals represented in the sample. Colors represent different tooth types, stacked according to their mesio-distal progression within the dentition. In b), 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.

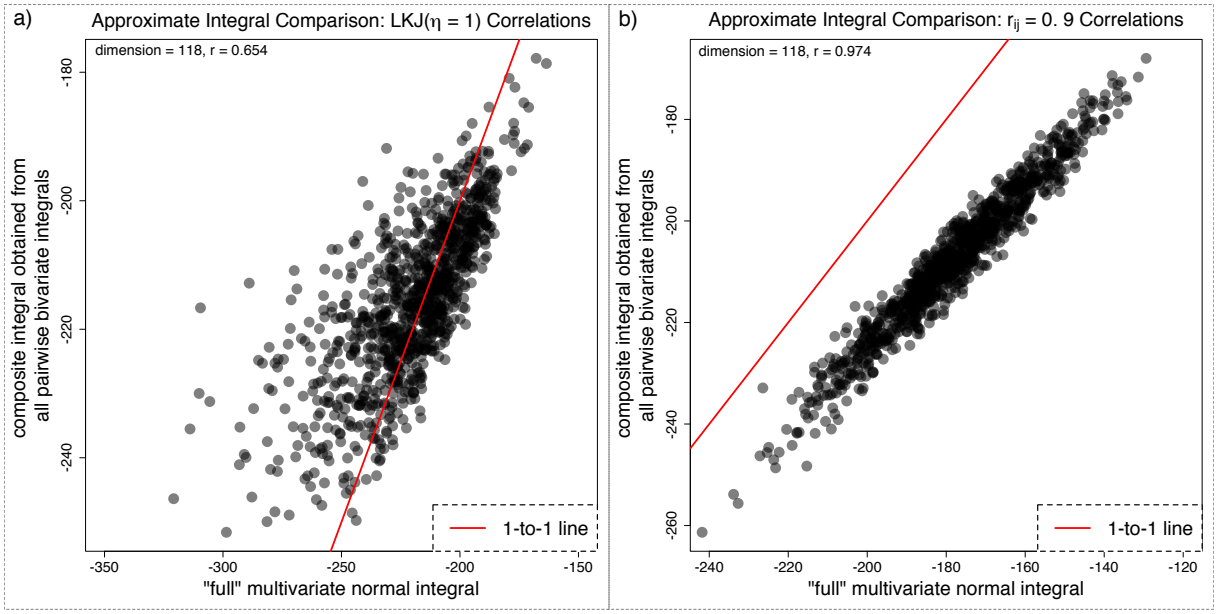

Figure 4.2: Visualizing relationships between the transformed bivariate integral of a multivariate normal and its full evaluation. In a), the integral of a multivariate normal with mean at the origin and 118 x 118 correlation matrix sampled from an LKJ(1) was evaluated with both methods between pairs 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.

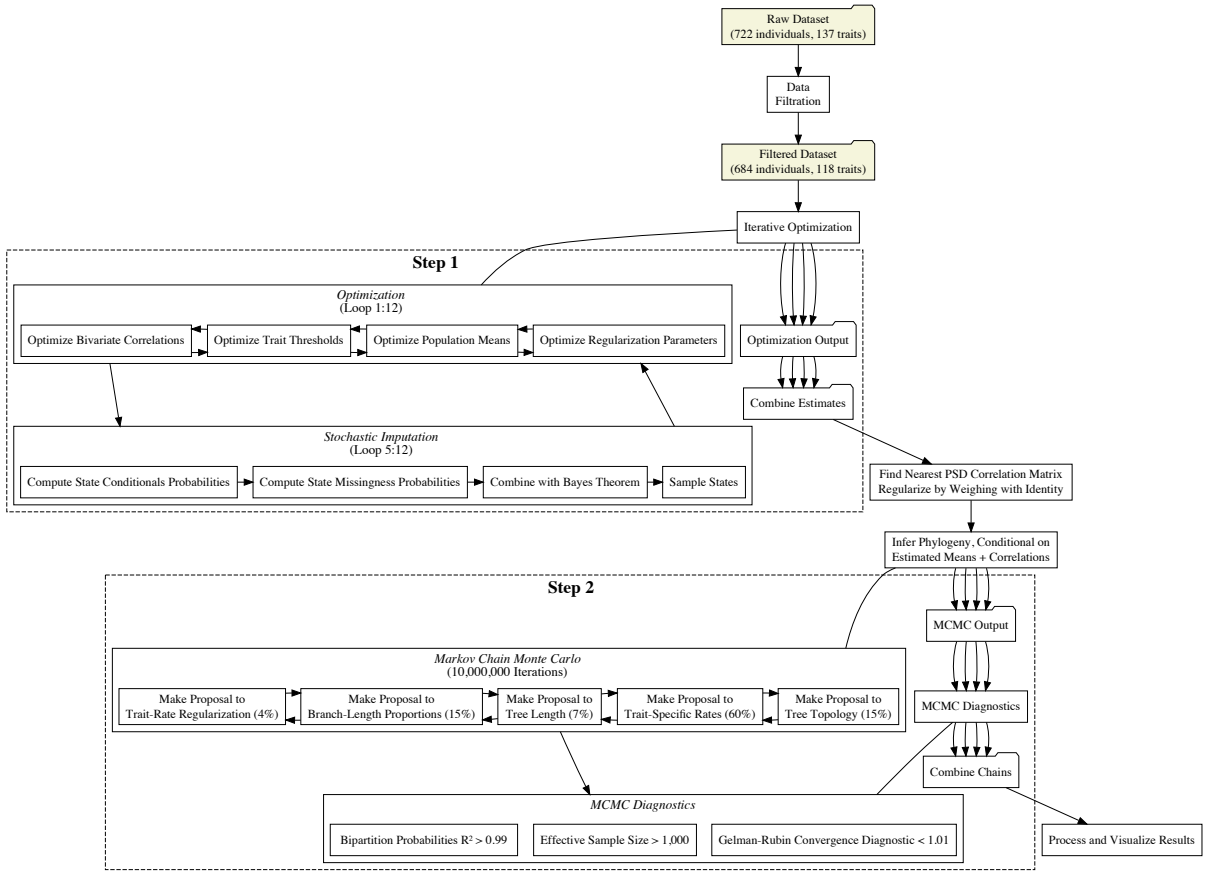

Figure 4.3: A flowchart depicting the order of analysis and other data and output processing steps performed during the TSAR-MBOP procedure.

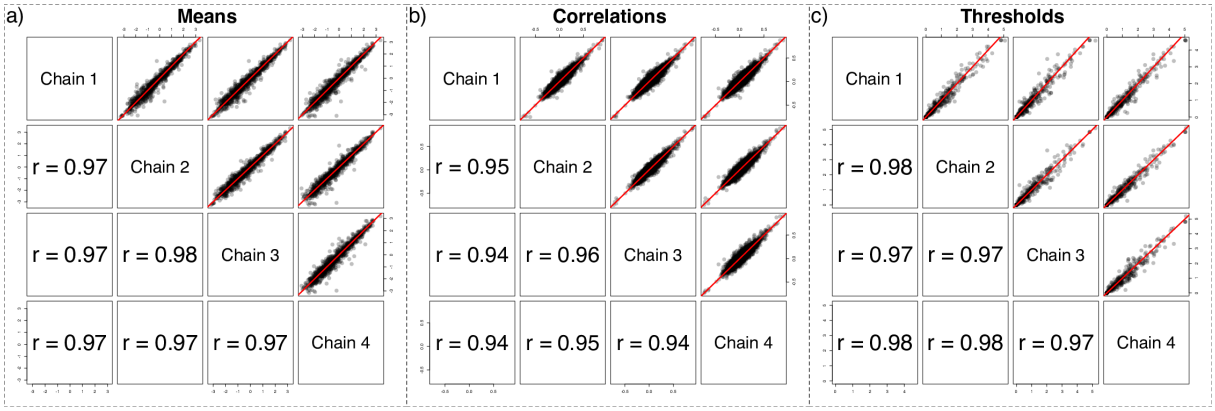

Figure 4.4: Output from the iterative optimization step of our two-step algorithm across four independent runs. In a), means are plotted in the upper right panels of the figure, with correlations between runs in the lower left panels. In b), within-group, between-liability correlation parameters are plotted. In c), threshold locations.

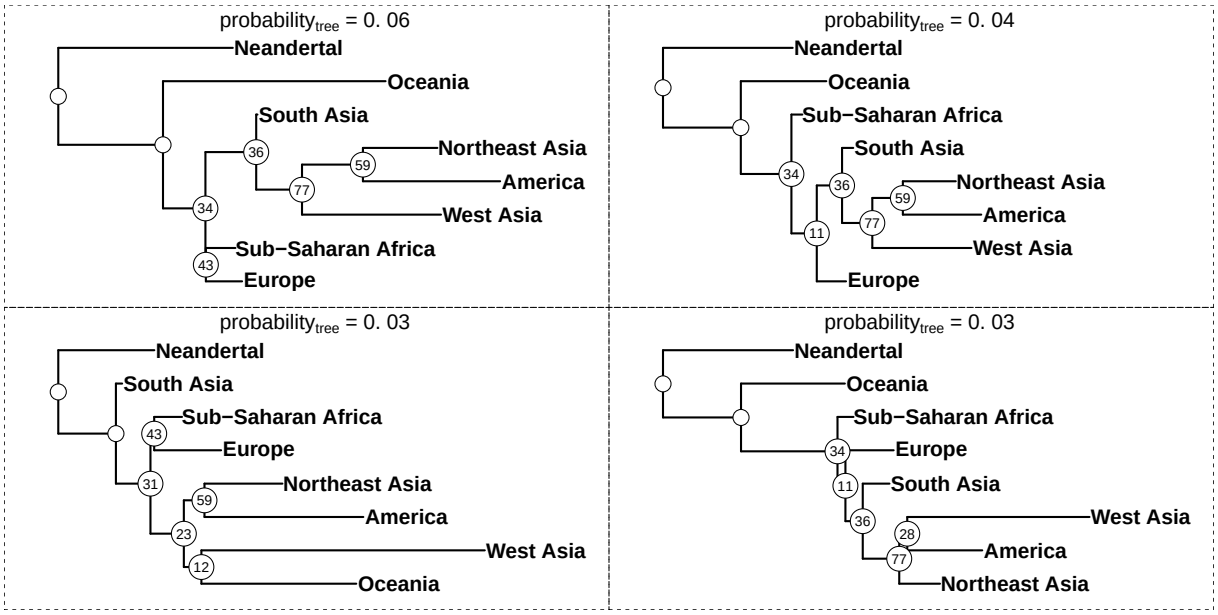

Figure 4.5: The four most probable trees from the posterior distribution of our Bayesian phylogenetic analysis, with nodal posterior probabilities plotted. Branch lengths are posterior mean estimates conditional on each tree topology and are proportional to the extent of morphological evolution on each branch. Trees were rooted along the Neandertal branch approximately $\frac{5}{8}^{ths}$ of the length to the internal node.

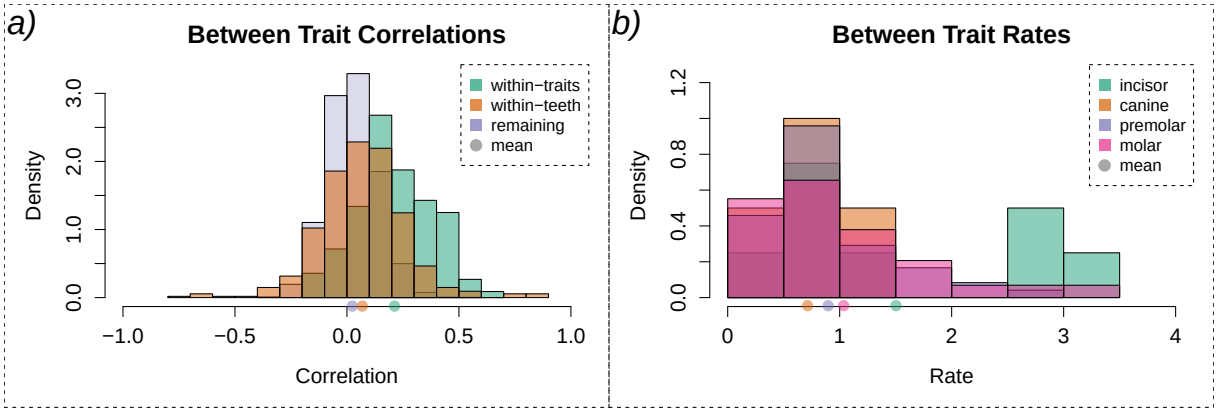

Figure 4.6: In a), histograms of correlations between traits within individual teeth, within traits across teeth, and for the remaining elements of the correlation matrix are plotted. Correlations are those from the evolutionary rate matrix, and so are interpretable as correlations of the evolutionary 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. Means in both panels are marked with color-coded filled circles.

