HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2016 April 04.

Published in final edited form as:

Ann Appl Stat. 2015 June ; 9(2): 969–991. doi:10.1214/15-AOAS821.

ASSESSING PHENOTYPIC CORRELATION THROUGH THE MULTIVARIATE PHYLOGENETIC LATENT LIABILITY MODEL

Gabriela B. Cybis[†,**], **Janet S. Sinsheimer**[‡], **Trevor Bedford**[§], **Alison E. Mather**[¶], **Philippe Lemey**[||], and **Marc A. Suchard**[‡]

[†]Federal University of Rio Grande do Sul

[‡]University of California, Los Angeles

[§]Fred Hutchinson Cancer Research Center

[¶]Wellcome Trust Sanger Institute

[||]KU Leuven

Abstract

Understanding which phenotypic traits are consistently correlated throughout evolution is a highly pertinent problem in modern evolutionary biology. Here, we propose a multivariate phylogenetic latent liability model for assessing the correlation between multiple types of data, while simultaneously controlling for their unknown shared evolutionary history informed through molecular sequences. The latent formulation enables us to consider in a single model combinations of continuous traits, discrete binary traits, and discrete traits with multiple ordered and unordered states. Previous approaches have entertained a single data type generally along a fixed history, precluding estimation of correlation between traits and ignoring uncertainty in the history. We implement our model in a Bayesian phylogenetic framework, and discuss inference techniques for hypothesis testing. Finally, we showcase the method through applications to columbine flower morphology, antibiotic resistance in *Salmonella*, and epitope evolution in influenza.

Keywords

Bayesian phylogenetics; Threshold model; Evolution; Genotype-phenotype correlation

Department of Statistics, Federal University of Rio Grande do Sul, Rua Bento Gonçalves 9500, Porto Alegre, RS, 91509-900, Brazil, gabriela.cybis@ufrgs.br
Rega Institute, KU Leuven, Minderbroedersstaat 10, 3000 Leuven, Belgium, philippe.Lemey@rega.kuleuven.be
Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109, tbedford@fhcrc.org
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, United Kingdom, am24@sanger.ac.uk
Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, 695 Charles E. Young Drive, Los Angeles, CA 90095-1766, janet@mednet.ucla.edu, msuchard@ucla.edu
Department of Biomathematics and Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, 695 Charles E. Young Drive, Los Angeles, CA 90095-1766, janet@mednet.ucla.edu, msuchard@ucla.edu
[**]Supported in part by the Fulbright Science & Technology fellowship.

SUPPLEMENTARY MATERIAL

Supplement A: Supplementary tables for applications
(). Point estimates and BCI's for correlation coefficients from section 3.

1. Introduction

Biologists are often interested in assessing phenotypic correlation among sets of traits, since it can help elucidate many biological processes. For example, correlation across the presence or absence of resistance to different antibiotics characterizes the recent evolutionary history of important pathogenic bacteria such as Salmonella. Phenotypic correlation may be a result of genetic constraints, in which traits are partially determined by the same or linked loci. Alternatively, the correlation may be evidence of selective effects, in which the same environmental pressure acts on two seemingly unrelated traits or the outcome of one trait affects selective pressure on the other. Studying these processes is one of the aims of comparative biology.

The purpose of this paper is to present a statistical framework for estimating phenotypic correlation among many traits simultaneously for combinations of different types of data. We consider combinations of continuous data, discrete data with binary outcomes, and discrete data with multiple ordered and unordered outcomes. We also provide inference tools to address specific hypotheses regarding the correlation structure.

Several comparative methods have been proposed to assess the phenotypic correlation between groups of traits (Felsenstein, 1985; Pagel, 1994; Grafen, 1989; Ives and Garland, 2010). These methods estimate correlations in trait data across multiple species while controlling for shared evolutionary history through phylogenetic trees. Yet their use is generally limited to fixed phylogenetic trees, specific types of data or small datasets.

Markov chains are a natural choice to model the evolution of discrete traits, allowing for correlation between them (Pagel, 1994; Lewis, 2001). In this case, the state space of the Markov chain includes all combinations of possible values for all the traits, and correlation is assessed through the transition probabilities between states. Thus, when the number of traits and possible outcomes for each trait increase, the number of parameters to be estimated in the rate matrix scales up rapidly.

For continuous data, a common approach for assessing phenotypic correlation is the independent contrasts method that models the evolution of multiple traits as a multivariate Brownian diffusion process along the tree (Felsenstein, 1985). Correlation between traits is assessed through the precision matrix of the diffusion process. This method has been extended to account for phylogenetic uncertainty by integrating over the space of trees in a Bayesian context (Huelsenbeck and Rannala, 2003). Recent developments increase the method's flexibility by allowing for different diffusion rates along the branches of the tree (Lemey *et al.*, 2010), more efficient likelihood computation, and thus, larger datasets (Pybus *et al.*, 2012).

Phylogenetic linear models and related methods naturally consider combinations of different types of data (Grafen, 1989; Ives and Garland, 2010). Developments in this area have led to flexible and efficient methods (Faria *et al.*, 2013; Ho and Ané, 2014). These models assess the effects of independent variables on a dependent trait that evolves along a tree. Although it is possible that the independent variables are phylogenetically correlated, the evolution of

these variables is not explicitly modeled. Thus, these models are not tailored to assess correlation between sets of traits evolving along the same phylogenetic tree.

An approach for assessing correlated evolution that can combine both binary and continuous data is the phylogenetic threshold model (Felsenstein, 2005, 2012). The threshold model is used in statistical genetics for traits with a discrete outcome determined by an underlying unobserved continuous variable (Wright, 1934; Falconer, 1965). Felsenstein (2005) proposed the use of this model in phylogenetics. In his model, the underlying continuous variable (or latent liability) undergoes Brownian diffusion along the phylogenetic tree. At the tips, a binary trait is defined depending on the position of the latent liability relative to a specified threshold. This non-Markovian model has the desirable property that the probability of transition from the current state to another can depend on time spent in that current state.

A possible interpretation for this model is that the binary outcome represents the presence or absence of some phenotypic trait, and the underlying continuous process represents the combined effect of a large number of genetic factors that affect this trait. During evolution, these factors undergo genetic drift, which is usually modeled as Brownian diffusion.

In its multivariate version, the threshold model allows for inference on the phenotypic correlation structure between a few continuous and binary traits. As with the independent contrasts method, this correlation can be assessed through the covariance matrix of the multivariate Brownian diffusion for the continuous latent liability.

In this paper we build upon the flexibility of the threshold model to create a Bayesian phylogenetic model for the evolution of binary data, discrete data with multiple ordered or unordered states and continuous data. We explore recent developments in models for continuous trait evolution that improve computational efficiency, and make the joint analysis of multiple traits feasible in the presence of possible phylogenetic uncertainty (Lemey *et al.*, 2010; Pybus *et al.*, 2012).

Importantly, our approach estimates the between trait correlation while simultaneously controlling for the correlation induced through the traits being shared by descent. As shown in one of our examples, failing to control for the evolutionary history can confound inference of correlation between traits, in analogy to false inference in association analysis when failing to control for population substructure or relatedness among individuals.

2. Methods

Consider a dataset of N aligned molecular sequences $\mathbf{S}$ from related organisms and an $N \times P$ matrix $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)^t$ of P-dimensional trait observations from each of the N organisms, such that $\mathbf{Y}_i = (y_{i1}, \ldots, y_{iP})$ for $i = 1, \ldots, N$. We model the sequence data $\mathbf{S}$ using standard Bayesian phylogenetics models (Drummond *et al.*, 2012) that include, among other parameters φ less germaine to our development here, an unobserved phylogenetic tree F. This phylogenetic tree is a bifurcating, directed graph with N terminal nodes $(\nu_1, \ldots, \nu_N)$ of degree 1 that correspond to the tips of the tree, $N - 2$ internal nodes $(\nu_{N+1}, \ldots, \nu_{2N-2})$ of degree 3, a root node ν_{2N-1} of degree 2 and edge weights $(t_1, \ldots, t_{2N-2})$ between nodes that

track elasped evolutionary time. Conditional on F, we assume independence between **S** and **Y**, and refer interested readers to, for example, Suchard, Weiss and Sinsheimer (2001) and Drummond *et al.* (2012) for detailed development of $p(\mathbf{S}, \varphi, F)$.

The dimensions of $\mathbf{Y}_i$ contain trait observations that may be binary, discrete with multiple states, continuous or a mixture thereof. Importantly, to handle the myriad of different data types, we assume that the observation of **Y** is governed by an underlying unobserved continuous random variable $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^t$, called a latent liability, where each row $\mathbf{X}_i = (x_{i1}, \ldots, x_{iD}) \in \mathbb{R}^D$ with $D \geq P$ depending on the mixture of data types. We assume that **X** arise from a multivariate Brownian diffusion along the tree F (Lemey *et al.*, 2010) for which we provide a more indepth description shortly. At the tips of F, the realized values of **Y** emerge deterministically from the latent liabilities **X** through the mapping function $g(\mathbf{X})$.

2.1. Latent Liability Mappings

When column j of **Y** is composed of binary data, these values map from a single dimension j' in **X** following a probit-like formulation in which the outcome is one if the underlying continuous value is larger than a threshold and zero otherwise. Without loss of generality, we take the threshold to be zero, such that

$$y_{ij} = g(x_{ij'}) = \begin{cases} 0 & \text{if } x_{ij'} \leq 0 \\ 1 & \text{if } x_{ij'} > 0. \end{cases} \quad (1)$$

Alternatively, if column j of **Y** assumes K possible discrete states $(s_1, \ldots, s_K)$, and they are ordered so that transitions from state s_k to s_{k+2} must necessarily pass through s_{k+1}, we use a multiple threshold mapping (Wright, 1934). Again, column j of **Y** maps from a single dimension j' in the latent liabilities **X**; however, the position of $x_{ij'}$ relative to the multiple thresholds $(a_1, \ldots, a_{K-1})$ determines the value of y_{ij} through the function

$$y_{ij} = g(x_{ij'}) = \begin{cases} s_1 & \text{if } x_{ij'} < a_1 \\ s_k & \text{if } a_{k-1} \leq x_{ij'} < a_k \quad \text{for } k = 2, \ldots, K-1 \\ s_K & \text{if } x_{ij'} \geq a_{K-1}, \end{cases} \quad (2)$$

where $a_2, \ldots, a_{K-1}$ in increasing values are generally estimable from the data if we set $a_1 = 0$ for identifiability. Let $\mathbf{A} = \{a_k\}$ track all of the non-fixed threshold parameters for all ordered traits.

When column j of **Y** realizes values in K multiple states, but there is no ordering between them, we adopt a multinomial probit model. Here the observed trait maps from $K-1$ dimensions in the latent liabilities **X**, and the value of y_{ij} is determined by the largest component of these latent variables,

$$y_{ij} = g(x_{ij'}, \ldots, x_{i,j'+K-2}) = \begin{cases} s_1 & \text{if } \quad 0 = \sup(0, x_{ij}, \ldots, x_{i,j+K-2}) \\ s_{k+1} & \text{if } \quad x_{ik} = \sup(0, x_{ij}, \ldots, x_{i,j+K-2}), \end{cases} \quad (3)$$

where, without loss of generality, the first state s_1 is the reference state.

Finally, if column j of $\mathbf{Y}$ contains continuous values, a simple monotonic transform from $\mathbb{R}$ suffices. For example, for normally distributed outcomes, $y_{ij} = g(x_{ij}) = x_{ij}$.

2.2. Trait Evolution

A multivariate Brownian diffusion process along the tree F (Lemey *et al.*, 2010) gives rise to the elements of $\mathbf{X}$. This process posits that the latent trait value of a child node v_k in F is multivariate normally distributed about the unobserved trait value of its parent node $v_{\text{pa}(k)}$ with variance $t_k \times \mathbf{\Sigma}$. In this manner, the unknown $D \times D$ matrix $\mathbf{\Sigma}$ characterizes the between-trait correlation and the tree F controls for trait values being shared by descent.

Assuming that the latent trait value at the root node v_{2N-1} draws *a priori* from a multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and variance $\tau_0 \times \mathbf{\Sigma}$ and integrating out the internal and root node trait values (Pybus *et al.*, 2012), we recall that the latent liabilities $\mathbf{X}$ at the tips of F are matrix normally distributed, with probability density function

$$p(\mathbf{X}|\mathbf{V}(F), \sum, \boldsymbol{\mu}_0, \tau_0) = \frac{\exp\left\{-\frac{1}{2}\text{tr}\left[\sum^{-1}(\mathbf{X}-\boldsymbol{\mu}_0)^t(\mathbf{V}(F)+\tau_0\mathbf{J})^{-1}(\mathbf{X}-\boldsymbol{\mu}_0)\right]\right\}}{(2\pi)^{NP/2}|\sum|^{N/2}|\mathbf{V}(F)+\tau_0\mathbf{J}|^{P/2}}, \quad (4)$$

where $\mathbf{J}$ is an $N \times N$ matrix of all ones and $\mathbf{V}(F) = \{v_{ij}\}$ is an $N \times N$ matrix that is a deterministic function of F. Let $d_F(u, w)$ equal the sum of edge weights along the shortest path between node u and node w in F. Then diagonal elements $v_{ii} = d_F(v_{2N-1}, v_i)$, the time-distance between the root node and tip node i, and off-diagonal elements $v_{ii'} = [d_F(v_{2N-1}, v_i) + d_F(v_{2N-1}, v_{i'}) - d_F(v_i, v_{i'})]/2$, the time-distance between the root and the most recent common ancestor of tip nodes i and i'.

We consider the augmented likelihood for the trait data $\mathbf{Y}$ and latent liabilities $\mathbf{X}$ and highlight a convenient factorization

$$p(\boldsymbol{Y}, \mathbf{X}|\mathbf{V}(F), \sum, \mathbf{A}, \boldsymbol{\mu}_0, \tau_0) = p(\boldsymbol{Y}|\mathbf{X}, \mathbf{A}) \times p(\mathbf{X}|\mathbf{V}(F), \sum, \boldsymbol{\mu}_0, \tau_0). \quad (5)$$

The conditional likelihood $p(\mathbf{Y} \mid \mathbf{X}, \mathbf{A}) = \mathbf{1}_{(\mathbf{Y}=g(\mathbf{X}))}$ in factorization (5) is simply the indicator function that $\mathbf{X}$ are consistent with the observations $\mathbf{Y}$. Consequentially, the augmented likelihood is a truncated, matrix normal distribution.

Figure 1 illustrates schematic representations of the latent liability model for all four types of data. In the figure, we include trees with $N = 4$ to 6 taxa, annotated with their observed traits $\mathbf{Y}$ at the tree tips and plot potential realizations of the latent liabilities $\mathbf{X}$ values along these trees that give rise to $\mathbf{Y}$.

We complete our model specification by assuming *a priori*

$$\sum^{-1} \sim \text{Wishart}(d_0, \mathbf{T}), \quad (6)$$

with degrees of freedom d_0 and rate matrix $\mathbf{T}$. For the non-fixed threshold parameters $\mathbf{A}$, we assume differences $a_k - a_{k-1}$ for each trait are *a priori* independent and Exponential(α) distributed, where α is a rate constant. Finally, we specify fixed hyperparameters ($\boldsymbol{\mu}_0$, τ_0, d_0, $\mathbf{T}$, α) in each of our examples.

2.3. Inference

We aim to learn about the posterior distribution

$$\begin{aligned} p(\textstyle\sum, F, \boldsymbol{\phi}, \mathbf{A} | \boldsymbol{Y}, \mathbf{S}) &\propto p(\boldsymbol{Y} | \textstyle\sum, F, \mathbf{A}) \times p(\textstyle\sum) \times p(\mathbf{A}) \times p(\mathbf{S}, \boldsymbol{\phi}, F) \\ &= (\textstyle\int p(\boldsymbol{Y}, \mathbf{X} | \sum, F, \mathbf{A}) \mathrm{d}\mathbf{X}) \times p(\sum) \times p(\mathbf{A}) \times p(\mathbf{S}, \boldsymbol{\phi}, F). \end{aligned} \quad (7)$$

We accomplish this task through Markov chain Monte Carlo (MCMC) and the development of computationally efficient transitions kernels to faciliate sampling of the latent liabilities $\mathbf{X}$. We exploit a random-scan Metropolis-with-Gibbs scheme. For the tree F and other phylogenetic parameters $\boldsymbol{\varphi}$ involving the sequence evolution, we employ standard Bayesian phylogenetic algorithms (Drummond *et al.*, 2012) based on Metropolis-Hastings parameter proposals. Further, the full conditional distribution of $\boldsymbol{\Sigma}^{-1}$ remains Wishart (Lemey *et al.*, 2010), enabling Gibbs sampling.

MCMC transition kernels for sampling $\mathbf{X}$ are more problematic; tied into this difficulty also lies computationlly efficient evaluation of Equation (4). Strikingly, the solution to the latter problem points to new directions in which to attack the sampling problem. As written, computing $p(\mathbf{X} \mid \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ to evaluate a Metropolis-Hasting acceptance ratio appears to require the high computational cost of $\mathcal{O}(N^3)$ involved in forming $(\mathbf{V}(F) + \tau_0 \mathbf{J})^{-1}$. Such a cost would be prohibitive for large N when F is random, necessiating repeated inversion. This is one reason why previous work has limited itself to fixed, known F. However, we follow Pybus *et al.* (2012), who develop a dynamic programming algorithm to evaluate density (4) in $\mathcal{O}(N)$ that avoids matrix inversion. Critically, we extend these algorithmic ideas in this paper to construct computationally efficient sampling procedures for $\mathbf{X}$.