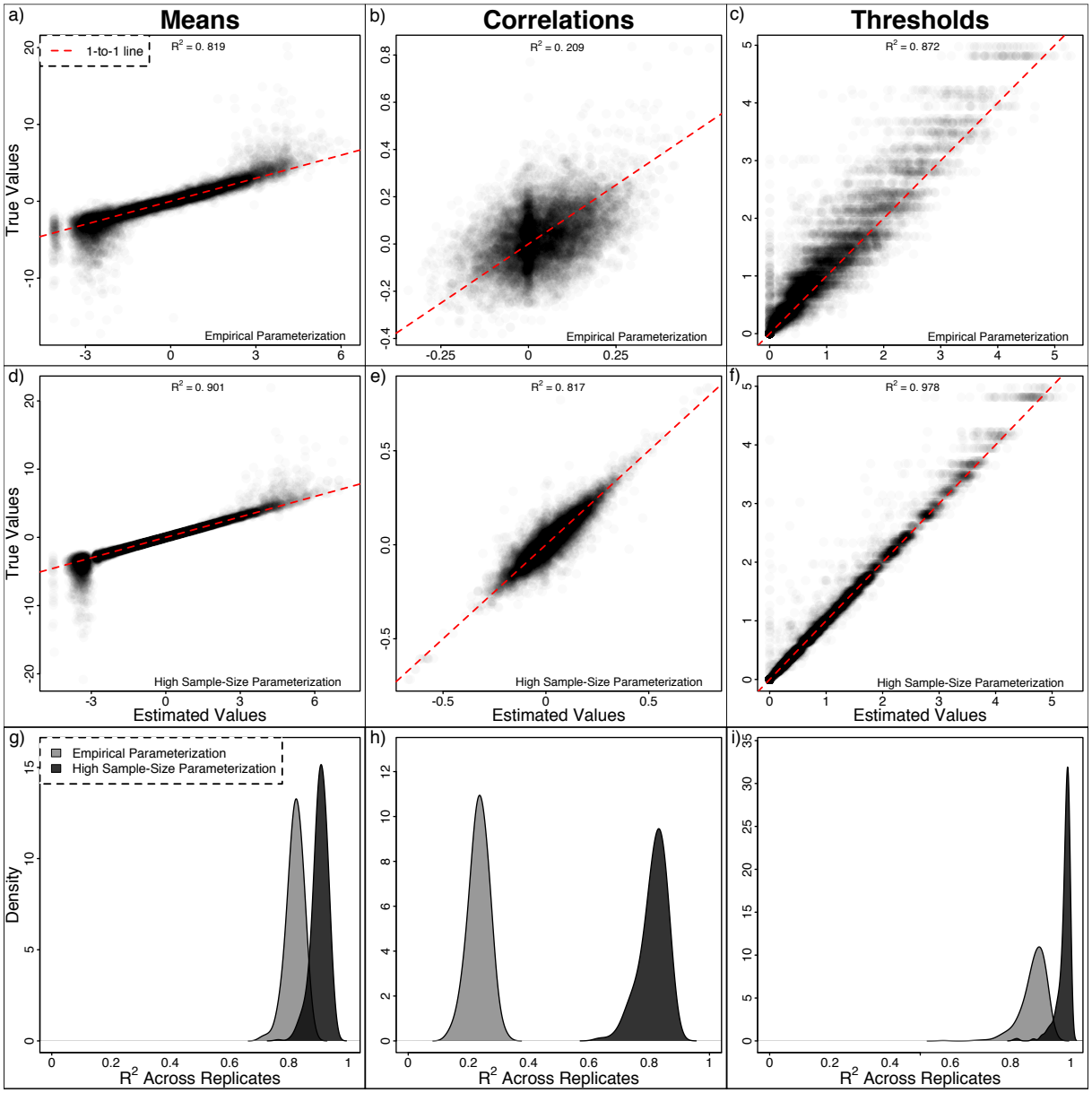

Figure 4.7: In a-c), estimates of population means, between-trait / within-group correlations, and threshold locations, respectively, are plotted together for 500 simulations under the lower-sampled, empirically parameterized condition. The one-to-one line is shown, and an overall R^2 is labeled. In d-f), the same are plotted for the "high-sample size" condition. Finally, in g-i), kernel density estimates are plotted for both conditions for replicate-specific R^2 values.

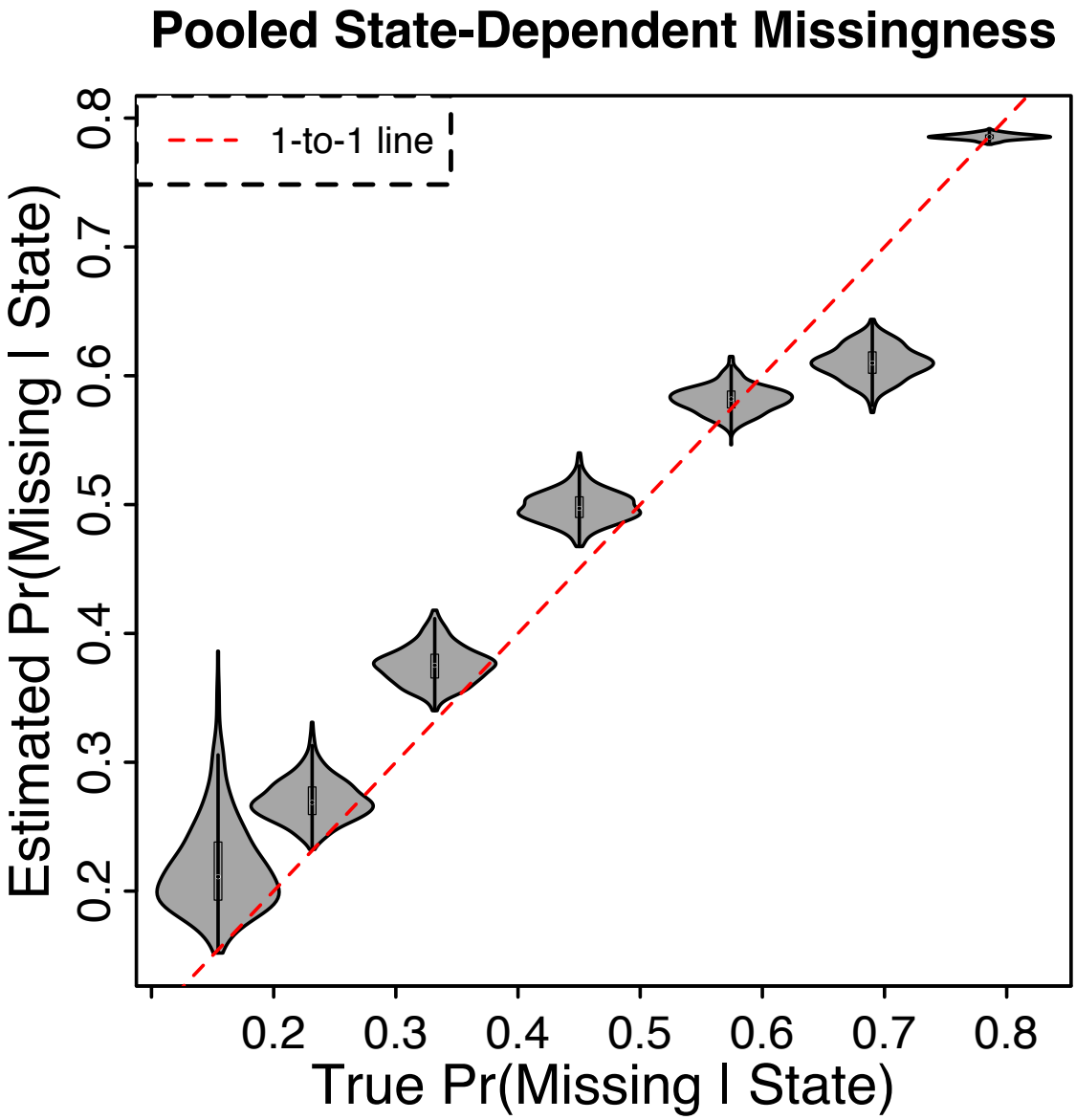

Figure 4.8: Estimates of state-dependent missingness during the last round of iterative optimization, averages across four independent chains. Violin plots show the distribution of these estimates across 500 replicate analyses of simulated data under the "empirically parameterized" condition. These estimates are pooled across traits for visual clarity, but analyses used trait-specific estimates as between-trait commensurability was questionable. A one-to-one line is plotted for ease of interpretation.

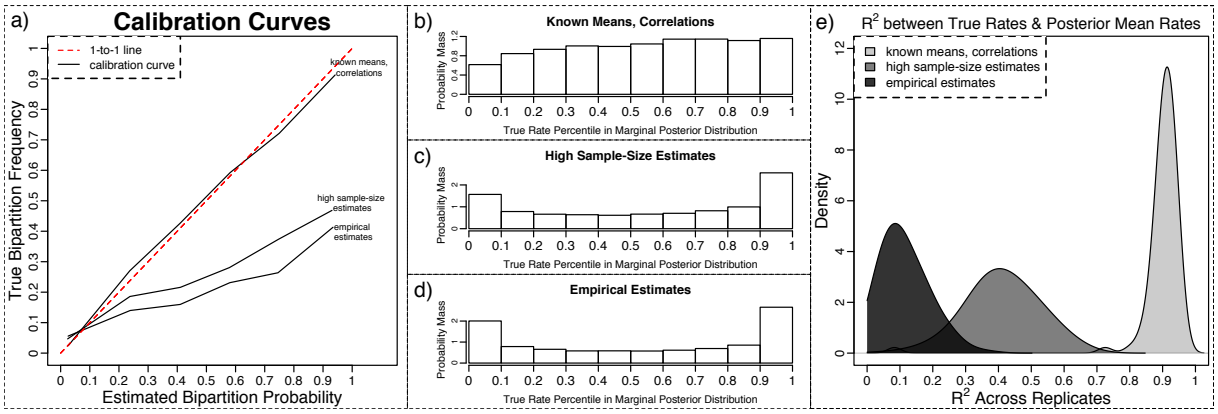

Figure 4.9: In a), calibration curves across the three sets of simulation conditions: where discrete data were simulated under conditions resembling our empirical dataset (labeled *empirical estimates*), where it was simulated under high-sample size conditions described in text (labeled *high sample-size estimates*, and where it was not simulated at all, and true means and correlations were used. These curves describe the relationship between an estimated bipartition probability and the true frequency with which it was found in the data-generating tree, with pooling done within sextiles. In b-d), percentile plots of true, data-generating trait-specific rates in the marginal posterior distribution for each of those rates are shown for each of the three conditions, with plot data pooled across replicates. In e), kernel density estimates of R^2 values between true-rates and the posterior means of these trait-specific rates are shown across replicates.

4.7 Tables

	NEAN	OCEAN	EUR	WAS	SAS	NEAS	SSAF	AMER
UI1.LC	25	9	37	5	16	8	37	54
UI1.SH	26	9	36	5	12	9	36	39
UI1.DSH	25	8	37	5	16	8	36	44
UI1.TD	25	11	32	4	8	7	32	35
UI2.SH	26	13	24	3	15	11	39	49
UI2.DSH	22	13	26	3	19	9	40	42
UI2.IG	22	16	17	3	15	8	36	36
UI2.TD	19	15	24	3	11	10	38	41
C.SH	22	12	26	6	13	10	43	53
C.DSH	22	12	29	6	19	10	47	60
C.TD	23	11	26	6	15	8	37	47
C.MR	19	11	21	5	12	9	33	37
UPM3.BMR	21	16	44	5	15	9	62	58
UPM3.LMR	17	15	43	5	16	9	61	57
UPM3.BMRF	19	16	44	5	13	9	62	56
UPM3.LMRF	17	15	43	5	14	9	61	57
UPM3.TM	24	19	45	6	21	12	64	88
UPM3.DAC	20	16	40	5	20	11	65	64
UPM3.MAC	20	16	43	6	20	12	62	69
UPM3.XC	17	18	45	6	21	12	67	83
UPM4.BMR	23	20	40	3	14	13	54	55
UPM4.LMR	19	16	42	3	14	12	52	51
UPM4.BMRF	20	19	42	3	13	13	52	53
UPM4.LMRF	19	17	43	3	12	11	51	49
UPM4.TM	22	23	44	4	20	14	58	90
UPM4.DAC	20	19	39	3	16	11	48	63
UPM4.MAC	20	21	35	3	14	12	52	61

UPM4.XC	17	21	44	4	20	14	59	88
UM1.ME	34	42	74	22	38	17	112	143
UM1.HY_RED	35	42	71	22	39	15	115	138
UM1.C5	22	34	63	20	30	14	84	91
UM1.CC	24	39	70	19	28	16	93	93
UM1.EE	21	39	66	19	33	15	102	106
UM1.BG	12	29	71	17	35	17	102	122
UM1.MPT	12	12	41	9	18	6	60	37
UM1.MAT	12	14	40	8	17	5	60	37
UM1.PROT	11	14	39	8	18	6	60	38
UM2.ME	22	36	53	2	21	16	80	105
UM2.HY	24	26	41	3	12	12	72	64
UM2.C5	21	37	54	5	16	14	70	89
UM2.CC	21	35	53	5	19	16	72	99
UM2.EE	21	30	51	5	18	14	73	96
UM2.BG	12	24	56	4	21	16	74	105
UM2.MPT	10	25	44	3	13	11	55	65
UM2.MAT	10	28	46	3	13	10	59	66
UM2.PROT	9	27	44	3	12	10	57	71
UM3.ME	13	13	17	4	16	12	43	40
UM3.HY	17	15	17	4	18	13	37	45
UM3.C5	16	19	19	4	17	13	42	45
UM3.CC	12	19	17	4	15	12	43	44
UM3.PA	17	20	17	4	17	11	44	48
UM3.EE	10	19	14	4	13	9	40	34
UM3.BG	11	9	15	4	16	13	41	47
UM3.MPT	7	16	18	4	14	10	31	46
UM3.MAT	7	16	18	4	15	10	29	40
UM3.PROT	9	15	18	4	16	10	30	39

UM3.PEG	13	13	22	0	17	13	52	41
LP3.PLC	31	16	37	16	20	11	59	71
LP3.AF	23	17	36	15	15	11	56	65
LP3.MP	29	18	36	16	20	11	58	70
LP3.MI	28	18	37	16	20	11	58	68
LP3.MH	29	17	37	16	20	11	58	70
LP3.XC	28	17	36	16	20	11	59	70
LP3.LF	29	18	37	16	20	11	57	71
LP3.DAR	18	16	31	14	14	11	51	50
LP3.MAR	14	16	35	13	17	11	54	50
LP3.MLG	25	16	36	16	19	11	59	72
LP3.MeLG	5	16	37	16	19	8	58	74
LP3.DLG	8	16	37	16	19	8	58	74
LP3.ASM	17	15	32	7	19	11	56	78
LP3.DLC	15	16	37	12	20	11	56	66
LP4.PLC	29	16	36	14	20	10	53	60
LP4.AF	22	17	35	15	19	9	52	55
LP4.MP	29	16	36	15	20	10	54	60
LP4.MI	27	17	36	14	20	10	53	59
LP4.MH	27	17	36	14	19	9	53	58
LP4.XC	29	18	36	14	20	10	54	56
LP4.LF	28	18	36	14	20	10	50	59
LP4.DAR	14	13	33	9	14	8	48	36
LP4.MAR	15	13	32	9	17	9	51	36
LP4.MLG	25	10	35	15	20	8	32	58
LP4.ASM	30	10	30	9	0	8	54	72
LM1.4CUS	45	33	56	17	29	11	89	95
LM1.DW	25	14	37	10	14	7	60	45
LM1.DTC	31	26	47	12	26	9	77	50