Pybus *et al.* (2012) propose a post-order tree traversal that visits each node u in F, starting at the tips and ending at the root. For the example tree in Figure 2, one possible post-order traversal proceeds through nodes $\{1 \to 2 \to 4 \to 3 \to 5\}$. Let $\mathbf{X}_u$ for $u = N+1, \ldots, 2N-1$ imply now hypothesized latent liabilities at the internal and root nodes of F. Then, at each visit, one computes the conditional density of the tip latent liabilities $\{\mathbf{X}\}_u^{\text{post}}$ that are descendent to node u given $\mathbf{X}_{\text{pa}(u)}$ at the parent node of u by integrating out the hypothesized value $\mathbf{X}_u$ at node u. For example, when visiting node $u = 4$ in Figure 2, one considers the conditional density of $(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{X}_5$. Each of these conditional densities are proportional to a multivariate normal density, so during the traversal it suffices to keep track of the partial mean vector $\mathbf{m}_u^{\text{post}}$, partial precision scalar p_u^{post} and remainder term ρ_u that characterize the conditional density. We refer interested readers to the Supplementary Material in Pybus *et al.* (2012) for further details.

Building upon this algorithm, we identify that it is possible and practical to generate samples from $p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ for tip v_i without having to manipulate $\mathbf{V}(F)$ via one additional pre-order traversal of F. This approach enables us to exploit $p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ as a proposal distribution in an efficient Metropolis-Hastings scheme to sample $\mathbf{X}_i$, since the distribution often closely approximates the full conditional distribution of $\mathbf{X}_i$.

To ease notation in the remainder of this section, we drop explicit dependence on $\mathbf{V}(F)$, $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_0$, τ_0 in our distributional arguments. Further, let $\{\mathbf{X}\}_u^{\text{pre}}$ collect the latent liabilities at the tree tips that are not descendent to node u for $u = 1, \ldots, 2N-1$, such that $\{\mathbf{X}\}_u^{\text{pre}} \cup \{\mathbf{X}\}_u^{\text{post}} = \mathbf{X}$ and $\{\mathbf{X}\}_u^{\text{pre}} \cap \{\mathbf{X}\}_u^{\text{post}} = \varnothing$. Notably, $\{\mathbf{X}\}_i^{\text{pre}} = \mathbf{X}_{(-i)}$ and $\{\mathbf{X}\}_{2N-1}^{\text{pre}} = \varnothing$. With these goals and definitions in hand, we find $p(\mathbf{X}_i \mid \mathbf{X}_{(-i)})$ recursively.

Consider a triplet of nodes in F such that node u has parent $\text{pa}(u) = w$ that it shares with sibling $\text{sib}(u) = v$. For example, in Figure 2, $u = 1$, $v = 2$ and $w = 4$ is one of two choices. Because of the conditional independence structure of the multivariate Brownian diffusion process on F, we can write

$$p(\mathbf{X}_u \mid \{\mathbf{X}\}_u^{\text{pre}}) = \int p(\mathbf{X}_u \mid \mathbf{X}_{\text{pa}(u)}) p(\mathbf{X}_{\text{pa}(u)} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}}, \{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}}) \, \mathrm{d}\mathbf{X}_{\text{pa}(u)}, \quad (8)$$

where Equation (8) returns the desired quantity when $i = u$ and the first term of the integrand is a multivariate normal density $\text{MVN}(\mathbf{X}_u; \mathbf{X}_{\text{pa}(u)}, (t_u \boldsymbol{\Sigma})^{-1})$ centered at $\mathbf{X}_{\text{pa}(u)}$ with precision $(t_u \boldsymbol{\Sigma})^{-1}$. The second term requires more exploration

$$\begin{aligned} p(\mathbf{X}_{\text{pa}(u)} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}}, \{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}}) &= \frac{p(\mathbf{X}_{\text{pa}(u)}, \{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}})}{p(\{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}})} \\ &\propto p(\{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}} \mid \mathbf{X}_{\text{pa}(u)}) p(\mathbf{X}_{\text{pa}(u)} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}}), \end{aligned} \quad (9)$$

where the normalization constant does not depend on $\mathbf{X}_{\text{pa}(u)}$ and we fortuitously have determined that the probability $p(\{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}} \mid \mathbf{X}_{\text{pa}(u)})$ is proportional to $\text{MVN}\left(\mathbf{X}_{\text{pa}(u)}; \mathbf{m}_{\text{sib}(u)}^{\text{post}}, p_{\text{sib}(u)}^{\text{post}} \sum^{-1}\right)$ during the post-order traversal.

Substituting Equation (9) in Equation (8) furnishes a set of recursive integrals down the tree

$$p(\mathbf{X}_u \mid \{\mathbf{X}\}_u^{\text{pre}}) \propto \int p(\mathbf{X}_u \mid \mathbf{X}_{\text{pa}(u)}) p(\{\mathbf{X}\}_{\text{sib}(u)}^{\text{post}} \mid \mathbf{X}_{\text{pa}(u)}) p(\mathbf{X}_{\text{pa}(u)} \mid \{\mathbf{X}\}_{\text{pa}(u)}^{\text{pre}}) \, \mathrm{d}\mathbf{X}_{\text{pa}(u)}. \quad (10)$$

To solve the set of integrals in (10), we recall that $p(\mathbf{X}_{2N-1} \mid \{\mathbf{X}\}_{2N-1}^{\text{pre}}) = p(\mathbf{X}_{2N-1})$ is MVN $(\mathbf{X}_{2N-1}; \boldsymbol{\mu}_0, (\tau_0 \boldsymbol{\Sigma})^{-1})$ and so define pre-order, partial mean vector $\mathbf{m}_{2N-1}^{\text{pre}} = \boldsymbol{\mu}_0$ and partial precision scalar $p_{2N-1}^{\text{pre}} = 1/\tau_0$. Since the convolution of multivariate normal random variables remains multivariate normal, we identify that $p(\mathbf{X}_u \mid \{\mathbf{X}\}_u^{\text{pre}})$ is $\text{MVN}\left(\mathbf{X}_u; \mathbf{m}_u^{\text{pre}}, p_u^{\text{pre}} \sum^{-1}\right)$ where pre-order, partial mean vectors and precision scalars unwind through

$$\begin{aligned} \mathbf{m}_u^{\text{pre}} &= \frac{p_{\text{sib}(u)}^{\text{post}}\mathbf{m}_{\text{sib}(u)}^{\text{post}} + p_{\text{pa}(u)}^{\text{pre}}\mathbf{m}_{\text{pa}(u)}^{\text{pre}}}{\mathbf{m}_{\text{sib}(u)}^{\text{post}} + \mathbf{m}_{\text{pa}(u)}^{\text{pre}}}, \text{ and} \\ \frac{1}{p_u^{\text{pre}}} &= t_u + \frac{1}{p_{\text{sib}(u)}^{\text{post}} + p_{\text{pa}(u)}^{\text{pre}}}, \end{aligned} \quad (11)$$

until we hit tip node i.

With a simple algorithm to compute the mean and precision of the full conditional $p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ at our disposal, we finally turn our attention toward a Metropolis-Hastings scheme to sample $\mathbf{X}_i$. The algorithm needs to generate samples only for the latent liabilities $\mathbf{X}_{i(-c)}$ corresponding to the discrete traits, since the map function $g(\cdot)$ fixes the latent liabilities $\mathbf{X}_{ic}$ for all the continuous traits. Thus we consider the proposal distribution $p(\mathbf{X}_{i(-c)} \mid \mathbf{X}_{ic}, \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$, which is obtained from $p(\mathbf{X}_i \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \tau_0)$ by further conditioning on the fixed liabilities $\mathbf{X}_{ic}$. This conditional distribution is $\text{MVN}(\mathbf{X}_{ic}; \mathbf{m}_i^{\text{cond}}, p_i^{\text{pre}}\mathbf{W}_{cc})$, where

$$\mathbf{m}_i^{\text{cond}} = \mathbf{m}_{i(-c)}^{\text{pre}} - \mathbf{W}_{cc}^{-1}\mathbf{W}_{c(-c)}\left(\mathbf{X}_{i(-c)} - \mathbf{m}_{i(-c)}^{\text{pre}}\right). \quad (12)$$

Here the vector $\mathbf{m}_{i(-c)}^{\text{pre}} = (\mathbf{m}_{i(-c)}^{\text{pre}}, \mathbf{m}_{ic}^{\text{pre}})$ is partitioned according to correspondence to continuous traits, as is the precision matrix for the diffusion process

$$\sum^{-1} = \begin{pmatrix} \mathbf{W}_{(-c)(-c)} & \mathbf{W}_{(-c)c} \\ \mathbf{W}_{c(-c)} & \mathbf{W}_{cc} \end{pmatrix}. \quad (13)$$

Several approaches compete for generating truncated multivariate normal random variables, including rejection sampling (Breslaw, 1994; Robert, 1995) and Gibbs sampling (Gelfand, Smith and Lee, 1992; Robert, 1995) possibly with data augmentation (Damien and Walker, 2001). For the examples we explore in this manuscript, the dimension D of $\mathbf{X}_i$ can be large, ranging up to 54 with $N = 360$ tips, with occasionally high correlation in $\boldsymbol{\Sigma}$. Gibbs sampling can suffer from slow convergence in the presence of high correlation between dimensions. Consequentially, we explore an extension of rejection sampling that involves a multiple-try Metropolis (Liu, Liang and Wong, 2000) construction. We simulate up to R draws $\mathbf{X}_i^{(r)} \sim p(\cdot \mid \mathbf{X}_{(-i)}, \mathbf{V}(F), \sum, \boldsymbol{\mu}_0, \tau_0)$. For draw $\mathbf{X}_i^{(r)}$, if $p(\mathbf{X}_i^{(r)} \mid \mathbf{Y}_i, \mathbf{A}) \neq 0$, then we accept this value as our next realization of $\mathbf{X}_i$. The Metropolis-Hastings acceptance probability of this action is 1. If all R proposals return 0 density, the MCMC chain remains at its current location.