LM1.MTC	32	27	46	12	24	10	75	55
LM1.PR	38	31	55	17	27	10	83	86
LM1.C5	41	32	56	17	27	11	86	79
LM1.C6	21	22	49	16	21	9	67	54
LM1.C7	32	30	56	17	28	11	84	86
LM1.EE	26	30	48	11	26	11	82	75
LM1.AF	33	18	40	13	18	9	74	50
LM2.DW	19	25	37	9	17	10	59	51
LM2.DTC	26	32	43	11	21	13	70	63
LM2.MTC	25	33	43	11	21	13	67	64
LM2.PR	32	31	42	12	17	13	64	76
LM2.C5	27	30	40	11	17	11	65	60
LM2.C6	18	23	40	11	17	10	54	40
LM2.C7	25	33	43	12	21	13	70	77
LM2.EE	18	27	38	8	17	13	63	67
LM2.AF	22	31	40	10	20	11	65	68
LM3.4CUS	0	8	19	3	14	12	26	11
LM3.DW	14	18	17	6	12	12	46	29
LM3.DTC	18	21	19	7	16	12	53	38
LM3.MTC	16	21	19	7	16	12	52	44
LM3.PR	21	23	16	7	14	11	44	50
LM3.C5	14	21	19	5	13	11	52	48
LM3.C6	9	21	19	5	14	11	50	42
LM3.C7	15	22	18	7	16	12	48	56
LM3.EE	11	13	14	3	12	9	46	35
LM3.AF	14	19	15	7	14	12	46	40
UPM3.BMxP_MR	18	14	42	4	13	8	60	51
UPM3.BMxP_DR	18	14	42	4	13	8	60	51
UPM3.LMxP_MR	17	13	44	5	13	8	60	51

UPM3.LMxP_DR	17	13	44	5	13	8	60	51
UPM4.LMxP_MR	18	17	42	3	14	12	49	48
UPM4.LMxP_DR	18	17	42	3	14	12	50	48
UPM4.BMxP_MR	17	19	40	3	14	11	53	51

Table 4.1: A table detailing the composition of the discrete dental data used in our empirical analysis. Elements of the table represent numbers of individuals in each column population with a definitive observation in each row trait. Population codes are as follow: NEAN (*Neandertal*), OCEAN (*Oceanian*), EUR (*European*), WAS (*West Asian*), SAS (*South Asian*), NEAS (*Northeast Asian*), SSAF (*Sub-Saharan African*), AMER (*American*). Details regarding trait codes key can be found in (Bailey, 2002).

4.8 Supplemental Figures

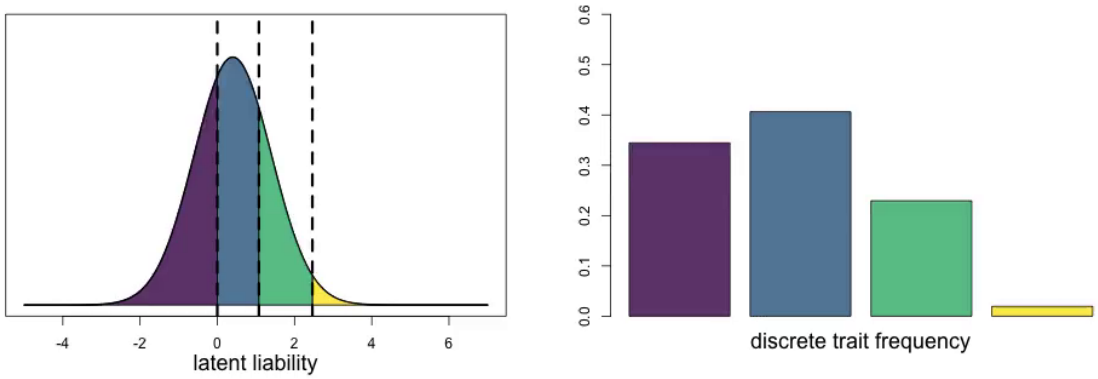

Figure S4.1: Visualizing the effect of smooth variation in a univariate mean liability on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

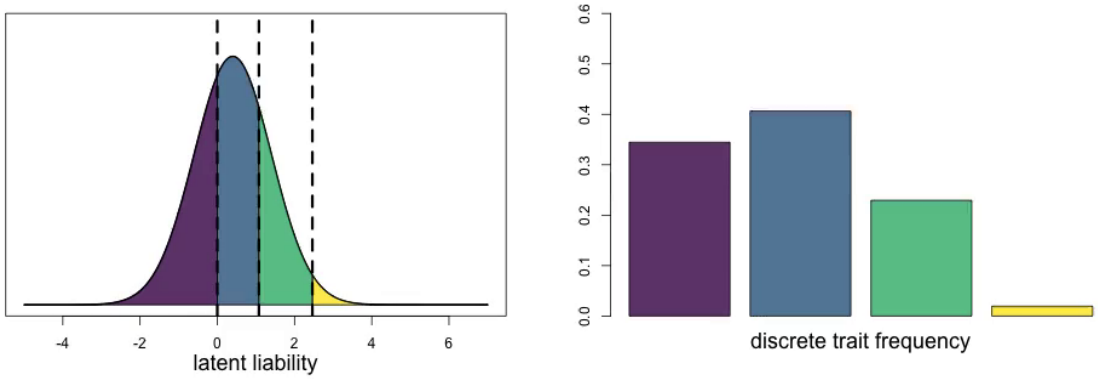

Figure S4.2: Visualizing the effect of smooth variation in the location of a threshold on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

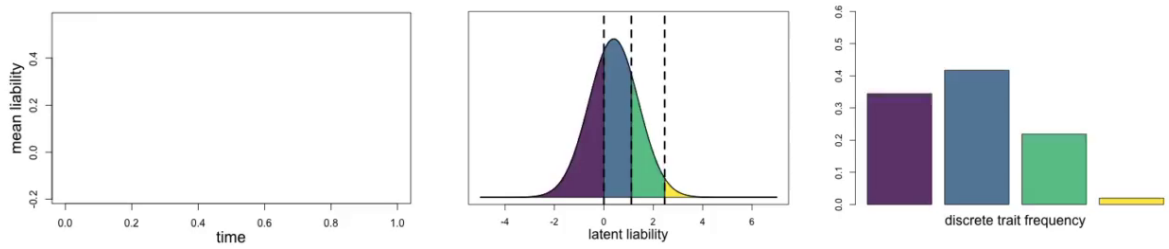

Figure S4.3: Visualizing the effect of a univariate Brownian motion process acting on a univariate mean liability on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

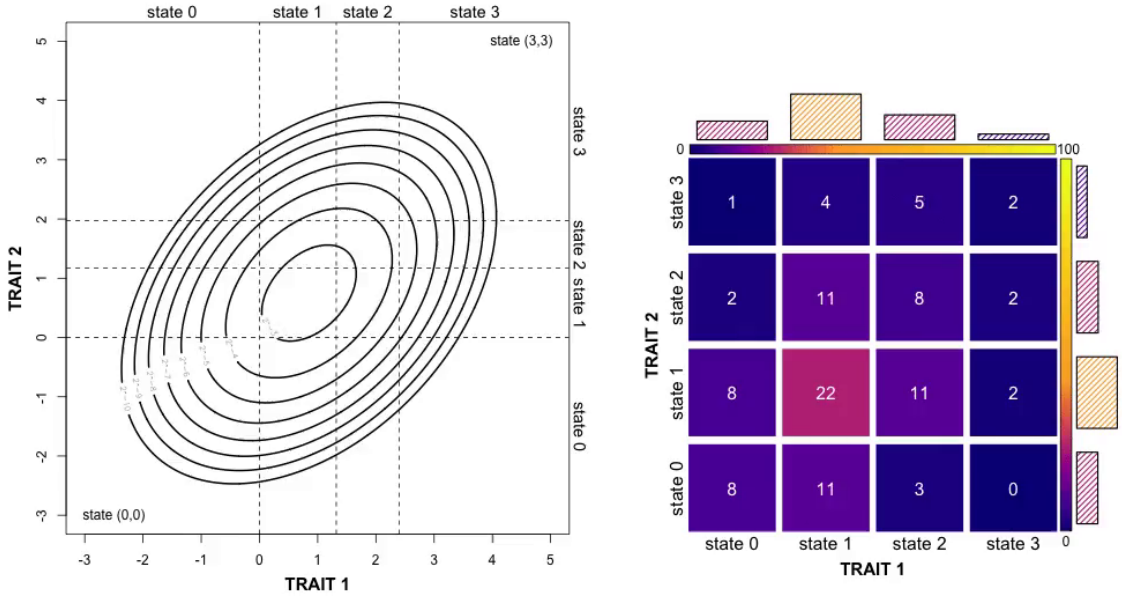

Figure S4.4: Visualizing the effect of smooth variation in a bivariate mean liability on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

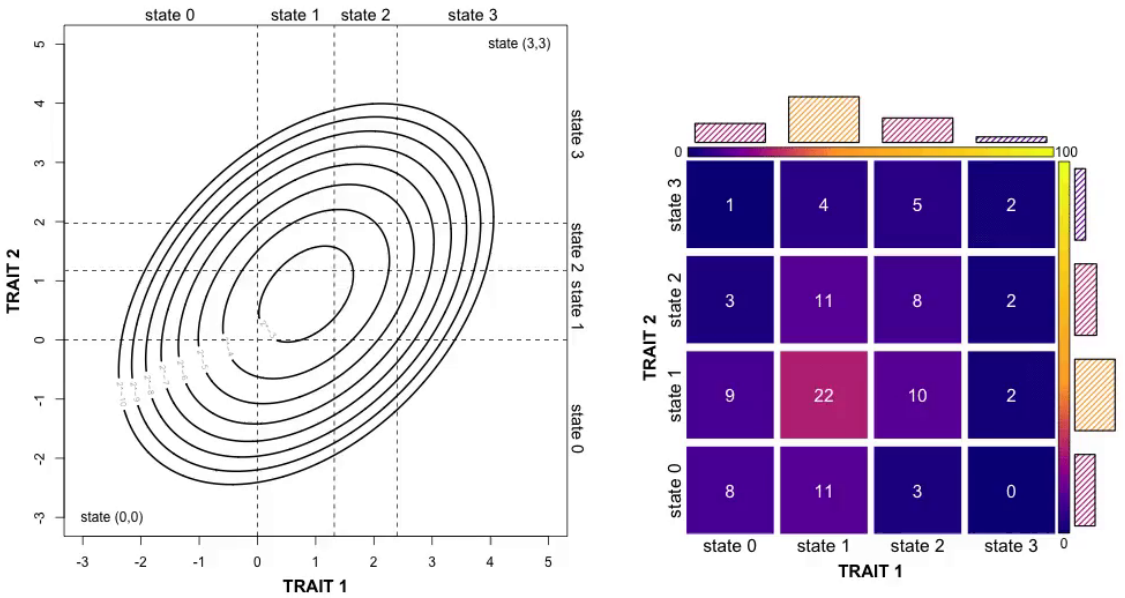

Figure S4.5: Visualizing the effect of smooth variation in the locations of two thresholds on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

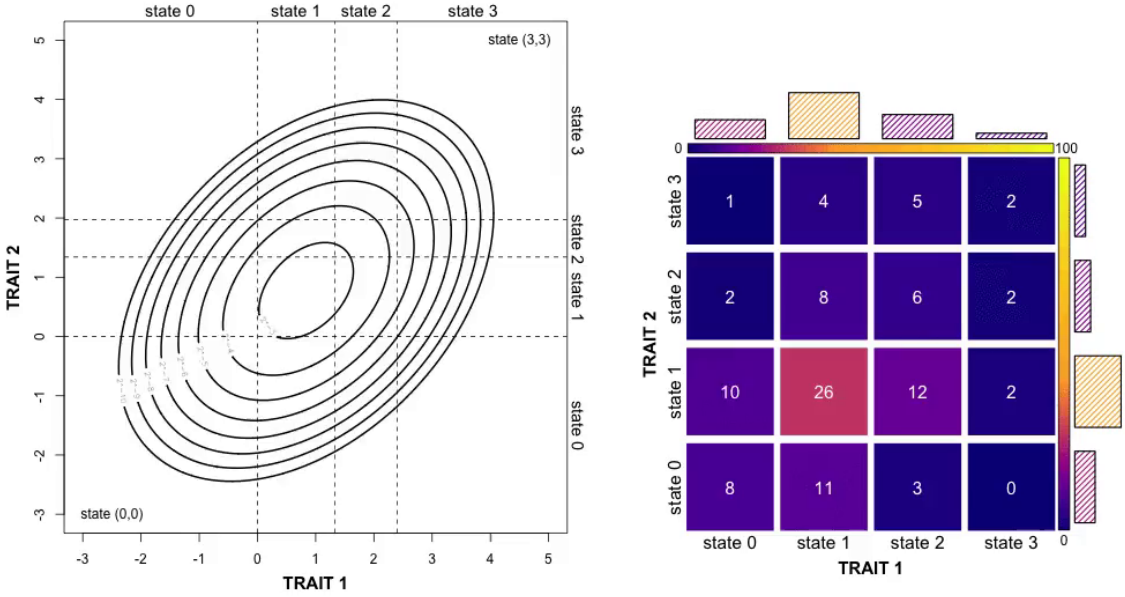

Figure S4.6: Visualizing the effect of smooth variation in a correlation coefficient describing the non-independent expression of two traits on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

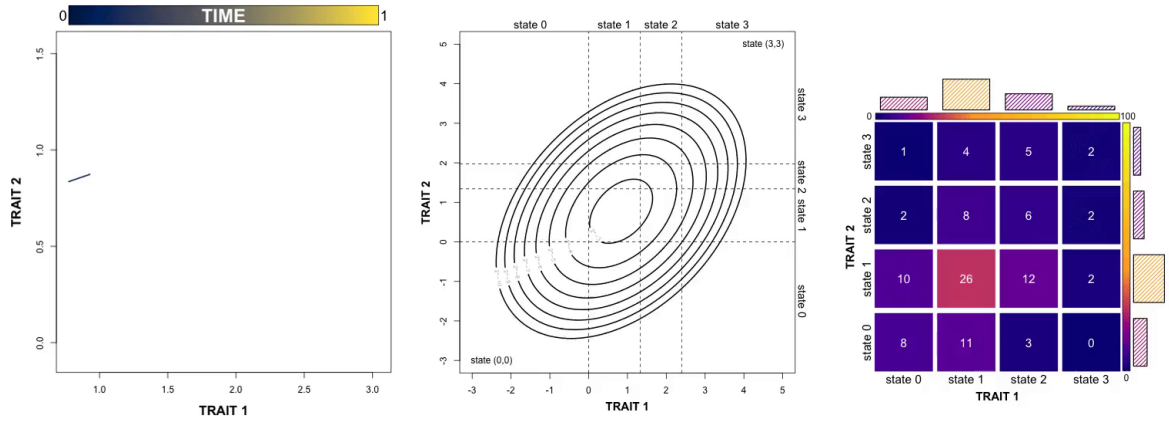

Figure S4.7: Visualizing the effect of a bivariate Brownian motion process acting on a bvariate mean liability on the ordinal trait frequencies of a hypothetical population. MP4 viewable with a compatible PDF Reader.