In our largest example, we evaluate one approach to select R. We start with a very large $R = 10000$ and observe that most proposals that lead to state changes occur in the first 20 attempts; after 100 attempts, the residual probability of generating a valid sample becomes negligible. Thus, we set $R = 100$ for future MCMC simulations. As MCMC chains converge towards the posterior distribution, the probably of generating a valid sample approaches the 75 – 85% range in our examples. Finally, we employ a Metropolis-Hastings scheme to

sample **A** in which the proposal distribution is a uniform window centered at the parameter's current value with a tunable length.

2.4. Correlation Testing and Model Selection

To assess the phenotypic relationship between two specific components of the trait vector **Y**, we look at the correlation of the corresponding elements in the latent variable **X**. One straightforward approach entertains the use of the marginal posterior distribution of pair-wise correlation coefficients $\rho_{jj'}$ determined from $\boldsymbol{\Sigma}$. As a simple rule-of-thumb, we designate $\rho_{jj'}$ significantly non-zero if > 99% of its posterior mass falls strictly greater than or strictly less than 0.

When scientific interest lies in formal comparison of models that involve more than pair-wise effects, we employ Bayes factors. Possible examples include identifying block-diagonal structures in $\boldsymbol{\Sigma}$, comparing the latent liability model to other trait evolution models and, as demonstrated in our examples, state-ordering of multiple discrete traits.

The Bayes factor that compares models M_0 and M_1 can be obtained as

$$B_{01}=\frac{p(\boldsymbol{Y},\mathbf{S}|M_0)}{p(\boldsymbol{Y},\mathbf{S}|M_1)}, \quad (14)$$

in which $p(\mathbf{Y}, \mathbf{S}|M)$ is the marginal likelihood of the data under model M (Jeffreys, 1935). Computing these marginal likelihoods is not straightforward, involving high dimensional integration. We adopt a path sampling approach which estimates these integrals through numerical integration.

To estimate the marginal likelihoods in (14), we follow Baele *et al.* (2012) in considering a geometric path $q_\beta(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$ that goes from a normalized source distribution $q_0(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \boldsymbol{\theta})$ to the unnormalized posterior distribution $p(\mathbf{Y}, \mathbf{S}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X}, \boldsymbol{\theta})$. Here both distributions are defined on the same parameter space, and $\boldsymbol{\theta} = \{\boldsymbol{\Sigma}, F, \boldsymbol{\varphi}, \mathbf{A}\}$ collects all model parameters. The path sampling algorithm employs MCMC to numerically compute the path integral

$$\log(p(\boldsymbol{Y},\mathbf{S}|M))=\int_0^1 E_{q_\beta}\left[\log(q_1(\boldsymbol{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta}))-\log(q_0(\boldsymbol{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta}))\right]\mathrm{d}\beta. \quad (15)$$

A natural choice for the source distribution is the prior $p(\mathbf{X}, \boldsymbol{\theta})$. However, due to truncations in the distribution of **X** induced by the map function $g(\cdot)$, the path from the prior to the unnormalized posterior is not continuous. Since continuity along the whole path is required for (15) to hold, we propose here a different destination distribution that leads to a continuous path. Let

$$q_0(\boldsymbol{Y},\mathbf{S};\mathbf{X},\boldsymbol{\theta})=p(\mathbf{X}|\boldsymbol{Y},\mathbf{A})\psi(\mathbf{X})p(\boldsymbol{\theta}), \quad (16)$$

where $p(\boldsymbol{\theta})$ is the prior, $p(\mathbf{X} \mid \mathbf{Y}, \mathbf{A}) = \mathbf{1}_{(\mathbf{Y}=g(\mathbf{X}))}$, and $\psi(\mathbf{X})$ is a function proportional to a conveniently chosen matrix normal distribution. The proportionality constant of $\psi(\mathbf{X})$ is selected to guarantee

$$\int p(\mathbf{X}|\boldsymbol{Y},\mathbf{A})\psi(\mathbf{X})\mathrm{d}\mathbf{X}=1, \quad (17)$$

and thus a normalized source distribution $q_0(\mathbf{Y}, \mathbf{S}; \mathbf{X}, \theta)$.

The choice of function $\psi(\mathbf{X})= \psi^*(\mathbf{X})/Q(\mathbf{Y}, \mathbf{A})$ is central to the success of this path sampling approach. We select the matrix normal distribution $\psi^*(\mathbf{X})$ so that all entries in $\mathbf{X}$ are independent, and consequently the proportionality constant is

$$Q(\boldsymbol{Y},\mathbf{A})=\prod_{i=1}^{N}\prod_{j=0}^{P}Q(y_{ij},\mathbf{A})=\prod_{i=1}^{N}\prod_{j=0}^{P}\int p(\mathbf{X}_{ij^*}|y_{ij},\mathbf{A})\psi^*(\mathbf{X}_{ij^*})\mathrm{d}\mathbf{X}_{ij^*}, \quad (18)$$

where $\mathbf{X}_{ij}^*$ are all the entries of the latent liability corresponding to y_{ij}.

For binary traits, $\mathbf{X}_{ij}^*$ is univariate, and $\psi(\mathbf{X}_{ij}^*)$ is proportional to a normal distribution whose mean $\bar{X}_{ij}^*$ and variance $\bar{\sigma}_{ij^*}^2$ match those of the posterior distribution of $\mathbf{X}_{ij}^*$. Considering that the map function $g(\cdot)$ restricts $\mathbf{X}_{ij}^*$ to be larger (or smaller) than 0, and that $\bar{X}_{ij}^*$ always belongs to this valid region, the proportionality constant for a binary trait is

$$Q(\boldsymbol{Y}_{ij},\mathbf{A})=\Phi\left(\frac{|\bar{X}_{ij^*}|}{\bar{\sigma}_{ij^*}}\right), \quad (19)$$

where $\mathbf{\Phi}(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

For traits with $K \geq 3$ ordered states, $\mathbf{X}_{ij}^*$ is also univariate, and we make the same choice for mean and variance parameters of $\psi^*(\mathbf{X}_{ij}^*)$. The map function depends on the threshold parameters $\mathbf{A}$, that must be fixed for this analysis. If $a_l(y_{ij})$ and $a_u(y_{ij})$ denote respectively the lower and upper threshold for the valid region mapped from y_{ij}, then the proportionality constant becomes

$$Q(y_{ij},\mathbf{A})=\Phi\left(\frac{a_u(y_{ij})-\bar{X}_{ij^*}}{\bar{\sigma}_{ij^*}}\right)-\Phi\left(\frac{a_l(y_{ij})-\bar{X}_{ij^*}}{\bar{\sigma}_{ij^*}}\right). \quad (20)$$

When y_{ij} assumes one of the extreme states s_1 and s_K, then the normalizing constant considers the appropriate open interval.

For discrete data with $K \geq 3$ unordered states, y_{ij} maps from $K-1$ dimensions in $\mathbf{Y}$. For simplicity, $\psi^*(\mathbf{X}_{ij}^*)$ is a standard multivariate normal distribution, and the proportionality constant is

$$Q(y_{ij},\mathbf{A})=\begin{cases} 2^{-(K-1)} & \text{if} \quad y_{ij}=s_1 \\ \frac{1-2^{-(K-1)}}{K-1} & \text{if} \quad y_{ij}=s_2,\cdots,s_K. \end{cases} \quad (21)$$

Finally, for continuous y_{ij} we simply have $\psi(\mathbf{X}_{ij}^*) = y_{ij}$.

Implementation—The methods described in this paper have been implemented in the software package BEAST (Drummond *et al.*, 2012).

3. Applications

We present applications of our model to three problems in which researchers wish to assess correlation between different types of traits while controlling for their shared evolutionary history.

3.1. Antimicrobial resistance in Salmonella

Development of multidrug resistance in pathogenic bacteria is a serious public health burden. Understanding the relationships between resistance to different drugs throughout bacterial evolution can help shed light on the fundamentals of multidrug resistance on the epidemiological scale.

We use the phylogenetic latent liability model to assess phenotypic correlation between resistance traits to 13 different antibiotics in *Salmonella*. We analyse 248 isolates of *Salmonella* Typhimurium DT104, obtained from animals and humans in Scotland between 1990 and 2011 (Mather *et al.*, 2013). For each isolate, we have sequence data and binary phenotypic data indicating the strains resistance status to each of the 13 antibiotics.

To assess which resistance traits are associated we examine the correlation matrix of the latent liabilities $\mathbf{X}$. Because the trait data are binary, the underlying latent variables $\mathbf{X}_i$ for this problem are $D = 13$-dimensional, with each entry corresponding to resistance to one antibiotic. To highlight the main correlation structure of $\boldsymbol{\Sigma}$, Figure 3 presents a heatmap of the significantly non-zero pair-wise correlation coefficients. This matrix contains only positive correlations, consistent with genetic linkage between resistance traits. Additionally, the significant correlations form a block-like structure. Table S1 presents posterior mean and 95% Bayesian credible interval (BCI) estimates for all correlations between resistance traits. Estimates of non-significant correlations tend to be slightly positive, with the exception of correlations involving resistance to ciprofloxacin.

Our analysis reveals a block of strong positive correlations between resistance traits to the antibiotics tetracycline, ampicillin, chloramphenicol, spectinomycin, streptomycin and sulfamethoxazole (sulfonamide), similar to those found using a simpler model (Mather *et al.*, 2012). We estimate a posterior probability > 0.9999 for positive correlation between all these resistance traits simultaneously. This block is consistent with the *Salmonella* genomic island 1 (SGI-1), a 43-kb genomic island conferring multidrug resistance. Among the drugs considered here, SGI-1 confers resistance to these 6 antibiotics (Boyd *et al.*, 2001).

Another pair of antibiotic resistance traits that we infer to be strongly correlated are gentamicin and netilmicin, with a 95% BCI of [0.80, 0.98]. These drugs are both aminoglycoside antibiotics, and the same genes may confer resistance to both antibiotics. These drugs also appear correlated with some of the resistance traits connected to SGI-1.

Although previous analysis of this dataset has revealed that most of the evolutionary history that these data capture was spent in human hosts, human-to-animal or animal-to-human

transitions do occur across the tree (Mather *et al.*, 2013). We investigate whether these interspecies transitions also correlate with antibiotic resistance. To do so, we include host species (animal/human) as a 14th binary trait under in latent liability model. None of the pair-wise correlations are significantly non-zero given our rule-of-thumb definition. Table S2 contains estimated correlations to the host trait.

3.2. Columbine flower evolution

The flowers of columbine genus *Aquilegia* have attracted several different pollinators throughout their evolutionary history. One question that remains is the exact role the pollinators play in the tempo of columbine flower evolution (Whittall and Hodges, 2007). Since different pollinator species demonstrate distinct preferences for flower morphology and color, we investigate here how these traits correlate over the evolutionary history of *Aquilegia*.

We analyse $P = 12$ different floral traits for $N = 30$ monophyletic populations from the genus *Aquilegia*. Of these traits, 10 are continuous and represent color, length and orientation of different anatomical features of the flowers. Additionally, we consider a binary trait that indicates presence or absence of anthocyanin pigment, and another discrete trait that indicates the primary pollinator for that population. As the prevailing hypothesis is that evolutionary transitions from bumblebee-pollinated flowers (Bb) to those primarily pollinated by hawkmoths (Hm) are obligated to pass through an intermediate stage of hummingbird-pollination (Hb) (Whittall and Hodges, 2007), we treat pollinators as ordered states, but we formally test alternative orderings. Taken together, this results in a latent liability model with $D = 12$ dimensions. As sequence data are not readily available for all the taxa included in this analysis, we consider for our analysis the same fixed phylogenetic tree used in Whittall and Hodges (2007). The ability to either condition on a fixed phylogeny F or integrate over a random F in a single framework presents a strength in a field that has traditionally focused on either genetic or phenotypic data alone and joint datasets are an emerging addition. Whittall *et al.* (2006) and Whittall and Hodges (2007) have published the original data, that are available at (http://bodegaphylo.wikispot.org).

To draw inference on the phenotypic correlation structure of these traits, we focus on the 12×12 variance matrix $\boldsymbol{\Sigma}$ of the Brownian motion process that governs the evolution of $\mathbf{X}$ on the tree. We report posterior mean and BCI estimates for all pair-wise correlations in $\boldsymbol{\Sigma}$ in Table S3. Figure 4 presents a heatmap of the posterior means of the correlations. Our analysis reveals a strong block correlation structure between the floral traits. We find one block of positive correlation between chroma of both spur and blade and the presence of anthocyanins. All other color and morphological traits in the analysis form a second block of positive correlation. Additionally, phenotypic correlation between the first and second trait blocks are all negative.

Whittall and Hodges (2007) highlight the relationship between changes in pollinators and increases in floral spur length. They argue that flowers with long spurs are only pollinated by animals with the long tongues required to access and feed on the nectar contained at the end of the spur. We estimate a positive correlation between pollinators and spur length, with a posterior mean of 0.76, and a 95% BCI of [0.60; 0.88], consistent with their findings.

The pollinator trait has $K = 3$ ordered states and, under the latent liability model, its outcome is determined by the relative position of one dimension in **X** to threshold parameters $a_1 = 0$ and a_2. Consequently, our estimate of a_2 is instrumental in determining the relative probabilities of the states in our model and the inferred trait state at the root of the tree. We estimate a_2 to have a posterior mean of 3.00 with a 95% BCI of [1.14; 5.34].

The bumblebee ↔ hummingbird ↔ hawkmoth (Bb-Hb-Hm) ordering is only one of several, and alternative hypotheses regarding pollinator adaptation have been proposed (van der Niet and Johnson, 2012). We examine whether the data support this ordering, or if there is another model with a better fit. We use Bayes factors to compare four different models for the pollinator trait: the Bb-Hb-Hm, Hb-Hm-Bb, Hm-Bb-Hb, and an unordered formulation. Note that there are only three possible orderings for a K =3 state ordered latent liability model since, for symmetric models such as Bb-Hb-Hm and Hm-Hb-Bb, inverting the order leads to equivalent models with inverted signs for the latent traits. The unordered model leads to a bivariate contribution to latent liability **X**. Table 1 presents the path sampling estimates for the marginal likelihood of each model and the corresponding Bayes factors. These comparisons indicate that, in agreement with Whittall and Hodges (2007), the data strongly support the Bb-Hb-Hm model.

Our latent liability model estimates correlation between traits while accounting for shared evolutionary history. To evaluate the effect that phylogenetic relatedness has on our estimates, we estimated the same correlation under a latent liability model with no phylogenetic structure. In this analysis, a star tree with identical distance between all taxa was used. Table S4 presents these correlation estimates and the corresponding 95% BCI. Comparing these results to the original latent liability analysis that accounts for shared evolutionary history, we noticed that most estimates were consistent between both analyses, with a mean absolute difference for posterior means of correlation of 0.11. However, for three of the pairwise correlations (anthocianins × orientation, orientation × blade length, spur length × spur hue) the BIC's for the model that does not account for shared evolution did not contain the posterior mean for the evolutionary model. In particular, the evolutionary model estimates a significantly weaker correlation between orientation and anthocianins (posterior mean of −0.45) than does the model that does not account for shared history, with a 95% BCI of [−0.78; −0.46].

3.3. Correlation within and across influenza epitopes

In influenza, the viral surface proteins hemagglutinin (HA) and neuraminidase provide the antigenic epitopes to which the host immune system responds. Rapid mutation of these proteins to evade immune response, known as antigenic drift, severely challenges the design of annual influenza vaccines. The epitope regions in these genes are particularly important to the drift process (Fitch *et al.*, 1991; Plotkin and Dushoff, 2003). In this context, we are interested in studying the phenotypic correlation among the amino acid sites of these epitopes, because the identification of correlated amino acids grants insight into the dynamics of antigenic drift in influenza.

The HA protein has five identified epitopes A–E, each containing around 20 amino acids. We focus on epitopes A and B, because these are the most immunologically stimulating for

most influenza strains (Bush *et al.*, 1999; Cox and Bender, 1995). We analyse sequence data for 180 strains of human H3N2 influenza dating from 1995 to 2012, obtained from the Influenza Research Database (http://www.fludb.org) and selected to promote geographic diversity. We use the amino acid information in epitope A and B for the latent liability part of the model, and the remaining sequence data in a standard phylogenetic approach to inform the tree structure.

Of the 40 amino acid sites in epitopes A and B of the HA protein, we find 17 to be variable in our sample. The number of unique amino acids in these sites varies between $K = 2$ and $K = 5$. Through a preliminary survey of a larger sample of influenza strains (900 samples) from the same period we find that all polymorphic sites for which the major allele frequency is < 99% are also variable in our 180 sequence sample, strongly suggesting that our limited dataset contains information about all the common variant sites in epitopes A and B during this period.