Chapter 5

Conclusion

Nikolai G. Vetr

5.1 Does it work?

Having finally arrived at the end of these chapters, we ask: does model-based inference of phylogeny from morphological data *work*, and if not, what hope remains of its success? We have endeavored here to inject a greater dose of biological realism into the character evolutionary models used for phylogenetic inference, and so it is natural to ask whether that additional realism adequately captured vital aspects of the data-generating process not present in simpler models, and if those improvements were sufficient to retrieve phylogeny accurately and precisely. The trivial answer is that, to the extent that a statistical phylogenetic model may make predictions about its output that vary on the basis of tree topology or other parameters, those output can be used to make inference about those parameters in both Maximum Likelihood / Optimization and Bayesian frameworks. Assuming the model used for inference is not too poorly specified, probabilistic inferences should be well calibrated (Chapters 2, 4), with estimated nodal posterior probabilities trustworthy, which is to say accurate reflections of our uncertainty with respect to phylogenetic model parameters.

Probabilities rarely feel intuitive, however (Griffiths and Tenenbaum, 2001). We see high (e.g. > 0.95) and low (e.g. < 0.05) values as all but guaranteed, and are disproportionately surprised when their corresponding hypotheses come into conflict with independent inferential results or simulated ground-truth. Our surprise is even greater in point-estimated contexts — when an algorithm whispers to us some hypothesis optimally satisfying the criteria we provide it, we balk at the thought of error, of inconsistency with our target reference (Collard and Wood, 2000; Gibbs et al., 2000). But no statistical procedure guarantees perfection, and estimation error is entirely expected outside the limit of infinite data. It is through simulation that we may learn to temper our own expectations about what sorts of conclusions we are able to reach, training ourselves to be more skeptical of inferential results that we may more easily forgive them when their falsity is revealed.

The methods developed here for inferring phylogeny from continuous and discrete morphological data appear to work when our expectations are so tempered. Plenty of

information resides in continuous characters (Chapter 2), and it appears they can retrieve trees consistent with those obtained from molecular data commensurate with that information (Chapter 3). These trees are almost certainly all *wrong*, which is to say that due to how diffusely the joint posterior distribution is spread, the reference or true tree is almost guaranteed never to have been sampled. But that is not to say that the inference procedure has failed us; perhaps more accurately, it is we who have failed the inference procedure, by failing to provide it with enough information that it not disappoint.

That said, it may be that fundamental limits will forever impede arbitrarily precise phylogenetic inference from morphological data. As described in the Introduction, morphological evolution is messy, and adequate models may quickly outpace in parameters both our computational resources and the extent of information we are able to squeeze from available datasets. Molecular sequences are a trove of (nearly) independently evolving loci, each possessing some small, insignificant signal of phylogeny, the pooling of which enables tremendously powerful inference under sophisticated phylogenetic models (Kapli et al., 2020). In contrast to these thousands or millions of characters, morphological datasets rarely breach a hundred, and non-independence between characters in the presence of noise — environmental, developmental, inferential, and measuremental — quickly reduces the contribution of each marginal addition (Chapter 2), especially if novel parameters must be introduced to represent more complex morphological evolutionary processes.

One may more easily imagine the impact of noise by imagining the rate matrix diagonalization procedure in Appendix A to not use Cholesky decomposition, but rather eigendecomposition. Absent noise, later eigenvectors (with smaller eigenvalues) correspond to equally informative axes of variation as earlier ones (Figure 5.1a-b, $\sigma^2 = 0$). If covariance patterns in the noise are equivalent to that of the multivariate Brownian rate matrix, we likewise have not damaged ourselves much, merely extending terminal branch lengths by some small, if non-trivial amount (Figure 5.1b). But if the noise is emitted from a distribution unlike the rate matrix, it may entirely obliterate the signal from those later eigenvectors if their eigenvalues are sufficiently small (Chapter 2, Figure

5.1a). The story may even be more tragic than that portrayed (Figure 5.1a), as the rate and noise matrices in these simulations are drawn from quite similar distributions, with the LKJ($\eta = 1$) concentrating most of its probability at high dimension near 0, implying the marginal distribution of each correlation coefficient to be beta(50, 50) stretched to the (-1,1) range. But morphological data at high dimension are likely not distributed according to flat correlation structure, and would feature far steeper scree plots than those implied by the flat LKJ (Figure 5.1c), potentially disassociating projected trait values on later eigenvectors even more steeply and so more steeply diminishing the contribution of each subsequent measured trait. New techniques to characterize the human and primate morphomes will bring with them increasingly larger alignments of character data, but whether we will be able to use those data to our own satisfaction remains to be seen.

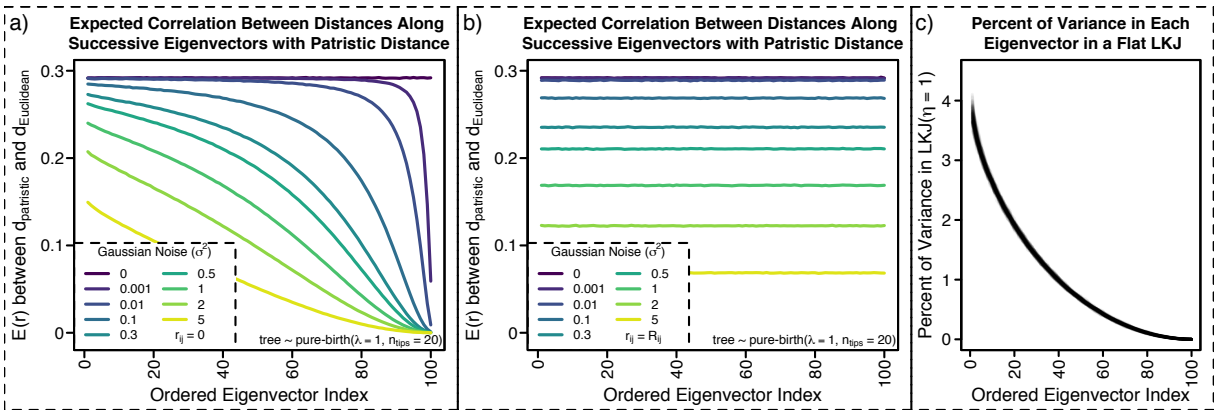

Figure 5.1: Visualizing the association, here given as Pearson's r, between Euclidean distances between tips along succesive eigenvectors of a multivariate Brownian rate matrix and patristic distances on the tree on which those data evolved. For these figures, trees were first sampled from a pure birth distribution with 20 tips and $\lambda = 1$ (tree height $\approx 2.6 \pm 0.8, \mu \pm \sigma$). Then continuous character evolution was simulated on the sampled tree according to multivariate Brownian motion with a 100 x 100 rate matrix drawn from an LKJ($\eta = 1$). Tip characters were then perturbed with normally distributed noise and projected onto the eigenvectors of the Brownian rate matrix, and correlations between the distances separating each of the tips on each of these new axes and their corresponding patristic distances evaluated. In a), we see the expected correlation when the noise distribution is iid standard normal, as distinct from each sampled rate matrix. In b), noise is drawn from a multivariate normal with correlation structure equal to that of the multivariate Brownian rate matrix, effectively extending each tip by some amount σ^2. In c), the proportion of total variance corresponding to each eigenvector is shown, resulting in a distribution of scree plots implied by the LKJ.

5.2 Limitunities

Of course, we have far from exhausted our abilities to specify and fit models of morphological character evolution, and many opportunities await those willing to work past present limits. Joint inference represents the lowest hanging of these fruits, not only with respect to the use of full likelihoods, absent conditioning on erroneous optima (Chapter 4) estimated with imprecise approximations to our desired functions, but also through the pooling of information across molecular and morphological partitions. Much as fossils can tell us about morphological evolutionary processes across extant data (Slater et al., 2012), so too can extant morphologies distributed across well-resolved molecular phylogenies tell us about the phylogenies of fossils. For those parts of the tree where nucleotide sequence (or other molecular) data are available, let *them* sing their signals loud and clear, hopefully harmonizing across their thousands or millions of voices. Firm inference where available could then help to scaffold our understanding of local morphological evolutionary processes, as well as their variation. In those parts of the tree where molecular data are absent or unobtainable, hierarchical model structure will provide adaptively informative prior distributions for morphological evolutionary model parameters, that information present in those morphologies not need to spread itself too thin learning the location and shape of nuisance.

Model averaging may also help us to use more sophisticated models in proportion to the justification for their use. Developing more efficient jump moves (Huelsenbeck et al., 2004) to move between high and low dimensional parameter spaces will allow us to sample from highly parameterized models according to their posterior probability, which will be lower according to the extent to which that model's marginal likelihood succeeds or fails to concentrate probability in favorable regions of the prior. With model averaging, phylogenetic inference can more organically explore the space of plausible models specified *a priori*, averaging inference of focal model parameters pooled across those models (such as tree topology) to obtain a joint posterior distribution that makes use of multiple attempts to represent the underlying data-generating process. To the extent that simpler processes are preferred, the ensemble may visit them more, leveraging their simplicity to make more

confident inference where possible.

On the model-specification front, the largest pitfalls of the preceding research are likely to pertain to assumptions of process homogeneity — the constancy of some set of model parameters, such as trait-specific rates or between-trait correlations — throughout the whole of our phylogenetic tree. Hierarchical model structure is the obvious solution. By specifying that some set of branch-specific model parameters are drawn from a common population of such parameters with some distribution, we can pool inference across branches while still accommodating branch-specific process heterogeneity. This strategy was employed in both chapters 3 and 4, as well as during several side-projects undertaken during this dissertation, both mentioned (e.g. in Figure 1.1) and unmentioned. Clever solutions exist for computing phylogenetic likelihoods in the face of such heterogeneity (e.g. Caetano and Harmon, 2019), but where such solutions are not fast forthcoming, data augmentation may also be used over internal node states to dismantle a complex problem into a series of simpler ones (though not without its own difficulties, e.g. with respect to making efficient proposals). In this work, we strove to use populations and species for whom non-reticulate tree structure might apply with minor apologies. But population dynamics at sub-specific scales are likely to feature ample admixture, especially in such versatile, mobile, and promiscuous species as are found in our own lineage (Prüfer et al., 2014; Durvasula and Sankararaman, 2020). Accommodating reticulate network structure in the graphs we use to describe the evolution of morphological character data under multivariate Brownian motion (Pickrell and Pritchard, 2012; Bastide et al., 2018) could be both intrinsically interesting and essential to accurate recovery of the underlying tree. And bounded Brownian motion processes, which are briefly mentioned in Appendix C, could also help to improve inference insofar as our taxa of interest might clash against hard limits in morphospace across the time scales separating them (Estes and Arnold, 2007).

Finally, there exist a paucity of posterior predictive summary statistics (Gelman et al., 1996) for use in model adequacy checks of morphological character evolutionary models, especially multivariate Brownian models. Elaboration of a plausible set of such statistics,

as well as an exploration of their properties, could help to both identify when and where evolutionary models fail, as well as provide suggestions for how a model might be modified in order to better capture properties of our applied empirical context.

Cliché though it may read, there remains much work to be done in the world of Bayesian morphological phylogenetics. In *this* work, we explored the performance of multivariate Brownian models in a variety of simulated contexts, as well as with application to three empirical datasets drawn from the faces and dentitions of various monkeys and apes. Along the way, we developed both novel computational algorithms (such as those described in the appendices), as well as useful data visualization tools to characterize the informativeness of diffuse joint posterior distributions. The empirical results presented here are not intended to serve as final words on any outstanding questions, as might be expected given their focus on data able to be matched to molecular sequences. But they should serve to contextualize data analyses bereft of such convenience for paleontological samples whose use in the face of future discovery is all but guaranteed.

Appendix A

Factorization of the Phylogenetic Likelihood

Below can be found a short derivation of one decomposition for the phylogenetic multivariate Brownian likelihood function, used internally to accelerate the calculation of phylogenetic likelihoods in RevBayes and other software. It involves two separate factorizations that allow for a very high dimensional multivariate normal density to be expressed as the product of univariate normal densities. These factorizations exploit properties of the multivariate Brownian motion rate matrix R and Phylogenetic Covariance Matrix P, as well as the relation between them.

The full phylogenetic likelihood can be written as

$$|2\pi(R \otimes P)|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T (R \otimes P)^{-1} (x-\mu)}$$

Which is the recognizable probability density function of a multivariate normal distribution. The $\otimes$ symbols represents the Kronecker product, which multiples R and P to form a covariance matrix. Meanwhile, x represents the states at the tips, stacked into a vector according to the order appropriate that in which the Kronecker product is taken. The symbol μ likewise represents the state at the root of the tree, repeated also in the appropriate order.

The two algorithms used here to diagonalize the matrices R and P are Cholesky factorization and the Felsenstein Pruning algorithm, respectively. Cholesky factorization provides the decomposition $R = LIL^T$, where L is the lower Cholesky factor and I is the identity. The Felsenstein pruning algorithm, meanwhile, decomposes the phylogenetic covariance matrix P into $C^{-1}QC^{-T}$, where Q is a matrix representing the branch lengths of each independent contrast, and C is the matrix that transforms the tip characters into independent contrasts. This decomposition is often represented as a post-order tree traversal, but can also be expressed and performed linear algebraically, operating purely on the phylogenetic covariance matrix.

Substituting these relations into the above expression, we obtain:

$$|2\pi(LIL^T \otimes C^{-1}QC^{-T})|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T(LIL^T \otimes C^{-1}QC^{-T})^{-1}(x-\mu)}$$

Using the properties of Kronecker products, we can rewrite this as:

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T((L\otimes C^{-1})(I\otimes Q)(L^T\otimes C^{-T}))^{-1}(x-\mu)}$$

Which can be further reordered and simplified:

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T(L^T\otimes C^{-T})^{-1}(I\otimes Q)^{-1}(L\otimes C^{-1})^{-1}(x-\mu)}$$

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T(L^{-T}\otimes C^T)(I\otimes Q)^{-1}(L^{-1}\otimes C)(x-\mu)}$$

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T(L^{-1}\otimes C)^T(I\otimes Q)^{-1}(L^{-1}\otimes C)(x-\mu)}$$

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}((L^{-1}\otimes C)(x-\mu))^T(I\otimes Q)^{-1}(L^{-1}\otimes C)(x-\mu)}$$

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}} e^{-\frac{1}{2}((L^{-1}\otimes C)x-(L^{-1}\otimes C)\mu))^T(I\otimes Q)^{-1}((L^{-1}\otimes C)x-(L^{-1}\otimes C)\mu)}$$

And since

$$C\mu = 0$$

and

$$(L^{-1} \otimes C)\mu = 0$$

We can further simplify this into:

$$((2\pi)^{nm}|LIL^T|^m|C^{-1}QC^{-T}|^n)^{-\frac{1}{2}}e^{-\frac{1}{2}((L^{-1}\otimes C)x)^T(I\otimes Q)^{-1}((L^{-1}\otimes C)x)}$$

As I and Q are both diagonal matrices, we obtain in the place of our original function a multivariate normal distribution pdf whose covariance matrix is a diagonal matrix, with trait vector x appropriately transformed. This allows us to compute the original density as the product of univariate normal densities (or sum of log densities), so long as we take care to propagate the appropriate quantities to the determinant.

Appendix B

More Efficient Proposals Over Correlation Matrices

B.1 Motivation

Correlation matrices are a special case of covariance matrix. They have all the latter's properties, including symmetry and positive semi-definiteness (PSD), in addition to the requirement of a unit diagonal. In Bayesian inferential contexts involving covariance matrices, it is common to factor covariance matrix C into the matrix product SRS, where S is a diagonal matrix of standard deviations and R is a correlation matrix. Priors can then be specified on these components separately. Under the phylogenetic multivariate Brownian motion model, the rate matrix is a covariance matrix, describing the shape of the multivariate normal distribution — scaled multiplicatively by the branch length — from which displacements to the location of some vector of traits is drawn.

A problem arises when making proposals to correlation matrices during the Metropolis-Hastings algorithm (Hastings, 1970) used to numerically approximate the Bayesian joint posterior distribution of model parameters. The problem involves the PSD constraint, which with increasing matrix dimensionality rapidly narrows the window of viability available to each correlation coefficient, conditional on the values observed at all other positions of the correlation matrix. This is implied by the narrowing marginal distribution of each coefficient in samples from an LKJ distribution, which can be described by a beta distribution stretched into the (-1,1) range with shape parameters $\eta - 1 + D/2$, where D is the dimension of the correlation matrix. It can also be demonstrated experimentally by sampling a correlation matrix from a flat LKJ distribution ($\eta = 1$), selecting one of its off-diagonal elements, and identifying the magnitude of the interval about that element for which the PSD condition is preserved. Averaging across many replicates, we obtain an estimate of the expected interval over which proposals to individual elements of a correlation matrix may be made for matrices of a given size, which we summarize graphically (Figure B.1).

As this interval shrinks, the Metropolis-Hastings algorithm slows tremendously with respect to its mixture over the posterior distribution of correlation matrices if naive proposal distributions are used that do not respect the constraint of positive semi-definiteness. Additionally, the number of parameters to be estimated increases quadratically with the

dimension of the correlation matrix, as a correlation matrix of dimension n has (n choose 2) off-diagonal elements. And tests for positive-semidefiniteness — for example, involving eigendecomposition and the surveyance of positive eigenvalues — can themselves become quite computationally cumbersome, especially at high dimensionality. As such, approaches that rely on the rejection of invalid proposals must either sample from more and more conservative distributions over increasingly many elements of the correlation matrix, or else make tremendously more proposals over the course of the MCMC.

Here, we describe a proposal distribution over correlation matrices that makes proposals to multiple correlation parameters simultaneously and is guaranteed to sample from the PSD space. It is asymmetric, and our description therefore also entails the calculation of forward and backward proposal probabilities. In addition, it is tunable on a per-trait-index basis, allowing for the optimization of acceptance probabilities contingent upon the narrowness of the target distribution. Finally, we describe its implementation with respect to the Cholesky factor of the correlation matrix, which is a commonly used internal representation of a correlation matrix for the purpose of computational efficiency (e.g. in transforming tip characters or specifying non-centered parameterizations in Gaussian Process Regression). The proposal distribution also samples a new Cholesky factor in $O(n^2)$ time complexity, removing the need to perform a more costly $O(n^3)$ Cholesky factorization. One minor complication of the below described proposal distribution involves its return of a Cholesky factor corresponding to a permuted correlation matrix, rather than one with the original row and column indexing. In our implementation, an optional argument can be toggled to return to the original order at the cost of some of the aforementioned efficiency. Instead, we prefer to permute the (e.g. multivariate normal) data whose density is being calculated, a much more efficient procedure that nevertheless requires slightly more bookkeeping.

Following our description of the proposal distribution, we conduct a short validation of its performance, first using it to sample via the Metropolis-Hastings algorithm from a flat LKJ distribution — uniform over all correlation matrices — and comparing these samples to those drawn directly from the LKJ. Then, we use the proposal distribution to

approximate a more informative distribution: the posterior distribution of the correlation matrix of a 10-dimensional multivariate normal random variable whose means and variances are known, conditional on 50 samples from that distribution and an LKJ($\eta = 1$) prior. We compare these to the same distribution independently approximated through more innocuous means: a uniform sliding window proposal over all choose(10, 2) pairwise elements of the correlation matrix. The former is run for 1E7 iterations and the latter for 5E7, with thinning occurring at an interval of 1E3 and 5E3, respectively. The first 20% of each chain is discarded as burn-in, and automatic tuning is performed across 50 sub-rounds of burn-in to target a per-trait (in our novel proposal distribution) and per-correlation (in the uniform sliding window proposal distribution) acceptance probability of 0.234 (Roberts et al., 1997). In practice, this resulted in acceptance probabilities ultimately falling in the interval (0.20, 0.24).

B.2 Description

The proposal distribution was inspired by the *extended onion method* of the 2009 paper by Lewandowski, Kurowicka, and Joe for sampling from the eponymous LKJ distribution. In this algorithm, the Cholesky factor of a correlation matrix is built iteratively up to the desired dimensionality n, one layer at a time, as one might grow an onion. Our proposal distribution simply unravels the last step of the extended onion method, corresponding to the construction of the correlations of the nth variable with the remaining n - 1 variables of the n x n correlation matrix. In the extended onion method, this final column of the upper Cholesky factor U is constructed through sampling a single beta distributed random variable y from a beta($n/2$, η), and a vector $u = (u_1, ..., u_{n\text{-}1})$ uniformly sampled from the surface of an n-dimensional hypersphere. The u are then rescaled by the square root of y to form the first n - 1 elements of the column of U. The square root of (1-y) then forms the last entry of that column, ensuring unit length.

Working backwards, we can straightforwardly identify the unique y and u necessary to produce some current U, conditional also on its (n - 1) x (n - 1) submatrix. We can then resample these values centered on these targeted states — disallowing any room around them will reproduce U, and resampling from $y \sim$ beta($n/2$, η) and $u \sim$ uniform on the unit hypersphere will generate a sample from the conditional LKJ(η). Between these extremes, we can resample y' from within a window of tunable width w centered around the target y in proportion to its probability density in beta($n/2$, η) through the cumulative distribution and quantile functions of the beta distribution, with forward and backward proposal probabilities equal to the cumulative probability contained within each window centered on the forward state y and backward state y', i.e. found by integrating the beta probability density function between each set of bounds, which has positive density in (0,1) and zero density elsewhere. Resampling u', meanwhile, can be accomplished by sampling n - 1 normal random variables with mean = u and variance = tuning parameter v, and then rescaling this vector to unit length. By rescaling, we permit infinitely many samples from the aforementioned normal distributions to correspond to the same u', all falling on the vector stretching from the origin through u'. Forward and backward proposal

probabilities can be obtained by integrating along this vector the multivariate normal probability distribution function with mean equal to the target state u or u' and covariance matrix diagonal with entries v. However, symmetry in the proposal distributions results in these integrals evaluating to the same value, and so the proposal ratio for this step can be set to 1.

As yet, the above proposal distribution samples only correlations of the final row and column of some correlation matrix $\mathrm{R} = U^T U$. Its aggressiveness is controlled by two tuning parameters, w and v, though in practice we find that setting w to some small value, such as 0.1, and tuning only v is sufficient to achieve desired acceptance probabilities on moderately informative target distributions, such as those used in the second experiment below. For more informative target distributions, tuning w may also become necessary. Working on only the last dimension of U can be inefficiently accommodated by recomposing R, permuting, and refactoring. However, we can more cheaply remain in upper Cholesky form through two $O(n^2)$ steps, a rank-one update and downdate involving a series of Givens rotations, implemented in the R-package *mgcv* (Wood, 2015) as the function `choldrop`, followed by the solution of a triangular system of equations to reinsert the removed column in the final place, accomplished with base-R (Team, 2013) function `backsolve`. To return the Cholesky factor to its original order, this procedure could be performed n - i times, where i is the index of the character deleted and reinserted. This is still more efficient than Cholesky factorization, but up to $2n$ times more costly than the small burden of permutation index bookkeeping.

B.3 Validation

Having implemented this proposal distribution in R, we seek to validate its performance in simulation, first by sampling uniformly from the space of correlation matrices by setting the acceptance probability to the ratio of proposal probabilities, implicitly sampling from an LKJ(1) with dimension 10. After having done so, we sample directly and independently from the analytical distribution, and visually inspect quantile-quantile plots of the marginal correlations and matrix determinants (Figure B.2).

Noting that these fall along the 1-to-1 line, we move on to explore sampling from a more tightly constrained distribution of correlation matrices, that defined by the joint posterior distribution of correlation matrices of a multivariate normal random variable with mean and variances known and equal to 1 and 10 x 10 correlation matrix drawn from an LKJ($\eta = 1$) (Figure B.3).

Specifically, this distribution represents the compromise between our flat LKJ($\eta = 1$) prior and the information contained in 50 samples from the above described multivariate normal, with target distribution density equal to the probability densities of our 50 draws with correlation matrix sampled and means and variances fixed. Comparing the samples obtained using the novel proposal distribution to those using the conventional sliding window, we first visually examining marginal trace plots and histograms and find that they closely overlap. Following that, we construct a similar figure as above, visualizing quantile-quantile plots and covariance patterns in both sets of samples and noting that they hew close to the 1-to-1 line, indicating that the same distribution is being sampled in both cases (Figure B.4).

Then, we coerce both arrays of correlation matrices to `mcmc.list` objects and evaluate Gelman and Rubin's Convergence Diagnostic (Gelman and Rubin, 1992) in the R package *coda* (Plummer et al., 2006), finding for all pairwise correlations upper 95% confidence interval values of 1.00 and a multivariate $\hat{R}$ of 1. Effective sample sizes (ESS) of marginal correlations for both chains together ranged between 3,793 and 14,483, though much of those were from the chain using our novel proposal distribution, whose ESS ranged between 3,279 and 8,052, rather than from the sliding window chain, whose ESS

ranged between 458 and 6,482, despite the latter having been tuned to have elementwise acceptance probabilities of 0.234 and being run for fivefold the number of iterations.

B.4 Conclusion

Thus, we have described, implemented, and validated a novel proposal distribution for correlation matrices. Our validation used a relatively small 10 x 10 correlation matrix — much greater improvements would be found with matrices of higher dimensionality. Often, Cholesky factorization counts among the most computationally steps of phylogenetic likelihood calculation in multivariate Brownian models, and circumventing this step by making proposals directly to the Cholesky form should enable tremendous gains to efficiency when performing inference of the character evolutionary processes governing the evolution of multiple traits. Even if one desires proposals directly on correlation matrices, making smarter proposals to multiple correlations simultaneously without the need for expensive PSD checks and rejection sampling should offer a substantial improvement to the use of more naive proposal distributions that do not intrinsically respect their constraints and properties.