We model these data with the latent liability model for multiple unordered states. For each amino acid site, we have $K - 1$ corresponding latent traits, yielding a total of $D = 32$ latent dimensions in $\mathbf{X}$. Without loss of generality, we take the amino acid observed in the oldest sequence of the sample as the reference state, and each entry of the latent liability column corresponds to one of the other amino acid variants for that site.

To assess the phenotypic correlation structure between sites in epitopes A and B, we estimate the correlation matrix associated with $\boldsymbol{\Sigma}$ of the latent liability $\mathbf{X}$. Figure 5 presents pairwise correlations for the significantly non-zero estimates. The arrangement of states follows the order of sites in the primary amino acid sequence, even though the sites are not necessarily contiguous in folded protein-space.

Our analysis suggests a group of 11 sites that are strongly correlated with each other. These sites have significant positive correlations to at least three other sites in the group. The group includes all the sites identified by Koel *et al.* (2013) as being the major determinants of antigenic drift that are polymorphic in our sample. We do not find preferential correlations within epitopes.

Table S5 presents a list with point estimates and 95% BCI of correlations whose credible intervals do not include zero. All correlations in this list are positive and point estimates range from 0.6 to 0.74. Since, for all sites the oldest variant was taken as the reference state, a positive correlation between two latent traits could be seen as association between novel amino acids in both sites. The strongest evidence for correlation was found between sites 158(E)K and 156(K)Q, with an estimated correlation coefficient of 0.74 (95% BIC of [0.40, 0.93]). Koel *et al.* (2013) identified these specific mutations in both sites as being the main drivers of major antigenic change taking place between 1995 and 1997. Mutations in sites 159 and 189 are another example of a pair of substitutions identified as driving major antigenic change taking place in the late 1980's. Even though the oldest sequence in our sample only dates back to 1995, correlation between these two sites remains strongly supported by our analysis, with an estimated correlation coefficient between 159(Y)F and 189(S)N of 0.69 (95% BIC of [0.27, 0.92]).

4. Discussion

We present the phylogenetic latent liability model as a framework for assessing phenotypic correlation between different types of data. Through our three applications, we illustrate the use of our methodology for binary data, discrete data with multiple ordered and unordered states, continuous data and combinations thereof. The applications exemplify current biological problems which our method can naturally address. Additionally, we show how the model can be used to reveal the overall phenotypic correlation structure of the data, and we provide tools to test hypotheses about individual correlations and for general model testing.

The threshold structure of the phylogenetic latent liability model renders it non-Markovian for the discrete traits. Both Felsenstein (2005, 2012) and Revell (2013) argue that this is actually a valuable property for many phenotypic traits for which the probability of transitioning between states should vary depending on the time spent at that state. Based on this argument, Revell (2013) investigates ancestral state reconstruction for univariate ordered traits under the threshold model, and finds consistent reconstructions for simulated data. For our model, it would be straightforward to perform ancestral state estimation for multivariate traits of all types considered, because the inference machinery is already implemented in BEAST.

A problem with many comparative biology methods for phenotypic correlation is the requirement for a fixed tree. Through sequence data, our model can account for the uncertainty of tree estimation by integrating over the space of phylogenetic trees, as we do for the influenza epitope and antibiotic resistance examples.

As a caveat for this type of model, Felsenstein (2012) points out a general lack of power, arguing that for realistically sized datasets confidence intervals would be too large. This issue could be magnified on discrete traits, since the correlations are an extra step removed from the data. In our applications, the size of our posterior credible intervals are relatively large for intervals constrained between −1 and 1. However, this did not prevent us from recovering general correlation patterns and identifying important correlations. Moreover, for the columbine flower example, we find no difference in average size of credible intervals for correlations including latent traits and those between two continuous traits.

Analytically integrating out continuous trait values at root and internal nodes to compute the likelihood of Brownian motion on a tree leads to significant improvement in efficiency of inference methods (Pybus *et al.*, 2012). This strategy computes successive conditional likelihoods by a post-order tree traversal in a procedure akin to Felsenstein's peeling algorithm (Felsenstein, 1981). Its effectiveness has been explored in similar contexts in univariate (Novembre and Slatkin, 2009; Blum *et al.*, 2004) and multivariate Brownian motion (Freckleton, 2012) and to estimate the Gaussian component of Lévy processes (Landis, Schraiber and Liang, 2013). A related post-order traversal approach improves computation in the context of phylogenetic regressions for some Gaussian and non-Gaussian models (Ho and Ané, 2014). Unfortunately, a similar solution is not available to marginalize the latent liability **X** at the tips of the tree in our model. Consequently this integration must be performed by MCMC. Integration for **X** is a critical part of our method, and for large

datasets, mixing becomes a problem. To address this issue, we present an efficient sampler that, at each iteration, updates all components of the multivariate latent variable **X** at one tip of the tree. This algorithm builds upon the dynamic programming strategy of Pybus *et al.* (2012) to obtain a truncated multivariate normal as the full conditional distribution of $\mathbf{X}_i$. Even though sampling from this truncated distribution requires an accept/reject step that could be highly inefficient, we find that as the chain approaches equilibrium, rejection rates become small.

Computational time for our method varies depending on the size and type of the dataset and on additional model specifications of phylogenetic inference. Our example with the shortest computational time is the columbine flower analysis, in which we used a fixed phylogenetic tree and only 2 of the traits required latent variables. This application ran at 0.02 hours per million states on a regular desktop computer, and the analysis was completed with parallel chains of 200 million states. On the other extreme, the influenza epitope analysis required the longest computational time, at 1.03 hours per million states and taking a couple of weeks to complete the analysis on independent chains. Computationally, the bottle neck in this analysis is the numerical integration over the latent traits; the analysis required a total 32 latent traits for 180 viral strains. Additionally, in this analysis, we jointly estimated the tree from sequence data.

In our analysis of influenza epitopes, we set the oldest amino acid observed for each site as the reference state, and for each of the remaining variants we assigned an entry in **X**. For the multiple unordered states model, this choice results in a reduction of dimensionality in the problem, but is done mainly to improve identifiability. However, this procedure breaks the symmetry of the model and complicates interpretability of correlations. In fact, a correlation between two entries of the latent trait **X** cannot be directly translated as a correlation between the states they represent, because variations in an entry of **X** are linked to all other states for that trait through the reference state. Despite this caveat, general statements about the correlation structure of the data can still be made based on the latent liability **X**, as we show in the influenza epitopes application.

In this context, different model choices could be used to change the interpretational links between correlations in **X** and in the data. Hadfield and Nakagawa (2010) briefly discuss a multinomial phylogenetic mixture model where a latent variable determines the probability of the multinomial outcome. They consider the common choice of constraining the latent variable to a simplex by setting the sum of its components to one. This makes the value of the latent trait immediately interpretable as probabilities, however it further complicates interpretability of the correlations. A possible alternative to address this issue is to model the evolution of **X** in the latent liability model with a central tendency such as the Ornstein-Uhlenbeck process. It remains to be investigated whether this change would improve identifiability, eliminating the need to impose constraints on the model.

Lartillot and Poujol (2011) have studied the correlation between continuous traits and parameters of the molecular evolution model, such as dS/dN ratio and mutation rate, by modelling the evolution of these parameters as a diffusion process along the tree. One possible extension to our method would be to incorporate the evolution of these parameters

in our model, allowing for the estimation of correlations between our continuous and discrete traits and these evolutionary parameters.

The Bayesian phylogenetic framework in which we integrate our model easily lends itself to combination of different models. These could be phylogenetic models for demographic inference (Minin, Bloomquist and Suchard, 2008), methods for calibrating trees or relaxed clock models (Drummond *et al.*, 2006). Additionally, we can explore the relaxed random walk (Lemey *et al.*, 2010) to get varying rates of trait evolution along different branches of the tree. The latent liability model can easily be associated with these existing models to provide comprehensive analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge the Scottish Salmonella, Shigella & C. difficile Reference Service for providing the Salmonella Typhimurium DT104 isolates and phenotypic resistance data. We thank Kenneth Lange, Christina Ramirez and Jamie Lloyd-Smith for providing constructive feedback on an earlier version of this manuscript.

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007–2013] under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864, Wellcome Trust grant 098051, National Institutes of Health grants R01 AI107034 and R01 HG006139 and National Science Foundation grants DMS 1264153 and IIS 1251151.

References

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution. 2012; 29:2157–2167. [PubMed: 22403239]

Blum MG, Damerval C, Manel S, François O. Brownian models and coalescent structures. Theoretical Population Biology. 2004; 65:249–261. [PubMed: 15066421]

Boyd D, Peters GA, Cloeckaert A, Boumedine KS, Chaslus-Dancla E, Imberechts H, Mulvey MR. Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella* enterica serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona. Journal of Bacteriology. 2001; 183:5725–5732. [PubMed: 11544236]

Breslaw J. Random sampling from a truncated multivariate normal distribution. Applied Mathematics Letters. 1994; 7:1–6.