B.5 Figures

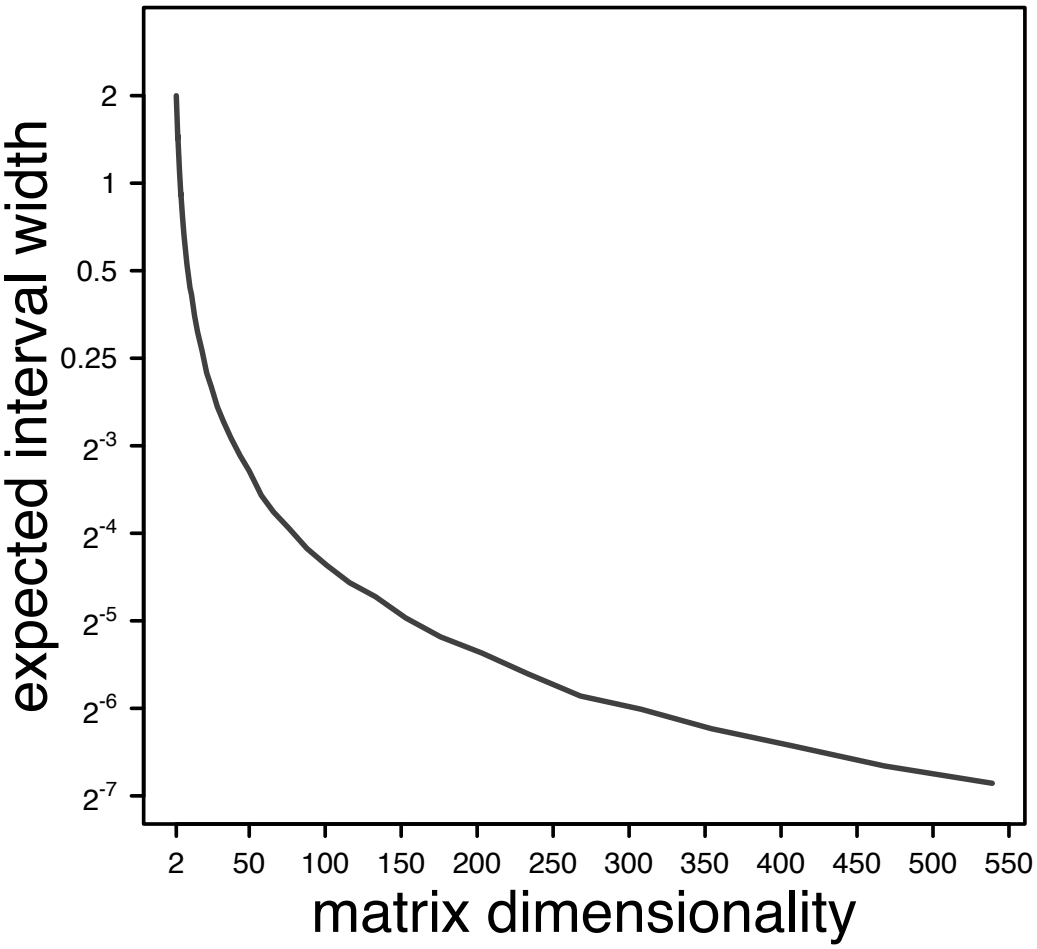

Figure B.1: Depicting the relationship between how much space, in expectation, is available in a random correlation matrix of given dimensionality for valid proposals to marginal correlation coefficients. Quantities approximated using Monte Carlo simulation using samples from an LKJ($\eta = 1$). A random correlation coefficient was then chosen and perturbed in both directions by increasing amounts until the PSD constraint was violated.

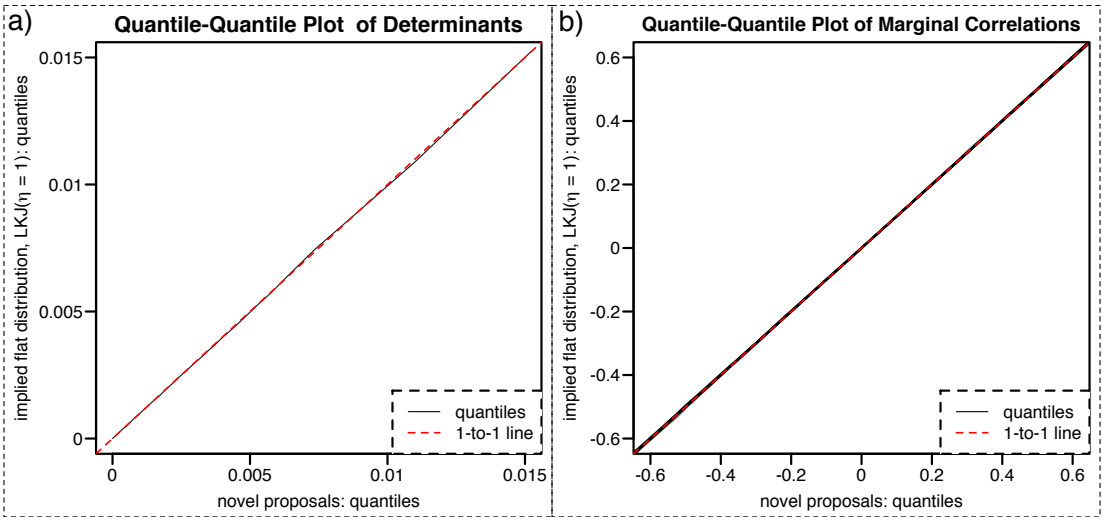

Figure B.2: Depicting an initial experiment in the correct specification of the novel proposal distribution described herein. Target distribution density ratio was set to 1, implicitly sampling 10 x 10 correlation matrices from an LKJ($\eta = 1$). Samples were then generated from this distribution directly, and the quantiles of marginal correlation coefficients and matrix determinants compared.

1	-0.224	-0.175	-0.406	0.295	-0.028	0.199	-0.219	-0.211	-0.275
-0.224	1	0.365	0.326	0.008	-0.004	-0.026	0.235	0.016	0.2
-0.175	0.365	1	0.49	-0.139	0.539	-0.497	0.678	0.572	0.372
-0.406	0.326	0.49	1	-0.493	0.626	-0.134	0.324	0.621	0.22
0.295	0.008	-0.139	-0.493	1	0.009	0.041	0.044	-0.387	0.2
-0.028	-0.004	0.539	0.626	0.009	1	-0.246	0.393	0.451	0.363
0.199	-0.026	-0.497	-0.134	0.041	-0.246	1	-0.882	0.038	-0.417
-0.219	0.235	0.678	0.324	0.044	0.393	-0.882	1	0.233	0.523
-0.211	0.016	0.572	0.621	-0.387	0.451	0.038	0.233	1	-0.079
-0.275	0.2	0.372	0.22	0.2	0.363	-0.417	0.523	-0.079	1

Figure B.3: The 10 x 10 covariance matrix used to generate samples from a multivariate normal distribution in the "informative target distribution" experiment. Generated by sampling from an LKJ($\eta = 1$) distribution after calling `set.seed(1)` in *R*.

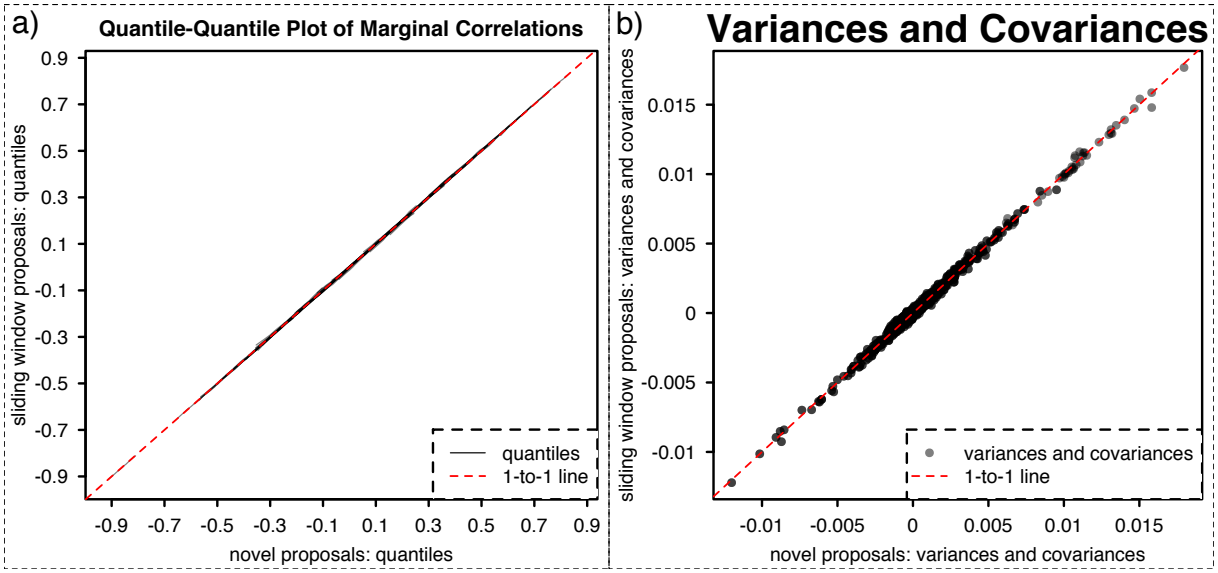

Figure B.4: A visual comparison of MCMC output obtained using both our novel correlation matrix proposal distribution and a more standard uniform sliding window proposal distribution. In a), quantiles of each marginal correlation coefficient are plotted. In b), the variances and covariances of each marginal correlation coefficient for both chains are plotted.

Appendix C

Reconstructing Ancestral Histories Under Multivariate Brownian Motion

C.1 Motivation

When simulating under a multivariate Brownian motion model of character evolution, the location of the resultant continuous character alignment is marginalized out during phylogenetic likelihood calculation with the Felsenstein Pruning Algorithm, because if both the root state and location are unknown, the full likelihood is completely non-identifiable with respect to either. Adding the same value to all tip characters changes the likelihood not at all, so it matters not what arbitrary value is assigned to the root of the tree. When those continuous characters are passed through an ordinalization filter, however, there arises a strong dependency between the location of the root state and the phylogenetic likelihood, and with it our ability to estimate focal model parameters. This is because the locations of the latent liabilities relative to the locations of the thresholds determine which discrete characters are expressed. Shifting the root state can dramatically change which discrete traits are expressed without a concordant shift in the locations of the thresholds, and so if we wished to simulate fictitious data in an empirically realistic manner, we might wish to fix the thresholds to their estimated values, leaving open the question of where and with what value to set the root.

The location of the root is still non-identifiable, so, as mentioned in the main text, we arbitrarily performed midpoint rooting. The most convenient values to fix its state to for the purpose of estimating between trait correlations, threshold locations, and tip means would be intermediate between the minimum and maximum thresholds, but empirical realism does not beget convenience. Instead, the most principled root state would be that implied by the root's location, midway between the most distantly separated tips, conditional on the observed values throughout the rest of the tree and the character evolutionary process. The multivariate Brownian bridge passing through the root, then, on its way between the tips.

C.2 Algorithm

Finding the parameters of this distribution is straightforward — one needs only to recall that under multivariate Brownian motion, the tips are drawn from a multivariate normal random variable whose mean is the root state and whose covariance matrix is the Kronecker product of the phylogenetic covariance matrix and the Brownian motion rate matrix. Because of the pulley principle, one can arbitrarily reroot the tree at any tip and equivalently say that the remaining tip data, along with the unobserved state at the midpoint root, are multivariate normal distributed with mean at the observed tip and covariance matrix the Kronecker product of the *new*, rerooted tree and the Brownian motion rate matrix. Depending on the order one took the Kronecker product, as well as the order of the columns in the phylogenetic covariance matrix, one needs to then permute our multivariate normal distribution's covariance matrix into block form, such that those indices corresponding to unobserved characters at the root are found in the upper left corner. Then, one case use known formulae for conditional distributions of multivariate normals to compute the conditional mean and covariance at the root, where if one has a multivariate normal distribution with mean μ and covariance matrix Σ, partitioned as

$$\mu_{all} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

and realizations partitioned as

$$x = \begin{bmatrix} x_{unobs} \\ x_{obs} \end{bmatrix}$$

The distribution of x_{unobs} given x_{obs} is itself multivariate normal, with means and covariance matrix:

$$\mu_1' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_{obs} - \mu_2)$$

$$\Sigma'_{11} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

This is the same relation exploited by the Felsenstein Pruning Algorithm to diagonalize the phylogenetic covariance matrix, only it computes conditional distributions conditional upon values at descendant nodes, rather than on tips also descended from sister branches.

C.3 Additional Applications

One character history of particular interest to paleontologists and neontologists alike might be the truncated biogeographic diffusion history separating taxa distributed across some geographic range, especially under hypotheses of correlation between geographic (e.g. latitudinal) and morphological variables. Fitting a Brownian motion model to such data can help to investigate latitudinal trends in body shape and size in a phylogenetically sensible way, elegantly avoiding pseudoreplication problems encountered when failing to take phylogeny or population history into account. But naively specifying a multivariate Brownian motion over the whole of R^n does not respect geographic boundaries that might impede an organism's travel, such as mountains, deserts, and oceans (Figure C.1). Preliminary work found that it was possible and efficient to perform data augmentation over internal geographic and morphological states with sufficient granularity as to avoid ocean-faring and mountain-hopping voyages by rejection sampling biogeographic diffusion histories that were consistent with these boundaries. As most such histories would be invalid at high granularity, one could instead rejection sample single locations at a time, conditioning on previously accepted locations. By first passing over states at internal nodes, one would only need to condition on the start and end of a branch to sample the history within it, also keeping count of the product of rejected proportions across subsequent sampled points when proposing novel histories. Parameters of the multivariate Brownian rate matrix, then, could be inferred using these augmented histories, simultaneously estimating the coevolution of latitude and morphology, in addition to the migratory routes lineages took in their journeys across the continent or globe.

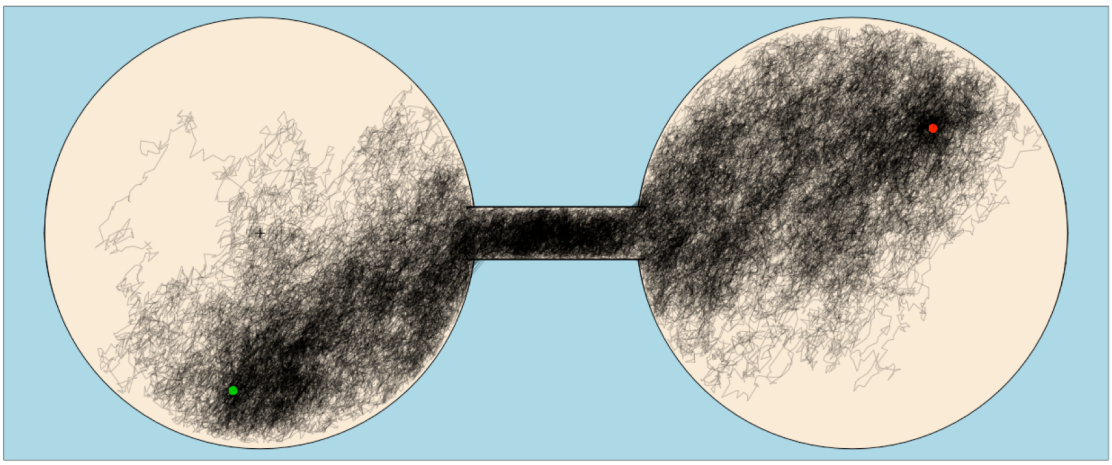

Figure C.1: An example of the truncated Brownian bridges between two tan islands surrounded by nigh-impassable ocean. Biogeographic diffusion began at the green dot and ended at the red dot, with multivariate Brownian paths of 0 correlation between the two sampled efficiently from their conditional distribution.

References

Adams, D., Collyer, M., and Kaliontzopoulou, A. (2019). Geometric Morphometric Analyses of 2D/3D Landmark Data.

Adams, D. C., Rohlf, F. J., and Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix*, 24(1):7.

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., and Willerslev, E. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748):4724–4733.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415.

Álvarez-Carretero, S., Goswami, A., Yang, Z., and Dos Reis, M. (2019). Bayesian estimation of species divergence times using correlated quantitative characters. *Systematic Biology*, 68(6):967–986.

Argue, D., Groves, C. P., Lee, M. S., and Jungers, W. L. (2017). The affinities of Homo floresiensis based on phylogenetic analyses ofácranial, dental, and postcranial characters. *Journal of Human Evolution*, 107:107–133.

Arnold, C., Matthews, L. J., and Nunn, C. L. (2010). The 10kTrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology: Issues, News, and Reviews*, 19(3):114–118.

Arnold, S. J., Bürger, R., Hohenlohe, P. A., Ajie, B. C., and Jones, A. G. (2008). Understanding the Evolution and Stability of the G-Matrix. *Evolution; international journal of organic evolution*, 62(10):2451–2461.

Bailey, S. E. (2002). *Neandertal Dental Morphology: Implications for Modern Human Origins*. PhD thesis, Arizona State University Tempe.

Bastide, P., Solís-Lemus, C., Kriebel, R., William Sparks, K., and Ané, C. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic biology*, 67(5):800–820.

Bates, D. and Maechler, M. (2019). Package 'Matrix'.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53):370–418.

Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. C. (2012). Modeling Stabilizing Selection: Expanding the Ornstein–Uhlenbeck Model of Adaptive Evolution. *Evolution*, 66(8):2369–2383.

Bivand, R., Ono, H., Dunlap, R., Stigler, M., and Bivand, M. R. (2020). Package 'classInt'.

Bliss, C. I. (1934). The method of probits. *Science*.

Blomberg, S. P., Garland Jr, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57(4):717–745.

Blows, M. W., Allen, S. L., Collet, J. M., Chenoweth, S. F., and McGuigan, K. (2015). The phenome-wide distribution of genetic variance. *The American Naturalist*, 186(1):15–30.

Bookstein, F. L. (1997). Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Medical image analysis*, 1(3):225–243.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., and De Maio, N. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650.

Bradley, B. J. (2008). Reconstructing phylogenies and phenotypes: A molecular view of human evolution. *Journal of Anatomy*, 212(4):337–353.

Brazeau, M. D. (2011). Problematic character coding methods in morphology and their effects. *Biological Journal of the Linnean Society*, 104(3):489–498.

Brecht, M. and Freiwald, W. A. (2012). The many facets of facial interactions in mammals. *Current Opinion in Neurobiology*, 22(2):259–266.

Brocklehurst, N. and Benevento, G. L. (2020). Dental characters used in phylogenetic analyses of mammals show higher rates of evolution, but not reduced independence. *PeerJ*, 8:e8744.

Brown, J. M., Hedtke, S. M., Lemmon, A. R., and Lemmon, E. M. (2009). When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. *Systematic Biology*, 59(2):145–161.

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21):161–173.

Butler, M. A. and King, A. A. (2004). Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695.

Caetano, D. S. and Harmon, L. J. (2019). Estimating correlated rates of trait evolution with uncertainty. *Systematic biology*, 68(3):412–429.

Carter, K., Worthington, S., and Smith, T. M. (2014). News and views: Non-metric dental traits and hominin phylogeny. *J. Hum. Evol.*, 69:123–128.

Chamberlain, A. T. and Wood, B. A. (1987). Early hominid phylogeny. *Journal of Human Evolution*, 16(1):119–133.

Cheverud, J. M. (1996). Developmental integration and the evolution of pleiotropy. *American Zoologist*, 36(1):44–50.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Collard, M. and Wood, B. (2000). How reliable are human phylogenetic hypotheses? *Proceedings of the National Academy of Sciences*, 97(9):5003–5006.

Collard, M. and Wood, B. (2001). How reliable are current estimates of fossil catarrhine phylogeny? An assessment using extant great apes and Old World monkeys. *Hominoid Evolution and Climatic Change in Europe: Volume 2: Phylogeny of the Neogene Hominoid Primates of Eurasia*, 2.

Collins, M. J., Nielsen–Marsh, C. M., Hiller, J., Smith, C. I., Roberts, J. P., Prigodich, R. V., Wess, T. J., Csapo, J., Millard, A. R., and Turner–Walker, G. (2002). The survival of organic matter in bone: A review. *Archaeometry*, 44(3):383–394.

Cybis, G. B., Sinsheimer, J. S., Bedford, T., Mather, A. E., Lemey, P., and Suchard, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The annals of applied statistics*, 9(2):969.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection Or the Preservation of Favoured Races in the Struggle for Life.* H. Milford; Oxford University Press.

Darwin, C. (1896). *The Descent of Man and Selection in Relation to Sex*, volume 1. D. Appleton.

Dembo, M., Matzke, N. J., Mooers, A. Ø., and Collard, M. (2015). Bayesian analysis of a morphological supermatrix sheds light on controversial fossil hominin relationships. *Proceedings of the Royal Society B: Biological Sciences*, 282(1812):20150943.

Dembo, M., Radovčić, D., Garvin, H. M., Laird, M. F., Schroeder, L., Scott, J. E., Brophy, J., Ackermann, R. R., Musiba, C. M., and de Ruiter, D. J. (2016). The evolutionary relationships and age of Homo naledi: An assessment using dated Bayesian phylogenetic methods. *Journal of Human Evolution*, 97:17–26.

Demir, F., Oktay, E. A., and Topcu, F. T. (2017). Smile and dental aesthetics: A literature review. *Med Sci*, 6(1):172–7.

Donoghue, M. J. and Moore, B. R. (2003). Toward an integrative historical biogeography. *Integrative and comparative biology*, 43(2):261–270.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis: With Applications in R*, volume 995. John Wiley & Sons.

Durvasula, A. and Sankararaman, S. (2020). Recovering signals of ghost archaic introgression in African populations. *Science advances*, 6(7):eaax5097.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Einstein, A. (1956). *Investigations on the Theory of the Brownian Movement*. Courier Corporation.

Estes, S. and Arnold, S. J. (2007). Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. *The American Naturalist*, 169(2):227–244.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471–492.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution*, 35(6):1229–1242.

Felsenstein, J. (1985a). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791.

Felsenstein, J. (1985b). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.

Felsenstein, J. (1988a). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1):445–471.

Felsenstein, J. (1988b). Phylogenies and Quantitative Characters. *Annual Review of Ecology and Systematics*, 19(1):445–471.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package), Version 3.5 c.* Joseph Felsenstein.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, Mass.

Felsenstein, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1427–1434.

Felsenstein, J. (2012). A Comparative Method for Both Discrete and Continuous Characters Using the Threshold Model. *The American Naturalist*, 179(2):145–156.

Felsenstein, J. and Felenstein, J. (2004). *Inferring Phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fischer, E. K., Ghalambor, C. K., and Hoke, K. L. (2016). Plasticity and evolution in correlated suites of traits. *Journal of evolutionary biology*, 29(5):991–1002.

Fisher, R. A. (1935). The case of zero survivors in probit assays. *Annals of Applied Biology*, 22(1):164–165.

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416.

Garcia-Cruz, J. and Sosa, V. (2006). Coding quantitative character data for phylogenetic analysis: A comparison of five methods. *Systematic Botany*, 31(2):302–309.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445–449.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gelman, A. and Yao, Y. (2020). Holes in Bayesian statistics. *arXiv preprint arXiv:2002.06467*.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Genz, A. and Bretz, F. (2002). Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4):950–971.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., Hothorn, T., and Hothorn, M. T. (2020). Package 'mvtnorm'.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. *Computing science and statistics: Proceedings of 23rd Symposium Interface*, 156.

Gibbs, S., Collard, M., and Wood, B. (2000). Soft-tissue characters in higher primate phylogenetics. *Proceedings of the National Academy of Sciences*, 97(20):11130–11132.

Glazko, G. V. and Nei, M. (2003). Estimation of divergence times for major lineages of primate species. *Molecular biology and evolution*, 20(3):424–434.

Goloboff, P. A., Mattoni, C. I., and Quinteros, A. S. (2006). Continuous characters analyzed as such. *Cladistics*, 22(6):589–601.

Gómez-Robles, A. (2019). Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Science advances*, 5(5):eaaw1268.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.

Griffiths, T. L. and Tenenbaum, B. (2001). Reconciling Intuition and Probability Theory. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, page 370. Lawrence Erlbaum Associates.

Grüneberg, H. (1955). Genetical studies on the skeleton of the mouse XV. Relations between major and minor variants. *Journal of Genetics*, 53(3):515.

Hanihara, T. (2008). Morphological variation of major human populations based on nonmetric dental traits. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 136(2):169–182.

Hansen, T. F. (1997). Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351.

Hansen, T. F. and Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4):1404–1417.

Harmon, L. (2018). Phylogenetic comparative methods: Learning from trees.

Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., and Near, T. J. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution: International Journal of Organic Evolution*, 64(8):2385–2396.

Harmon, L. J., Schulte, J. A., Larson, A., and Losos, J. B. (2003). Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, 301(5635):961–964.

Harvati, K., Frost, S. R., and McNulty, K. P. (2004). Neanderthal taxonomy reconsidered: Implications of 3D primate models of intra-and interspecific differences. *Proceedings of the National Academy of Sciences*, 101(5):1147–1152.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Heath, T. A. and Moore, B. R. (2014). Bayesian inference of species divergence times. *Bayesian phylogenetics: methods, algorithms, and applications*, pages 277–318.

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343.

Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., and Barroso-Ruíz, C. (2014). The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512(7514):306–309.

Hlusko, L. J., Carlson, J. P., Chaplin, G., Elias, S. A., Hoffecker, J. F., Huffman, M., Jablonski, N. G., Monson, T. A., O'Rourke, D. H., and Pilloud, M. A. (2018). Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk. *Proceedings of the National Academy of Sciences*, 115(19):E4426–E4432.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic biology*, 63(5):753–771.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Holland, B. R. (2013). The rise of statistical phylogenetics. *Australian & New Zealand Journal of Statistics*, 55(3):205–220.

Howell, P. G. T. (1987). The variation in the size and shape of the human speech pattern with incisor-tooth relation. *Archives of Oral Biology*, 32(8):587–592.

Howells, W. W. (1973). Cranial variation in man: A study by multivariate analysis of patterns of difference among recent human populations. *Peabody Museum of Archaeology and Ethnology, Harvard Univ.*

Howells, W. W. (1989). Skull shapes and the map: Craniometric analyses in the dispersion of modern Homo. *Papers of the Peabody museum of Archaeology and Ethnology*, 79.

Howells, W. W. (1995). Who's who in skulls: Ethnic identification of crania from measurements. *Papers of the Peabody Museum of Archaeology and Ethnology*, 82.

Hubbard, A. R., Guatelli-Steinberg, D., and Irish, J. D. (2015). Do nuclear DNA and dental nonmetric data produce similar reconstructions of regional population history? An example from modern coastal K enya. *American journal of physical anthropology*, 157(2):295–304.

Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular biology and evolution*, 21(6):1123–1133.

Huelsenbeck, J. P. and Rannala, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic biology*, 53(6):904–913.

Irish, J. D., Bailey, S. E., Guatelli-Steinberg, D., Delezene, L. K., and Berger, L. R. (2018). Ancient teeth, phenetic affinities, and African hominins: Another look at where Homo naledi fits in. *Journal of Human Evolution*, 122:108–123.

Irish, J. D., Guatelli-Steinberg, D., Legge, S. S., de Ruiter, D. J., and Berger, L. R. (2013). Dental morphology and the phylogenetic "place" of Australopithecus sediba. *Science*, 340(6129):1233062.

Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography*, 7:186–190.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., and Poon, A. F. (2016). Ancestral reconstruction. *PLoS computational biology*, 12(7):e1004763.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132.

Kaplan, D. (2004). *The Null Ritual. What You Always Wanted to Know about Significance Testing but Were Afraid to Ask*. Sage Publications Thousand Oaks, CA.

Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, pages 1–17.

Kenkel, M. B. (2015). Package 'pbivnorm'.

Klingenberg, C., Duttke, S., Whelan, S., and Kim, M. (2012). Developmental plasticity, morphological variation and evolvability: A multilevel analysis of morphometric integration in the shape of compound leaves. *Journal of evolutionary biology*, 25(1):115–129.

Klingenberg, C. P. (2008). Morphological integration and developmental modularity. *Annual review of ecology, evolution, and systematics*, 39:115–132.

Klingenberg, C. P. (2014). Studying morphological integration and modularity at multiple levels: Concepts and analysis. *Phil. Trans. R. Soc. B*, 369(1649):20130249.

Kluge, A. G. and Farris, J. S. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Biology*, 18(1):1–32.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.

Lande, R. (1976). Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, 30(2):314–334.

Lande, R. (1979). Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry. *Evolution*, 33(1):402–416.