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. Science. 1999; 286:1921–1925. [PubMed: 10583948]

Cox, NJ.; Bender, CA. Seminars in Virology. Vol. 6. Elsevier; 1995. The molecular epidemiology of influenza viruses; p. 359-370.

Damien P, Walker SG. Sampling truncated normal, beta, and gamma densities. Journal of Computational and Graphical Statistics. 2001:10.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLoS Biol. 2006; 4:e88. [PubMed: 16683862]

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 2012; 29:1969–1973. [PubMed: 22367748]

Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics. 1965; 29:51–76.

Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013:368.

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution. 1981; 17:368–376. [PubMed: 7288891]

Felsenstein J. Phylogenies and the comparative method. American Naturalist. 1985; 125:1–15.

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

Felsenstein J. A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist. 2012; 179:145–156.

Fitch WM, Leiter J, Li X, Palese P. Positive Darwinian evolution in human influenza A viruses. Proceedings of the National Academy of Sciences. 1991; 88:4270–4274.

Freckleton RP. Fast likelihood calculations for comparative analyses. Methods in Ecology and Evolution. 2012; 3:940–947.

Gelfand AE, Smith AF, Lee TM. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. Journal of the American Statistical Association. 1992; 87:523–532.

Grafen A. The phylogenetic regression. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 1989; 326:119–157. [PubMed: 2575770]

Hadfield J, Nakagawa S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology. 2010; 23:494–508. [PubMed: 20070460]

Ho LST, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology. 2014; 3:397–402. [PubMed: 24500037]

Huelsenbeck JP, Rannala B. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution. 2003; 57:1237–1247. [PubMed: 12894932]

Ives AR, Garland T. Phylogenetic logistic regression for binary dependent variables. Systematic biology. 2010; 59:9–26. [PubMed: 20525617]

Jeffreys H. Some tests of significance, treated by the theory of probability. In Proceedings of the Cambridge Philosophical Society. 1935; 31:203–222. Cambridge Univ Press.

Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, Skepner E, Lewis NS, Spronken MI, Russell CA, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science. 2013; 342:976–979. [PubMed: 24264991]

Landis MJ, Schraiber JG, Liang M. Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. Systematic Biology. 2013; 62:193–204. [PubMed: 23034385]

Lartillot N, Poujol R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution. 2011; 28:729–744. [PubMed: 20926596]

Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Molecular Biology and Evolution. 2010; 27:1877–1885. [PubMed: 20203288]

Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology. 2001; 50:913–925. [PubMed: 12116640]

Liu JS, Liang F, Wong WH. The multiple-try method and local optimization in Metropolis sampling. Journal of the American Statistical Association. 2000; 95:121–134.

Mather AE, Matthews L, Mellor DJ, Reeve R, Denwood MJ, Boerlin P, Reid-Smith RJ, Brown DJ, Coia JE, Browning LM, et al. An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations. Proceedings of the Royal Society B: Biological Sciences. 2012; 279:1630–1639. [PubMed: 22090389]

Mather A, Reid S, Maskell D, Parkhill J, Fookes M, Harris S, Brown D, Coia J, Mulvey M, Gilmour M, et al. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. Science. 2013; 341:1514–1517. [PubMed: 24030491]

Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution. 2008; 25:1459–1471. [PubMed: 18408232]

Novembre J, Slatkin M. Likelihood-based inference in isolation-by-distance models using the spatial distributions of low frequency alleles. Evolution. 2009; 63:2914–2925. [PubMed: 19624728]

Pagel M. 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. 1994; 255:37–45.

Plotkin JB, Dushoff J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proceedings of the National Academy of Sciences. 2003; 100:7152–7157.

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences. 2012; 109:15066–15071.

Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012; 3:217–223.

Revell LJ. Ancestral character estimation under the threshold model from quantitative genetics. Evolution. 2013

Robert CP. Simulation of truncated normal variables. Statistics and Computing. 1995; 5:121–125.

Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Molecular Biology and Evolution. 2001; 18:1001–1013. [PubMed: 11371589]

van der Niet T, Johnson SD. Phylogenetic evidence for pollinator-driven diversification of angiosperms. Trends in Ecology & Evolution. 2012; 27:353–361. [PubMed: 22445687]

Whittall JB, Hodges SA. Pollinator shifts drive increasingly long nectar spurs in columbine flowers. Nature. 2007; 447:706–709. [PubMed: 17554306]

Whittall JB, Voelckel C, Kliebenstein DJ, Hodges SA. Convergence, constraint and the role of gene expression during adaptive radiation: floral anthocyanins in Aquilegia. Molecular Ecology. 2006; 15:4645–4657. [PubMed: 17107490]

Wright S. An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics. 1934; 19:506. [PubMed: 17246735]

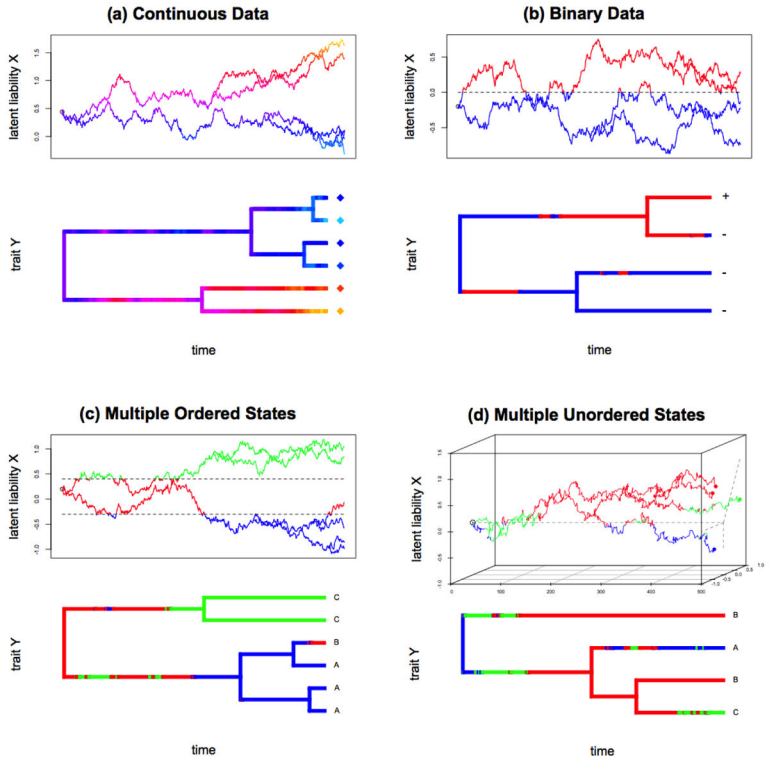

Fig 1.

Realizations of the evolution of latent liabilities **X** and observed trait **Y** for different types of data. Both tree and Brownian motion plots are color coded according to the trait **Y**. Realization **(a)** represents a continuous trait, **(b)** represents discrete binary data, **(c)** represents discrete data with multiple ordered states, and **(d)** represents discrete data with multiple unordered states, for which the latent liabilities X is multivariate. ** This figure was created using code modified from R package phylotools (Revell, 2012).

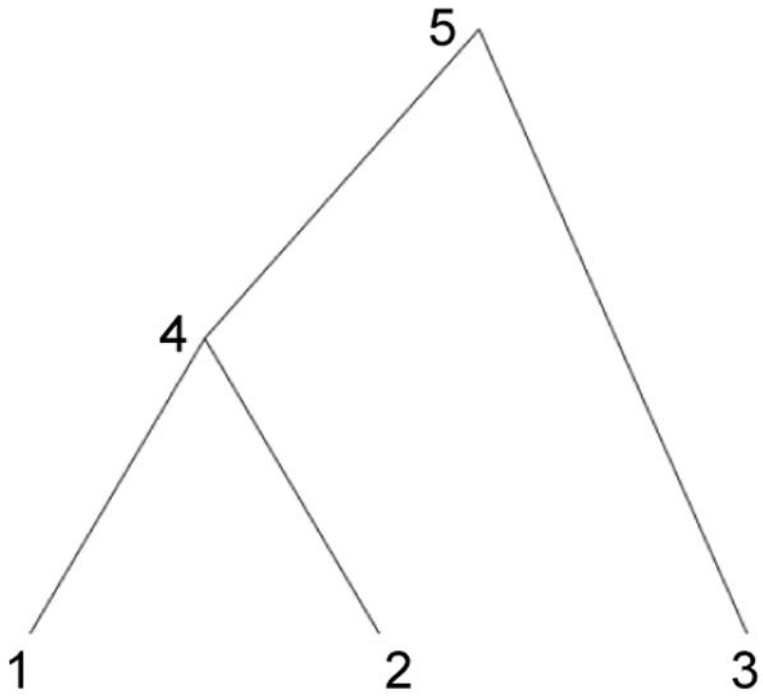

Fig 2.

Example $N=3$ tree to illustrate pre-and post-order traversals for efficient sampling of latent liabilities $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)^t$.