Lee, M. S. and Palci, A. (2015). Morphological phylogenetics in the genomic age. *Current Biology*, 25(19):R922–R929.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., and Cavalli-Sforza, L. L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104.

Maddison, W. P. (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Systematic Biology*, 40(3):304–314.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., and Tandon, A. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

May, M. R. and Moore, B. R. (2020). A Bayesian Approach for Inferring the Impact of a Discrete Character on Rates of Continuous-Character Evolution in the Presence of Background-Rate Variation. *Systematic Biology*, 69(3):530–544.

Mayr, E. (1988). *Toward a New Philosophy of Biology: Observations of an Evolutionist*. Number 211. Harvard University Press.

Mendes, F. K., Fuentes-González, J. A., Schraiber, J. G., and Hahn, M. W. (2018). A multispecies coalescent model for quantitative traits. *Elife*, 7:e36482.

Mitteroecker, P. and Bookstein, F. (2007). The conceptual and statistical relationship between modularity and morphological integration. *Systematic biology*, 56(5):818–836.

Mitteroecker, P. and Gunz, P. (2009). Advances in geometric morphometrics. *Evolutionary Biology*, 36(2):235–247.

Mitteroecker, P., Gunz, P., Bernhard, M., Schaefer, K., and Bookstein, F. L. (2004). Comparison of cranial ontogenetic trajectories among great apes and humans. *Journal of Human Evolution*, 46(6):679–698.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology letters*, 17(4):508–525.

Mulligan, C. J. and Szathmáry, E. J. (2017). The peopling of the Americas and the origin of the Beringian occupation model. *American journal of physical anthropology*, 162(3):403–408.

Neaux, D., Sansalone, G., Ledogar, J. A., Ledogar, S. H., Luk, T. H., and Wroe, S. (2018). Basicranium and face: Assessing the impact of morphological integration on primate evolution. *Journal of human evolution*, 118:43–55.

Nee, S., May, R. M., and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311.

Nichol, C. R. (1989). Complex segregation analysis of dental morphological variants. *American Journal of Physical Anthropology*, 78(1):37–59.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310.

Nixon, K. C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15(4):407–414.

Omland, K. E. (1999). The assumptions and challenges of ancestral state reconstructions. *Systematic biology*, 48(3):604–611.

O'Reilly, J. E., Puttick, M. N., Pisani, D., and Donoghue, P. C. (2018). Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology*, 61(1):105–118.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342):37–45.

Pagel, M. (1999a). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Pagel, M. (1999b). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic biology*, 48(3):612–622.

Pagel, M. and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, 167(6):808–825.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Parins-Fukuchi, C. (2017). Use of Continuous Traits Can Improve Morphological Phylogenetics. *Systematic biology*, 67(2):328–339.

Parsons, K. J., Márquez, E., and Albertson, R. C. (2011). Constraint and opportunity: The genetic basis and evolution of modularity in the cichlid mandible. *The American Naturalist*, 179(1):64–78.

Pellis, S. M., Pellis, V. C., Reinhart, C. J., and Thierry, B. (2011). The use of the bared-teeth display during play fighting in Tonkean macaques (Macaca tonkeana): Sometimes it is all about oneself. *Journal of Comparative Psychology*, 125(4):393.

Pemberton, T. J., DeGiorgio, M., and Rosenberg, N. A. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes, Genetics*, 3(5):891–907.

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., and Rumpler, Y. (2011). A molecular phylogeny of living primates. *PLoS genetics*, 7(3):e1001342.

Petersdorf, M., Weyher, A. H., Kamilar, J. M., Dubuc, C., and Higham, J. P. (2019). Sexual selection in the Kinda baboon. *Journal of human evolution*, 135:102635.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC evolutionary biology*, 5(1):50.

Pickrell, J. K. and Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8(11):e1002967.

Pickrell, J. K. and Reich, D. (2014). Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics*, 30(9):377–389.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Molecular biology and evolution*, 25(7):1253–1256.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., and De Filippo, C. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

Rabosky, D. L., Grundler, M., Anderson, C., Title, P., Shi, J. J., Brown, J. W., Huang, H., and Larson, J. G. (2014). BAMM tools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, 5(7):701–707.

Rannala, B. (2015). The art and science of species delimitation. *Current Zoology*, 61(5):846–853.

Rathmann, H. and Reyes-Centeno, H. (2020). Testing the utility of dental morphological trait combinations for inferring human neutral genetic variation. *Proceedings of the National Academy of Sciences*, 117(20):10769–10777.

Rathmann, H., Reyes-Centeno, H., Ghirotto, S., Creanza, N., Hanihara, T., and Harvati, K. (2017). Reconstructing human population history from dental phenotypes. *Scientific reports*, 7(1):1–9.

Rein, T. R. and Harvati, K. (2014). Geometric Morphometrics and Virtual Anthropology: Advances in human evolutionary studies. *Anthropologischer Anzeiger*, 71(1-2):41–55.

Reis, M. D., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., and Yoder, A. D. (2018). Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Systematic biology*, 67(4):594–615.

Reusch, T. and Blanckenhorn, W. U. (1998). Quantitative genetics of the dung fly Sepsis cynipsea: Cheverud's conjecture revisited. *Heredity*, 81(1):111–119.

Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, 3(2):217–223.

Revell, L. J. (2014). Ancestral Character Estimation Under the Threshold Model from Quantitative Genetics. *Evolution*, 68(3):743–759.

Revell, L. J. (2020). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Revell, L. J., Harmon, L. J., Langerhans, R. B., and Kolbe, J. J. (2007). A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard Anolis cristatellus. *Evolutionary Ecology Research*, 9(2):261–282.

Reyes-Centeno, H., Rathmann, H., Hanihara, T., and Harvati, K. (2017). Testing modern human out-of-Africa dispersal models using dental nonmetric data. *Current Anthropology*, 58(S17):S406–S417.

Rink, W. J., Schwarcz, H. P., Smith, F. H., and Radovĉiĉ, J. (1995). ESR ages for Krapina hominids. *Nature*, 378(6552):24–24.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Robinson, D. and Foulds, L. (1981a). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Robinson, D. F. and Foulds, L. R. (1981b). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular biology and evolution*, 20(10):1692–1704.

Rodrigue, N., Lartillot, N., Bryant, D., and Philippe, H. (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene*, 347(2):207–217.

Rodrigue, N., Philippe, H., and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular biology and evolution*, 23(9):1762–1775.

Roff, D. A. (1995). The estimation of genetic correlations from phenotypic correlations: A test of Cheverud's conjecture. *Heredity*, 74(5):481–490.

Roff, D. A. (1996). The evolution of genetic correlations: An analysis of patterns. *Evolution*, 50(4):1392–1403.

Rogers, J., Raveendran, M., Harris, R. A., Mailund, T., Leppälä, K., Athanasiadis, G., Schierup, M. H., Cheng, J., Munch, K., and Walker, J. A. (2019). The comparative genomics and complex population history of Papio baboons. *Science Advances*, 5(1):eaau6947.

Rogers, J. S. (1991). A comparison of the suitability of the Rogers, modified Rogers, Manhattan, and Cavalli-Sforza and Edwards distances for inferring phylogenetic trees from allele frequencies. *Systematic Biology*, 40(1):63–73.

Rohlf, F. J. (2002). Geometric morphometrics and phylogeny. *Morphology, shape and phylogeny*, 64.

Ronquist, F. and Sanmartín, I. (2011). Phylogenetic methods in biogeography. *Annual Review of Ecology, Evolution, and Systematics*, 42:441–464.

Roseman, C. C. (2016). Random genetic drift, natural selection, and noise in human cranial evolution. *American journal of physical anthropology*, 160(4):582–592.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., and Soodyall, H. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358(6363):652–655.

Schliep, K. P. (2011). Phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Schluter, D., Price, T., Mooers, A. Ø., and Ludwig, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, 51(6):1699–1711.

Scott, G. R. (1973). Dental morphology: A genetic study of American white families and variation in living Southwest Indians. *Ph. D. Dissertation, Arizona State University*.

Scott, G. R., Pilloud, M. A., Navega, D., d'Oliveira, J., Cunha, E., and Irish, J. D. (2018a). rASUDAS: A new web-based application for estimating ancestry from tooth morphology. *Forensic Anthropology*, 1(1):18–31.

Scott, G. R., Turner II, C. G., Townsend, G. C., and Martinón-Torres, M. (2018b). *The Anthropology of Modern Human Teeth: Dental Morphology and Its Variation in Recent and Fossil Homo Sapiens*, volume 79. Cambridge University Press.

Shea, B. T. (1983). Paedomorphosis and neoteny in the pygmy chimpanzee. *Science*, 222(4623):521–522.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Skelton, R. R. and McHenry, H. M. (1992). Evolutionary relationships among early hominids. *Journal of Human Evolution*, 23(4):309–349.

Slater, G. J., Harmon, L. J., and Alfaro, M. E. (2012). Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution: International Journal of Organic Evolution*, 66(12):3931–3944.

Sodini, S. M., Kemper, K. E., Wray, N. R., and Trzaskowski, M. (2018). Comparison of genotypic and phenotypic correlations: Cheverud's conjecture in humans. *Genetics*, 209(3):941–948.

Sokal, R. R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438.

Springer, M. S., Meredith, R. W., Gatesy, J., Emerling, C. A., Park, J., Rabosky, D. L., Stadler, T., Steiner, C., Ryder, O. A., and Janečka, J. E. (2012). Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PloS one*, 7(11):e49521.

Steel, M. and Penny, D. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular biology and evolution*, 17(6):839–850.

Strait, D. S., Grine, F. E., and Moniz, M. A. (1997). A reappraisal of early hominid phylogeny. *Journal of human evolution*, 32(1):17–82.

Stringer, C. B. (1987). A numerical cladistic analysis for the genus Homo. *Journal of Human Evolution*, 16(1):135–146.

Swofford, D. L. and Berlocher, S. H. (1987). Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Systematic zoology*, 36(3):293–325.

Team, R. C. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Thierry, B. (1985). Patterns of agonistic interactions in three species of macaque (Macaca mulatta, M fascicularis, M tonkeana). *Aggressive Behavior*, 11(3):223–233.

Thierry, B. (2007). Unity in diversity: Lessons from macaque societies. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 16(6):224–238.

Thierry, B., Demaria, C., Preuschoft, S., and Desportes, C. (1989). Structural convergence between silent bared-teeth display and relaxed open-mouth display in the Tonkean macaque (Macaca tonkeana). *Folia primatologica*, 52(3-4):178–184.

Thorpe, R. S. (1984). Coding Morphometric Characters for Constructing Distance Wagner Networks. *Evolution*, 38(2):244–255.

Trinkaus, E. (1987). The Neandertal face: Evolutionary and functional perspectives on a recent hominid face. *Journal of Human Evolution*, 16(5):429–443.

Tuffley, C. and Steel, M. (1997). Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of mathematical biology*, 59(3):581–607.

Turner, C. I., Nichol, C., and Scott, G. (1991). Scoring produces for key morphological traits of the permanent dentition: The Arizona State University dental anthropology system. *Advances in dental anthropology*, pages 13–31.

Varón-González, C., Whelan, S., and Klingenberg, P. (2020). Estimating Phylogenies from Shape and Similar Multidimensional Data: Why It Is Not Reliable. *Systematic Biology*.

Wagner, P. J. (2012). Modelling rate distributions using character compatibility: Implications for morphological evolution among fossil invertebrates. *Biology Letters*, 8(1):143–146.

Waitt, D. E. and Levin, D. A. (1998). Genetic and phenotypic correlations in plants: A botanical test of Cheverud's conjecture. *Heredity*, 80(3):310–319.

Weaver, T. (2018). Neutral theory and the evolution of human physical form: An introduction to models and applications. *J. Anthropol. Sci*, 96:7–26.

Weaver, T. D., Roseman, C. C., and Stringer, C. B. (2007). Were neandertal and modern human cranial differences produced by natural selection or genetic drift? *Journal of human evolution*, 53(2):135–145.

Weaver, T. D. and Stringer, C. B. (2015). Unconstrained cranial evolution in Neandertals and modern humans compared to common chimpanzees. *Proceedings of the Royal Society B: Biological Sciences*, 282(1817):20151519.

Wen, D. and Nakhleh, L. (2018). Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. *Systematic Biology*, 67(3):439–457.

Wen, D., Yu, Y., and Nakhleh, L. (2016). Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLoS Genetics*, 12(5):e1006006.

Wiens, J. J. (2001). Character Analysis in Morphological Phylogenetics: Problems and Solutions. *Systematic Biology*, 50(5):689–699.

Wiens, J. J. and Donoghue, M. J. (2004). Historical biogeography, ecology and species richness. *Trends in ecology & evolution*, 19(12):639–644.

Williams, F. L., Godfrey, L. R., Sutherland, M. R., and Culich, A. (2001). Diagnosing heterochronic perturbations in the craniofacial evolution of Homo (Neandertals and modern humans) and Pan (P-troglodytes and P-paniscus). *AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY*.

Wood, S. (2015). Package 'mgcv'. *R package version*, 1:29.

Worthington, S. (2017). Selection of Character Coding Method Is Not Phylogenetically Neutral: A Test Case Using Hominoids. *Folia Primatologica*, 88(5):385–400.

Wright, A. M. and Hillis, D. M. (2014). Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *PLoS ONE*, 9(10):e109210.

Wright, A. M., Lloyd, G. T., and Hillis, D. M. (2016). Modeling Character Change Heterogeneity in Phylogenetic Analyses of Morphology through the Use of Priors. *Systematic Biology*, 65(4):602–611.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular evolution*, 39(3):306–314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9):367–372.

Young, R. L. and Badyaev, A. V. (2006). Evolutionary persistence of phenotypic integration: Influence of developmental and functional relationships on complex trait evolution. *Evolution*, 60(6):1291–1299.

Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87.

Zhang, Z., Nishimura, A., Bastide, P., Ji, X., Payne, R. P., Goulder, P., Lemey, P., and Suchard, M. A. (2019). Large-scale inference of correlation among mixed-type biological traits with Phylogenetic multivariate probit models. *arXiv preprint arXiv:1912.09185*.