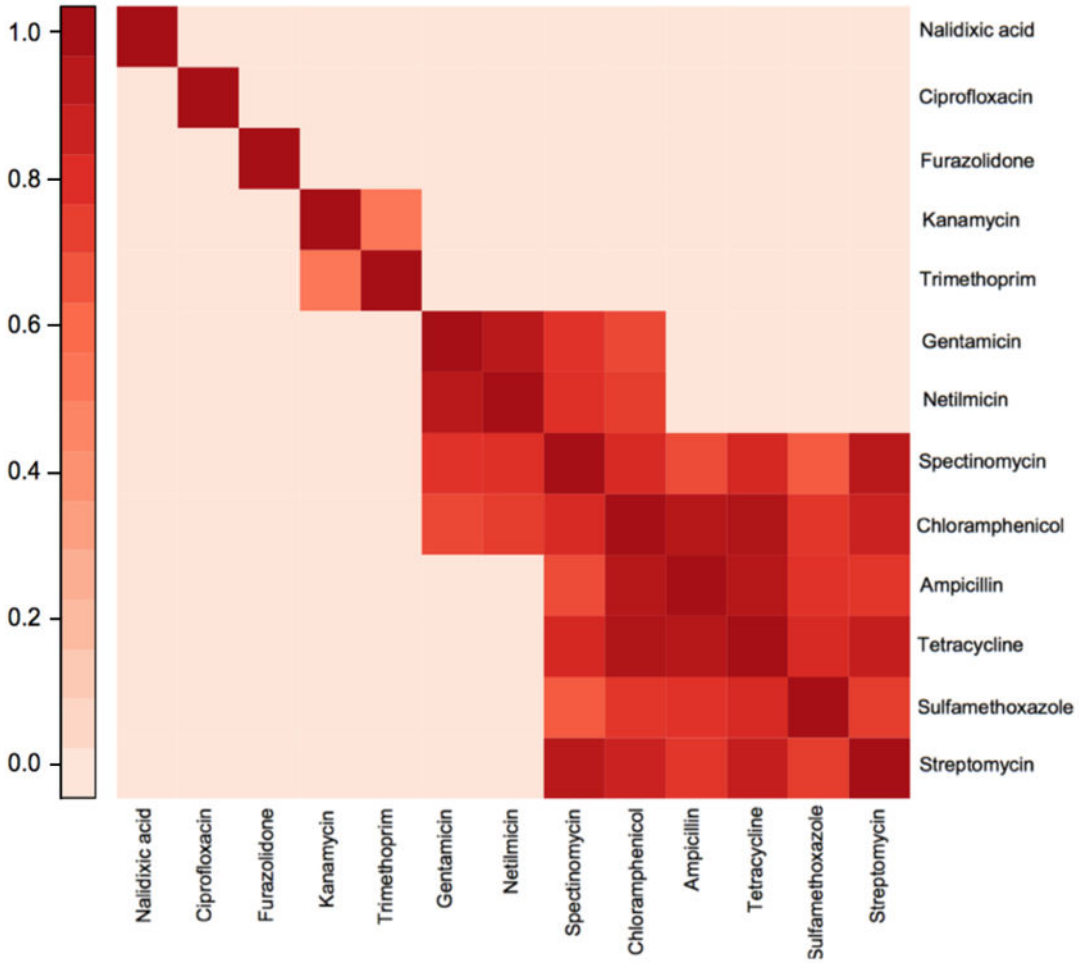

Fig 3.

Heatmap of posterior means for significantly non-zero correlations between antibiotic resistance traits for the latent liability model. Darker colors indicate stronger positive correlation.

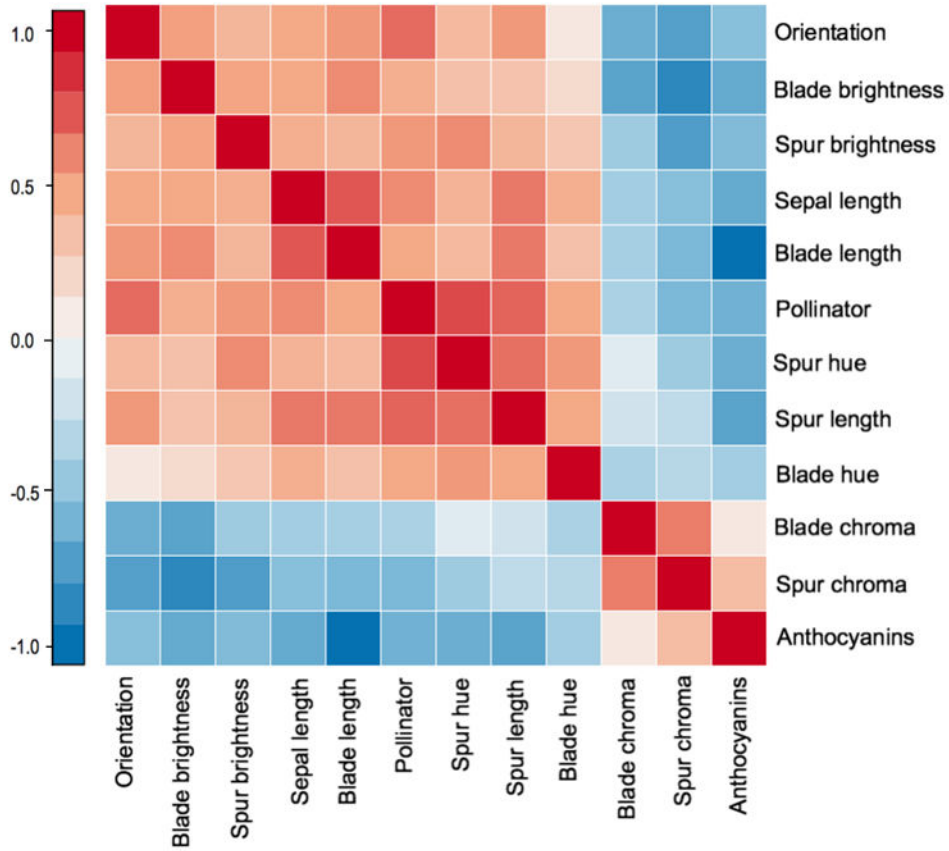

Fig 4.

Heatmap of the posterior mean for the phenotypic correlation of columbine floral traits in the latent liability model. Darker colors indicate stronger correlations; shades of red for positive correlation and blue for negative correlation.

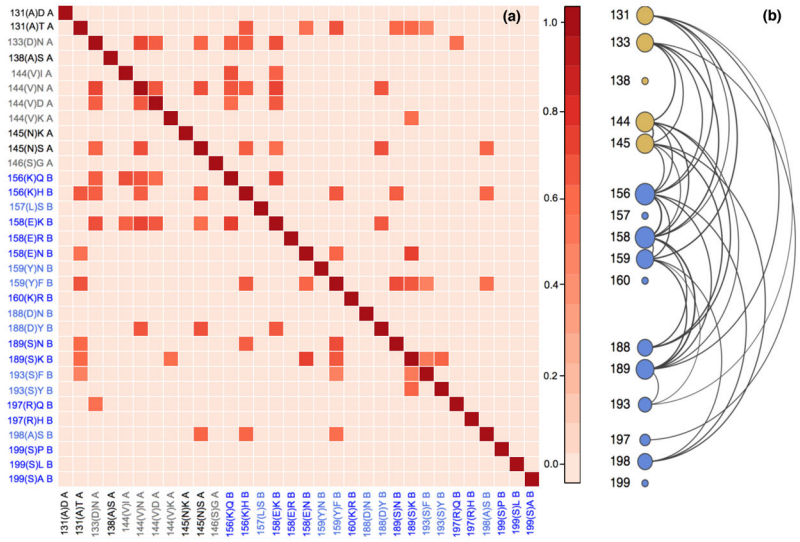

Fig 5.

(a) Heatmap of the posterior mean for the non-zero phenotypic correlation of amino acids in H3N2 epitopes A and B in the latent liability model. Darker colors indicate stronger correlation. We list the sites as follows: the number of the amino acid site in the aligned sequence; the one letter code for the reference amino acid for the site, in parentheses; the code for the amino acid corresponding to the latent trait; and the epitope to which the site belongs.(b) Network representation of the correlation structure of antigenic sites. Yellow nodes represent sites from epitope A, and blue ones from epitope B. Edges represent significant correlations, edge thickness represent correlation coefficient, and node sizes are proportional to network centrality.

Table 1

Model selection for the ordering of bumblebee (Bb), hummingbird (Hb) and hawkmoth (Hm) pollinators in Columbine flowers.

		log Bayes Factor		
Order	**log Marginal Likelihood**	**Hm-Bb-Hb**	**Hb-Hm-Bb**	**unordered**
Bb-Hb-Hm	−11.2	9.4	14.2	24.8
Hm-Bb-Hb	−20.6	-	4.8	15.3
Hb-Hm-Bb	−25.4	-	-	10.5
unordered	−36.0	-	-	-